Personal Imaging

by

Steve Mann

B.Sc. (physics) , McMaster University (1986)
B.Eng. (electrical), McMaster University (1989)
M.Eng. (electrical), McMaster University (1991)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

Author .....
the Program in Media Arts and Sciences, School of Architecture and Planning
August 8, 1997

Certified by .....
Rosalind W. Picard
NEC Development Professor of Computers and Communications
Thesis Advisor

Accepted by .....
Stephen A. Benton
Allen Professor of Media Arts and Sciences, Chair, Departmental Committee on Graduate Studies

Personal Imaging

by

Steve Mann

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
on August 8, 1997, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, I propose a new synergy between humans and computers, called "Humanistic Intelligence" (HI), and provide a precise definition of this new form of human-computer interaction.

I then present a means and apparatus for reducing this principle to practice. The bulk of this thesis concentrates on a specific embodiment of this invention, called Personal Imaging, most notably, a system which I show attains new levels of creativity in photography, defines a new genre of documentary video, and goes beyond digital photography/video to define a new renaissance in imaging, based on simple principles of projective geometry combined with linearity and superposition properties of light.

I first present a mathematical theory of imaging which allows the apparatus to measure, to within a single unknown constant, the quantity of light arriving from each direction, to a fixed point in space, using a collection of images taken from a sensor array having a possibly unknown nonlinearity. Within the context of personal imaging, this theory is a contribution in and of itself (in the sense that it was an unsolved problem previously), but when also combined with the proposed apparatus, it allows one to construct environment maps by simply looking around.

I then present a new form of connected humanistic intelligence in which individuals can communicate, across boundaries of time and space, using shared environment maps, and the resulting computer-mediated reality that arises out of long-term adaptation in a personal imaging environment.

Finally, I present a new philosophical framework for cultural criticism which arises out of a new concept called 'humanistic property'. This new philosophical framework has two axes, a 'reflectionist' axis and a 'diffusionist' axis. In particular, I apply the new framework to personal imaging, thus completing a body of work that lies at the intersection of art, science, and technology.

Thesis Advisor: Rosalind W. Picard
Title: NEC Development Professor of Computers and Communications

This page intentionally left blank.

Doctoral Dissertation Committee

R. W. Picard .....
NEC Development Professor of Computers and Communications

Thesis Advisor

Berthold K.P. Horn .....
Professor

Thesis Reader

Marvin Minsky .....
Toshiba Professor Of Media Arts And Sciences

Thesis Reader

This page intentionally left blank.

Acknowledgments

The direction taken in this dissertation, unlike that of a typical PhD, is something that evolved from a personal hobby in new forms of photographic exploration. Indeed, when I applied to MIT, part of my application was a portfolio of work — pictures I had generated from my apparatus, as well as pictures of my wearable computer and personal imaging "rig" — in short, a portfolio describing what I wanted to do for a PhD dissertation.

Many professors would generally wish their students to further their own research agenda. However, something remarkably different about my thesis advisor, Rosalind W. Picard, was her willingness to allow me to continue to explore my own personal interest — that which I call 'personal imaging'. The degree of latitude that she has extended to me, and the rest of her students is something that can only be described as truly altruistic. I would also like to express my thanks to Rosalind for a tremendous amount of interaction, support, encouragement, and general input, pertaining to this thesis as well as all other aspects of my life throughout my studies at MIT. Rosalind has also been a tremendous role model.

I also wish to thank the other members of my thesis committee: B.K.P. Horn and Marvin Minsky who have offered much advice and constructive criticism which has shaped my thinking and improved my work.

My parents, mostly through RTTY (radioteletype with email, talk, www, etc.) offered me patience, support, encouragement, and love, as well as cyranic advice from and into my visual field of view, without which life's stress may well have made this thesis impossible. I thank my brother, Richard for his assistance in much of the debugging of systems (such as the early pushbroom "duster" softwear, font tables for it, etc.), and, more recently, advice from afar. The title of, and general layout of this thesis was his suggestion or at the very least arose out of a discussion with him. He has also contributed extensively to WearComp2, and somewhat to WearComp7 and WearComp8. I thank my "rig" for keeping us all together, at least "virtually". My sister, Beth, provided much in the way of artistic direction in the early days of my "dusting" efforts, thus contributing to the artistic vision of the lightspace concept.

Thanks also to my extended family. My grandparents taught me, as a young child, the skills I needed to accomplish the building of the many prototypes of my various inventions. My grandfather taught me how to weld and work with sheet metal, as well as all the skills of the machine shop, and my grandmother taught me how to knit and sew so that I could make my own smarter clothes, as well as modify existing clothing.

My wife, Betty, for 14 years of "5, 4, 3, 2, 1, didn't sync/lost packet", thanks for being my "other half" and still not getting "dusted" away and for helping me catch each packet of happiness, however brief. Betty also both funded and contributed to the design of WearComp7 and WearComp8.

Many thanks are due also to other members of the MIT community, including Charles Wyckoff, Kim Vandiver, and Bill McRoberts of the Edgerton Center who helped sustain my fascination with electronic flash, photography, and the like.

Olivier Faugeras, through his course, and our after-class discussions had much to say toward helping me decide what to focus on (and his associate, Quang-Tuan Luong, for suggesting that the *noniterative* nature of my proposed video orbits work be emphasized more strongly).

Michael Artin instilled in me a love of algebra, Lie groups, and tolerated my desire to apply exact science (algebra) to the inexact world. Gilbert Strang, instilled his love of linear algebra, Victor Guilleman, his love of harmonic analysis, and Irving Segal, his love of metaplectomorphisms of the position-momentum (or time-frequency) space which give rise to what I call "chirplets".

Hiroshi Ishii helped keep me on a scholarly track and helped me in many other ways. Ted Adelson sustained my natural scientific curiosity.

Shawn Becker and Kenneth Russell both had the uncanny ability to sift through thousands of lines of old code I'd poorly written, some translated from FORTRAN, and zero in on the bugs. Shawn, Ken, and Chris Graczyk were all instrumental in helping bring Video Orbits from a mess into release 1.0. Kris Popat (who instilled in me the notion that ideals are more important than ideas and that having an internal "compass" is more important than following rules), Arno Klein (and Arnold Klein who I hope doesn't have his cancelled prints show until many years to come), Dave

Tames, Obed Torres, Warren Sack, John Wang, Nassir Navab, Ujjaval Desai, Chris Graczyk, Walter Bender, Fang Liu, have all contributed something, and will no doubt continue to stay in touch. I also feel I must thank Krzysztof Wodiczko for asking me the question "what does it mean" when he looked over some of my early lightpaintings. In particular, Wodiczko's existential, interrogative, and situationist view of my art helped immensely. Thanks also to Ron MacNeil for suggesting that my 'lightspace' formulation was a new language, but that I really should be focusing on trying to "write some poetry in that language". Leila Kinney provided a context within the history, theory, and criticism of art. Thanks to Chuck Oman for pointing out references [1] and [2]. Thad Starner and Flavia Sparacino helped me get the cursor-control software running with X-windows on the SGI Reality Engine for the passive version of finger-tracking. Jeremy Levitan offered much time and help in using the 3-D printer and 3-D CAD tools, as well as expertise in the machine shop.

Joe Paradiso suggested I document my early wearable computing efforts in an 'experiential first-person account' (which formed my IEEE Computer article, which then became incorporated into this thesis).

Thanks to Lee "elwin" Campbell for use of his lens distortion correction software. Thanks also to elwin, Rosalind, Dan Gruhl, Jennifer Healey, Bill and Ruth Mann, Betty, Chris, Tom Minka, and the many others who helped with getting the condenser banks up to the roof of building 54 to power my flashlamp, and for help in getting this experiment set up (example of homomorphic imaging). Thanks also to the many others, among them Richard Fletcher and Flavia, who helped with various pictures.

Those at the List Visual Arts Center, such as Katherine Klein, Jennifer Riddell and Jonathan Roll, through helping me with my upcoming show, have provided much in the way of meaningful discussion of my work, thus contributing to this thesis.

Patricia Flanagan, and the other librarians, Linda Martinez, and Rae Jean Wiggins, provided much input on the humanistic aspects of personal imaging. Likewise, those with MIT Press: Ed Barrett, Robert Prior, Jeremy Grainger, and Scott Mcginley provided much in the way of thoughts, ideas, situationist connections, and an overall perspective that contributed to this work.

Matt Reynolds, KB2ACE helped me upgrade my outbound ATV channel.

And 73 to N4RVE and W1GSL; thanks for many useful suggestions about communications, and hope that we continue to CQ.

Thanks to Eric Paulos and John Canny for pointing out reference [3], and making some other important connections between my work and the work of others.

Joseph Segman provided many useful connections with regard to incorporating scale into the Weyl-Heisenberg group, which led to a better understanding of the chirplet transform and the background upon which Video Orbits was based, as well as an appreciation of symplectic techniques in optical design. Thanks also to David Mumford and Alan Yuille, for useful discussion, and for inviting me to lecture at Harvard where I met Joseph.

Martin Bichsel engaged in much discussion about my "lightspace" theory and some similar ideas he had.

Adam Oranchak provided his expertise in industrial design and sculpture toward making a better fit for many of my new rigs as well as updating the fit on a good number of my old rigs that I had since outgrown. Oranchak also provided the "third hemisphere" metaphor that I often now use. Thanks to David Brin for a long discussion at dinner one night, which re-enforced some aspects of my 'diffusionist' philosophy. Also his assertion that "hobbyists will take over the world", although perhaps exaggerated for effect, captures so well the spirit of what I believe will be the next generation of human endeavour facilitated through readily available worldwide communications (the Linux operating system being a prime example of this).

Peter Anders pointed out further important connections to art history. Julia Scher (who invited me to lecture at Mass College of Art, where the class discussion afterwards significantly affected the outcome of Chapters 6-9 of this thesis), Liz Canner, and many others in the art community, as well as many folks at MIT such as Mitch Resnick and Sherry Turkle (who each invited me to lecture, where the class discussion afterwards significantly helped shape the content of Chapters 6-9) contributed much toward the artistic vision of this thesis. They, as well as Freedom Baird, Christine Southworth, and Brian Bradley helped greatly by providing constructive criticism of my

documentary video work, which contributed toward my attempt to define the 'personal verité' genre presented in chapter 7.

Thanks to the many others who either invited me to lecture on personal imaging, or provided much in the way of useful feedback after such lectures. Those such as Alan Siegel, Randy Pausch, and Daniel Shurman who have provided honest, candid constructive criticism must certainly be acknowledged, for this has invariably made the work reported in this thesis that much stronger. Thanks in advance, to those, such as Safwat Zaky and Ron Baecker, who have already begun to take a mentoring role and provide advice in terms of future directions.

Thanks to Alex Drukarev and Jeanne Wiseman at HP Labs, Palo Alto for asking that I take a summer off from my studies and work with them on an exciting imaging effort and Paul and Claire Hubel of HP Labs for the use of the "Claire" image sequence.

Dr. Simon Haykin, my M. Eng advisor, who inspired my love of "Radar Vision", Dr. Carter who helped me interface my lightpainting pushbroom to my 6502 wearable, and Kent Nickerson who helped with some of my miniature personal radar units, all three from McMaster University, each had a significant impact on this work. Much of the early work on biosensors and wearable computing was done with, or at least inspired by work I did with Dr. Ghista, and later refined with input from Dr. DeBruin, both of McMaster University. Dr. Max Wong of McMaster university supervised my undergrad project designing an RF link between two 8085 wearable computers which I had assembled for my "photographer's assistant" project.

Thanks to Ron Lancaster, for his enthusiasm in math class, helping keep me out of (and into) trouble, for donating lots of "goodies" like that seemingly bottomless box of 6 conductor phone wires.

Grace, Eleni, and all the others who lasted past the first "dust", thanks for not being "dusted" away by me.

Thanks to Antonin Kimla who gave me the stepping relays to build my first wearable lightpainting computer, and later the funding to do it right (with solid state components).

Jeff Eleveld and Graham Lovell were both of much help and inspiration in the early days of my "wearables", wireless, and "dusting" efforts.

Most recently, I wish to thank the many anonymous reviewers of my publications, as well as various "shepherds" (volunteers, working to assist in author's preparation of paper submissions). Most notably, the guidance and advice from Steve Feiner, Don Norman, and Carole Goble not only helped toward producing better papers, but also carried over into significantly shaping and improving this dissertation.

Many useful comments have come from the thousands of people I have met, in my day-to-day interactions, through the apparatus — either face-to-face (on the street, etc.), or through the net of which my body is, in some ways, a part. These people — too numerous to mention or even identify — have responded with comments ranging from harsh criticism to insightful ideas.

HP labs supported me for six long years during my doctoral work without complaint. BT also provided some support. The artistic portions of this work were funded, in part, through the Council for the Arts at MIT. Kodak provided me with a high-resolution digital camera to tear apart and build into one of my WearComp/WearCam rigs. BelTronics donated 24.360GHz microwave components for my wearable BlindVision project, and C-K and M/A-Com both donated other microwave components.

Thanks to Larry Smarr of University of Illinois for use of the NCSA Supercomputing facility, to Chris Barnhart for special-purpose processing hardware, to Bran Ferren of Disney for the FT-623, to Wyckoff for photometric instrumentation, to Shurman for help with the X-frame pack version of the personal imaging system, to Robert Kinney and Rich Landry for help in recently updating my ThinkTank/VibraVest apparatus, to HP labs, Thought Technologies Inc., VirtualVision, Compaq, Kopin, ViA, Ed Gritz, Miyota, Virtual Research, and Sony for lending or donating additional equipment that made my experiments possible and quite literally made the completion of this thesis possible, for this is probably the first thesis dissertation to by typed primarily on a wearable computer system (done on WearComp5, WearComp6, and WearComp7). The day before this thesis was due, when there was a power failure throughout most of the city for most of the evening, the ample supply of charged batteries proved essential to its timely completion.

Thanks to University of Toronto for special arrangements to take the self-portrait (through Betty's eyes) of Fig 4-1, where photography is normally strictly prohibited.

Finally, I thank God for letting there be light, in the wide sense, including those portions of the electromagnetic wave spectrum that have kept me in touch with my loved ones.

Contents

Chapter 1

What we'll be

1.1 Overview of thesis, its contributions, and their philosophical context

The contribution of the thesis spans three broad areas:

1. Art and philosophy
2. Theory (mathematical and scientific underpinnings)
3. Applications (technological contributions in the form of various inventions).

1.1.1 Overview of thesis

This thesis has two main parts:

- PART A: Scientific and mathematical underpinnings, which are addressed in chapters 2,3,4,5,6.
- PART B: Artistic and philosophical considerations which are addressed in chapters 6,7,8,9.

Chapter 6 is at the intersection of both parts, and is where contributions in the "softer" science of human-computer interaction, which is as much an art as it is a science, are presented. Engineering contributions (technology, applications, and inventions) transcend the boundary between these two parts, and are thus presented throughout the entire thesis.

A body of work that lies at the intersection of ART, SCIENCE, and TECHNOLOGY would seem most appropriate for a degree in Media **Arts** and **Sciences** at the Massachusetts Institute of **Technology**.

Indeed, it is my belief that we are entering a pivotal era in which disciplinary boundaries are beginning to fall, as tools of connectivity like the Internet break down barriers of communication between what were once disparate and specialized groups.

In some ways artists and scientists have always shared a similar desire to find a kind of inner truth, often free of the constraints of the engineer who is often more fettered (e.g. required to make something work) than is the artist or scientist. The scientist seeks an objective inner truth — answers to basic questions of what is, while the artist seeks answers to a more subjective inner truth. The scientist finds beauty in truth, and the artist finds truth in beauty.

However, engineers have often misunderstood, or even sometimes had contempt for artists[1], or at the very least, the two groups have had little in the way of interaction, while at the same time

[1] This is not to say that *all* engineers hate artists — many of them work with artists or are artists themselves, but this disdain is a general observation that has even been sustained, for example, through ritualized mockery of art and "artsies" in the socialization process of engineering freshmen. Much of the socializing process of becoming an engineer involves its own mechanism for transmission of culture, in the tradition of "Lady Godiva" and "Cold Iron" [4]. In some traditions, such as Carnegie Mellon University, the engineers refer to the artists as the "fruits", and the artists counter this attack by referring to the engineers as the "vegetables".

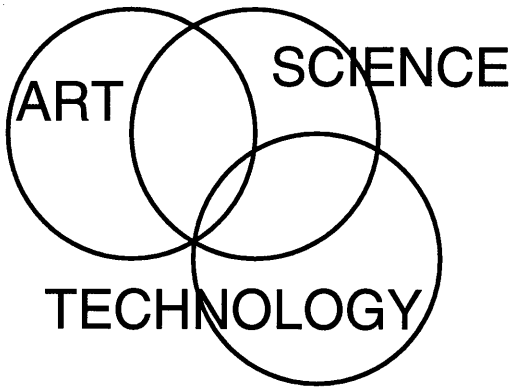


Figure 1-1: Artists and scientists seek a common goal of basic inner truth — the scientist seeks an objective truth, while the artist seeks a more subjective truth. Engineers have often misunderstood and therefore failed to respect or appreciate artists, although engineers have tolerated, and in many cases eagerly collaborated with scientists.

engineers and scientists often worked together, or at the very least, had mutual respect or tolerance for one another in a common practical and objective set of goals. This situation is depicted in Fig 1-1. This disparity was not always present. In the era of the Renaissance, art, science, and technology were inextricably intertwined.

Recently a number of organizations have appeared, many on the Internet, in support of individuals at the intersection of art, science, and technology. The appearance of the journal "Leonardo", named in honor of daVinci — artist, scientist, and inventor — perhaps best captures the essence of this new synergy. Prestigious professional conferences such as Ars Electronica and ISEA (International Symposium on Electronic Art) mark the beginning of a new era — a 'new renaissance' where those who do not wish to be categorized as artist, or as scientist, or as engineer, can have a voice in a world that once turned them to silence.

As we enter a new millenium, we, through enhanced capability to acquire and transfer information, will witness a great wealth of knowledge and richness of thought that arises from new synergies happening at the boundaries between what were once disparate fields of study.

Therefore, one of my goals in writing this thesis is to provide, by way of example, a body of work where the boundaries created by individual disciplines of study, which are often erected in the interests of empire-building rather than a true quest for knowledge, do not apply.

The organization of this thesis, by chapters, into this 'new renaissance' of thinking, is depicted in Fig 1-2.

1.1.2 Specific contributions of this thesis

"PART A" of this thesis, comprised of Chapters 2, 3, 4, 5, (6), proposes a new conceptual/mathematical theory called 'homomorphic imaging', — the general philosophy that a camera may (and in fact should) be regarded as an array of photometric measuring instruments, and that likewise a light source may similarly be regarded as the inverse of a photometric/homomorphic camera.

My theory of how one may characterize how scenes or objects respond to light, which I present in Chapter 2, is based on the very simple observation, made by daVinci, that pertains to bundles of light rays, linearity, and superposition.

While the theory I put forth in Chapter 2, called "Lightspace", is too unwieldy to implement in practice, certain special cases of it are of great practical utility. In Chapter 3, I put forth a special case of the lightspace theory, namely how one may use an ordinary camera (with unknown nonlinear response function) as a photometric measuring instrument. I name this theory the 'Wyckoff principle' in honor of Charles Wyckoff, who developed the extended response photographic film that provided me the inspiration for this work. The Wyckoff principle sets forth a framework for self-calibration of the camera into an array of light-measuring instruments, based on analyzing pictures that differ only in exposure. Furthermore, it also provides a means of combining differently exposed images to extend

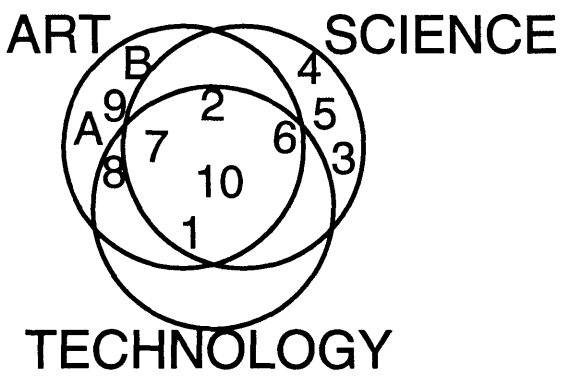


Figure 1-2: We are at a pivotal era, where new technologies are so dramatically enhancing our ability to acquire and disseminate information, that the boundaries between specialized fields of study are beginning to crumble. The World Wide Web, with its associative gestalt "memory" is fostering a new connected humanistic intelligence that will give rise to a new renaissance in which we will no longer need to consider ourselves typecast into a particular specialty. A goal of this thesis is to present a body of work that lies at the intersection of art, science, and technology. The chapter numbers are indicated in roughly which one or more of these three areas each chapter lies.

the dynamic range of this measurement process. Even if we are not interested in photometry, but merely would like a beautiful photograph, the Wyckoff principle affords us with means and apparatus for providing a picture of much richer tonal range and much greater dynamic range than previously possible. Furthermore, the Wyckoff principle defines a new form of homomorphic filtering which affords us with far greater creative and artistic capability.

Once we, through self-calibration, transform the camera into an array of directional light meters (what photographers call "spotmeters"[2]) pointing in different directions, but all measuring rays of light passing through the same point in space, then we are prepared for some new results in "pencigraphic imaging" — new direct featureless methods of estimating the projective coordinate transformation relating multiple pictures of the same scene or object. These new results, presented in Chapter 4, are based on a general philosophy that I call "Video Orbits". The Video Orbits theory asserts that in certain special but important cases, images exist within the same orbit of a group of coordinate transformations, and that we can therefore use this group structure to define and propagate relationships across long cascades of image sequences. Most notably, in video, the coordinate transformations from frame to frame are in the neighbourhood of the identity (e.g. the Lie algebra of the continuous group of coordinate transformations), but by constantly relating this infinitesimal orbit to the true group structure, a means of assembling very large wide-sweeping panoramas and environment maps was developed. Furthermore, a repetitive (but non-iterative, in the sense that no sophisticated nonlinear optimization strategy is needed) variant of Video Orbits allows for larger jumps between frames.

Chapter 5 combines the work of Chapter 3 (the Wyckoff principle) with Video Orbits, giving rise to a what I call the 'projectivity+gain' group of coordinate transformations. In particular, Chapter 5 generalizes the concept of estimation of the parameters of a Lie group of coordinate transformations to include transformations in both the domain and range of images, where we regard the light falling on the image sensor as a real-valued function of one real variable, q, the quantity of light. In particular, given that we measure $f(q)$ in one picture, and $f(kq((Ax+b)/(cx+d)))$ in another picture, the new result of Chapter 5 allows us to estimate, simultaneously, both a homography of the plane as well as a homomorphic gain change, that is, it allows us to estimate the parameters k, A/d, b/d, and c/d.

Chapter 6 presents a new form of real-world visual interaction which I call 'mediated reality'. Mediated reality affords us with the ability to augment, diminish, or otherwise alter our visual perception of the world, and to do so in our day-to-day lives (while shopping, standing in line at

[2]Throughout this thesis, I regularly use the metaphors of photography, sometimes in the traditional sense of film, as a point of departure from which my new theories evolve.

the bank, riding the bus, etc), not just in a lab or other special environment. Mediated reality also suggests new forms of "interaction" with everyday objects — a reality metaphor of sorts.

Chapter 6 sits upon the boundary between "PART A" and "PART B" of this thesis, as it provides the link from the conceptual/mathematical contributions of Chapters 2, 3, 4, and 5 to the artistic and philosophical contributions of Chapters 7, 8, and 9. Chapter 6 presents a wide variety of various embodiments of my 'WearCam' invention.

Mediated reality also affords us with the ability to allow others to alter our visual perception of reality, and therefore gives rise to a new form of communication and social interaction. I have made many new discoveries through the exploration of mediated reality, for example, I observe that coordinate transformations in the neighbourhood of the identity have more lasting relative aftereffects and flashbacks, and that the aftereffect and flashbacks are reversed rather than merely incapacitating in a diffuse and inexplicable sense.

These explorations in altered perception of reality gave rise to a new kind of documentary video process. Firstly, because of some of the coordinate transformations, such as a visual world rotated 90 degrees, I discovered a new style of shooting, most notably, in face-to-face social interactions. This gave rise to a very closely cropped face shot, sustained over the conversation. Secondly, because I am "living" in the 2-D world of the documentary, as I am creating it, a unique cinematographic style arose, in which the camera began to function as a true extension of my mind and body, and served as a device to record my exact visual experience, no more and no less.

With a lighter-weight version of WearCam, comprising only a single processing channel (one camera, one display, and one video processing "engine"), I also discovered a "photographic mindset" through long-term adaptation. First I noticed that I started to lose my ability to perceive 3d depth, and began to see the world as two dimensional. This effect seemed to persist even when I removed the apparatus, and would revisit me in the form of 2d "flashbacks", so that I began to see the world in two ways, much like we see the Necker cube illusion in two possible ways. This discovery (which also contributed toward my attempt at defining a new genre of documentary video) is described in Chapter 7.

This loss of depth perception, in and of itself, gave rise to some important discoveries, such as the "fingermouse" — using the finger as a pointing device, as also described in Chapter 7.

Many social and philosophical considerations also arise from this work, and are presented in Chapters 7, 8, and 9.

The art+philosophical contribution

> The purpose of art is to lay bare the questions which have been hidden by the answers.
>
> –James Baldwin

> Computers are useless. They only give answers.
>
> –Pablo Picasso

In "PART B", the artistic contribution, I have tried to emphasize the "fine arts", at their highest scholarly level, as opposed to commercial art, graphic arts, or art whose goal is to entertain.

It is my hope that by leaving the artistic and philosophical contribution to the end (by having it in "PART B") that, by the very nature of art (e.g. as interrogative rather than conclusive), the thesis will close with the raising of many important questions upon which others will build.

1.1.3 Summary of specific contributions of this thesis

1. Humanistic Inteligence, a new philosophical framework for computing. HI, which forms a broad umbrella under which personal imaging lies, is defined in this chapter (Chapter 1), and used in Chapters 6, 7, and 8 (especially in 8).

2. WearComp: means and apparatus upon which HI is realized. Presented in Chapters 1 and 6. Also used in Chapters 7, 8, and 9.

3. Smart clothing: A possible new framework for WearComp that's truly wearable. Presented in Chapter 1.

4. Personal imaging (both a general framework, as well as the apparatus upon which it is realized — the latter being a special case of WearComp). Presented in Chapters 1 and 6. Also used in Chapters 7, 8, and 9.

5. Lightspace (theoretical framework for personal imaging) Presented in Chapter 2; further developed and used in the remaining chapters.

6. Homomorphic Imaging and the Wyckoff principle, presented in Chapter 3.

7. Video orbits: conceptual (mathematical) framework for personal imaging, presented in Chapter 4.

8. Joint estimation of homography and homomorphic gain (parameters of the 'projectivity+gain' group). Presented in Chapter 5.

9. Tetherless augmented reality, with means and apparatus, presented in Chapter 6.

10. Mediated reality, which is also tetherless, as well as means and apparatus to implement it, presented in Chapter 6. Mediated reality is an important conceptual and interactional framework for personal imaging.

11. Towards defining a new genre of visual art based on personal imaging, ingredients of which are presented throughout all chapters.

12. Towards defining a new genre of documentary video based on personal imaging, presented in Chapter 7.

13. The 'WearComp' and 'WearCam' inventions are described from an idealist's perspective, in the context of the balance between individual intelligence and collective intelligence (with an emphasis on personal video versus video surveillance), in Chapter 8.

14. Reflectionism: a new philosophical construct for cultural criticism and interrogative art, with emphasis on video surveillance and matters related to personal imaging. Presented in Chapter 9.

15. Diffusionism: a new philosophical construct for cultural criticism and interrogative art, with emphasis on video surveillance and matters related to personal imaging. Presented in Chapter 9.

1.2 What we'll be

In what remains of this chapter, I provide a brief overview of some of the more new and exciting aspects of this work. These pertain to a new synergy between human and machine, where we "become one with the machine" — perhaps one might go so far as to say that we become the machine. Thus the rest of this chapter is hopefully a glimpse into the future of "what we'll be".

1.3 General disclaimer

In the following section, and throughout this thesis, I will be describing a new synergy between human and machine. **While I have made effort to ensure the accuracy and workability of the material contained in this thesis, I shall, under no circumstances, be liable for incidental or consequential damages or related expenses resulting from the use of this information. If errors or omissions are found, please notify me.**

1.4 'WearComp', a first step toward personal imaging

I will provide a brief introduction to a particular form of computational apparatus I call 'WearComp' — an alternative to today's laptop computers and PDAs.

That the WearComp project resulted in a form of computation quite different from laptops and PDAs owes much to the fact that when I envisioned and developed it, originally as a hobbyist, there was no such thing as a laptop computer or a PDA. Thus 'WearComp' evolved on a completely different path, with no pre-conceived notion of what portable computing should be.

To understand the entire motivation behind 'WearComp', one must first consider the rationale for it — for the WearComp project at first seemed like a ridiculous notion, in and of itself. Thus I defer explanation of WearComp, until I first justify it by summarizing its philosophical underpinnings in the following section.

1.5 Steps toward Humanistic Intelligence (HI)

I define a new synergy between humans and computers. It is called "Humanistic Intelligence" (HI), and is characterized by three aspects:

1. A goal of HI is to create a machine that dramatically assists the human, through a form of synergy, the design of which recognizes the strengths and weaknesses of each, so that the two function in a complimentary rather than competitive way. (This first goal of HI is closely related to so-called "intelligence amplification".) This assistance need not be attained strictly from the machine itself, but may also be attained from the intelligence of one or more other human beings, through the facility of the machine. For example, in the original photographer's assistant application, there was typically another human providing information over the wireless communications network[3]. The way that this assistance is inextricably intertwined with the user is, most notably, of very short latency such that it appears to be an extension of one's own capability. This short latency itself has two facets, which I illustrate by way of example, using the way that the human mind communicates with its peripherals (parts of the body):

   (a) Because of the "constancy of user-interface" our brain has to parts of our own body, we have adapted, over many years, to experience these peripherals as very immediate. This user-interface is not "user-friendly" in the traditional MacIntosh sense, but, rather, it takes many years to learn. Instead it is "user-friendly" in a different sense, that is, it is consistent, so that one need expend very little mental energy or mental delay to use it, although this immediacy only develops after a period of many years of use. I call this subcriterion 'first brain ephemeral'.

   (b) We do not experience or perceive delays when our brain issues commands to parts of our own bodies. We do not perceive that parts of our own bodies have a "mind of their own" (e.g. are held-up "waiting for I/O"). In the context of this thesis, I will propose that the apparatus be thought of as a 'second brain', so I will call this subcriterion 'second brain ephemeral'.

   Thus because of both constancy of user-interface, and through it's temporal immediacy/responsiveness, the machine appears as a true extension of the user's mind and body. I refer to this criterion as the '**ephemeral** criterion'.

[3] Much of what we will see later pertains to new forms of communication between humans, facilitated by the machine. These new forms of 'humanistic intelligence' are related to the principle of the so-called cyranoid [3], but are also quite different. A person talking to someone equipped with the form of humanistic intelligence I propose in this thesis experiences a mixture of personalities and opinions (that of the person in their immediate vicinity mixed with that of the one or more remote humans), rather than just the personality of one remote human as is the case with a cyranoid [3].

2. Physically, the human and machine are inextricably intertwined, to seamlessly fit together into a single unit, in order to meet the ephemeral criterion stated above. This inextricable intertwining has two purposes, the first social, and the second personal.

   (a) The social aspect is that others would not perceive the machine as a separate entity. This means, for example, that if one enters a department store or the like, where one is typically asked to leave one's bag or briefcase at the counter, that the apparatus should be so-designed that one is not required to leave behind one's 'second brain' at the counter, for this would impair constancy of user-interface (e.g. the ephemeral criterion). This sub criterion may be achieved either by making the apparatus covert, or by situating the apparatus within our *prosthetic territory* [5]. I refer to this sub criterion as the 'social eudaemonic criterion'.

   (b) The personal aspect also pertains to this long-term adaptation. If we ourselves regard the apparatus as part of our own day-to-day lifestyle — part of our own existence, then we will begin to treat it as such, and think of it as part of ourselves, which also involves an altering of human perception. This sub criterion may be achieved by making the device *comfortable*. I refer to this sub criterion as the 'personal eudaemonic criterion'.

   I refer to this criterion as the '**eudaemonic** criterion'.

3. The apparatus "empowers" the user (e.g. puts the user in control). By this, I mean that the user and his/her intellect are in the feedback loop of the important high-level processes of the combined (human and machine) intelligence. A very simple example of user-empowerment is an automatic camcorder with electronic viewfinder and full manual override such that the interface to the override is ergonomically well-designed. Although much of the processing ("thought"), such as decisions regarding exposure settings, is implemented in the "second brain" (the machine) the user is still inside the feedback loop by virtue of the fact that the electronic viewfinder mediates his/her perception of reality in accordance with decisions that the system has made. Thus just as in the ephemeral criterion where the machine does not exhibit a "mind of its own" through delays in responsiveness, here the machine does not exhibit a "mind of its own" through the theft of control from the user. By theft of control I mean the taking of control away from a user who wants more control. Indeed, the second brain can and should have its own "intelligence", and this in fact may empower the user (e.g. the fully automated camera frees the user to concentrate more on higher level compositional and cinematographic aspects), but it should not "enslave" the user by removal of the mechanism for controllability or observability. Thus there are two sub criteria associated with this criterion:

   (a) The apparatus should afford as much control to the user as is reasonably possible/practical. I call this sub criterion the 'existential controllability criterion'

   (b) The apparatus should inform the user of its operation and operational status as much as is reasonably practical/possible. I call this sub criterion the 'existential observability criterion'

   In order to better understand this criterion, I consider some counter-examples (e.g. systems that violate it). An extreme example of this violation is the synergy of enslavement arising from the remotely-controlled pain-giving device attached to prisoners [6] to make them into obedient "cyborgs". This third goal of HI borrows from existential philosophy the principle of self-determination and mastery over one's own destiny, as well as from humanistic psychology, the principle of self-actualization [7][8]. I refer to this last of the three criteria as the '**existential** criterion'.

The eudaemonic and existential criteria also provide a certain degree of personal assertiveness, which I will discuss in Chapters 7, 8, and 9. In particular, issues pertaining to the more "personal" aspects of 'second brain', e.g. that it should be and can be put beyond the power of subpoena, just as first brain already is beyond the power of anyone extracting information from it (except through extreme

measures such as torture, which only extract partial information) are presented in the nature of a manifesto on humanistic intelligence, in Chapter 8.

I should note that humanistic intelligence is somewhat different than the notion of emulating human thought by computer (e.g. Artificial Intelligence (AI) [9]), or by replacing humans with computers, and instead points toward empowering and assisting of the individual toward achieving humanistic goals.

The bulk of this thesis concentrates on a specific example of HI, called Personal Imaging, most notably, a system for attaining new levels of creativity in photography, defining a new genre of documentary video, and going beyond digital photography to define a new renaissance in imaging based on projective geometry as well as the linearity and superposition properties of light observed by Leonardo daVinci [10]. In this system, the computer need not necessarily have high-level visual intelligence, and may, in fact, be relatively un-intelligent; it is the task of the human operator to provide high-level intelligence and artistic direction.

1.6 The "WearComp" project

Imagine you haul around a travel companion or assistant in a large light-tight wooden trunk, which you open up only for occasional brief interactions with the person. How could you possibly expect such a person to be helpful? In some sense, Today's multimedia portables are like assistants we carry in sealed light-tight boxes; it's not surprising that they're often more of a burden than a help.

'WearComp' was characterized by the following three criteria:

1. **Ephemeral.** Time in processing queue (CPU) and I/O queue (user) is negligible. In Today's computing framework, this condition can be made even stronger: interactional and operational delays are nonexistent, or very small, as the computational apparatus is constant, both in operation and interaction (e.g. the computer screen/viewfinder is visible at all times, not just when the computer is being "used"). In this sense it meets both the 'first brain ephemeral' and 'second brain ephemeral' criteria.

2. **Eudaemonic.** The apparatus is worn-and-ready rather than carried-and-dormant. It was originally worn (e.g. part of the prosthetic territory) on clothing, and has recently been moved into and under clothing (e.g. covert).

3. **Existential.** The user has full control over the apparatus, e.g. while walking around, etc. (for example, the user can type, send/receive email, acquire video, process images, etc., while walking around during ordinary day-to-day situations).

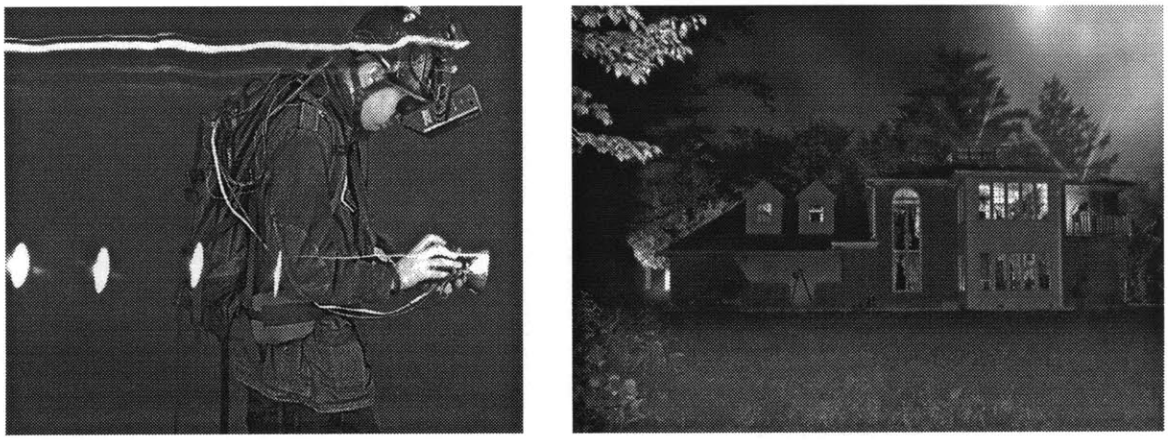
A 1970s WearComp, with a 1980 display system, is shown in Fig 1-3(a). (More details of this system will be presented in later chapters, e.g. See Fig 3-10.)

WearComp arose from an interest in the visual arts, in particular, still-life and landscape imaging — combining multiple exposures of a static scene to a variety of different light sources (Fig 1-3(b)) — but it also has widespread use, beyond its original "lightpainting" application.

By moving the screen off the lap and up to the eyes (Fig 1-4), one is able to simultaneously talk to someone and take notes without breaking eye-contact. When the apparatus is miniaturized into a normal pair of eyeglasses (Fig 1-4(d)), this capability is quite useful in business meetings and other ordinary day-to-day interactions, where an unusual appearance might be detrimental to social interaction.

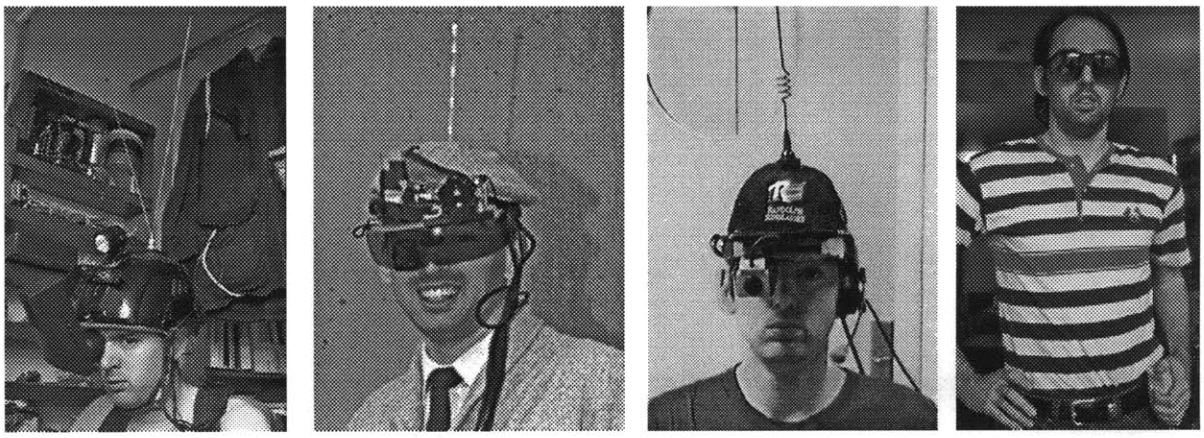
With truly portable computing, including wireless Internet connection and a simple input device (Fig 1-5) we will be able to talk to people without the distraction of looking away at a note pad or the like.

With a reasonable amount of light, images may also be incorporated into the note-taking process in a natural manner, without distracting the other person. Even in low light (for example talking to someone outside after dark), a small flash (Fig 1-5(a)) may be used during a conversation without breaking eye contact (the only distraction being the light from the flash itself).



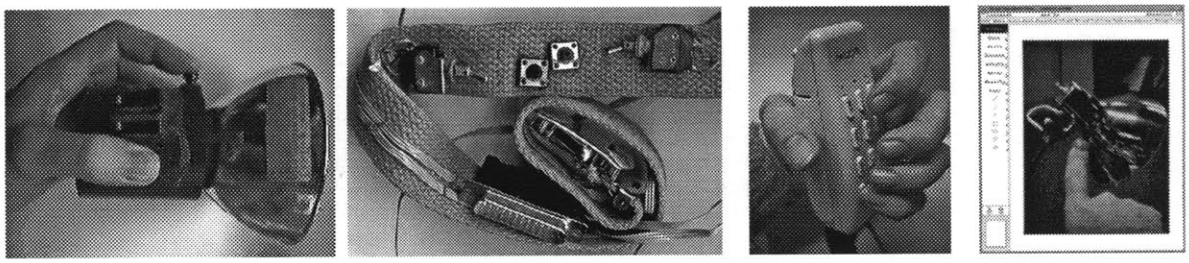
(a) (b)

Figure 1-3: (a) Early personal imaging computer (WearComp) designed and built by author for exploring new concepts in imaging/lighting. The WearComp invention consisted of a computer system that was battery powered and had wireless communications capability so that I was free to roam about, untethered. At the time, battery operated tetherless computing was a new modality of computing, as the *laptop computer* had not yet been invented. This apparatus, however, differed from present-day laptop computers and PDAs in the sense that I could interact with it while walking around, doing other things. The computer system pictured here dates from the 1970s, with a display from 1980. The display (CRT, which I typically attached somewhere where I could easily look into it without having to hold it to my eye by hand) presented both text and images. In my hand is an electronic flash lamp which allowed me to capture images in total darkness, and, more importantly, to capture, over time, a representation of how a scene or object responded to light. On the flashlamp head was an array of pushbutton switches which controlled the computer, camera, etc. (b) An early "lightpainting" made with the help of the WearComp system as a photographer's assistant. This image was generated from four differently illuminated pictures (original data acquired onto photographic emulsion, followed by postprocessing using the 'lightspace' theory to be presented in Chapters 2 and 3). The unique capabilities of a completely tetherless wearable {computer,imaging system,lighting kit} have allowed me to create expressive images that could not be created by any other means. The images transcend the boundary between photography, painting, and computer graphics. The artistic aspects of this imaging process will be presented in Chapter 7. (C) Steve Mann, 1984.



(a) (b) (c) (d)

Figure 1-4: (a) Author wearing early WearComp system from the 1970s/early 1980s, together with a 1980 display. The 1.5 inch CRT was cumbersome, requiring a well-fitted helmet to support its weight. Typically two or three antennae, operating in different frequency bands, were used to allow simultaneous transmission and reception of data, voice, or video. Alternate versions of the communications apparatus included a slightly less cumbersome clothing-based antenna array (seen hanging behind me) comprised of wires sewn directly into the clothing. This was necessary to clear doorways and ceilings during indoor use. Note the base station (a 1970s/early 1980s imaging apparatus) on the left side of the top shelf behind me, in the upper left area of the picture. (b) With the advent of consumer camcorders, miniature CRTs became available, making possible a late 1980s eyeglass-mounted multimedia computer. Here I used a 0.6 inch CRT facing down (angled back to stay close to the forehead). This apparatus was later transferred to optics salvaged from an early 1990s VirtualVision television set. (As with all of these rigs, they are in a constant state of flux, changing and evolving over time, so dates are not precise, but span a time interval.) The unit was still somewhat cumbersome, but could be worn comfortably for several hours at a time. Note whip antenna in hat (for network communications). (c) Modern commercial display product made by Kopin, together with commercially available cellular communications. With the advent of cellular and other commercial communications options, it is no longer necessary to obtain a radio license to experience 'online living'. Unlike my earlier prototypes, this system was assembled from "off-the-shelf" components. (d) Prototype "normal-looking" system (WearComp8) currently still under development.



(a) (b) (c) (d)

Figure 1-5: **Hand-held keyboards, mice, controls, etc.:** (a) Early prototype built by author (comprising one microswitch for each finger, and 3 possible microswitches for thumb) into the handle of an electronic flash lamp, allowing simultaneous one-handed control of computer, camera, and flashlamp. This form of input device is typical of my 1970s and 1980s embodiments of the 'WearComp' project. See for example the input device in Fig 3-10. (b) Covert belt-based input device operated by right hand, reaching behind back. Typically this device may be hidden underneath an untucked T-shirt or the like. The units that look like toggle switches are really spring-loaded extremely light-touch lever rockers. (c) Modern off-the-shelf mouse/keyboard combination made by Handykey Corporation. Mouse consists of tilt sensor inside housing. (d) Virtual mouse: camera in eyeglasses tracks finger which controls cursor, allowing author to look at a Luxo lamp through the glasses and draw its outline on the computer screen. The evolution of this invention is described in Chapters 6 and 7. *Self-referential note: pictures in (a), (c), and (d) were all shot by the apparatus, while wearing the apparatus, using the devices that are pictured, as devices to issue the commands which took the pictures.*

The current personal imaging prototype [11], equipped with head-mounted display, camera(s), microphones, and various other sensors, linked through wireless communications, enables computer-assisted forms of interaction in ordinary day-to-day situations, such as while walking, shopping, or meeting people. This apparatus, with its multitude of functions and affordances, replaces or subsumes the following consumer electronics devices:

- cellular telephone (e.g. sound input over i-phone or the like).
- pager (replaced with RTTY and email directly into eyeglasses)
- personal sound system (I can fit more music on my hard drive than what I can carry with me in the way of standard CDs or cassettes).
- wristwatch: having an X clock in front of my eye at all times, I can have a general awareness of time, such that when I want to check the time while conversing with others, I do not need to appear rude or seem like I am trying to get rid of them. In fact others have no way of knowing whether or not I am checking the time.
- heart rate monitor: because the apparatus is in close proximity to my body, it can take measurements of my biological information, such as heart rate, respiration, etc.
- video camera: While wearing the apparatus, I have, in a sense *become* a video camera, so I do not need to carry a camera. (This will be discussed more in Chapter 7).
- still camera: I have no need to carry a still camera either, since the apparatus is capable of making photometric measurements from which high-quality still pictures can be generated. (This procedure will be presented in Chapters 2,3,4,5.)
- personal safety device (gun, mace, or the like): With the rise in crime, an alternative to carrying a gun or other weapon is to use information and accountability as a form of self defense. It is said that in the future, wars will be fought with information. Wearing a telematic camera which is monitored by friends and relatives who have the potential to capture and record the image stream, as well as the ability to summon help, one becomes a much less desirable target/victim for would-be thieves or muggers. As department store owners and the like arm themselves with cameras, it makes sense that individuals should also be similarly protected. The social implications of this philosophy will be discussed in Chapter 8.

1.7 Smart Clothing: developing computers to wear

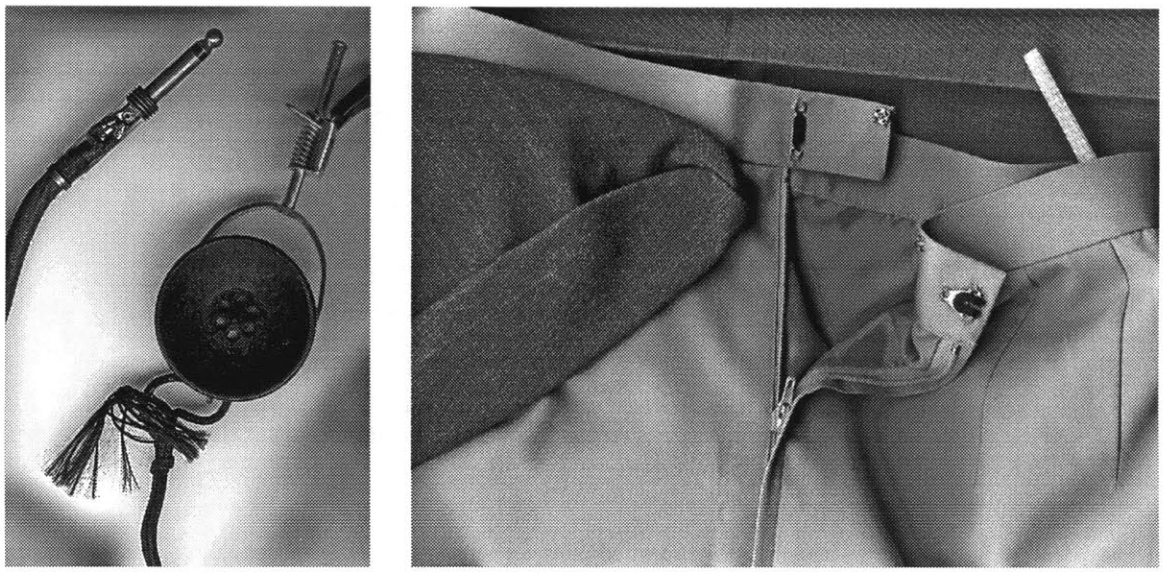
I use the term 'smart clothing' to denote variations of WearComp that are built directly into clothing, and are characterized by (or at least an attempt at) making components distributed rather than lumped, whenever possible/practical.

Smart clothing was inspired by the need for comfortable devices that I could wear for extended periods of time. The inspiration for smart clothing arose out of noticing that many of the early "wearables"[4] (e.g. headsets typically used with early "crystal radios") were far more comfortable than the newer headsets, and could often be worn for many hours[5]. Most notably, one such early headset had no headband, but instead was sewn into a cloth hat meant to be worn underneath a helmet, and thus could be worn very comfortably while sleeping.

Of particular interest is the cords used in early headsets (Fig 1-6(a)), early telephones, early patch cords, etc., which often felt much more like rope than like wire. The notion that cloth be rendered *conductive*, through the addition of metallic fibers woven in with the cloth, is one thing

[4] My father, in the early days of radio, had attempted to build a portable wearable radio system from vacuum tubes. Much of the inspiration toward smart clothing came from looking at his early headsets, passed down from previous generations in the family.

[5] It was not until many years later, with the advent of the SONY "Walkman", that comfortable headsets were rediscovered.



(a) (b)

Figure 1-6: Some simple examples of cloth which has been rendered *conductive*. (a) Cords on early headsets, telephones, etc., often felt more like rope than wire. (b) A recent generation of 'nontransparent clothing' made from bridged-conductor two-way fabric which I call 'BC2' (or includes BC2, or fabric that is rendered BC2 through addition of conductive members). Pictured here: Leftmost, a commercially manufactured nontransparent sock (the odd shape is owing to the lack of stretchability of the fabric), which I determined had approximately 20-40dB attenuation from 300MHz to 3GHz; Lower rightmost, a commercially manufactured nontransparent skirt which provided 40-50dB attenuation; Upper rightmost (behind the skirt) is a piece of fabric from 'smart clothing', with 70dB attenuation. Note that this clothing looks just like ordinary clothing, and people are not likely to see it any differently unless, for example, they are trying to covertly see through it using holographic radar or the like.

that makes possible RF shielded clothing (Fig 1-6(b)), marked in response to the growing fear of the health effects of long-term exposure to radio-frequency exposure. When I discuss the new see-through-clothing security cameras in Chapter 8, a new reason for this "nontransparent clothing" will also become evident.

One possible future direction for wearable computing is suggested by some of the 'smart clothing' I developed in the 1980s. Smart clothing is clothing that's equipped with electronic circuits. Although many of my smart clothing prototypes were merely simple or trivial forms of wearable "computers" built into clothing to switch small lights on and off for entertainment/amusement at high school dances and the like (e.g. in response to the beat of the music, as in Fig 1-7)), some of the later versions (mid 1980s) began to incorporate more advanced features, such as antenna arrays sewn into clothing or the like.

The LED shirt itself evolved from a fashion item I wore to dance clubs, parties, etc., into a useful speech-controlled LED lightpaintbrush (Fig 1-7(e)) that became a further element of the lighting toolkit I will describe in Chapter 2.

Currently, I am trying to improve this approach to using clothing itself as a connectivity medium. I experimented with two approaches to making "smart fabric": additive and subtractive. In additive, I start with ordinary cloth and sew fine wires or conductive threads into the clothing. I implemented the subtractive form using conductive cloth, of which I have identified four[6] kinds which I call BC1, IC1, BC2, IC2: conductive one direction, and conductive in both directions, either bare or insulated, respectively. See Fig 1-7(a). Note that BC2 can have two variants, those that are totally bare, and those that are bare to each other (e.g. bridged together) but then insulated, so BC2 is also short for "Bridged Conductors, 2-way".

[6] If I were weaving this myself, I would introduce a fifth, made from loops of cloth, to use as a privacy/chaff layer for secure WearComp.

Conductive materials have been used in certain kinds of drapery for many years[7], the conductive members woven in for appearance and stiffness, rather than electrical functionality. BC1 is the most common such variety. Ordinary cloth I call C0 (conductors in zero directions).

Smart clothing may have multiple layers, e.g. BC2 as RF shield, followed by one of the following possibilities:

- two BC1 layers with C0 to insulate them,
- two IC1 layers oriented at right angles,
- a single IC2 layer,

the first two being equivalent, while the last requiring additional incisions to be made to disconnect unwanted extra connectivity in both dimensions where insulation is removed with solvent. Either of these three options allow components to be "wired" together into something that's unobtrusive even to the new see-through-clothing security cameras (I measured some BC2, and found it to provide approximately 60dB of protection over a wide range of frequencies). Connections to 'smart clothing' are shown in Fig 1-7(c,d).

Recently, there has even been some commercial effort toward producing so-called "wearable computers". However, these are extremely "lumped" and uncomfortable, and we have yet to see any of the manufacturers actually use them in day-to-day living.

1.7.1 A constant and intimate user-interface

Clothing is with us almost all the time, and seems like the natural place to put computing.

Once 'personal imaging' is incorporated into our wardrobe, and we use it constantly, our computer system will enjoy the same first-person perspective as we do, and will begin to take on the role of an independent processor, much like a second brain[8]. As it "sees" the world from our perspective, it will "learn" from us — from this first-person perspective — even during moments when we are not "using" it in the traditional sense of "use".

To have a computer really "see" is a far-reaching goal — something that may require years of research. However, personal imaging is an important first step toward perceptually intelligent clothing with situational awareness.

1.8 The 'Personal Visual Assistant (PVA)' for the visually challenged

The use of the spatial filtering capability of the apparatus, as an assistant to the partially sighted, has been suggested [12], and will be presented in Chapter 6, where 'mediated reality' is introduced (e.g. see, for example, Fig 6-3). This tetherless apparatus is worn over the eyes, and, in real time, computationally augments, diminishes, or alters visual perception in day-to-day situations [13].

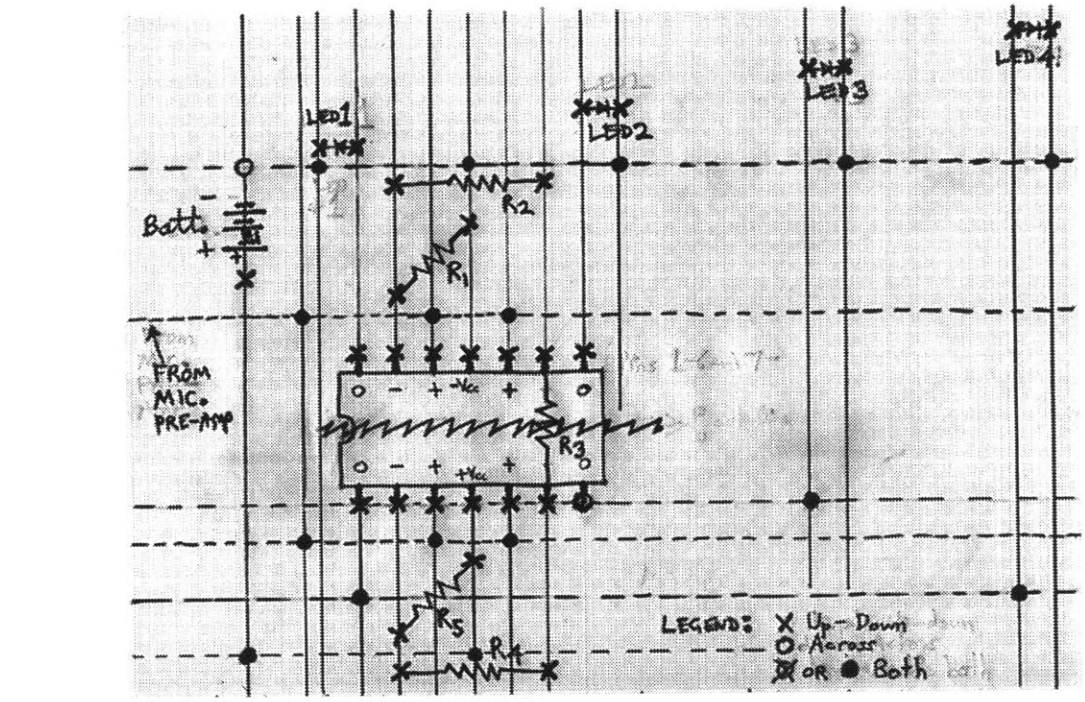
In a "fully mediated reality" experience, the only visual stimulus experienced by the wearer is that which comes from the computer screens — the glasses are not of the see-through variety.

1.9 The 'visual memory prosthetic'

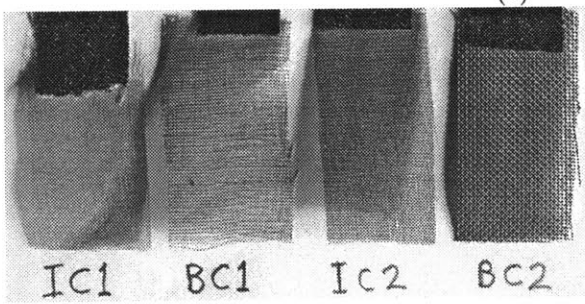
The 'visual memory prosthetic' is another possible application of compute-clothing that exemplifies sustained use of computing. While the PVA was based on spatial visual filtering [13], the 'visual memory prosthetic' is based on temporal visual filtering.

[7] Indeed, if you happen to wear electronic devices with exposed wiring, leaning against certain drapes can entail a shocking and unexpected experience.

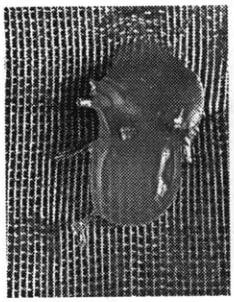
[8] Adam Oranchak (a leading industrial designer from New York, with a strong interest in the arts) describes the apparatus as a "third hemisphere" of the brain.



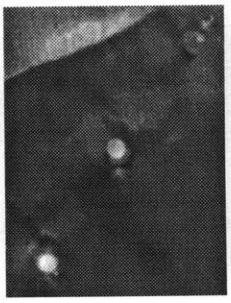
(a)



(b)



(c)



(d)

(e)

Figure 1-7: The 'smart clothing' effort as possible future generation of WearComp. (a) Early version of L.E.D. shirt comprised 4 L.E.D.s which flashed in response to sound (music, in the context of a high-school dance, or the like). This drawing illustrates a new notation developed for smart clothing circuit diagrams. The "X" and "O" notation borrows from the tradition of depicting arrows in and out of the page (e.g. "X" denotes connection to top layer which is oriented in the up-down direction, while "O" denotes connection to bottom "across" layer). The "sawtooth" denotes a cut line where enough of the fabric is removed that the loose ends will not touch. Optional lines were drawn all the way from top to bottom (and dotted or "hidden" lines across) to make it easier to "read" the diagram. (b) Four kinds of conductive fabric (see main text of article for description). (c) Back of a recent LED shirt showing where one of the LEDs is soldered directly to type-BC1 fabric (the joint has been strengthened with a blob of glue). Note the absence of wires leading to or from the glue blob, since the fabric itself acts as conductor. Typically one layer of BC1 is put inside the shirt, while the other is outside the shirt. Alternatively, either an undergarment is used, or a spacer of type-C0 between the two layers. (d) Three LEDs on type-BC1 fabric, bottom two lit, top one off. (e) A ten-LED shirt driven by wearable "computer". Note various photographic instruments I am also wearing (such as light meter worn around neck). (C) Steve Mann, 1985; thanks to Renatta Barrera for assistance.

1.9.1 'Edgertonian Eyes': Flashbacks and freeze-frames

Early on, I experimented with a variety of different *visual filters* [13], as I walked around in my day-to-day activities. Each of these filters provided a different visual reality. For one such filter, I experimented by applying a repeating freeze-frame effect to WearCam (with the cameras' own shutters set to 1/10000 second). With this video *sample and hold*, I found that quasi-periodic spatiotemporal patterns (such as railings, writings on rapidly moving automobile tires, etc) would appear to freeze at certain speeds, as objects do under the stroboscopic lights of Harold Edgerton [14].

Of greater interest than just being able to see things I would otherwise miss, was the fact that sometimes the effect would cause me to remember certain things much better. There was something very visceral about having an image frozen in space in front of my eyes. I found, for example, that I would often remember people much better with this *freeze-frame* effect. Often the frozen image would remain in my memory much longer than the moving one. Much more will be said about these so-called 'visual filters', in Chapter 6.

1.9.2 Visual Clew

We've all no doubt been lost at one time or another. Maybe you enter a new city or large shopping complex, then can't find your way back to the car or subway stop at the end of the day.

One way I overcome such visual amnesia and save myself from getting lost in a large shopping complex is by transmitting a sequence of images to my WWW page[9], and then if (when) I get lost, I browse my own WWW page to find my way back to where I started. An advantage of having the image stream on the WWW is that friends and relatives with wearable WWW browsers can catch up to me later — in some sense this forms a sort-of 'shared visual memory'.

1.10 Painting with looks: building environment maps by looking around

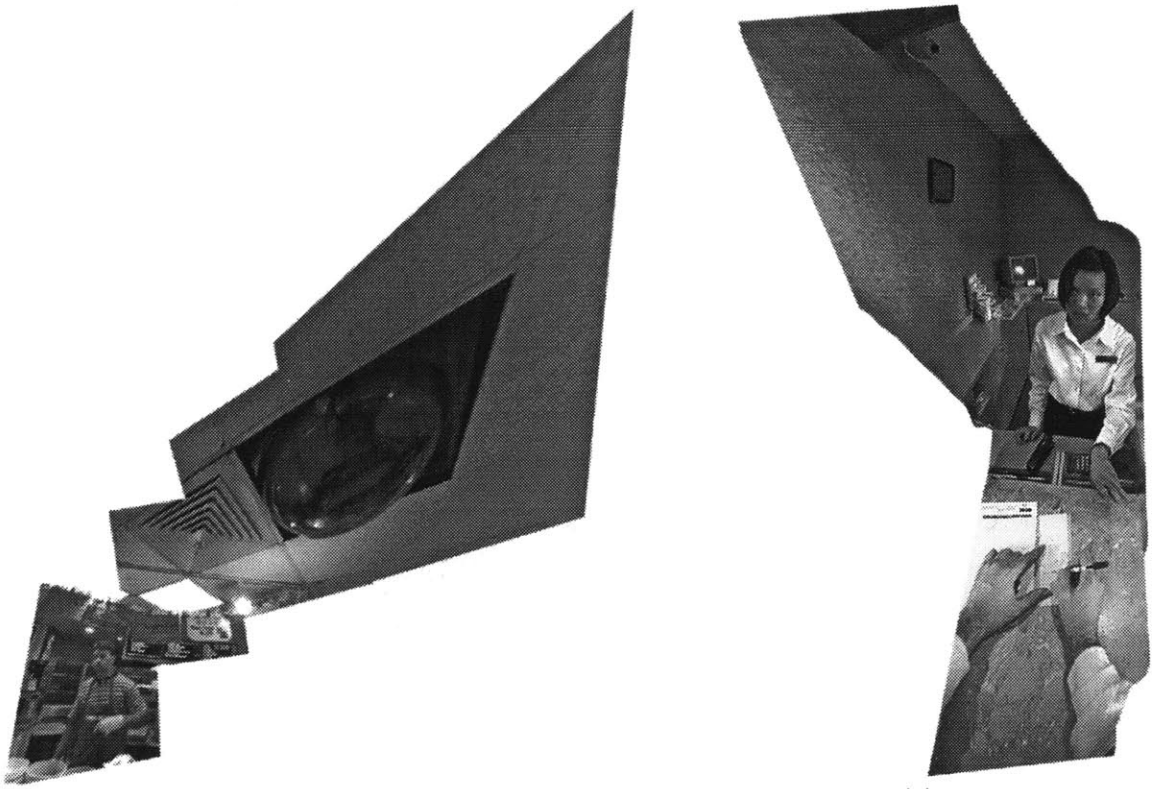
An important goal of personal imaging is for the apparatus to make sense of the world by looking around in it. Although it would certainly be desirable that the apparatus be intelligent and capable of "seeing" (e.g. through a solution of the machine vision problem), a much simpler, and more attainable goal is that of human communication through imagery, and in particular, constructing environment maps by looking around. An environment map is a collection of images, seamlessly "stitched" together, into some unified representation of the quantity of light that has arrived from each angle in space, over the range of angles for which there exist measurement data. Examples of environment maps appear in Fig 1-8.

In constructing these environment maps, the computer performs basic calculations which it is good at, while the human operator makes higher level decisions about artistic content, etc..

As described earlier, the human becomes at one with the machine (in this case the camera) through a long-term adaptation process, so that, as one experiences one's life through the apparatus (living in a computer-mediated world), the subject matter of interest is automatically captured by the human operator. Note that in this simple case, there is no Artificial Intelligence, but instead, there is a synergy between human and machine indicative of the artistic, expressive and humanistic approach to computing that's characteristic of this thesis.

This humanistic part of the goal is to build a tool to enable a human to experience the photographic world of light and shadow, and thereby, through immersion in this world over an extended period of time, develop an ability to capture details of importance in a scene. Personal imaging, which is facilitated through wearable computing, image processing, machine vision, and computer-mediated reality, will enable the user to effortlessly capture high-quality images of a new genre (to

[9] Some aspects of this personal navigation invention arose out of discussions with Rosalind W. Picard on a bike trip in June 1994, when I was wearing one of my rigs, transmitting images back to my WWW site, and we happened to get lost along the way.



(a) (b)

Figure 1-8: Environment maps (a) made from 4 input images. Image boundaries are clearly visible. Note the unified perspective and the lack of distortion (e.g. lines on the ceiling tiles are almost perfectly straight, despite the extreme perspective). (b) made from 226 input images. This composite image illustrates the nature of first-person perspective. Note that both my hands are visible in the picture. Because the apparatus is wearable, it provides a new point of view, while at the same time, capturing what is important in a scene. The methodology for generating these kinds of environment maps will be presented in Chapters 4 and 5.

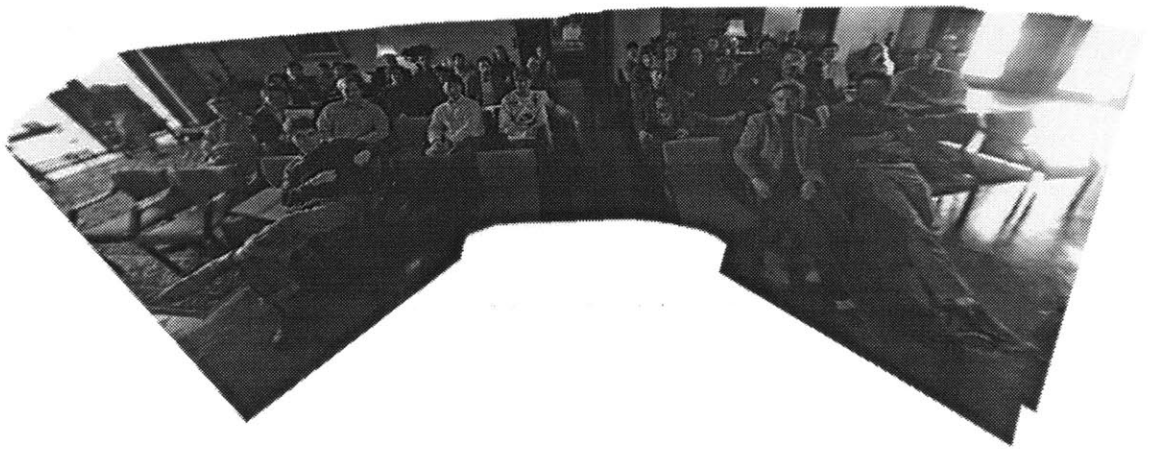


Figure 1-9: This image depicts a group of people to whom I am lecturing. Here I am able to quickly sweep out the important details of this scene (namely all of the participants), while leaving out areas of the room where nobody is seated. Note how the image is close-cropped, leaving out the two empty chairs in the center, while at the same time, extending out to include the feet of those sitting to the left and right of the empty chairs. This natural selection of subject matter happens often without conscious thought or effort, and is characteristic of the symbiotic relationship between human and machine that arises when the two become inextricably intertwined through a constancy of user-interface extending over a time period of many years.

be presented in Chapter 7), characterized by not only enhanced tonal range and spatial resolution, but also by the ability to include and exclude areas of interest (Fig 1-9).

1.10.1 Homographic modeling

A wearable object-recognizer was presented in [15]. The wearable object recognizer may also insert virtual name tags into the wearer's visual reality; the name tags will appear to stabilize on people even though the image flowfield is in motion due to the wearer looking around. This new user-interface to the real-world will be described further in Chapter 6.

Using the 'video orbits' framework [16] that will be described in Chapters 3, 4, and 5, the virtual image of a rigid planar patch [17] may be superimposed into the wearer's visual field of view, creating the illusion that a person is wearing a name tag or other visual information floating in space as a virtual image (Fig 1-10). The homography of the plane is estimated and tracked throughout, so that even when the objects being recognized fall outside the camera's field of view, tracking continues by the homography alone.

1.10.2 Seeing 'eye-to-eye'

With two personal imaging systems, two people can stay in touch, sending data, voice, and video to each other. One example application is exchange of viewpoints, where each person sees the other person's point of view overlaid on a portion of their own point of view. This gives rise to new forms of communication and interaction which will be described in Chapters 6 and 7. Through the use of shared environment maps (e.g. See Fig 6-9), we will see that personal imaging can help us allow others to not only experience our point of view vicariously, but will also allow us to allow others to mediate our perception of reality. Such mediation may range from simple annotation of objects in our "reality stream", to completely altering our perception of reality.

1.10.3 'SafetyGlasses' and 'SafetyNet'

In Chapter 9, I will describe how instead of just two people, we may have a networked community of individuals wearing networked compute-clothing. A personal safety device will be proposed, comprising a device built such that others can not easily determine whether or not it is a camera,

Figure 1-10: **Six frames of video from processed image sequence:** Apparatus recognizes cashier and superimposes previously entered shopping list on her. When I turn my head to the right, the list moves to the left on my screen, following the flowfield of the video imagery coming from my camera. Note that the tracking (initially triggered by automatic object-recognition) continues even when the cashier is completely outside my visual field, because tracking is sustained by other objects in the room, such as the counters, walls, row of fluorescent lights, and the 20 or 30 video surveillance cameras installed on the ceiling. In this way, the illusion is that the list of items I have purchased from this cashier appears to be attached to the cashier. (Unlike typical augmented reality in which registration is a major hurdle, here mediated reality, to be described in Chapter 6, is used to attain sub-pixel accuracy of registration between real and virtual worlds.) This general functionality is handy during a refund explanation, where a clear recollection of the facts is desirable. Issues pertaining to a new balance between individuals and establishments, that such a technique provides, will be discussed in Chapter 8.

and whether or not it might comprise a networked neighbourhood watch in the form of humanistic intelligence. Furthermore, a fear of danger might be triggered by a 'maybe I'm in distress' signal from the wearer, for example, when uncertain whether a situation calls for emergency action or not, the wearer could just enter into a casual interaction over the Internet, or the clothing itself could trigger the signal. For example, in my clothing I have a heart rate monitor and a footstep activity meter in my shoes. Heart rate divided by the rate of footsteps could give a 'saliency' index, which might be a cue to danger. For example someone pulling out a gun and asking for cash might cause heart rate to increase and footsteps to slow down, which is contrary to the usual patterns of physiological behaviour[10].

A version of the personal imaging workstation, equipped with biosensors[11], is illustrated in Fig 1-11.

In actual practice, intelligent signal processing might be used to make inferences as to possible states of danger, or the like, responding appropriately through a network of individuals looking out for each other's safety and well-being. A community of individuals networked in this way would look out for each others' safety in the form of a 'neighbourhood watch'.

In Chapters 8 and 9, I propose 'Safety nets' as an alternative [15] to the proliferation of government surveillance cameras installed throughout many cities, such as in the UK. Even in the US, the government is experimenting with the installation of ubiquitous video surveillance to keep watch over citizens' activities (e.g. 200 cameras are currently being installed in Baltimore as an experiment). However, rather than requesting government surveillance, citizens might look out for one another using clothing-based Internet-connected computing. This would reduce tax dollars and provide a future more like David Brin's "Earth" than George Orwell's "1984"[12], bringing us toward a small town/global village where we communicate with each other. This philosophy will be described in

[10] The idea for a wearable computer with a multi-channel analog to digital converter, for measuring and responding to biological signals, arose out of my work with Dr. Ghista at McMaster university, where we developed portable instruments used in actual clinical studies. This work was further developed, some years later, in connection with a course I took at McMaster, under the direction of Dr. DeBruin. That such an apparatus might be used to make inferences about human emotions (e.g. to correlate measurements of biological signals with human emotion) arose out of work with my advisor, Rosalind. W. Picard, at M.I.T.. Many of the recent directions in this work arose out of discussions with Picard.

[11] Early versions of wearable computers with biological sensors were based on my home-brew analog to digital converters. More recently, these systems have been based on the ProComp 8-channel analog to digital converter. To the best of my knowledge, the first embodiment of this recent generation of wearables based on the ProComp was WearComp5. Subsequently, I assisted Jennifer Healey in the procurement of components to build a version of WearComp5. More recently, however, the ProComp has been used in conjunction with WearComp6, or other similar PC104-based wearable computers [18].

[12] Orwell predicts a future of cameras and microphones distributed throughout the environment, e.g. two-way TV

Figure 1-11: Personal Imaging workstation equipped with sensors for measuring biological signals. In the upper right are my "smart sunglasses" with built in video cameras and display system. These look like ordinary sunglasses when worn (wires are concealed inside the eyeglass holder). Immediately to the left of the glasses (top center of picture) is a commercial display unit called the Private Eye, which may be used instead of the "smart sunglasses" if imaging capability is not required. At the far left is an 8 channel analog to digital converter together with a collection of biological sensors, both manufactured by Thought Technologies Limited, of Canada. At the lower right is an input device called the "twiddler", manufactured by HandyKey, and to the left of that is a Sony Lithium Ion camcorder battery with custom-made battery holder. In the lower central area of the image is the computer, equipped with special-purpose video processing/video capture hardware (visible as the top stack on this stack of PC104 boards). To the left of the computer, is a serial to fiber-optic converter that provides communications to the 8 channel analog to digital converter over a fiber-optic link. Its purpose is primarily one of safety, to isolate high voltages used in the computer and peripherals (e.g. the 500 volts or so present in the smart sunglassses) from the biological sensors which are in close proximity, typically with very good connection, to the body of the wearer.

more detail in Chapter 8.

1.11 Chapter summary

The overall layout of the thesis has been presented in the context of a new synergy between art, science, and technology, and it was argued that this synergy is most apropos for a PhD degree in the Media **Arts** and *Sciences* section of the Massachusetts Institute of *Technology*.

The notion of 'humanistic intelligence' has been put forth together with means and apparatus to realize it (WearComp). WearComp suggests a different context for computing, in particular, devices that will be completely covert, totally comfortable, and will interface to us as naturally as ordinary clothing and eyeglasses. A possible future direction for WearComp was suggested, based on the notion of 'smart clothing' — clothing equipped with computational capability of sorts.

'Personal imaging' has also been proposed, and arises from a special case of WearComp called WearCam. Practical applications of personal imaging were also suggested, for example, the resulting wearable tetherless computer-mediated reality was presented as a prosthetic device.

The boundaries between seeing and viewing, and between remembering and recording will blur. Shared visual memory will begin to enlarge the scope of what the visual memory prosthetic currently provides. Connected humanistic intelligence will afford us with new ways of collaborating, for it will be possible to 'remember' details of something or someone that one never saw. For example, we might allow others to alter our perception of reality so that, for example, a spouse who sees an old friend remotely through the apparatus might annotate our 'reality stream' and remind us to say hello to an old friend, who we then greet by name even though we never met the person before.

Once computing becomes part of the user, interaction with the computer will become much more natural. Not only does this improve the ability to do traditional computing tasks while standing or walking, but it also suggests a future in which our computer systems will function much like a second "brain". Obviously this could and probably will evolve toward situational awareness and perceptual intelligence — towards an ability for the apparatus to "see" from the first-person perspective of the wearer, and to assist in day-to-day interactions, by being constantly attentive to our environment. However, more importantly, within the context of this thesis, a new framework for photographic/video personal documentary and completely new ways for a person to interact with their computer system will arise, and a whole new set of needs and issues will need to be addressed, technically, scientifically, and socially, as we take a first step towards personal imaging.

In the future, "mental tools" and personal electronics will not be what we carry, but what we'll be.

sets, while Brin predicts a future in which many citizens wear cameras and are networked together. Brin argues that "cameras are coming", one way or the other, and that privacy as we know it will disappear. He argues that the kind of privacy loss one experiences in a small town is "less evil" than that experienced in an Orwellian society.

Chapter 2

Beyond digital photography: A new imaging renaissance.

Personal imaging is an attempt to (1) re-situate the camera in a new way — as an extension of the mind and body rather than an entity that we might carry with us, and (2) allow us to capture a personal account of reality. The latter goal is towards both personal documentary and an expressive (artistic and creative) form of imaging arising from the ability to capture a rich multidimensional description of a scene, and then "render" an image from this description at a later time.

This latter goal is not to alter the scene content, as is the goal of much in the way of digital photography [19] (e.g. through such programs as Adobe's PhotoShop or the like), but, rather, to manipulate the tonal range, apparent scene illumination, or the like, to faithfully, but expressively, capture an image of objects actually present in the scene.

In much the same way that daVinci's paintings portray realistic scenes, but with inexplicable light and shade (e.g. the shadows often appear to correspond to no single possible light source), a goal of personal imaging is a new renaissance in tonal range, light-and-shadow, etc., for which the wearable computer and associated imaging apparatus described in Chapter 1 was first designed and built.

Accordingly, a general framework for understanding some simple but important properties of light is put forth. Later this framework will be related to the specifics of the wearable computer and personal imaging apparatus, but for the purposes of this chapter, the theory will be put forth in the abstract, as well as with the aid of some simple thought experiments.

2.1 Introduction

I present a philosophical (conceptual) framework that describes a model of the way that light interacts with a scene or object. I call this framework "lightspace". I first show how any of a variety of typical light sources (including those found in the home, office, and photography studio) can be mathematically represented in terms of a collection of primitive elements that I call 'spotflashes'. Due to the linearity and additivity properties of light intensity, I then show that any lighting situation (e.g. combination of sunlight, fluorescent light, etc) may be expressed in terms of a collection of 'spotflashes'. Lightspace captures everything that can be known about how a scene will respond to each of all possible spotflashes, and therefore, by this decomposition, to any possible light source. Note that the presentation in this chapter is not meant to be of direct practical utility, nor is it meant to replace the principles of computer graphics which might be more efficient to actually implement.

In Chapter 3, I will address more practical issues, and use lightspace as a framework to derive some new algebraic relationships between multiple pictures of the same scene or object, as arise from the wearable personal imaging apparatus.

2.2 The "plenoptic function"

We begin by asking what potentially can be learned from measurements of all the light rays present in a particular region of space. Adelson asks this question:

> What information about the world is contained in the light filling a region of space? Space is filled with a dense array of light rays of various intensities. The set of rays passing through any point in space is mathematically termed a *pencil.* Leonardo da Vinci refers to this set of rays as a "radiant pyramid" [20]

Leonardo expressed a similar idea, realizing the significance of this complete visual description:

> The body of the air is full of an infinite number of radiant pyramids caused by the objects located in it[1]. These pyramids intersect and interweave without interfering with each other during their independent passage throughout the air in which they are infused. [10]

We can also ask how we might benefit from being able to capture, analyse, and re-synthesize these light rays. In particular, *black and white (greyscale)* photography captures the pencil of light at a particular point in spacetime (x, y, z, t) integrated over all wavelengths (or integrated together with the spectral sensitivity curve of the film). Color photography captures three readings of this wavelength-integrated pencil of light each with a different spectral sensitivity (color). An earlier form of color photography, known as *Lippman photography* [21][22] decomposes the light into an infinite[2] number of spectral bands, providing a record of the true spectral content of the light at each point on the film.

A long exposure photograph captures a time-integrated pencil of light. Thus a black and white photograph captures the pencil of light at a specific spatial location (x, y, z), integrated over all (a particular range of) time, and over all (a particular range of) wavelengths. Thus the idealized (conceptual) analog camera is a means of making uncountably many measurements at the same time (e.g. measuring many of these light rays at once).

2.2.1 The spot-flash-spectrometer

Let us, for the moment, suppose that we wish to measure (and record) the energy in a single one of these rays of light, at a particular wavelength, at a particular instant in time [3]. We select a point in space (x, y, z) and place a flashmeter at the end of a collimator (Fig 2-1) at that location. We select the wavelength of interest by adjusting the prism [4] which is part of the collimator. We select the time period of interest by activating the trigger input of the flashmeter. In practice, a flashmeter integrates the total quantity of light over a short time period, such as 1/500 of a second, but we can envision an apparatus where this time interval can be made arbitrarily short, while the instrument is made more and more sensitive[5]. Note that the collimator and prism serve to restrict our measurement to light traveling in a particular direction, at a particular wavelength, λ.

There are seven degrees of freedom in this measuring apparatus[6]. I denote these by $\theta, \phi, \lambda, t, x, y,$ and z, where the first two degrees of freedom are derived from a unit vector that indicates the direction we are aiming the apparatus, and the last three denote the location of the apparatus in space (or

[1] Perhaps more correctly, by the interaction of light with the objects located in it

[2] While we might argue about infinities, in the context of quantum effects of light, or the like, I use the term "infinite" in the same conceptual spirit as daVinci used it, that is, without regard to practical implementation, or actual information content.

[3] neglecting any uncertainty effects due to the wavelike nature of light, and any precision effects due to the particle-like nature of light

[4] In practice a blazed grating (diffraction grating built into a curved mirror) might be used, since it selects a particular wavelength of light more efficiently than a prism, though I use the familiar triangular icon to denote this splitting up of the white light into a rainbow of wavelengths.

[5] neglecting the theoretical limitations of both sensor noise and the quantum (photon) nature of light

[6] Note that in a transparent medium one can move along a ray of light with no change, so that measuring the lightspace along a plane will suffice thus making the measurement of it throughout the entire volume redundant. In many ways, of course, the lightspace representation is conceptual, rather than practical.

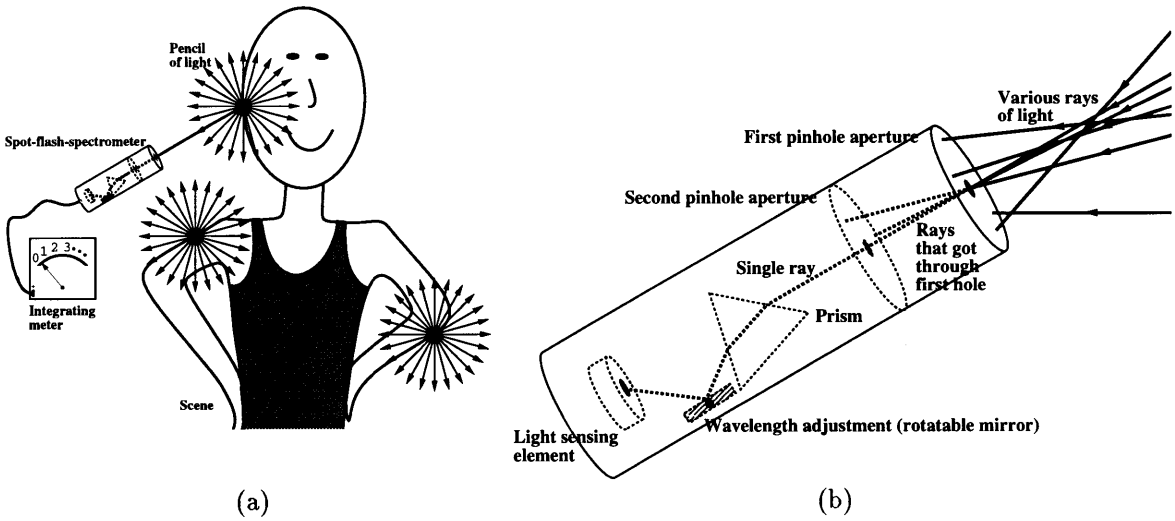


Figure 2-1: Every point in an illuminated 3-D scene radiates light in all directions. Conceptually, at least, we can characterize the scene, together with the way it is illuminated, by measuring all of these rays of light in the space around the scene. At each point in space, we measure the amount of light traveling in every possible direction (direction being characterized by a unit vector which has two degrees of freedom). Since objects have various colors, and, more generally, various spectral properties, so too will the rays of light reflected by them, so that wavelength is also a quantity which we wish to measure. (a) Measurement of one of these rays of light. (b) Detail of measuring apparatus comprising omnidirectional point sensor in colimating apparatus. We will call this apparatus a 'spot-flash-spectrometer'.

the last four denote the location in 4-space, if you prefer to think that way). At each point in this seven-dimensional space we obtain a reading that indicates the quantity of light at that point in the space. This quantity of light might be found, for example, by observing an integrating voltmeter connected to the light sensing element at the end of the collimator tube. We will call this entire apparatus a 'spot-flash-spectrometer', as it is similar to the flash spotmeter that photographers use to measure light bouncing off a single spot in the image, typically over a narrow (e.g. 1 degree or so) beam spread, and short (e.g. 1/500 sec) time interval.

Suppose that we could obtain a complete set of these measurements (e.g. measurements of the uncountably[7] many rays of light present in the space around the scene). This complete description is a real-valued function of seven real variables, which completely characterizes the scene to the extent that from it we would be able to later synthesize all possible natural-light (e.g. no flash or other artificially imposed light sources allowed) pictures (still pictures or motion pictures) that could have been taken of the scene. Adelson calls this function the "plenoptic function" [20].

Suppose, for example, that we now knew the plenoptic function defined over the setting[8] of Dallas, November 22, 1963. From this plenoptic function, we would be able to synthesize all possible natural-light pictures of the presidential entourage with unlimited accuracy and resolution; we could synthesize motion pictures of the grassy knoll at the time that the president was shot, and we could know everything about this event that could be obtained by visual means (e.g. by the rays of light present in this setting). In a sense we could extract more information than if we had been there, for we could synthesize extreme close-up pictures of the gunman on the grassy knoll, and magnify still more to show the serial number on his gun, without any risk of getting shot by him. We could generate a movie at any desired frame rate, such as 10000 frames per second, and watch the bullet come out the barrel of the gun, examining it in slow motion, to see what markings it might have on it while it is traveling through the air, even though this information might not have been of interest

[7] Again, I use the term "uncountable" in a conceptual spirit. If the reader prefers to visualize the rationals — dense in the reals but countable, or prefers to visualize a countably infinite discrete lattice, or a sufficiently dense finite sampling lattice, this will still convey the general spirit of light in the daVinci sense.

[8] A *setting* is a time-span and space-span, or, if you prefer, a region of (x, y, z, t) 4-space.

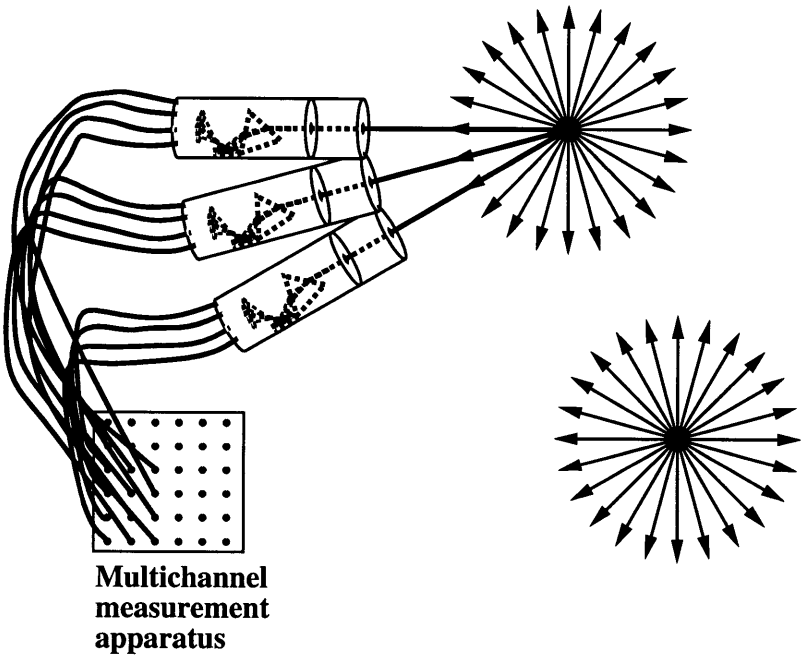


Figure 2-2: A number of spotmeters arranged to simultaneously measure multiple rays of light. Here the particular situation depicted has the measuring instruments measuring rays at 4 different wavelengths, traveling in 3 different directions, but the rays all pass through the same point in space. If we had uncountably many measurements over all possible wavelengths and directions at one point, we would have an apparatus capable of capturing a complete description of the pencil of light at that point in space.

(or even thought of) at the time that the plenoptic function was measured.

In order to speed up the measurement of a "plenoptic function", we consider a collection of measuring instruments combined into a single unit. Some examples might include:

- A collimator that has many light sensing elements placed inside, around the prism, so that each one measures a particular wavelength. This device could simultaneously measure many wavelengths over a discrete lattice. We call such an instrument a 'spot-spectrometer' or 'spot-flash-spectrometer'.

- A number of spot-spectrometers operating in parallel at the same time, to simultaneously measure more than one ray of light. Rather than placing them in a row (simple linear array), there is a nice conceptual interpretation that results if the collimators are placed so that they all measure light rays passing through the same point (Fig 2-2). With this arrangement, all the information gathered from the various light-sensing elements pertains to the same pencil of light.

In our present case, we are interested in an instrument that would simultaneously measure an uncountable number of light rays coming in from an uncountable number of different directions, and measure the spectral content (e.g. make measurements at an uncountable number of wavelengths) of each ray. Though this is impossible in practice, the human eye comes pretty close, with its 100 million or so light sensitive elements. Thus we will denote this collection of spot-flash-spectrometers by the human-eye icon ('eyecon') depicted in Fig 2-3. The important difference to keep in mind, however, when making this analogy, is that the human eye only captures three spectral bands (e.g. represents all spectral readings as three real numbers denoting the spectrum integrated with each of the 3 spectral sensitivities) whereas the proposed collection of spot-spectrometers captures all spectral information of each light ray passing through the particular point where it is positioned, at every instant in time, so that a multichannel recording apparatus could be used to capture this information.

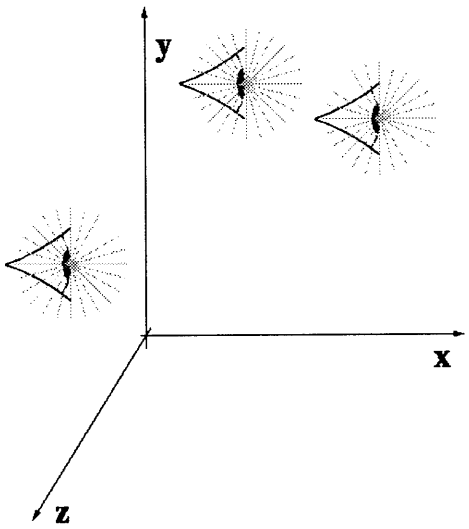


Figure 2-3: An uncountable number of spot-spectrometers arranged (as in Fig 2-2) to simultaneously measure multiple rays of light is denoted by the human eye icon ('eyecon') because of the similarity to the human visual system. An important difference, though, is that in the human visual there are only three spectral bands (colors), whereas in our version there are an uncountable number of spectral bands. Another important difference is that our collection of spot-spectrometers can "see" in all directions simultaneously, whereas the human visual system does not allow one to see rays coming from behind. Each eyecon represents an apparatus that records a real-valued function of four real variables, $f(\theta, \phi, \lambda, t)$, so that if the 3-D space were packed with uncountably many of these, then the result would be a recording of the plenoptic function, $f(\theta, \phi, \lambda, t, x, y, z)$.

2.3 The "spotflash" primitive

So far, I have said a great deal about rays of light. Now let us consider an apparatus for generating one. If we take the light measuring instrument depicted in Fig 2-1 and replace the light sensor with a flashtube (a device capable of creating a brief burst of white light that radiates in all directions), we obtain a similar unit that functions in reverse. The flashtube emits white light in all directions (Fig 2-4), and the prism (or diffraction grating) causes these rays of white light to break up into their component wavelengths. Only the ray of light that has a certain specific wavelength will make it out through the holes in the two apertures. The result is a single ray of light that is localized in space (by virtue of the selection of its location, in time (by virtue of the instantaneous nature of electronic flash), in wavelength (by virtue of the prism), and in direction (azimuth and elevation).

Perhaps the closest actual realization of a spotflash would be a pulsed dye-laser[9] which can create short bursts of light of selectable wavelength, confined to a narrow beam.

As with the spotmeter, there are seven degrees of freedom associated with this light source: azimuth, θ_l; elevation, ϕ_l, wavelength, λ_l, time, t_l; and spatial position, (x_l, y_l, z_l).

2.3.1 Building a conceptual lighting toolbox: Using the spotflash to synthesize other light sources

The spotflash is a primitive from which other light sources may be built. We will construct a hypothetical toolbox containing various lights built up from a number of spotflashes.

[9]Though lasers are most known for their coherency, in this chapter, we ignore the coherency properties of light, and are interested only in the fact that lasers generally shine a ray of monochromatic light along a single direction.

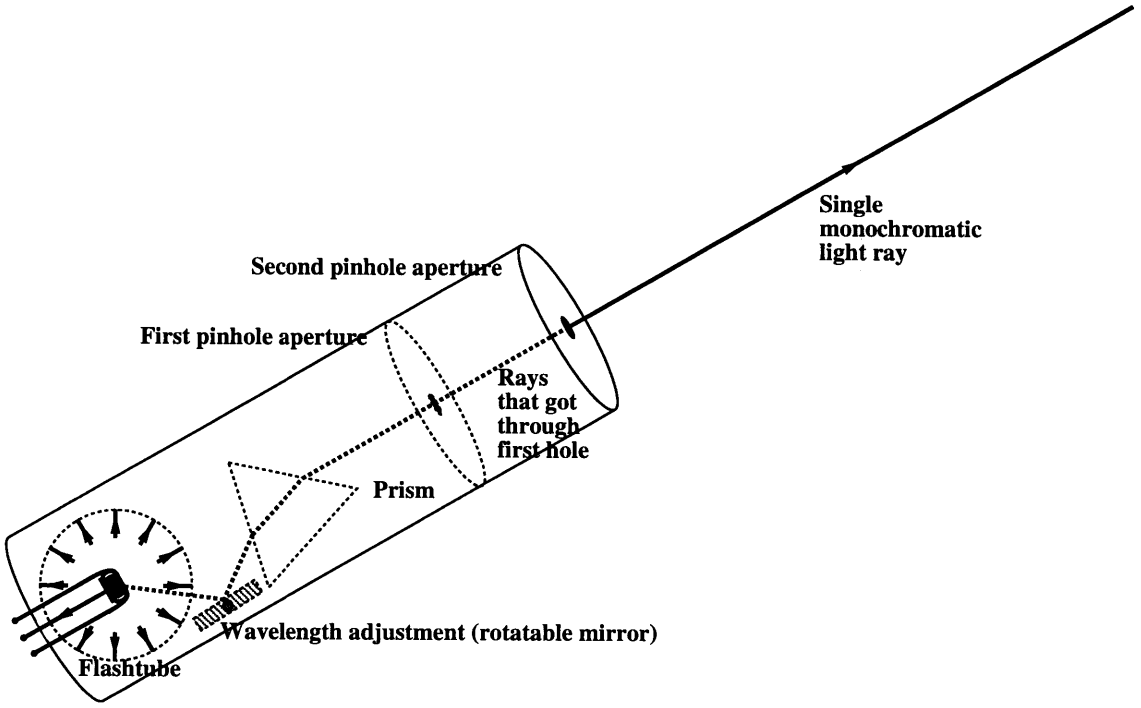


Figure 2-4: Monochromatic flash spotlight source of adjustable wavelength. I will refer to this light source as a 'spotflash', as it is similar to a colored spotlight that is flashed for a brief duration. (Note the integrating sphere around the flashlamp. It is reflective inside, and has a small hole through which light can emerge.)

White spotflash

The ideal spotflash is infinitesmally[10] small, so that we can pack arbitrarily many of them into as small a space as desired. If we pack uncountably many spotflashes close enough together, and have them all shine in the same direction, we can set each one at a slightly different wavelength, so that they will act collectively to produce a single ray of light that contains all wavelengths. Now imagine that we connect all of the trigger inputs together so that they all flash simultaneously at each of the uncountably many component wavelengths. We will call this light source the 'white-spotflash'. The white-spotflash produces a brief burst of white light confined to a narrow beam. Now that we have built a white-spotflash, we put it into our conceptual toolbox for future use.

Fan beam (pencil of white light)

If we pack uncountably many white-spotflashes together into the same space so that they fan out in different directions, but the light rays all exist in the same plane, and all pass through the same point, then, if we fire all of the white-spotflashes at the same time, we obtain a sheet of directed light that all emanates from a single point. We call this light source the 'fan beam', and place it into our conceptual toolbox for future use. This arrangement of white-spotflashes resembles the arrangement of flash spotmeters in Fig 2-2.

Flash point source (bundle of white light)

If we pack uncountably many white-spotflashes together into the same space so that they fan out in all possible directions but pass through the same point, then we obtain a 'flash point source' of light. Now that we have constructed a 'flash point source', we place it in our conceptual toolbox for

[10] Again, the same caveat applies to "infinitesimal" as to "infinite" and "uncountable".

future use. Light sources that approximate this ideal 'flash point source' are particularly common. The best example I know of is Harold Edgerton's *microflash point source* which is a small spark gap that produces a flash of white light, radiating in all directions, and lasting approximately 1/3 of a microsecond. Any bare electronic flashtube (e.g. with no reflector) is a reasonably close approximation to a 'flash point source'.

Point source

If we take a flash point source and fire it repeatedly over and over again[11] we obtain a flashing light. If we allow the time period between flashes to approach zero, we obtain a light that stays on continuously. We have now constructed a continuous source of white light that radiates in all directions. We call it a 'point source' and place it in the conceptual toolbox for future use.

In practice, if we could use a microflash point source that lasts one third of a microsecond, and flash it with a three megahertz trigger signal (three million flashes per second) it would light up continuously[12].

The point source is much like a bare light bulb, or a household lamp with the shade removed, continuously radiating white light in all directions, but from a single point in (x, y, z) space.

linelight

If we take either uncountably many point sources and arrange them along a line in 3-space (x, y, z), or if we take a lineflash and flash it repeatedly so that it stays on, we obtain a linear source of light that I call the 'linelight', which we place in the conceptual toolbox for future use. This light source is quite similar to the long fluorescent tubes that are used in office buildings.

sheetlight

A sheetflash fired repetitively, so that it stays on, produces a continuous light source that I call a 'sheetlight'. Videographers often use a light bulb placed behind a white cloth to create a light source similar to the 'sheetlight'. Now that we have "constructed" yet another form of light, the 'sheetlight', let us place it in our conceptual lighting toolbox for future use.

volume light

Uncountably many 'sheetlights' stacked on top of one another forms a 'volume light', which we now place into our conceptual toolbox. Some practical examples of volumetric light sources include the light from luminous gas like the sun, or a flame. Note that I have made the non-realistic assumption that each of these constituent sheetlights is transparent. (Do not forget that lightspace, as it is presented in this chapter, is a concept rather than a practical entity.)

seen through the glass, lightly (on "impossible" light sources)

This assumption — that rays of light can pass through the 'sheetlight' instrument itself — is no small assumption.

In fact, photographer's softboxes (perhaps the practical closest approximation to 'sheetlight') are far from transparent. Typically a large cavity behind the sheet is needed to house a more conventional light source. Now suppose that what I desire is a picture that is illuminated by a sheet light that is located between the camera and the object being photographed. That is, what I would like is a picture of an object as it appears, while looking through the sheetlight. One way of obtaining such a picture is to average over the light intensity falling on an image sensor (e.g. through

[11] Or, alternatively, we can think of this arrangement as a row of flash point sources arranged along the time axis, and fired together in (x, y, z, t) 4-space.

[12] Of course this "practical" example is itself somewhat hypothetical, as the flash really takes time to "recycle" itself to be ready for the next flash. In this thought experiment, recycle time is neglected. Alternatively, imagine a xenon arc lamp that stays on continuously.

a long-exposure photograph, or through making a video and then homomorphically averaging all the frames together, as will be described in Chapter 3), while moving a 'line light' across directly in front of the object. The 'line light' is moved (say, from left to right), directly in front of the camera, but because it is in motion, it is not seen by the camera — the object itself gets averaged out over time. A picture taken in this manner is shown in Fig 2-5. As indicated in the figure caption, the light source itself may be constructed to radiate in some directions more than others, and this radiation pattern may even change (evolve) as the light source is moved from left to right. An approximate (e.g. discrete) realization of a linelight that can evolve as it moves from left to right will be discussed in Chapter 3 (e.g. see Fig 3-10)

It should also be noted that the linelight, which is made from uncountably many point sources (or a finite approximation), may also have fine structure. In particular, each of these point sources may be such that it radiates unequally in various directions. A simple example of a picture that was illuminated with an approximation to a linelight, appears in Fig 2-6[13].

Dimensions of light

In Fig 2-7 I illustrate some of these light sources, categorizing them by the number of dimensions (degrees of freedom) that they have in both 4-space (t, z, y, z), and 7-space $(\theta, \phi, \lambda, t, x, y, z)$.

The aremac[14] and controllable light sources

We now have, in our conceptual toolbox, various hypothetical light sources, such as a point source of white light, an infinitely long slender lamp (e.g. a line that produces light), an infinite sheet of light (a plane that produces light), (from which could be — although only conceptually due to self-occlusion — constructed an infinite 3-D volume that produces light). We already saw pictures taken with some practical examples of approximations to some of these light sources — pictures that show how we can expand our creative horizons by using "impossible" light sources (e.g. light sources that do not exist in reality, but through some "trickery" of lightspace, we are able to synthesize a picture as it would have appeared had it been taken with such a light source).

We now imagine that we have full control of each light source. In particular, the 'light volume' (volumetric light source) is composed of uncountably many 'spotflashes' (infinitesimal rays of light).

If, when we construct the various light sources, we retain control of the individual spotflashes, rather than connecting them together to fire in unison, we obtain a 'controllable light source'.

For example, if we assemble a number of spotflashes of different wavelength, as we did to form the white spotflash, but this time we retain control (e.g. we have a voltage on each spotflash), we could select any desired spectral distribution (e.g. color). We call the resulting source a 'controllable spotflash'. The 'controllable spotflash' takes a real-valued function of one real variable as its input, and from this input, produces, for a brief instant, a ray of light that has a spectral distribution corresponding to that input function.

The 'controllable spotflash' subsumes the 'white spotflash' as a special case. The white spotflash corresponds to a 'controllable spotflash' driven with a wavelength function that is constant.

Assembling a number of controllable spotflashes at the same location but pointing in all possible directions in a given plane, and maintaining separate control of each spotflash, provides us with a source that can produce any pencil of light, varying in intensity and spectral distribution, as a function of angle. I call this apparatus the 'controllable flash-pencil', and it takes as input, a real-valued function of two real variables.

---

[13] Once I indicate how the image was generated, it will become quite obvious why the *directionality* arises. Here I had three models stand in a rail boxcar, open at both sides, but stationary on a set of railway tracks. On an adjacent railway track, a train with headlamps moved across behind the models, during a long exposure which effectively integrated, over time. However, the thought exercise — thinking of this process as a single static long slender lightsource, composed of uncountably many point sources that each radiate over some fixed solid angle to the right, helps us to better understand the principle of lightspace.

[14] I formed this word by reversing the order of the letters in the word "camera" (e.g. spelled backwards). We will see later, that the aremac is the functional opposite of a camera in the sense that it converts image data into a bundle of light rays, as opposed to the camera which converts a bundle of light rays into image data.

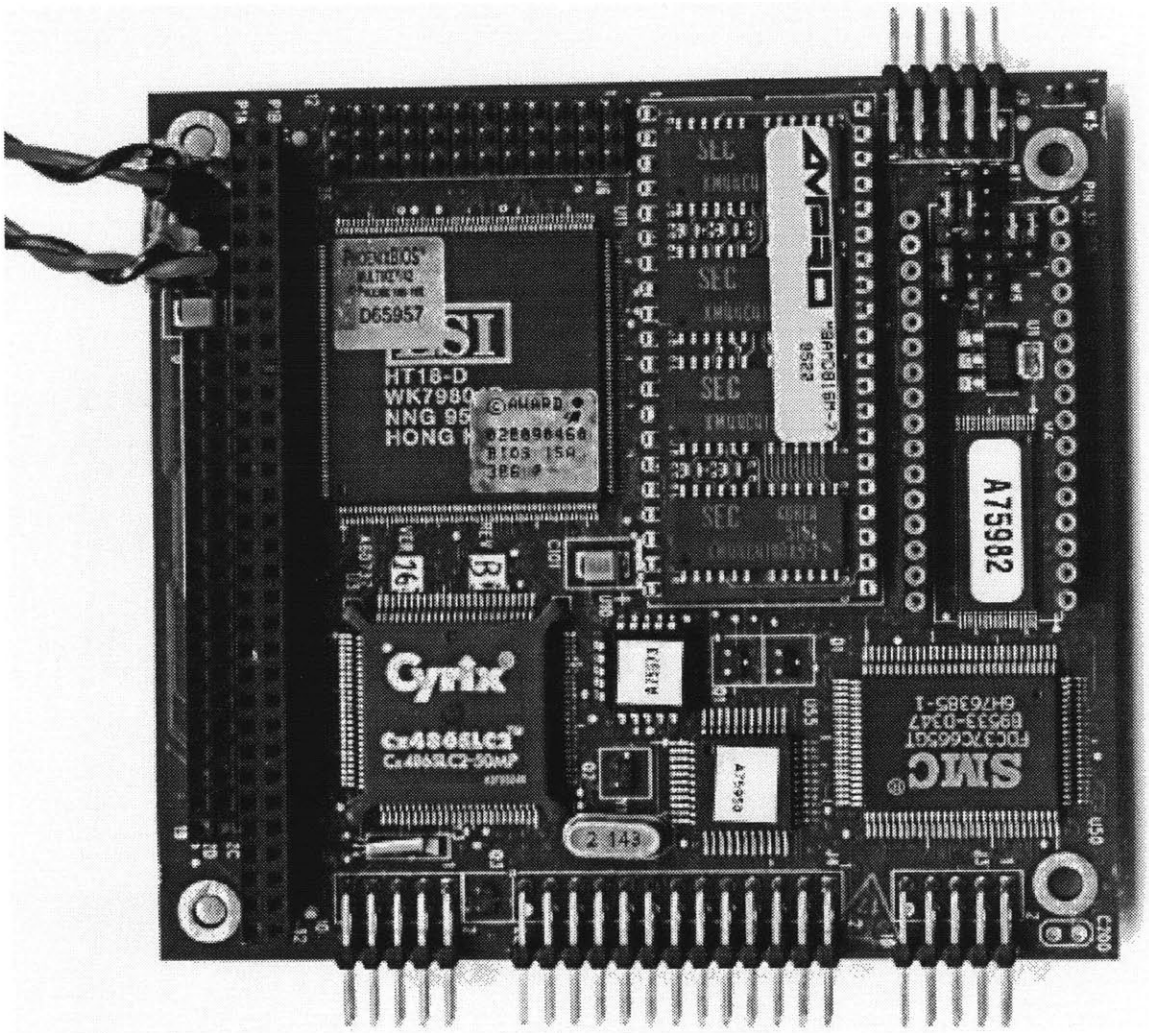


Figure 2-5: **Now we see, as though, through a glass, lightly.** Imagine that there is a plane of light (e.g. a glass or sheet that produces light itself). Imagine now that this light source is also totally transparent, and that it is placed between you and some object. The resulting light is very soft upon the object, providing a uniform illumination without distinct shadows. Such a light source does not exist in practice, but may be simulated by homomorphically combining multiple pictures (as will be described in Chapter 3), each taken with a linear source of light ('linelight'). Here a linelight was moved from left to right. Note also that the linelight need not radiate equally in all directions. If it is constructed so that it will radiate more to the right than to the left, a nice and subtle shading will result, giving the kind of light we might expect to find in a Vermeer painting (very soft yet distinctly coming from the left). The lightspace framework provides a means of synthesizing such "impossible" light sources — light sources that could never exist in reality. Having a "toolbox" containing such light sources affords one with great artistic and creative potential.

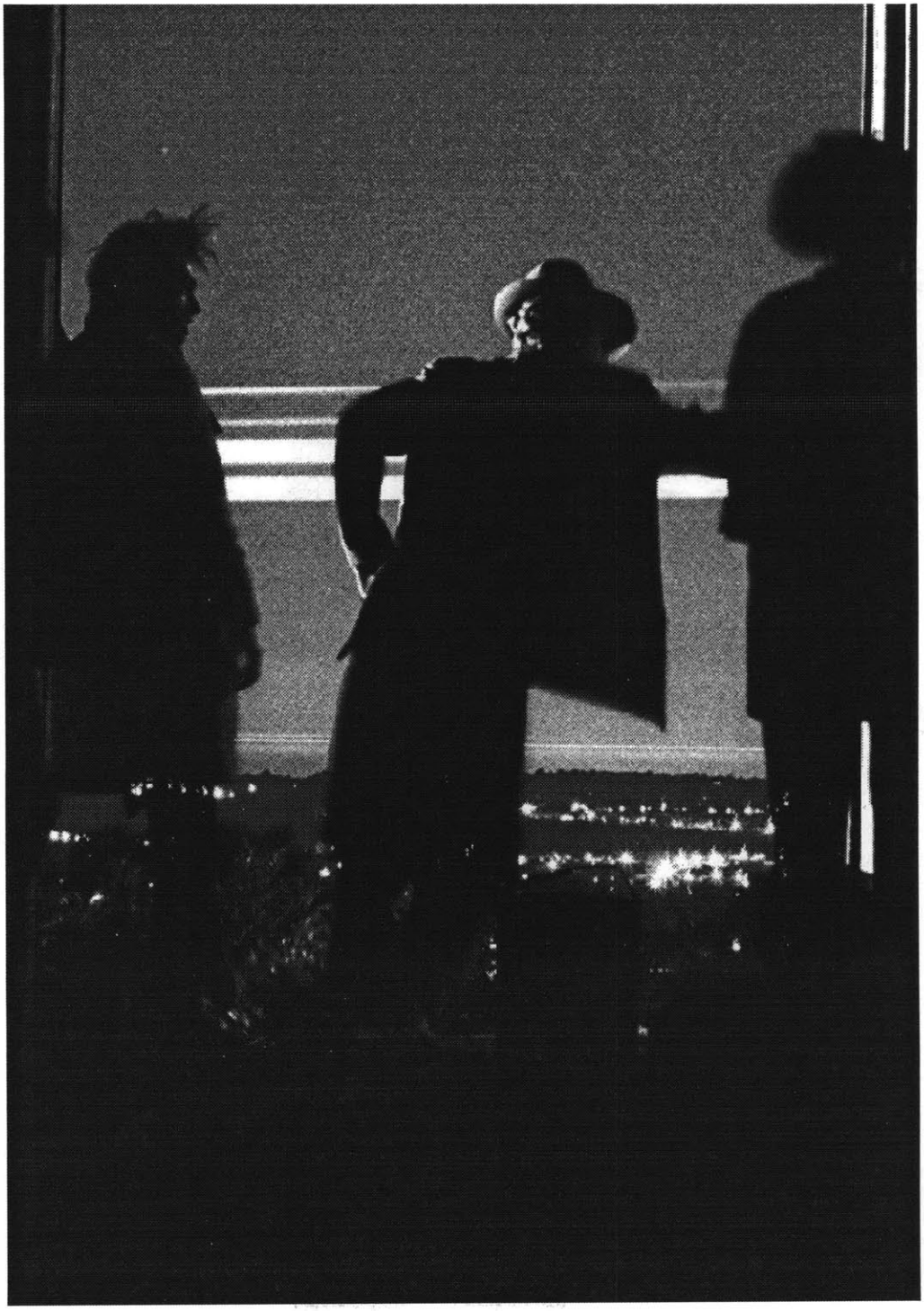


Figure 2-6: Subject matter illuminated, from behind, by linelight. This picture is particularly illustrative because the light source itself (notice the two thick bands, and two thinner bands in the background which are linelights) is visible in the picture. However we see that the three people standing in the open doorway, illuminated by the linelight, are lit on their left side more than on their right side. Also notice how the doorway is lit more on the right side of the picture than on the left side. This *directionality* of the light source is owing from the fact that it is effectively composed of point sources which each radiate mostly to the right. (C) Steve Mann, 1984.

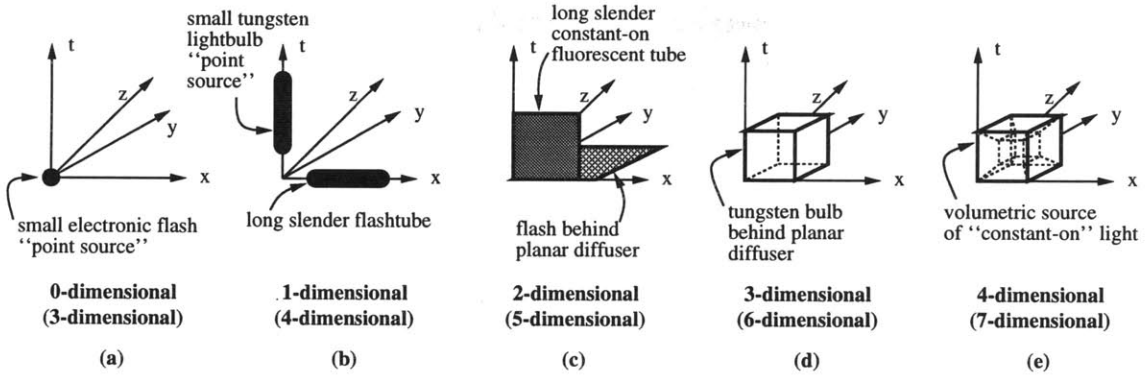


Figure 2-7: A taxonomy of light sources in 4-space. The dimensionality in the 4-space (x, y, z, t) is indicated below each set of examples, while the dimensionality in the new 7-space is indicated in parentheses. (a) A flash point source located at the origin gives a brief flash of white light that radiates in all directions (θ, ϕ) over all wavelengths, λ, and is therefore characterized as having 3 degrees of freedom. A fat dot is used to denote a practical real-world approximation to the point flash source, which has a nonzero flash duration, and a nonzero spatial extent. (b) Both the point source and the lineflash have 4 degrees of freedom. Here the point source is located at the spatial (x, y, z) origin, and extends out along the t axis, while the lineflash is aligned along the x axis. A fat line of finite length is used to denote a typical real-world approximation to the ideal source. (c) A flash behind a planar diffuser, and a long slender fluorescent tube are both approximations to these light sources that have 5 degrees of freedom. (d) Here a tungsten bulb behind a white sheet gives a dense planar array of point sources that is confined to the plane $z = 0$, but spreads out over the six remaining degrees of freedom. (e) A volumetric source, such as might be generated by light hitting particles suspended in the air, radiates white light from all points in space and in all directions. It is denoted as a *hypercube* in 4-space, and exhibits all 7 degrees of freedom in 7-space.

Assembling a number of controllable spotflashes at the same location but pointing in all possible directions in 3-D space, and maintaining separate control of each of them, provides us with a source that can produce any pattern of flash emanating from a given location. I call this light source a 'controllable flash point source'. It is driven by a control signal that is a real-valued function of three real variables, θ_l, ϕ_l, and λ_l.

So far I have said that a flashtube can be activated to flash or to stay on constantly, but, more generally, its output can be varied rapidly and continuously, through the application of a time-varying voltage[15].

the aremac

Similarly, if we apply time-varying control to the 'controllable flash point source' we obtain a controllable point source that I call the 'aremac'. The aremac is capable of producing any bundle of light rays that pass through a given point. It is driven by a control signal that is a real-valued function of four real variables, θ_l, ϕ_l, and λ_l, and t_l. The aremac subsumes the 'controllable flash point source', and the 'controllable spotflash' as special cases. Clearly it also subsumes the 'white spotflash', and the 'flash point source' as special cases.

The aremac is the exact reverse of the idealized pinhole camera. The idealized pinhole camera[16] absorbs and quantifies incoming rays of light and produces a real-valued function of four variables (x,y,t,λ) as output. The aremac takes as input the same kind of function that the idealized pinhole camera gives as output.

The closest approximation to the aremac that one may typically come across is the video projector. A video projector takes as input, a video signal (3 real-valued functions of three variables,

[15] An ordinary tungsten-filament light bulb can also be driven with a time-varying voltage, but due to the time required to heat or cool the filament, it responds quite sluggishly to the control voltage, and the electronic flash is much more in keeping with the spirit of the ideal time-varying lightsource. Indeed, visual artist Joe Davis has shown that the output intensity of an electronic flash can be modulated at video rates, so that it can be used to transmit video to a photoreceptor at some remote location.

[16] The idealized pinhole camera does not exist in practice. The closest practical approximation would be a motion picture camera that implemented the Lippman photography process.

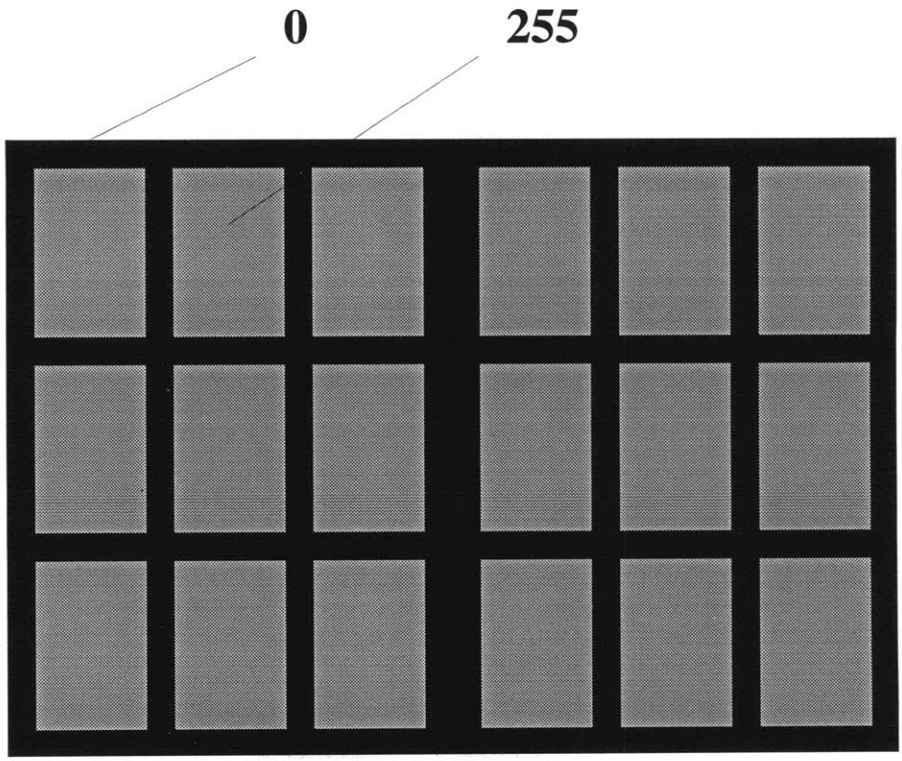


Figure 2-8: Using a computer screen to simulate the light from a window on a cloudy day. All of the regions on the screen that are shaded correspond to areas that should be set to the largest numerical value (typically 255), while the solid (black) areas denote regions of the screen that should be set to the lowest numerical value (0). The light coming from the screen would then light up the room in the same way as a window of this shape and size. This trivial example illustrates the way in which the computer screen can be used as a controllable light source.

x,y, and t). Unfortunately, its wavelength is not controllable, but, it can still produce rays of light in a variety of different directions, under program control, to be whatever color is desired within its limited color gamut, and these colors can evolve with time, at least up to the frame/field rate.

A linear array of separately addressable aremacs produces a 'controllable linesource'. Stacking these one above the other (and maintaining separate control of each) produces a 'controllable sheet-source'. Now if we take uncountably many 'controllable sheetsources' and place them one above the other, maintaining separate control of each, we arrive at a light source that is controlled by a real-valued function of seven real variables, θ_l, ϕ_l, λ_l, t_l, x_l, y_l, and z_l. We call this light source the 'plenoptic aremac'

The 'plenoptic aremac' subsumes all of the light sources that we have mentioned so far. In this sense it is the most general light source — the only one we really need in our conceptual lighting toolbox.

An interesting subset of the plenoptic aremac is the computer screen. Computer screens typically comprise over a million small light sources spread out over a time-varying 2-D lattice (or, if you prefer, think of it as a 3-D lattice in (t_l, x_l, y_l)). Each light source can be separately adjusted in its light output. A greyscale computer screen is driven by a signal that is an integer-valued function of three integer variables: t_l, x_l, and y_l, but the number of values that the function can assume (typically 256) and the fine-grained nature of the lattice (typically over 1000 pixels in the spatial direction, and 72 frames per second in the temporal direction) make it behave almost indistinguishably from a hypothetical device driven by a real-valued function of three real variables. Using the computer screen as a light source, we can, for example, synthesize the light from a particular north-facing window (or cloudy-day window), by displaying the light pattern of that window. This light pattern might be an image array as depicted in Fig 2-8. Taking a picture in a darkened room, where the only source of light is this computer screen, which is displaying the image of Fig 2-8, will give us a

picture similar to what we would obtain using a real window as a light source, when the sky outside is completely overcast. However, because the light from each point on the screen radiates in all directions, we cannot synthesize the sunlight streaming in through a window (because the screen cannot send out a ray that is confined to a particular direction of travel). Thus we cannot use the computer screen to take a picture of a scene as it would appear illuminated by light coming through a window on a sunny day (e.g. with parallel rays of light illuminating the scene or object being photographed).

A plenoptic video system (a device that can produce a spatially organized set of light rays where the direction as well as the position can be controlled) is perhaps the closest thing to the 'plenoptic aremac' that exists in practice. We could use a plenoptic video system to synthesize a wider variety of light sources, such as the sunlight streaming in through a window.

2.4 Plenoptic × plenoptic imaging ("Lightspace")

So far, we have considered the camera as a mechanism for simultaneously measuring many rays of light passing through a single point, that is, measuring a portion of the plenoptic function. As I have mentioned from time to time, knowing the plenoptic function allows us to reconstruct natural-light pictures of a scene, but not pictures taken by our choice of lighting. For example, with the plenoptic function of the setting of Dallas, November 22, 1963, we cannot construct a picture equivalent to one that was taken with a *fill-flash*[17]. The reason for this limitation lies in the structure of the "plenoptic function".

Though the "plenoptic function" is a very information-rich scene description, and might appear to give us far more visual information than we could ever hope to use, an even more complete scene characterization, which I call "lightspace", is now suggested.

Lightspace attempts to characterize everything that can be known about the way that light can interact with a scene. Knowing the lightspace of the setting of Dallas, November 22, 1963, for example, would allow us to synthesize a picture that had been taken with flash, or to synthesize a picture taken on a completely overcast day (despite the fact that the weather was actually quite clear that day), obtaining, for example, a completely shadow-free picture of the gunman on the grassy knoll, (even though, in reality, he had been standing there in bright sunlight, with harsh shadows).

We define lightspace as the set of all plenoptic functions measured with each possible ray of light that we can shine onto the scene. Thus the lightspace consists of a plenoptic function located at every point in the seven dimensional space $(\theta_l, \phi_l, \lambda_l, t_l, x_l, y_l, z_l)$. Thus, equivalently, the lightspace is a real-valued function of 14 real variables. The lightspace may be evaluated at a single point in this 14-D space using a spot-flash-spectrometer and spotflash, as depicted in Fig 2-9.

2.4.1 Upper-triangular nature of lightspace along two dimensions (Fluorescent and phosphorescent objects)

Not all light rays sent out will return. Some may pass by the scene and travel off into 3-space. The lightspace corresponding to these rays will thus be zero. Therefore, it is clearly possible that a ray of light sent out at a particular point in 7-space, $(\theta_l, \phi_l, \lambda_l, t_l, x_l, y_l, z_l)$ does not arrive back at the location of the sensor in 7-space, $(\theta, \phi, \lambda, t, x, y, z)$.

A good practical example of zero-valued regions of lightspace arises when the light reading is taken before the ray is sent out. This situation is depicted mathematically as $t < t_l$ (See also Fig 2-10(a).) Similarly, if we shine red light ($\lambda = 700nm$) on the scene, and look through a blue filter ($\lambda = 400nm$), we would not expect to see any response. In general, then, the lightspace will be zero whenever $t < t_l$ or $\lambda < \lambda_l$.

However, if we flash a ray of light at the scene, and then look a few seconds later, we may still pick up a non-zero reading. Consider, for example, a glow-in-the-dark toy (or clock), a computer

[17] Even on a bright sunny day, a small flash helps to fill in some of the shadows and results in a much improved picture. Such a flash is called a *fill-flash*.

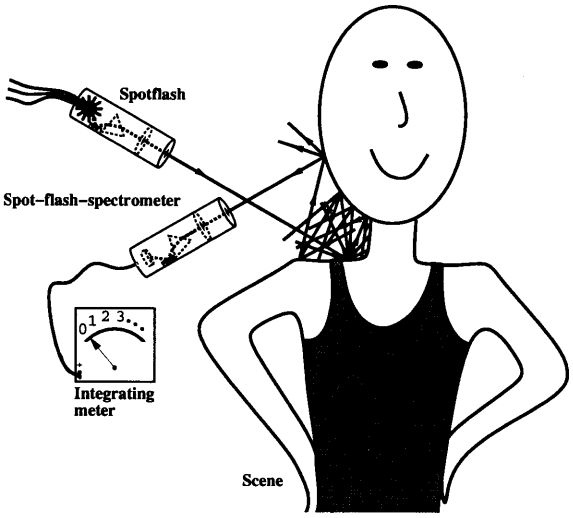


Figure 2-9: Measuring one point in the lightspace around a particular scene, using a spot flash-spectrometer and a spotflash. The measurement provides a real-valued quantity that indicates how much light comes back along the direction (θ, ϕ), at the wavelength λ, and time t, to location (x, y, z), as a result of flashing a monochromatic ray of light in the direction (θ_l, ϕ_l), having a wavelength of λ_l, at time t_l, from location (x_l, y_l, z_l).

screen, or a TV screen. Even though it might be turned off, it will glow for a short time after it is excited by an external source of light, due to the presence of phosfluorescent materials. Thus the objects can absorb light at one time, and re-radiate it at another.

Similarly, some objects will absorb light at short wavelengths (such as ultraviolet or blue light) and re-radiate at longer wavelengths. Such materials are said to be *fluorescent.* A fluorescent red object might, for example, provide a nonzero return to a sensor tuned to $\lambda = 700nm$ (red), even though it is illuminated only by a source at $\lambda = 400nm$ (blue).

Thus, along the time and wavelength axes, lightspace is 'upper triangular'[18] (Fig 2-10(b)).

2.5 Lightspace subspaces

In practice, the lightspace is too unwieldy to work with directly and is, instead, only useful as a conceptual framework in which to pose other practical problems. In particular, as I mentioned earlier, rather than using a spotflash and spot-flash-spectrometer to measure lightspace, we will most likely use a camera. A video camera, for example, can be used to capture an information-rich description of the world, and, in a sense, provide many measurements of the lightspace.

In practice, the measurements we make with a camera will be more crude than those made with the precise instruments (spotflash and spot-flash-spectrometer), but the camera makes a large number of measurements in parallel (at the same time). The crudeness of each measurement is expressed by integrating the lightspace together with some kind of 14-D blurring function. For example, a single greyscale picture, taken with a camera having an image sensor of dimension 480 by 640 pixels, taken with a particular kind of illumination, may be expressed as a collection of 480*640=307200 crude measurements of the light rays passing through a particular point. In particular, each of these measurements corresponds to a particular sensing element of the image array, which is sensitive to a range of azimuthal angles, θ and elevational angles, ϕ. Each reading is also sensitive to a very broad range of wavelengths, and the shutter speed dictates the range of time that the measurements are sensitive to. Thus the blurring kernel will completely blur the λ axis, sample the time axis at a single point, and somewhat blur the other axes. A color image of the same dimensions will provide three times as many readings, each one blurred quite severely along the wavelength axis, but not so severely as the greyscale image readings. A color motion picture will

[18] a term borrowed from linear algebra that denotes matrices with entries of zero below the main diagonal

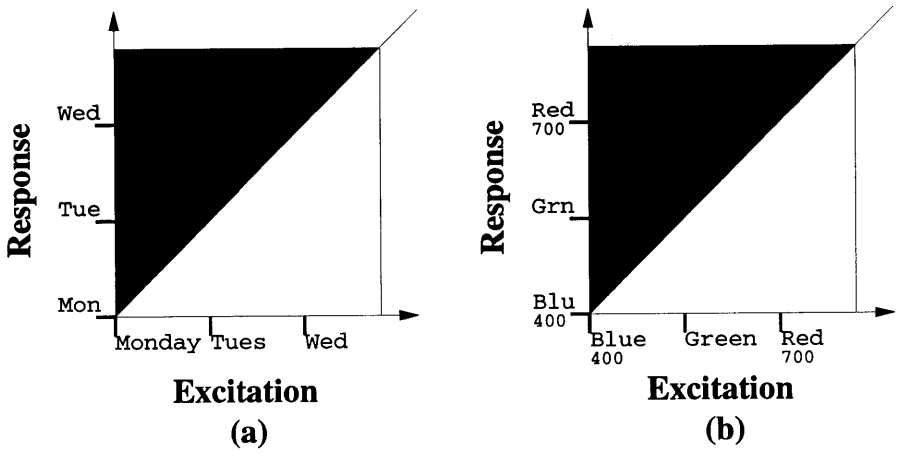


Figure 2-10: In practice, not all light rays that are sent out will return to the sensor. (a) In particular, if we try to measure the response before the excitation, we would expect a zero result. (b) Many natural objects will radiate red light as a result of blue excitation, but the reverse is not generally true. (These are 'density plots', e.g. black indicates true.)

represent a blurring and repetitive sampling of the time axis. A one second (30 frames per second) movie then provides us with $3*30*480*640 = 27648000$ measurements (e.g. 27000K of data), each blurred by the nature of the measurement device (camera).

We can trade some resolution in the camera parameters for resolution in the excitation parameters, by, for example, having a flash activated every second frame so that half of the frames are naturally-lit, and the other half are flash-lit. Using multiple sources of illumination in this way, we could attempt to characterize the lightspace of the scene in a crude way. Each of the measurements is given by:

$$\begin{aligned} l_k = \int\int\int\int\int\int\int\int\int\int\int\int\int\int & \\ \mathcal{L}(\theta_l, \phi_l, \lambda_l, t_l, x_l, y_l, z_l, \theta, \phi, \lambda, t, x, y, z) & \\ \mathcal{B}_k(\theta_l, \phi_l, \lambda_l, t_l, x_l, y_l, z_l, \theta, \phi, \lambda, t, x, y, z) & \\ d\theta_l d\phi_l d\lambda_l dt_l dx_l dy_l dz_l d\theta d\phi d\lambda dt dx dy dz & \end{aligned} \tag{2.1}$$

where $\mathcal{L}$ is the lightspace and $\mathcal{B}_k$ is the blurring kernel of the kth measurement apparatus (incorporating both excitation and response).

We may re-write this measurement (in the Lebesgue sense rather than the Reimann sense to avoid writing out all the integral signs):

$$l_k = \int \mathcal{L}(\theta_l, \phi_l, \lambda_l, t_l, x_l, y_l, z_l, \theta, \phi, \lambda, t, x, y, z) d\mu_k \tag{2.2}$$

where μ_k is the measure associated with the blurring kernel of the kth measuring apparatus.

We will refer to a collection of such blurred measurements as a lightspace subspace, for it fails to capture the entire lightspace. The lightspace subspace, rather, slices through portions of lightspace (decimates or samples it), and then blurs those slices.

2.6 'Lightvector' subspace

A vector, v, of length L, may be regarded as a single point in a multi-dimensional space, $\mathbb{R}^L$. Similarly, a real-valued greyscale picture defined on a discrete lattice of dimensions M (height) by N (width), may also be regarded as a single point in $R^{M \times N}$, a space that I will call 'imagespace'. Any image can be represented as a point in imagespace because the picture may be unraveled into

one long vector, row by row.[19] Thus, if we linearize, using the procedure of Section 3.5, then, in the ideal noise-free world, all of the linearized elements of a Wyckoff set, $L_n = r^{-1}W_n$, are linearly related to each other through a simple scale factor:

$$L_n = k_n L_0 \tag{2.3}$$

where L_0 is the linearized reference exposure. In actual practice, image noise prevents this from being the case, but conceptually, the camera and Wyckoff film provide us with a means of obtaining the one dimensional subspace of the imagespace.

2.6.1 "Practical" example: 2-d lightvector subspace

Suppose we take pictures of a particular static scene, with a fixed camera position (e.g. on a tripod), but vary only the lighting. Suppose we further restrict the light sources to not change in shape, position, orientation, or color, but allow them to only vary in intensity.

For the moment consider two light sources, each of adjustable intensity. The most common practical example of two light sources arises when a photographer uses an on-camera flash to supplement the natural light in the scene. In this example, we could vary the effective natural light intensity by adjusting the exposure time (e.g. by making one exposure for 1 second, and another for 2 seconds, we've doubled the amount of light arriving at the image sensor). We can vary the flash intensity (by its main capacitor or thyristor control). Since we have two degrees of freedom (shutter speed and flash output), we can obtain images from various linear combinations of the two light sources (Fig 2-11).

Consider now a particular example, namely a discrete image sensor of dimension 480 by 640 (pixels). The images that this sensor provides are 480 by 640 arrays, and each element of the array is monotonically proportional to the quantity of light falling on a particular element of the sensor. Assuming that each element of the array is a continuously varying (real) quantity, we have, after linearizing the array, for each image, a point in $\mathbb{R}^{480\times640} = \mathbb{R}^{307200}$.

Now if we take a variety of pictures where only the exposure (shutter speed) is varied, we have only one degree of freedom. With an infinitesmally short shutter speed (e.g. zero exposure), all of the elements of the array will be zero, and so, this image will lie at the origin of the 307200-D space. A picture taken with an exposure of 1 second will land at some arbitrary point in the 307200-D space. A picture taken with an exposure of 2 seconds will lie further out along the same axis defined by the first exposure. More generally, any number of images that belong to a Wyckoff set (differ only in exposure) lie along the same axis (Fig 2-12). Similarly a picture of the same scene, with the same camera location, but taken with a flash (using a short enough exposure that the only light arriving back at the image sensor is that due to the flash), produces an image array with different numerical values. This array of different numerical values lies somewhere else in $\mathbb{R}^{307200}$. A picture taken with more or less flash illumination lies along the same coordinate axis as the first flash picture (Fig 2-12).

Because light intensity is additive, if we take a picture with a combination of natural light and flash (e.g. if we leave the shutter open for an extended period of time, and fire the flash sometime during that exposure), we obtain a point in $\mathbb{R}^{307200}$ that lies on the plane defined by the two (natural light and flash) pure lightvectors. Various combinations of natural light with flash are depicted in this way in Fig 2-13.

I have made the assumption that, for all practical purposes, the flash duration is infinitesmally short, so that if a picture of the scene as illuminated by only the flash, is desired, then we can obtain it by using a very short exposure (e.g. short enough that no natural light will affect the image). However, to the extent that this might not be exactly true[20], we may apply a coordinate transformation that shears the lightvector subspace. In particular, we, in effect, subtract out the

[19] This is indeed the way that a picture (or any 2-D array) is typically stored in a file on a computer – sequentially in one dimension.

[20] In practice, this happens with large studio flash systems because the capacitors are so big that the flash duration actually starts getting a little long, on the order of 1/60 second or so.

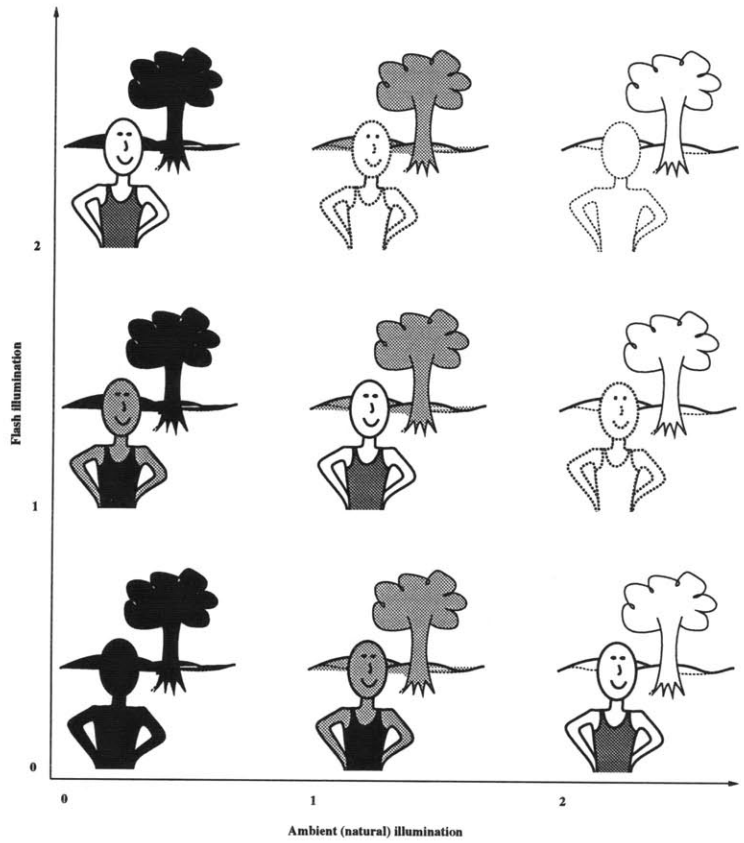


Figure 2-11: Hypothetical depiction of a collection of pictures taken with different amounts of ambient (natural) light and flash. One axis denotes the quantity of natural light present (exposure), and the other denotes the quantity of flash, activated during the exposure. The flash adds to the total illumination of the scene, and affects primarily the foreground objects in the scene, while the natural light exposure affects the whole scene. These pictures form a 2-D image subspace, which may be regarded as being formed from the dimensions defined by two "lightvectors". For argument's sake we might define the two "lightvectors" as being the image at coordinates (1,0) and the image at (0,1). These two images form a basis set that may be used to generate all nine depicted here in this figure, as well as any other linear combination of the two lightvectors.

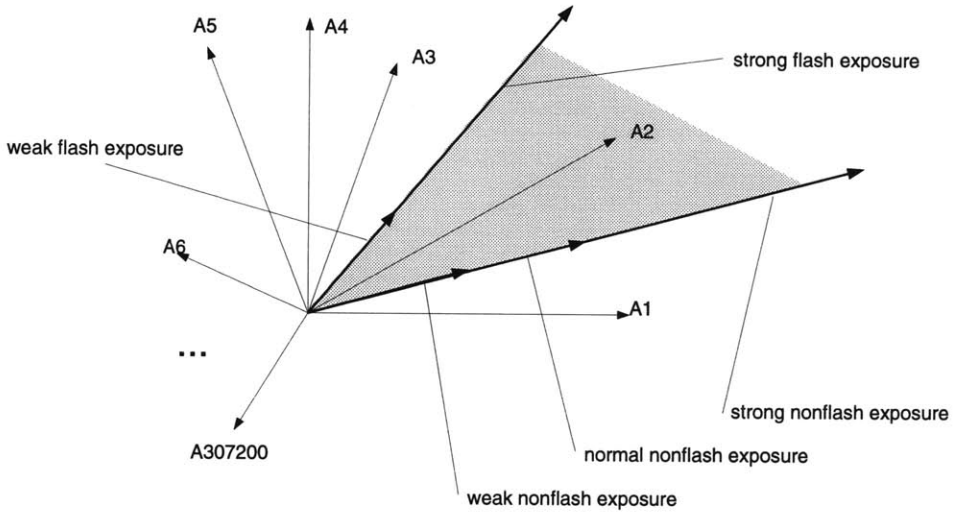


Figure 2-12: A picture sampled on a discrete 480 by 640 lattice may be regarded as a single point in $\mathbb{R}^{480\times640} = \mathbb{R}^{307200}$. A Wyckoff set formed by 3 pictures, differing only in exposure, is depicted as 3 collinear arrows coming out from the origin. These 3 pictures were taken with no flash (just the natural light present in the scene). Another two pictures taken with a flash, and with such a high shutter speed that they are representative of the scene as lit by only the flash, are also depicted, as two colinear arrows from the origin. The 2-D subspace formed from these two Wyckoff sets (natural light set and flash set) is depicted as a shaded grey planar region.

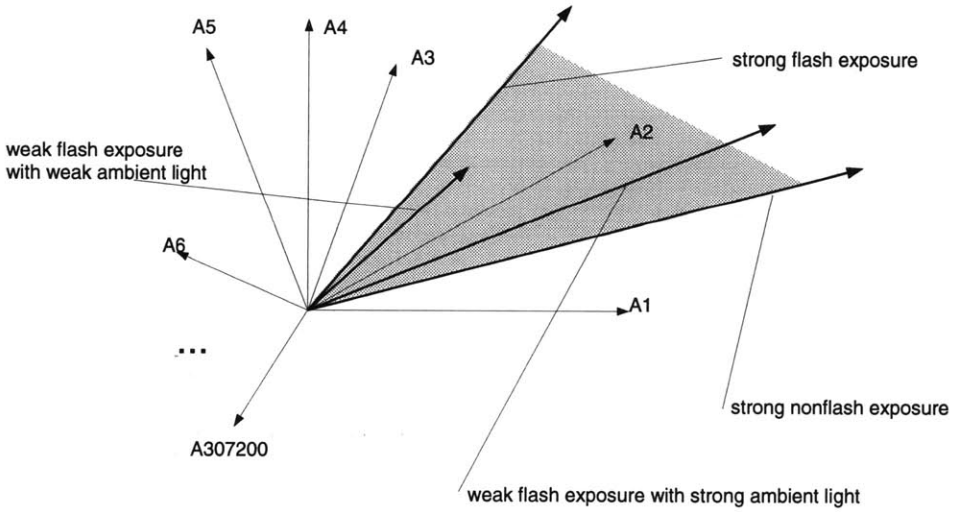


Figure 2-13: The 2-D subspace formed from various pictures each taken with different combinations of the two light sources (natural and flash). All of these pictures exist as points in $\mathbb{R}^{307200}$ that lie within the planar (2-D) subspace of Fig 2-12. Note that since light intensity cannot be negative, that this subspace is confined to the portion of the plane bounded by the two pure lightvectors (natural light and flash).

effect of the natural light, and obtain a "total-illumination" axis, so that the images of Fig 2-11 move to new locations (Fig 2-14) in the imagespace.

In real practice

In actual practice, the coordinate transformation depicted in these hypothetical cartoon characters may not be directly applied to actual pictures, because actual pictures have undergone some unknown nonlinearity. Actual pictures do not represent the quantity of light falling on the image sensor or the like, but, rather, represent some unknown (but typically monotonic) function of this light quantity. The human visual system is very forgiving of distortion in the greyscale levels of images; unlike sound recordings where nonlinear distortion is very objectionable, nonlinear distortion in pictures is quite acceptable, even desirable. Typically pictures look better when the contrast is high, and often they look better when highlight details are clipped, and shadow details are similarly lost. Photographers often describe images where the contrast is too low as lacking "punch" or "kick", even though they may contain more information about the scene than the pictures they have come to prefer.

What we desire, for the purposes of this thesis, however, is a means of capturing, from ordinary pictures, as accurately and completely as possible, the manner in which the scene responds to light. In the next chapter, we will see how this is accomplished through the process of using ordinary cameras as measuring instruments to determine, up to a single unknown scale factor, the quantity of light falling on the image sensor, due to a particular source of illumination. Even when the scene has an extremely high dynamic range (as most scenes do), we will still be able to determine the response to light.

Even if we only desire a nice looking picture in the end, the picture may be generated by suitable deliberate degradation and nonlinear distortion of the final measurement space, once we have done the processing in the linearized measurement space. I refer to such processing (e.g. processing on a calculated value of the quantity of light falling on the image sensor, followed by returning to a normal picture) as 'homomorphic imaging'.

Because cameras seldom actually measure light, in a way that is directly usable in the context of lightspace, that the linearization procedure presented in Chapter 3 is so important. In particular, I will propose the 'Wyckoff principle' as a framework for determining (up to a single unknown constant, which is linear on the response function of the camera and therefore of no consequence from a visual point of view) and working with the quantity of light falling on the image sensor.

2.7 Chapter summary

I have described a conceptual framework called 'lightspace', which characterizes the response of arbitrary scenes or objects to light. Mathematically, lightspace is the *vector outer product* (tensor product) of the plenoptic function with itself.

I have made use of the linearity and superposition properties of light intensity to define the structure of lightspace.

Since the lightspace of a scene is a complete description of the way that it responds to light, from the lightspace, we can synthesize any picture that might have been taken of the scene using any artificial light of our choice.

I have also identified the 'lightvector subspace' resulting from pictures taken with various fixed light sources that vary only in the quantity of light that they produce.

While the concept of lightspace is a hypothetical framework in which practical measurement spaces must exist as subspaces (due to the large amounts of data involved), in the next chapter, we will see how the properties of lightspace may be used in some practical applications. In particular, the general philosophy of Chapter 3 will be to use an array of image sensors, with lens (which happens to take the form of a video camera built into the computer glasses described in Chapter 1), as a collection of light meters.

The unknown nonlinearity of the sensor array and digitizer will be determined, up to a constant scale factor, and thus an ordinary video camera will be turned into a set of light-measuring instruments.

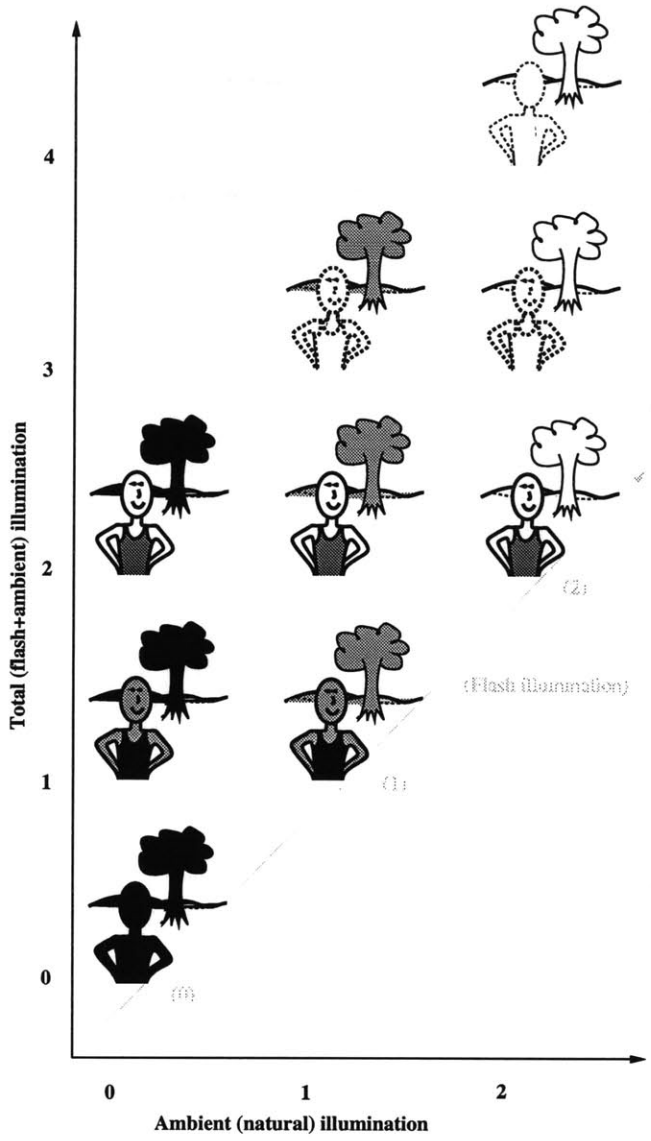


Figure 2-14: A coordinate-transformed 2-D image subspace, formed from the dimensions of two lightvectors (natural light and on-camera flash). The new axes, natural-light and total-light, define respectively the light levels of the background and foreground objects in the scene.

Later in this thesis, these concepts will be explored in a practical imaging situation, involving automatic gain control (AGC).

Chapter 3

Homomorphic Imaging: The camera and the range of light

In this chapter, I report a theoretical framework I completed in 1992 [23]. While much has happened since then, this framework remains as the basis upon which personal imaging and the rest of this thesis has evolved.

Briefly, what I mean by 'homomorphic imaging' is the process of applying the linearity and superposition [1] properties of light to images, by regarding the image, $f(q)$, as some nonlinear (typically unknown) function, f, of the quantity of light falling on the image sensor, q.

The word "homomorphic" already has at least two meanings:

1. its mathematical (algebraic) meaning; a "homomorphism" is a mapping from a set upon which there exists a group structure, to a new set upon which there also exists a group structure, such that the mapping preserves the group structure: $f(q_0)\Box f(q_1) = f(q_0 \circ q_1)$, where $\circ$ is the law of composition in the domain, and $\Box$ is the law of composition in range of f.

2. its meaning within the signal processing community; "homomorphic filtering" refers to the process of applying a nonlinearity, followed by linear filtering, followed by an inverse nonlinearity. The most common example is that of applying a logarithmic function, followed by linear filtering, followed by applying an antilogarithmic (exponential) function [24].

My use of this term combines both of these meanings, in the sense that the quantity of light, q_1, q_2, ..., falling on the image sensor, due to separate sources, e.g. numbered 1, 2, ..., is additive, while the corresponding pictures resulting from this light falling on the image sensor may be combined in a certain way that gives a result equivalent to that which would be taken when all of the combined light falls on the image sensor, $q_1 + q_2 + \ldots$. Thus I denote $f(q_1 + q_2 + \ldots) = f(q_1)\Box f(q_2)\ldots$ as this means of combining multiple, differently illuminated pictures of the same scene or object.

In addition to affording a law of composition of multiple pictures, differing only in exposure or illumination, I propose new forms of filtering, that is, I suggest that the most natural range in which images should be filtered is the linear range of the light that originally fell on the image sensor.

Stockham [24] proposed that, prior to filtering images, one should take their logarithms, perform the filtering, and then take antilogs of the result. My empirical observation has been that it is actually preferable, when f is not known, to do exactly the opposite of what Stockham has advocated, namely I found it preferable to first compute an antilog function, then work with (e.g. whether filtering or combining images) $\exp(f(q))$, etc., and then apply a logarithmic function to the result.

My empirical findings of this nature also match our intuition regarding film, that is, that the density of the film is generally proportional (at least over a certain region) to the logarithm of the

[1] In this thesis, I use the term "linearity" to denote that $f(kq) = kf(q) \forall k \in \Re$ and the term "superposition" to denote that $f(q_0 + q_1) = f(q_0) + f(q_1)$. Some authors use the term "linearity" to denote both these properties, but I prefer to make the distinction.

exposure, so that if the film scanner reports density, then we have (approximately) the logarithm of q. (Or in the case of an electronic camera, the entire process may also be so modeled.)

Although I found it better to do exactly the opposite of what Stokham advocated, in the absence of knowing f, I found a simple means of determining f, and suggest that the estimate of f^{-1} is the preferable function to use.

3.1 Introduction

Most everyday scenes have a far greater dynamic range than can be recorded on a photographic film or electronic imaging apparatus (whether it be a digital still camera, consumer video camera, or eyeglass-based personal imaging apparatus as described in Chapter 1). However, a set of pictures, that are identical except for their exposure, collectively show us much more dynamic range than any single picture from that set.

The dark pictures show us highlight details of the scene that would be washed out in a "properly exposed" picture, while the light pictures show us some shadow detail that would also not appear in a "properly exposed" picture.

I propose a means of combining differently exposed pictures to obtain a single picture of extended dynamic range, and improved color fidelity. Given a set of digital pictures, I produce a single picture which is, for all practical purposes, 'undigital', in the sense that it is a floating point image, with the kind of dynamic range we are accustomed to seeing in typical floating point representations, as opposed to the integer images from which it was generated.

The method is completely automatic; it requires no human intervention, and it requires no knowledge of the response function of the imaging device. It works reliably with images from a digital camera of unknown response, or from a scanner with unknown response, scanning an unknown film type.

3.2 Being undigital

Digital photography allows us to do many things we cannot do with traditional analog photography. However, being digital is not desirable in and of itself – it is desirable for what it facilitates (instant feedback, ability to rapidly transmit high quality anywhere in the world, ease of manipulation, etc).

Digital imaging imposes certain limitations on the ways we think about images. Ideally, what we want is not bits, but, rather, a mathematical or parametric representation of the continuous underlying intensity variations projected onto an image plane, represented in a form that allows for easy transmission, storage, and analysis.

As the spatial resolution of digital images has improved over the years, we are approaching a level where the image may be regarded as essentially continuous – it is essentially free of *pixels*. Thus high resolution digital images give us the spatial continuity of analog photography, together with the ability to view pictures right away, transmit them over wireless links, analyze them computationally, etc.

However, while there may be so many pixels that we can, for all practical purposes, assume the image is a function of two real coordinates, each of these pixels is still typically represented as a triplet of integers each of which can assume typically only 256 different values, for each color channel[2]. So-called *24 bit color*, also known as *full color*, *true color direct visual*, etc., is not as "full" or "true" as these names imply. In particular, these images are also typically manipulated using 8-bit precision arithmetic. Any simple manipulations in an image editing program, such as Photoshop, quickly degrade the quality of the images, introducing gaps in the histograms that grow with each successive computation.

[2] Although many cameras and digitizers capture higher definition images internally, they generally only produce 8 bits per color channel on the output. For example, the Kodak PhotoCD scanner scans at 14 bits (for each color channel) internally, but then applies a nonlinearity to this data, and provides the user with only 8 bits per color channel.

The purpose of this chapter is to examine the recovery of the 'true image', a real-valued quantity of light projected onto a flat surface. We regard the 'true image' as a collection of **analog** photometric quantities that might have been measured with an array of linearized lightmeters having floating-point precision, and thus, being essentially, for all practical purposes, 'undigital'.

Of course, all images that are stored on a computer are digital. A floating point number is digital. But a double-precision (64 bit) floating point number is close to analog in spirit and intent.

With the growing word size of desktop computational hardware, floating point arithmetic is becoming more practical for large images. The new DEC 3000 (Alpha) computer has a word size of 64 bits, and can easily handle images as double precision arrays. Double precision is nothing new. For years, languages like FORTRAN have supported floating point arithmetic, used widely by the scientific community, but floating point calculations are not supported in any of the popular image manipulation software such as Photoshop or Live picture. Capturing images that are essentially unlimited in dynamic range, and, while digitally represented, behave as analog images, allows us to capture and surpass the benefits traditionally offered by truly analog image formats like film.

3.3 What is a camera

In this chapter, I will show how an image may be regarded as a collection of photometric measurements, and a camera as an array of light meters. Part of this work will involve dealing with the fact that in most cameras, there is an unknown nonlinearity. Therefore, in the context of the proposed philosophy, I might say that each of these measurements (pixels) are made with a light meter (sensor element) that has some unknown nonlinearity followed by a quantization to a measurement having typically 8-bit precision.

3.3.1 Dynamic range and amplitude resolution

Many everyday scenes contain a tremendous dynamic range. For example, the scene might be a dimly lit room, with a window in the background; through the window we might observe a beautiful blue summer sky with puffy white clouds. Yet a picture that is exposed for the indoor scene will render the window as a white blob, blooming out into the room, where we can scarcely discern the shape of the window, let alone, see beyond it. Of course, if we exposed for the sky outside, the interior would appear completely black.

Cameras (whether analog or digital) tend to have a very limited dynamic range. It is possible to extend the dynamic range by various means. For example, in the case of photographic emulsion, the film can be made thicker, but there are tradeoffs (e.g. thicker emulsion results in increased scattering, which results in decreased spatial resolution). Nyquist showed how a signal can be reconstructed from a sampling of finite resolution in the domain (e.g. space or time), but assumed infinite dynamic range. On the other hand, if we have infinite spatial resolution, but limited dynamic range (even if we have only 1 bit of image depth), Curtis and Oppenheim [25] showed that we can also obtain perfect reconstruction. This tradeoff between image resolution, and image *depth* is also at work in a slightly different way in image *halftoning*.

Before the days of digital image processing, Charles Wyckoff formulated a multiple layer photographic emulsion [26][27]. The Wyckoff film had three layers that were identical in their spectral sensitivities (each was roughly equally sensitive to all wavelengths of light), and differed only in their overall sensitivities to light (e.g. the bottom layer was very *slow*, with an ISO rating of 2, while the top layer was very *fast* with an ISO rating of 600).

A picture taken on Wyckoff film can both record a dynamic range of one to a hundred million and capture very subtle differences in exposure. Furthermore, the Wyckoff picture has very good spatial resolution, and thus **appears** to overcome the resolution/depth tradeoff, by using different color dyes in each layer, which have a specular density as opposed the diffuse density of silver. Wyckoff printed his *greyscale* pictures on color paper, so the *fast* (yellow) layer would print blue, the medium (magenta) layer would print green, and the *slow* (cyan) layer would print red. His result was a *pseudo-color* image similar to those used now in data visualization systems to display floating point arrays on a computer screen of limited dynamic range.

Wyckoff's most well-known pictures are perhaps his motion pictures of nuclear explosions – one could clearly see the faint glow of a bomb just before it exploded (which would appear as blue, since it only exposed the fast top layer), as well as the details in the highlights of the explosion (which appeared white since they exposed all 3 layers – the details discernible primarily on account of the slow bottom layer).

3.3.2 Combining multiple pictures of the same scene

The idea of computationally combining differently exposed pictures of the same scene to obtain extended dynamic range has been recently proposed [28], where the images were assumed to have been taken from roughly the same position in space, with possibly different camera orientations (pan, tilt, rotation about optical axis), and different zoom settings. In this chapter, I describe, in further detail, the computational means of combining differently exposed pictures into a floating-point image array, and assume a simpler case, namely that all pictures are taken from a camera at a fixed location in space and a fixed orientation, with a fixed focal length lens. This simpler case corresponds to pictures that differ only in exposure.

I refer to a collection of pictures that differ only in exposure as a *Wyckoff set*, in honor of Charles Wyckoff, who was the first to exploit such a set of pictures collectively, by using pseudocolor to display what was essentially an image containing no color information[3].

Photographers, through a procedure called *exposure bracketing* (trying a variety of exposure settings and later selecting the one exposure that they most prefer) also produce Wyckoff sets but generally with the intent of later merely selecting the best image from the set, without exploiting the potential value of using the differently exposed images collectively.

3.4 Exposure bracketing of digital images

Whenever the dynamic range of the scene exceeds the range of the recording medium (which is almost always) photographers tend to expose for areas of interest in the scene. For example, a scene containing people is usually exposed to show the most detail in them (Fig. 3-1) at the expense of details elsewhere in the scene. Additionally, in our case, a picture was taken immediately afterward (Fig. 3-2), with four times the exposure time, so that the surrounding contextual details of the scene would show up nicely.

Ideally, only one picture would be needed to capture the entire dynamic range of the scene, and we wouldn't even need to worry about whether the picture was overexposed or underexposed because we could lighten or darken it later on, by simply using the appropriate 'lookup operator'. By 'lookup operator', I mean any spatially invariant nonlinearity: $g(x,y) = g(f(x,y))$. A 'lookup operator' is the continuous analog of a *lookup table*. Gamma correction is an example of a 'lookup operator'.

However, due to various noise sources, such as quantization noise, a 'lookup operator' will only be able to compensate for a very limited amount of overexposure or underexposure. For example, we will never recover the detail in the faces of the people from Fig. 3-2. The increased exposure has caused this information to be lost by the combined effect of saturation and noise. Similarly, nothing can be done to recover the shadow details in the darker portions of Fig. 3-1, because these areas have pixel values that are uniformly zero. Even in slightly brighter areas, where there is variation in the pixels, this variation is subject to extreme quantization noise. For example, in dark areas where the pixel values fluctuate between zero and one, there is only one bit of precision. A camera with a small number of bits of depth (such as a one-bit camera), but which has very high spatial resolution, may be used to capture a continuous tone image [25]. Indeed, a stat camera, used in a photo mechanical transfer (PMT) machine, is able to capture images that appear to be continuous-tone (due to the *halftoning screen*), even though the film can only record two distinct levels. This is possible because

[3] In this regard, Wyckoff may also be regarded as the father of so-called pseudocolor — using color images to display greyscale image information that has a much greater dynamic range than could otherwise be displayed on the color display medium.

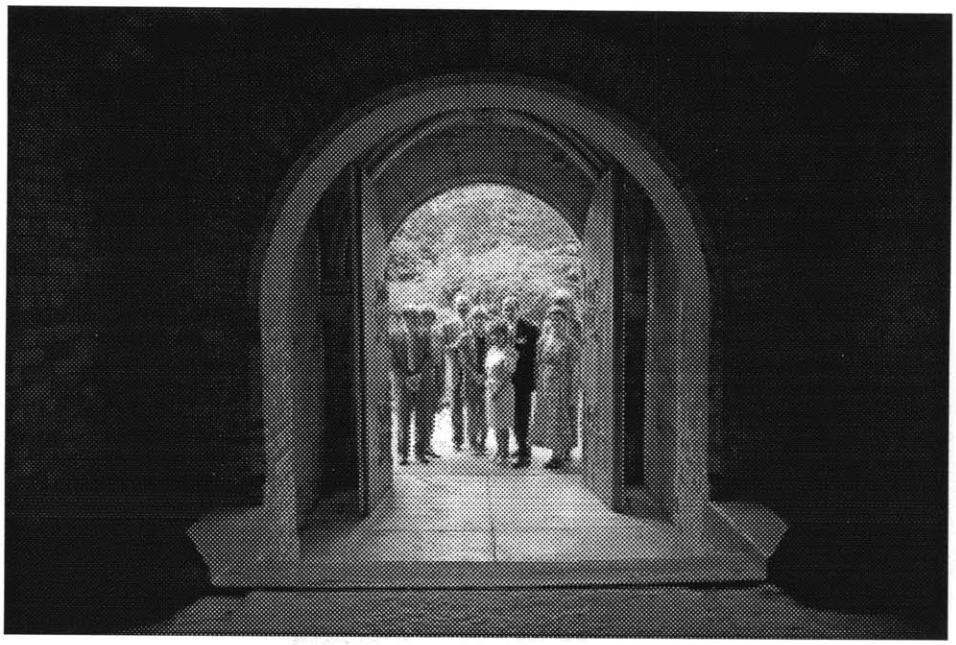


Figure 3-1: The Mann family standing outside an old building with the camera inside. Here the exposure was selected so that the people would show up nicely.

the film has essentially unlimited spatial resolution, and is recording through a screen of much lower (e.g. 85dpi) spatial resolution.

However, in most digital photography and video applications, spatial resolution is much lower than we would like. We do not have the luxury of tremendously high spatial resolution that photomechanical transfer systems have, and so we are not at liberty to trade spatial resolution for improved dynamic range.

Therefore, I propose the use of exposure bracketing as an alternative, whereby I make the tradeoff along the time axis, exchanging reduced frame-rate for improved dynamic range, rather than reduced spatial resolution for improved dynamic range. In particular, often a still image is all that is desired from a video camera, and in many other digital video applications, all that is needed is a few frames per second, from a camera capable of producing 30 frames per second or more.

3.5 Self-calibrating camera: Enforcing linearity

The numerical quantity appearing at a pixel in the image is seldom linearly related[4] to the quantity of light falling on the corresponding sensor element. In the case of an image scanned from film, the density of the film varies nonlinearly with the quantity of light to which it is exposed. Furthermore, the scanner will most likely introduce a further unknown nonlinearity. In this section, I propose two broad classes of methods for estimating the unknown nonlinearity, given pictures that differ only in exposure. In both cases, we know nothing about the scene content, except that it remained constant while the two or more differently exposed images were captured.

3.5.1 Non parametric self-linearizing methods

I propose a simple algorithm for finding the pointwise nonlinearity of the entire process, f, that maps the light q projected on a point in the image plane to the pointwise value in the picture, $f(q)$,

[4] In fact, quite often, photographers **desire** a nonlinear relationship: the nonlinearities tend to make the image look better when printed on media that have limited dynamic range.

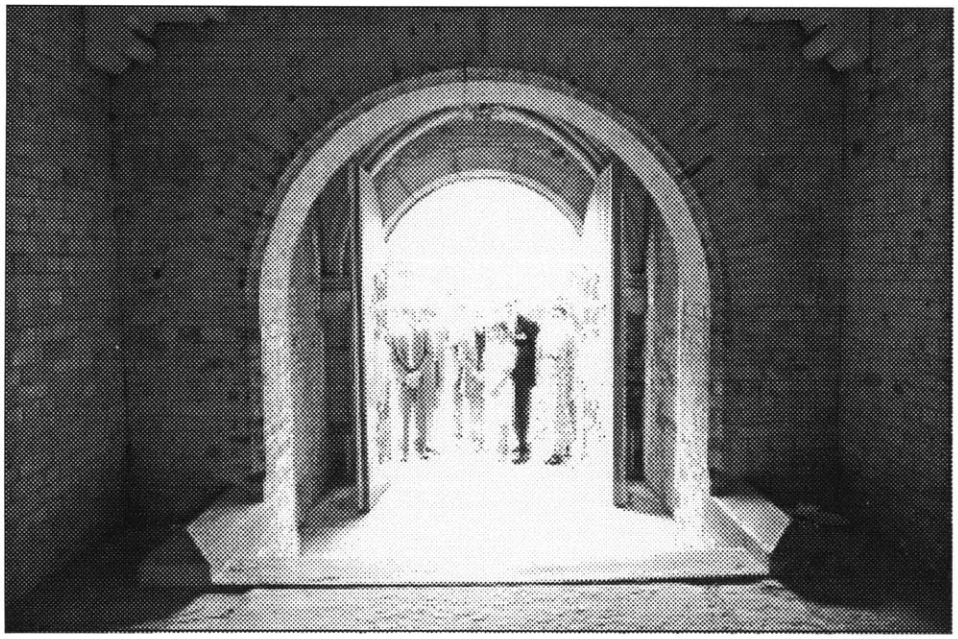


Figure 3-2: The exposure was increased by a factor of $k = 4$, compared to Fig. 3-1; as a result, the interior of the building is nicely visible.

up to a constant scale factor. I ignore, until Section 3.5.2, the fact that each pixel can only assume a finite number of values, the fact that there are a finite number of pixels in the image, and the effects of image noise. The algorithm proceeds as follows[5]:

1. Select a relatively dark pixel from image a, and observe both its location, (x_0, y_0), and its numerical value, f_0. We do not know the actual quantity of light that gave rise to f_0, but I will call this unknown quantity q_0. Since f_0 is the result of some unknown mapping, f, applied to the unknown quantity of light, q_0, I denote $a(x_0, y_0)$ by $f(q_0)$.

2. Locate the corresponding pixel in image b, namely $b(x_0, y_0)$. We know that k times as much light gave rise to $b(x_0, y_0)$ as to $a(x_0, y_0)$. Therefore $b(x_0, y_0) = f(kq_0)$. For convenience, I denote $b(x_0, y_0)$ by $f(q_1)$, so that $f(q_1) = f(kq_0)$. Now search around in image a for a pixel that has the numerical value $f(q_1)$, and make a note of the coordinates of the found pixel. Call these coordinates (x_1, y_1), so that we have $a(x_1, y_1) = f(q_1)$.

3. Look at the same coordinates in image b and observe the numerical quantity $b(x_1, y_1)$. We know that k times as much light fell on $b(x_1, y_1)$ as did on $a(x_1, y_1)$. Therefore $b(x_1, y_1) = f(kq_1)$. For convenience, I denote $b(x_1, y_1)$ by $f(q_2)$. So far we have that $f(q_2) = f(kq_1) = f(k^2 q_0)$. Now search around in image a for a pixel that has the numerical value $f(q_2)$ and note these coordinates (x_2, y_2).

4. Continuing in this fashion, we obtain the nonlinearity of the image sensor at the points $f(q_0)$, $f(kq_0)$, $f(k^2 q_0)$, ..., $f(k^n q_0)$.

Now we can construct points on a plot of $f(q)$ as a function of q, where q is the quantity of light measured in arbitrary (reference) units. I illustrate this process diagrammatically (Fig 3-3(a)), where I have introduced a plot of the numerical values in the first image, $a = f(q)$ against the numerical values in the second image, $b = f(kq) = g(f(q))$, which I call the 'range-range' plot, as the axes are both the range of f, with a constant domain ratio, k.

It is helpful to consider, for a moment, the 'range-range' plot, as we will later begin to appreciate that it is an important entity in itself. I will denote the curve traced out by this plot by the function $g(f)$, defined by

$$g(f(q(x, y))) = f(kq(x, y)) \tag{3.1}$$

[5] This algorithm is for illustrative purposes, and is meant to convey the essence of the nonparametric self-linearizing approach used, rather than to be of practical utility.

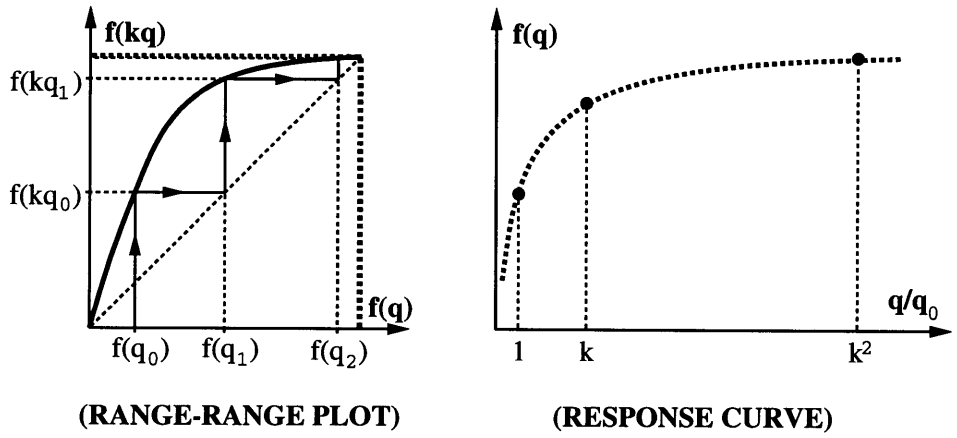


Figure 3-3: Procedure for finding the pointwise nonlinearity of an image sensor from two pictures differing only in their exposures. (RANGE-RANGE PLOT) Plot of pixel values in one image against corresponding pixel values in the other, which I call the 'range-range' plot. (RESPONSE CURVE) Points on the response curve, found from only the two pictures, without any knowledge about the characteristics of the image sensor. If we use a logarithmic exposure scale (as most photographers do) then the samples fall uniformly on the $log(q/q_0)$ axis.

or more simply, $g(f(q)) = f(kq)$, where $q(x, y)$ is the quantity of light received in a first exposure, and $kq(x, y)$, the quantity of light received in a second exposure, is k times that of the first exposure. In traditional film cameras, k would most likely be 2^n, $n \in \mathbb{Z}$, but in electronic cameras, k may vary continuously. In particular, if we wish more points on the curve $f(q)$, we may use a smaller value for k. (Of course, if we have a parametric model for f, we may interpolate points on f, but we will also see later, that we can apply a parametric model directly to the range-range data.) A fully general and continuous solution going from the range-range plot to f, based on integral equations, is itself an interesting problem for future study.

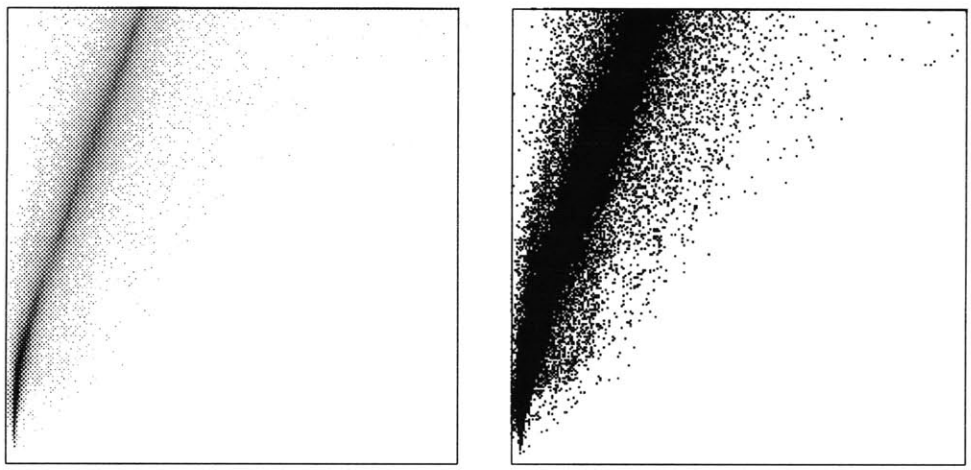
Once the camera is calibrated, we may use the response curve to properly combine sets of pictures like the ones in Fig. 3-1 and 3-2. The pictures that are used to calibrate the camera need not be the same ones used to make the composite. In fact, had we used a smaller value for k to calibrate the camera (e.g. 1.4 or 2 instead of 4), we would have obtained more sample points on the response curve (Fig 3-3(b)).

In general, estimating a function, $f(q)$, from g (e.g. a graph of $f(q)$ versus $f(kq)$), is a difficult problem. However, we can place certain restrictions on f. For example, I typically perform the estimation with the constraint that f is semi-monotonic[6] (increases or remains constant with increasing q). Since the response curve is semi-monotonic, so is the plot depicted in Fig 3-3(a). We can also impose that $f(0) = 0$ by taking a picture with the lens cap on, and subtracting the resulting pixel value from each of the two (or more) images. This step will insure that the plot of Fig 3-3(a) passes through the origin.

3.5.2 Nonparametric self-calibration in the presence of quantization and other noise

In practice, the pixel values are quantized, so that the range-range plot is really a *staircase function*. It is still semi-monotonic, since it is a quantized version of a continuous semi-monotonic function.

[6]The only practical situation that would likely violate this assumption, is where a negative film is being used, the sun is in the picture, and the sun's rays are concentrated on the film for a sufficiently long time to burn a hole through a negative film. The result is a print where the brightest object in the scene (the sun) appears black.



(a) (b)

Figure 3-4: Cross histogram of the images in Figs. 3-1 and 3-2. The cross-histogram of two images is itself an image. Since the two images have a depth of 8 bits, the cross histogram is a 256 × 256 image regardless of the sizes of the two images from which it is obtained. The bin count at the origin (lower left corner) indicates how many pixels were black (had a value of zero) at the same location in both images. (a) Cross histogram displayed as an image. Darker areas correspond to greater bin counts. (b) All non-empty bins are shown as black. Ideally, there should only be a slender "staircased" curve of non empty bins, but due to noise in the images, the curve fattens.

In addition to quantization effects, we also have noise, which may be due to a variety of causes, such as thermal noise in the image sensor, grain in the film, slight misregistration of the images, or slight changes in camera position, scene content, and lighting. We consider a 'joint histogram'[7] of the two images (Fig. 3-4(a)), which is the discrete equivalent of the 'range-range' plot of Fig 3-3. It is a 256 by 256 array since each pixel of the two images can assume 256 distinct values. Due to noise, we see a fat ridge, rather than a slender "staircase". Ideally there should be no points off of the staircase, defined by quantizing the range-range plot, but in practice there is a considerable number of such non-empty bins (Fig. 3-4(b)). In order to estimate the range-range function, $g(f)$, from these cross histograms, one simple algorithm is to find the peaks (indices of the highest bin count) along each row of the cross histogram, or along each column, which may be used as lookup tables to convert an image from the first range to the second range or vice versa (depending on whether computing along rows or along columns). However, this simplistic approach is undesirable for a variety of reasons. Obviously, only integer values will result, so that converting one range to another will result in loss of precision (e.g. most likely differing pixel values will end up being converted to identical pixel values).

We may regard the process of selecting the maximum bin count across each row or column as just one example of a moment calculation, and then consider other moments. For example, the first moment (center of gravity along each row or down each column) typically gives us a non-integer value for each entry of this lookup table. (It is interesting to note that if the cross histogram were a joint probability distribution function, this method would amount to Bayes least squares, e.g. we are going through the same calculations with our deterministic histograms as we would do if we were working with probabilities[8].) Calculating moments across rows or down columns is somewhat successful in "slenderizing" the cross histogram into a range-range curve. However, it still does not enforce monotonicity, and I have found that the resulting curves are not monotonic.

[7] I could also use the term 'joint histogram' to emphasize the similarity to *probability distribution functions*, but have decided not to emphasize that similarity because of some of the confusion that has arisen.

[8] Also, one might be tempted to make a connection to Paul Viola's cross-entropy methods. However, the problem here is far more constrained than the more general method of Viola, and it was found that the method presented here works better.

3.5.3 Non-parametric self-calibration algorithm

In order to impose the monotonicity constraint, I typically proceed as follows:

1. Establish an upper bounding curve, g_u, using a "ratchet effect" as follows:
   (a) Define $g_u(0) = 0$
   (b) Compute moments along each row or column (depending on whether the mapping is from range $f(q)$ to range $f(kq)$ or vice versa). Without loss of generality I will describe the algorithm for columns (e.g. to determine g that takes us from $f(q)$ to $f(kq)$ as opposed to g^{-1} that takes us the other way). Call the moment of the nth column m_n.
   (c) Set $g_u(1) = \max(g_u(0), m_1)$
   (d) Proceed recursively, setting $g_u(n) = \max(g_u(n-1), m_n)$, as n is increased, going left to right across columns of the joint histogram, until the maximum value of n is reached.
2. Establish a lower bounding curve, g_l, using a similar "ratchet effect", but starting at N, the maximum value of n, and initializing $g_l(N) = M$, where M is the maximum possible pixel value (typically 255). This is done by decreasing n, moving us from right to left on the cross histogram, and it is done by selecting $\min(g_l(n), m_{n-1})$.
3. The final result, $g(f(q))$, is computed from the average: $g = (g_u + g_l)/2$.

3.5.4 Non-parametric self-calibration example

I now present a simple practical test-case in which I determine some range-range curves of the sensor array inside one of the personal imaging rigs I designed and built (as described in Chapter 1). This example illustrates how the theory developed in the previous section is typically used in practice. In order to generate the cross histograms, I captured images differing only in exposure, the exposure being changed by adjusting the integration time ("shutter" speed) of the sensor array. Rather than just using two differently exposed images, I used a total of five differently exposed images[9]. (See Fig 3-5.)

In this case, the duration of the exposures were known (1/4000, 1/2000, 1/1000, 1/500, and 1/250 of a second), so that cross histograms were generated for each possible combination of these image pairs, and, knowing that some were redundant (e.g. there were four curves for $k = 2$, three curves for $k = 4$, two curves for $k = 8$, but only one curve for $k = 16$), the cross histograms were averaged together in cases where there were more than one, and the averaged cross histogram was slenderized, using the proposed algorithm presented in the previous subsection. These slenderized cross histograms are shown as a family of curves in Fig 3-6(a). From the five images, four curves were generated above the diagonal, and four more below the diagonal, the latter being generated by reversing the order of the image sequence. The diagonal, which represents any image compared against itself, is known to be the identity function $g(f) = f$. Thus there are nine curves in Fig 3-6(a). Through the process of interpolating (and extrapolating) between (and beyond) these curves, a function $g(f(q)) = f(kq)$ may be returned for any desired value of k. Thus we have a complete non-parametric characterization of the response function particular sensor array and digitization hardware, and in this sense, we say that the camera is *calibrated.*

Some alternatives to direct estimation of g from the cross histograms involve fitting to some parametric curve, such as a spline or the like. In the next subsection, I present one such parametric choice for g which is motivated by the tradition of photographic emulsion.

[9] Each of these images was gathered by signal averaging (capturing 16 times, and then averaging the images together) to reduce noise. This step is probably not necessary with most normal-sized cameras; noise from my sensor array was very high because I used the smallest sensor I could obtain, and built this into an ordinary pair of sunglasses, in such a way that the opening through which light entered was very small. Primarily because the device needed to be unobtrusive, the image quality was very poor. However, as we will see in subsequent chapters, this poor image quality can be mitigated by various new image processing techniques.

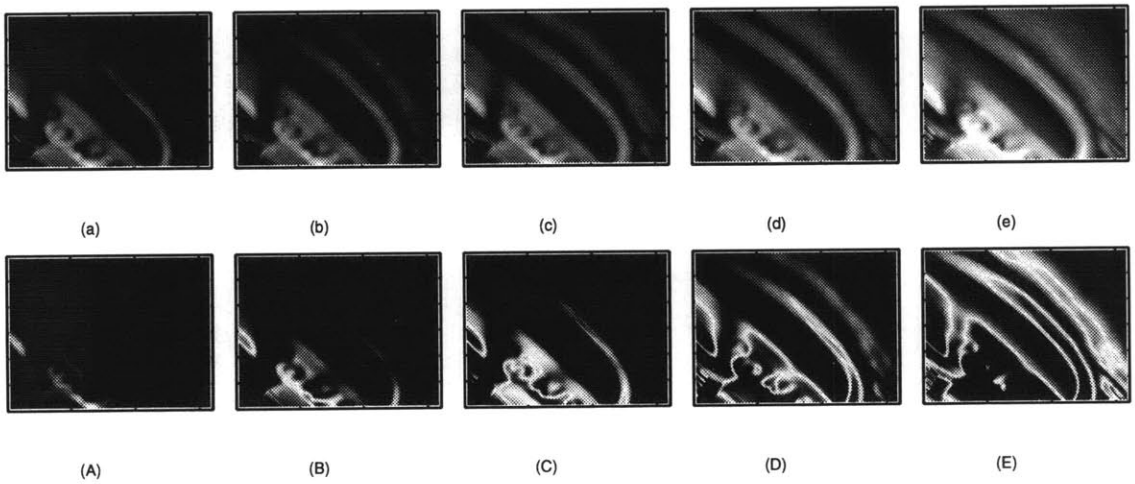


Figure 3-5: (a-e) Collection of differently exposed images used to calibrate the author's eyeglass-based personal imaging system. These images differ only in exposure. (A-E) Certainty images corresponding to each image. The certainty images, $c(f(x,y))$ are calculated by evaluating f with the derivative of the estimated response function. Areas of higher certainty are white and correspond to the mid tones, while areas of low certainty are black and correspond to highlights and shadows, which are clipped or saturated at the extrema (toe or shoulder of the response curve) of possible exposures. Another interpretation of the proposed method of combining multiple images of different exposure is to think of the result as a weighted sum of exposure adjusted images (adjusted to the same range), where the weights are the certainty images.

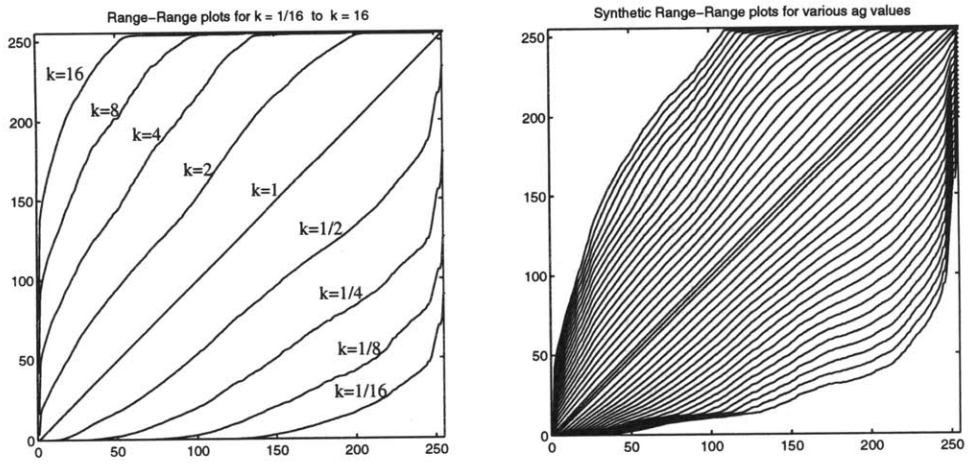


Figure 3-6: Range-range plots, $g(f(q(x,y))) = f(kq(x,y))$, characterizing the specific eyeglass-mounted CCD sensor array and wearable digitizer combination designed and built by author. (a) plots estimated from cross histograms of differently exposed pictures of the same scene, using the proposed non parametric self-calibration algorithm. (b) family of curves generated for various values of k, by interpolating between the nine curves in (a).

3.5.5 Parametric self-linearizing methods

We may be willing to place even stronger restrictions on the response curve. When examining the response curves of photographic emulsions, it is common to use logarithmic coordinates, that is, to consider the *D log E* (density versus log exposure) curve. Density is already a logarithmic unit, so it is not required to take its logarithm when making the plot. Also note that the use of density rather than transmissivity takes care of the sign-change that is a result of the fact that films are typically "negative" (e.g. that photographic negatives are darker in areas of high light exposure and vice versa).

The *D log E* curve of most typical photographic emulsions is linear over a relatively wide region, which suggests the commonly used empirical law for the response function of film [27]:

$$f(q) = \alpha + \beta q^{\gamma} \tag{3.2}$$

where q is the quantity of light falling on the image plane.

I have proposed other parametric response functions, for example, one given by $f = \alpha + exp(1/2 + \arctan(\gamma \log(q) + \beta)/\pi)$, which attempts to capture the toe and shoulder regions of the response curve, but for the purposes of this chapter of the thesis, I will use the simpler model of (3.2), mainly because it is most suitable for illustrative purposes.

The constant α characterizes the density of unexposed film which is typically not zero (e.g. areas of the film that were never exposed to light are not completely clear, but, rather, have some degree of "fog", in addition to the fact that the base upon which the film is made also has a slight absorption of light). The density of film that had zero exposure to light prior to development, also known as *Dmin* (short for "minimum density"), can be measured from a portion of the film at the edges (sprocket holes), or from an unexposed frame (e.g. as may be generated by deliberately taking a picture with lens cap on) or film leader.

This model may also be applied to electronic cameras, again subtracting off α, by using a picture taken with the lens cap on.

The range-range plot would then take the form

$$g(f(q)) = k^{\gamma} f \tag{3.3}$$

where k is the ratio of exposures relating the two pictures. Thus to find the value of the linear constant, k^{γ}, in $f = k^{\gamma} f$ we simply apply linear regression to points known in the range-range space. From k^{γ} we can obviously find the film's contrast parameter, γ.

3.6 Self-calibrating camera: Enforcing superposition

Another way that the camera can calibrate itself is that it may enforce superposition. To do this, we may use two light sources, light 1 and light 2, and enforce additivity of the two lights. This works by creating a lightvector subspace as described in Chapter 2. Let the projection of the scene on the image plane when only lamp 1 is turned on be L_1, and the projection of the scene as illuminated by only lamp 2 be L_2. Let the unknown response of the imaging system be f, so that a picture taken with only lamp 1 on will be $W_1 = f(L_1)$, and a picture taken with only lamp 2 on, will be $W_2 = f(L_2)$. From these two pictures, together with a third picture taken with both lamps on together, $W_{12} = f(L_1 + L_2)$, we can derive an estimate for f, $\tilde{f}$, by using the fact that incoherent light is additive[10]. If we plot $\tilde{f}^{-1}(W_{12})$ as a function of two variables, $\tilde{f}^{-1}(W_1)$ and $\tilde{f}^{-1}(W_2)$, we would like to have the plot be, as close as possible, that of a plane passing through the origin and the points $(0, 1, 1)$ and $(1, 0, 1)$. (The plane described here is that of the lightvector subspace described

[10] Bichsel [29] has proposed a similar way of calibrating the response of one pixel of a camera by manually adjusting two light sources, and making note of the reading with each one turned on separately, and with both turned on together.

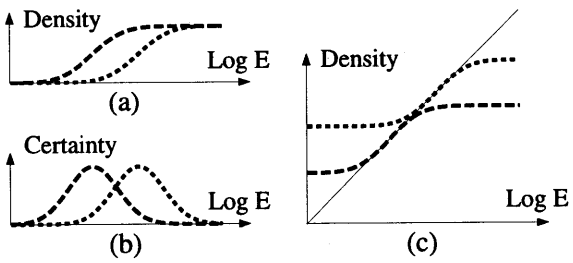


Figure 3-7: Response curves of the Wyckoff set (note the log scale as opposed to the scale of Fig 3-3 (RESPONSE CURVE) which was linear). (a) Response curves corresponding to two different exposures, depicted as though they were taken on two different films. The dashed line may be thought of as either a longer exposure or the faster layer of a 2-layer Wyckoff film, while the dotted line may be regarded as a shorter exposure or the slow layer of a 2-layer Wyckoff film. (b) "Certainty functions" calculated by differentiating the two response curves. (c) Hypothetical curves re-aligned as they will be when the images are later combined. The response of the ideal composite is indicated by the thin solid line. Using more exposure bracketing (or more layers on a Wyckoff film), we can extend this response indefinitely.

in Chapter 2.) Thus we find $\tilde{f}$ by minimizing the error

$$e = |\tilde{f}^{-1}(W_1(x,y)) + \tilde{f}^{-1}(W_2(x,y)) - \tilde{f}^{-1}(W_{12}(x,y))| \tag{3.4}$$

over all locations in the image, (x, y).

As in Section 3.5, we constrain this search for f to the space of monotonic functions. Also, as in Section 3.5, we take one picture with both lights turned off, and subtract this from each of the three pictures, so that we can constrain the response curve, f, to pass through the origin, $f(0) = 0$, regardless of any other light sources that might be present in the scene and any bias in the imaging sensor array (or of any initial density of the film, *Dmin*, and any bias in the film scanning apparatus).

The above constraints are always imposed, whether or not we are using an empirical law, but, in addition to these, we may wish to fit the data to an empirical response curve, in which case the task of estimating the function f is transformed into a task of estimating the parameters of the empirical function.

3.7 Combining images of different exposure

At this point we have found the response curve (by fitting to the data in the range-range plot, as in Fig 3-4), and can shift the response curve to the left or right to get the curves of the two or more exposures (Fig. 3-7(a)). In the shadow areas (areas of low exposure, E) the same quantity of light in the scene has had a more pronounced effect on the dashed-exposure, so that the shadow detail in the scene will still be on a portion of the dashed line that is relatively steep. The highlight detail will saturate this exposure, but not the dotted-exposure.

In general, for parts of the film that are exposed in the extremes (greatly overexposed or greatly underexposed), detail is lost – we can no longer distinguish small changes in light level since the resulting changes in film density are so small that they fall below the noise floor (e.g. we are operating on the flat parts of Fig 3-7). On the other hand, steep portions of the response curves correspond to detail that can be more accurately recovered, and are thus desirable operating points. In these regions, small changes in light will cause large changes in the measured value of the response function, and even if the measurements are highly quantized (e.g. only made with 8 bit precision), small differences in the measured quantities will remain discernible.

Thus we are tempted to plot the derivatives of these hypothetical response curves (Fig. 3-7(c)), which I call the *certainty functions.*

At first glance, one might be tempted to make a composite from two or more differently exposed pictures by manually combining the light regions from the darker pictures and the dark regions from the lighter pictures (e.g. manually selecting the middle of Fig. 3-1 and pasting on top of Fig. 3-2). However, we wish to have the algorithm automatically combine the images. Furthermore,

Figure 3-8: 'Crossover image' corresponding to the two pictures in Figs. 3-1 and 3-2. Black denotes pixel locations where Fig. 3-1 is the more "certain" of the two images, and thus where Fig. 3-1 should contribute to the composite. White denotes pixel locations where Fig. 3-2 is the more "certain" of the two images, and thus where it should contribute to the composite. In practice, we take a weighted sum of the images rather than the abrupt switchover depicted in this figure.

the boundary (Fig. 3-8) between light regions and dark regions is, in general, not a smooth shape, and would be difficult to trace out by hand. Pasting this irregular region of Fig. 3-1 into Fig. 3-2, amounts to choosing, at each point of the composite, the source image that has the higher *certainty* of the two. However, abrupt changes resulting from suddenly switching from one image to another occasionally introduce unpleasant artifacts, so instead, we compute a weighted average. Every pixel of the composite, whether shadow or highlight, or in the transition region, is drawn from all of the input images, by weighting based on the certainty functions. This provides a gradual transition between the images, where the shadow detail comes **primarily** from the lighter image, and the highlight detail comes **primarily** from the darker image.

The extended-response image array from the two pictures of Figs. 3-1 and 3-2 is a floating point array which has more than 256 distinct values, and therefore cannot be displayed on a conventional 8-bit display device.

3.8 Dynamic range; 'dynamic domain'

Tekalp, Ozkan, and Sezan [30], Irani and Peleg [31], and Mann and Picard [32] have proposed methods of combining multiple pictures that are identical in exposure, but differ in camera position. The result is increased *spatial resolution.* When one of these images is too big to fit on the screen, we look at it through a small movable viewport, scrolling around and exploring one part of the 'image domain' at a time.

In this chapter, the composite image is a *floating point* array, and is therefore too *deep* for conventional screen depths of 24 bits (8 bits for each color channel), so I constructed a slider control to allow the user to interactively look at only part of the 'image range' at a time. The user slides the control back and forth depending on the area of interest in the composite image. This control is to screen range as the scrolling window is to screen domain – showing the vast tonal range one piece at a time. Of course we were able to obtain the underexposed view much like Fig 3-1, by sliding the control left, and the overexposed view much like Fig 3-2 by sliding the control right.

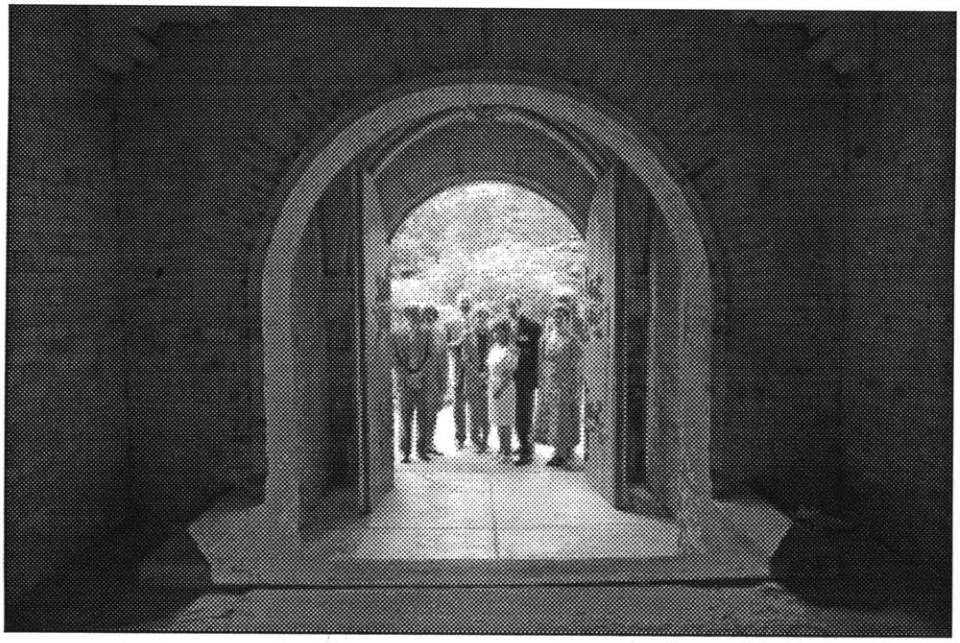


Figure 3-9: Wyckoff composite, derived from Fig. 3-1 and Fig. 3-2, reduced in contrast and then quantized to 8 bit image depth.

When an image is too big to fit on the screen, one can also subsample its domain to make it fit on the screen. Analogously, I applied the appropriate range-subsampling (quantization to 8 bits) to our floating-point composite image for screen display, or print (Fig 3-9). Before quantization, I applied a nonlinearity which restored the appearance of the image to the familiar tonal scale to which photographers are accustomed, and I added the appropriate amount of *dither*[11]. It is worth mentioning that the final nonlinearity before quantization selects the tonal range of interest. We can regard its derivative (the 'certainty function') as depicting the 'Wyckoff spectrum' (which regions of greyvalue are emphasized and by how much) analogous to a conventional bandpass filter which selects the frequencies of interest. The elements of a Wyckoff set, having equally spaced certainty functions of identical shape, are analogous to a bank of *constant Q* filters.

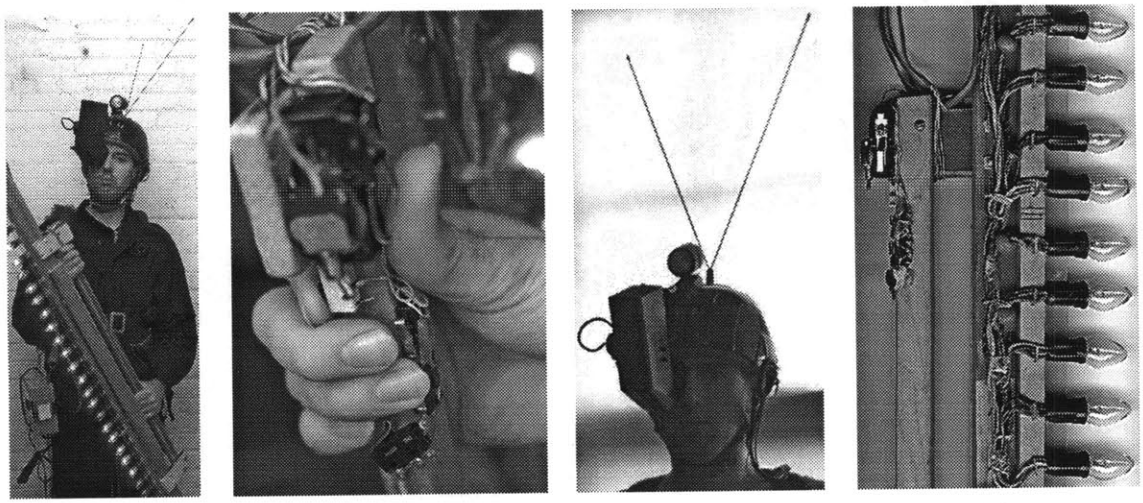
If all that is desired is a single print, why not just try to formulate a super-low-contrast film or image sensor? The superiority of the Wyckoff composite lies in the ability to control the process of going to the low contrast medium. For example, we might apply a homomorphic [24] filtering operation to the final composite, which would bring out improved details at high spatial frequencies, while reducing the unimportant overall changes in density at low spatial frequencies.

3.9 Combining pictures of differing illumination

During the 1970s, I developed means and apparatus for "lightpainting", where I was interested in using movable and wearable light sources as extensions of my own body — tools for the production of visual art, and various imaging experiments for the purposes of characterizing the response of objects to light. The apparatus comprised a 6502-based wearable computer system with various peripherals in the form of control devices and outputs (light sources). See Fig 3-10. This apparatus, which I called the "lightspacer" was used to capture a lightvector subspace, as shown in Fig 3-11.

The information captured from this process has been parameterized on two planes, a light plane and an image plane. The light plane parameterizes the direction from which rays of light enter into

[11] The dither did not seem to have a perceivable effect on an 8 bit image, but when reducing a Wyckoff composite to 5 bits or less, the dither appeared to make a noticeable improvement.



(a) (b) (c) (d)

Figure 3-10: Lightspace computer (WearComp1) and output device, called the 'pushbroom duster', which I developed in the 1970s, pictured here together with a 1980 display device. Such an apparatus is used for "sweeping" out large portions of 'lightspace' in a wholesale fashion, intended for use in conjunction with a motion picture camera, video camera, or 35mm still camera with motor drive. The 'pushbroom duster' can sequence lamps individually, or produce patterns of lightspace. Furthermore, when the lightpattern is to appear right in the picture (as in Fig 2-6 where we see the actual light source), either directly or through reflection in chrome objects or the like, it is often desirable to control the shape of the light source for certain effects. For example, in a later version of the apparatus based on a 6502 microprocessor (WearComp2, completed in 1981), a font table was designed for the apparatus so that the light source could assume the shape of printed characters (much as a dot-matrix printer creates characters from a single row of print pins). (a) Author pictured with wearable computer which I integrated into a welded-steel frame worn on my shoulders (note belt around my waist which directed some of the weight onto my hips). Power converter hung over one shoulder. Two or three antennae, operating at different frequencies, were typically used to allow simultaneous transmission and reception of data, voice, or video. This system, which replaced an electromechanical wearable lightspace computer I had built earlier, allowed for complete interaction while walking around doing other things. (b) Close up of the uppermost end of the lightpaintbrush handle, showing the end that's held in my right hand. The collection of six spring-lever switches, one for each finger and two for thumb, permits input of data, as well as control of the lightpainting programs. (c) Close up of my 1980 display (upgrade to the late 1970s display). (d) Close-up showing the linear array of lamps. Lamps are spaced 1.5 inches apart (e.g. for 2/3 dpi "printing" in 3-D space). Some components of this apparatus have been updated to keep it operational (e.g. re-fitting on new helmet as I outgrew the original helmet, as well as upgrade to a peripheral interface adapter instead of the original TTL logic interface).

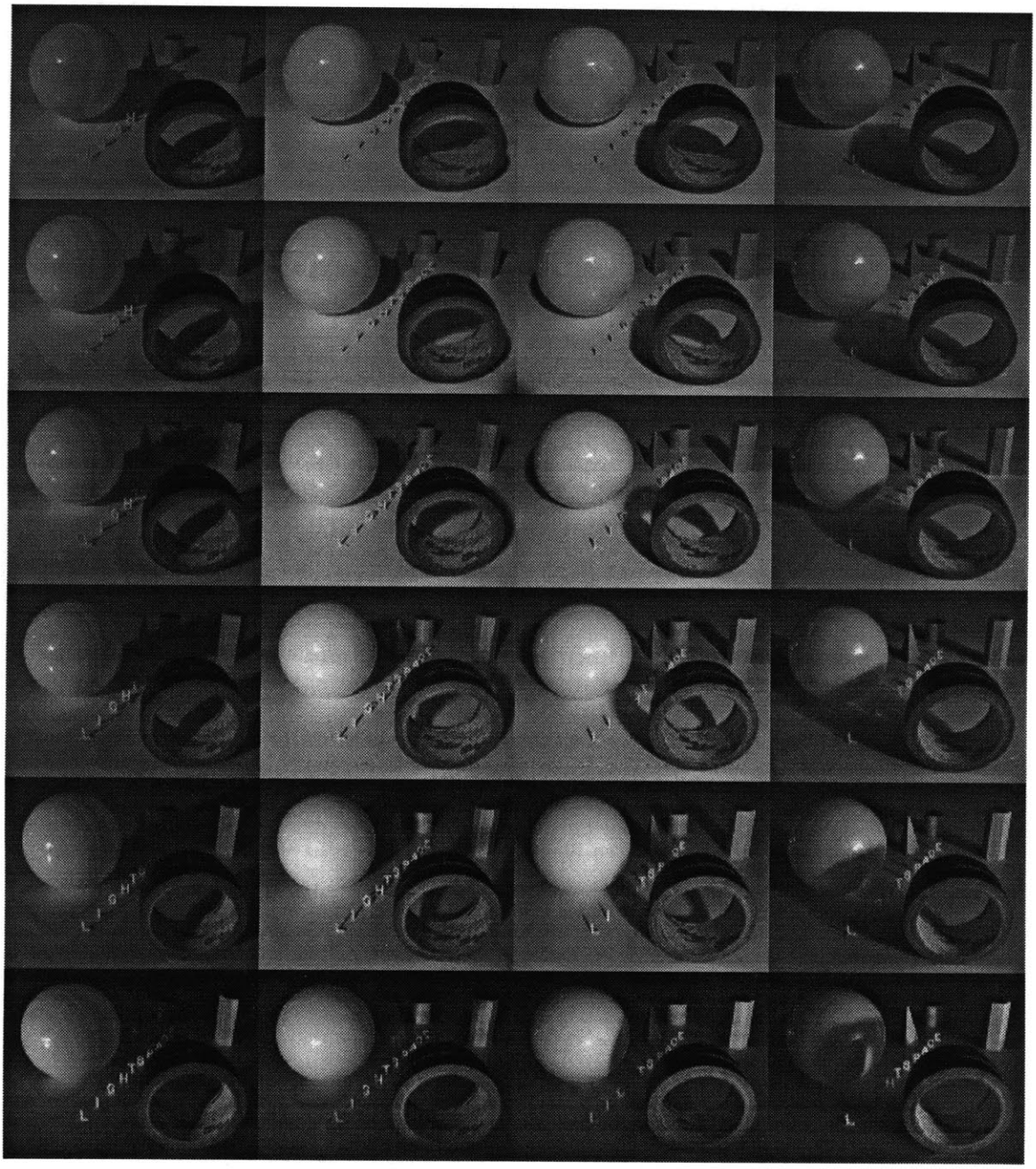
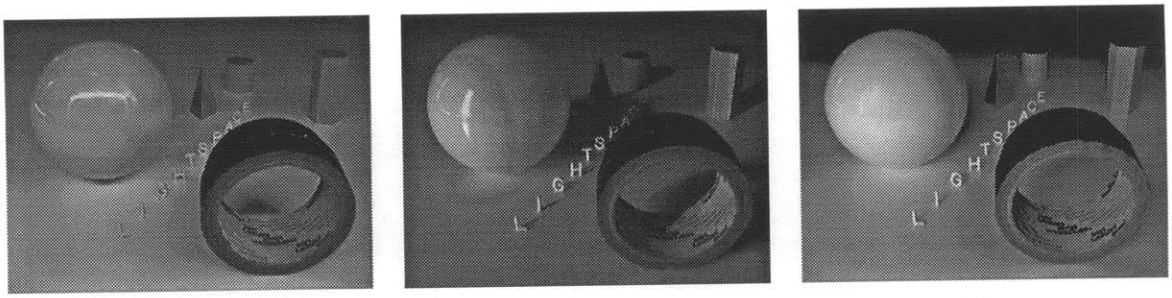


Figure 3-11: Lightvector subspace acquired from a system similar to that depicted in Fig 3-10. As the row of lamps is swept across (sequenced), it traces out a plane of light ("sheetlight" as described in Chapter 2). The resulting measurement space is a four dimensional array, parameterized by two indices (azimuth and elevation) describing rays of incoming light, and two indices (azimuth and elevation) describing rays of outgoing light. Here this information is displayed as a block matrix, where each block is an image. The indices of the block indicate the lightvector, while the indices within the block are pixel coordinates.



(a) (b) (c)

Figure 3-12: Homomorphic superposition over various surfaces in lightspace. (a) Here I synthesize the effect of a scene illuminated with long horizontal slender light source (e.g. as would be visible if it were lit with a bare fluorescent light tube), reconstructing shadows that appear sharp perpendicular to the line of the lamp, but soft across it. Notice also the slender line-like highlight in the specular sphere. (b) Here I synthesize the effect of a vertical slender light source. (c) Here I synthesize the effect of two light sources, so that the scene appears as if lit by a vertical line source, as well as a star-shaped source to the right of it, but both sources coming from the left of the camera. The soft yet highly directional light is in some way reminiscent of a Vermeer painting, yet all of the input images were taken by the harsh but moving light source of the pushbroom-like apparatus.

the scene, while the image plane parameterizes directions from which rays of light leave the scene. This four dimensional subspace of the 14 dimensional lightspace is enough to synthesize a picture of the scene as it would appear if it were illuminated by any desired shape of light source that lies in the light plane. For example, a picture of how the scene would look under a long slender-shaped light source (like that produced by a long straight fluorescent light tube) may be obtained by linearizing the lightspace measurements (as described in Sections 3.5 and 3.6), then integrating over the desired light shape, then undoing the linearization process. In reality, these measurements are made over a discrete sampling lattice (finite number of lamps, finite number of pixels in each photometrically linearized camera). The Wyckoff principle allows us to neglect the effects of finite word length (quantization in the quantity of light reported at each sensor element), thus the measurement space depicted in Fig 3-11 may be regarded as a continuous real-valued function of four integer variables. Thus rather than integrating over the desired light shape, we would sum (homomorphically) over the desired lightvector subspace. This summation corresponds to taking a weighted sum of the images themselves. Examples of these summations are depicted in Fig 3-12. I have also developed a method of capturing, recording, computing, and displaying this information, in real time, with a natural and intuitive display user-interface [33]. Thus not only is it possible to record, in a practical way, the manner in which scenes or objects respond to light, but it is also possible to present this information in a tangible way.

3.10 Wyckoff analysis and synthesis filterbanks

We can regard the Wyckoff film (or exposure bracketing) as performing an *analysis* by decomposing the light falling on the sensor into its 'Wyckoff layers'. The proposed algorithm provides the *synthesis* to *reconstruct* a floating point image array with the dynamic range of the original light falling on the image plane. This *analysis-synthesis* concept is illustrated in Fig 3-13.

The analysis-synthesis concept suggests the possibility of using the Wyckoff layer decomposition as a "Wyckoff filter" that could, treat the shadows, midtones, and highlights of an image differently. For example, we might wish to sharpen the highlights of an image without affecting the midtones and shadows.

The Wyckoff filter provides a new kind of filtering – 'amplitude domain' filtering – as opposed to the classic *Fourier domain*, *spatial domain*, *temporal domain*, and *spatiotemporal* filters. We envision a generalized Nyquist-like theory for reconstruction from 'amplitude samples', to augment classic sampling theory.

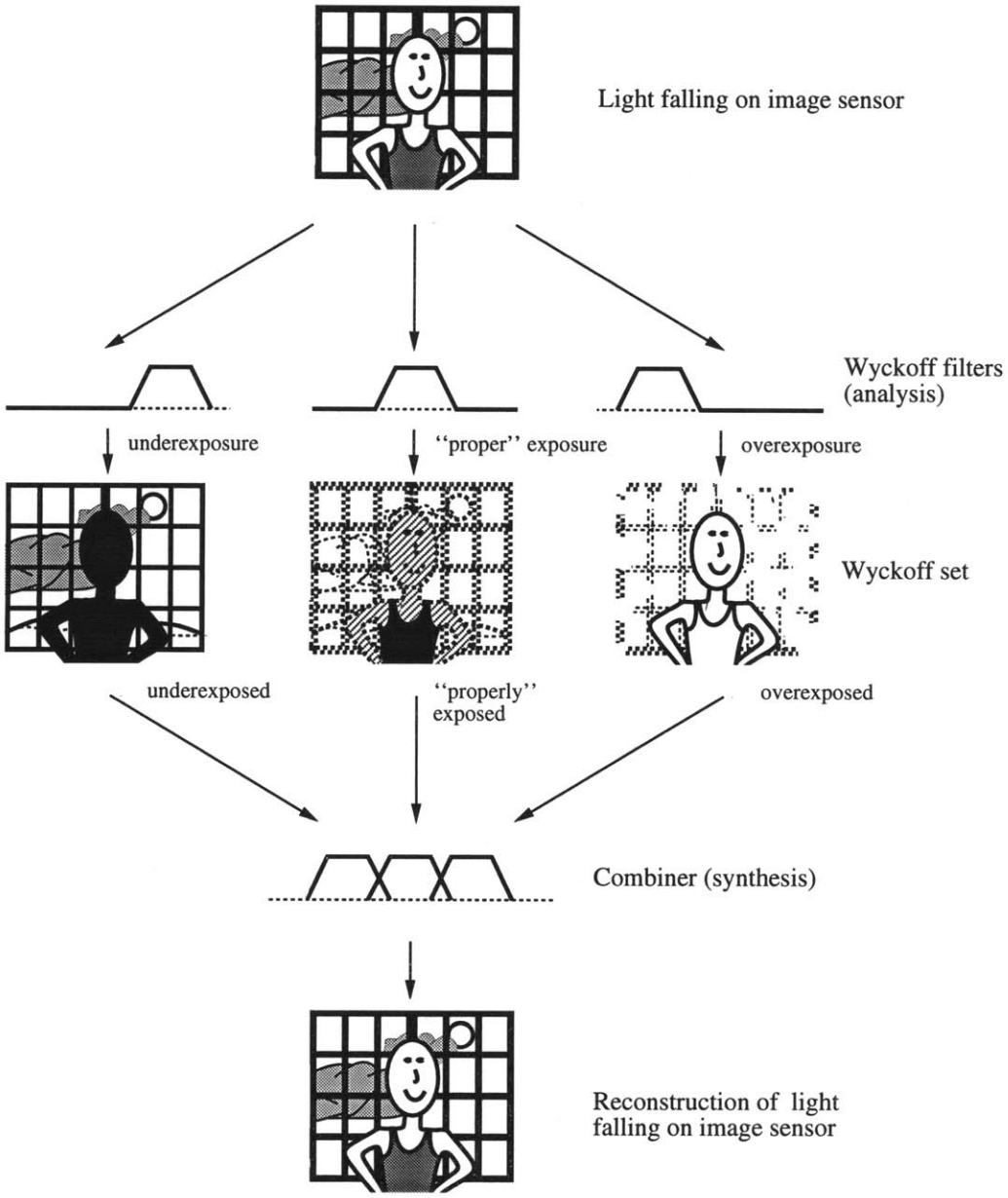


Figure 3-13: The layers of a Wyckoff film decompose the light falling on the film into differently exposed images. Each of these images may be regarded as a filtered version of the light falling on the image sensor. These 'Wyckoff filters' act as a filterbank to capture overlapping portions of the exposure "spectrum", and perform an analysis of the light falling on the image sensor. The set of pictures can then be used to obtain perfect reconstruction of the original light intensity falling on the image sensor.

3.11 Homomorphic linearity, superposition, and the range of light

The concept presented in this chapter is part of the larger framework called 'lightspace' [34], presented in Chapter 2.

Regarding an image of size $M \times N$ pixels as a point or vector in $\mathbb{R}^{MN}$ allowed us to consider each of a set of differently exposed images, prior to nonlinearities and quantization, as collinear vectors in $\mathbb{R}^{MN}$.

Furthermore, we saw that when we obtained multiple pictures of the same scene differing only in lighting, they spanned a subspace of $\mathbb{R}^{MN}$, which I called the 'lightvector subspace'. From any set of 'lightvectors' (pictures of a scene taken with particular lighting) that span a particular 'lightvector subspace' we can synthesize pictures taken with any combination of the light sources, by using the homomorphic superposition principle.

In Chapter 2, we saw how a simple example, consisting of two pictures, one taken with flash and the other without (e.g. two differently illuminated pictures), spanned a two dimensional picture space. An example of such a picture space is illustrated in Fig 3-14. In that picture, I did not merely combine the two basis pictures $F_0(x,y) = f(q_0(x,y))$ and $F_1(x,y) = f(q_1(x,y))$ with linear operations (scaling and superposition) applied directly to the pictures:

$$\alpha F_0 + \beta F_1 = \alpha f(q_0(x,y)) + \beta f(q_1(x,y)) \tag{3.5}$$

but, rather, I did this *homomorphically*. By *homomorphically*, I mean that I performed the operations on q itself:

$$f(\alpha f^{-1}(F_0(x,y)) + \beta f^{-1}(F_1(x,y))) = f(\alpha q_0 + \beta q_1) \tag{3.6}$$

which I refer to as 'homomorphic superposition'. The 'homomorphic superposition' provides an image that would have arisen had the scene been illuminated by natural (ambient) lighting of strength α and a flashlamp of strength β.

The importance of homographic superposition (as opposed to linear superposition) is illustrated in Fig 3-15, where I illustrate, with actual images, the difference between a linear and homomorphic additivity.

In addition to homomorphic superposition, one may use the Wyckoff principle to combine multiple, differently illuminated images in other more expressive or creative ways. Some other possibilities are suggested in Chapter 11.

Furthermore, we can work with multiple pictures rather than just two. Samples from the lightvector subspace of three differently illuminated pictures appears in Fig 3-16.

3.11.1 From lightvectors to lightmodules

To the extent that a multichannel image (such as color, having three channels: R,G,B), having L channels is a collection of L vectors, then for each of a set of multiple channel pictures differing only in lighting, we can associate L vectors. I call the set of L vectors a 'lightmodule'.

It has been shown [23] that a set of 'lightmodules' (which I call a 'lightmodule subspace') also spans a useful space. For example, a set of color pictures of a scene differing only in lighting, taken with white lights at various places in the scene, was used to synthesize the result of having taken a picture with colored lights at these same locations.

3.12 Chapter summary

We have presented a means of combining multiple digital images that differ only in their exposure, to arrive at an extended-response floating point image array. The method proceeds as follows:

1. From the set of pictures (or from another set of pictures taken with the same camera) determine the camera's pointwise response function using the "self-calibration" method of Section 3.5.

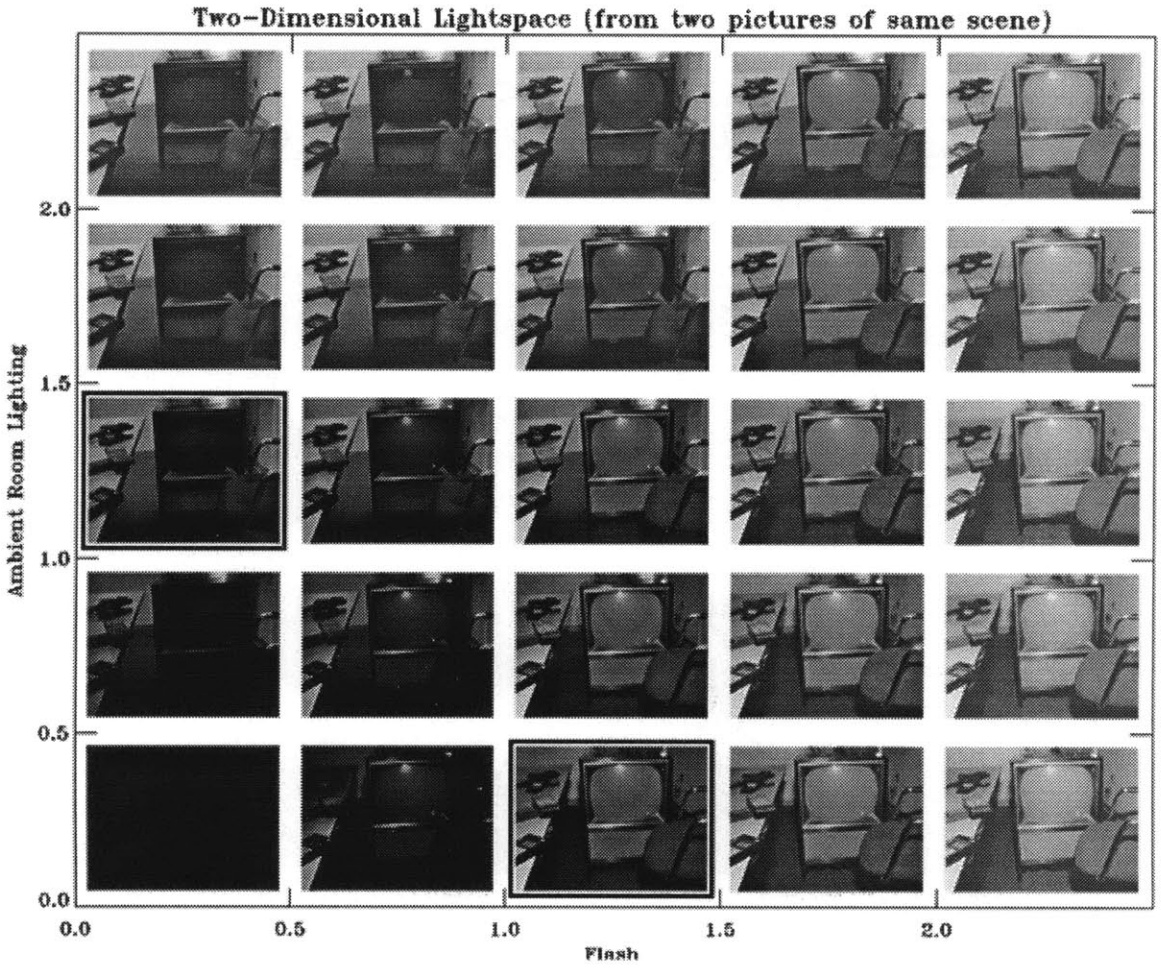


Figure 3-14: Twenty five points sampled from a two dimensional homomorphic lightvector subspace generated from two images, $F_0(x,y) = f(q_0(x,y))$ and $F_1(x,y) = f(q_1(x,y))$. The two input images are denoted by the heavy black outline around each one (located at 1.0 on the Ambient Room Lighting axis and at 1.0 on the Flash axis). Notice that the point at $(0,0)$ is uniformly black (corresponding to uniformly zero light falling on the image sensor). This 5 by 5 block matrix of images was generated according to $f(\alpha f^{-1}(F_0(x,y)) + \beta f^{-1}(F_1(x,y))) = f(\alpha q_0 + \beta q_1)$, where $q_0(x,y)$ was an estimate of the actual quantity of light falling on the image sensor due to the ambient illumination, and $q_1(x,y)$ was an estimate of the actual quantity of light falling on the image sensor due to the flashlamp. Thus each picture, at coordinates (α, β), is the actual picture that would have been captured from that same camera, had the scene been illuminated with α units of ambient room lighting and β units of flash.

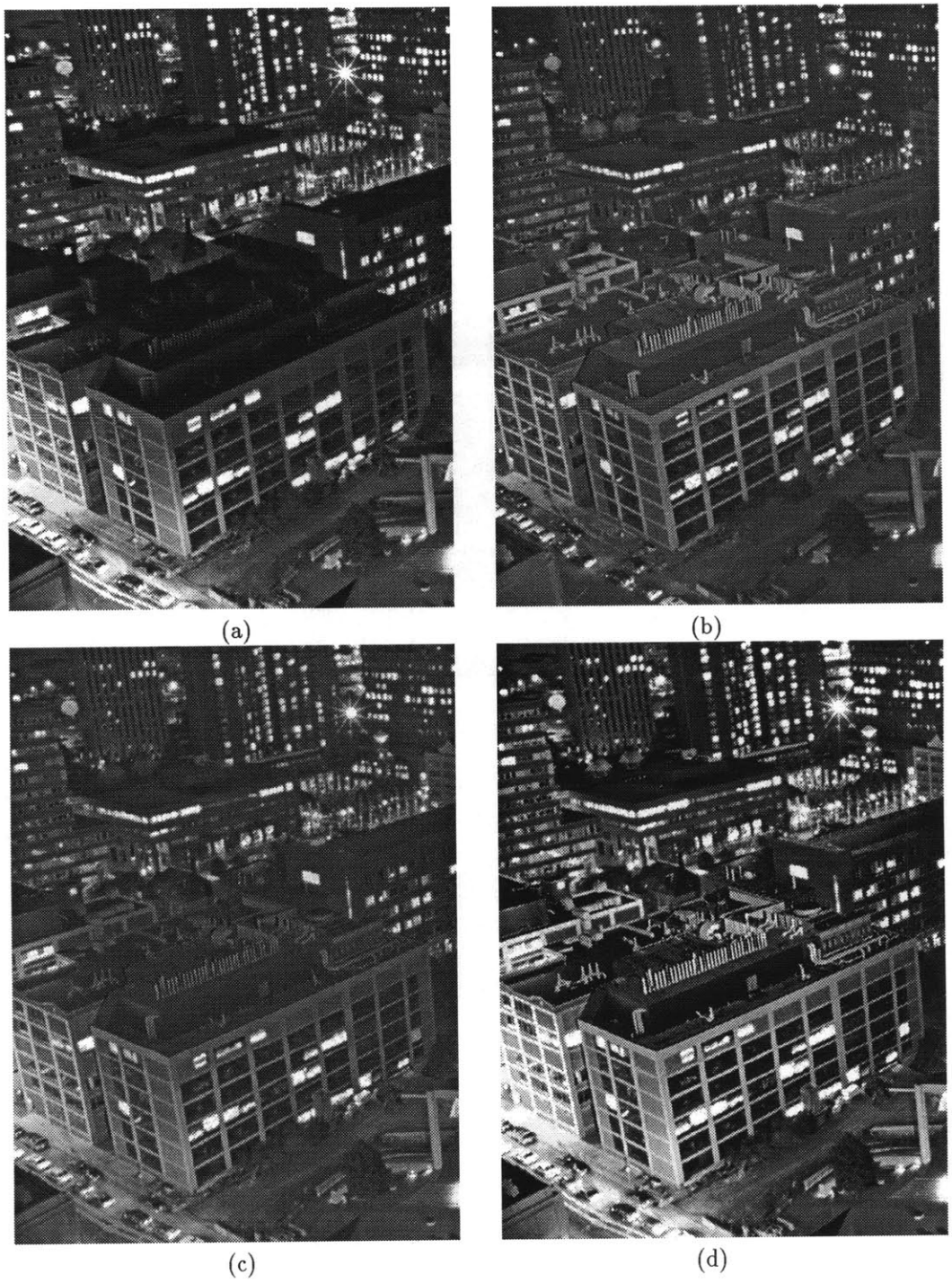


Figure 3-15: 'Homomorphic superposition': (a) Picture of the Cambridge cityscape taken under its natural ("ambient") light. (b) Picture taken with electronic flash (FT-623 operating at 16kJ in 30 inch highly polished chrome reflector) shows nice detail on the rooftops of the buildings where very little of the city's various light sources are shining. (Note thin shadow to left of each building, as flash was to right of camera.) Notice how some portions of the picture in (a) are better represented, while other portions in (b) are better. (c) Linear combination of above two images. Notice undesirable "muted" highlights. (d) Homomorphic superposition (see text) provides an image with much better contrast and tonal fidelity.

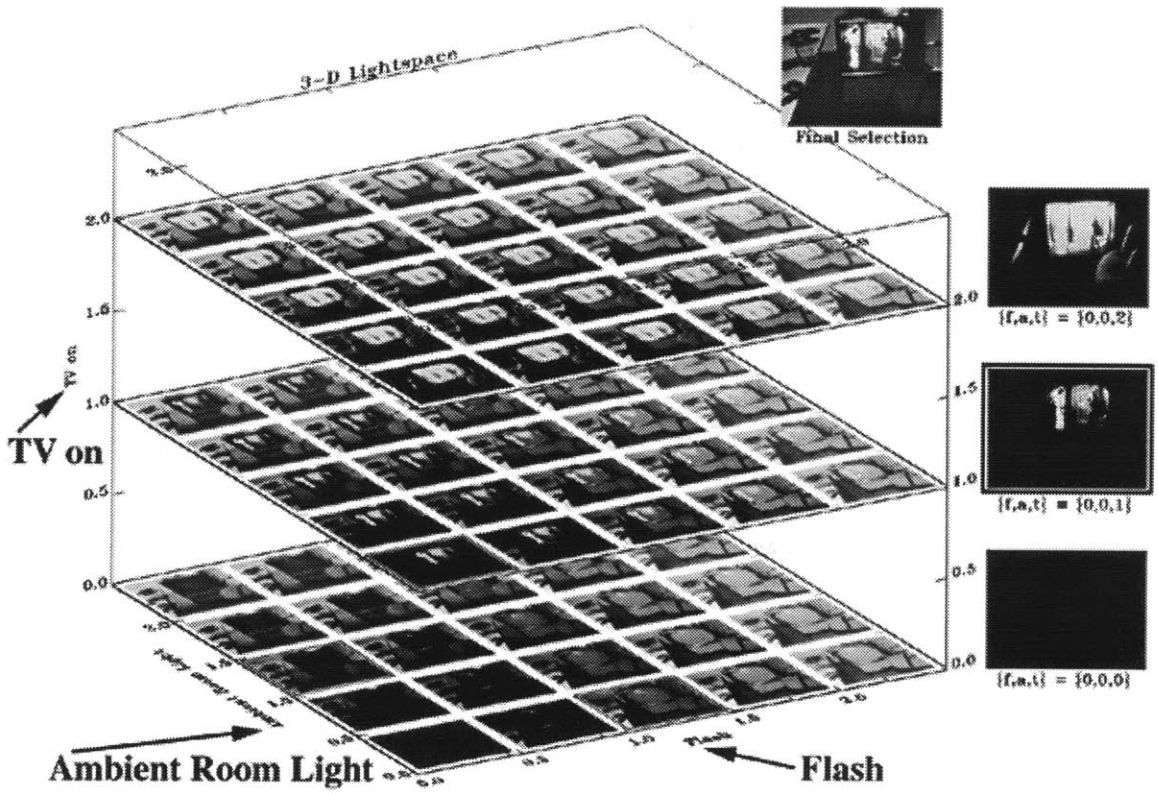


Figure 3-16: Here a 3-D lightvector subspace is illustrated. Each of these 75 pictures corresponds to a point in this sampling of the lightvector space on a 3 by 5 by 5 lattice. The three unit lightvectors, again indicated by a heavy black outline, are located at 1.0 Ambient Room Light, 1.0 Flash, and 1.0 TV on. (Figure generated before the advent of PostScript fonts, but subsequently, axes have been re-labeled for readability at this small size.) The last axis was generated by taking one picture with no flash, and also with the lamp in the room turned off, so that the only source of light was the television set itself. The new axis (TV on) is depicted by itself to the right of this 3-D space, where there are 3 points along this axis. The origin at the bottom is all black, while the middle picture is the unit (input image) lightvector, and the top image is synthesized by calculating $f(2q_2)$, where $q_2(x, y)$ is the quantity of light falling on the image sensor due to the television set being turned on. Near the top of the figure, I have indicated my Final Selection, which is the picture I selected out according to my personal taste, from among all possible points in this 3-D space.

2. Linearize the images (undo the nonlinear response of each), if desired, or map the response curves onto one desired final response curve.

3. Compute the *certainty function* by differentiating the response function. The certainty function of each image is found by appropriately shifting this one certainty function along the exposure axis.

4. Compute the weighted sum of these images, weighting by the *certainty functions*.

The composite may be explored interactively or contrast-reduced and quantized, for a conventional display device. Furthermore, we can regard the Wyckoff film (or exposure bracketing) as performing an *analysis* by decomposing the light falling on the sensor into its 'Wyckoff layers'. The proposed algorithm provides the *synthesis* to *reconstruct* a floating point image array with the dynamic range of the original light falling on the image plane.

3.13 Beyond homomorphic imaging

In addition to homomorphic filtering, the new forms of analysis and synthesis filterbanks suggest the possibility of a 'Wyckoff filter' that could, for example, blur the highlights of an image while sharpening the midtones and shadows. Wyckoff filters work in the 'amplitude domain', in contrast to Fourier filters which work in the frequency domain, or spatio-temporal filters which work in the space and time domains.

To carry this analogy further, we can also imagine something analogous to the Nyquist sampling theorem, which might tell us how many differently exposed pictures of a scene we would need to capture to provide an accurate measurement of $q(x, y)$, or how many differently illuminated pictures the scene we might require to describe the manner in which the scene responds to arbitrary lighting (e.g. one might imagine various lighting interpolation functions, etc.).

Chapter 4

Projective geometry and the domain of light

In the early days of personal imaging, a specific location was selected from which a measurement space or the like was constructed. From this single vantage point, a collection of differently illuminated/exposed images was constructed using the wearable computer and associated illumination apparatus. However, this approach was often facilitated by transmitting images from this single specific location (base station) back to the wearable computer, and vice-versa. Thus, when I developed the eyeglass-based computer display/camera system, it was natural to exchange viewpoints with another person (namely the operator of the base station). This mode of operation ("seeing eye-to-eye") made the notion of perspective very apparent, and thus projective geometry is at the heart of personal imaging.

Personal imaging situates the camera such that it provides a unique first-person perspective, that is, in the case of the eyeglass-mounted camera, the machine captures the world from the same perspective as its host (human).

In this chapter, I emphasize the importance of projective geometry, and present some new results that are germane to the principles of personal imaging, and are used in such applications as "painting with looks" (building environment maps by looking around), wearable tetherless computer-mediated reality, and the new genre of personal documentary that arises from this new perspective.

4.1 Introduction

I present direct featureless methods for estimating the 8 parameters of an "exact" projective (homographic) coordinate transformation to register pairs of images, together with the application of seamlessly combining a plurality of images of the same scene, resulting in a single image (or new image sequence) of greater resolution or spatial extent. The approach is "exact" for two cases of static scenes: (1) images taken from the same location of an arbitrary 3-D scene, with a camera that is free to pan, tilt, rotate about its optical axis, and zoom or (2) images of a flat scene taken from arbitrary locations. The featureless projective approach generalizes inter-frame camera motion estimation methods which have previously used an *affine* model (which lacks the degrees of freedom to "exactly" characterize such phenomena as camera pan and tilt) and/or which have relied upon finding points of correspondence between the image frames. The featureless projective approach, which operates directly on the image pixels, is shown to be superior in accuracy and ability to enhance resolution. The proposed methods work well on image data collected from both good-quality and poor-quality video under a wide variety of conditions (sunny, cloudy, day, night). These new fully-automatic methods are also shown to be robust to deviations from the assumptions of static scene and no parallax.

Many problems require finding the coordinate transformation between two images of the same scene or object. Whether to recover camera motion between video frames, to stabilize video images,

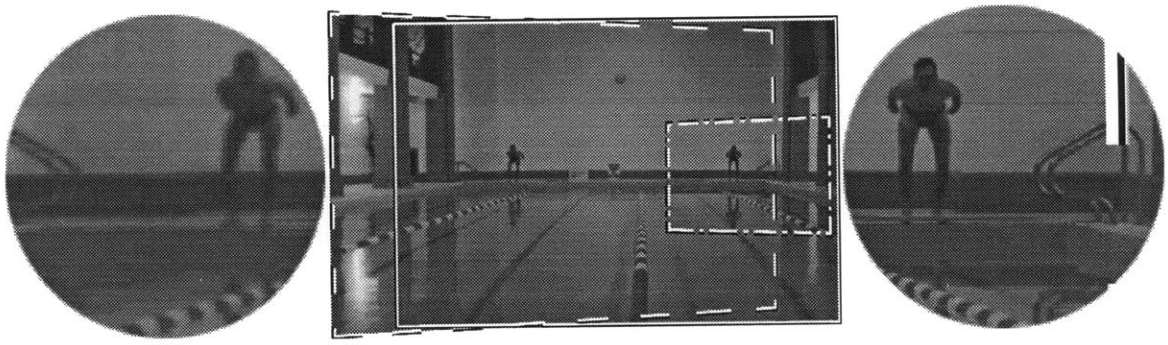


Figure 4-1: Image composite made from three pictures (moving between two different locations) in a large room: one was taken looking straight ahead (outlined in a solid line), one was taken panning to the left (outlined in a dashed line), and the third was taken panning to the right with substantial zoom-in (outlined in a dot-dash line). The second two have undergone a coordinate transformation to put them into the same coordinates as the one outlined in a solid line (the *reference frame*). This composite, made from NTSC-resolution images, occupies about 2000 pixels across and, in places, shows good detail down to the pixel level. Note increased sharpness in regions visited by the zooming-in, compared to other areas. (See magnified portions of composite at sides.) This composite only shows the result of combining three images, but in the final production, many more images were used, resulting in a high resolution full-color composite showing most of the room. (Figure reproduced from [28], courtesy of IS&T.)

to relate or recognize photographs taken from two different cameras, to compute depth within a 3-D scene, or for image registration and resolution enhancement, it is important to have both a precise description of the coordinate transformation between a pair of images or video frames, and some indication as to its accuracy.

Traditional *block matching* (e.g. as used in *motion estimation*) is really a special case of a more general *coordinate transformation*. In this chapter I demonstrate a new solution to the *motion estimation* problem using a more general estimation of a coordinate transformation, and propose techniques for automatically finding the 8-parameter projective coordinate transformation that relates two frames taken of the same static scene. I show, both by theory and example, how the new approach is more accurate and robust than previous approaches which relied on affine coordinate transformations, approximations to projective coordinate transformations, and/or the finding of point correspondences between the images. The new techniques take as input two frames, and automatically output the 8 parameters of the "exact" model, to properly register the frames. They do not require the tracking or correspondence of explicit features, yet are computationally easy to implement.

Although the theory I present makes the typical assumptions of static scene and no parallax, I show that the new estimation techniques are robust to deviations from these assumptions. In particular, I apply the direct featureless projective parameter estimation approach to image resolution enhancement and compositing, illustrating its success on a variety of practical and difficult cases, including some that violate the non-parallax and static scene assumptions.

An example image composite, made with featureless projective parameter estimation, is reproduced in Fig 4-1, where the spatial extent of the image is increased by panning the camera while compositing (e.g. by making a *panorama*) and the spatial resolution is increased by zooming the camera and by combining overlapping frames from different viewpoints.

4.2 Background

Hundreds of papers have been published on the problems of motion estimation and frame alignment. (For review and comparison, see [35].) In this section I review the basic differences between coordinate transformations and emphasize the importance of using the "exact" 8-parameter projective coordinate transformation.

Model	Coordinate transformation from $\mathbf{x}$ to $\mathbf{x}'$	Parameters
Translation	$\mathbf{x}' = \mathbf{x} + \mathbf{b}$	$\mathbf{b} \in \mathbb{R}^2$
Affine	$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$	$\mathbf{A} \in \mathbb{R}^{2\times 2}, \mathbf{b} \in \mathbb{R}^2$
Bilinear	$x' = q_{x'xy}xy + q_{x'x}x + q_{x'y}y + q_{x'}$ $y' = q_{y'xy}xy + q_{y'x}x + q_{y'y}y + q_{y'}$	$bfq_* \in \mathbb{R}$
Projective	$\mathbf{x}' = \frac{\mathbf{A}\mathbf{x}+\mathbf{b}}{\mathbf{c}^T\mathbf{x}+1}$	$\mathbf{A} \in \mathbb{R}^{2\times 2}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^2$
Relative-projective	$\mathbf{x}' = \frac{\mathbf{A}\mathbf{x}+\mathbf{b}}{\mathbf{c}^T\mathbf{x}+1} + \mathbf{x}$	$\mathbf{A} \in \mathbb{R}^{2\times 2}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^2$
Pseudo-perspective	$x' = q_{x'x}x + q_{x'y}y + q_{x'} + q_\alpha x^2 + q_\beta xy$ $y' = q_{y'x}x + q_{y'y}y + q_{y'} + q_\alpha xy + q_\beta y^2$	$\mathbf{q}_* \in \mathbb{R}$
Biquadratic	$x' = q_{x'x^2}x^2 + q_{x'xy}xy + q_{x'y^2}y^2 + q_{x'x}x + q_{x'y}y + q_{x'}$ $y' = q_{y'x^2}x^2 + q_{y'xy}xy + q_{y'y^2}y^2 + q_{y'x}x + q_{y'y}y + q_{y'}$	$bfq_* \in \mathbb{R}$

Table 4.1: Image coordinate transformations discussed in this chapter

4.2.1 Coordinate transformations

A coordinate transformation maps the image coordinates, $\mathbf{x} = [x, y]^T$ to a new set of coordinates, $\mathbf{x}' = [x', y']^T$. The approach to "finding the coordinate transformation" relies on assuming it will take one of the forms in Table 4.1, and then estimating the parameters (2 to 12 parameters depending on the model) in the chosen form. An illustration showing the effects possible with each of these forms is shown in Fig. 4-3.

The most common assumption (especially in motion estimation for coding, and optical flow for computer vision) is that the coordinate transformation between frames is translation. Tekalp, Ozkan, and Sezan [30] have applied this assumption to high-resolution image reconstruction. Although translation is the least constraining and simplest to implement of the seven coordinate transformations in Table 4.1, it is poor at handling large changes due to camera zoom, rotation, pan and tilt.

Zheng and Chellappa [36] considered the image registration problem using a subset of the affine model — translation, rotation and scale. Other researchers [31][37] have assumed affine motion (six parameters) between frames. For the assumptions of static scene and no parallax, the affine model exactly describes rotation about the optical axis of the camera, zoom of the camera, and pure shear, which the camera does not do, except in the limit as the lens focal length approaches infinity. The affine model cannot capture camera pan and tilt, and therefore cannot properly express the "keystoning" and "chirping" we see in the real world. (By "chirping" I mean the effect of increasing or decreasing spatial frequency with respect to spatial location, as illustrated in Fig 4-2.) Consequently, the affine model attempts to fit the wrong parameters to these effects. Even though it has fewer parameters, I find that the affine model is more susceptible to noise because it lacks the correct degrees of freedom needed to properly track the actual image motion.

The 8-parameter *projective* model gives the desired 8 parameters that exactly account for all possible zero-parallax camera motions; hence, there is an important need for a featureless estimator of these parameters. To the best of my knowledge, the only algorithms proposed to date for such an estimator are [28], and shortly after, [38]. In both of these, a computationally expensive nonlinear optimization method was presented. In the earlier [28], a direct method was also proposed. This direct method uses simple linear algebra, and is non-iterative insofar as methods such as Levenberg-Marquardt and the like are in no way required. The proposed method instead uses repetition with the correct law of composition on the projective group, going from one pyramid level to the next by application of the group's law of composition. Because the parameters of the projective coordinate transformation had traditionally been thought to be mathematically and computationally too difficult to solve, most researchers have used the simpler affine model or other approximations to the projective model. Before I propose and demonstrate the featureless estimation of the parameters of the "exact" projective model, it is helpful to discuss some approximate models.

Going from first order (affine), to second order, gives the 12-parameter 'biquadratic' model. This model properly captures both the chirping (change in spatial frequency with position) and

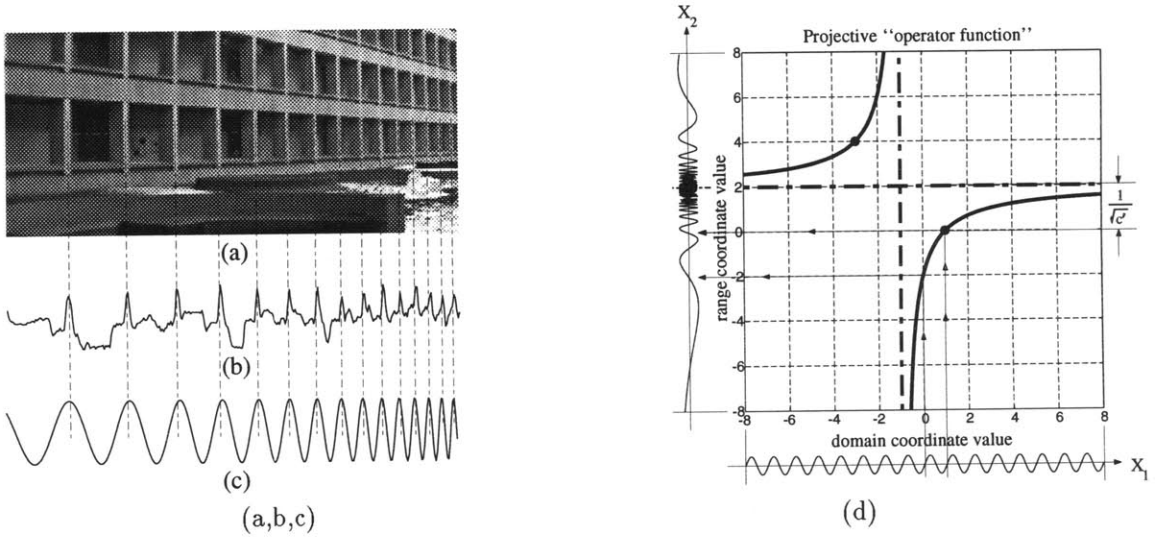


Figure 4-2: **The 'projective chirping' phenomenon.** (a) A real-world object that exhibits periodicity generates a projection (image) with "chirping" — 'periodicity-in-perspective'. (b) Center raster of image. (c) Best-fit projective chirp of form $\sin(2\pi((ax+b)/(cx+1)))$. (d) Graphical depiction of exemplar 1-D projective coordinate transformation of $\sin(2\pi x_1)$ into a 'projective chirp' function, $\sin(2\pi x_2) = \sin(2\pi((2x_1-2)/(x_1+1)))$. The range coordinate as a function of the domain coordinate forms a rectangular hyperbola with asymptotes shifted to center at the *vanishing point* $x_1 = -1/c = -1$ and 'exploding point', $x_2 = a/c = 2$, and with 'chirpiness' $c' = c^2/(bc-a) = -1/4$.

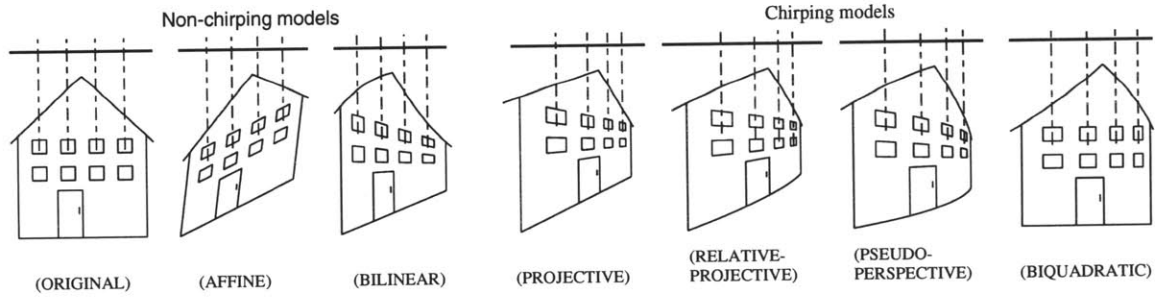


Figure 4-3: Pictorial effects of the six coordinate transformations of Table 4.1, arranged left to right by number of parameters. Note that translation leaves the ORIGINAL house figure unchanged, except in its location. Most importantly, only the four rightmost coordinate transformations affect the periodicity of the window spacing (inducing the desired "chirping" which corresponds to what we see in the real world). Of these four, only the PROJECTIVE coordinate transformation preserves straight lines. The 8-parameter PROJECTIVE coordinate transformation "exactly" describes the possible image motions ("exact" meaning under the idealized zero-parallax conditions).

converging lines (keystoning) effects associated with projective coordinate transformations, but does not constrain chirping and converging to work together (the example in Fig 4-3 being chosen with zero convergence yet substantial chirping, illustrates this point). Despite its larger number of parameters, there is still considerable discrepancy between a projective coordinate transformation and the best-fit biquadratic coordinate transformation. Why stop at 2nd order? Why not use a 20-parameter 'bicubic model? While an increase in the number of model parameters will result in a better fit, there is a tradeoff, where the model begins to fit noise. The physical camera model fits exactly in the 8-parameter projective group; therefore, we know that "eight is enough." Hence, it seems reasonable to have a preference for approximate models with exactly eight parameters.

The 8-parameter bilinear model is perhaps the most widely-used [39] in the fields of image processing, medical imaging, remote sensing, and computer graphics. This model is easily obtained from the biquadratic model by removing the four x^2 and y^2 terms. Although the resulting bilinear model captures the effect of converging lines, it completely fails to capture the effect of chirping.

The 8-parameter *pseudo-perspective* model [40] and an 8 parameter 'relative-projective' model both do, in fact, capture both the converging lines and the chirping of a projective coordinate transformation. The pseudo-perspective model, for example, may be thought of as first, removal of two of the quadratic terms ($bfq_{x'y^2} = q_{y'x^2} = 0$), which results in a ten parameter model (the 'q-chirp' of [41]) and then constraining the four remaining quadratic parameters to have two degrees of freedom. These constraints force the "chirping effect" (captured by $q_{x'x^2}$ and $q_{y'y^2}$) and the "converging effect" (captured by $q_{x'xy}$ and $q_{y'xy}$) to work together in the "right" way to match, as closely as possible, the effect of a projective coordinate transformation. By setting $q_\alpha = q_{x'x^2} = q_{y'xy}$, the chirping in the x-direction is forced to correspond with the converging of parallel lines in the x-direction (and likewise for the y-direction).

Of course, the desired "exact" eight parameters come from the projective model, but they have been perceived as being notoriously difficult to estimate. The parameters for this model have been solved by Tsai and Huang [42], but their solution assumed that features had been identified in the two frames, along with their correspondences. The main contribution of this chapter is a simple featureless means of automatically solving for these 8 parameters.

Other researchers have looked at projective estimation in the context of obtaining $3-D$ models. Faugeras and Lustman [43], Shashua and Navab [44], and Sawhney [45] have considered the problem of estimating the projective parameters while computing the motion of a rigid planar patch, as part of a larger problem of finding 3-D motion and structure using parallax relative to an arbitrary plane in the scene. Kumar *et al.* [46] have also suggested registering frames of video by computing the flow along the *epipolar* lines, for which there is also an initial step of calculating the gross camera movement assuming no parallax. However, these methods have relied on feature correspondences, and were aimed at 3-D scene modeling. My focus is not on recovering the 3-D scene model, but on aligning 2-D images of 3-D scenes. Feature correspondences greatly simplify the problem; however, they also have many problems. The focus of this chapter is simple featureless approaches to estimating the projective coordinate transformation between image pairs.

4.2.2 Camera motion: common assumptions and terminology

Two assumptions are typical in this area of research. The first assumption is that the scene is constant – changes of scene content and lighting are small between frames. The second assumption is that of an ideal pinhole camera – implying unlimited depth of field with everything in focus (infinite resolution) and implying that straight lines map to straight lines[1]. Consequently, the camera has three degrees of freedom in 2-D space and eight degrees of freedom in 3-D space: translation (X, Y, Z), zoom (scale in each of the image coordinates x and y), and rotation (rotation about the optical axis, pan, and tilt. These two assumptions are also made in this chapter.

In this chapter, an "uncalibrated camera" refers to one in which the principal point[2] is not

[1] When using low cost wide-angle lenses, there is usually some barrel distortion which we correct using the method of [47].

[2] The principal point is where the optical axis intersects the film.

	Scene assumptions	Camera assumptions
Case 1:	arbitrary 3-D	free to zoom, rot., pan, and tilt, fixed COP
Case 2:	planar	free to zoom, rot., pan, and tilt, free to trans.

Table 4.2: The two "no parallax" cases for a static scene. Note that the first situation has 7 degrees of freedom (yaw, pitch, roll, translation in each of the 3 spatial axes, and zoom), while the second has four degrees of freedom (pan, tilt, rotate, and zoom). Both, however, are represented within the 8 scalar paramters of the projective group of coordinate transformations.

necessarily at the center (origin) of the image and the scale is not necessarily isotropic[3] I assume that the zoom is continually adjustable by the camera user, and that we do not know the zoom setting, or whether it changed between recording frames of the image sequence. I also assume that each element in the camera sensor array returns a quantity that is linearly proportional to the quantity of light received[4]. With these assumptions, the exact camera motion that can be recovered is summarized in Table 4.2.

4.2.3 Video orbits

Tsai and Huang [42] pointed out that the elements of the projective *group* give the true camera motions with respect to a planar surface. They explored the group structure associated with images of a 3-D rigid planar patch, as well as the associated *Lie algebra*, although they assume that the correspondence problem has been solved. The solution presented in this chapter (which does not require prior solution of correspondence) also relies on projective group theory. I briefly review the basics of this theory, before presenting the new solution in the next section.

Projective group in 1-D coordinates

A group is a set upon which there is defined an associative law of composition (*closure, associativity*), which contains at least one element (*identity*) who's composition with another element leaves it unchanged, and for which every element of the set has an *inverse*.

A *group* of operators together with a *set* of operands form a so-called *group operation*[5].

In this chapter, coordinate transformations are the operators (group), and images are the operands (set). When the coordinate transformations form a group, then two such coordinate transformations, $\mathbf{p_1}$ and $\mathbf{p_2}$, acting in succession, on an image (e.g. $\mathbf{p_1}$ acting on the image by doing a coordinate transformation, followed by a further coordinate transformation corresponding to $\mathbf{p_2}$, acting on that result) can be replaced by a single coordinate transformation. That single coordinate transformation is given by the *law of composition* in the group.

The *orbit* of a particular element of the set, under the group operation [49] is the new set formed by applying to it, all possible operators from the group.

In this chapter, the orbit is a collection of pictures formed from one picture through applying all possible projective coordinate transformations to that picture. I refer to this set as the 'video orbit' of the picture in question. Image sequences generated by zero-parallax camera motion on a static scene contain images that all lie in the same video orbit.

For simplicity, I review the theory first for the projective coordinate transformation in one dimension[6].

Suppose we take two pictures, using the same exposure, of the same scene from fixed common location (e.g. where the camera is free to pan, tilt, and zoom between taking the two pictures).

[3] Isotropic means that magnification in the x and y directions is the same. Our assumption facilitates aligning frames taken from different cameras.

[4] This condition can be enforced over a wide range of light intensity levels, by using the Wyckoff principle [27][48].

[5] also known as a *group action* or *G-set* [49].

[6] In this 2-D world, the "camera" consists of a center of projection (pinhole "lens") and a line (1-D sensor array or 1-D "film").

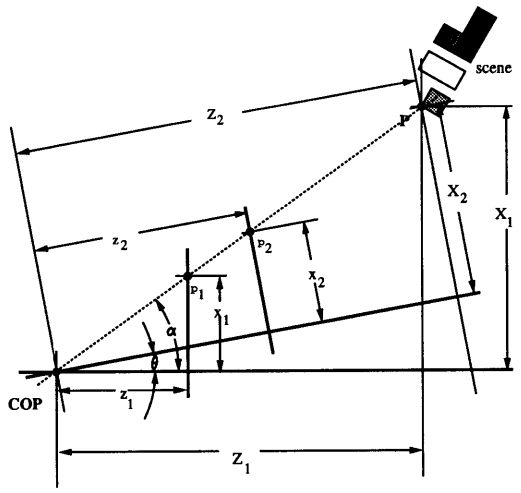


Figure 4-4: Camera at a fixed location. An arbitrary scene is photographed twice, each time with a different camera orientation, and a different principal distance (zoom setting). In both cases the camera is located at the same place (COP) and thus captures the same pencil of light. The dotted line denotes a ray of light traveling from an arbitrary point, **P**, in the scene, to the **COP**. Heavy lines denote both camera optical axes in each of the two orientations as well as the image sensor in each of its two pan and zoom positions. The two image sensors (or films) are in front of the camera to simplify mathematical derivations.

Both of the two pictures capture the same pencil of light[7], but each one projects this information differently onto the film or image sensor. Neglecting that which falls beyond the borders of the pictures, each picture captures the same information about the scene, but records it in a different way. The same object might, for example, appear larger in one image than in the other, or might appear more squashed at the left and stretched at the right than in the other. Thus we would expect to be able to construct one image from the other, so that only one picture should need to be taken (assuming its field of view covers all the objects of interest) in order to synthesize all the others. We first explore this idea in a make-believe "Flatland" where objects exist on the 2-D page, rather than the 3-D world in which we live, and where pictures are real-valued functions of one real variable, rather than the more familiar real-valued functions of two real-variables.

For the two pictures of the same pencil of light in Flatland, I define the common COP at the origin of our coordinate system in the plane. In Fig. 4-4 I have depicted a single camera that takes two pictures in succession as two cameras shown together in the same figure. Let $Z_k, k \in \{1,2\}$ represent the distances, along each optical axis, to an arbitrary point in the scene, P, and let X_k represent the distances from P to each of the optical axes. The principal distances are denoted z_k. In the example of Fig. 4-4, we are *zooming in* (increased magnification) as we go from frame 1 to frame 2.

The geometry of Fig. 4-4 defines a mapping from x_1 to x_2, given by [50][13]:

$$\begin{aligned} x_2 &= z_2 \tan(\arctan(x_1/z_1) - \theta), \ \forall x_1 \neq o_1 \\ &= (ax_1 + b)/(cx_1 + 1), \ \forall x_1 \neq o_1 \end{aligned} \tag{4.1}$$

where $a = z_2/z_1$, $b = -z_2 \tan(\theta)$, $c = \tan(\theta)/z_1$, and $o_1 = z_1 \tan(\pi/2 + \theta) = -1/c$, is the location of the singularity in the domain ('appearing point' [13]). I should emphasize here that if we set $c = 0$ we arrive at the affine group. (Recall, also, that c, the degree of perspective, has been given the interpretation of a chirp-rate [50].)

Let $\mathbf{p} \in \mathbf{P}$ denote a particular mapping from x_1 to x_2, governed by the three parameters $\mathbf{p}' = [z_1, z_2, \theta]$, or equivalently by a, b and c from (4.1).

[7]We neglect the boundaries (edges or ends of the sensor) and assume that both pictures have sufficient field of view to capture all of the objects of interest.

Proposition 1 *The set of all possible operators,* $\mathbf{P}_1$*, given by the coordinate transformations (4.1),* $\forall a \neq bc$*, acting on a set of 1-D images, forms a group-operation.*

Proof: A pair of images produced by a particular camera rotation and change in principal distance (depicted in Fig. 4-4) is an operator that takes any function g on image line 1, to a function, h on image line 2:

$$\begin{aligned} h(x_2) &= g(x_1) = g((-x_2 + b)/(cx_2 - a)), \ \forall x_2 \neq o_2 \\ &= g \circ x_1 = g \circ \mathbf{p}^{-1} \circ x_2 \end{aligned} \tag{4.2}$$

where $\mathbf{p} \circ x = (ax + b)/(cx + 1)$ and $o_2 = a/c$. As long as $a \neq bc$, each operator, $\mathbf{p}$, has an inverse, namely that given by composing the inverse coordinate transformation:

$$x_1 = (b - x_2)/(cx_2 - a), \ \forall x_2 \neq o_2 \tag{4.3}$$

with the function $h()$ to obtain $g = h \circ \mathbf{p}$. The identity operation is given by $g = g \circ e$, where e is given by $a = 1$, $b = 0$, and $c = 0$.

In complex analysis, (see for example, Ahlfors [51]) the form $(az + b)/(cz + d)$ is known as a linear fractional transformation. Although our mapping is from $\mathbb{R}$ to $\mathbb{R}$ (as opposed to theirs from $\mathbb{C}$ to $\mathbb{C}$), I can still borrow the concepts of complex analysis. In particular, a simple group-representation is provided using the 2×2 matrices, $\mathbf{p} = [a, b; c, 1] \in \mathbb{R}^2 \times \mathbb{R}^2$. Closure[8] and associativity are obtained by using the usual laws of matrix multiplication followed with dividing the resulting vector's first element by its second element. □

Proposition 1 says that an element of the $(ax + b)/(cx + 1)$ group can be used to align any two frames of the (1-D) image sequence provided that the COP remains fixed.

Proposition 2 *The set of operators that take nonsingular projections of a straight object to one another form a group,* $\mathbf{P_2}$*.*

A "straight" object is one which lies on a straight line in Flatland[9].

Proof: Consider a geometric argument. The mapping from the first (1-D) frame of an image sequence, $g(x_1)$ to the next frame, $h(x_2)$ is parameterized by the following: camera translation perpendicular to the object, t_z; camera translation parallel to the object, t_x; pan of frame 1, θ_1; pan of frame 2, θ_2; zoom of frame 1, z_1; and zoom of frame 2, z_2. (See Fig. 4-5.) We want to obtain the mapping from x_1 to x_2. Let's begin with the mapping from X_2 to x_2:

$$x_2 = z_2 \tan(\arctan(X_2/Z_2) - \theta_2) = \frac{a_2 X_2 + b_2}{c_2 X_2 + 1} \tag{4.4}$$

which can be represented by the matrix $\mathbf{p}_2 = [a_2, b_2; c_2, 1]$, so that $x_2 = \mathbf{p}_2 \circ X_2$. Now $X_2 = X_1 - t_x$ and it is clear that this coordinate transformation is inside the group, for there exists the choice of $a = 1$, $b = -t_x$, and $c = 0$ that describe it: $X_2 = \mathbf{p}_t \circ X_1$, where $\mathbf{p}_t = [1, -t_x; 0, 1]$. Finally, $x_1 = z_1 \tan(\arctan(X_1/Z_1) - \theta) = \mathbf{p}_1 \circ X_1$. Let $\mathbf{p}_1 = [a_1, b_1; c_1, 1]$. Then $\mathbf{p} = \mathbf{p}_2 \circ \mathbf{p}_t \circ \mathbf{p}_1^{-1}$ is in the group by the law of composition. Hence, the operators that take one frame into another, $x_2 = \mathbf{p} \circ x_1$, form a group. □

Proposition 2 says that an element of the $(ax + b)/(cx + 1)$ group can be used to align any two images of linear objects in flatland, regardless of camera movement.

Proposition 3 *The two groups* $\mathbf{P_1}$ *and* $\mathbf{P_2}$ *are isomorphic; a group-representation for both is given by the* 2×2 *square matrix* $[a, b; c, 1]$*.*

[8] Also know as *law of composition* [49]

[9] An important difference to keep in mind, with respect to pictures of a flat object, is that in Flatland, if you take a picture of a picture, that is equivalent to a single picture for an equivalent camera orientation and position. However, with 2-D pictures in a 3-D world, a picture of a picture is, in general, not necessarily a simple perspective projection (however, if you continue taking pictures you do not get anything new beyond the second picture). However, the 2-D version of the group representation contains both cases.

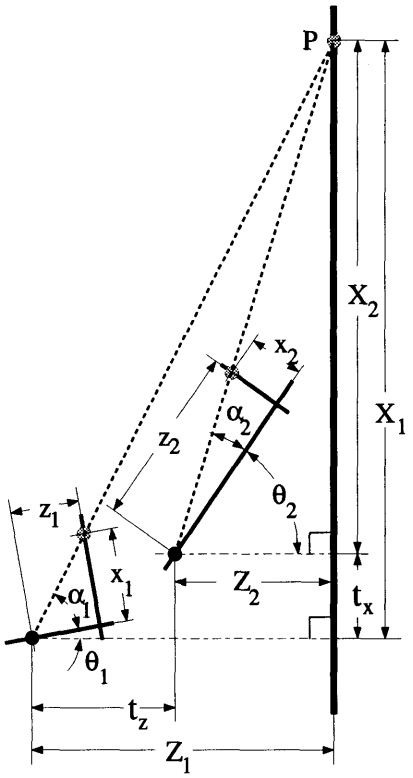


Figure 4-5: Two pictures of a flat (straight) object. The point P is imaged twice, each time with a different camera orientation, a different principal distance (zoom setting), and different camera location (resolved into components parallel and perpendicular to the object).

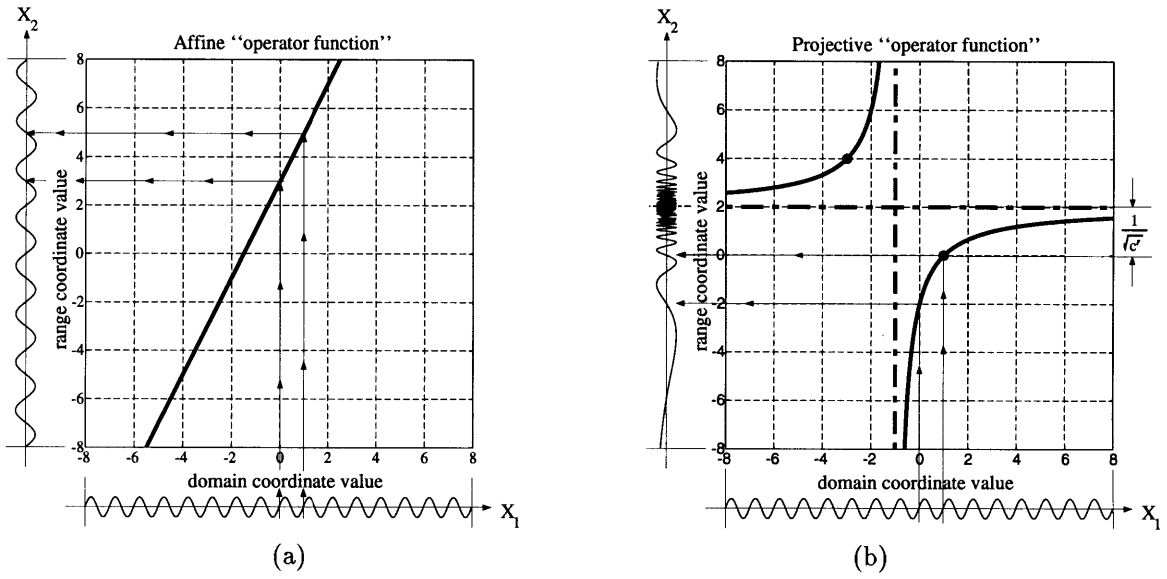


Figure 4-6: Comparison of 1-D affine and projective coordinate transformations, in terms of their 'operator functions', acting on a sinusoidal image. Note that whether the function is enlarged, or "chirped", the same information remains present, but is simply recorded in a different way by the new function. (a) Orthographic projection is equivalent to affine coordinate transformation, $y = ax + b$. In this example, $a = 2$ and $b = 3$. (b) Perspective projection for a particular fixed value of $\mathbf{p}' = \{1, 2, 45°\}$. Note that the plot is a rectangular hyperbola like $x_2 = 1/(c'x_1)$ but with asymptotes at the shifted origin $(-1, 2)$. Here $g(x_1) = \sin(2\pi x_1)$. The arrows indicate how a chosen cycle of this sine wave is mapped to the corresponding cycle of the 'P-chirp', $h(x_2)$.

Isomorphism follows because $\mathbf{P}_1$ and $\mathbf{P}_2$ have the same group representation[10]. The $(ax+b)/(cx+1)$ operators in the above propositions form the *projective group* $\mathbf{P}$ in Flatland.

Previously I emphasized the fact that the affine operator that takes a function space G to a function space H may itself be viewed as a function. Let us now construct a similar plot for a member of the group of operators, $\mathbf{p} \in \mathbf{P}$, in particular, the operator $\mathbf{p} = [2, -2; 1, 1]$ which corresponds to $\mathbf{p}' = \{1, 2, 45°\} \in \mathbf{P}_1$. We have also depicted the result of mapping $g(x_1) = \sin(2\pi x_1)$ to $h(x_2)$. When G is the space of Fourier analysis functions (harmonic oscillations), then H is a family of functions known as P-chirps [50], adapted to a particular *vanishing point*, o_2 and 'normalized chirp-rate', $c' = c^2/(bc - a)$ [13]. Fig. 4-6(b), is a *rectangular hyperbola* (e.g. $x_2 = \frac{1}{c'x_1}$) with an origin that has been shifted from $(0, 0)$ to (o_1, o_2).

A member of this group of coordinate transformations: $x' = (ax + b)/(cx + d)$, $\forall ad \neq bc$ (where the images are functions of one variable, x) is denoted by $p_{a,b,c,d}$, and has inverse $p_{-d,b,,c,-a}$. The law of composition is given by $p_{e,f,g,h} \circ p_{a,b,c,d} = p_{ae+cf,be+df,ag+cd,bg+d^2}$. In almost all practical engineering applications, $d \neq 0$, so I will divide through by d, and denote the coordinate transformation $x' = (ax + b)/(cx + 1)$ by $x' = p_{a,b,c} \circ x$. When $a \neq 0$ and $c = 0$, the projective group becomes the affine group of coordinate transformations, and when $a = 1$ and $c = 0$, it becomes the group of translations.

Of the coordinate transformations presented in the previous section, only the projective, affine, and translation operations form groups.

The equivalent two cases of Table 4.2 for this hypothetical "flatland" world of 2-D objects with 1-D pictures correspond to the following. In the first case a camera is at a fixed location, and free to zoom and pan. In the second case, a camera is free to translate, zoom, and pan, but the imaged object must be flat (i.e., lie on a straight line in the plane). The resulting two (1-D) frames taken

[10] For 2-D images in a 3-D world, the isomorphism no longer holds. However, the group still *contains* and therefore represents both cases.

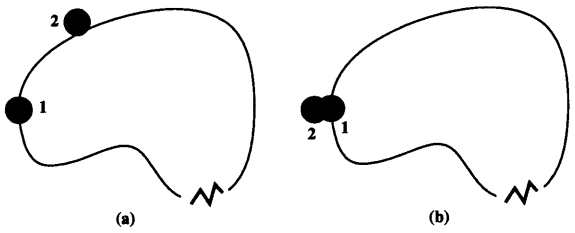


Figure 4-7: Video orbits. (a) The orbit of frame 1 is the set of all images that can be produced by acting on frame 1 with any element of the operator group. Assuming that frames 1 and 2 are from the same scene, frame 2 will be close to one of the possible projective coordinate transformations of frame 1. In other words, frame 2 "lies near the orbit of" frame 1. (b) By bringing frame 2 along its orbit, we can determine how closely the two orbits come together at frame 1.

by the camera, are related by the coordinate transformation from x_1 to x_2, given by [50]:

$$\begin{aligned} x_2 &= z_2 \tan(\arctan(x_1/z_1) - \theta), \ \forall x_1 \neq o_1 \\ &= (ax_1 + b)/(cx_1 + 1), \ \forall x_1 \neq o_1 \end{aligned} \tag{4.5}$$

where $a = z_2/z_1$, $b = -z_2 \tan(\theta)$, $c = \tan(\theta)/z_1$, and $o_1 = z_1 \tan(\pi/2 + \theta) = -1/c$, is the location of the singularity in the domain. We should mention that c, the degree of perspective, has been given the interpretation of a chirp-rate [50].

The coordinate transformations of (4.5) form a group operation. This result, and the proof of this group's isomorphism to the group corresponding to nonsingular projections of a flat object are given in [32].

Projective group in 2-D coordinates

The theory for the projective, affine, and translation groups also holds for the familiar 2-D images taken of the 3-D world. The 'video orbit' of a given 2-D frame is defined to be the set of all images that can be produced by applying operators from the 2-D projective group to the given image. Hence, I restate the coordinate transformation problem: Given a set of images that lie in the same orbit of the group, I wish to find for each image pair, that operator in the group which takes one image to the other image.

If two frames, say, f_1 and f_2, are in the same orbit, then there is an group operation $\mathbf{p}$ such that the mean-squared error (MSE) between f_1 and $f_2' = \mathbf{p} \circ f_2$ is zero. In practice, however, I find which element of the group takes one image "nearest" the other, for there will be a certain amount of parallax, noise, interpolation error, edge effects, changes in lighting, depth of focus, etc. Fig. 4-7 illustrates the operator $\mathbf{p}$ acting on frame f_2, to move it nearest to frame f_1. (This figure does not, however, reveal the precise shape of the orbit, which occupies an 8-D space.)

Summarizing, the 8-parameter projective group captures the exact coordinate transformation between pictures taken under the two cases of Table 4.2. The primary assumptions in these cases are that of no parallax, and of a static scene. Because the 8-parameter projective model is "exact," it is theoretically the right model to use for estimating the coordinate transformation. Examples presented in this chapter demonstrate that it also performs better in practice than the other proposed models.

4.3 Framework: motion parameter estimation and optical flow

To lay the framework for my new results, I will review existing methods of parameter estimation for coordinate transformations. This framework will apply to both existing methods as well as the new

methods. The purpose of this review is to bring together a variety of methods that appear quite different, but which actually can be described in a more unified framework which I present here.

The framework I give breaks existing methods into two categories: feature-based, and featureless. Of the featureless methods, I consider two subcategories: 1) methods based on minimizing MSE (generalized correlation, direct nonlinear optimization) and 2) methods based on spatiotemporal derivatives and optical flow. Note that variations such as *multiscale* have been omitted from these categories; multiscale analysis can be applied to any of them. The new algorithms I propose in this chapter (with final form given in Sec. 4.4) are featureless, and based on (multiscale if desired) spatiotemporal derivatives.

Some of the descriptions of methods below will be presented for hypothetical 1-D images taken of 2-D "scenes" or "objects". This simplification yields a clearer comparison of the estimation methods.

The new theory and applications will be presented subsequently for 2-D images taken of 3-D scenes or objects.

4.3.1 Feature-based methods

Feature-based methods [52][17] assume that point correspondences in both images are available. In the projective case, given at least three correspondences between point pairs in the two 1-D images, I will find the element, $\mathbf{p} = \{a, b, c\} \in \mathbf{P}$ that maps the second image into the first. Let $x_k, k = 1, 2, 3, \ldots$ be the points in one image, and let x'_k be the corresponding points in the other image. Then: $x'_k = (ax_k + b)/(cx_k + 1)$. Re-arranging yields $ax_k + b - x_k x'_k c = x'_k$, so that a, b, and c can be found by solving $k \geq 3$ linear equations in 3 unknowns:

$$\begin{bmatrix} x_k & 1 & -x'_k x_k \end{bmatrix} \begin{bmatrix} a & b & c \end{bmatrix}^T = \begin{bmatrix} x'_k \end{bmatrix} \tag{4.6}$$

using least squares if there are more than three correspondence points. The extension from 1-D "images" to 2-D images is conceptually identical; for the affine and projective models, the minimum number of correspondence points needed in 2-D is three and four respectively.

A major difficulty with feature-based methods is finding the features. Good features are often hand-selected, or computed, possibly with some degree of human intervention [53]. A second problem with features is their sensitivity to noise and occlusion. Even if reliable features exist between frames (e.g. line markings on a playing field in a football video, see Sec. 4.5.2) these features may be subject to signal noise and occlusion (e.g. running football players blocking a feature). The emphasis in the rest of this chapter will be on robust featureless methods.

4.3.2 Featureless methods based on generalized cross-correlation

The purpose of this subsection is for completeness: we'll consider first what is perhaps the most most obvious approach (generalized cross-correlation in 8-D parameter space) in order to motivate a different approach provided in Sec 4.3.3, the motivation arising from ease of implemention and simplicity of computation.

Cross-correlation of two frames is a featureless method of recovering translation model parameters. Affine and projective parameters can also be recovered using generalized forms of cross-correlation.

Generalized cross-correlation is based on an inner-product formulation which establishes a similarity metric between two functions, say, g and h, where $h \approx \mathbf{p} \circ g$ is an approximately coordinate-transformed version of g, but the parameters of the coordinate transformation, $\mathbf{p}$ are unknown.[11] We can find, by exhaustive search (applying all possible operators, $\mathbf{p}$, to h), the "best" $\mathbf{p}$ as the one which maximizes the inner product:

$$\int_{-\infty}^{\infty} g(x) \frac{\mathbf{p}^{-1} \circ h(x)}{\int_{-\infty}^{\infty} \mathbf{p}^{-1} \circ h(x) dx} dx \tag{4.7}$$

[11] In the presence of additive white Gaussian noise, this method, also known as "matched filtering", leads to a maximum likelihood estimate of the parameters [54].

where I have normalized the energy of each coordinate-transformed h before making the comparison. Equivalently, instead of maximizing a similarity metric, we can minimize some distance metric, such as MSE, given by $\int_{-\infty}^{\infty}(g(x) - \mathbf{p}^{-1} \circ h(x))^2 - Dx$. Solving (4.7) has an advantage over finding MSE when one image is not only a coordinate-transformed version of the other, but is also an amplitude-scaled version, as generally happens when there is an automatic gain control or an automatic iris in the camera.

In 1-D, the orbit of an image under the affine group operation is a family of *wavelets* (assuming the image is that of the desired "mother wavelet", in the sense that a wavelet family is generated by 1-D affine coordinate transformations of a single function) while the orbit of an image under the projective group of coordinate transformations is a family of 'projective chirplets' [55][12], the objective function (4.7) being the cross-chirplet transform. A computationally efficient algorithm for the cross-wavelet transform has recently been presented [58]. (See [59] for a good review on wavelet-based estimation of affine coordinate transformations.)

Adaptive variants of the chirplet transforms have been previously reported in the literature [60]. However, there are still many problems with the adaptive chirplet approach; thus, for the remainder of this chapter, we consider featureless methods based on spatiotemporal derivatives.

4.3.3 Featureless methods based on spatiotemporal derivatives

Optical flow ('translation flow')

When the change from one image to another is small, optical flow [61] may be used. In 1-D, the traditional optical flow formulation assumes each point x in frame t is a translated version of the corresponding point in frame $t + \Delta t$, and that Δx and Δt are chosen in the ratio $\Delta x/\Delta t = u_f$, the translational flow velocity of the point in question. The image brightness $E(x,t)$ is described by:

$$E(x,t) = E(x + \Delta x, t + \Delta t), \ \forall (x,t), \tag{4.8}$$

where u_f is the translational flow velocity of the point in In the case of pure translation, u_f is constant across the entire image. More generally, though, a pair of 1-D images are related by a quantity, $u_f(x)$ at each point in one of the images.

Expanding the right hand side of (4.8) in a Taylor series, and canceling 0th order terms gives the well-known optical flow equation: $u_f E_x + E_t + h.o.t. = 0$, where E_x and E_t are the spatial and temporal derivatives respectively, and $h.o.t.$ denotes higher order terms. Typically, the higher order terms are neglected, giving the expression for the optical flow at each point in one of the two images:

$$u_f E_x + E_t \approx 0 \tag{4.9}$$

Weighing the difference between "Affine fit" and "affine flow"

A comparison between two similar approaches is presented, in the familiar and obvious realm of linear regression versus direct affine estimation, highlighting the obvious differences between the two approaches. This difference, in weighting, motivates new weighting changes which will later simplify implementations pertaining to the new methods.

Given the optical flow between two images, g and h, I wish to find the coordinate transformation to apply to h to register it with g. We now describe two approaches based on the affine model[13]: (1) finding the optical flow at every point, and then fitting this flow with an affine model ('affine fit'), and (2) rewriting the optical flow equation in terms of an affine (not translation) motion model ('affine flow').

Wang and Adelson have proposed fitting an affine model to the optical flow field [62] between two 2-D images. I briefly examine their approach with 1-D images; the reduction in dimensions

[12] Symplectomorphisms of the time-frequency plane [56][57] have been applied to signal analysis [55], giving rise to the so-called q-chirplet [55], which differs from the projective chirplet discussed here.

[13] The 1-D affine model is a simple yet sufficiently interesting (non-Abelian) example selected to illustrate differences in weighting.

simplifies analysis and comparison to 'affine flow'. Denote coordinates in the original image, g, by x, and in the new image, h, by x'. Suppose that h is a dilated and translated version of g, so $x' = ax+b$ for every corresponding pair (x', x). Equivalently, the affine model of velocity (normalizing $\Delta t = 1$), $u_m = x' - x$, is given by $u_m = (a-1)x + b$. We can expect a discrepancy between the flow velocity, u_f, and the model velocity, u_m, due to either errors in the flow calculation, or to errors in the affine model assumption, so I apply linear regression to get the best least-squares fit by minimizing:

$$\varepsilon_{fit} = \sum_x (u_m - u_f)^2 = \sum (u_m + E_t/E_x)^2 \tag{4.10}$$

The constants a and b that minimize ε_{fit} over the entire patch are found by differentiating (4.10), and setting the derivatives to zero. This results in what I call the "affine fit" equations:

$$\begin{bmatrix} \sum_x x^2, \sum_x x \\ \sum_x x, \sum_x 1 \end{bmatrix} \begin{bmatrix} a-1 \\ b \end{bmatrix} = - \begin{bmatrix} \sum_x x E_t/E_x \\ \sum_x E_t/E_x \end{bmatrix} \tag{4.11}$$

Alternatively, the affine coordinate transformation may be directly incorporated into the brightness change constraint equation (4.8). Bergen *et al.* [63] have proposed this method, which I will call 'affine flow', to distinguish it from the 'affine fit' model of Wang and Adelson (4.11). Let us show how 'affine flow' and 'affine fit' are related. Substituting $u_m = (ax+b) - x$ directly into (4.9) in place of u_f and summing the squared error:

$$\varepsilon_{flow} = \sum_x (u_m E_x + E_t)^2 \tag{4.12}$$

over the whole image, differentiating, and equating the result to zero, gives a linear solution for both a and b:

$$\begin{bmatrix} \sum_x x^2 E_x^2, \sum_x x E_x^2 \\ \sum_x x E_x^2, \sum_x E_x^2 \end{bmatrix} \begin{bmatrix} a-1 \\ b \end{bmatrix} = - \begin{bmatrix} \sum_x x E_x E_t \\ \sum_x E_x E_t \end{bmatrix} \tag{4.13}$$

To see how this result compares to the 'affine fit' I rewrite (4.10)

$$\varepsilon_{fit} = \sum_x (\frac{u_m E_x + E_t}{E_x})^2 \tag{4.14}$$

and observe, comparing (4.12) and (4.14) that 'affine flow' is equivalent to a weighted least-squares fit, where the weighting is given by E_x^2. Thus the 'affine flow' method tends to put more emphasis on areas of the image that are spatially varying than does the 'affine fit' method. Of course, one is free to separately choose the weighting for each method in such a way that 'affine fit' and 'affine flow' methods both give the same result. Both my intuition and our practical experience tends to favor the 'affine flow' weighting, but, more generally, perhaps we should ask "What is the best weighting?" Lucas and Kanade [64], among others, have considered weighting issues, though the rather obvious difference in weighting between fit and flow doesn't appear to have been pointed out previously in the literature. The fact that the two approaches provide similar results, yet have drastically different weightings, suggests that we can exploit the choice of weighting. In particular, we will observe in Sec 4.3.3 that we can select a weighting that makes the implementation easier.

Another approach to the 'affine fit' involves computation of the optical flow field using the multiscale iterative method of Lucas and Kanade, and *then* fitting to the affine model. An analogous variant of the 'affine flow' method involves multiscale iteration as well, but in this case the iteration and multiscale hierarchy are incorporated directly into the affine estimator [63]. With the addition of multiscale analysis, the 'fit' and 'flow' methods differ in additional respects beyond just the weighting. My intuition and experience indicates that the direct multiscale 'affine flow' performs better than the 'affine fit' to the multiscale flow. Multiscale optical flow makes the assumption that blocks of the image are moving with pure translational motion, and then, paradoxically, the affine fit refutes this pure-translation assumption. However, 'fit' provides some utility over 'flow' when it is desired to segment the image into regions undergoing different motions [65], or to gain robustness

by rejecting portions of the image not obeying the assumed model.

'Projective fit' and 'projective flow': new techniques

Analogous to the "affine fit" and "affine flow" of the previous section, I now propose the two new methods: 'projective fit' and 'projective flow'. For the 1-D affine coordinate transformation, the graph of the range coordinate as a function of the domain coordinate is a straight line; for the projective coordinate transformation, the graph of the range coordinate as a function of the domain coordinate is a rectangular hyperbola (Fig 4-2(d)). The 'affine fit' case used linear regression; however, in the projective case I use 'hyperbolic regression.' Consider the flow velocity given by (4.9) and the model velocity:

$$u_m = x' - x = \frac{ax+b}{cx+1} - x \tag{4.15}$$

and minimize the sum of the squared difference as was done in (4.10):

$$\varepsilon = \sum_x (\frac{ax+b}{cx+1} - x + \frac{E_t}{E_x})^2 \tag{4.16}$$

As discussed earlier, the calculation can be simplified by judicious alteration of the weighting, in particular, multiplying each term of the summation (4.16) by $(cx+1)$, and solving, gives:

$$\left(\sum_x \phi(x)\phi^T(x)\right)[a,b,c]^T = \sum_x (x - E_t/E_x)\phi(x) \tag{4.17}$$

where the *regressor* is $\phi = [x, 1, xE_t/E_x - x^2]^T$.

For 'projective-flow' ('p-flow'), I substitute $u_m = \frac{ax+b}{cx+1} - x$ into (4.12). Again, weighting by $(cx+1)$ gives:

$$\varepsilon_w = \sum (axE_x + bE_x + c(xE_t - x^2E_x) + E_t - xE_x)^2 \tag{4.18}$$

(the subscript w denotes weighting has taken place) resulting in a linear system of equations for the parameters:

$$\left(\sum \phi_w \phi_w^T\right)[a,b,c]^T = \sum (xE_x - E_t)\phi_w \tag{4.19}$$

where $\phi_w = [xE_x, E_x, xE_t - x^2E_x]^T$. Again, to show the difference in the weighting between projective flow and projective fit, we can rewrite (4.19):

$$(\sum E_x^2 \phi\phi^T)[a,b,c]^T = \sum E_x^2 (xE_x - E_t)\phi \tag{4.20}$$

where ϕ is that defined in (4.17).

The unweighted projectivity estimator

If we do not wish to apply the ad-hoc weighting scheme, we may still estimate the parameters of projectivity in a simple manner, still based on solving a linear system of equations. To do this, we write the Taylor series of u_m:

$$u_m + x = b + (a - bc)x + (bc - a)cx^2 + (a - bc)c^2x^3 + \cdots \tag{4.21}$$

and use the first 3 terms, obtaining enough degrees of freedom to account for the 3 parameters being estimated. Letting $\varepsilon = \sum(-h.o.t.)^2 = \sum((b + (a - bc - 1)x + (bc - a)cx^2)E_x + E_t)^2$, $\mathbf{q}_2 = (bc - a)c$, $\mathbf{q}_1 = a - bc - 1$, and $\mathbf{q}_0 = b$, and differentiating with respect to each of the 3 parameters of $\mathbf{q}$, setting the derivatives equal to zero, and verifying with the second derivatives, gives the linear system of

equations for 'unweighted projective flow':

$$\begin{bmatrix} \sum x^4 E_x^2 & \sum x^3 E_x^2 & \sum x^2 E_x^2 \\ \sum x^3 E_x^2 & \sum x^2 E_x^2 & \sum x E_x^2 \\ \sum x^2 E_x^2 & \sum x E_x^2 & \sum E_x^2 \end{bmatrix} \begin{bmatrix} q_2 \\ q_1 \\ q_0 \end{bmatrix} = - \begin{bmatrix} \sum x^2 E_x E_t \\ \sum x E_x E_t \\ \sum E_x E_t \end{bmatrix} \tag{4.22}$$

In Sec. 4.4 I will extend this derivation to 2-D images.

4.4 Multiscale implementations in 2-D

In the previous section, two new techniques, 'projective-fit' and 'projective-flow' were proposed. Now I describe these algorithms for 2-D images. The brightness constancy constraint equation for 2-D images [61] which gives the flow velocity components in the x and y directions, analogous to (4.9) is:

$$\mathbf{u_f}^T \mathbf{E_x} + E_t \approx 0 \tag{4.23}$$

As is well-known [61] the optical flow field in 2-D is underconstrained[14]. The model of *pure translation* at every point has two parameters, but there is only one equation (4.23) to solve, thus it is common practice to compute the optical flow over some neighborhood, which must be at least two pixels, but is generally taken over a small block, 3×3, 5×5, or sometimes larger (e.g. the entire image, as in this chapter).

Our task is not to deal with the 2-D translational flow, but with the 2-D projective flow, estimating the eight parameters in the coordinate transformation:

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{\mathbf{A}[x,y]^T + \mathbf{b}}{\mathbf{c}^T[x,y]^T + 1} = \frac{\mathbf{Ax} + \mathbf{b}}{\mathbf{c}^T\mathbf{x} + 1} \tag{4.24}$$

The desired eight scalar parameters are denoted by $\mathbf{p} = [\mathbf{A}, \mathbf{b}; \mathbf{c}, 1]$, $\mathbf{A} \in \mathbb{R}^{2\times 2}$, $\mathbf{b} \in \mathbb{R}^{2\times 1}$, and $\mathbf{c} \in \mathbb{R}^{2\times 1}$.

Analogous to (4.14), we have, in the 2-D case:

$$\varepsilon_{flow} = \sum \left(\mathbf{u}\mathbf{m}^T \mathbf{E_x} + E_t\right)^2 = \sum \left((\frac{\mathbf{Ax} + \mathbf{b}}{\mathbf{c}^T\mathbf{x} + 1} - \mathbf{x})^T \mathbf{E_x} + E_t \right)^2 \tag{4.25}$$

Where the sum can be weighted as it was in the 1-D case:

$$\varepsilon_w = \sum \left((\mathbf{Ax} + \mathbf{b} - (\mathbf{c}^T\mathbf{x} + 1)\mathbf{x})^T \mathbf{E_x} + (\mathbf{c}^T\mathbf{x} + 1)E_t \right)^2 \tag{4.26}$$

Differentiating with respect to the free parameters $\mathbf{A}, \mathbf{b}$, and $\mathbf{c}$, and setting the result to zero gives a linear solution:

$$\left(\sum \phi\phi^T\right) [a_{11}, a_{12}, b_1, a_{21}, a_{22}, b_2, c_1, c_2]^T = \sum (\mathbf{x}^T \mathbf{E_x} - E_t)\phi \tag{4.27}$$

where $\phi^T = [E_x(x,y,1), E_y(x,y,1), xE_t - x^2E_x - xyE_y, yE_t - xyE_x - y^2E_y]$

4.4.1 'Unweighted projective flow'

As with the 1-D images, we make similar assumptions in expanding (4.24) in its own Taylor series, analogous to (4.21). If we take the Taylor series up to 2nd order terms, we obtain the biquadratic model mentioned in Sec. 4.2.1. As mentioned in Sec. 4.2.1, by appropriately constraining the twelve parameters of the biquadratic model we obtain a variety of 8-parameter approximate models. In my algorithms for estimating the 'exact unweighted' projective group parameters, I use one of these

[14] Optical flow in 1-D did not suffer from this problem.

approximate models in an intermediate step.[15]

The Taylor series for the bilinear case gives:

$$\begin{aligned} u_m + x &= q_{x'xy}xy + (q_{x'x} + 1)x + q_{x'y}y + q_{x'} \\ v_m + y &= q_{y'xy}xy + q_{y'x}x + (q_{y'y} + 1)y + q_{y'} \end{aligned} \tag{4.28}$$

Incorporating these into the flow criteria yields a simple set of eight linear equations in eight unknowns:

$$\left(\sum_{x,y}(\phi(x,y)\phi^T(x,y))\right)\mathbf{q} = -\sum_{x,y} E_t\phi(x,y) \tag{4.29}$$

where $\phi^T = [E_x(xy, x, y, 1), E_y(xy, x, y, 1)]$.

For the relative-projective model, ϕ is given by

$$\phi^T = [E_x(x, y, 1), E_y(x, y, 1), E_t(x, y)] \tag{4.30}$$

and for the pseudo-perspective model, ϕ is given by

$$\phi^T = [E_x(x, y, 1), E_y(x, y, 1), (x^2E_x + xyE_y, xyE_x + y^2E_y)] \tag{4.31}$$

In order to see how well the model describes the coordinate transformation between 2 images, say, g and h, one might *warp*[16] h to g, using the estimated motion model, and then compute some quantity that indicates how different the resampled version of h is from g. The MSE between the reference image and the warped image might serve as a good measure of similarity. However, since we are really interested in how the *exact model* describes the coordinate transformation, we assess the goodness of fit by first relating the parameters of the approximate model to the exact model, and then find the MSE between the reference image and the comparison image after applying the coordinate transformation of the exact model. A method of finding the parameters of the exact model, given the approximate model, is presented in Sec 4.4.1.

'Four point method' for relating approximate model to exact model

Any of the approximations above, after being related to the exact projective model, tend to behave well in the neighborhood of the identity, $\mathbf{A} = \mathbf{I}, \mathbf{b} = \mathbf{0}, \mathbf{c} = \mathbf{0}$. In 1-D, I explicitly expanded the model Taylor series about the identity; here, although I do not explicitly do this, I shall assume that the terms of the Taylor series of the model correspond to those taken about the identity. In the 1-D case we solve the 3 linear equations in 3 unknowns to estimate the parameters of the approximate motion model, and then relate the terms in this Taylor series to the exact parameters, a, b, and c (which involves solving another set of 3 equations in 3 unknowns, the second set being nonlinear, although very easy to solve).

In the extension to 2-D, the estimate step is straightforward, but the relate step is more difficult, because we now have eight nonlinear equations in eight unknowns, relating the terms in the Taylor series of the approximate model to the desired exact model parameters. Instead of solving these equations directly, I now propose a simple procedure for relating the parameters of the approximate model to those of the exact model, which I call the 'four point method':

1. Select four ordered pairs (e.g. the four corners of the bounding box containing the region under analysis, or the four corners of the image if the whole image is under analysis). Here suppose, for simplicity, that these points are the corners of the unit square: $\mathbf{s} = [s_1, s_2, s_3, s_4] = [(0,0)^T, (0,1)^T, (1,0)^T, (1,1)^T]$.

[15] Use of an approximate model that doesn't capture chirping or preserve straight lines can still lead to the true projective parameters as long as the model captures at least eight degrees of freedom.

[16] The term *warp* is appropriate here, since the approximate model does not preserve straight lines.

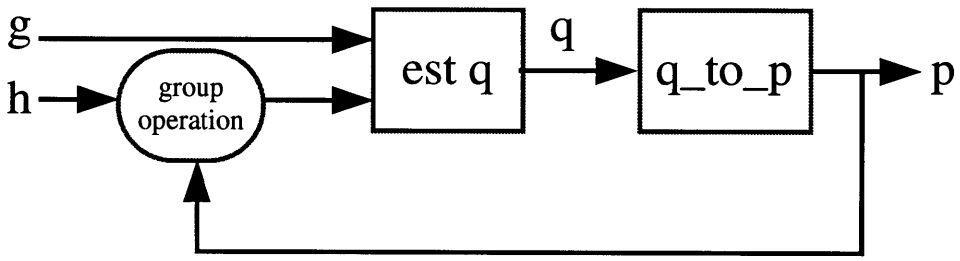


Figure 4-8: Method of computation of eight parameters $\mathbf{p}$ between two images from the same pyramid level, g and h. The approximate model parameters $\mathbf{q}$ are related to the exact model parameters $\mathbf{p}$ in a feedback system.

2. Apply the coordinate transformation using the Taylor series for the approximate model (e.g. (4.28)) to these points: $\mathbf{r} = \mathbf{u}_m(\mathbf{s})$.

3. Finally, the correspondences between $\mathbf{r}$ and $\mathbf{s}$ are treated just like features. This results in four easy to solve linear equations:

$$\begin{bmatrix} x'_k \\ y'_k \end{bmatrix} = \begin{bmatrix} x_k, y_k, 1, 0, 0, 0, -x_k x'_k, -y_k x'_k \\ 0, 0, 0, x_k, y_k, 1, -x_k y'_k, -y_k y'_k \end{bmatrix} \begin{bmatrix} a_{x'x}, a_{x'y}, b_{x'}, a_{y'x}, a_{y'y}, b_{y'}, c_x, c_y \end{bmatrix}^T \quad (4.32)$$

where $1 \leq k \leq 4$. This results in the exact eight parameters, $\mathbf{p}$.

We remind the reader that the four corners are **not** feature correspondences as used in the feature-based methods of Sec. 4.3.1, but, rather, are used so that the two featureless models (approximate and exact) can be related to one another.

It is important to realize the full benefit of finding the exact parameters. While the "approximate model" is sufficient for small deviations from the identity, it is not adequate to describe large changes in perspective. However, if we use it to track small changes incrementally, and each time relate these small changes to the exact model (4.24), then we can accumulate these small changes using the *law of composition* afforded by the group structure. This is an especially favorable contribution of the group framework. For example, with a video sequence, we can accommodate very large accumulated changes in perspective in this manner. The problems with cumulative error can be eliminated, for the most part, by constantly propagating forward the true values, computing the residual using the approximate model, and each time relating this to the exact model to obtain a goodness-of-fit estimate.

Algorithm for 'unweighted projective flow': overview

Below is an outline of the algorithm; details of each step are in subsequent sections.

Frames from an image sequence are compared pairwise to test whether or not they lie in the same orbit:

1. A Gaussian pyramid of three or four levels is constructed for each frame in the sequence.

2. The parameters $\mathbf{p}$ are estimated at the top of the pyramid, between the two lowest-resolution images of a frame pair, g and h, using the iterative method depicted in Fig. 4-8.

3. The estimated $\mathbf{p}$ is applied to the next higher-resolution (finer) image in the pyramid, $\mathbf{p} \circ g$, to make the two images at that level of the pyramid nearly congruent before estimating the $\mathbf{p}$ between them.

4. The process continues down the pyramid until the highest-resolution image in the pyramid is reached.

4.4.2 Multiscale iterative implementation

The Taylor-series formulations I have used implicitly assume smoothness; the performance is improved if the images are blurred before estimation. To accomplish this, I do not downsample critically after low-pass filtering in the pyramid. However, after estimation, I use the original (unblurred) images when applying the final coordinate transformation.

The strategy I present differs from the multiscale iterative (affine) strategy of Bergen *et al.* in one important respect beyond simply an increase from six to eight parameters. The difference is the fact that we have two motion models, the 'exact motion model' (4.24) and the 'approximate motion model', namely the Taylor series approximation to the motion model itself. The approximate motion model is used to iteratively converge to the exact motion model, using the algebraic *law of composition* afforded by the exact projective group model. In this strategy, the exact parameters are determined at each level of the pyramid, and passed to the next level. The steps involved are summarized schematically in Fig. 4-8, and described below:

1. Initialize: Set $h_0 = h$ and set $\mathbf{p}_{0,0}$ to the identity operator.
2. Iterate ($k = 1 \ldots K$):
   (a) **ESTIMATE:** Estimate the 8 or more terms of the approximate model between two image frames, g and h_{k-1}. This results in approximate model parameters $\mathbf{q}_k$.
   (b) **RELATE:** Relate the approximate parameters $\mathbf{q}_k$ to the exact parameters using the 'four point method'. The resulting exact parameters are $\mathbf{p}_k$.
   (c) **RESAMPLE:** Apply the *law of composition* to accumulate the effect of the $\mathbf{p}_k$'s. Denote these composite parameters by $\mathbf{p}_{0,k} = \mathbf{p}_k \circ \mathbf{p}_{0,k-1}$. Then set $h_k = \mathbf{p}_{0,k} \circ h$. (This should have nearly the same effect as applying $\mathbf{p}_k$ to h_{k-1}, except that it will avoid additional interpolation and anti-aliasing errors you would get by resampling an already resampled image [39]).

Repeat until either the error between h_k and g falls below a threshold, or until some maximum number of iterations is achieved. After the first iteration, the parameters $\mathbf{q}_2$ tend to be near the identity since they account for the residual between the "perspective-corrected" image h_1 and the "true" image g. We find that only two or three iterations are usually needed for frames from nearly the same orbit.

A rectangular image assumes the shape of an arbitrary quadrilateral when it undergoes a projective coordinate transformation. In coding the algorithm, I pad the undefined portions with the quantity NaN, a standard IEEE arithmetic value, so that any calculations involving these values automatically inherit NaN without slowing down the computations. The algorithm (in Matlab on an HP 735) takes about six seconds per iteration for a pair of 320x240 images.

4.4.3 Exploiting commutativity for parameter estimation

There is a fundamental uncertainty [66] involved in the simultaneous estimation of parameters of a noncommutative group, akin to the Heisenberg uncertainty relation of quantum mechanics. In contrast, for a commutative[17] group (in the absence of noise), we can obtain the exact coordinate transformation.

Segman [67] considered the problem of estimating the parameters of a commutative group of coordinate transformations, in particular, the parameters of the affine group [68]. His work also deals with noncommutative groups, in particular, in the incorporation of scale in the Heisenberg group[18] [69].

[17] A commutative (or *Abelian*) group is one in which elements of the group commute, for example, translation along the x-axis commutes with translation along the y-axis, so the 2-D translation group is commutative.

[18] While the Heisenberg group deals with translation and frequency-translation (modulation), some of the concepts could be carried over to other more relevant group structures.

Estimating the parameters of a commutative group is computationally efficient, e.g., through the use of Fourier cross-spectra [70]. I exploit this commutativity for estimating the parameters of the noncommutative 2-D projective group by first estimating the parameters that commute. For example, we improve performance if we first estimate the two parameters of translation, correct for the translation, and then proceed to estimate the eight projective parameters. We can also simultaneously estimate both the isotropic-zoom and the rotation about the optical axis by applying a log-polar coordinate transformation followed by a translation estimator. This process may also be achieved by a direct application of the Fourier-Mellin transform [71]. Similarly, if the only difference between g and h is a camera pan, then the pan may be estimated through a coordinate transformation to cylindrical coordinates, followed by a translation estimator.

In practice, I run through the following 'commutative initialization' before estimating the parameters of the projective group of coordinate transformations:

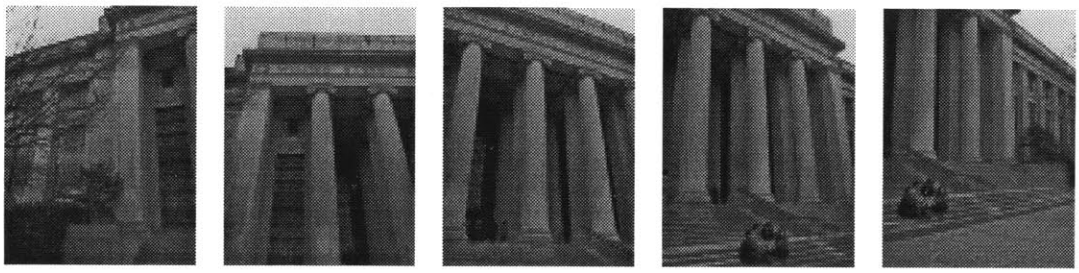
1. Assume that h is merely a translated version of g.
   (a) Estimate this translation using the method of Girod [70].
   (b) Shift h by the amount indicated by this estimate.
   (c) Compute the *MSE* between the shifted h and g, and compare to the original MSE before shifting.
   (d) If an improvement has resulted, use the shifted h from now on.
2. Assume that h is merely a rotated and isotropically zoomed version of g.
   (a) Estimate the two parameters of this coordinate transformation.
   (b) Apply these parameters to h.
   (c) If an improvement has resulted, use the coordinate-transformed (rotated and scaled) h from now on.
3. Assume that h is merely an "x-chirped" (panned) version of g, and, similarly, 'x-dechirp' h. If an improvement results, use the 'x-dechirped' h from now on. Repeat for y (tilt.)

Compensating for one step may cause a change in choice of an earlier step. Thus it might seem desirable to run through the commutative estimates iteratively. However, my experience on lots of real video indicates that a single pass usually suffices, and in particular, will catch frequent situations where there is a pure zoom, a pure pan, a pure tilt, etc, both saving the rest of the algorithm computational effort, as well as accounting for simple coordinate transformations such as when one image is an upside-down version of the other. (Any of these pure cases corresponds to a single parameter group, which is commutative.) Without the 'commutative initialization' step, these parameter estimation algorithms are prone to get caught in local optima, and thus never converge to the global optimum.

4.5 Performance and Applications

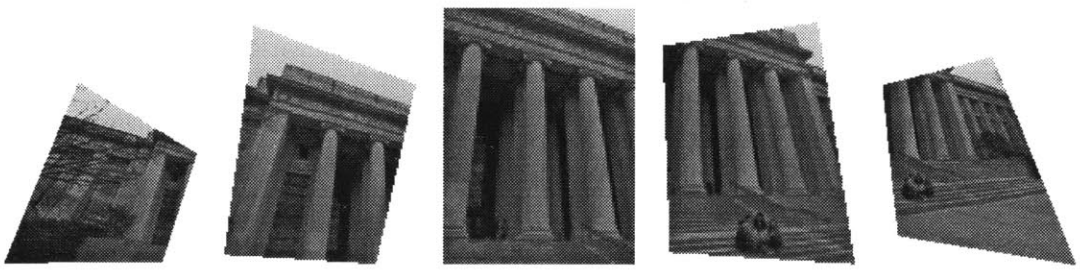
Figure 4-9 shows some frames from a typical image sequence. Figure 4-10 shows the same frames brought into the coordinate system of frame (c), that is, the middle frame was chosen as the *reference frame*.

Given that we have established a means of estimating the projective coordinate transformation between any pair of images, there are two basic methods we use for finding the coordinate transformations between all pairs of a longer image sequence. Because of the group structure of the projective coordinate transformations, it suffices to arbitrarily select one frame and find the coordinate transformation between every other frame and this frame. The two basic methods are:



(a) (b) (c) (d) (e)

Figure 4-9: Frames from original image orbit, sent from my personal imaging apparatus. Note camera is mounted sideways so that it can "paint" out the image canvas with a wider "brush", when sweeping across for a panorama. Thus the visual field of view that I experienced was rotated through 90 degrees. Much like George Stratton did with his upside-down glasses, I adapted, over an extended period of time, to experiencing the world rotated 90 degrees. (Adaptation experiments will be covered in Chapter 6.)



(a) (b) (c) (d) (e)

Figure 4-10: Frames from original image video orbit after a coordinate transformation to move them along the orbit to the reference frame (c). The coordinate-transformed images are alike except for the region over which they are defined. Note that the regions are not parallelograms; thus, methods based on the affine model fail.

1. **Differential parameter estimation**: The coordinate transformations between successive pairs of images, $\mathbf{p}_{0,1}$, $\mathbf{p}_{1,2}$, $\mathbf{p}_{2,3}$, ..., estimated.

2. **Cumulative parameter estimation**: The coordinate transformation between each image and the reference image is estimated directly. Without loss of generality, select frame zero (E_0) as the reference frame and denote these coordinate transformations as $\mathbf{p}_{0,1}$, $\mathbf{p}_{0,2}$, $\mathbf{p}_{0,3}$, ...

Theoretically, the two methods are equivalent:

$$\begin{aligned} E_0 &= p_{0,1} \circ p_{1,2} \circ \ldots \circ p_{n-1,n} E_n \quad \text{differential method} \\ E_0 &= p_{0,n} E_n \quad \text{cumulative method} \end{aligned} \tag{4.33}$$

However, in practice, the two methods differ for two reasons:

1. **cumulative error**: In practice, the estimated coordinate transformations between pairs of images register them only approximately, due to violations of the assumptions (e.g. objects moving in the scene, center of projection not fixed, camera swings around to bright window and automatic iris closes, etc.). When a large number of estimated parameters are composed, cumulative error sets in.

2. **finite spatial extent of image plane**:

   Theoretically, the images extend infinitely in all directions, but, in practice, images are cropped to a rectangular bounding box. Therefore, a given pair of images (especially if they are far from adjacent in the orbit) may not overlap at all; hence it is not possible to estimate the parameters of the coordinate transformation using those two frames.

The frames of Fig 4-9 were brought into register using the differential parameter estimation, and "cemented" together seamlessly on a common canvas. "Cementing" involves piecing the frames together, for example, by median, mean, or trimmed mean, or combining on a subpixel grid [32]. (Trimmed mean was used here, but the particular method made little visible difference.) Fig 4-11 shows this result ("projective/projective"), with a comparison to two non-projective cases. The first comparison is to "affine/affine" where affine parameters were estimated (also multiscale) and used for the coordinate transformation. The second comparison, "affine/projective," uses the six affine parameters found by estimating the eight projective parameters and ignoring the two "chirp" parameters $\mathbf{c}$ (which capture the essence of tilt and pan). These six parameters $\mathbf{A}$, $\mathbf{b}$ are more accurate than those obtained using the affine estimation, as the affine estimation tries to fit its shear parameters to the camera pan and tilt. In other words, the affine estimation does worse than the six affine parameters within the projective estimation. The affine coordinate transform is finally applied, giving the image shown. Note that the coordinate-transformed frames in the affine case are parallelograms.

4.5.1 Subcomposites and the support matrix

Two situations have so far been dealt with:

1. The camera movement is small, so that any pair of frames chosen from the video orbit have a substantial amount of overlap when expressed in a common coordinate system. (Use differential parameter estimation.)

2. The camera movement is monotonic, so that any errors that accumulate along the registered sequence are not particularly noticeable. (Use cumulative parameter estimation.)

In the example of Fig 4-11, any cumulative errors are not particularly noticeable because the camera motion is progressive, that is, it does not reverse direction, or loop around on itself. Now let us look

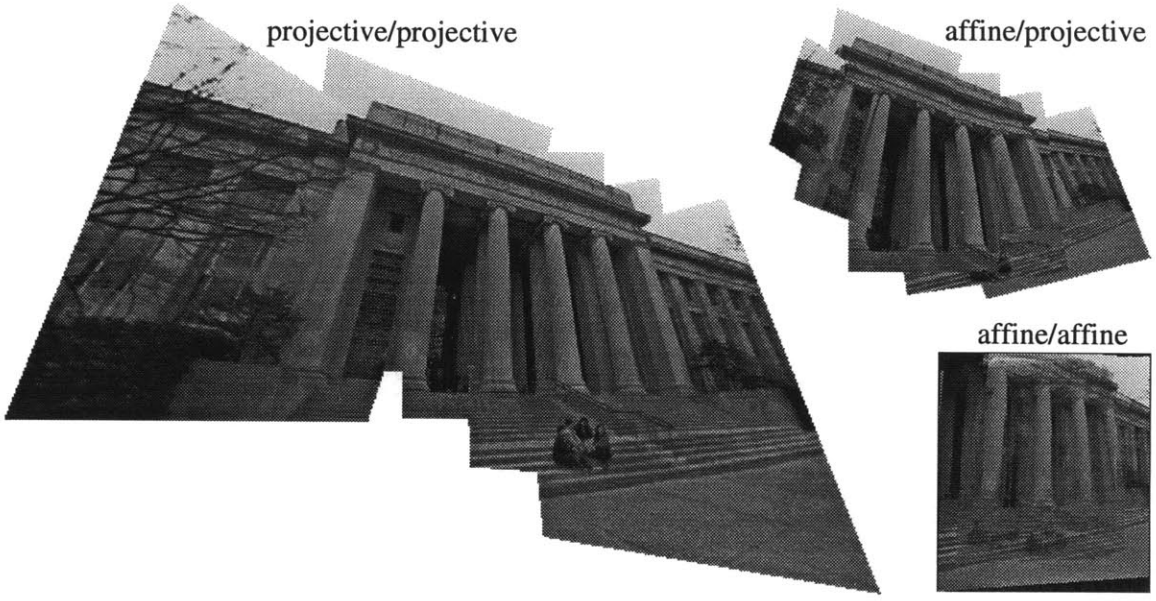


Figure 4-11: Frames of Fig 4-10 "cemented" together on single image "canvas", with comparison of affine and projective models. Note the good registration and nice appearance of the projective/projective image despite the noise in the amateur television receiver, wind-blown trees, and the fact that the rotation of the camera was not actually about its center of projection. Note also that the affine model fails to properly estimate the motion parameters (affine/affine), and even if the "exact" projective model is used to **estimate** the affine parameters, there is no affine coordinate transformation that will properly register all of the image frames.

Figure 4-12: The Hewlett Packard "Claire" image sequence, which violates the assumptions of the model (the camera location was not fixed, and the scene was not completely static). Images appear in TV raster-scan order.

at an example where the camera motion loops back on itself and small errors, due to violations of the assumptions (fixed camera location and static scene), accumulate.

Consider the image sequence shown in Fig 4-12. The composite arising from bringing these 16 image frames into the coordinates of the first frame exhibited somewhat poor registration due to cumulative error; I use this sequence to illustrate the importance of subcomposites.

The 'differential support matrix[19]', for which the entry $\mathbf{q}_{m,n}$ tells us how much frame n overlaps with frame m when expressed in the coordinates of frame m, for the sequence of Fig 4-12 appears in Fig 4-13.

Examining the support matrix, and the mean-squared error estimates, the local maxima of the support matrix correspond to the local minima of the mean-squared error estimates, suggesting the subcomposites[20]: $\{7,8,9,10,6,5\}$, $\{1,2,3,4\}$, and $\{15,14,13,12\}$. It is important to note that when

[19] The 'differential support matrix' is not necessarily symmetric, while the 'cumulative support matrix' for which the entry $bfq_{m,n}$ tells us how much frame n overlaps with frame m when expressed in the coordinates of frame 0 (reference frame) is symmetric.

[20] Researchers at Sarnoff also consider the use of sub composites, and refer to them as *tiles* [72][73]

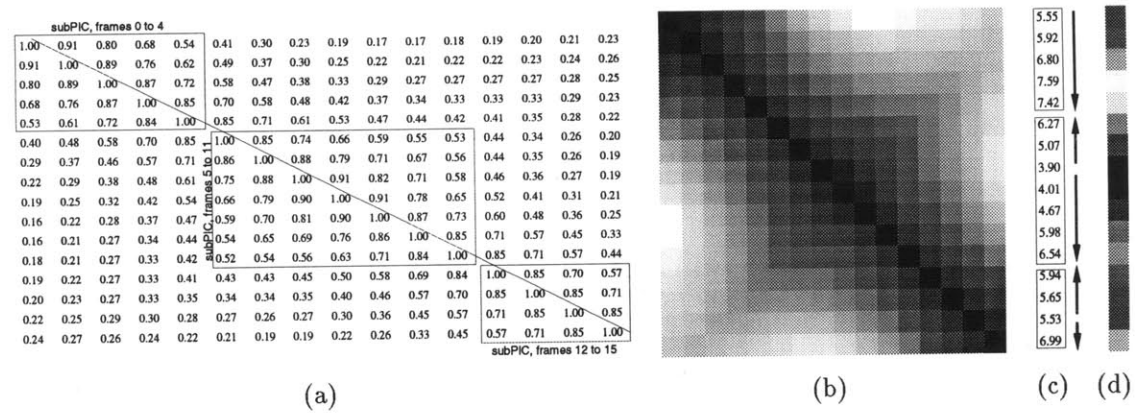


Figure 4-13: Support matrix and mean-squared registration error defined by image sequence in Fig 4-12 and the estimated coordinate transformations between images. (a) entries in table. The diagonals are one since every frame is fully supported in itself. The entries just above (or below) the diagonal give the amount of pairwise support. For example, frames 0 and 1 share high mutual support (.91). Frames 7, 8, and 9 also share high mutual support (again .91). (b) corresponding *density plot* (more dense ink indicates higer values). (c) mean-square registration error (d) corresponding density plot

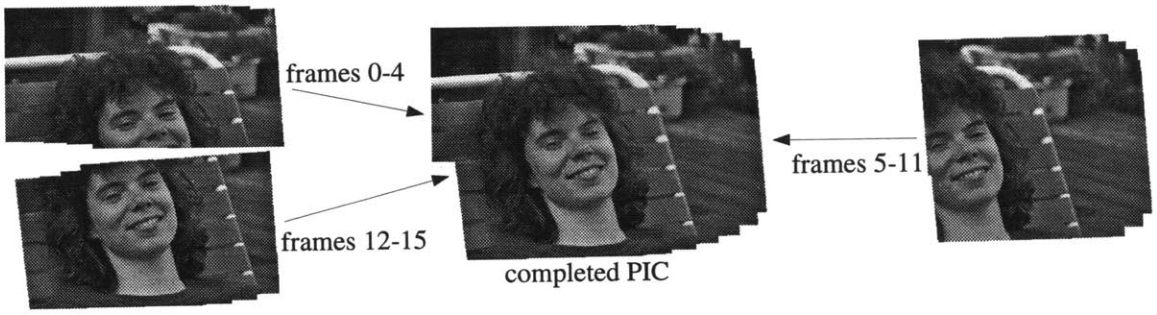


Figure 4-14: Subcomposites are each made from subsets of the images that share high quantities of mutual support and low estimates of mutual error, and then combined to form the final composite. (PIC stands for Pencigraphic Image Composite.)

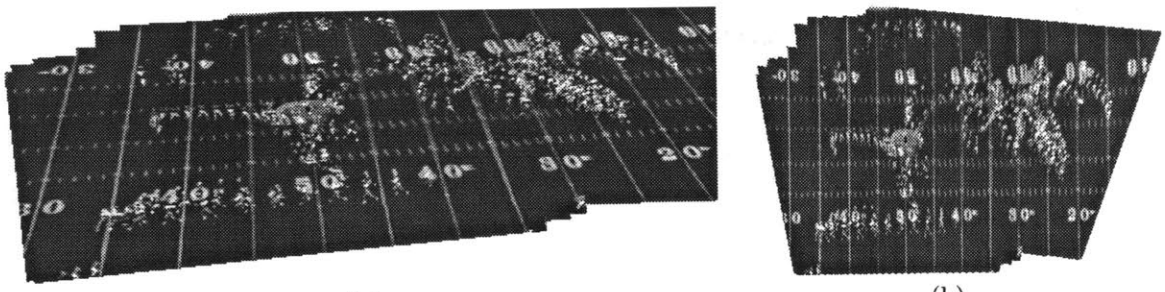
the error is low, if the support is also low, the error estimate might not be valid. For example if the two images overlap in only one pixel, then even if the error estimate is zero (e.g. perhaps that pixel has a value of 255 in both images) the alignment is not likely good.

The selected subcomposites appear in Fig 4-14. Estimating the coordinate transformation between these subcomposites, and putting them together into a common frame of reference results in a composite (Fig 4-14) about 1200 pixels across, where the image is sharp despite the fact that the person in the picture was moving slightly and the camera operator was also moving (violating the assumptions of both static scene and fixed center of projection).

4.5.2 Flat subject matter and alternate coordinates

Many sports such as football or soccer are played on a nearly flat field that forms a rigid planar patch over which the analysis may be conducted. After each of the frames undergoes the appropriate coordinate transformation to bring it into the same coordinate system as the reference frame, the sequence can be played back showing only the players (and the image boundaries) moving. Markings on the field (such as numbers and lines) remain at a fixed location, which makes subsequent analysis and summary of the video content easier. This data makes a good test case for the algorithms because the video was noisy and the players caused the assumption of static scene to be violated.

Despite the players moving in the video, the proposed method successfully registers all of the images in the orbit, mapping them into a single high-resolution image composite of the entire playing field. Figure 4-15(a) shows 16 frames of video from a football game combined into a single image



(a) (b)

Figure 4-15: Image composite made from 16 video frames taken from a television broadcast sporting event. Note the "Edgertonian" appearance, as each player traces out a stroboscopic-like path. The proposed method works robustly, despite the movement of players on the field. (a) Images are expressed in the coordinates of the first frame. (b) Images are expressed in a new useful coordinate system corresponding to none of the original frames. Note the slight distortion, due to the fact that football fields are not perfectly flat, but, rather, are raised slightly in the center.

composite, expressed in the coordinates of the first image in the sequence. The choice of coordinate system was arbitrary, and any of the images could have been chosen as the reference frame. In fact, a coordinate system other than one chosen from the input images could also be used. In particular, a coordinate system where *parallel lines never meet*, and periodic structures are "dechirped" (Fig 4-15(b)) lends itself well to machine vision and player-tracking algorithms [74]. Even if the entire playing field was never visible in any one image, collectively, the video from an entire game will likely reveal every square yard of playing surface at one time or another, hence enabling us to make a composite of the entire playing surface.

4.6 Chapter summary

We proposed and demonstrated featureless estimation of the projective coordinate transformation between two images. Not just one method, but various methods were proposed, among these, "projective fit" and "projective flow" which estimate the projective (homographic) coordinate transformation between pairs of images, taken with a camera that is free to pan, tilt, rotate about its optical axis, and zoom. The new approach was also formulated and demonstrated within a multiscale iterative framework. Applications to seamlessly combining images in or near the same orbit of the projective group of coordinate transformations were also presented. The proposed approach solves for the 8 parameters of the "exact" model (the projective group of coordinate transformations), is fully automatic, and converges quickly. The approach was also explored together with the use of sub composites, useful when the camera motion loops back on itself.

The proposed method was found to work well on image data collected from both good-quality and poor-quality video under a wide variety of conditions (sunny, cloudy, day, night). It has been tested with a head-mounted wireless video camera, and performs successfully even in the presence of noise, interference, scene motion (such as people walking through the scene), lighting fluctuations, and parallax (due to movements of the wearer's head). It remains to be shown which variant of the proposed approach is optimal, and under what conditions.

Chapter 5

The domain and range of light: Estimating parameters of the homomorphic projectivity group of transformations

5.1 Overview

In this chapter, I present a framework I completed in 1993 [28], in which I have combined my 'homomorphic imaging' work (which we saw in Chapter 3) with 'video orbits' (presented in Chapter 4). The result, which I refer to as 'homographic projectivity', suggests that a camera, rotated about its center of projection, may be used as a measuring instrument. Thus the 'painting with video' metaphor of Chapter 4 will now provide not just a picture of increased spatial extent, but will provide a means to 'paint' a set of *photometric measurements*, onto an empty 'canvas', as we sweep the camera around. These measurements describe, up to a single unknown scalar constant (for the entire canvas), the quantity of light arriving from each corresponding direction in space.

5.1.1 Turning AGC from a bug into a feature

In Chapter 3, much was said about and done with differently exposed images. In this chapter, the images will differ not only in exposure, but also in projection, that is, they will be related by a projective coordinate transformation (as described in Chapter 4) as well as a homographic gain adjustment (as described in Chapter 3).

Consider a static scene and fixed center of projection, about which a camera is free to zoom, pan, tilt, and rotate about its optical axis. With an ideal camera, the resulting images are in the same orbit of the projective group-action, and each pixel of each image provides a measurement of a ray of light passing through a common point in space. Unfortunately (from the standpoint of Chapter 4) most modern cameras have a built in automatic gain control (AGC), automatic shutter, or auto-iris, which, in many cases cannot be turned off. Many modern digitizers to which cameras are connected have their own AGC which also cannot be disabled. With AGC, the characteristic response function of the camera varies, making it impossible to accurately describe one image as a projective coordinate transformed version of another. This chapter proposes not only a solution to this problem, but a means of turning AGC into an asset. It turns out that AGC is the very entity that produces a variety of differently exposed images, for as we point the camera at bright objects, the gain decreases so that darker objects in the periphery may be grossly underexposed, while when the camera is centered on dark objects, the periphery is grossly overexposed. Since the same objects often appear in the periphery of both overexposed and underexposed images, we obtain, without

expending any conscious thought or effort, pictures of overlapping subject matter in which the same subject matter is available at a wide variety of different exposures. Therefore, even in cases where AGC could be disabled, we will most likely chose not to turn it off.

5.2 Introduction

Suppose we take two pictures, using the same settings (in manual exposure mode), of the same scene, from a fixed common location (e.g. where the camera is free to zoom, pan, tilt, and rotate about its optical axis between taking the two pictures). Both of the pictures capture the same pencil of light (I neglect the boundaries of the sensor array and assume that both pictures have sufficient field of view to capture all of the objects of interest), but each one projects this information differently onto the film or image sensor. Neglecting that which falls beyond the borders of the pictures, the images are in the same orbit of the projective group of coordinate transformations. The use of projective (homographic) coordinate transformations to automatically (without use of explicit features) combine multiple pictures of the same scene into a single picture of greater resolution or spatial extent, was first described in 1993 [28]. These coordinate transformations were shown to capture the essence of a camera at a fixed center of projection (COP) in a static scene.

Note that the projective group of coordinate transformations is not Abelian and there is thus some uncertainty in the estimation of the parameters associated with this group of coordinate transformations [16]. However, we may first estimate parameters of Abelian subgroups (for example, the pan/tilt parameters, perhaps approximating them as a 2-D translation so that Fourier methods [70] may be used). Estimation of zoom (scale) together with pan and tilt, would incorporate non-commutative parameters (zoom and translation don't commute), but could still be done using the *multiresolution Fourier transform* [75][76], at least as a first step, followed by an iterative parameter estimation procedure over all parameters. An iterative approach to the estimation of the parameters of a projective (homographic) coordinate transformation between images was suggested in [28], and later in [32] and [38].

Lie algebra is the algebra of symmetry, and pertains to the behaviour of a group in the neighbourhood of its identity. With typical video sequences, coordinate transformations relating adjacent frames of the sequence are very close to the identity. Thus we may use the Lie algebra of the group when considering adjacent frames of the sequence, and then use the group itself when combining these frames together. Thus, for example, to find the coordinate transformation, p_{09}, between $F_0(x, y)$, Frame 0 and $F_9(x, y)$, Frame 9, we might use the Lie algebra to estimate p_{01} (the coordinate transformation between Frame 10 and 11) and then estimate p_{12} between frames F_1 and F_2, and so on, each one being found in the neighbourhood of the identity. Then to obtain p_{09}, we use the **true law of composition** of the group: $p_{09} = p_{01} \circ p_{12} \circ \ldots \circ p_{89}$.

5.2.1 Ideal spotmeter

Recall the ideal spotmeter, presented in Chapter 2. It is a perfectly directional lightmeter which measures the quantity of light, q, arriving from the direction in which it is pointed (parameterized by its azimuth, θ, and its elevation, ϕ). In Chapter 3, we saw how an ordinary camera can be linearized by simply analyzing a collection of differently exposed images. Here we will therefore see how a camera with time-varying gain (e.g. one with AGC) can be made to function as closely as possible, to a collection of the idealized spotmeters of Chapter 2, similar to that depicted in Fig 2-2, although typically with only three color channels (red, green, and blue). I refer to the measurement of the quantity of light in each ray of light, passing through a given point in space, as the 'pencigraph'[1].

Panoramic photography attempts to record a large portion of the 'nonmetric pencigraph' onto a single piece of film (often by rotating the camera while sliding the film through a slit). By *nonmetric*, I mean that even if we know the exact direction of arrival corresponding to each pixel in

[1] This terminology originated from when I derived the theory for 1-D images in a 2-D space, where we have a *pencil* of rays passing through the center of projection of the camera.

the panoramic picture, it does not tell us the actual quantity of light arriving from that direction (e.g. it does not give us a *linearized* measurement).

In Chapter 4 we saw how the nonmetric pencigraph could also be estimated from a collection of pictures all taken from the same point in space (but with with differing camera orientations and lens focal lengths).

The basic philosophy of Chapter 4 is that the camera may be regarded as an array of (nonmetric) spotmeters, measuring rays of light passing through the center of projection (COP). To each pair of pixel indices of the sensor array in a camera, we may associate an azimuth and elevation. Eliminating lens distortion [47] makes the images obey the laws of projective geometry, so that they are (to within image noise and cropping) in the same orbit of the projective group action. (Lens distortion may also be simply absorbed into the mapping between pixel locations and directions of arrival.)

In Chapter 3 we found that trying to use a pixel from a camera as a lightmeter raised many interesting and important problems. The output of a typical camera, f, is not linear with respect to the incoming quantity of light, q. For a digital camera, the output, f, is the pixel value, while for film, the output might be the density of the film at the particular location under consideration. I assumed that the output was some unknown but monotonic function of the input. (Monotonicity, a weaker constraint than linearity, is what I mean by 'nonmetric pencigraph' — our knowledge of the quantity of light received is in terms of the nonmetric quantity $f(q)$, not q itself.)

Methods (both parametric and non parametric) of estimating this nonlinearity, from pictures that differ only in exposure, were presented in Chapter 3, and these methods facilitated the use of an ordinary camera as an array of metric spotmeters, measuring, to within a single unknown constant, the quantity of light arriving from each direction in space.

In this chapter, I will select one of these methods developed in Chapter 3, one of the parametric method for estimating f, that was based on automatically determining the parameters of the classic response curve [27] given in Equation 3.2 from pictures that differ only in exposure. This method will be the one that I emphasize here in Chapter 5, where it is combined with the Video Orbits methodology of Chapter 4. However, any of the methods of Chapter 3 may be similarly combined with Video Orbits.

5.2.2 AGC

If what is desired is a picture of increased spatial extent or spatial resolution, the nonlinearity is not a problem, so long as it is not image dependent. However, most low-cost cameras have a built in automatic gain control (AGC), electronic level control, auto iris, or some other form of automatic exposure which cannot be turned off or disabled. (For simplicity I refer to all of these methods of automatic exposure control as AGC, whether or not they are actually implemented using gain adjustment.) This means that the unknown response function, $f(q)$, is image dependent, and will therefore change over time, as the camera framing changes to include brighter or darker objects.

Although AGC was a good invention for its intended application, serving the interests of most camera users who merely wish to have a properly exposed picture without having to make adjustments to the camera, it has previously thwarted attempts to estimate the projective coordinate transformation between frame pairs. Examples of an image sequence, acquired using a camera with AGC, appear in Fig 5-1.

The purpose of this chapter is to propose a joint estimation of the projective coordinate transformation and the tone-scale change. Each of these two may be regarded as a "motion estimation" problem if we extend the concept of "motion estimation" to include both 'domain motion' (motion in the traditional sense) as well as 'range motion' (Fig 5-2).

5.3 Joint estimation of both domain and range coordinate transformations

As in [32], we consider one dimensional "images" for purposes of illustration, with the understanding that the actual operations are performed on 2-D images. The 1-D projective-Wyckoff group is

Figure 5-1: The 'fire-exit' sequence, taken using a camera with AGC. (a)-(j) frames 10-19: as the camera pans across to take in more of the open doorway, the image brightens up showing more of the interior, while, at the same time, clipping highlight detail. Frame 10 (a) shows the writing on the white paper taped to the door very clearly, but the interior is completely black. In frame 15 (f) the paper is completely obliterated — it is so "washed out" that we cannot even discern that there is a paper present. Although the interior is getting brighter, it is still not discernible in frame 15 (f), but more and more detail of the interior becomes visible as we proceed through the sequence, showing that the fire exit is blocked by the clutter inside. (A)-(J) 'certainty' images (as described in Chapter 3) corresponding to (a)-(j).

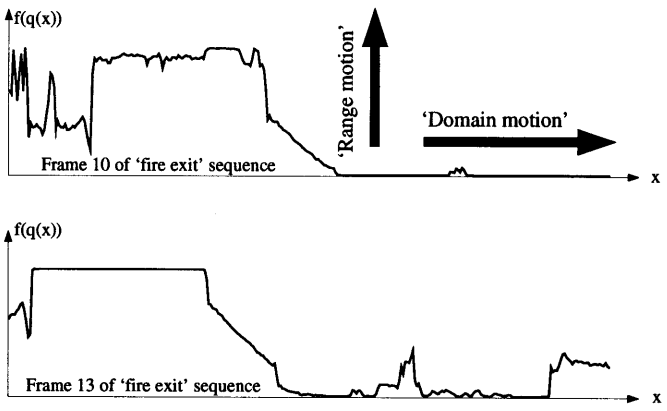


Figure 5-2: One row across each of two images from the 'fire exit' sequence. 'Domain motion' is motion in the traditional sense (e.g. motion from left to right, zoom, etc.), while 'Range motion' refers to a tone-scale adjustment (e.g. lightening or darkening of the image). In this case, the camera is panning to the right, so domain motion is to the left. However, when panning to the right, the camera points more and more into the darkness of an open doorway, causing the AGC to adjust the exposure. Thus there is some "upwards" motion of the curve as well as "leftwards" motion. Just as panning the camera across causes information to leave the frame at the left, and new information to enter at the right, the AGC causes information to leave from the top (highlights get clipped) and new information to enter from the bottom (increased shadow detail).

defined in terms of the group of projective coordinate transformations, taken together with the one-parameter group of image darkening/lightening operations: $p_{a,b,c,k} \circ f(q(x)) = g(f(q(\frac{ax+b}{cx+1}))) = f(kq(\frac{ax+b}{cx+1}))$ where g characterizes the lightening/darkening operation.

The law of composition is defined as: $(p_{abc}, p_k) \circ (p_{def}, p_l) = (p_{abc} \circ p_{def}, p_k \circ p_l)$ where the first law of composition on the right hand side is the usual one for the projective group, and the second one is that of the one-parameter lightening/darkening subgroup.

Two successive frames of a video sequence are related through a group-action that is near the identity of the group, thus one may think of the Lie algebra of the group as providing the structure locally. As in previous work [32] an approximate model which matches the 'exact' model in the neighbourhood of the identity is used. For the projective group, the approximate model has the form $q_2(x) = q_1((ax+b)/(cx+1))$.

For the 'Wyckoff group' (which is a one parameter group isomorphic to addition over the reals, or multiplication over the positive reals), the approximate model may be taken from Eq 3.2, by noting that

$$g(f(q)) = f(kq) = \alpha + \beta(kq)^\gamma = \alpha - k^\gamma\alpha + k^\gamma\alpha + k^\gamma\beta q^\gamma = k^\gamma f(q) + \alpha - \alpha k^\gamma \tag{5.1}$$

Thus we see that $g(f)$ is a "linear equation" (is affine) in f. This affine relationship suggests that linear regression on the cross histogram between two images would provide an estimate of α and γ, while leaving β unknown, which is consistent with the fact that the response curve may only be determined up to a constant scale factor.

From (5.1) we have that the (generalized) brightness change constraint equation is:

$$g(f(q(x,t))) = f(kq(x,t)) = f(q(x+\Delta x, t+\Delta t)) = k^\gamma f(q(x,t)) + \alpha - \alpha k^\gamma = k^\gamma F(x,t) + \alpha(1-k^\gamma) \tag{5.2}$$

where $F(x,t) = f(q(x,t))$. Combining this equation with the Taylor series representation:

$$F(x+\Delta x, t+\Delta t) = F(x,t) + \Delta x F_x(x,t) + \Delta t F_t(x,t) \tag{5.3}$$

where $F_x(x,t) = (df/dq)(dq(x)/dx)$, at time t, and $F_t(x,t)$ is the frame difference of adjacent frames, we have:

$$k^\gamma F + \alpha(1-k^\gamma) = F + \Delta x F_x + \Delta t F_t \tag{5.4}$$

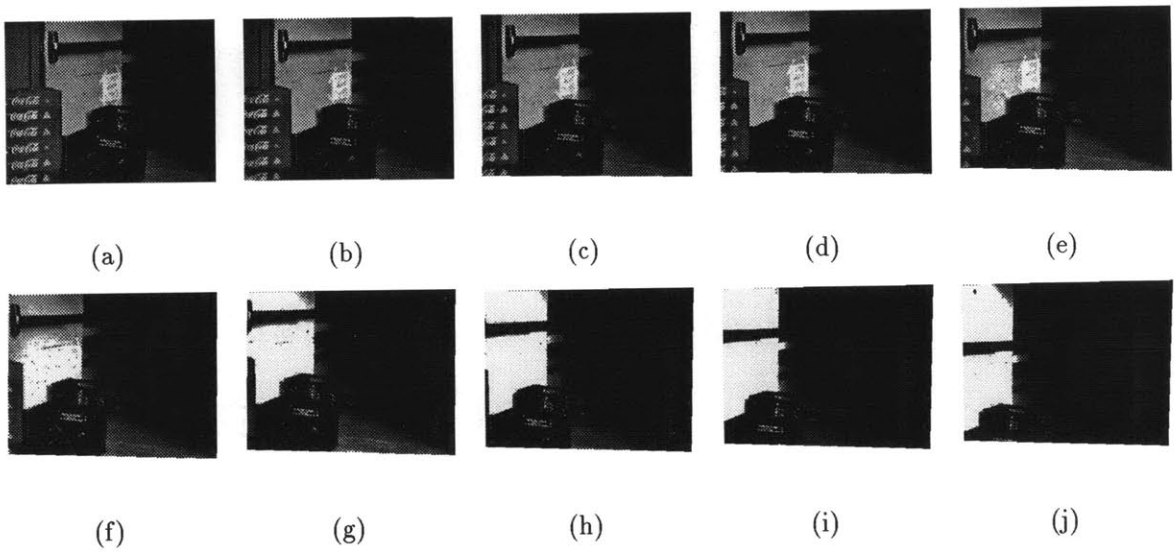


Figure 5-3: 'Homomorphic homographies': All images are expressed in spatiotonal coordinates of the first image in the sequence.

Thus, the brightness change constraint equation becomes:

$$F + uF_x + F_t - k^\gamma F - \alpha(1 - k^\gamma) = \epsilon \tag{5.5}$$

where I have normalized $\Delta t = 1$.

Substitution of the approximate model, as was used in Equation 4.22 (e.g. that of Equation 4.21) into (5.5) gives:

$$F + (q_2x^2 + q_1x + q_0)F_x + F_t - k^\gamma F + \alpha - \alpha k^\gamma = \epsilon \tag{5.6}$$

Minimizing $\sum \epsilon^2$ yields a linear solution in parameters of the approximate model:

$$\begin{bmatrix} \sum x^4F_x^2 & \sum x^3F_x^2 & \sum x^2F_x^2 & \sum x^2FF_x & -\sum x^2F_x \\ \sum x^3F_x^2 & \sum x^2F_x^2 & \sum xF_x^2 & \sum xFF_x & -\sum xF_x \\ \sum x^2F_x^2 & \sum xF_x^2 & \sum F_x^2 & \sum FF_x & -\sum F_x \\ \sum x^2FF_x & \sum xFF_x & \sum FF_x & \sum F^2 & -\sum F \\ \sum x^2F_x & \sum xF_x & \sum F_x & \sum F & -\sum 1 \end{bmatrix} \begin{bmatrix} q_2 \\ q_1 \\ q_0 \\ 1-k^\gamma \\ \alpha - \alpha k^\gamma \end{bmatrix} = - \begin{bmatrix} \sum x^2F_xF_t \\ \sum xF_xF_t \\ \sum F_xF_t \\ \sum FF_t \\ \sum F_t \end{bmatrix}$$

The parameters of the approximate model are related to those of the exact model, in the same way as in Chapter 3 (e.g. using the same kind of feedback process illustrated in Fig 4-8).

Using this new mathematical result enables images to be brought not just into register in the traditional 'domain motion' sense, but also brought into the same tonal scale through homomorphic gain adjustment. I refer to the combination of a spatial coordinate transformation combined with a tone scale adjustment as a 'spatiotonal transformation'. In particular, it is the spatiotonal transformation of 'homomorphic homography' that is of interest (e.g. homographic coordinate transformation combined with homomorphic gain adjustment). This form of spatiotonal transformation is illustrated in Fig 5-3 where all the images are transformed into the coordinates of the first image of the sequence, and in Fig 5-4 where all the images are transformed into the coordinates of the last frame in the image sequence. Here, because a tele lens was used, the perspective is not as dramatic as it was in some of the image sequences of Chapter 3, but it is still quite pronounced. Thus I have succeeded in simultaneously estimating the underlying gain change as well as the projective coordinate transformation relating successive pairs of images in the image sequence.

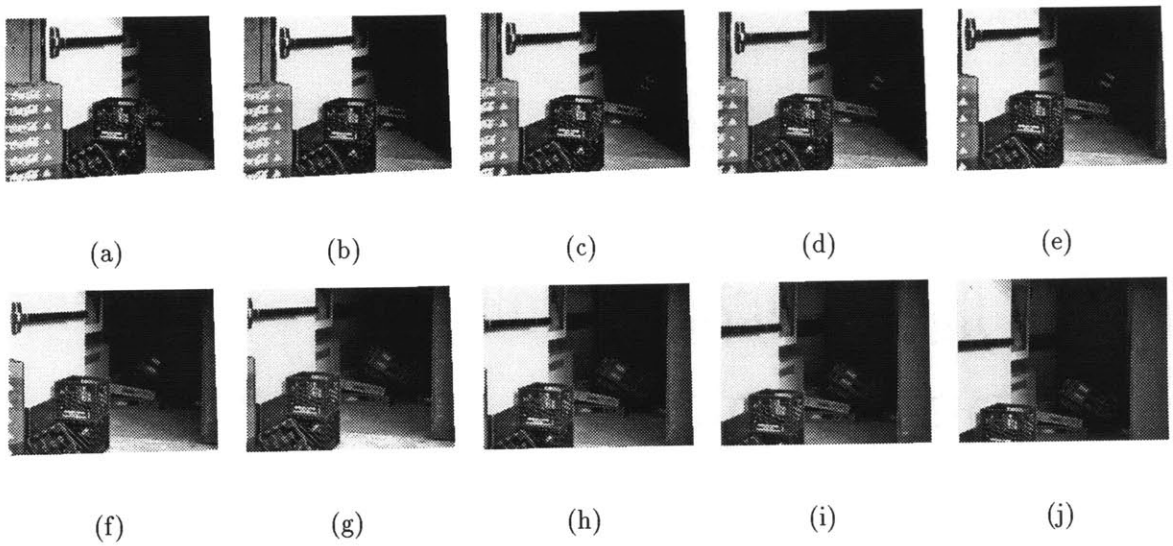


Figure 5-4: 'Homomorphic homographies': All images are expressed in spatiotonal coordinates of the last image in the sequence.

5.4 The big picture

To construct a single floating-point image of increased spatial extent and increased dynamic range, each pixel of the output image is constructed from a weighted sum of the images whose coordinate-transformed bounding boxes fall within that pixel. The weights in the weighted sum are the so-called 'certainty functions' [48], which are found by evaluating the derivative of the corresponding 'effective response function' at the pixel value in question. While the response function, $f(q)$, is fixed for a given camera, the 'effective response function', $f(k_i(q))$ depends on the exposure, k_i, associated with frame, i, in the image sequence. By evaluating $f_q(k_i(q_i(x,y)))$, we arrive at the so-called 'certainty images' (Fig 5-1). Lighter areas of the 'certainty images' indicate moderate values of exposure (mid-tones in the corresponding images), while darker values of the certainty images designate exposure extrema — exposure in the *toe* or *shoulder* regions of the response curve where it is difficult to discern subtle differences in exposure.

The pencigraphic estimate may be explored interactively on a computer system (Fig 5-5), but the simultaneous wide dynamic range and ability to discern subtle differences in greyscale are lost once the image is reduced to a tangible form (e.g. a hardcopy printout).

Paper and the range of light

Print paper typically has a dynamic range which is much less than photographic emulsion (film)[2]. Standard photographic processes may be used to mitigate this effect to a limited extent. For example, through careful choice of chemical developers used in processing the film, a lateral inhibition effect can be produced (very similar to the lateral inhibition that happens in the human visual system) that violates the monotinicity constraint we have emphasized so strongly, and instead returns an image where the greyscale value at a given point depends not only on the quantity of light arriving at that point, but also on the greyscale values of neighbouring pixels. Alternatively, for negative film, contrast masks may be made which help in the photographic printing process.

More recently, computational methods of reducing the dynamic range of images have been explored [24]. Just as these methods were applied to the images of Chapter 3, they may be applied to the metric pencigraphs of this chapter. As with the photographic lateral inhibition, these methods also relax the monotinicity constraint.

[2] The dynamic range of some papers is around 100:1, while that of many films is around 500:1.

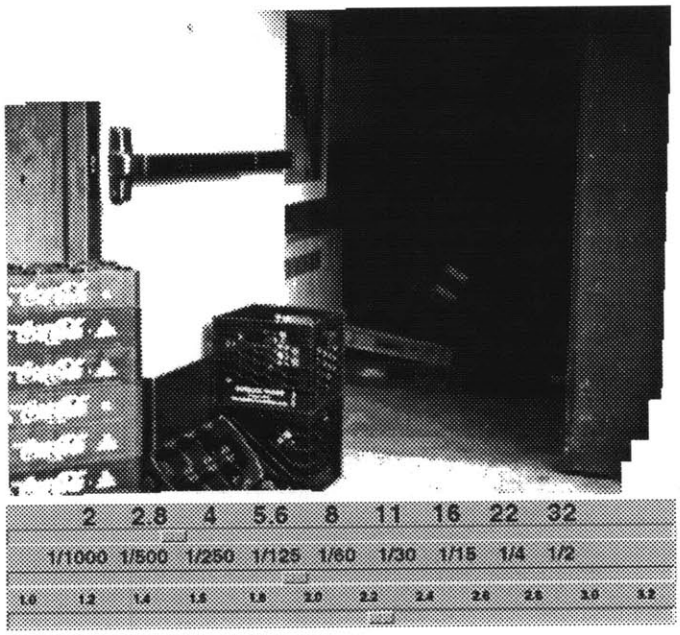


Figure 5-5: Floating point pencigraphic image constructed from the fire-exit sequence. The dynamic range of the image is far greater than that of a computer screen or printed page. The pencigraphic information may be interactively viewed on the computer screen, however, not only as an environment map (with pan, tilt, and zoom), but also with control of 'exposure' and contrast. With a 'virtual camera' we may move around in the pencigraph, both spatially and tonally.

Thus, in order to print a pencigraph, it may be preferable to relax the monotonicity constraint, and perform some local tone-scale adjustments (Fig 5-6). This relaxation of photometric constraints is even more apparent as the dynamic range of the scene is increased somewhat, as we see in Fig 5-7. In this figure, we also see the effect of perspective, which is more visible here than it was in the image composite of Fig 5-6.

Even if the end goal is a picture of limited dynamic range (as in Fig 5-5), perhaps where the artist wishes to deliberately wash out highlights and mute down shadows for expressive purposes, the proposed philosophy is still quite valid. It is preferable to capture a measurement space, recording the quantity of light arriving at each angle in space, and then from that measurement space, synthesize the tonally degraded image. Within the spirit of this approach, one attempts to capture as much information about the scene as possible, produce a pencigraphic estimate, and then "put expression" into that estimate (by throwing away information in a controlled fashion) to produce a final picture.

5.5 Chapter summary

The procedure for self-calibrating a camera (to within a constant scale factor) has been exploited for capturing pencigraphic measurements, in particular, treating the camera as an array of photometric measuring instruments. This has been accomplished by proposing and implementing a global motion estimation algorithm which considers jointly global "motion" in the domain and range of the functions undergoing "motion". Dynamic range has been extended by combining differently exposed images where the AGC, rather than thwarting motion estimation algorithms as is generally otherwise the case, actually provides both more information from the scene and information about the camera's unknown response function.

Figure 5-6: Fixed-point image made by tone-scale adjustments that are only locally monotonic, followed by quantization to 256 greylevels. Note that we can see clearly both the small piece of white paper on the door (and even read what it says — "COFFEE HOUSE CLOSED"), as well as the details of the dark interior. Note that we could not have captured such a nicely exposed image using an on-camera "fill-flash" to reduce scene contrast, because the fill-flash would mostly light up the areas near the camera (which happen to be the areas that are already too bright), while hardly affecting objects at the end of the dark corridor which are already too dark. Thus, one would need to set up additional photographic lighting equipment to obtain a picture of this quality. This image demonstrates the advantage of a small lightweight personal imaging system, built unobtrusively into a pair of eyeglasses, in that an image of very high quality was captured by simply looking around, without entering the corridor. This might be particularly useful if trying to report a violation of fire-safety laws, while at the same time, not appearing to be trying to capture an image. Note that this image was shot from some distance away from the premises (using a miniaturized tele lens I built into my eyeglass-based system) so that the effects of perspective, although still present, are not as immediately obvious as with some of the other extreme wide-angle image composites presented in this thesis.

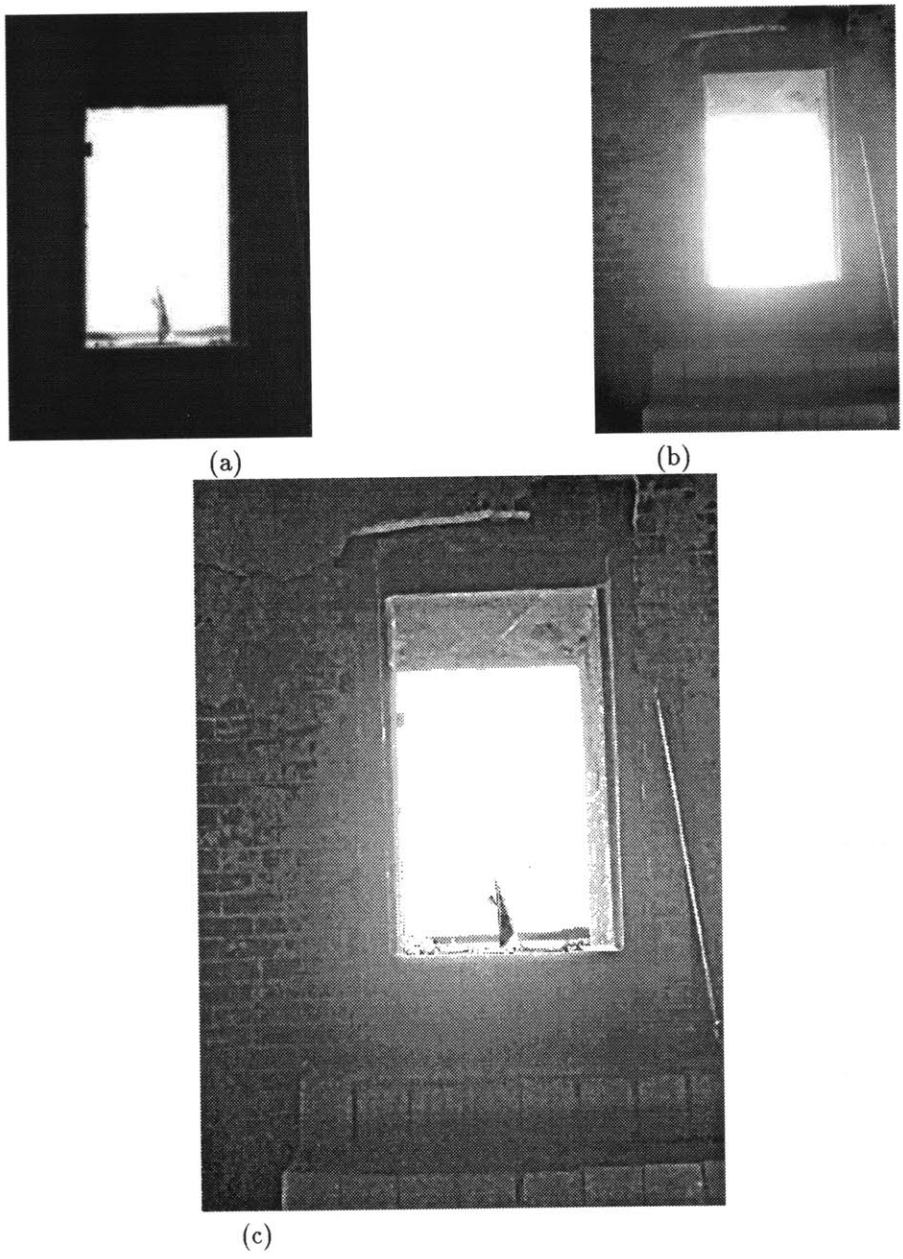


Figure 5-7: (a) Looking out onto the ocean from the inside of a dark abandoned fortress, but using an exposure appropriate for outdoors (note sailboat on the ocean but only the silhouette of the small access hatch is visible). (b) Exposure for the inside of the fortress leaves the hatch completely washed out, so that even its shape is not visible, let-alone anything beyond. (c) Combined image reduced to printed page with nonmonotic processing: here I have captured a dynamic range of approximately a hundred million to one, and reduced it to the printed page. Intricate detail is visible in the brightly backlit sail of the sailboat on the ocean outside the window to a dark fortress. Details of the inside of the dark fortress are also visible. Note the non-monotonic tone-scale due to spatial filtering; note, in particular, that parts of the sail are black, while parts of the inside of the fortress are white, even though the darkest part of the sail is thousands of times brighter than the brightest part of the inside of the fortress. Although I could have used a fill-flash here, I decided not to, in order to illustrate the range of light. Figure courtesy of the Society of Imaging Science and Technology, 1993; (C) Steve Mann, 1992.

Chapter 6

Life through the screen: Reconfigured Eyes in the age of wearable, tetherless computer-mediated reality

Virtual reality allows us to experience a new visual world, but deprives us of the ability to see the actual world in which we live. Indeed, many VR game spaces are enclosed with railings or the like so that players will not fall down, since they have replaced their reality with a new space, and are therefore blind to the real world.

Augmented reality attempts to bring together the real and virtual. The general spirit and intent of Augmented Reality (AR) is to *add* virtual objects to the real world. A typical AR apparatus consists of a video display with partially transparent visor, upon which computer-generated information is *overlaid*.

In this chapter, Mediated Reality (MR) is proposed. Mediated reality forms the basis for modern implementations of WearComp. We will also see how it forms a basis for humanistic intelligence, and in the following chapters, how MR forms a basis for a possible new genre of documentary video, and various other contributions based on MR.

MR differs from typical AR in two respects:

1. The general spirit of MR, like typical AR, includes *adding* virtual objects, but also includes the desire to *take away*, *alter*, or more generally to visually 'mediate' real objects. Thus MR affords the apparatus the ability to augment, diminish, or otherwise alter our perception of reality.

2. Typically, an AR apparatus is tethered to a computer workstation which is connected to an AC outlet, or constrains the user to some other specific site (such as a workcell, helicopter cockpit, or the like). What is proposed (and reduced to practice) is a collection of various means and apparatus which facilitate the augmenting, diminishing, or altering of the visual perception of reality in the context of ordinary day-to-day living.

MR uses a body-worn apparatus where both the *real* and *virtual* objects are placed on an equal footing, in the sense that both are presented together via a synthetic medium (video display).

Successful implementations have been realized by *viewing* the real world using a head-mounted display (HMD) fitted with video camera(s), body-worn processing, and/or bidirectional wireless communications to one or more remote computers. This portability enabled various forms of the apparatus to be tested extensively in everyday circumstances, such as while riding the bus, shopping, banking, and various other day-to-day interactions.

The proposed approach shows promise in applications where it is desired to have the ability to

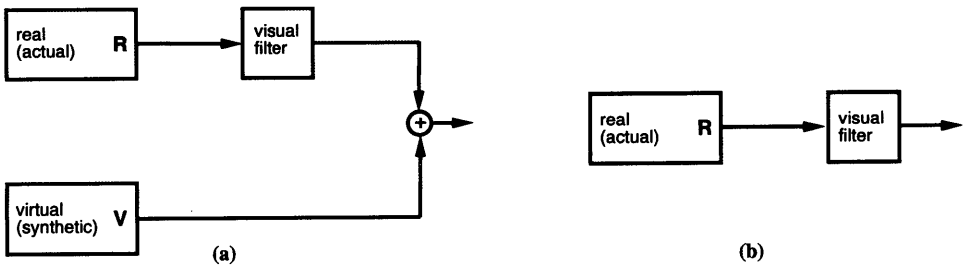


Figure 6-1: **Two equivalent interpretations of mediated reality (MR):** (a) In addition to the ability to add computer-generated (synthetic) material to the wearer's visual world, there is potential to alter reality, if desired, through the application of a 'visual filter'. The coordinate transformation embodied in the 'visual filter' may either be inserted into the virtual channel as well, or the graphics may be rendered in the coordinate system of the filtered reality channel, so that the real and virtual channels are in register. (b) The 'visual filter' need not be a *linear system.* In particular, the 'visual filter' may itself embody the ability to create computer-generated objects and therefore subsume the "virtual" channel.

reconfigure reality. For example, color may be deliberately diminished or completely removed from the real world at certain times when it is desired to highlight parts of a virtual world with graphic objects having unique colors. The fact that vision may be *completely* reconfigured also suggests utility to the visually handicapped.

6.1 Introduction

Ivan Sutherland, a pioneer in the field of computer graphics, described a head-mounted display with half-silvered mirrors so that the wearer could see a virtual world superimposed on reality [77][78], giving rise to "Augmented Reality (AR)".

Others have adopted Sutherland's concept of a Head-Mounted Display (HMD) but generally without the see-through capability. An artificial environment in which the user cannot see through the display is generally referred as a Virtual Reality (VR) environment. One of the reasons that Sutherland's approach was not more ubiquitously adopted is that he did not merge the virtual object (a simple cube) with the real world in a meaningful way. Feiner's group was responsible for demonstrating the viability of AR as a field of research, using sonar (Logitech 3D trackers) to track the real world so that the real and virtual worlds could be registered [79][80]. Other research groups [81] also contributed to this development. Some research in AR arises from work in telepresence [82].

AR, although lesser known than VR, is currently used in some specific applications. Helicopter pilots often use a see-through visor that superimposes virtual objects over one eye, and the F18 fighter jet, for example, has a beam-splitter just inside the windshield that serves as a heads-up display (HUD), projecting a virtual image that provides the pilot with important information.

The general spirit of AR is to *add* computer graphics or the like to the real world. A typical AR apparatus does this with beam splitter(s) so that the user sees directly through the apparatus while simultaneously viewing a computer screen.

The goal of this chapter is to consider a wireless (untethered) apparatus worn over the eyes that, in real time, computationally *reconfigures* reality in addition to adding to it. This 'mediation' of reality may be thought of as a *filtering* operation applied to reality and then a combining operation to insert *overlays* (Fig 6-1(a)). Equivalently, the addition of computer-generated material may be regarded as arising from this filtering operation itself (Fig 6-1(b)).

A means of *mediating* (augmenting, enhancing, deliberately diminishing, or otherwise altering) reality, in real time, through an apparatus worn over the eyes, will first be described using an idealized implementation based on a hypothetical 'lightspace glass', and later in a more practical implementation, using video camera(s), a head-mounted video display, and a combination of body-worn and untethered remote processing hardware. In either case (idealized or practical), the entire

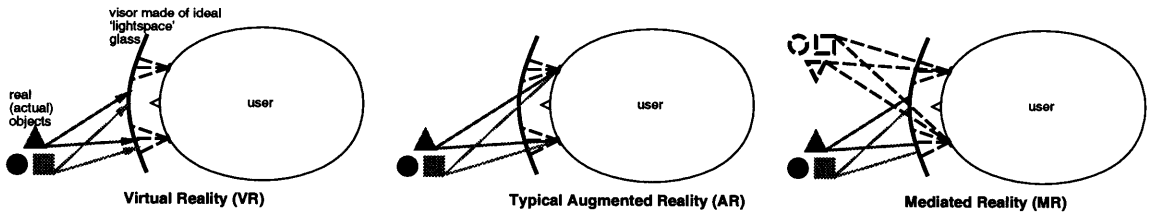


Figure 6-2: Consider a hypothetical glass that absorbs and quantifies every ray of light that hits it, and is also capable of generating any desired rays of light. Such a glass, made into a visor, could produce a virtual reality (VR) experience by ignoring all rays of light from the real world, and generating rays of light that simulate a virtual world. Rays of light from real (actual) objects indicated by solid shaded lines; rays of light from the display device itself indicated by dashed lines. The device could also produce a typical augmented reality (AR) experience by creating the 'illusion of transparency' and also generating rays of light to make computer-generated "overlays". Furthermore, it could 'mediate' the visual experience, allowing the perception of reality itself to be altered. In this figure, a non-useful (except in the domain of psychophysical experiments) but illustrative example is shown: objects are *left-right reversed* before being presented to the viewer.

apparatus will be referred to as a 'Reality Mediator' (RM).

6.1.1 'lightspace glass'

In what follows, a simplified model [33] – *rays* of light – will be used. This simplified model neglects both wave-like properties (such as diffraction) and particle-like properties (such as quantum effects) of light. (The model also assumes the existence of the abstract notion of *instantaneous* frequency). A special window that can both measure and produce rays of light is hypothesized as a conceptual framework upon which the concept of 'mediated reality (MR)' is developed.

Consider a hypothetical window that would absorb and quantify every ray of light incident upon it. The information obtained from such a window — the location on the glass where each ray of light struck, its direction of arrival, its wavelength, and the exact instant in time it struck — would be sufficient to reconstruct all the light passing into the window. Suppose this hypothetical window could also produce any ray of light desired — that it could send out any number of light rays from specified locations on the glass, in specified directions, and of specified wavelengths. This hypothetical 'lightspace glass', would be essentially an ideal plenoptic video camera and ideal plenoptic video display in one.

The closest approximation to this hypothetical lightspace glass might be the holographic video cameras and *holovideo* displays that have actually been built [83][84] (the two have been connected together), though the apparatus occupied some large optical benches and a room full of equipment. It would not even fit in a small room, let alone be small enough to wear in a pair of eye glasses. Nevertheless, perhaps in years to come, these technologies could become feasible.

The use of a hypothetical 'lightspace glass' to measure the way a scene responds to light has been previously proposed [33], together with a more practical realization over a very limited domain ('lightspace' of a static monochromatic scene), though this apparatus has not been realized for moving scenes, and certainly not in a small enough package to be body-worn. However, the 'lightspace glass' has been a useful abstraction for the purposes of understanding the concepts underlying MR. Note that if it were possible to create a sufficiently realistic implementation of 'lightspace glass', it could hypothetically be used to surround an object, as well as all of the support circuitry for the glass itself, and render that object invisible, by virtue of its ability to absorb any ray of light before it got to the object and then re-produce that ray correctly at the other side of the object.

6.1.2 'lightspace glasses'

Suppose that we were to make a visor from this glass. Clearly, it could be used as a VR display, because the *plenoptic camera* functionality of it could absorb and quantify all the incoming rays of light and then simply ignore this information, while the *holovideo* portion of the glass could create a virtual environment for the user. (See Fig 6-2 (VR).)

The glass would also have the capability of functioning like an ordinary window in the sense that the *plenoptic camera* functionality of it could absorb and quantify all the rays of light incident upon it, and then the *holovideo* portion of it could send exactly those same rays of light out the other side. For the moment, assume that the ideal nature of the glass permits it to perfectly sustain the 'illusion of transparency'.

This 'illusion of transparency' would have many uses of its own. Obviously it could be used to make the visor function like a pair of sunglasses darkening rays of light coming out the other side. Because of the computer control, the darkening could even vary, in accordance with some gradient that would be darker up where the sun was, resulting in 'smart sunglasses'. The 'smart sunglasses' would use machine vision to track the sun and adjust the position of the darkening mask.

Many conventional sunglasses have a fixed gradient, typically being darker at the top than at the bottom, which creates an annoying artificial percept of motion when the wearer tilts his or her head back or forward, even though nothing in the scene is moving. "Smart sunglasses" would eliminate this problem by fixing the darkening mask with respect to the scene rather than the wearer.

Now in addition to creating the illusion of allowing light to pass right through, the visor could also create new rays of light, having nothing to do with the rays of light coming into it. The combined 'illusion of transparency' and the new light would provide the wearer with a typical AR experience (Fig 6-2 (AR)).

In an AR environment, graphics sometimes fail to stand out from the real objects. For example, when looking through the glasses at a brightly colored scene, there may not exist a unique color to use for the overlays. Suppose, however, that the glass, instead of creating an illusion of transparency, creates an illusion of being *achromat transparent.* Being *achromat transparent* means that each incoming ray of light is absorbed and quantified, and its wavelength is ignored. A ray from the same location, is sent out in the same direction, at the same time, but with a flat (grey) spectrum. This would make the user colorblind to real objects, making the real world appear less "busy" when combined with some colorful computer-generated overlays where color could be used, more effectively, to accentuate the virtual objects. This would prevent computer-generated objects from being "lost" in the clutter of the real world.

Using practical (non-plenoptic) implementations of MR (to be described in Sec 6.2), I have found color-reduced reality mediation to be quite useful. For example, when I am comfortably seated on a commercial airline or commuter train and wish to read text on my screen (e.g. read email), I like to "tone down" my surroundings so they take on a lesser role. I do not wish to be blind to my surroundings, as is someone who is reading a newspaper (newspapers can easily end up covering most of a person's visual field).

This form of reality mediation allows me to focus primarily on the virtual world which might, for example, be comprised of email, a computer source file, and other miscellaneous work, running in emacs19, with colorful text, where the text colors are chosen so that no black, white, or grey text (text colors that would get 'lost' in the new reality) is used. My experience is like reading a newspaper printed in brightly colored text on a transparent material, behind which the world moves about in black and white. I am completely aware of the world behind my "newspaper" but it does not distract from my ability to read the "paper".

Alternatively, the real world could be left in color, but the color mediated slightly so that unique and distinct colors could be reserved for virtual objects and graphics overlays. In addition to this 'chromatic mediation', other forms of 'mediated reality' are often useful.

Registration between real and virtual worlds

Alignment of the real and virtual worlds is very important, as indicated in the following quote [85]:

> Unfortunately, registration is a difficult problem, for a number of reasons. First, the human visual system is very good at detecting even small misregistrations, because of the resolution of the fovea and the sensitivity of the human visual system to differences. Errors of just a few pixels are noticeable. Second, errors that can be tolerated in Virtual Environments are not acceptable in Augmented Reality. Incorrect viewing parameters, misalignments in the Head-Mounted Display, errors in the head-tracking

system, and other problems that often occur in HMD-based systems may not cause detectable problems in Virtual Environments, but they are big problems in Augmented Reality. Finally, there's system delay: the time interval between measuring the head location to superimposing the corresponding graphic images on the real world. The total system delay makes the virtual objects appear to "lag behind" their real counterparts as the user moves around. The result is that in most Augmented Reality systems, the virtual objects appear to "swim around" the real objects...

Until the registration problem is solved, Augmented Reality may never be accepted in serious applications.

(emphasis added)

The problem with many implementations of AR is that even once registration is attained, if the glasses slip down your nose, ever so slightly, the real and virtual worlds will not generally remain in perfect alignment.

Using the 'illusory transparency' approach, the illusion of transparency is perfectly coupled with the virtual world once the signals (video or perhaps in the future, *holovideo*) corresponding to the real and virtual worlds are put into register and combined into one signal. Not all applications lend themselves to easy registration at the signal level, but those that do (such as the finger-pointing principle to be discussed in Chapter 7) call for the 'illusory transparency' approach of mediated reality as opposed to an augmented reality overlay. In mediated reality, when the glasses slip down the nose a little (or a lot for that matter), both the real and virtual worlds slip down together in a unified way, and remain in perfect register. Since they are both the same medium (e.g. video or the like), once registration is attained between the real and virtual video signals themselves, the registration problem remains solved regardless of how the glasses might slide around on the wearer, or how the wearer's eyes are focused or positioned with respect to the glasses.

Another important point is that even with perfect registration, when using a see-through visor (with beam splitter or the like), real objects may lie in a variety of different depth planes, while virtual objects are generally flat (in each eye that is), to the extent that their distance is at a particular focus (apart from the variations in binocular disparity of the virtual world). This is not too much of a problem when all of the objects are far away as is often the case in aircraft (e.g. in the fighter jets using HUDs), but in many other applications (such as in a typical building interior) the differing depth planes destroy the illusion of unity between real and virtual worlds.

With the 'illusory transparency' approach, however, the real and virtual worlds exist in the same medium and therefore are not only registered in location but also in depth, since the depth limitations of the display device affect both the virtual and real environments in exactly the same way.

Transformation of the perceptual world

Even without any graphics overlays, mediated realities are still interesting and useful. For example, the colorblinding glasses in themselves might be useful to an artist trying to study relationships between light and shade. While it is certain that the average person might not want any part of this experience, especially given the cumbersome nature of the current realization of the RM, without question, there are at least a small number of users who would be willing to wear an expensive and cumbersome apparatus in order to see the world in a different light. Consider, for example, the artist who travels halfway around the world to see the morning light in Italy. As the cost and size of the RM decreases, no doubt there would be a growing demand for glasses that alter (enhance or diminish) tonal range, allowing artists to manipulate contrast, color, etc..

MR glasses could (in principle) be used to synthesize the effect of ordinary glasses, but with a computer-controlled prescription that would modify itself, while conducting automatically scheduled eye tests on the user.

The RM might also, for example, reverse the direction of all outgoing light rays to allow the wearer to live in an "upside-down" world (Fig 6-2 (MR)), perhaps being useful for experiments in psychology. Although the vast majority of RM users of the future will no doubt have no desire to

live in an upside-down, left-right-reversed, or sideways rotated world, these visual worlds serve as illustrative examples of extreme reality mediation.

In his 1896 paper [86], George Stratton reported on experiments in which he wore eyeglasses that inverted his visual field of view. Stratton argued that since the image upon the retina was inverted, it seemed reasonable to examine the effect of presenting the retina with an "upright image".

His "upside-down" glasses consisted of two lenses of equal focal length, spaced two focal lengths, so that rays of light entering from the top would emerge from the bottom, and vice-versa. Stratton, upon first wearing the glasses, reported seeing the world upside-down, but, after an adaptation period of several days, was able to function completely normally with the glasses on.

Dolezal [1] (page 19) describes "various types of optical transformations", such as the *inversion* explored by Stratton, as well as *displacement*, *reversal*, *tilt*, *magnification*, and *scrambling*. Kohler [2] also discusses "transformation of the perceptual world".

Each of these "optical transformations" could be realized by selecting a particular *linear time-invariant system* as the visual filter in Fig 6-1. (A good description of *linear time-invariant systems* may be found in a communications or electrical engineering textbook such as [87].)

The optical transformation to greyscale, described earlier, could also be realized by a 'visual filter' (Fig 6-1 (a)) that is a linear time-invariant system, in particular, a *linear integral operator* [88] (page 669) that, for each ray of light, collapses all wavelengths into a single quantity giving rise to a ray of light, having a flat spectrum, emerging from the other side.

Of course, the 'visual filter' of Fig 6-1 (b) may not, in general, be realized through a *linear system*, but there exists an equivalent *nonlinear* filter arising from incorporating the generation of virtual objects into the filtering operation.

One final and somewhat amusing note on 'lightspace glasses' is in order. When I am wearing a practical implementation of the RM, (with head-mounted display) in my day-to-day social interactions in public, people often complain of a loss of eye contact with me. However, if I were wearing the hypothetical 'lightspace glasses' I would also be wearing my own personal plenoptic video display, because these glasses would be able to produce any collection of light rays. This would allow me to present any desired 3D image to those who look into the glasses. In particular, the loss of eye-contact that people complain about when they try to talk to me could be eliminated by using the glasses to generate a *holovideo* of my eyes. Of course if I were sleeping through a boring meeting, I could still present others with a *holovideo* of wide open eyes dancing in attentive saccades rather than the actual view of closed eyes.

6.2 Non-plenoptic realizations of MR

6.2.1 'Video transparency'

The idealized *plenoptic video* camera is difficult to make in practice, since one would require a microscopic lattice of photocells or the like with unrealizably fast response. Even a discrete realization of a plenoptic video camera made from a dense array of miniature video cameras is costly and bulky. The *holovideo* display is even more so.

However, since the visor may be relatively well fixed with respect to the wearer, there is not really a great need for full parallax plenoptic video. The fixed nature of the visor conveniently prevents the wearer from having *look-around* with respect to the visor itself (e.g. look-around is accomplished when the user and the visor move together to explore the space). Thus two views, one for each eye, suffice to create a reasonable 'illusion of transparency'.

Others [81][82][89] have also explored video-based 'illusory transparency', augmenting it with virtual overlays. Nagao, in the context of his hand-held TV set with single hand-held camera [89] calls it "video see-through".

It is worth noting that whenever 'illusory transparency' is used, as in the work of [81][82][89] reality will be 'mediated', whether or not that mediation was intended. At the very least this mediation takes on the form of limited dynamic range and color gamut, as well as some kind of distortion, which may be modeled as a 2D coordinate transformation. Since this mediation is inevitable, it is worthwhile to attempt to exploit it, or at least plan for it, in the design of the

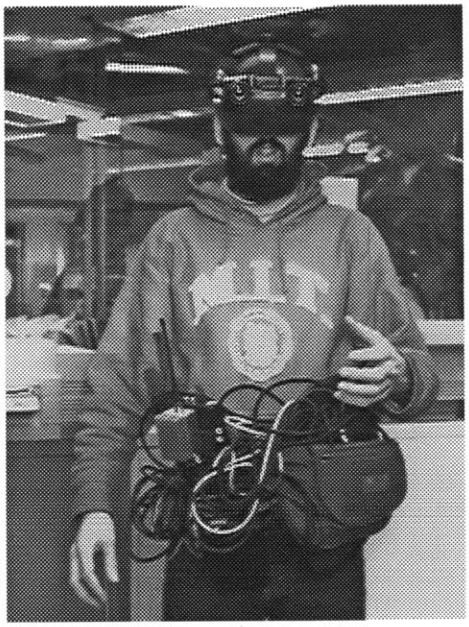


Figure 6-3: 'Reality mediator' as of late 1994, showing a color stereo head-mounted display (VR4) with two cameras mounted to it. The inter-camera distance and field of view match approximately my interocular distance and field of view with the apparatus removed. The components around my waist comprise radio communications equipment (video transmitter and receiver). The antennas are located at the back of the head-mount to balance the weight of the cameras, so that the unit is not front-heavy. (C) Steve Mann, 1994.

apparatus. A 'visual filter' may even be used to attempt to mitigate the distortion.

A practical color stereo 'reality mediator (RM)' may be made from video cameras and display. One example, made from a display having 480 lines of resolution, is depicted in Fig 6-3.

It is desired to have the maximum possible 'visual bandwidth', even if the RM is going to be used to conduct experiments on *diminished reality*. For example, the apparatus of Fig 6-3 may be used to experience colorblindness and reduced resolution by applying the appropriate 'visual filter' to select the desired degree of degradation in a controlled manner that can also be automated by computer. (For example, color can be gradually reduced over the course of a day, under program control, to the extent that I can become color blind but not even realize it.)

I mounted the cameras the correct interocular distance apart, and used cameras that had the same field of view as the display devices. With the cameras connected directly to the displays, the illusion of transparency was realized to some degree, at least to the extent that each ray of light that entered the apparatus (e.g. was absorbed and quantified by the cameras) appeared to emerge at roughly the same angle (by virtue of the display).

Although I had no depth-from-focus capability there was, enough depth perception remaining on account of the stereo disparity for me to function somewhat normally with the apparatus. Depth-from-focus is what is sacrificed in working with a non-plenoptic RM.

A first step in using a reality mediator is to wear it for a while to become accustomed to its characteristics. Unlike in typical beam-splitter implementations of augmented reality, transparency, if desired, is synthesized, and therefore only as good as the components used to make the RM.

I wore the apparatus in identity map configuration (cameras connected directly to the displays) for several days. I could easily walk around the building, up and down stairs, through doorways, to and from the lab, etc. I did, however, experience difficulties in scenes of high dynamic range, and also in reading fine print (such as a restaurant menu or a department store receipt printed in faint ink when the ribbon was near the end of its useful life).

The unusual appearance of the apparatus was itself a hindrance in my daily activities (for example when I wore it to a formal dinner), but after some time people appeared to become accustomed to seeing me this way.

The attempt to create an illusion of transparency was itself a useful experiment because it established some working knowledge of what can be performed when vision is *diminished* or *degraded* to RS170 resolution and field of view is somewhat limited by the apparatus.

Knowing what can be performed when reality is mediated (e.g. diminished) through the limitations of a particular HMD (e.g. the VR4) would be useful to researchers who are designing VR environments for that HMD, because it establishes a sort of *upper bound* on "how good" a VR environment could ever hope to be when presented through that particular HMD. A reality mediator may also be useful to those who are really only interested in designing a traditional beam-splitter-based AR system because RM could be used as a development tool, and could also be used to explore new conceptual frameworks.

6.2.2 Mediated presence

McGreevy [90] also explored the use of a head-mounted display directly connected to two cameras, although his apparatus had no computational or processing capability. His head-mounted camera/display system was a very simple form of reality mediator, where the mediation was fixed (e.g. since there was no processing or computer control).

He was interested in showing that despite the fact that his two subjects had essentially immediate response (essentially no delay) and also had the luxury of perfect (e.g. un-mediated) touch, hearing, smell, etc., they still had much difficulty adapting to the mediated visual reality in which they were placed. This showed that no matter how good a VR or telepresence simulation could be, there would still be significant limitations imposed by the interface between that simulation and the user – the HMD.

However, in his experiments he also noted that the system deprived the user of both color and stereo vision. Even though it had 2 cameras: "The lack of stereo vision provided in the head-mounted display prompted both of his experimental subjects to use alternative cues to make spatial judgements [90]"[1] (Both subjects would move their heads back and forth to perceive depth from the induced motion parallax.)

Something McGreevy appeared less interested in, which is far more important in the context of this thesis, is the notion of a deliberately diminished reality, which underscores the important difference between mediated reality and augmented reality.

Most notably, living in a deliberately diminished reality allows us to come closer to what others will experience when they share this experience of reality (either at a later time, a different place, or both). In Chapter 7, I will explain how living in a depth-deprived (2-D) world that is also deliberately diminished in other ways gives rise to a proposed new genre of cinematography and other related experiences.

In order to be able to experiment with diminished reality in a controlled manner, it is desirable to first attain a system that can come close to passing reality through with more bandwidth than desired, so that the exact desired bandwidth can be found experimentally. The system of Fig 6-3 overcame some of the problems associated with McGreevy's system – having 34 times more 'visual bandwidth', it was just at the point where it was possible to conduct much of my daily life through this illusion of transparency. I could degrade my RM down to the level of McGreevy's system, and beyond, in a computer-controlled fashion to find out how much 'visual bandwidth' I needed to conduct my daily affairs. (The 'visual bandwidth' is calculated as the number of pixels times two for stereo, times another three for color, although one might argue that there is redundancy because left and right, as well as color channels are quite similar to one another.) I found, for various tasks, a certain point at which it was possible to function in the RM. In particular, I found that anything below about 1/8 of my system's full bandwidth (about four times that of McGreevy's system) made most tasks very difficult, or impossible.

Once it is possible to live within the shortcomings of the RM's ability to be 'transparent', new and interesting experiments can also be performed.

[1] I have been unsuccessful in contacting McGreevy to determine how he routed the signals from two cameras into a non-stereo display.

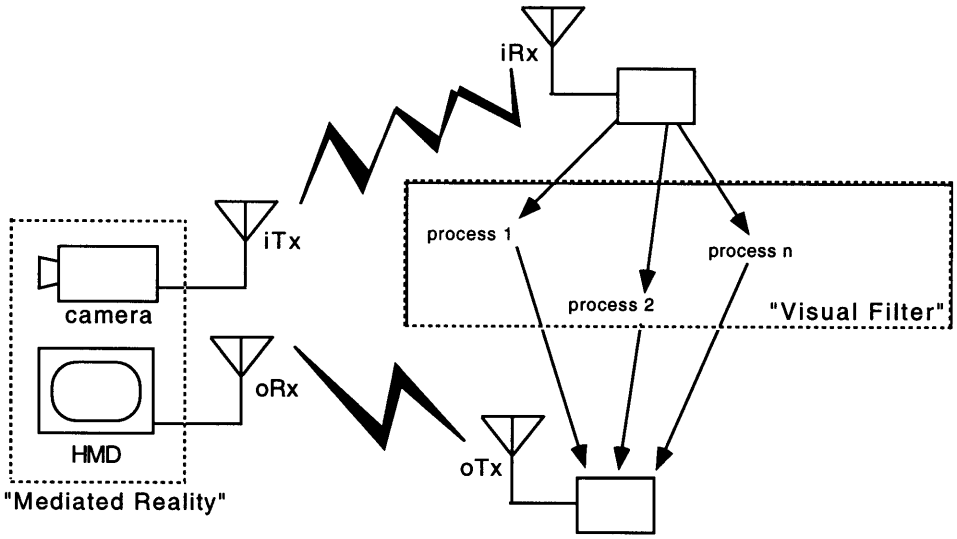


Figure 6-4: **Simple implementation of a 'reality mediator (RM)'.** The camera sends video to one or more computer systems over a high-quality microwave communications link, which I refer to as the 'inbound channel'. The computer system(s) send back the processed image over a UHF communications link which I refer to as the 'outbound channel'. Note the designations "i" for inbound (e.g. iTx denotes inbound transmitter), and "o" for outbound. 'visual filter' refers to the process(es) that mediate(s) the visual reality and possibly insert "virtual" objects into the reality stream.

6.2.3 Video mediation

Once the apparatus is worn long enough to be comfortable with the 'illusory transparency', mediation of the reality can begin.

In the early days of this research (up until the early to mid 1990s) I attained a "virtual WearComp" system by wearing the I/O portion of the computer (display and input devices including camera), while situating the actual computer remotely.

This remote situation of the computer was attained by establishing a full-duplex video communications channel between the display/camera portion of the reality mediator and the host computer(s). In particular, a high-quality communications link (which I call the 'inbound-channel') is used to send the video from my cameras to the remote computer(s), while a lower quality communications link (the 'outbound channel') is used to carry the processed signal from the computer back to my HMD. This apparatus is depicted in a simple diagram (Fig 6-4). Ideally both channels would be of high-quality, but the machine-vision algorithms were found to be much more susceptible to noise than was my own vision (e.g. I could still find my way around in a "noisy" reality, and still interact with "snowy" virtual objects).

Originally, communication was based on antennas that I had installed on the roof of the tallest building in the city but later I found that if I moved one of the antennas to another rooftop that the inbound/outbound channel separation was dramatically improved. The apparatus provided good coverage on the university campus, moderate coverage over a good part of the city, and some coverage from a nearby city.

With this remote computational approach, I was able to simulate the operation of future more recent generations of WearComp, even though they were, at the time, not technologically feasible in a self-contained body-worn package.

To a very limited extent, looking through a camcorder provides a 'mediated reality' experience, because we see the real world (usually in black and white, or in color but with a very limited color fidelity) together with virtual text objects, such as shutter speed and other information about the camera. If, for example, the camcorder has a black and white viewfinder, the 'visual filter' (the colorblindness one experiences while looking through the viewfinder with the other eye closed) is unintentional in the sense that the manufacturer would rather have provided a full-color viewfinder. This is a very trivial example of a mediated reality environment where the filtering operation is

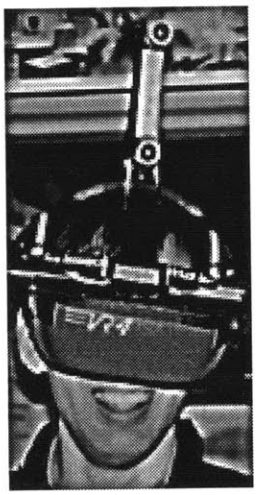


Figure 6-5: **Living in a "Rot 90" world:** It was found to be necessary to rotate both the cameras rather than just rotate each one. Thus, it would not seem possible to fully adapt to, say, a prism that rotated the image of each eye, but the use of cameras allows the up-down placement of the "eyes". The parallax, now in the up-down direction, affords a similar sense depth as we normally experience with eyes spaced from left to right together with left-right parallax.

unintentional but nevertheless present.

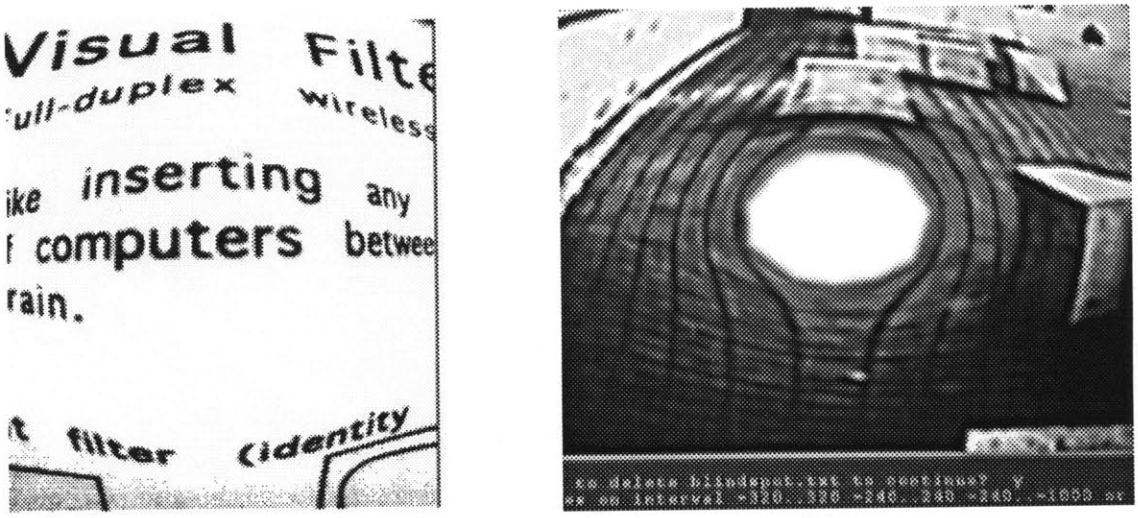
Although the colorblinding effect of looking through a camcorder may be undesirable most of the time, there are times when it is desirable. The 'diminished reality' it affords may be a desired artifact of the 'reality mediator' (for example in the case where the user chooses to remove color from the scene either to "tone-down" reality or to accentuate the perceptual differences between light and shade). This simple example points out the fact that a 'mediated reality' system need not function as just a 'reality enhancer', but rather, it may enhance, alter, or deliberately *degrade* reality.

Stuart Anstis [91], using a camcorder that had a "negation" switch on the viewfinder, experimented with living in a "negated" world. He walked around holding the camcorder up to one eye, looking through it, and observed that he was unable to learn to recognize faces in a negated world. His negation experiment bore a similarity to Stratton's inversion experiment mentioned in Sec 6.1.2, but the important difference within the context of this chapter is that Anstis experienced his mediated visual world through a video signal. In some sense both the regular eyeglasses that people commonly wear, as well as the special glasses researchers have used in prism adaptation experiments [2][1] are reality mediators, but it appears that Anstis was the first to explore, in detail, an electronically mediated world.

6.2.4 The reconfigured eyes

Using my 'reality mediator', I repeated the classic experiments like those of Stratton and Anstis (e.g. living in an upside-down or negated world), as well as some new experiments, such as learning to live in a world rotated 90 degrees. However, in this sideways world, I found that I could not adapt to having each of the images rotated by 90 degrees separately, but had to rotate the cameras together (Fig 6-5).

The video-based RM (e.g. Fig 6-3) permits me to experience any coordinate transformation that can be expressed as a mapping from a 2D domain to a 2D range, in real time (30frames/sec = 60fields/sec) in full color, because a full-size remote computer (e.g. SGI Reality Engine) is used to perform the coordinate transformations. This apparatus allows me to experiment with various computationally-generated coordinate transformations both indoors and outdoors, in a variety of different practical situations. Examples of some useful coordinate transformations appear in Fig 6-6.



(a) (b)

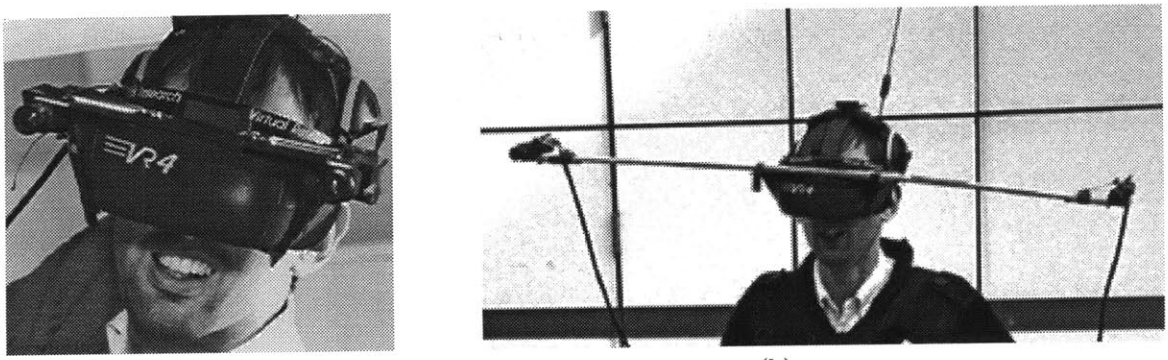
Figure 6-6: **Living in coordinate-transformed worlds:** Color video images are transmitted, coordinate-transformed, and then received back at 30 frames per second – the full frame-rate of the VR4 display device. (a) This 'visual filter' would allow a person with very poor vision to read (due to the central portion of the visual field being hyper-foveated for a very high degree of magnification in this area), yet still have good peripheral vision (due to a wide visual field of view arising from demagnified periphery). (b) This 'visual filter' would allow a person with a *scotoma* (a blind or dark spot in the visual field) to see more clearly, once having learned the mapping. The visual filter also provides edge enhancement in addition the coordinate transformation. Note the distortion in the cobblestones on the ground and the outdoor stone sculptures.

Researchers at Johns Hopkins University have been experimenting with the use of cameras and head-mounted displays for helping the visually handicapped. Their approach has been to use the optics of the cameras for magnification, together with the contrast adjustments of the video display to increase apparent scene contrast [92]. They also talk about using *image remapping* in the future:

> One of the most exciting developments in the field of low vision is the Low Vision Enhancement System (LVES). This is an electronic vision enhancement system that provides contrast enhancement... *Future enhancements* to the device include text manipulation, autofocus and *image remapping.*

(quote from their WWW page [92], emphasis added). Their research effort suggests the utility of the real-time visual mappings (Fig 6-6) previously already implemented using the apparatus of Fig 6-3.

The idea of living in a coordinate transformed world has been explored extensively by other authors [2][1], using optical methods (such as prisms and the like). Much could be written about my experiences in various electronically coordinate transformed worlds, but a detailed account of all of the various experiences is beyond the scope of this chapter. Of note, however, I observed that visual filters differing slightly from the identity (e.g. rotation by a few degrees) had a more lasting impression on me when I removed my apparatus (e.g. left me incapacitated for a greater time period upon removal of the apparatus), than visual filters that were far from the identity (e.g. rotation by 180 degrees – upside-down). Furthermore, the visual filters close to the identity tended to leave me with an opposite aftereffect (e.g. I'd consistently reach too high after taking off the RM where the images had been translated down slightly, or reach too far 'clockwise' after removing the RM that had been rotating images a few degrees counterclockwise). Visual filters far from the identity (such as reversal or upside-down mappings) did not leave me with an opposite aftereffect: I would **not** see the world as being upside down upon removing upside-down glasses. I think of this phenomenon as being analogous to learning a second language (either a natural language or computer language). When the second language is similar to the one we already know, we make more mistakes switching back and forth than when the two are distinct. When two (or more) adaptation spaces were distinct, for example, in the case of the identity map and the rotation operation ('rot 90'), I could sustain



(a) (b)

Figure 6-7: **Giant's eyes: extended baseline.** (a) With a 212mm baseline, I could function in most everyday tasks, but would see crosseyed at close conversational distances. (b) With a 1m baseline, I could not function in most situations, but had a greatly enhanced sense of depth for distant objects (e.g. while looking out across the Charles river). Wires from the cameras go down into my waist bag containing the rest of the apparatus. Inbound transmit antenna is just visible behind my head.

a dual adaptation space and switch back and forth between the identity operator and the 'rot 90' operator without one causing lasting aftereffects in the other.

Regardless of how much care is taken in creating the illusion of transparency, there will be a variety of flaws, not the least of which is limited resolution, lack of dynamic range, limited color (mapping from the full spectrum of visible light to three responses of limited color gamut), and improper alignment and placement of the cameras. In Fig 6-3, for example, the cameras are mounted *above* the eyes. Even if they are mounted in front of the eyes, they will extend, putting me in the visual world of some hypothetical organism that has eyes that stick out of its head some 3 or 4 inches (except in the case of a blind person in the future when sufficient technological advances permit using cameras for artificial eyes). Thus some adaptation is almost always needed.

After wearing my apparatus for an extended period of time, I eventually adapted, despite its flaws, whether these be unintended (e.g. limited dynamic range, limited color gamut, etc.), or intended (e.g. deliberately presenting myself with an upside-down image). In some sense I subsumed the visual reconfiguration induced by the apparatus into my brain, so that the apparatus and I act as a single unit. Manfred Clynes uses the example of a person riding a bicycle to describe this sort of synergism [93] where, after sufficient adaptation time, conscious effort is no longer needed in order to use the machine. He refers to this state as *cyborgian* [94]. Thus, perhaps by long-term adaptation to a reality mediator, one becomes a *cyborg*.

Giant's Eyes

I found that having the cameras above the display (as in Fig 6-3) induced some parallax error for nearby objects, so I tried mounting the cameras at the sides of my head (Fig 6-7(a)). This gave me an interocular distance of approximately 212mm, resulting in an enhanced sense of depth. Objects appeared smaller and closer than they really were – the world looked like a size-reduced scale-model of reality. While walking home that day (wearing the apparatus), I felt that I had to duck down to avoid hitting what appeared to be a low tree branch. However, my recollection from previous walks home had been that there were no low branches on the tree, and, removing my RM, I noticed that the tree branch that appeared to be within arm's reach was several feet in the air. After some time I got used to this enhanced depth perception, and then tried mounting the cameras on a 1 meter baseline. Crossing the street, I had the illusion of small toy cars moving back and forth very close to my nose, and I had the feeling that I could just push them out of my way, but my better judgement served to make me wait until there was a clearing in the traffic before crossing the road to get to the river. Looking out across the river, I had the illusion that the skyscrapers on the other side were within my arm's reach in both distance and height.

'Slowglasses'

Suppose that we had a hypothetical glass of very high refractive index. (Science fiction writer Bob Shaw refers to such glass as *slowglass* [95]. In Shaw's story, a murder is committed and a piece of slowglass is found at the scene of the crime – the glass being turned around as curious onlookers wait for the light present during the crime to emerge from the other side.) Every ray of light that enters one side of the glass comes out the other side unchanged, but simply delayed. A visor made from *slowglass* would present the viewer with a full-parallax delayed view of a particular scene, playing back with the realism of the idealized *holovideo* display discussed in Sec 6.1.2.

A practical (non-plenoptic) implementation of this illusion of *delayed transparency* was created using the reality mediator (Fig 6-3) with a video delay.

As is found in any poor simulation of virtual reality, wearing 'slowglasses' induces a similar dizziness and nausea to reality. After experimenting with various delays one will develop an appreciation of the importance of moving the information through the RM in a timely fashion to avoid this unpleasant delay.

'Edgertonian' Eyes

Instead of a fixed delay of the video signal, I experimented by applying a repeating freeze-frame effect to it (with the cameras' own shutters set to 1/10000 second). With this video *sample and hold*, I found that nearly periodic patterns would appear to freeze at certain speeds. For example, while looking out the window of a car, periodic railings that were a complete blur without my RM would snap into sharp focus with the RM. Slight differences in each strut of the railing would create interesting patterns that would dance about revealing slight irregularities in the structure. (Regarding the nearly periodic structure as a true periodic signal plus noise, the noise is what gave rise to the interesting patterns). Looking out at another car, traveling at approximately the same speed as me, I could read the writing on the tires, and easily count the number of bolts on the wheel rims. Looking at airplanes, I could see the number of blades on the spinning propellers, and, depending on the sampling rate of my RM, the blades would appear to rotate slowly backwards or forwards, in much the same way as objects do under the stroboscopic lights of Harold Edgerton [14]. By manually adjusting the processing parameters of my RM, I could see many things that escape normal vision.

Virtual 'smart strobe'

By applying machine vision (some rudimentary intelligence) to the incoming video, the RM should be able to decide what sampling rate to apply. For example, it should recognize a nearly periodic or cyclostationary signal and adjust the sampling rate to lock onto the signal much like a phase-locked loop. A sufficiently advanced RM with eye tracking and other sensors might make inferences about what you'd like to see, and, for example, when looking at a group of airplanes in flight would freeze the propeller on the one you were concentrating on.

Wyckoff's world

One of the problems with the RM is the limited dynamic range of CCDs. One possible solution is to operate at a higher frame rate than needed, while underexposing, say, odd frames and overexposing even frames. The shadow detail may then be derived from the overexposed stream, the highlight detail from the underexposed stream, and the midtones from a combination of the two streams. The resulting extended-response video may be displayed on a conventional HMD by using Stockham's *homomorphic filter* [24] as the 'visual filter'. The principle of extending dynamic range by combining differently exposed pictures is known as the Wyckoff principle [48], in honor of Charles Wyckoff. Using a Wyckoff composite, I could be outside on bright sunny days and see shadow detail when I looked into open doorways to dark interiors, as well as see detail in bright objects like the sun.

The Wyckoff principle is also useful in the context of night vision because of the high contrasts encountered at night. In the Wyckoff world, one can read rating numbers printed on a bright

(a) (b)

Figure 6-8: **Partially mediated reality:** (a) Half MR: My right eye is *completely* immersed in a mediated reality environment arising from a camera on my right, while my left eye is free to see unmediated real-world objects. (b) Substantially less than half MR: My left eye is *partially* immersed in a mediated reality environment arising from a camera also on my left. (C) Betty and Steve Mann, Jul. 1995.

mercury vapor arc lamp and also see into the darkness off in the distance behind the lamp, neither brightness extreme of which is visible to the naked eye.

6.2.5 Conclusion of Sec 6.2

With high-quality cameras and display devices, the illusion of transparency was found to be sufficiently good that I was able to function comfortably in many of my day-to-day tasks. Further improvements in the technology suggest the possibility for creating an even more comfortable illusion of transparency. Once this illusion is assimilated and accepted (e.g. after long-term adaptation), for all practical purposes, the cameras behave as the eyes, except now signals may be both extracted from them, and inserted into where they were connected. Such an assimilation of this illusion is discussed in 'The Eudaemonic Eye' [96]. Furthermore, the signal path may now be conveniently interrupted so that a 'visual filter' may be installed, to some degree, behaving as though it were positioned between the eye and the brain.

6.3 Partially mediated reality

Artificial Reality is a term defined by Myron Krueger to describe video-based, computer-mediated interactive media [77]. His apparatus consisted of a video display (screen) with a camera above it that projected a 2D outline of the user together with sprite-like objects. Myron Krueger's environment is a partially mediated reality, in the sense that within the screen the reality is mediated, but the user is also free to look around the room and see unmediated real objects. For example, the part of the user's visual field that shows his or her "reflection" (a left-right reversed video image of the camera is superimposed with computer graphic objects) is a mediated-reality zone, while the periphery (e.g. the user's own feet which can be seen by looking straight down) is outside this mediation zone.

The Artificial Life Interactive Video Environment (ALIVE) [97] is similar to Myron Krueger's environment. In the ALIVE, a user sees him/her self in a "magic mirror" created by displaying a left-right reversed video image from a camera above the screen. Virtual objects, appear, for example, a virtual dog will come over and greet the user. ALIVE is also a partially mediated reality.

6.3.1 Monocular mediation

A camera and a display device completely covering only one eye (Fig 6-8(a)) can be used to create a partially mediated reality. In the apparatus of Fig 6-8(a) my right eye sees a green image (processed NTSC on a VGA display) which becomes fused with the unobstructed (full-color) view through my left eye.

Often the mediated and unmediated zones are in poor register and I cannot fuse them. The poor register may even be deliberate, e.g. I often like to have my right eye in a rotated ('rot 90') world even though this means that I cannot see in stereo in a meaningful way. However, I can still switch my concentration back and forth. I am able to selectively decide to concentrate on one or the other of these two worlds.

An RM made from a camera and a Virtual Vision television set permits a mediation of even lesser scope to take place. Not only does it play into just one eye, but the field of view of the display only covers part of that eye (Fig 6-8(b)). The visor is transparent so that both eyes can see the real world (although my left eye is partially blocked). With these glasses, I might see an object with both eyes, through the transparent visor, and then look over to the 'mediation zone' where my left eye sees, "through" the illusion of transparency in the display. Again, I can switch my attention back and forth between the mediated reality and ordinary vision. I see a double-vision effect (e.g. when I look at someone's face through the glasses of Fig 6-8(b), I often see two replicas of their face, the one that is mediated, and the one that is not). This doubling effect, due to imperfect registration between the mediated and unmediated zones, may or may not be a problem depending on how the RM is used. For example, if I present the mediated world as grey, it remains distinct from the unmediated world, and I am able to mentally switch back and forth between seeing directly, and living in the mediated world, even though the two overlap almost exactly. I often even have the camera present the images in 'rot 90' and then switch my concentration back and forth, having a dual adaptation space. Thus, depending on the application or intent, there may be desire to register or to deliberately misregister the possibly overlapping direct and mediated zones.

6.4 Seeing 'eye-to-eye'

With two personal imaging systems, configured as reality mediators of the kind depicted in Fig 6-8(b), I would set the output frequency of one to the input frequency of the other, and vice versa, so that someone else would see through my eyes and me through the other person's eyes. The Virtual Vision glasses allowed me to concentrate mainly on what was in my own visual field of view (because of the transparent visor), but at the same time have a general awareness of the other person's visual field. This 'seeing eye-to-eye' as I called it, allowed for an interesting form of collaboration. Seeing eye-to-eye through the apparatus of Fig 6-3 requires a *picture in picture* process (unless one wishes to endure the nauseating experience of looking *only* through the other person's eyes), usually having the wearer's own view occupy most of the space, while using the apparatus of Fig 6-8(b) does not require any processing at all.

Usually when we communicate (e.g. by voice or video) we expect the message to be received and concentrated on, while when 'seeing eye-to-eye' there is not the expectation that the message will *always* be seen by the other person. Serendipity was the idea, where each of us would sometimes pay attention and sometimes not.

In Chapter 9, I will present the notion of larger communities of online individuals, connected wirelessly through various forms of wearable devices.

6.5 Life through the screen: Reconfigured eyes in the age of the Internet

Due to the nature of the reality mediator, the exact information I have within my visual field of view is typically available at one or more remote sites. Since I live my life *through the screen*, the resulting video is much more indicative of my experience, than, say, a camera attached to a sports helmet or the like. In fact, I have used this "life through the screen" property of MR to create documentary videos, and am hoping that this will give rise to a new genre of personal documentary cinema verité. This work will be discussed in Chapter 7.

Furthermore, I often maintained a station log (recording of transmitted video), for a variety of reasons:

1. Firstly, to revisit and debug technical difficulties with the system.
2. Secondly, to defend the radio license against possible allegations of inappropriate use (e.g. to prove that a callsign was transmitted regularly, etc.). (Many radio stations keep a station log — to serve as evidence.)
3. Thirdly, as a scientific record of the adaptation process.

The transmission and recording of video raises some interesting privacy concerns. In fact, one of my reasons for envisioning, designing, and building the apparatus was to re-situate the video camera in a disturbing and disorienting fashion in order to challenge our pre-conceived notions of video surveillance in society. These social and privacy issues are dealt with in Chapters 8 and 9.

6.5.1 Shared environment maps

Personal imaging provides some new methods for individuals to interact and communicate through the use of shared environment maps. In particular, the material presented in Chapter 5, may be used in the context of 'painting with looks' — building environment maps by looking around. When these are transmitted to a WWW page or the like, others can remotely experience the space (e.g. navigate around in the environment map) in much the same manner as one navigates around in a QuickTime VR or other environment map, except that it is in real-time, e.g. being navigated while it is still being created.

Shared environment maps (Fig 6-9) will help us allow others to not only experience our point of view vicariously, but will also allow us to allow others to mediate our perception of reality. Such mediation may range from simple annotation of objects in our "reality stream", to completely altering our perception of reality.

6.5.2 The visual memory prosthetic

I use my apparatus as an artist's sketch pad of sorts, useful for taking down visual "notes", and helping me overcome my visual memory disability. Supplementing part of the brain with computer memory accessible on the Internet, there is no reason why one should ever need to forget what someone looks like. In addition to video input on my wearable apparatus, I have a variety of other devices, such as biosensors. With biosensors, I hope it will be possible to have a visual memory aid that has an awareness of my *affective state* [98] and operates without conscious thought or effort [94].

Others have more recently used wearable computers for enhanced memory [99], but for text rather than pictures. Such text-based memory aids include a program that continually runs in the background and helps the user remember text that he or she has previously typed [99].

6.6 Wearable Interactive Video Environment (WIVE)

6.6.1 Equipment repair

In a current collaborative project with Thad Starner, a wearable system is being developed (Fig 6-10) to allow a technician to repair a piece of equipment and see a computer graphics repair manual with drawings superimposed on the real world. Because of the ease with which exact registration is possible using active light sources or specially prepared markers, the registration problem may be solved at the video signal level, resulting in an environment where the real and virtual worlds fuse together as one.

The approach presented here might also be useful within the context of work done by other researchers, such as Knowledge-based Augmented Reality for Maintenance Assistance (KARMA) [79][80], seeing architectural anatomy of buildings [100], and combining ultrasound with virtual reality in obstetrics [81].

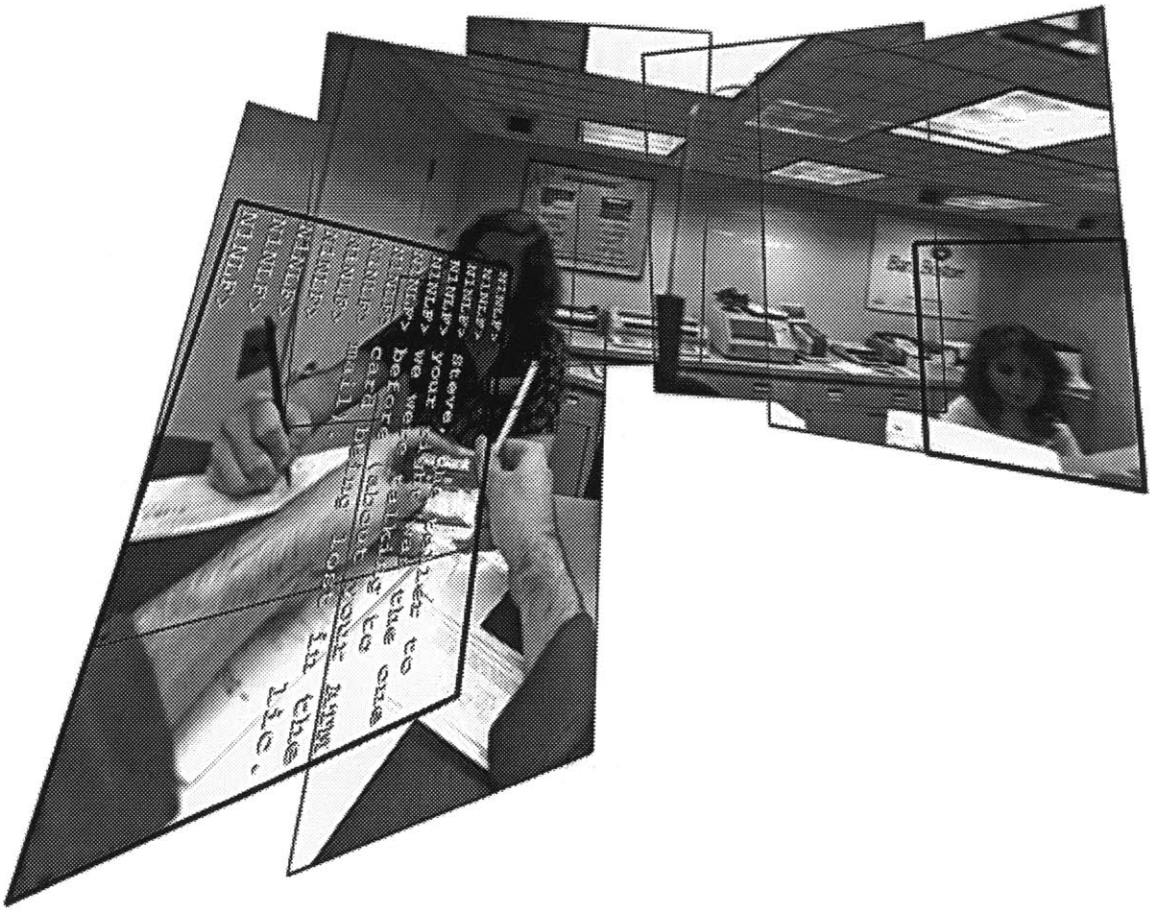


Figure 6-9: Shared environment maps are one obvious application of a computer-networked personal imaging workstation. Images transmitted from my "Wearable Wireless Webcam" may be seamlessly "stitched" together onto a WWW page so that others can see my point of view. However, because the communication is bidirectional, others can send me messages, for example, allowing me to recognize people I've never met before. In addition to simply allowing others to annotate my "reality stream", we might also allow others to alter our perception of reality. Thus personal imaging allows the individual to go beyond a cyranic [3] experience, toward a more symbiotic relation to a networked humanistic intelligence within a mediated reality environment [13]. (C) Steve Mann, 1995. Picture rendered at higher-than-normal screen resolution for use as cover of a journal.

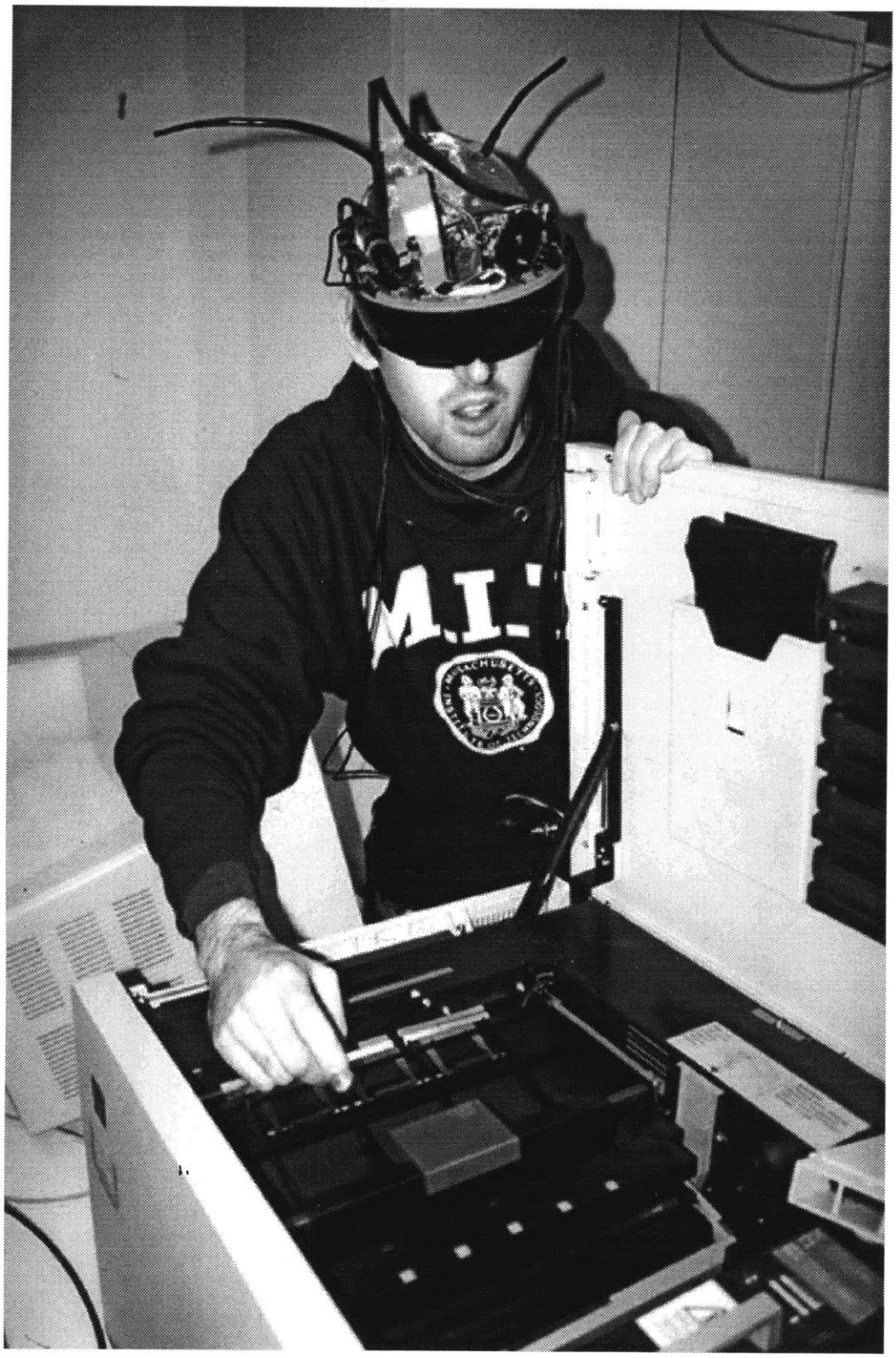
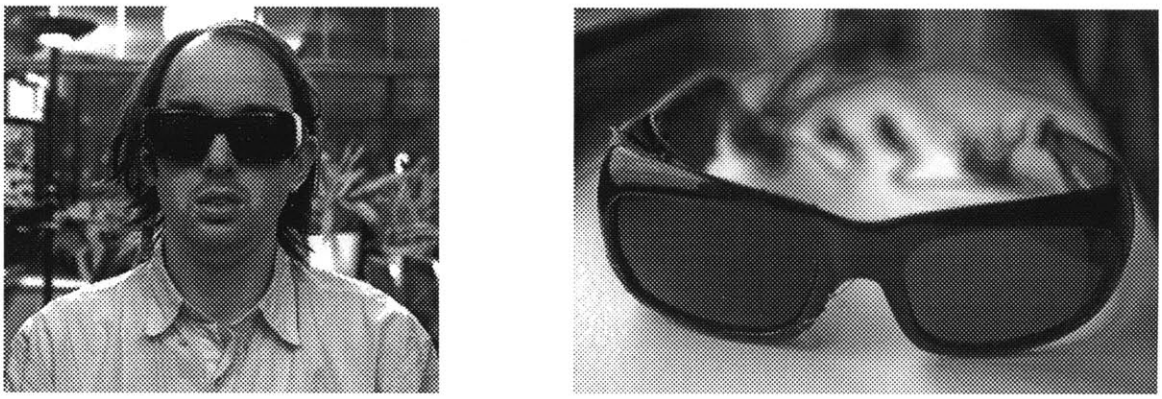


Figure 6-10: Equipment repair (in this case, a laser printer being serviced), in a 'mediated-reality' environment, using apparatus designed and built by author. Note the colored tape installed in the printer and on my fingertip, which is recognized by the camera in my eyeglasses.



(a) (b)

Figure 6-11: WearComp7: Covert embodiment of the WearComp invention. This invention comprises a complete multimedia computer, with cameras, microphones, and earphones, all built into an ordinary pair of sunglasses together with electronics items that either fit into a very large shirt pocket (e.g. for one embodiment using a fullsize computer, a large pocket on the back of a special undershirt was built of re-enforced fabric specially for the computational apparatus) or are distributed around and sewn into a special undergarment (Fig 6-14(b)). Note that long hear was needed to hide the wire going from the glasses to the undergarment, etc.. We are at a pivotal era in which miniaturization of components will soon make all of this apparatus "disappear" into the clothing altogether. The rig pictured here is currently running the Linux 2.0 operating system, with XFree86 (variant of X-windows).

6.7 The covert reality-mediator

One of the problems with early embodiments of WearComp was its obtrusiveness. This made it difficult to use in all facets of life (e.g. especially in places where photography is strictly prohibited or the like, or where there might be criminal activity or others excessively paranoid about such an apparatus, or simply in the sense that its appearance was very untidy).

Most notably, if personal imaging is to be useful in all facets of life, it needs to be based on embodiments of WearComp that do not look unusual. Accordingly, the current state of WearComp (which I call WearComp7) exists as a miniaturized multimedia computer built unobtrusively into a pair of ordinary sunglasses (Fig 6-11), with the remaining components either carried in a small box that fits in the pocket of a specially designed shirt, or spread out over a specially prepared undergarment (Fig 6-14(b)). The current embodiment of WearComp, pictured in Fig 6-12(e), which I refer to as WearComp8, is still under development, and is particularly unobtrusive such that it may be worn in a variety of ordinary situations without seeming like it is out of the ordinary.

In Chapter 7, I will describe my attempt at establishing a new genre of documentary video based on completely covert personal imaging workstations configured to operate as reality mediators.

6.8 Synthetic synesthesia for a sixth or seventh sense

But slow, what light through yonder window breaks.

Tail lights are red, headlamps are blue
Time seems to fly when I spend it with you
Let me have peace, nature serene
Lie still in the grass, the trees ever so green

–(c) Steve Mann, 1987

TimeWarp

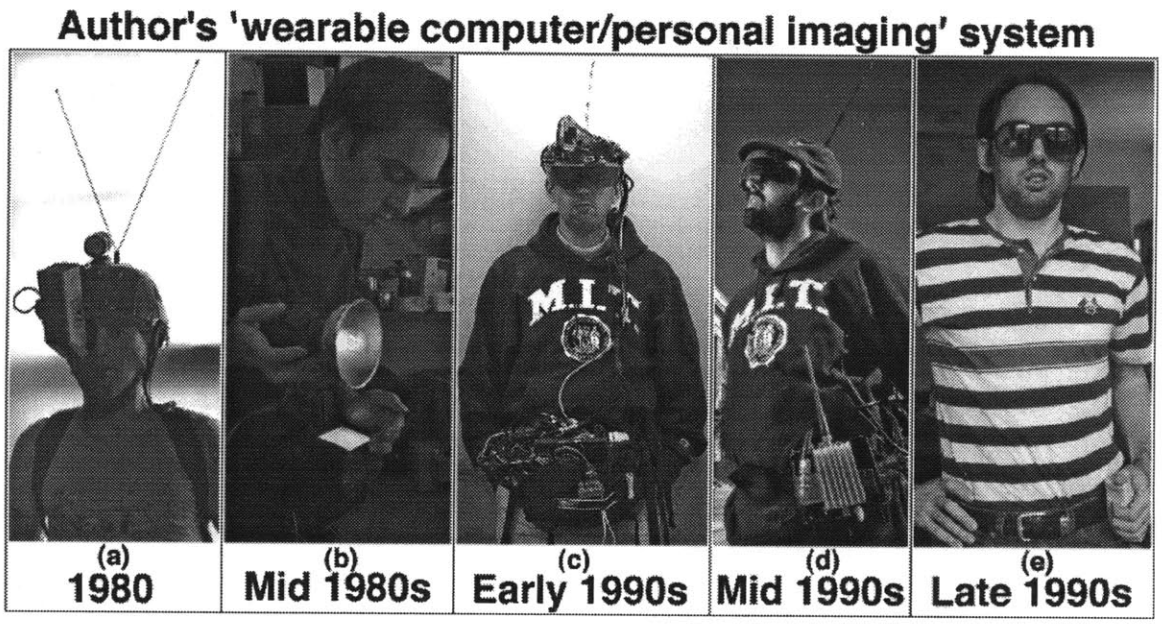


Figure 6-12: Evolution of the 'WearComp' invention. WearComp8, currently still under development, is pictured at far right. Computational apparatus is on a special undergarment (the 'underwearable'), and is arranged mostly on the back of the shirt so that it is not obtrusive in face-to-face interaction. In Chapter 7, we will understand the utility of this unobtrusiveness in the context of a personal investigative documentary.

In just eight hours your thesis is due
I am the clock and I'm watching you

The hurrier you go the aheader I get
But when you wait in the rain you're gonna get wet

'Cause when you're waiting my hands stand still
Time is a concept you just can't kill

–(c) Steve Mann, 1990

Mediated reality may include, in addition to video, an audio reality mediator, or, more generally, a 'perceptual reality mediator'. This generalized mediated perception system may include deliberately induced synesthesia[2]. Examples I have explored include seeing sounds, hearing light, etc., but the most interesting pertain to the addition of a new sense (depending on who's opinion one takes, we already have five or six senses, so the new one would be the sixth or seventh).

The new sense that I have explored most extensively is that of radar. In particular, I developed a number of vibrotactile wearable radar systems in the 1980s, of which there were three primary variations:

- 'CorporealEnvelope': baseband output from the radar system was envelope-detected to provide a vibrotactile sensation which was proportional to the overall energy of the return [3]. This provided the sensation of an extended 'envelope' around the body, in which one could feel objects at a distance. In later (late 1980s) embodiments of 'CorporealEnvelope', envelope

[2]Synesthesia [101][102] is manifest as the crossing of sensory modalities, as, for example, the ability (or as some might call a disability) to taste shapes, see sound, etc..

[3]Strictly speaking the actual quantity measured in early systems was that of a single homodyne channel, which only approximated energy. Later in some systems this was done properly with separate I and Q channels.

detection was done after splitting the signal into three or four separate frequency bands, each driving a separate vibrotactile device, so that each would convey a portion of the Doppler spectrum (e.g. each corresponding to a range of velocities of approach). In another late 1980s embodiment, variously colored lamps were used, attached to the wearer's eyeglasses to provide a visual synesthesia of the radar sense. In one particular embodiment, red, green, and blue lamps were used, such that objects moving toward the wearer illuminated the blue lamp, while objects moving away illuminated the red lamp. Objects not moving relative to the wearer, but located near the wearer appeared green. This work was inspired by using the metaphor of the natural Doppler shift colors one might experience while approaching the speed of light, or equivalently, what one might experience if light were slowed down to the speeds that we encounter in our day-to-day life. (See 'slowlight' quote above.) In a much more recent (1996) version of 'corporeal envelope', I used seven vibrotactile elements, each conveying a portion of the Doppler spectrum, together with one of my early 24.360GHz wearable radars.It is not difficult to imagine a continuum of vibrotactile elements that would convey a continuous Doppler spectrum.

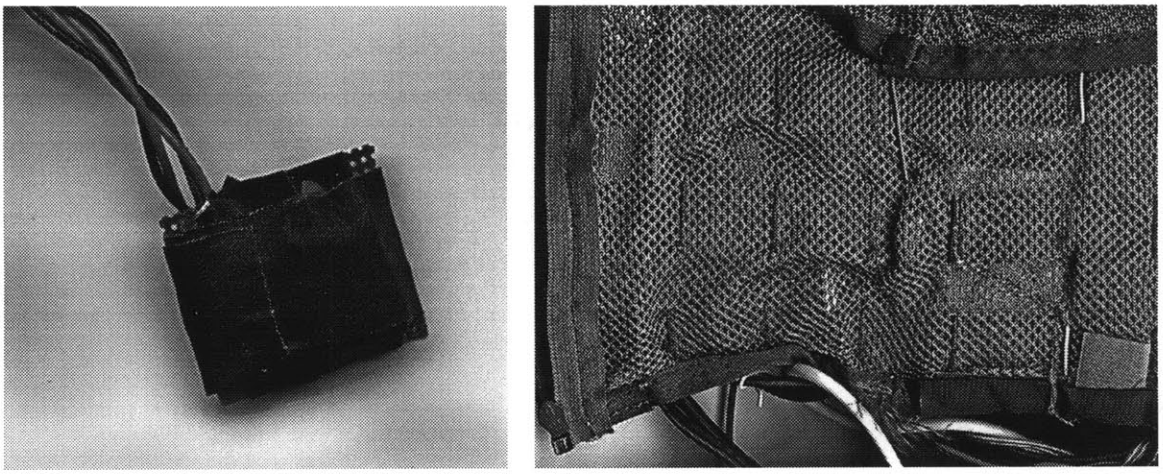
- 'VibroTach' (vibrotactile tachometer): the speed of objects moving toward or away from the wearer was conveyed, but not the magnitude of the Doppler return (e.g. it was not possible to distinguish between objects of small radar cross section and those of large radar cross section). This was done by having a Doppler return drive a motor, so that the faster an object moved toward or away from the wearer, the faster the motor would spin. The first VibroTach was built as an art installation, to drive a clock (see "TimeWarp" quote above). Upon making a wearable version of "TimeWarp", it was noticed that one could "feel" the motion of objects at a distance. Holding onto the clock and moving back and forth created a surreal sensation as though strings were attached to objects in the room, and that these strings were passing through the clock, and hence through my hands. The spinning motor could be felt as a vibration having frequency proportional to that of the dominant Doppler return (a picture, although somewhat surreal on account of the expressive nature in which I captured it, showing the specially modified clock, appears in Fig 6-13). I also explored the use of multiple vibrotactile transducers (typically permanent-magnet loudspeakers) to simulate motion without movement in the same way that a light-chaser simulates a visual motion percept by turning lights on and off in the proper sequence (in fact I used a light-chaser circuit similar to that implemented in WearComp1 described in Chapter 3, but with vibrotactile units instead of lights). The initial reason for using multiple units around the body was to be able to perceive the difference between clockwise (e.g. toward the body) and counter-clockwise (away from the body) motion of radar targets. Early 1980s versions of VibroTach used synchronous AC motors (driven directly from the plates of a backpack-based audio amplifier, e.g., with output transformer removed), while later versions tended to be characterized mostly by the use of DC motors (driven by solid-state devices).

- 'Electric Feel Sensing': the entire doppler signal (not just a single dominant speed or amplitude) was conveyed to my body. Thus if there were two objects approaching me at different speeds, I could discern them separately from a single vibrotactile sensor (e.g. both at the same point on my body). Various embodiments of 'electric feel sensing' included direct electrical stimulation of the body using an automobile spark coil, and later a trigger coil from an electronic flashlamp, as well as the use of a single broadband vibrotactile device. An example of the latter included a backpack-based 6 by 9 inch elliptical loudspeaker mounted in a wooden cabinet, driven by an audio amplifier made from transistors salvaged from a surplus electronic cash register.

One of the problems with this work was the processing, which, in the early 1980s embodiments was done using a wearable analog computer. However, Today's wearable computers, capable of computing the chirplet transform[4] in real time, suggest a fully digital VibraVest (Fig 6-14(b)).

[4]The chirplet transform [55] characterizes the acceleration signature of Doppler returns, so that objects can be prioritized, e.g. those accelerating faster toward the wearer can be given higher priority, predicting eminent collision, etc..

Figure 6-13: Twisted Time: An early embodiment of VibroTach was discovered while handling a clock into which I had built a radar system. I built the clock to tell distance rather than time, for an art installation in the 1980s. (The Doppler spectrum was effectively integrated with a gain that resulted in 1 minute of clock advancement for each 12 inches of distance, thus carrying the clock 60 feet down a long corridor advanced it approximately 1 hour.) Most notably, a surreal experience resulted from handling the clock. The resulting synthetic synesthesia could best be described as a feeling as though there were strings attached to every object in the room, and that all those strings passed through the clockwork, so that in some sense, I could feel the movement of these objects at a distance. In this picture, I have attempted to convey the surreal effect of holding onto this special clock, through the use of an oblique slice in a spatiotemporal volume. (c) Steve Mann, 1989.



(a) (b)

Figure 6-14: VibroTach and VibraVest: (a) early 'vibrotactile tachometer' called 'VibroTach'. As objects approached me more quickly, the motor (with eccentric weight) would spin faster, so that I could feel it vibrate more strongly and violently. A collision at high speed could be averted by the almost painful sensation I experienced just prior to impact. (b) recent embodiment of the vibrotactile vest/wearable computer ('ThinkTank') invention. This tank top contains a complete internet-connected multimedia personal imaging workstation (running Linux 2.0 and XFree86) that can be concealed under an ordinary shirt. The proximity to the skin allows one to easily feel the vibrotactile actuators. This system also forms the basis for the 'underwearable' computer described previously.

A full discussion of all the non-video variations of mediated reality is beyond the scope of this thesis, but the above example is meant simply to indicate that video was not the only modality explored.

6.9 Chapter summary

A new philosophy of human-computer interaction, called "Mediated Reality" (MR) has been proposed. MR differs from Virtual Reality (VR) in the sense that it affords a perception of reality itself, while VR isolates the user from the real world. MR differs from typical Augmented Reality (AR) in the sense that MR allows reality to be not only augmented, but also diminished or otherwise altered, if desired. Another important distinction between the proposed framework (MR) and existing AR frameworks is its tetherless nature. Most notably, I have presented two enabling inventions that make MR tetherless: (1) a means and apparatus for simulating a powerful wearable computer, through the use of a full-duplex radio communications link, and (2) a method of tracking head position based on the use of the photometric measurement capability of the apparatus (e.g. using video cameras). In this sense, the role of photometric capability of the apparatus (e.g. the video cameras) is twofold (1) to take in "reality", which is mediated and then presented to the wearer, and (2) to perform the head-tracking function that is normally done by a head-tracker (hence something that would tether the wearer to a particular location).

A completely unobtrusive version of the reality meditator was presented, which can facilitate personal imaging applications in ordinary day-to-day situations which might otherwise be difficult with the encumbering and unsightly nature of earlier embodiments. Furthermore, it was demonstrated that other forms of perceptual mediation are possible, not just visual.

While a full practical implementation of MR is several years away, current implementations could be useful in specific domains. In particular, in applications where registration is extremely critical yet can be solved at the signal level, or where it is desirable to be able to alter as well as augment reality, MR has been shown to have great promise.

Chapter 7

Personal Imaging as a first step towards a new genre of documentary: personal verité

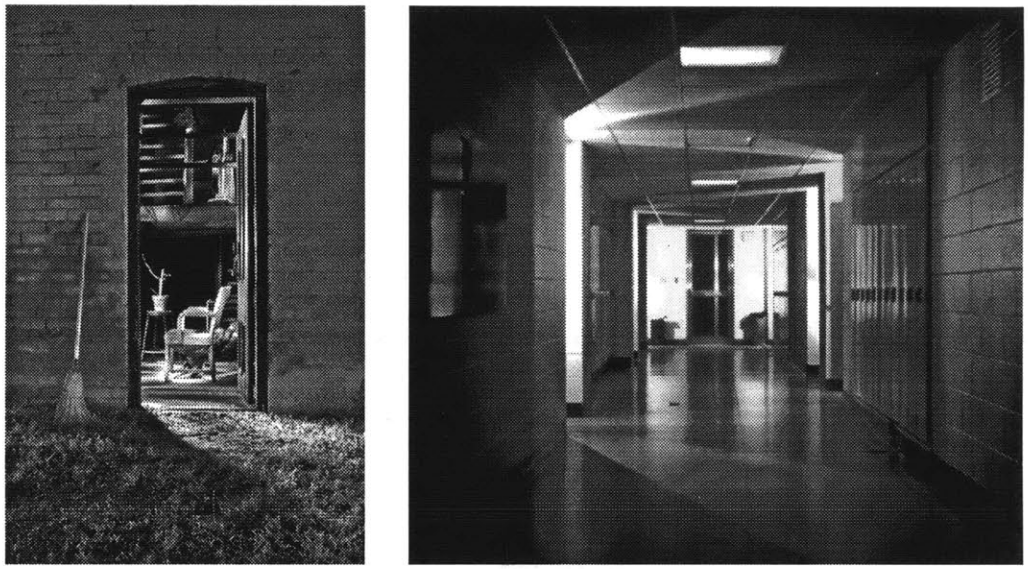
> The technologies for recording events lead to a curious result... Vicarious experience, even for those who were there. In this context "vicarious" means to experience an event through the eyes (or the recording device) of another. Yet here we have the real experiencer and the vicarious experiencer being the same person, except that the real experiencer didn't have the original experience because of all the activity involved in recording it for the latter, vicarious experience... we are so busy manipulating, pointing, adjusting, framing, balancing, and preparing that the event disappears... But there is a positive side to the use of recording devices: situations where the device intensifies the experience. Most of the time this takes place only with less sophisticated artifacts: the sketch pad, the painter's canvas... Those who benefit from these intensifying artifacts are usually artists... with these artifacts, the act of recording forces us to look and experience with more intensity and enjoyment than might otherwise be the case. –Don Norman[103]

7.1 Introduction: Evolution from new photographic genre to new cinematographic genre

The original motivation behind the WearComp project was an attempt to define a new genre of imaging characterized by unprecedented control over lighting, and, in particular, to create a tool that would allow reality to be experienced with greater intensity and enjoyment that might otherwise be the case. We already saw in Chapters 2 and 3, how the ability to control the shape and distribution of light sources allows one to create images with "impossible" lighting; images that could not have been created with a single exposure to any kind of light-producing apparatus were created using homomorphic linearity and superposition.

In some sense, lightspace is a new language of imaging [34]. However, as with any artistic medium, one must "write some poetry" in that language, to make it meaningful; lightspace would not be complete without a brief look at some attempts at creating expressive and meaningful images, using the affordances of the "toolbox" created in Chapter 2. Fig 7-1 depicts two early attempts at creating expressive images using the lightpainting technique.

The early embodiments of the WearComp [104] invention, originally motivated by these applications in the visual arts, were characterized by a heavy and obtrusive nature. However, WearComp has more recently evolved into a completely unobtrusive rig concealed in a pair of ordinary sunglasses and ordinary undergarments (as described in Chapter 6), making possible a new kind of documentary video, which I call 'personal verité' (short for 'personal documentary cinema verité').



(a) (b)

Figure 7-1: Some early attempts at creating expressive and meaningful images, using the affordances of the lighting "toolbox" created in Chapter 2. (a) In addition to having the rich tonal range characteristic of images described in Chapter 3, the image also has an expressive characteristic that could not be obtained in a single image exposure to conventional light sources. Notice how the broom appears to be its own light source (e.g. self-illuminated), while the open doorway appears to contain a light source emanating from within. The rich tonal range and details of the door itself, although only visible at a grazing viewing angle, is indicative of the affordances of lightspace. (b) hallways offer a unique perspective, which can also be illuminated expressively. (C) Steve Mann, sometime in the mid 1980s.

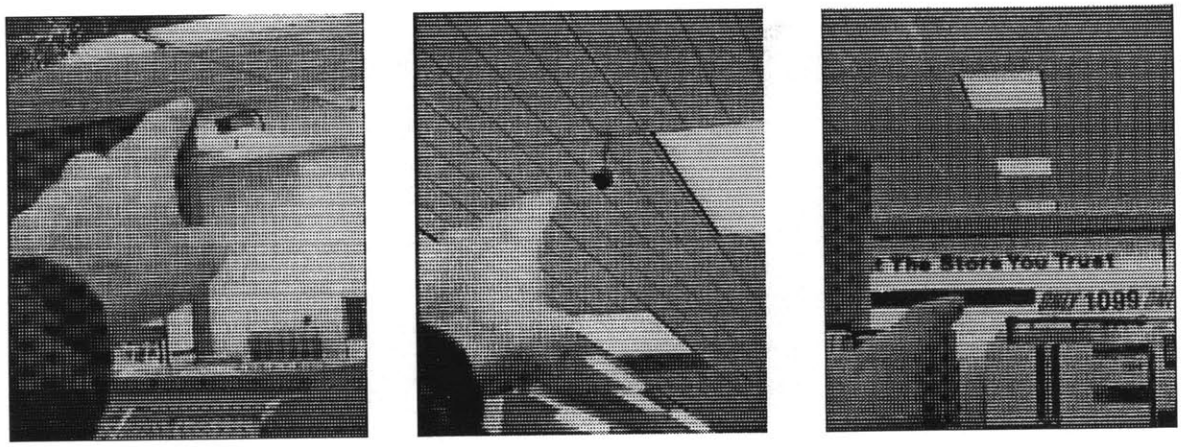
It should be noted that the methodology of 'personal verité' differs from current investigative journalism, in the sense that the long-term adaptation process, as described in Chapter 6 (e.g. often taking place over a period of many years) makes the camera behave as a true extension of the mind and body, and that the 'diminished reality' is exploited fully, in the capturing of a much richer and more truthful perception of reality. An example of the latter is quite evident in my documentaries, when, for example, I am asked to sign a bank withdrawal slip or the like. Because of the greatly diminished reality, I must bring my head very close to the written page (distance depending on the size of the lettering), in order to see it. A side effect of doing so is that I produce video in which the audience can also see the fine print, whereas shooting in a traditional investigative documentary style, this would not be so.

7.1.1 From "fly on the wall" documentary to 'fly in the eye' personal documentary

Clearly, today's digital cameras fail to meet my requirements for hands free operation and realtime processing/ transmission capabilities. Another, perhaps less obvious yet important missing element from many of today's digital cameras is the electronic viewfinder. Again, it is quite obvious to anyone skilled in the photographic arts, that not only should there be a means of previewing previously captured images, but the importance of having the human in the loop of the creative process should be regarded as absolutely essential.

The WearCam viewfinder goes beyond merely setting the camera correctly because there is a human completely in the loop (e.g. I will trip and fall if it is not set correctly, since I will not be able to see properly), so that it transcends being a mere compositional tool, toward allowing the camera to "become" the eye of the wearer.

The nature of WearCam is such that the combination of camera, processor, and display results in a visual adaptation process, causing it to function as a true extension of the body and mind, that is, after an extended period of time, it does not seem to the wearer to be an external entity. Images



(a) (b) (c)

Figure 7-2: Living in a 2-D world, through long-term adaptation. Fingerpointing from the perspective of *life through the screen.* Adaptation was in the "rot90" coordinate transformation described in the text.

captured on WearCam have a certain expressive quality that arises from having the wearer be "in the loop" — part of the complicated feedback process that synergizes both human and machine into a single unit.

The small size of WearComp (negative size if one considers what it replaces — cellphone, pager, walkman, camcorder, Personal Data Assistant (PDA), etc...), provides for a single unified point of contact — all of the personal electronics we normally carry become integrated into this one unit. Clearly the unit is capable of email, "i-phone" (internet telephony), and other forms of multimedia communication, so that it eliminates (or will eliminate, in time, once other people have such units) the need to carry a cellular phone or pager.

The system also has sound capability, so that it can function as a personal sound system, thus subsuming that form of existential media which provides the wearer musical self-determination. However, new forms of interaction — interaction that goes beyond the aggregate functionality of all of the consumer electronics we normally carry — is possible because they are all subsumed into a single unit.

It is this very utility that might make the apparatus a part of ordinary day-to-day living, which would make personal verité truly possible, as a spontaneous and natural form of personal documentary.

7.2 Living in a 2-D world

The deconfigured eye – on becoming a camera

Because of long-term adaptation to the apparatus, experiencing the world through it over a period of many years, I noticed that I began to lose my perception of 3-D depth, (a typical video camera lacks depth from stereo, depth from focus, etc.) and that, in fact, I experienced the world photographically. In this sense, I developed a "photographic mindset" in which I attained an enhanced sense of awareness of light and shade, and of simple renaissance perspective — life through the screen as a long-term modality of thought. I found that this effect persisted, even when I removed the apparatus, and would revisit me in the form of 2-D "flashbacks", so that I began to see the world in two ways, much like we see the Necker cube illusion in two possible ways. This discovery gave rise to the fingerpointing process (Fig 7-2) where I found that I would point at objects as though I were seeing them in 2-D plane projection, and others often told me that I was not really pointing at the object (as they saw it). The notion of attaching a light to my finger arose out of various expressive lightpainting efforts, where the world is imagined as 2-D, while a light source, attached to my finger, is moved around in 3-D space (Fig 7-3).

(a) (b)

Figure 7-3: Examples of tracing out a locus of points in 3-D space that are mapped onto a 2-D image. Here a small light source, attached to my finger, takes the form of a pointing device, which is used to outline objects in 3-D space, but falling upon their 2-D projection. (a) One of my early lightpaintings using this technique. (b) Image which won best color entry, in the National Fuju Film competition, 1986. Here the method is perfected somewhat. (C) Steve Mann, 1985.

7.2.1 Drawing in the air

Video environments like Myron Krueger's and the ALIVE are useful because they recognize the user's gestures. Similarly, the RM can be used to allow a body-worn computer[1] to recognize one's own gestures. For example, I might draw in free space (Fig 7-4), using my finger as a mouse to outline actual objects in the scene. In order to track my finger, I attach a small IR LED so that it will be brighter than anything else in view and then threshold the images to obtain a cluster of pixels that corresponds to my finger (the pointing device), or I attach some colored tape whose color is unique from the background.

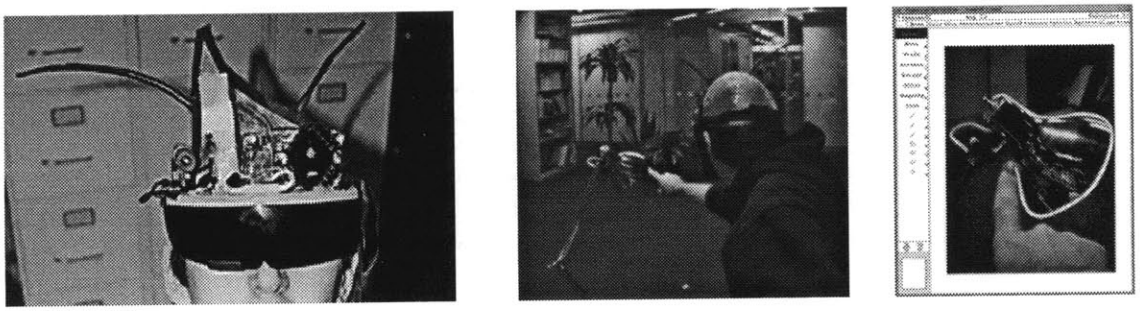
Because I am drawing right on top of the video stream, registration is, for all practical purposes, exact to within the pixel resolution of the devices. In this work, the apparatus used differs from that of Fig 6-3. In particular, there is only one channel instead of two, because the goal is to annotate images with no regard to depth.

7.3 A new cinematographic reality

In 1945, Vannevar Bush described a wearable camera (Fig 7-5) that would record whatever the wearer was looking at onto microfilm [105]. In many ways Bush's proposed camera foretold the so-called point-of-view (POV) cameras, that are used extensively in sports (e.g. mounted in the helmet of a football player), as well as the headcams of David Letterman (e.g. "monkeycam"), or the hidden body-mounted cameras used in TV shows like *60 minutes*, as well as by individuals who are trying to protect themselves from harassment or from false harassment charges.

However, these current body-mounted cameras do not provide exactly the same field of view that the wearer experiences. Bush attempted to address this problem by proposing that a small square in the eyeglass would be used to sight the camera. However this small square is really not in the

[1] The compute power is remote but in a virtual sense is worn on my body via the full-duplex video communications link.



(a) (b) (c)

Figure 7-4: **Drawing in the air:** (a) Monocular reality mediator used in fingertracking. The video signal from the camera is fed to the short antenna (*inbound transmit* channel, operating at microwave frequencies), while the *outbound receive* signal (UHF frequencies) arrives via the longer antenna and is fed to my right-eyed display (modified Virtual Vision system using a high-resolution CRT instead of the original LCD). (b) View of apparatus (note copper mesh cap on my head acting as ground plane for the antennae) and object being outlined (Luxo lamp). (c) What I see through the glasses: by moving my finger around in the space between the camera (taking the role of my eye) and the object, I generate the outline of the object, which I see in red, as it is being generated, while the rest of the scene is displayed in grey (although in this black and white figure we do not see the color of the outline). The rest of the scene remains gray so that distinct colors may be reserved for virtual objects such as the line I am drawing.

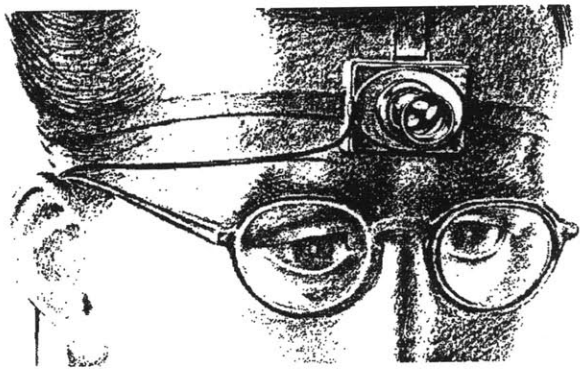


Figure 7-5: Vannevar Bush's camera: original caption reads "A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJECT"

spirit of MR, and so does not close the loop through the wearer. In addition to imprecise sighting as the glasses slide around, if the exposure were wrong, the wearer would not be immediately aware. Hence there is a need to depend on an automatic gain control which is seldom as good as having human in the loop. Using the RM, however, puts the wearer in the loop because it acts as *both* a recording and a seeing device — imperfect adjustment of the camera unfavorably mediates the wearer's vision in a way that causes him or her to adjust it toward a more optimal setting.

Recording the output of my reality mediator gives rise to an interesting method of cinematography, similar in some ways to the cameras mentioned above, but also quite distinct in other ways. The obvious difference is that if my camera is set wrong, I will not see properly, and will trip and fall.

Once I have worn my RM for some time, and become fully accustomed to experiencing the world through it, there is a certain synergy between me and the machine that is not experienced with a head-mounted camera alone. More subtle differences between a recording made from the output of an RM and that made from a conventional body worn camera include the way that when I am talking to two people the closing of the loop forces me to turn my head back and forth as I talk to one person, and then to the other. This need arises from my limited peripheral vision.

After wearing the apparatus for some time, I learn how to compensate for deficiencies such as limited peripheral vision and limited dynamic range. The resulting video [11] is much more like having extracted a signal from my eye than the signal arising from the traditional point-of-view methods, because I live with the RM over an extended period of time and, after time, it begins to behave as though it truly were an extension of my own body. What others see is exactly what I see, no more or no less. In some sense I've become what Manfred Clynes terms a *cyborg*.

7.4 Painting with looks: Creative/expressive applications of personal video imaging

The new featureless algorithm for estimating the projective coordinate transformations between images, presented in Chapters 4 and 5, allows images to be composited together over a wide range of viewing angles, as arises from a personal documentary when one sweeps out a particular gaze pattern. Such images have been used expressively, where the irregular-shaped boundaries (Fig 7-6) capture the gaze pattern of the wearer.

'Painting with looks' provides a user interface which is even more natural than the "point and click" user interface of modern cameras. Furthermore, 'painting with looks' affords the user total control of the process, and makes the process of capturing an image more engaging and fulfilling.

7.5 Personal documentary: 'ShootingBack'

The final creative application of WearCam is in personal documentary. Wearing the apparatus in day-to-day life has resulted in the ability to spontaneously capture events of interest, but in a more natural way than in other forms of personal documentary.

One such personal documentary, "ShootingBack", was a meta-documentary of sorts — it was a documentary about making a documentary. In "ShootingBack" the author carries a standard camcorder into organizations characterized by totalitarian surveillance (e.g. department stores where photography is prohibited yet surveillance is used extensively). Typically a conversation with the representative of the totalitarian surveillance organization, where the representative states (in response to a query as to why surveillance cameras are being used) something to the effect that only criminals would be afraid of cameras, or that the author must be trying to steal something if he were afraid of cameras, or that cameras should be of no concern. Then, pulling the camcorder out of a satchel, the viewer sees the eyecup of the camcorder moving up toward the "eye" (which is actually the eyeglass-based camera), and the resulting response from a representative (Fig 7-7). One might rightly ask why I carry a camcorder with me, because in some sense, *I am a camera*. However, the purpose of the camcorder is twofold: (1) to allow me to record the experience of a documen-

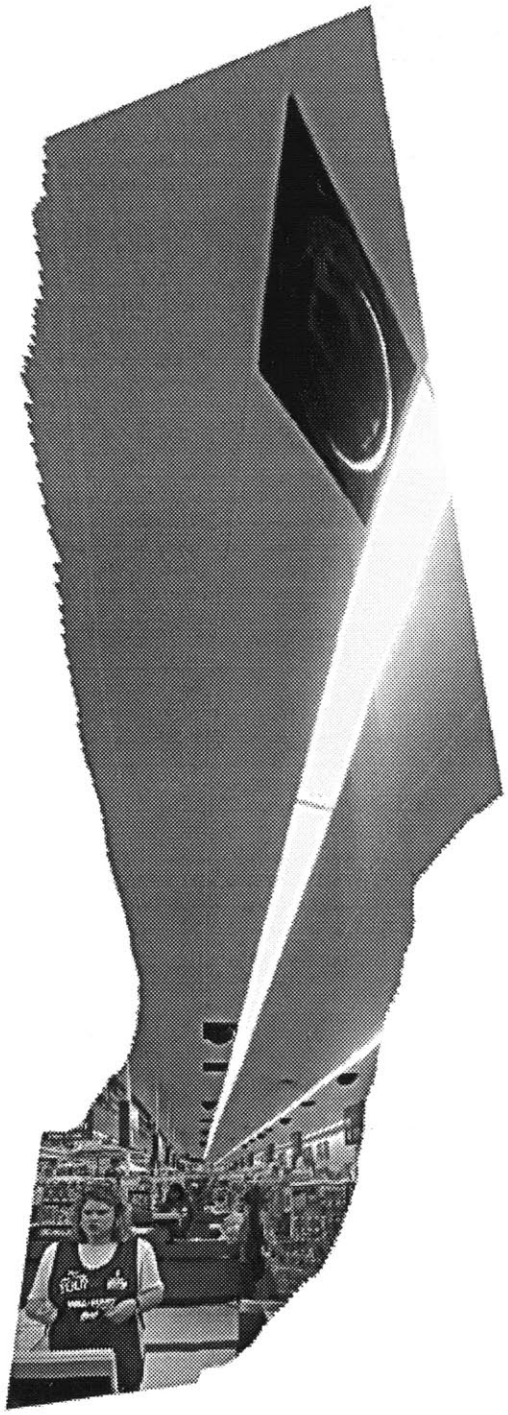


Figure 7-6: 'Painting with looks' produces expressive images by combining WearComp/ WearCam (as described in Chapters 1 and 6) and 'pencigraphic image compositing' (as described in Chapters 4 and 5). The irregularly shaped boundary conveys a sense of the gaze pattern of the wearer. Since WearComp/ WearCam functions as a true extension of the wearer's mind and body, the images often capture, quite accurately, objects of interest in the scene (e.g. highest resolution happens where wearer focuses maximal attention). Here 117 pictures have been seamlessly "stitched" together, to create an extreme wide-angle yet rectilinear sense of strong renaissance perspective. The use of extreme distortionless perspective (not possible within the realm of conventional photography), creates a sense that the mysterious ceiling domes of wine-dark opacity are looming overhead.

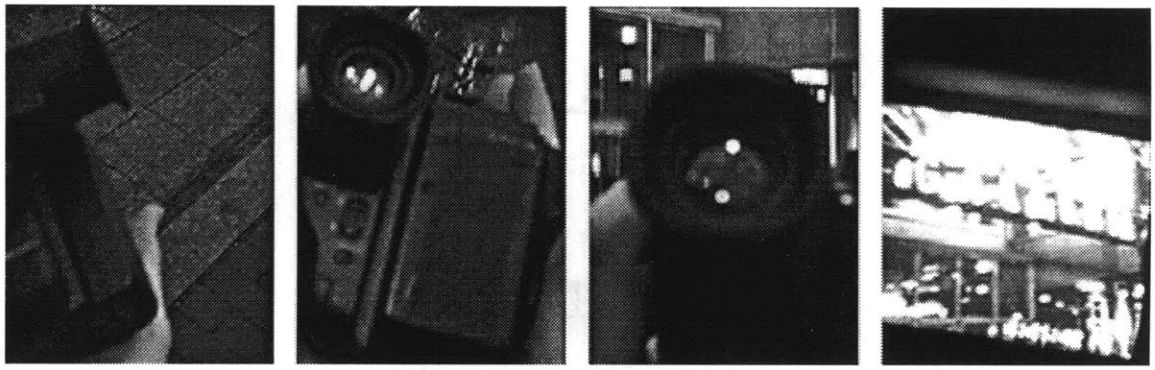


Figure 7-7: 'ShootingBack' is a meta-documentary (a documentary about making a documentary). Here the viewer sees the process of shooting with a conventional camcorder, as recorded by WearCam. Because, after time, the wearer forgets that he is wearing the WearCam apparatus, the process of using the camcorder is captured in a natural manner giving the viewer the impression that he/she is living the life of the documentary artist.

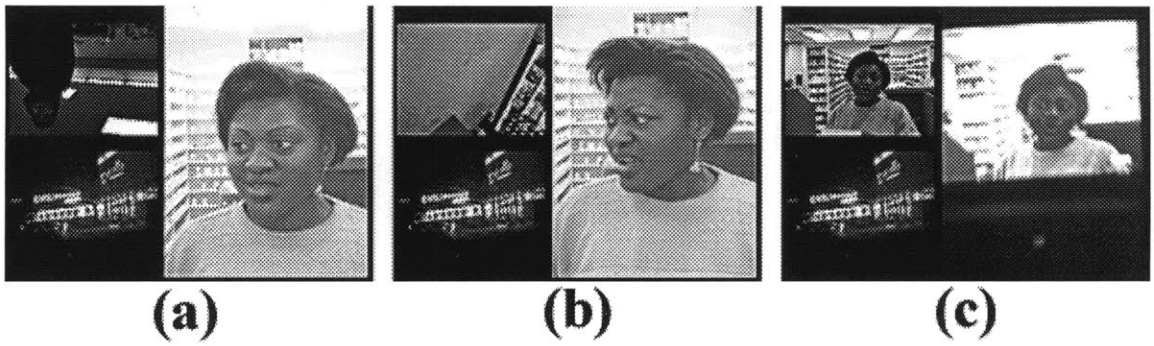
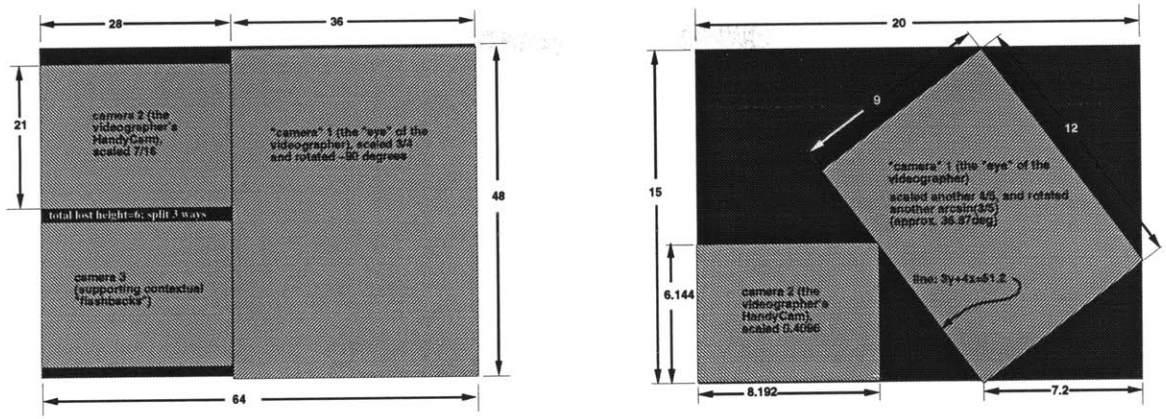


Figure 7-8: Split-screen format shows both the view from the viewer's "eye", as well as the output of the camcorder. Here a representative of an organization using totalitarian video surveillance responds to the author's camcorder. The offset between the representative's gaze and the center of projection is owing to a defect in the apparatus. In order to make it completely covert, yet able to see into the viewfinder of an ordinary camera, many of the design constraints made it difficult to minimize offset. This deficiency has been fixed in WearComp8.

tary videomaker from the perspective of the participant, and (2) as a confrontational entity. In this latter sense, although the main recording apparatus (WearComp) was completely unobtrusive, the second camera (camcorder) served to provide a high level of involvement in the activities of the subjects. In this sense, ShootingBack was much more like the original cinema Verité of French anthropologist Jean Rouch and sociologist Edgar Morin than the "direct cinema" variant of Robert Drew and Richard Leacock which stressed a low level of filmmaker involvement in the activities of the subjects[106].

Because there is also the output from the camcorder, in 'ShootingBack', a split-screen effect, reminiscent of the movies "The Boston Strangler" (directed by Richard Fleischer, USA 1968) and "The Thomas Crown Affair" (directed by Norman Jewison, USA 1968) is used. However, keeping to the integrity and reality-metaphor of the documentary, the natural window sizes are used at all times (e.g. nothing is cropped as is often done in other uses of split-screen). (See Fig 7-8) The effect is perhaps more reminiscent of the split-screen spatial division multiplexers used for "4-up" recordings of surveillance cameras than it is of other cinematography.

Most of ShootingBack is shot in "rot90" (90 degree rotated) format, as this brings the viewer in closest proximity to the person being shot. The aspect ratio of the human face fits well this 3:4 format (as opposed to the usual 4:3 ratio of TV). Two formats are used in 'ShootingBack' (Fig 7-9), the alternative format (Fig 7-9(b)) being used for full-body shots, where the length of someone standing up is situated diagonally across the frame.



(a) (b)

Figure 7-9: Approximate dimensions for screen layouts used in 'ShootingBack'. Units are arbitrary (selected to express ratios using integer quantities). (a) "Eye" view is rotated 90 degrees for "portrait" orientation, while standard camcorder fits in as a window of reduced size. This leaves space for one more window of reduced size which is filled with "flashbacks" (previously recorded/remembered imagery). (b) Diagonal format for full-body shots.

7.6 Chapter summary

A new genre of documentary video, based on long-term adaptation to a personal imaging system, has been proposed. This genre, called personal verité, was found to be successful in documenting personal day-to-day life, in such a way that after a number of years of wearing the apparatus, (on and off over the years, often for a couple of weeks at a time) the camera began to function as a true extension of my own mind and body. Furthermore, it gave rise to changes in my behaviour which caused this synergy to reach new heights. For example, I adapted into seeing the world in 2-D, and this in turn led to other discoveries, such as the use of the finger as a painting and pointing entity for cursor control and expressive/artistic purposes. This "life through the screen" framework builds on the theory and practice of Chapter 6, and points toward a new 'photographic mindset' which is of great use in the visual photographic and cinematographic arts. Finally, the covert variation of the reality mediator described in Chapter 6 was used to shoot a "meta documentary", that is, a documentary about making a documentary about video surveillance. This documentary demonstrated the new method of shooting, and was successful where other cinematographic efforts (Lady in the Lake, etc.) have failed, even though ShootingBack was a much more rigorous test of this methodology, in the sense that it required I operate an optical instrument naturally, in a sometimes hostile environment. Its success can be attributed to this very adaptation process, for the camera has literally become my eye. ShootingBack also explored some new multiple window shooting formats quite different from other previous work by others such as The Thomas Crown Affair or The Boston Strangler.

Excerpts from the ShootingBack project, which combined personal video documentary with 'painting with looks', may be viewed at http://wearcam.org/shootingback.html

Chapter 8

A humanistic intelligence manifesto: Striking a balance with excessive environmental intelligence

> Those who desire to give up Freedom in order to gain Security, will not have, nor do they deserve, either one.— Thomas Jefferson
>
> There is no place for the privacy factor when public safety is concerned [107]. — John Fitzgerald, Supervisor, Transportation Operations, U.S. Postal Service, New York.

A full understanding of personal imaging would not be complete without an understanding of some of the humanistic, philosophical, and artistic[1] motivations behind it. In particular, the WearComp invention, upon which WearCam is based, has a broad philosophical motivation behind it, which I refer to as 'humanistic intelligence'.

The purpose of this chapter is to present a humanistic manifesto of sorts, from an idealist, e.g., "what *should* be", and pragmatic, e.g., with proposed solutions[2] perspective, while Chapter 9 will present personal imaging more in the critical/interrogative tradition of the "Fine Arts", raising questions and awareness through performance, without necessarily offering solutions, scientific analysis, or interpretations of results.

In particular, although we can all recognize the utilitarian benefits of environmental intelligence, e.g., reduced crime, public safety, reduced losses due to shoplifting, lower prices because of reduced losses, as well as convenience in the case of environmentally intelligent environments and the like, there are some problematic aspects that I will focus on in this chapter.

8.1 A problem statement: tangible and intangible aspects

Accordingly, I first identify two problematic aspects of excessive environmental intelligence:

1. The first aspect is the immediate tangible risk that relates to a potential for direct physical harm to individuals, or direct harm to society as a whole, e.g., the risk that rapid proliferation of environmental intelligence might pave the way for a totalitarian regime to systematically

[1] Here I mean "art" in the interrogative and critical tradition of "fine-arts", as opposed to commercial art, or art that is meant to entertain.

[2] Rather than confront the problem head-on, some solutions will be proposed in the form of a shifting of the problem space.

suppress opposition. An example of this first aspect is where cameras purportedly installed for monitoring traffic have been used to round up, detain, and execute peaceful student demonstrators, for example, as happened in China's Tiananmen square in 1989. This first aspect pertains to an imbalance of power that threatens basic principles of liberty, freedom, and democracy.

2. The second aspect is more subtle and more difficult to appreciate. This second aspect relates to a principle I call 'self-ownership', based on the principle of self-determination and mastery over one's own destiny, and asserts that the individual, by default, should own his/her personal information. This personal information may be broken down into two sub-categories:

   (a) Intellectual property: that which we, through deliberate effort, create. Such works are currently protected by patents, copyrights, trademarks, etc..

   (b) 'Humanistic property': that which we produce as a matter of our own existence in, and interaction with, the world. 'Humanistic property' includes our physical likeness (facial appearance, fingerprints, etc.) as well as information determined from our actions, e.g., a list of the books we've checked out of the library, our telephone conversations, the number of condoms we've purchased in the past year, etc..

As an extreme analogy, e.g., to lengthen the "problem vector" in order to see which way it is pointing, consider the following: The difference between the effects arising from these two problems with excessive environmental intelligence is somewhat like the difference between murder and rape. While the first involves obvious tangible physical damage, such as that causing the termination of someone's life, the damage in the second is often less measurable and instead more psychological, that is, even if there is no physical damage, the crime is still dispicable.

A comparison of these two aspects, and the two subsets of the second aspect, is presented in Table 8.1, along with, in the third column of the table, the proposed solutions that will be discussed later in this chapter, and in the fourth column of the table, is the proposed cultural criticism — a collection of interrogative art "solutions" — that will be discussed in Chapter 9.

8.1.1 Humanistic property versus intellectual property

The thesis of this section is that humanistic property deserves at least as much protection, if not more protection, than intellectual property. I will begin by discussing the second of the two major problems associated with environmental intelligence, that of 'self ownership', and then, in Section 8.1.3, I will return to the first aspect.

In particular, self ownership emphasizes the value of 'humanistic property' — that which we "give off", as opposed to only protecting that which we "give"[3]. Specifically, intellectual property — the fruits of our labour (our "blood and sweat" so to speak) is very well protected in our society, while humanistic property — our "heart and soul" lacks such protection.

Humanistic property is, by its very nature of being the product of unintentional creation, generally not put forth as a commodity for sale. One who steals humanistic property, steals that which was not for sale at any price.

It is useful to distinguish between the two through another simple analogy. The analogy is between a shoplifter/softlifter (analogous to the copyright violator, which I will call the "Pirate") and a thief who steals someone's family heirlooms (analogous to someone who steals humanistic property, which I will call the "Rapist"[4]). The former has stolen something that has willingly been put up for sale, merely depriving the victim of some profits, while the latter has stolen something that the victim might not have been willing to sell at any price.

[3] Goffman [108] makes the distinction between deliberate and undeliberate actions of the individual in a social setting

[4] Again, this word is very extreme, but so is "Pirate", and we accept it quite readily in our day-to-day language, in the context of inappropriate copying of software, music, or other works of art.

Taxonomy of problems of excessive environmental intelligence and solutions			
Problem	**Extreme analogy** (lengthening the "problem vector" so we can see which way it is "pointing").	**Practical solution**	**Art "solution"** (Philosophy of cultural criticism of the problem)
Imbalance (lack of freedom, democracy, etc. as a result of the strong possibility for corruption that results from localization of omniscient power)	**Tangible damage** (e.g. physical injury or murder)	**Balance**: accountability forall	**Diffusionism**: shifting the problem axes through cultural engineering to make people accept ubiquitous cameras ∀
'Self ownership' loss	**Intangible damage**	**Protection mechanisms for 'self ownership'**	**Reflectionism**: directly confronting the problem through cultural engineering to make people oppose ubiquitous surveillance cameras.
Intellectual property loss	Financial damage without necessarily causing physical injury (e.g. use of a willing prostitute's service but then refusing to pay for it)	Patents, copyrights, trademarks, trade-secrets.	"Softwear license agreement" and other performances to be described in Chapter 9.
Humanistic property loss	Psychological damage without necessarily causing physical injury (e.g. as is caused by rape)	New laws, new forms of nontransparent clothing. Humanistic media as protective elements.	"My Manager" and other performances described in Chapter 9.

Table 8.1: The two problems of excessive environmental intelligence gathering infrastructure. The first, more obvious one receives a great deal of attention. The second, more subtle one, which itself is broken into two parts reveals that only one of these two parts has received the attention it deserves.

Relating this analogy to the earlier analogy, theft of intellectual property is like enlisting the services of a willing prostitute but then later refusing to pay, while theft of humanistic property is analogous to rape, where the service was not offered for sale to begin with. See Table 8.1.

The need for protecting intellectual property arose out of technological advancements making it possible to easily reproduce works of art [109], literary works, (e.g. through the invention of the printing press), and other forms of intellectual property.

Part of the reason humanistic property has been overlooked is that the technological threats against it are much more recent than those that created a need for intellectual property laws and the like. It is the recent proliferation of information gathering and distribution systems that has made it possible to steal humanistic property, and to profit directly or otherwise attain indirect benefit from this theft[5].

8.1.2 Threats to humanistic property

We are all no doubt aware of the vast quantity of surveillance cameras used to reduce crime, and ensure public safety, etc.. Phil Patton [110] discusses the surveillance dilemma, making reference to the ubiquitous "ceiling domes of wine-dark opacity", making mention that "many department stores use hidden cameras behind one-way mirrors in fitting rooms", and in general, that there is much more video surveillance than we might at first think. Much has been written about this topic, so I will concentrate, instead, on some of the more recent and more serious threats to humanistic property arising from the security and intelligence communities working closely with researchers in new forms of human-computer interaction — areas that have been largely overlooked from the context of privacy concerns. Some of this surveillance infrastructure could certainly capture a part of you that you did not intend to make available to others.

Cameras in private places

Recently, researchers have proposed the use of cameras in bedrooms, or bathrooms, that would watch the occupant and "try to be helpful" (for example they would "turn on the coffeemaker" automatically when they "saw" that the occupant was getting out of bed or finishing up in the shower [111]. The use of cameras in private spaces is not new (e.g. Holy Cross Hospital's installation of hidden cameras in the women's change rooms, or Sheraton's use of hidden cameras in employee locker rooms [112]), but it is, perhaps, more disturbing that they be introduced with a particular function or purpose, as this would provide them with a "justification" for existance in these places. In many cases proponents of surveillance justify cameras on the basis of theft (or suspected employee drug use, as in the Sheraton case), but as new technology evolves, it should not evolve without critical thought and discussion.

Cameras for monitoring children's bedrooms are frequently sold (the video equivalent of the audio "crib monitors" of earlier technology). Recently, researchers proposed an experimental kid's bedroom called the "kidsroom": "The room is also equipped with a microphone...and hidden video cameras" [113], used as an interactive space. However, much of this technology is developed without open debate as to whether or not it really increases the quality of life, or what long-term psychological effect it might have on children.

Environmental intelligence

A new emerging problem, that of so-called "environmental intelligence", has yet to be addressed. In the future, as more and more spaces are equipped with intelligence, we should call some of these

[5] I am taking the liberty of using these strongly judgemental words such as "steal" and "theft". Such strong wording, however, is already present in the context of intellectual property. We even readily accept terms like "software piracy" which make an analogy between someone who copies a floppy disk, and someone who seizes control of ocean-going vessels, often killing all those on board. It was not uncommon for pirates to pour acid on the hands and faces of their victims and tie them to liferafts so the bodies would float ashore as a warning to others not to tread on their turf. Thus an analogy between such gruesome mass-murder, and copying software ought to raise certain questions about our social value system.

(a) (b) (c) (d)

Figure 8-1: Examples of environmental intelligence. (a) "Intelligent highways" and ubiquitous surveillance. Systems like this are used for both traffic monitoring, and "public safety". In Baltimore, for example, the government is installing approximately 200 cameras throughout the city as an experiment to keep watch over the general activities of its citizenry. Often these systems are totalitarian in the sense that no ordinary citizens have the privilege of looking at the video output. (b) Smart ceilings (fifteen ceiling domes or dark windows visible here) monitor people in the space, purportedly for their benefit. Although this may be true to some extent (e.g. gives rise to lower prices, etc.), there is a very subtle and important element (namely humanistic property) that is being overlooked in the rational cost/benefit analyses that lead to decisions to install such systems. Often sophisticated machine vision algorithms are used to track shoppers' activities [114] and make inferences about possible suspicious behaviour. (c) Machines with dark windows monitor users' activities, purportedly for the users' own protection, although organizations are often secretive about the exact nature of these systems (hence the use of very dark glass to hide the apparatus behind it). (d) 'Smart toilets', with dark windows, provide an awareness of the user's state to a miniature computer system or the like contained inside the box with the window, to assist the user in flushing the toilet. Environmental intelligence, such as "electronic plumbing control" is often proposed under the assumption that the user is malicious or incompetent, even in simple everyday tasks. In addition to being "vandal-resistant", electronic plumbing control features provision for covert surveillance (functionality that is generally not subject to inspection by the users, by virtue of its tamperproof nature), thus environmental intelligence used to deter vandalism also gives rise to possible new threats to humanistic property.

into question and open debate. Even today, there is a wide variety of environmental intelligence around us (Fig 8-1). Part of the motivation for environmental intelligence is remote observability (surveillance) and controllability (obedience):

> "Electronic control **improves supervision**: Locating the status panel for a touch-activated system in an office allows instructors to **monitor** shower usage without having to see the shower area... instructor can **control traffic flow** by limiting the timing cycle, **locking** out reactivation or simply deactivating the showers" [115] [emphasis added]

Bradley provides environmental intelligence throughout a restroom or locker room environment through their "Bradlink" network, connected remotely to what Bradley refers to as the "Surveillance Center" [115].

Also the system is closed to inspection/verification by the user:

> "Bradley fixtures with ACCU-ZONE control have no surface mounted parts such as knobs, handles or buttons to tempt **vandals**. Just as important, the **control system is hidden** from view, **securely protected**" [115] [emphasis added]

Some of the most salient aspects of electronic plumbing (supervision, monitoring, locking) make certain pre-conceived assumptions about their users (e.g. "vandals"). Terminology like "control traffic flow" suggests a very utilitarianist perspective.

See-through-clothing security cameras

In modern society, we wear clothing both for self-expression as well as to define a boundary around us, providing both warmth and privacy. One threat to such privacy is cameras that can see through

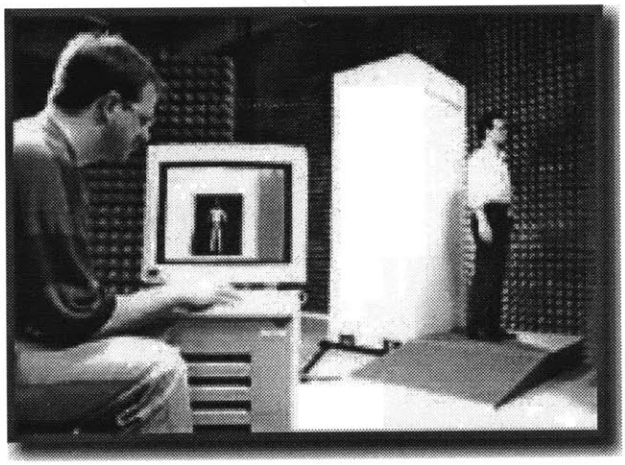


Figure 8-2: "Pacific Northwest's prototype surveillance system is fully functional. A high-speed computer controls holographic data collection and wideband processing and image display." [116] Because clothing is totally transparent in certain frequency ranges such as 10-300GHz, this system produces and records images of the naked human body. Thus its proposed use in security and covert surveillance applications has raised widespread concerns within the privacy community [117].

clothing (Fig 8-2) and present the operator with a picture in which people appear naked.

The assurances provided by proponents of the technology seem weak at best

> Privacy issues are a potential hurdle for the technology. To overcome objections, security personnel envision the images being viewed by same-sex security officers. ... The system could be used to secure mass transit systems, government buildings, public gathering places, and the offices and factories of companies concerned about theft or security. ...

especially given the very secretive nature of most security organizations, who are not likely to put themselves open to review by the general public.

Privacy advocates [117] have raised the question of police pornographers collecting pictures of naked people. Here although no tangible physical damage is likely to arise from this technology, there is something less tangible that is still taken by it, pertaining to human dignity. Combined with the secrecy typical of security organizations, the technology also creates an imbalance between security forces and individuals. Discussion of this imbalance will be the subject of the next section.

Non-transparent clothing (as described in Chapter 1) might provide protection from this technology. Other direct means of protection from theft of humanistic property will be discussed in Section 8.2.

An important question I raise, in the context of this thesis, is that concerning environmental intelligence-gathering infrastructure which is not subject to review by its users (e.g. networked devices, running intellectually encrypted operating systems, sealed in tamper proof housings, watching us through dark windows designed so that we cannot see through the windows to determine what is inside).

Thus, although these systems may "simplify" our daily life, in the context of rational cost/benefit analyses, there is something that they also take away — something much less tangible — that which I call 'humanistic property'.

8.1.3 Threats to balance, symmetry, freedom, and democracy

> The difference between stupidity and genius is that genius has limits. – Mortimer Adler

KOYAANISQATSI
ko.yan.iss.qatsi n. [Hopi, life out of balance] 1. crazy life. 2. life disintegrating. 3. a situation that calls for change.

In addition to its theft of the less tangible 'humanistic property', excessive environmental intelligence also has the more immediate and tangible potential to cause actual physical damage to individuals or to society as a whole.

Although the widespread use of machine vision, in particular, in various environmental intelligence and surveillance applications, remains largely unquestioned, occasional critical studies have been made on this subject. One of the more enlightening of such studies is that by French philosopher Paul Virilio [118]. This work traces its roots from the futurist movement of the early 20th century, detailing its photographic origins, where a connection between eugenics, racial biology, physiognomy, and police photography is made.

The "composite portraiture" invented by Francis Galton, who was also the founder of eugenics (a field of study that influenced the fanaticism in Nazi Germany), was instrumental in the technological development of police photography, physiognomy, and phrenology [119]. These technologies, and the futurist movement itself (which eventually contributed toward the midset leading up to fascism), failed to live up to their promise of a better world.

Many of the reasons behind these social failures can be traced to a purely utilitarianist perspective in which the worth of an individual is determined by how useful that individual is to society, or in milder forms, simply through only looking at rational cost/benefit analyses. An analysis strictly on the basis of rational cost/benefit calculations, which ignores the human element (namely humanistic property as well as long-term balance in society) could pave the way for a "surveillance superhighway" — an infrastructure for ubiquitous surveillance of many people by a few who are in control of it.

Soul theft and likeness piracy

Modern image capture for identification cards is evolving without open debate, toward systems which often entail the keeping of a database of individuals' likenesses on file. This could have a chilling effect for the future, as more and more humanistic property is collected for various uses in central databases. (See Appendix A, **"A self-ownership manifesto pertaining to picture ID cards"** .) In many cases (such as driver's licenses in California and welfare for recipients in New York), this also involves fingerprinting, which for many, calls to mind practices of a totalitarian police state or prison.

Totalitarian video surveillance: the lack of symmetry

By 'totalitarian video surveillance', what is meant is a regime or organization which uses video surveillance, yet prohibits the use of video recording by ordinary citizens (or members, visitors, etc.). Examples include most department stores (e.g. Fig 8-1(b)) where it is typical to forbid customers from keeping their own video record. Many public spaces have also become spaces of totalitarian surveillance at times (e.g. when perpetrators of police brutality or massacres such as in Tianmenen Square forbid citizens from documenting the event, or attempt to seize materials of a documented police beating or the like).

At best, totalitarian surveillance systems create a situation that is out-of-balance; surveillance combined with the secretive nature of most organizations that use surveillance technology:

> One of the fundamental contrasts between free democratic societies and totalitarian systems is that the totalitarian government [or organization] relies on secrecy for the regime but high surveillance and disclosure for all other groups, whereas in the civic culture of liberal democracy, the position is approximately the reverse. [120]

The symmetry axis

Much has been written about the erosion of privacy in our society, and what can be done to stop it, whether through activism, the passing of laws, or by some other direct means. In this section, I put forth a "symmetry" thesis: instead of attempting to solve the privacy problem directly, I propose a shifting of the problem axes — a re-definition of what is meant by privacy.

The symmetry axis arises from a common mis-conception where the small town is equated to the police state (in the sense that you lose "privacy" in both):

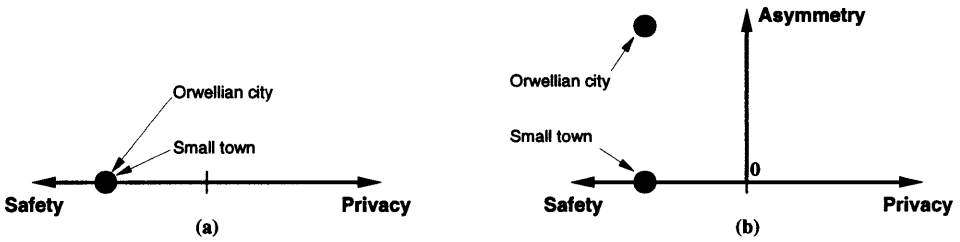


Figure 8-3: The symmetry axis in the privacy argument. Note that without this axis, we might fail to notice a distinction between a friendly small town or a close-knit family and a police state.

> Cameras make the world a smaller place, kind of like a small town. You give up privacy in exchange for safety. In a small town, if you were suffering from a heart attack and collapsed on the floor of your kitchen, chances are better that someone would come to your rescue. Perhaps a neighbour would come over to borrow some sugar, and, since your door would be unlocked, would just come right in and see you had collapsed and come to your aid.

Although this analogy makes perfect logical sense, there is an important missing element, namely the symmetry axis.

On the *safety versus privacy* axis, the small town of the past, and the Orwellian future appear very similar. However, if we look along a different dimension, characterized by symmetry, the small town and the Orwellian future are exact opposites. In a small town, the sheriff knows what everyone's up to, but everyone also knows what the sheriff is up to.

In particular, the privacy problem is multidimensional — it is not just that privacy has been eroded, but it is that the balance of power has shifted away from individuals toward organizations[6]. I emphasize the importance of a new axis in the privacy space, namely that of symmetry (Fig 8-3).

If we consider the two axes of "personal disclosure" (disclosure by individual entities) versus "collective disclosure" (disclosure by government and other large organizations), as depicted in Figure 8-4, we can see that the asymmetry = 0 axis is the line given by:

$$\text{personal disclosure} = \text{collective disclosure} \tag{8.1}$$

and that rotating the entire page 225 deg gives us the symmetry versus privacy axes.

The evolution from small town, to big city, is illustrated in Fig 8-5. Accordingly, an object of this thesis is to propose that the amount of surveillance an individual would be placed under should vary in direct proportion to how much damage that individual can inflict on others or on society. Thus police, armed government officials, and powerful figures, should be placed under high surveillance, while ordinary law-abiding citizens should be placed under less surveillance. Thus if there is going to be a form of surveillance, it should not be controlled by police (unless we are going to live in a police state).

8.2 A proposed solution: Accountability for all

> Power tends to corrupt, absolute power corrupts absolutely. -Lord Acton (1834–1902)

Imagine a future in which we have perfect visual memory, so that when we meet someone, a new

[6] Organizations have always had greater power than individuals, but, owing to computer technology, the balance has shifted even further from individual empowerment. For example, in today's world we are often not talking face-to-face with a decision-maker, but, rather, often just a clerk who is subservient to or pretends to be subservient to a higher and unquestionable authority manifesting itself through a computer network or the like. This phenomenon will be discussed at length in Chapter 9.

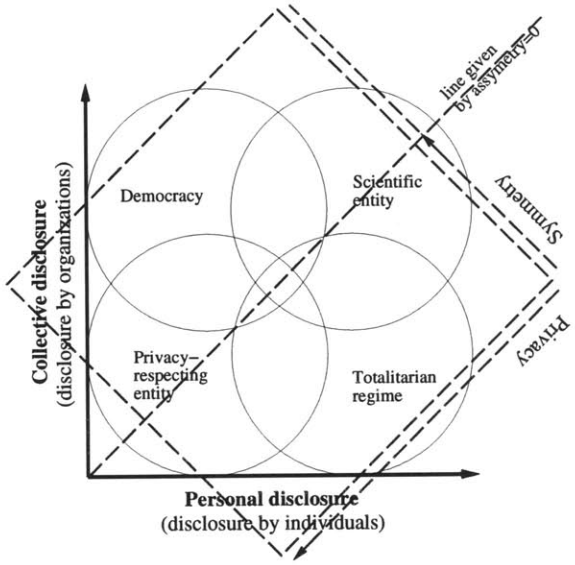


Figure 8-4: Consider the two axes of "personal disclosure" (disclosure by individual entities) versus "collective disclosure" (disclosure by government and other large organizations). Privacy is characterized by nondisclosure (e.g. the origin of the plot). A totalitarian regime is characterized by organiations that rely on secrecy for the regime but high surveillance and disclosure individuals within the regime (lower right), whereas in the civic culture of liberal democracy (upper left), the position is approximately the reverse. There is a long-standing tradition within the scientific community of disclosure for all. Members of the scientific community tend to disclose their findings widely, whether they are part of a large organization or an individual entity. Thus the research results are open to peer review. A complete society based on this scientific principle would be far preferable to a totalitarian society in which large organizations maintain secrets, yet require ordinary citizens to be under close scrutiny. If we rotate the page 225 deg, we obtain the privacy versus symmetry plot. The line personal disclosure = collective disclosure is the privacy axis defined by asymmetry = 0 and is perpendicular to the symmetry axis.

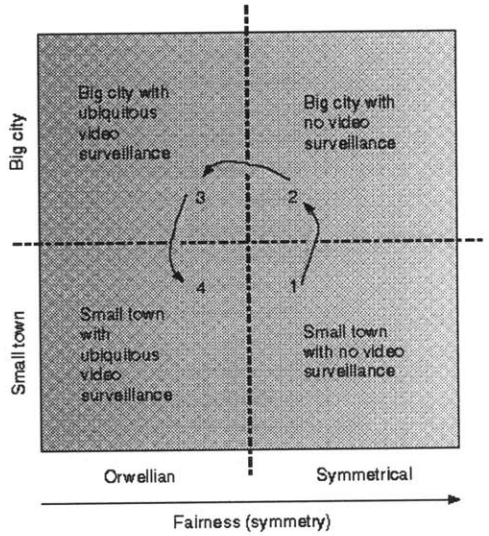


Figure 8-5: The evolution from small town (1) to big city (2), in which people became anonymous and invisible to one another, by virtue of the city's vastness. However, as crime became a problem in big cities, and technology made it possible, video surveillance also became typical of big cities (3). This situation is reminiscent of Bentham's panopticon [121] where we don't see one another on account of the vastness of the city (and because crime is high enough that, out of paranoia, we no longer talk to strangers), yet we are seen by a central "guard" (as in cities such as Liverpool or Baltimore which have a government-owned centralized video intelligence gathering facility where the video is not available for inspection by ordinary citizens). Then as the cost of video surveillance fell, and crime continued to spread from big cities to small towns, video surveillance found its way into the small towns as well (4).

"third hemisphere"[7] of our brain automatically runs a background check on the person and brings up relevant information, such as finger file, WWW page, etc. Such a system would make the big city more like the small town, rather than like Bentham's panopticon. Indeed, this is one of the goals of personal imaging — to create balance, and accountability for all.

What follows is a brief manifesto in support of personal imaging directed, in a humourous spirit, at representatives of totalitarian video surveillance organizations:

8.2.1 Who's afraid of personal imaging

WHO'S AFRAID OF PERSONAL IMAGING (WEARABLE WIRELESS WEBCAM)

So who is afraid of a camera connected to a radio transmitter? True it does reduce our privacy slightly. It's one more camera in a world already full of cameras. Given a choice between hidden cameras in my workplace, and cameras mounted on people in my workplace I'd choose the latter. Both are intrusions into my privacy, but the latter is far less intrusive, and far more symmetrical. With the latter, you use the simple rule: when somebody's looking, you're on camera, when nobody's looking you're not on camera. You can still pick your nose when nobody's looking. In the toilet stall or department store changeroom, nobody else is present so you're not on camera. Privacy equals seclusion. Observation needs company.

If we envision a society in which fixed-cameras of all kinds are prohibited, and only wearable cameras are allowed, and assume, further, that wearable cameras are cheap enough that everyone could afford one, such a society may well be more private than the one in which we are currently living. In fact, if we were all wearing cameras we could certainly reduce crime. Crimes would be solved by cooperation among individuals. In a sense we would be witnesses with augmented visual memory, and augmented visual communications skills. These augmentations would eliminate the need for surveillance cameras, and it would not be necessary to have fixed cameras (hidden or not).

Now it is true that criminals would still try to avoid being seen by anyone, but the fact that everyone had such good "memory" would make it much harder to avoid getting caught.

Some of these privacy issues become clearer when we consider Fig 8-6, a simple taxonomy of cameras.

A shopkeeper who has his fire exits chained shut is afraid of WearCam – such a shopkeeper is afraid of anyone with a good memory for that matter.

A trojan bank (e.g. like the trojan horse) is a bank run by criminals but masquerading as a branch of a major bank. The criminals lease a storefront, sometimes in a shopping mall, under a false name, and "seed" the bank with cash. You or I, delighted that there is a new branch of our bank close by, enter to make a withdrawal, not knowing that the trojan bank reads off our account number, PIN, etc.. Sometime later, the employees of the bank disappear without a trace, along with lots of cash withdrawn from legitimate branches of the same bank, perhaps through legitimate ATMs, by people wearing ski masks so as to avoid ATM securicam scrutiny. Employees working in these trojan banks would be afraid of WearCam.

A criminal wearing a stolen policeman's uniform would be afraid of WearCam. Of course, the law enforcement officers who beat Rodney King would be afraid of WearCam. What good would surveillance recordings of the Rodney King beating be if only police could access them?

People who work in nursing homes and abuse the elderly who live there would be afraid if those elderly wore NetCam. WearCam images sent to friends and relatives are far more

[7] The analogy between the WearComp apparatus and a new, third hemisphere of the brain, is due to Adam Oranchak [122]

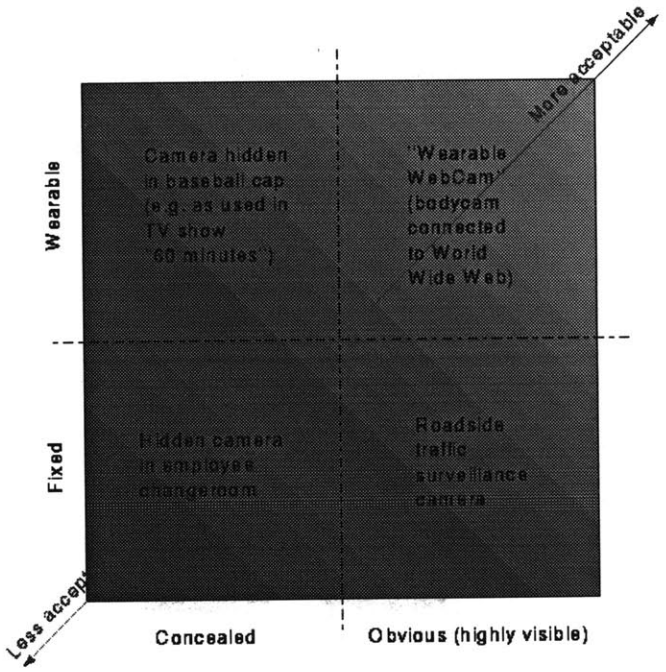


Figure 8-6: A taxonomy comparing early (highly visible) embodiments of the personal imaging system with other cameras. In particular, an argument is made that since privacy is the "quality or state of being apart from company or observation", that if we combine "company" with "observation", we arrive at the notion that we should at least have privacy when we are with out company (i.e. alone). Thus the camera hidden in the environment may be more objectionable than that which is hidden on a person (e.g. more recent versions of the personal imaging apparatus).

useful than images from the establishment's own surveillance cameras. Establishments have been known to "accidentally" lose image recordings, and sometimes "forget" to start the recording apparatus.

Priests who abuse children in the church would be afraid of children who wore NetCam, as would abusive school teachers. Already there is surveillance in many schools. Why not also have some students wearing video cameras? Currently, teachers are using video cameras to catch students throwing tennis balls, spitting on tables in the cafeteria, or sneaking out of classes. With WearCams, an abusive teacher would face similar possibilities of getting caught, and could be punished under law, or by parents filing civil action suits.

People falsely accused of crimes would be less likely to "fall" down jailhouse stairs while awaiting trial if they wore NetCam in jail.

Law enforcement is a tough job. Much of the time police fail to receive the credit that they deserve. These people risk their lives to ensure our safety. However, there is the occasional corrupt police officer. It is the occasional corrupt police officer who would be afraid of WearCam, while the majority of police officers would welcome WearCam as they welcome anyone observing their upright behaviour.

Thus it would seem reasonable to have a law that would make interference with someone's WearCam a criminal offence. It would seem reasonable to not only allow, but to encourage even those suspected of crime to wear WearCam while in jail, awaiting trial if they wanted to. In fact, it should be regarded as a human rights violation to deny anyone the right to self surveillance. After all, why would an organization deny such a person such a right except for purposes of concealing bad treatment or other misconduct such as torture.

THE END OF VIDEO SURVEILLANCE?

The distributed nature of the NetCam augmented memory data would make it less subject to a totalitarian control than video surveillance. Video surveillance will always be upon us. Quite likely, the establishment, with its use of video surveillance, will have the upper hand, for they have the advantage of fixed camera geometry calibrated within the environment, the ability to do motion detection (e.g., when nobody is present, all pixels remain the same), and better communications (hard-wired closed-circuit). However, the ubiquitous use of wearable NetCams will tip the balance a little toward the center, toward symmetry on the Surveillance Superhighway. While the taxi drivers, law enforcement officers, shopkeepers, and government will continue to have surveillance, now the passengers, suspects, shoppers, and citizens will be able to look back at the former on a more balanced and equal footing.

8.2.2 Proposed direct solution to the theft of humanistic property

We have already seen how humanistic intelligence can strike a balance with environmental intelligence, to move us toward a condition of symmetry, and thus address the first of the two problematic aspects of environmental intelligence (e.g., as summarized in Table 8.1). Finally, humanistic intelligence is presented as a solution to the second of these problems, theft of humanistic intelligence. One way of preventing abuse of personal information is to keep it to ourselves, through privacy enhancing technologies like encryption. Of course the ultimate cryptographic machine is one which we never let out of our sight, or allow anyone to use, and which we have complete control over. WearComp comes closest to being a personal computer system that is difficult for others to tamper with or compromise.

Furthermore, humanistic intelligence provides people with additional forms of personal empowerment. In the following, I present two examples of how even a small amount of "intelligence", owned operated and controlled by an individual, may help strike a balance, and reclaim some of the humanistic property that might have otherwise been eroded or lost by modern technology.

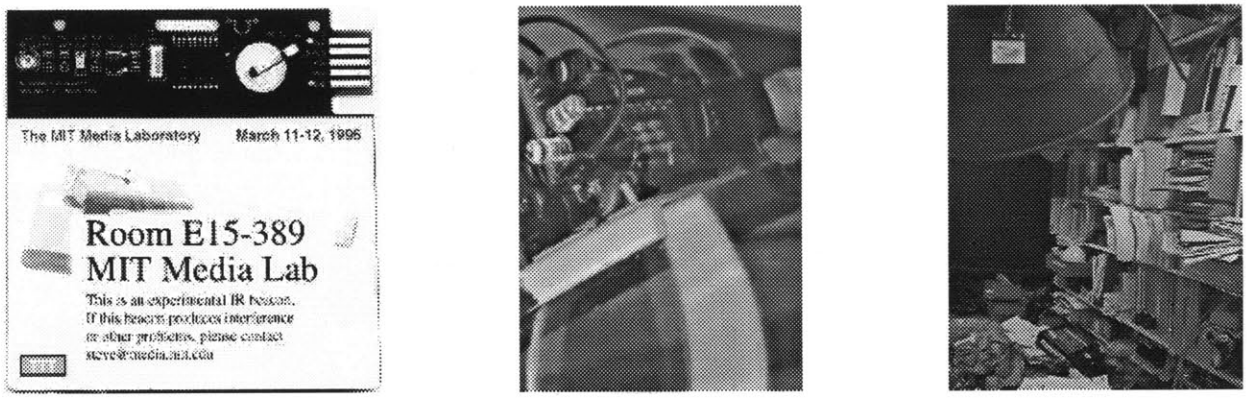
Turning the system inside-out

Many forms of environmental intelligence such as networked access control or location monitoring systems rely on a "smart" element built into the architecture (card reader or IR receiver) and a "dumb" element (card or beacon) carried or worn by the user. The "smart" element is networked to a central computer system, while the "dumb" element has no communications or networking capability whatsoever.

Suppose, however, that we swap the two. Suppose that the user carries or wears the "smart" element, and the building architecture is endowed with the "dumb" element. Thus, for example, the user might wear the infa red (IR) receiver, and have this connected to his/her 'smart clothing', while numerous beacons would be distributed throughout the building. This means that there is no need to network the beacons, no need to wire the building. The system relies on the communications infrastructure each user wears.

However, now the location of the user is known to the user's clothing, and thus the user has control over who can and cannot know his/her location. A user might, for example, define an access control list comprising faculty advisor, thesis advisor, colleagues, etc.. The user's clothing would automatically encrypt the user's location (as determined by the last beacon "seen" by the user's clothing) and transmit this information to the desired recipients. Any interception of this communications would be unintelligible to those not on the access control list.

In a prototype system, the author deployed a number of "room tags" — name tags fixed at known locations in the building (Fig 8-7(a)), and fixed an IR receiver to his glasses (Fig 8-7(b)), and connected this to the clothing-based computer. In addition to giving the user control over his/her personal whereabouts, such a system may also be used to provide location-dependent computer-induced flashbacks (Fig 8-7), adding new dimensions to the visual memory prosthetic. For example, entering my office, when presented with a visual "flashback" (display of image captured from when last I had entered my office) I was surprised to find that the lost sweater I had been looking for that



(a) (b) (c)

Figure 8-7: **A paradigm reversal for person-tracking** (a) 'Room tag' (one of many name tags that was deployed throughout the building). (b) Paradigm reversal: tags and receivers swap places. Receiver is attached to author's eyeglasses and plugged into 'smart clothing'. (c) Computer-induced flashback triggered by revisiting same room as before. Note name tag is visible in upper left of picture. Induced visual 'memory' of sweater, sitting on desk in lower left corner of picture, proved to be helpful when the author entered the office after having lost the sweater (e.g., "remembering" (knowing) when it was last there proved helpful in tracking it down).

day had been sitting on my desk only a short while ago. Thus I 'remembered' that I had worn it that day, and that therefore I must have lost it somewhere nearby.

This form of interaction where the user himself/herself (together with his/her WearComp) becomes the computer network, was first proposed in [15] and later in [123].

The personal database

Our clothing of the future may some day be interoperable and interconnected, so that it keeps track of our physical condition and allows us to decrypt this information for evaluation by a doctor or other professional of our choosing. Further description of the 'smart underwear' prototype, and anecdotes on the author's experience designing, building, and using it appears in [124].

The natural place for our medical records is right in our clothing. Having a patient wear his or her entire medical history would solve much of the medical records privacy problems we face today. With various biosensors, the most current and up-to-date information would be readily available within the very clothing that's taking the measurements[8]. This approach would eliminate the need for, and the possible abuses that can arise with, a central database of medical records, and would eliminate the need for a person to venture through bureaucratic procedures to access his or her own medical information. It would also eliminate the problems associated with smart cards, as clothing is almost always worn, while cards may be misplaced and inaccessible in times of emergency care. Epidemilogical research would still be possible with the patient's data — participating patients could make the data accessible to organizations doing the research, but this would be done through a query to each participating patient's online 'smart clothing' each time the data were needed, so that the patient's clothing would be kept "in the loop", that is, access logs would be automatically generated in the 'smart clothing', so that patients could trace the history/usage of their data at a later date if desired.

In defense of subpoena

People have often felt uneasy entrusting to my "second brain" certain personal pieces of information, fearing that even if I made my best efforts to keep this information in confidence, it could be compromised through a subpoena.

[8] Of course, one would want to have one's medical records replicated (backed up) in the clothing of selected friends and relatives, to prevent data loss in the event of clothing failure.

Even if I were to protect it with ordinary encryption, its security may still be compromised through the process of torture or other forms of duress.

In the same way that our first brain is protected by the Constitution (the Fifth) from contempt of court for refusing to answer (self incriminate), I desired to establish a means of protecting the "second brain" from self-incrimination.

The first solution was to construct it in such a manner that it would self-destruct if forcibly removed from the body, in the same way that "first brain" self destructs if forcibly removed from the body (e.g. if one's head is severed from the rest of the body, the information contained in it is not retrievable, as far as we know within the context of today's technology).

This second brain could also be constructed so that forcibly removing it from the wearer would result in death of the wearer (e.g. so that severing of the second brain would, like the first brain, be considered murder).

My second, alternative solution is to encrypt it with a key that I do not know myself, and then back up the data remotely on a variety of other systems to make it indestructible. To do this, I may encrypt data with public keys of other members of my online community. In this way, the community could recognize when I am under duress, and thus my inability to access my second brain would not be seen in contempt of whatever duress I might be placed under. The underlying principle here, which I call 'subservience empowerment', will be discussed in more detail in Chapter 9.

8.3 Chapter summary

A new principle, called 'humanistic property' was proposed and defended. Furthermore, it was related to existing concepts such as balance (symmetry, freedom, democracy), and intellectual property. (This taxonomy was summarized in Table 8.1.)

Traditional privacy issues were discussed in light of humanistic intelligence, as was the new proposed entity. In particular, it was realized that there is currently a deficiency with respect to mechanisms for the protection of humanistic property, so humanistic intelligence was proposed as a protective medium.

The issue of symmetry was raised, as a dimension of equal importance to privacy. Most notably, it was suggested that a balanced society is one in which the amount of surveillance an individual should be placed under should vary in proportion to how much damage that individual can inflict on others or on society. Accordingly, police, armed government officials, and powerful figures should be placed under high surveillance, while ordinary law-abiding citizens should be placed under low surveillance. It was argued that much of the surveillance we have today is contrary to this principle, and thus takes a first step toward totalitarianism. Indeed, a strong connection between our large cities, and Bentham's Panopticon design for a prison was made, in the sense that we often do not "see" one another in the crowd (and are even afraid to talk to one another — to talk to "strangers"), yet we are potentially seen by the networks of video surveillance cameras.

This video 'surveillance superhighway' contributes not only to the immediate tangible possibility of losingn balance by threatening freedom, liberty, and democracy, but also to the loss of something much less tangible, yet equally important — to the theft of our humanistic property.

'Humanistic intelligence' ("intelligence" inextricablyl intertwined with the individual human) was proposed as a means of striking a balance with the proliferation of excessive environmental intelligence gathering infrastructure.

Chapter 9

Artistic and philosophical considerations: Tactical and interrogative performances based on personal imaging

9.1 Introduction

I will propose 'Reflectionism' and 'Diffusionism' as new philosophical and tactical frameworks for deconstructing the video 'Surveillance Superhighway'.

9.1.1 Problem statement

The recent proliferation of video surveillance cameras, interconnected with high speed computers and central databases is moving us toward a high-speed 'surveillance superhighway', as cameras are used throughout entire cities (such as Liverpool and Baltimore) to monitor citizens in all public areas. As businesses work alongside governments to build this superhighway, and expand it into private areas as well, there is a growing need to develop new methodologies of questioning these practices.

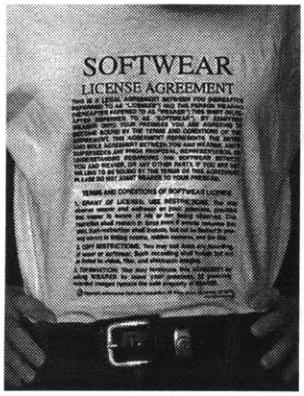
The goal of this chapter is to present a body of work that was created to stimulate inquiry into both surveillance, and the rhetoric used to justify its use.

'Reflectionism' is proposed as a new philosophical and tactical framework which takes the situationist tradition of appropriating the tools of the "oppressor[1]" one step further by also targeting that methodology directly against the oppressor, members and representatives[2] of which become part of the audience of the work+performance.

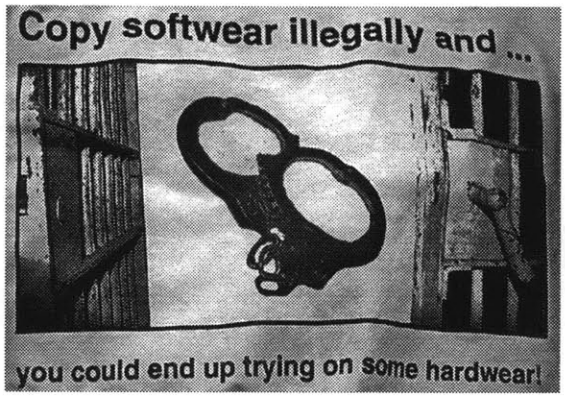
Applied to surveillance, 'reflectionism' attempts to "mirror" the principles of visual surveillance, such as the principle that it often arises from a higher and unquestionable authority.

[1] The term "oppressor", although widely used in the art community might be a little too strong for use in this context. However, we can still think, in the wide sense, of various problems as arising from actions or inactions (apathy) of various individuals.

[2] Again, the term "oppressor" needs to be taken in a wide-sense, to include milder forms of policy creation, or simply the lack of raising objection to certain policies and practices. In particular, the audience of a reflectionist performance may include representatives who are not directly responsible for the unacceptable practices of the large bureaucratic organizations that they are part of. Thus the performance is not meant to be fair and objective, but, rather, is meant to raise social awareness of a more widespread phenomenon. Unlike other social ills, such as the relocation of innocent citizens in Nazi Germany, here there is not one single identifiable "oppressor". However there is still the common diffusion of responsibility where clerks and other individuals absolve themselves of blame by simply declaring that someone higher up their hierarchy, not they, is the "oppressor".



(a)



(b)

(c)

Figure 9-1: Pieces I constructed to assert the principle of self-ownership. Given society's utilitarianist tendency to put our "blood and sweat" ahead of our "heart and soul", these pieces question the apparent notion that copyrighted material (intentional works) should have more protection than people and the data they give off (unintentional works). (a) Parody of the classic "shrink wrap" software license agreement, placing restrictions on those who might photograph the wearer (including restrictions when shirt is removed, e.g. restrictions on use of hidden cameras in fitting rooms). (b) Parody of software piracy poster put on back of T-shirt, including material to which I own the copyright. (c) Parody of a "Say No! to software piracy If you don't own it, DON'T use it!" poster.

9.2 Safe and secure, but at what price?

The perceived "success" (e.g. crime reduction, deterrence, etc.) of video cameras at the bank has led to their use in department stores, first at the cash register and then throughout the store, monitoring the general activities of shoppers. "Success" there has led to governments using ubiquitous surveillance throughout entire cities to monitor the general activities of citizens. In Baltimore, throughout the downtown area, the government is installing 200 cameras as part of an experiment [125], which, if "successful" would mean other cities would also be so-equipped. Businesses such as Sheraton hotel have used hidden video cameras in their employee locker rooms [112], claiming they were installed for the employees' own protection, to keep them from using drugs. The use of hidden cameras by both businesses and governments is increasing dramatically. Again, they do so with claims that the cameras are for the benefit of those under surveillance, which may be true to some extent, but there is another aspect that proponents of surveillance need to be challenged on, and we are naive to accept their "for YOUR protection" rhetoric blindly.

Although most typical uses of video surveillance are associated with claims toward a better future, (e.g., claims that there would be reduction in crime, trains would run on time, or there would be an increase in morale), an object of this chapter is to ask the question "at what price", and to stimulate further inquiry into some of these issues.

Embodied in the work presented in this chapter, is an assertion of 'acquisitional privacy' which challenges the right of organizations to capture/record images of an individual, regardless of what promises are given regarding end use. Tacit in my assertion is the notion of self-ownership[3]. Some self-ownership pieces are illustrated in Fig 9-1.

A further goal of this chapter is to present a body of performance and related cultural criticism aimed at calling into question totalitarian visual surveillance. Recall from Chapter 8 that totalitarian visual surveillance refers to a state of being in which individuals are "seen" by a remote and unobservable entity (be it man or machine) but do not "see" each other through the apparatus, much like Jeremy Bentham's *Panopticon* [121]. Examples of totalitarian video surveillance include department stores where extensive video surveillance is used, yet photography is prohibited. Of all forms of surveillance, totalitarian surveillance is particularly disturbing, as representatives of the video

[3]By self ownership, I mean that the same protections (e.g. copyright) governing the fruits of our labour (that which we intentionally put forth as a commodity) could also be applied to aspects of ourselves, such as our physical appearance, and other information that we generate unintentionally, just through our natural existence.

'surveillance superhighway' refuse to accept the accountability they demand — furthering us toward a Panopticon society in which we're treated more like prisoners than members of a community.

9.3 The five horsemen of the surveillance superhighway

Important to the thesis of this chapter are the following ways in which agents and representatives of the video surveillance superhighway defend their infrastructure:

1. Secrecy: cameras are often hidden in whole or in part (e.g. in dark domes so that we don't see which way they are pointing or even whether or not a camera is present). The security profession is itself also often not subject to open debate or peer-review;

2. Rhetoric: "public safety", "loss prevention", or "For YOUR protection you are being videotaped";

3. Constancy: department store clerks don't follow you around with camcorders, but, rather, video surveillance is present in a "matter of fact" manner, as part of the architecture's *prosthetic territory*;

4. Appeal to a higher and unquestionable authority: "I trust you and know you would never shoplift, but my manager installed the cameras", or "We trust you, but our insurance company requires the cameras";

5. Criminalization of the critic: "Why are you so paranoid; you're not trying to steal something are you?".

I refer to these five defenses as the 'five horsemen of the surveillance superhighway', and this seemingly impenetrable argument formed the inspiration behind the performance pieces presented in the next section. In particular, the material of this next section addresses these five points in a unique way, and attempts to challenge what might otherwise be an undisputable argument. The connection will be made formally in Section 9.5.

9.4 'Reflectionism'

I propose 'reflectionism' as a new philosophical framework for questioning social values. The reflectionist philosophy borrows from the situationist [126][127] movement in art, in particular, an aspect of the situationist movement called détournement[4] in which artists often appropriate tools of an "oppressor", re-situating them in a disturbing and disorienting fashion. 'Reflectionism' attempts to take this tradition one step further, by not only appropriating the tools of the "oppressor" but by also turning those same tools against the oppressor[5] as well. I coined the term 'reflectionism' because of the "mirrorlike" symmetry that is its end goal, and because the goal is also to induce deep thought ("reflection") through the construction of this "mirror". 'Reflectionism' is that which allows society to confront itself or to see its own absurdity.

In applying 'reflectionism' to the surveillance problem, one goal is to allow representatives of the 'surveillance superhighway' to see its absurdity and to confront the reality of what they have done, whether through their direct action, or through their inaction (blind obedience to higher and unquestionable authority).

[4] [Détournement]...is the art of appropriating common objects or images from their usual cultural contexts and resituating them in an incongruous, disturbing, and disorienting fashion in order to confront, question, or challenge society's stereotypes or biases.

[5] As mentioned earlier, this word needs to be taken in its broader and lighter sense.

9.4.1 'WearCam' as tactic for holding a "mirror" up to society

My 'WearComp'/'WearCam' invention described in Chapter 1,6,7,8 (Fig 6-12) formed a basis upon which to build the prosthetic camera, 'WearCam', which, as described in previous chapters of this thesis, was worn rather than carried, and could be operated with both hands free, and while doing other things [11].

In particular, an aspect of this invention which is germane to this chapter, is the manner in which the video recording/transmission functionality of the apparatus appeared as *incidental* rather than *intentional.* By *incidental,* I mean that if I enter an establishment for the purpose of recording video, it is not evident to representatives of that establishment that I had entered the establishment for the purpose of recording video, not just because the apparatus was less visible than a traditional camera, but, more importantly, because the apparatus did not require the appearance of intentionality.

In this way, the apparatus provided a "mirrorlike" symmetry between myself and those placing me under surveillance (e.g. shopkeeper's security guards), in the sense that it was possible to violate the privacy of representatives of an organization placing me under surveillance (e.g. representative of a department store or the like) without violating their solitude (e.g. without an unusual form of interaction as might be the case with a hand-held video camera where intentionality is very obvious). Thus the 'reflectionist' goal of apparent nonselectivity was attained.

In particular, the apparatus provided a means of taking pictures of representatives of establishments placing us under surveillance, in a manner in which they could not determine whether or not such pictures were being taken (just as we never know whether or not a department store surveillance camera is actually capturing an image of us at any given time).

As discussed in previous chapters, a wireless connection to the Internet provides offsite backup of all image data, facilitating another aspect of the reflectionist philosophy, namely to put at least some copies of the pictures beyond the potentially destructive reach of totalitarianist officials through making these pictures available to all (or at least to thousands of people so that it would be virtually impossible to destroy all copies). There is a common belief that art is devalued through reproduction [109], but in the case of many of these performance pieces, the value is increased by that very reproduction, because the value arises from the picture's very indestructibility.

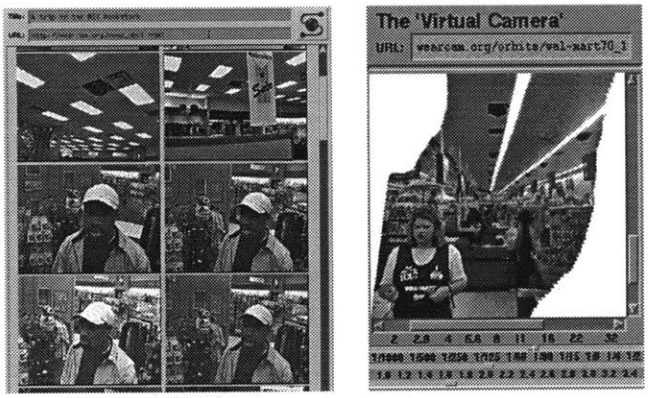
This indestructibility is a reasonably accurate 'reflectionism' of the surveillance superhighway itself, in the sense that just as an individual cannot rob a bank and then destroy the video record (because the video is recorded or backed up offsite, or is otherwise beyond the bank robber's reach), my apparatus of détournement (e.g. Wearable Wireless Webcam [11] or the like) puts the images beyond the reach of members of the establishment, because of the Internet connection, which allows for offsite backup of all images at various sites around the world.

As presented in Chapter 6 and 7, a side-effect of transmitting images to remote locations is the possibility of having multiple processors work together at various remote sites, to enhance the images by regarding each image as a collection of photometric measurements, and combining these measurements together (as described in Chapters 4 and 5) to reduce noise, extend dynamic range and tonal resolution, and increase spatial resolution and extent.

In one such enhancement, images were combined together into a seamless photometric composite (Fig 7-6) which provided a still image as a visual record of my gaze pattern, where the irregularly-shaped image boundary as well as the exceptionally high definition, often in excess of that attainable by photographic means provided a unique form of expressive personal imaging.

More recently, with the advent of the WWW (World Wide Web), the principle of offsite (off-body) backup was further enhanced. Once a collection of images was distributed via the WWW, it was further beyond the destructive capabilities of those (such as department store security guards) who might have attempted to seize or destroy it, as I no longer even knew how many copies of my transmitted pictures might have been made.

Evidence that might, for example, depict that a department store has illegally chained shut its fire exits, is not only beyond their ability to seize or destroy, but is also within easy reach of the fire marshall, who, following my directions via cellular phone from the department store, need only have a standard desktop computer with WWW browser in order to see first-hand what my call pertains to.



(a) (b)

Figure 9-2: Wearable Wireless Webcam (a) Raw unedited feed from camera presented as a sequence of still images read from left to right, top to bottom. Here, at the MIT bookstore, a person claiming to be a representative of the bookstore is informing me that I am not allowed to take pictures in this establishment, but the individual declined to identify himself, and did not appear to be wearing a name tag or any other ID. (However, I believe he was not just a random person trying to pretend to be an official of the establishment because I have been confronted by him on many different occasions during my shooting in this establishment). (b) Virtual camera allows remote viewer to select effective camera direction independent of the direction my gaze, and effective shutter speed, exposure, etc., independent of the actual exposure selection at time of shooting, because photometric image information (as described in Chapter 5, e.g. see Fig 7-6) is captured, rather than merely an ordinary picture.

WearCam on the WWW, thus extends this 'personal safety' infrastructure and further deters representatives of an otherwise totalitarian regime from being abusive in the sense that, in addition to the indestructible evidence of their hostile actions, my friends and relatives are quite likely to be watching, in real time, at any given moment.

This process is a form of "personal documentary" or "personal video diary" as described in Chapter 7, with no editing — 'Wearable Wireless WebCam' challenges the 'editing' tradition by transmitting, in real time, life as it happens, from the perspective of the surveilled (Fig 9-2).

9.4.2 I didn't take the picture, and I don't know who did

Perhaps the most important aspect of the imaging philosophy of Chapters 3,4,5, when considered in the context of this chapter, is the fact that they are merely measurements — not a picture until a picture is rendered from this data. Thus some interesting questions arise. Is it forbidden to take a light reading in a department store such as the MIT bookstore (the Coop), where photography is prohibited? When this data is sent to a remote site, and someone uses it to remotely navigate in the space that I am in, what happens if they use a virtual camera to capture an image, from a remotely navigated environment map. How does a department store enforce a prohibition of photography over a videoconferencing link, especially when the person at the other end of the link is communicating through an anonymous chain of packet retransmitters?

Because I am merely capturing measurements of light (based on the photometric image composite, which represents the quantity of light arriving from any angle to a particular point in space), which are then yet to be "rendered" into a picture, I may choose to leave it up to a remote viewer, operating a telematic 'virtual camera' (Fig 9-2(b)) to make the choices of framing of the picture (spatial extent), camera orientation, shutter speed, exposure, etc.. In this way I may absolve myself of responsibility for taking pictures in establishments (such as department stores) where photography is prohibited, for I am merely a robot at the mercy of a remote operator who is the actual photographer (the one to make the judgement calls and actually push the virtual shutter release button). In this manner, an image results, but I have chosen to not know who the photographer is. Indeed, an important purpose of these personal documentaries has been to challenge representatives of the video surveillance superhighway who also prohibit photography and video.

In these personal documentaries, such as 'ShootingBack' [128], there were typically two audiences,

one audience to which I performed, and another remote audience. Members of the remote audience knew they were an audience because they were entering a traditional "gallery". Even though it was virtual in the sense that it was on the Internet, it was still traditional in the sense that the interaction was analogous to a real-world gallery or museum. The other audience comprised those who were physically present in front of the WearCam apparatus (e.g. representatives of the surveillance superhighway, and other customers/patrons of their establishment).

The physically-present audience, at first, does not realize that they are an audience. On one level, they might be regarded as the "enemy" (e.g. they are being "shot at" in the sense of ShootingBack), while on another level, the performance is directed at them — to educate them, teaching being an act of love and human compassion — so that they are regarded, through sympathetic introspection, as fellow members of the "oppressed".

ShootingBack was a meta documentary (a documentary about making a documentary). Since I am a camera, in some sense, I do not need to carry a camera, but in ShootingBack, I did anyway. This second camera, an ordinary hand-held video camera, which I carried in a satchel, served as a prop, with which to confront members of organizations who place us under surveillance. First, before pulling the camera out of my satchel, I would ask them why they had cameras pointing at me, to which they would typically reply "why are you so paranoid", or "only criminals are afraid of the cameras". All this, of course, was recorded by my WearComp/WearCam apparatus concealed in an ordinary pair of sunglasses. Then I would open up my satchel and pull out the hand-held video camera and point it at them in a very obvious manner. Suddenly they had to swallow their own words. In some sense, ShootingBack got *the pot calling the kettle black.*

Personal anecdotes

To further the reflectionist symmetry, I also experimented with wearing some older more obtrusive versions of WearComp/WearCam, which I described to paranoid department store security guards as 'personal safety devices for reducing crime'. Their reactions to various forms of the apparatus were most remarkable. On one occasion, an individual came running over to me asking me what the device I was wearing was for. I told him that it was a personal safety device for reducing crime, and that, for example, if someone were to attack me with a gun or knife, it would record the incident and transmit video to various remote sites around the world. I found that by taking charge of the situation, and throwing the same rhetoric back at them, that even though photography was strictly prohibited, I could very overtly and obviously take pictures in their establishment, telling them in plain wording that I was doing so, and that there was nothing that they could do or say about it. I found that there was a big difference in the way that they responded to a hand-held video camera, and to a device that was presented to them as a machine "for purposes of personal safety and reducing crime". In particular, my approach, which essentially forced them to either swallow their words or their policy, left them tongue-tied, unable to apply their "photography prohibited" policy, confused, bewildered, in what I believed was a state of deep thought — at least they finally began to think about the consequences of their blind obedience.

In another incident, I happened to be carrying one of my older (more obtrusive) rigs in a backpack, and entered a small bagel shop. I bought a bagel and sat down at a table to discover a camera on a shelf, pointed at me. The conversation went as follows:

- me: "What's that" (pointing to camera)?
- owner: "It's just a video camera, nothing to worry about!"
- me: "Is it taking pictures of me?"
- owner: "It's just videotaping the shop, nothing to worry about!"
- me: "Why?"
- owner: "In case someone comes in here with a gun".

At that point, I opened up my backpack and took out my heavy old eyeglasses (with stereo camera (two cameras) and two large viewfinders serving as computer screens), walked over to him, and looked him in the eye. A conversation resulted, as follows:

- owner: "What's that (pointing to my rig)?".
- me: "It's just a video camera, nothing to worry about!".
- owner: "Are you videotaping me?"
- me: "I don't know; the images from my camera are being transmitted to various remote sites around the world. I'm not sure how many remote [World Wide Web] sites there are [viewing this video], or in how many different countries they are located, or if people at those remote sites might be saving some of the video images, but nothing to worry about!"
- owner: ""Why"
- me: "In case someone attacks me with a gun — in case I'm assaulted or attacked — personal safety!".

The above example was typical of a reflectionist performance piece — after the conversation, the owner was put into a state of deep thought (reflection) as he suddenly came to grips with the situation he'd created by installing a video camera. Had I merely complained about the camera, he might have written me off as a paranoid luddite, but it is difficult to call someone a luddite when they are wired into the internet over a 56kbps wireless link running through an antenna sticking out of their hat.

In some sense by calling me paranoid, it would be the pot calling the kettle black. Thus reflectionism affords one with the ability to criticize video surveillance without being written off or easily dismissed, because the absurdity of an argument is exposed by agreeing with it.

WearCam Concept

A problem with Wearable Wireless Webcam was that people were often too enamored by the technology itself to see the underlying philosophical concept of reflectionism, so a "low-tech" embodiment of the new philosophy was needed to isolate the concept from its realization.

The following are experiments that I have conducted and purposely taken to the extreme in order to (a) illustrate a point and (b) experience reactions and observations first hand. It is not likely that the average reader would go to these extremes but some more subtle variations of these experiments will still provide similar insight or reactions. They are presented, in the tradition of conceptual art, in the form of a "recipe" or list of instructions, especially since some of them are easy enough to implement that motivated readers will be able to repeat these performances.

Maybe Camera: You cannot patent a mere "idea", but, rather, the idea must first be *reduced to practice*. Similarly, you cannot copyright an idea, it must first manifest itself as some *tangible* form. Conceptual art, however, provides us with a means where the idea itself is the contribution. Accordingly, I propose the following:

- Take one piece 1/8 inch black or dark acrylic, cut to 3 by 4 inches.
- Obtain a bulky sweatshirt in your size.
- Print the words: "For YOUR protection, a video record of you and your establishment *may* be transmitted and recorded at remote locations. ALL CRIMINAL ACTS PROSECUTED!" in large letters, on the front of the shirt. Lay out the lettering so as to leave room for the acrylic between the words "locations" and "ALL" ("locations" to be at the end of one line of text, and "ALL" to begin the next line of text). (See Fig 9-3.)
- Affix the acrylic securely to the shirt.
- Wear the completed shirt into a department store or other location where

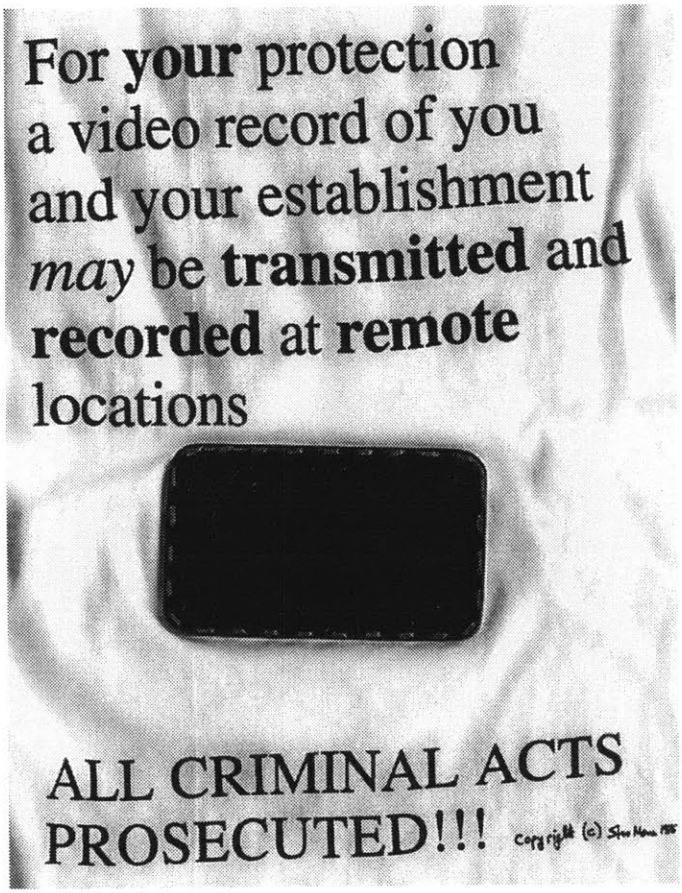


Figure 9-3: Construction of an embodiment of the 'Maybe Camera' concept. The ostensible altruism contained in the message (note how the word "your" is boldfaced) is a direct reflectionism of the signs typically posted on the entrances to department stores and the like (e.g. "For **your** protection, you are being videotaped").

- video surveillance is used but
- photography is strictly prohibited (this criterion can be determined experimentally even before the shirt is made, by entering the proposed establishment with a 35mm camera or the like, and taking pictures within said establishment in a somewhat obvious manner).

The above piece (Fig 9-4) is called 'Maybe Camera — Who's Paranoid?'. Another variation of 'Maybe Camera' involves making a large number of these shirts, but putting a real camera and transmitter (someone with repeater in a backpack provides uplink to a car parked outside the shop, which in turn wirelessly uplinks to the Internet) into one of the shirts, and having a large group wearing them on the surveillance superhighway (Fig 9-5).

Probably Camera: Depending on the level of paranoia, if 'Maybe Camera...' is not "understood" by your audience, then perhaps the following conceptual/performance/reflectionist piece would be:

- Obtain one miniature (12 inches in diameter or smaller) ceiling dome of wine-dark opacity, together with a camera and pan-tilt-zoom mechanism suitable for that dome.
- Affix dome to backpack, facing backwards, cutting appropriate mounting hole in backpack, leaving sufficient space, and installing appropriate housing for camera and pan-tilt-zoom mechanism. Leave the camera out for the time being.
- Insert a small battery powered computer equipped with video capture hardware, and means of controlling the function of the pan-tilt-zoom controls automatically.

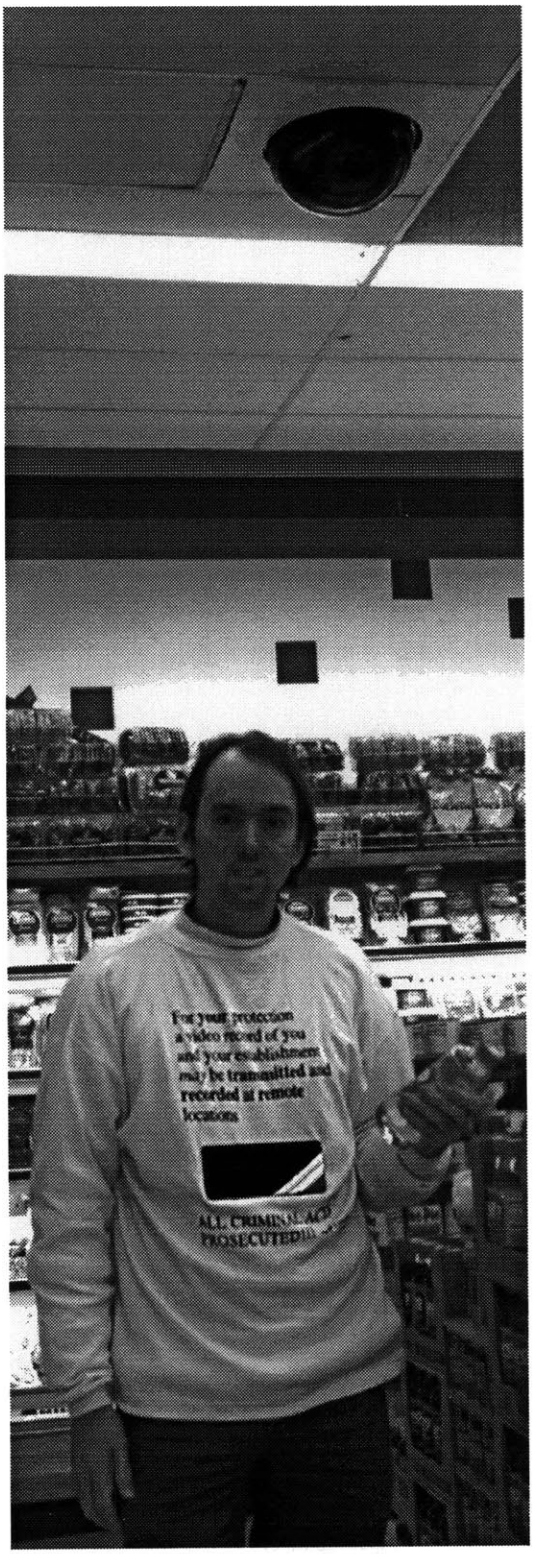


Figure 9-4: Author wearing a realization of "Maybe Camera — Who's Paranoid?". Just as I don't know what's in the mysterious ceiling dome of wine-dark opacity above my head, and must therefore be on my best behaviour at all times, so too, the shopkeeper doesn't know what's inside my shirt, and likewise must be on his best behaviour at all times as well.

Figure 9-5: 'Firing Squad': An early performance of 'safety net' — a maybe wireless network of individuals maybe looking out for one another's safety. One of us is wearing a camera with transmitter, but none of us know who has the real shirt (e.g. the one that works). Therefore, none of us are guilty of knowingly taking pictures within this establishment.

- Insert into the pack, means of wireless communication to/from the Internet, or to/from an Internet gateway/server.
- Prepare software to allow the function of the apparatus to be controlled remotely via a WWW page, with ability to capture and display images from the camera if the camera is present. Make this WWW page world-accessible and known to various people around the world.
- Leave the work area and have someone else do the final assembly in your absence, according to the following instructions: Roll two dice, and:
  - If the total comes to two or three, insert into the dome a small light bulb, affixed to the pan-tilt-zoom sensor but connected to it in no way, together with sufficient ballast into the pack to make up the difference in weight between the bulb and the camera, so that the wearer could not determine this difference by weight.
  - If the dice total exceeds three, insert the camera, properly mounting it and connecting it to video digitizer. Verify its operation using a Web browser of your choice.
- Wear backpack together with shirt ('Maybe Camera...'), into a record store, preferably "Tower Records", where ceiling domes of wine-dark opacity are used. If asked if it is a camera, or what it is, indicate that you're not certain, but point out the domes upon their ceiling and indicate the similarity, so that perhaps it could be a light fixture. (Security guards at Tower records have informed me that their ceiling domes of wine-dark opacity are "light fixtures".)

The above piece is called 'Probably Camera — Who's Paranoid?'.

Probably Camera and Maybe Camera can be worn together of course, since one uses the front of the body, while the other uses the back.

No Camera: A conceptual piece, involving video time-delay[6], to symbolize the disjointness between cause and effect that video recording creates is now described:

- Place pinhole camera and microphone into baseball cap, and record video from an establishment where photography, filming, and the like is strictly prohibited, but where video surveillance is used, and where there are documented cases of hidden cameras having been used. While recording video, talk to members of establishment, including manager. Ask whether or not they use video surveillance, and if so, why they are videotaping you without your permission. Ask what their "ceiling domes of wine-dark opacity" are, if any are present.
- Leave this establishment, and return with the following, but without the camera:
  - Flat-panel television screen affixed to shirt.
  - Source of previously recorded video material.
  - Means of switching between previously recorded material and standard broadcast television channels.
- Play the previously recorded video on the television screen, and if you are informed that photography, filming, or the like, is prohibited, indicate that there is NO CAMERA, and that what you are wearing is merely a television. Switch through the various channels, indicating that one of them (the one playing the previously recorded material) looks like it "must be a local channel — a VERY local channel".

The piece is called 'No Camera — Who's Paranoid?'.

9.4.3 'My Manager': Empowerment through subservience

Obsequious servility is often the hallmark of a sycophant trying to please his/her boss, manager, or the like. However, what happens when the obsequiosity is actually directed at one or more third parties, and sometimes when that boss or manager does not even really exist or is not at all involved, is something that I call 'subservience empowerment'.

Desiring or pretending to be under the control of a higher and unquestionable authority is actually a very effective means of empowerment. As an example of 'subservience empowerment', consider the following dialog between a customer and a used-car sales clerk:

```
customer:  Would you accept $2000?
clerk:     Let me check with my manager [clerk disappears into back
           room, has a coffee and glances over a newspaper all alone,
           because he's actually the manager himself, and it is actually
           his decision].
clerk:     [a few minutes later, emerging from his solitude]
           I'd love to give you the car for $2000 but my manager won't
           budge on the price.
```

Détournement of the official apparatus of externality and contractual obligation

One of my own subservience empowerment performances, called 'Self Ownership' involved joining an existing modeling agency (this was around 1985, actually as a "cyborg" performance artist) and signing away the rights to my likeness, such that I was then forbidden to pose for pictures that were kept on file by others. (This was in the context of an employee ID card for when I worked as a part time electronics technician in the Summer of 1985.)

[6] Other artists have also experimented with video time-delay but in different contexts. For example, Dan Graham uses video time delay together with mirrors, etc., to create a delay between cause and effect. His *video feedback* involves both senses of the word "feedback": (1) the cameras "sees" the screen which is displaying the output from the camera, and (2) the users who see themselves on the screen adjust their behaviour according to this psychological "feedback".

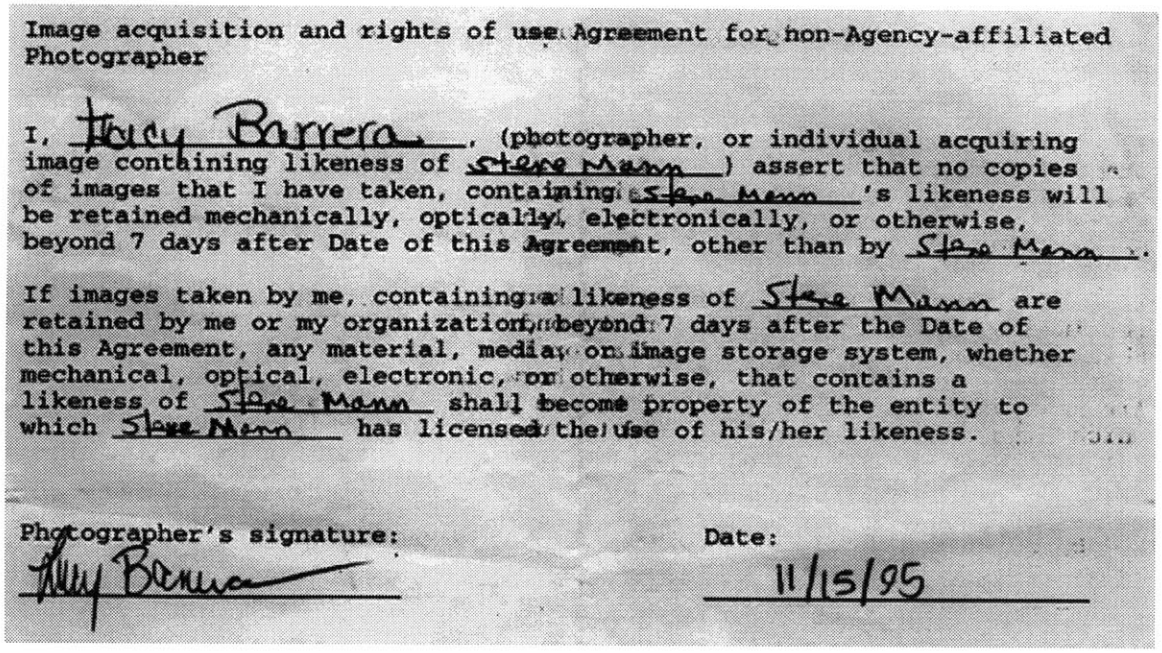


Image acquisition and rights of use Agreement for non-Agency-affiliated Photographer

I, Tracy Barrera, (photographer, or individual acquiring image containing likeness of Steve Mann) assert that no copies of images that I have taken, containing Steve Mann's likeness will be retained mechanically, optically, electronically, or otherwise, beyond 7 days after Date of this Agreement, other than by Steve Mann.

If images taken by me, containing a likeness of Steve Mann are retained by me or my organization beyond 7 days after the Date of this Agreement, any material, media, or image storage system, whether mechanical, optical, electronic, or otherwise, that contains a likeness of Steve Mann shall become property of the entity to which Steve Mann has licensed the use of his/her likeness.

Photographer's signature: Tracy Barrera

Date: 11/15/95

Figure 9-6: Contract I routinely had signed by the MIT Card Office, as required by Agency before I was permitted to pose for an ID card picture. (The MIT ID card office preferred to sign this contract rather than simply print an ID card without picture.) Members of Agency (e.g. students) previously sign away the rights to their likeness being stored in any database to the Agency, so that they are prohibited from being photographed before this contract is signed. (Photograph of contract Copyright (c), Steve Mann, 1995)

A more recent performance of 'Self Ownership' involved requiring that the MIT ID card office agree to similar terms (e.g. agency requires deletion of image from database after ID card is made) prior to my appearing for an ID card picture.

A "modeling agency" was constructed, and participants (including the author) signed up with this "modeling agency" and signed away the rights of their likeness to the agency.

Members, having signed a contract with the Agency were not permitted to be photographed by non-agency approved photographers. There is a clause in the contracts that allows them to be photographed by anyone provided that the model retain all image data (whether for their portfolio or for their ID). Family members were exempt (e.g. in the context of family snapshots) — the contract was directed at organizations making a systematic database of people's likenesses.

Under the ID Card infrastructure, students who have joined the Agency would be violating the terms of their contract if they were "captured" in the college ID Card database. This is because, an ID Card database was defined as a form of publication, according to the contract that these students have signed with the Agency.

Whether or not the data is in fact published in the traditional sense is irrelevant to the "Agency". Just the fact that image data would be "online" (e.g. on a computer system which has the capability of being accessed by or disseminated to more than one person) would make it a violation of the Agency's contract.

In order to enforce the deletion of the pictures, representatives of the ID card office were required to sign a form (Fig 9-6), from the Agency, indicating that they would delete images from the database, and in the event of not having deleted images from the database, the ownership of all manner of hardware and software upon which unauthorized copies of Agency members' likenesses were found would be transferred to the Agency.

This goal of this approach was to prevent the MIT Card office from keeping a copy of participants' likenesses, without leaving the onus on the participants to justify a reason for not wishing to be part of the face picture database and the host of possible covert uses that being part of such a database potentially entails.

WearComp and 'My Manager'

The WearComp invention proposed in this thesis can facilitate 'subservience empowerment' in an even more compelling way. A performance, called 'My Manager, which was based on WearComp, is now described.

'My Manager', borrows from the Stelarc/Elsenaar tradition in performance art[7]. 'My Manager' allows participants to, via Radio TeleTYpe (RTTY), become managers and remotely contribute to the creation of a documentary video in an environment under totalitarian surveillance (where photography, video, etc., other than by the totalitarian regime is prohibited).

In 'My Manager', I am metaphorically just a puppet on a "string" (to be precise, a puppet on a wireless data connection). I might, for example, dutifully march into the establishment in question, go over to the stationery department, select a pencil for purchase, and march past the magazine rack without stopping to browse the magazines, because I can't stop to browse the magazines. In this example, I have been sent on an errand to purchase a pencil for a higher and unquestionable authority. When challenged by the department store's clerks or security guards, as to the purpose of the cameras I am wearing, I indicate that what I am wearing is a company uniform, and that my manager requires me to wear the apparatus (the uniform) so that she can make sure that I do not stop and read magazines while I am performing errands on company time. Sometimes I remark: "I trust you, and I know you would never falsely accuse me of shoplifting, but my manager is really paranoid, and she thinks shopkeepers are out to get her employees by falsely accusing them of shoplifting"[8]. Thus a goal of this performance was for the individual to assert a right to wear/be a "black box" hazard recorder to be used for possible forensic purposes related to personal safety.

Just as representatives in an organization absolve themselves of responsibility for their surveillance systems by blaming surveillance on managers or others higher up their official hierarchy, I absolve myself of responsibility for taking pictures of these representatives without their permission because it is the remote manager(s) together with the thousands of viewers on the World Wide Web who are taking the pictures.

The subjects of the pictures, for example, department store managers, who had previously stated that "only criminals are afraid of video cameras", or that the use of video surveillance is beyond their control, either implicate themselves of their own accusations by showing fear in the face of a camera, or acknowledge the undesirable state of affairs that can arise from cameras that function as an extension of a higher and unquestionable authority.

If their response is one of fear and paranoia, I hand them a form, entitled RFD (Request For Deletion) which they may use to make a request to have their pictures deleted from my manager's database (I inform them that the images have already been transmitted to my manager and cannot be deleted by me). The form asks them for name, social security number, and the reason for which they'd like to have their images deleted, and requests that they sign a section certifying that the reason is not one of concealing criminal activity, such as hiding the fact that their fire exits are illegally chained shut. (See appendix B) This form is, itself, a reflectionist effort, for it has turned the machinery of bureaucracy back on the representative of the surveillance superhighway.

Through 'reflectionism' the department store attendant/representative sees himself/herself in the bureaucratic "mirror" which I have created through being a puppet on a (wireless) "string". 'My Manager' forces attendants/maintainers/supporters of the video 'Surveillance Superhighway', with all of its rhetoric and bureaucracy, to realize or admit that they are "puppets" for a brief instant, and confront the reality of what their blind obedience leads to.

---

[7] Both Stelarc [129] and Elsenaar [130] explore body control systems which use electrical stimulation to cause their muscles to move in response to an external input.

[8] There are well-documented cases where security guards have abused their ability to stalk victims for purposes of rape and murder, and where shopkeepers have falsely accused their customers of shoplifting, so my assertion is not as absurd as it might seem. In one well known murder case: "on march 16th, 1991, 15 year old Latasha Harlins entered a Korean owned grocery store to purchase a carton of orange juice. Soon Ja Du, the store owner, accused her of shoplifting even as Latasha attempted to pay for the juice. After a struggle in which Du tried to grab her book bag Latasha placed the juice back on the counter. As Latasha turned to go, Du shot her in the back of the head, killing her."

9.4.4 WearCam as 'cyborgian primitive'

In the following experiments, I have purposefully taken a principle to its extreme to show just how far out-of-balance the surveillance superhighway has gone. In particular, a camera is constructed as a permanent fixture of the body in order to challenge+balance+reflect the elements of the video surveillance superhighway and the way that they have defended themselves from being questioned by becoming permanent fixtures of our architecture and urban-planning infrastructure.

An earlier version of 'cyborgian primitive' (performed in the early and mid 1980s) involved growing hair through fine mesh in a skull cap, and then "locking" it on the other side (hair locking may be accelerated by teasing in bee's wax to cause the hair to tangle together permanently.). After the use of conductive/metallic hair dyes (to help make the hair form part of a ground-plane for a transmitter), the hair was found to be sufficiently "damaged", that it locked quite easily into a fuzzy metallic-grey mess. The skull cap then formed a substrate upon which to mount other devices. In this manner, I could not reasonably be asked to remove the apparatus, because that would require shaving off my hair. This necessary subversion of the body meant that there was a reasonable barrier to requests by others that the apparatus be removed.

'Cyborgian primitive' is related to the 'modern primitives' movement, in the sense that it involves physical modification of the body, but it is more general in that it may also involve modification of the brain as well. A more recent variant of 'cyborgian primitive' depended on modifying the brain rather than the body. These experiments were based on the 'mediated reality' of Chapter 6 which I have used as a method of conducting scientific experiments in visual perception, as well as for prosthetic purposes. As a prosthetic, the apparatus described in [13] (Fig 6-12), allowed me to computationally augment, diminish, or otherwise alter the perception of reality for the purposes of attaining a heightened sense of awareness, to see better, or to compensate for a visual deficiency that is of a form that cannot be corrected with ordinary (pure-refractive optical) eyeglasses.

As was noted by other scientists, I found, as described in Chapter 6, that often, an adaptation to the apparatus occurred, and that, after some time, a dependence on the apparatus was developed, such that removal of the apparatus would result in the inability to see properly, including the experience of nausea, dizziness, and disorientation. In this way, the device evolved into a necessary prosthetic, so that I could not reasonably be asked to remove it.

With this deliberate modification of the visual cortex, to develop alternate neural pathways, through the process of certain kinds of very long term visual adaptation, one may attain a permanent or semi-permanent bonding with the apparatus, in the sense that others cannot reasonably ask that it be removed.

In the spirit of 'reflectionism', WearCam is made to function as a true extension of the body, as a third eye (or second pair of eyes in the case of some of the two-camera systems presented in Chapter 6.

The incorporation of components into clothing is also another way to create a barrier to requests that the apparatus be removed.

> Clothing and cloth...prostheses (to cover and protect, to extend and support the body) such objects often become, after years of use, integrated so inextricably with one's *psychic body* that they cannot be replaced or removed without subversion of the physical body itself. The same holds true for objects that function as prostheses of the mind [5].

Beyond the fact that a totalitarianist asking that the device be removed is asking the wearer to violate or subvert his or her own body there is also the obvious legal responsibility he or she must accept for the prospect of an abrupt exposure to the previously defined neural pathways, and the possibility of any brain damage or onset of *flashbacks* that might result from a sudden re-instantiation of the old (temporarily or semi-permanently weakened) neural paths.

Thus when asked to remove the apparatus, if in fact it even could be removed (e.g. if it were not permanent or semi-permanent) one might merely present the totalitarianist attendant with a form to sign accepting all responsibility for any damage. This use of forms (e.g. an individual presenting officials with forms) is itself a 'reflectionist' gesture.

A joint mental and physical (permanent/semipermanent head cap) bonding was recently used in a self-ownership piece called 'primitive identity'. In this piece, I defined my sense of self in all

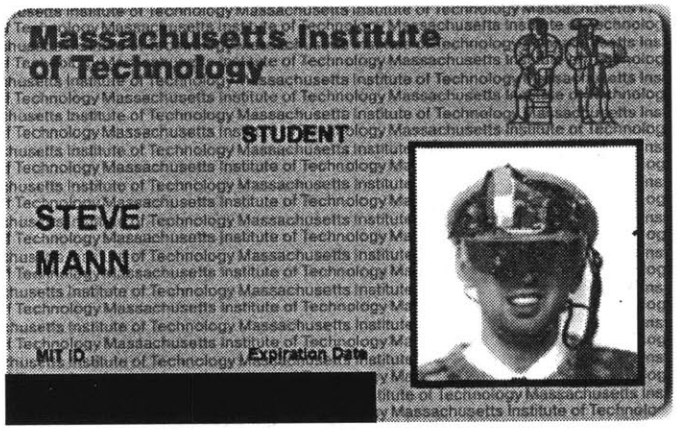


Figure 9-7: 'Primitive identity' established through a permanent/semi-permanent bonding with a prosthetic device. In this piece, the older (more primitive) apparatus of Fig 6-12(c) was chosen because it was the most amenable to a longer term 'cyborgian bonding'. Bonding with a past device was found to be easy — the mapping was re-learned quickly.

manner of official portraiture (e.g. Fig 9-7), regardless of any requirements that eyeglasses and the like may not be worn during such portraiture.

9.5 Reflections of the five horsemen of the surveillance superhighway

I will now conclude the presentation of the reflectionist philosophy with a brief summary that ties together the performances with each of the ways that proponents of the surveillance superhighway defend their position.

1. Secrecy: Completely covert embodiments of the WearComp/WearCam invention have been created, and used in places where photography is strictly prohibited, to document video surveillance apparatus together with representatives of organizations responsible for it. This work was presented in Chapter 7. Furthermore various 'reflectionist' performance pieces were made to "mirror" the structure of environmental surveillance. These included 'probably camera', a wearable device constructed to look like a camera, but in such a way that others were not sure which way it was pointing, and 'maybe camera', constructed so that others could not readily determine whether or not the device was a camera.

2. Rhetoric: The rhetoric of personal safety was used to reflect the rhetoric of public safety.

3. Constancy: The devices were wearable, rather than carried, so that it could not be discerned by others, whether they were in actual use or not, at any given time. The devices were part of the wearer's *prosthetic territory* (e.g. clothing).

4. Appeal to higher and unquestionable authority: Various performances, such as 'My Manager', used the principle of 'subservience empowerment'.

5. Criminalization of the critic: Inferences pertaining to the possible criminal intent of those who questioned personal imaging devices were made (e.g. in the RFD of Appendix B). These and other inferences included possible violations of fire safety, such as fire exits illegally chained shut or fire exits blocked (e.g. Fig 5-6).

Figure 9-8: Recent picture I took of myself (at left — I triggered the exposures through remote control of another imaging apparatus using radioteletype) together with the rest of a 'Safety Net', engages in a collective diffusionist performance piece. Some of these units are NETworked via TCP/IP, to work together to increase safety and reduce crime, but not all these units are on, transmitting images. Can you tell which ones are operational? (Four of the units I built myself, three with Internet connectivity, two actually running while the picture was being taken.) In some sense, this collection of 'maybe cameras' (e.g. one doesn't know which ones are transmitting) is a reflectionist questioning of Bentham's Panopticon. Just as surveillance puts the prisoner on his or her best behaviour at all times, the 'maybe camera' principle could be used to put police officers, department store owners, and public officials on their best behaviour at all times.

9.6 'Diffusionism' as second-choice in case 'reflectionism' fails

In the event that my 'reflectionist' philosophy should fail to have the desired impact (e.g. should it fail to raise sufficient awareness to make a meaningful reduction in the inappropriate use of video surveillance), an alternative philosophy, 'diffusionism', is proposed.

The goal of 'diffusionism' is to subvert the totalitarian nature of surveillance through a proliferation of wearable 'maybe cameras'.

As Foucault notes, it is not essential that the guard in the tower be watching a particular prisoner, or that there even be a guard in the tower; it is only necessary that the prisoner not know whether or not there is a guard in the tower who could be watching. Similarly, to subvert Panopticon, it would not be essential that the guard be watched, but just that there be a possibility that the guard could be spotted by a "prisoner" at some time.

To this diffusionist end, I have created a wireless communications infrastructure capable of supporting a networked community of hobbyists wearing a similar apparatus. During one performance piece, I, together with a group of others willing to participate, went out shopping one day, wearing such apparatus (thus those at the department store needed to confront not just one, but many of us). The quick snapshot I originally took of this group was of such popularity but poor image quality, that I re-did the shot[9] (Fig 9-8).

Part of the 'diffusionist' goal is enhanced by finding every-day uses for a wearable camera such as its use by those with visual memory disability to automatically recognize people [12] (we all suffer from some degree of visual memory deficit), as well as in wearable, tetherless computer-mediated reality.

While one might be inclined to think that the inevitable commercialization of this invention may mark the détournement of this détournement, diffusionism is put forth as a détournement of

[9] Some of the group members pictured in 9-8 are different than those in the original picture of the first performance.

a détournement of a détournement (as in the equation diffusionism=détournement^3). In this sense, diffusionism will attempt to appropriate the machinery of mass production and mass market — the very tools that threaten to appropriate the tools of reflectionism.

To this end, the goal is to turn WearCam into a useful and commercially viable everyday object, to help us see better, avoid getting lost (automatic directions combined with object recognition, GPS and video overlays), and recognize and remember people better. Thus these very utilitarian applications of WearCam may serve as a détournement of utilitarianism itself.

Summarizing, on the chronology of reflectionism and diffusionism in comparison to other movements in the history of art: futurist → dadaist → surrealist → situationist → reflectionist → diffusionist.

9.7 Chapter summary

'Reflectionism' has been proposed as a détournement of video surveillance with the hope of raising some serious questions and debate as to the merit of video surveillance — "safety and security, but at what price?".

The goal of 'reflectionism' is to 'reflect back' the principles of visual surveillance, such as the principle that it often arises from a higher and unquestionable authority. Reflectionism confronts representatives of the 'Surveillance Superhighway', forcing them to think about ("reflect on") their actions or inactions (blind obedience).

While the commercialization of my 'WearComp' invention seems inevitable, and may mark the détournement of this détournement, perhaps there is hope for a détournement of a détournement of a détournement (as in the equation reflectionism=détournement^3).

Thus, should 'reflectionism' fail, an alternative principle, 'diffusionism', is proposed. The goal of 'diffusionism' is to make visual image capture and transmission, or at least its potential, ubiquitous and diffuse it throughout the general population, through the appropriation of the tools of commercialization and mass production.

While reflectionism attempts to slow down what might be an inevitable engine of "progress", diffusionism attempts to outrun it — to gain speed toward diffusing the power of observation and informational control that might otherwise be given only to those within the totalitarian regime.

Both 'reflectionism' and 'diffusionism' borrow (either as negative or positive examples) from many other traditions in art. Chronologically, these are ordered as follows: futurist → dadaist → surrealist → situationist → reflectionist → diffusionist.

In much the same way that Shakespeare's Tragedies hold a mirror to what is mysterious and unquestionable in our lives ("Hamlet's mirror", telling us "The purpose of playing" was to "hold the mirror up to nature"), my goal in formulating the reflectionist philosophy is to hold a mirror up to the absurdity of the vision that futurists put forth for us. And if they cannot see their own absurdity, then, diffusionism remains as a fallback plan.

Chapter 10

Summary and conclusions

The purpose of art is to lay bare the questions which have been hidden by the answers.

–James Baldwin

'Personal Imaging' was proposed, under the broader umbrella of 'humanistic intelligence', as a body of work at the intersection of art, science, and technology. This synergy of ART, SCIENCE, and TECHNOLOGY — particularly apropos for a dissertation in the "Media **Arts** and **Sciences**" program of the "Massachusetts Institute of **Technology**" — attempted to capture a new imaging renaissance afforded by the tools of humanistic intelligence. Three precisely defined criteria were set forth to define an apparatus exhibiting (or more precisely, *enabling*) humanistic intelligence. Most notably, the proposed 'WearComp' apparatus was characterized by, and attained, these three very closely intertwined and overlapping criteria:

- The **ephemeral** criterion:
  1. Instantaneous response of the human to the machine was attained through long-term adaptation. It was found that for many different configurations, after using the apparatus for an extended period of time, that it functioned as an extension of the mind and body, in the sense that conscious effort was not required in order to use it. Thus the **'first brain ephemeral'** criterion was met.
  2. Instantaneous response of the machine to the human was attained through the development of realtime processing methodologies. To this end, a new method of simulating a future generation of WearComp using a full-duplex wireless communications facility was developed. Other contributions to achieve this goal include some near realtime video and image processing algorithms. Thus the **'second brain ephemeral'** criterion was met.
- The **eudaemonic** criterion: Physically, the human and machine were inextricably intertwined, to seamlessly fit together into a single unit. Both facets of this eudaemonic criterion were met:
  1. The **social** facet was met, in the sense that the computational apparatus was situated so that others would not perceive the machine as a separate entity (and in the more recent versions, the apparatus was made covert so that it would not be perceived as separate at all). It was found, for example, that upon entering establishments such as certain department stores or the like, where one is typically asked to leave one's bag or briefcase at the counter (and in fact, where photography is strictly prohibited), that the apparatus succeeded in being interpreted as part of the body.
  2. The personal facet of the eudaemonic criterion was also met, in the sense that the apparatus was so successful in altering human perception, that an actual dependency developed, in which it became a necessary prosthetic device.

Although the user-interface was not necessarily easy-to-learn and in fact, required years to master, the combination of human and machine was able to reach higher levels of achievement

in various goals (such as photographic creativity, personal cinematographic documentary, enhanced collaborative efforts). This aspect was facilitated through making the apparatus wearable, so that it could be present and active constantly, over an extended period of time.

- The **existential** criterion: The apparatus was found to provide a certain kind of personal empowerment. In addition to that afforded by the above two criteria (the empowerment of a well-learned user-interface for which there are obvious barriers to others to ask that it be removed), one of the things that grew out of this effort was a framework for 'existential technology'. This 'existential technology' provided self-determination and mastery over one's own destiny, in the form of both observability and controllability:

  1. Most notably, a variety of user-interfaces were proposed to empower the user, through putting the user in the feedback loop of most major functions of the apparatus. Mediated reality was itself an example of an 'informative' user-interface providing the user with complete control in the observability sense. 'Saltach' was another example of such a user-interface — one in which the user is kept informed of the status of the apparatus. Therefore, the **'existential observability criterion'** was met.
  2. Mediated reality was suggested as a framework for creating a new genre of documentary video, characterized by control of the camera that is shared between 'first brain' and 'second brains'. Thus even when automatic exposure, automatic image enhancement, and other degrees of automation were implemented, I always had control of the situation, so that these degrees of automation themselves, which exemplified existential observability, also behaved as extensions of my mind and body over which I exerted full control. In this manner, they attained the **'existential controllability criterion'**.

The specific focus of this thesis was on personal imaging, which was attained through a particular embodiment of WearComp, which I have called WearCam. This embodiment was found to give rise to a new creative genre of imaging, which went beyond the traditional goals of photographic imaging (and in that sense, digital photography falls into the traditional goal with it's traditional mindset). WearCam also took a first step toward defining a new genre of documentary video.

Furthermore, a new conceptual/mathematical framework was put forth for personal imaging, based on the linearity and superposition properties of light. This framework was based on the notion that operations should be done in 'lightspace' — in the range of light falling on the image sensor — and is based on a proposed self-linearizing camera calibration procedure. The new framework stressed the importance of the range of light. Just as cameras are frequently calibrated to correct for spatial distortion (e.g. barrel distortion, etc.), they should also be calibrated to correct for tonal distortion, giving rise to their use as photometric measuring instruments. Even when all we desire is an ordinary picture, e.g., when we are not interested in the scientific measurment process, it was found that the proposed framework provided much utility in the visual arts.

This new framework was put forth also in the context of projective geometry. To this end, an effort was made, by combining daVince's interpretation of light with projective geometry, to establish a new renaissance in imaging, toward a simple yet powerful tool that can be used in a personal imaging system, or with ordinary video or still cameras. In particular, a problem that others have attempted to solve, but not solved, namely a featureless method of estimating the projective coordinate transformation between pairs of images, was proposed, and then later this 'Video Orbits' methodology was also combined with the new 'lightspace' framework.

Lastly, a new philosophical framework for art and cultural criticism was proposed, and applied to personal imaging. This framework consisted of first identifying a new problem (e.g., something to apply the criticism to), and then developing two new philosophical constructs of criticism, "reflectionism" and "diffusionism". Both of these were applied, in the tradition of the fine-arts, to the domain of personal imaging.

Appendix A

Form-698 — Request For Deletion (RFD)

The Request For Deletion (RFD) form comprises a single sheet of paper, two sides. (See Fig A-1, and Fig A-2.)

REQUEST FOR DELETION
(RFD)

In the interest of employee safety, our employees are required to wear uniforms equipped with protective media to discourage others from exposing them to dangerous situations or environments (e.g. establishments where fire exits are chained shut illegally or the like), or to falsely accuse our employees of crimes (such as shoplifting and the like).

Our employee uniforms capture images, photometric measurements, and or other measurement information and the like, which may have been and may continue to be transmitted and recorded at remote locations. Furthermore, our employees are required to document, via more traditional photographic means, any incident in which there is a perceived or suspected safety hazard, or any incident in which there might be potential for a crime to be committed in the future (such as when an employee presents a company credit card, when an employee makes a purchase but is not given a receipt that provides proof of the purchase such that a false shoplifting accusation could be forthcoming, or when cash is being handled by one of our employees). For YOUR protection, our employees are also required to photograph each person they interact with, as well as maintain a recording of the conversation, in order to pre-empt any disputes regarding return of merchandise.

If you feel that one of our employees has documented something within your establishment that you do not wish to remain on file in our image archives, or if you feel that your likeness should not remain on file in our archives you may submit a REQUEST FOR DELETION (RFD) to our employee who will forward it to our Company Headquarters.

Your RFD, if properly completed in full, will be presented to a committee, and a decision will be made as to whether to expunge said image(s) or to flag said images as noteworthy (e.g. by submitting an RFD, you should be aware that it may in fact cause your likeness to be flagged as suspicious or of special interest to the permanent archives).

Part I: Declaration of reason for RFD (please circle only one):

A National security:
You must be a government establishment or have government affiliation (such as government funding) to select this option.

B Company confidential:
B1: A trade secret has been inadvertently documented by our employee.
B2: Strategic marketing plans have been inadvertently documented by our employee.
B3: Other company confidential _____
(please describe, use additional page if necessary)

C other _____ (please describe, use additional page if necessary)

Part II: Declaration of abstinence from willful destruction of evidence of a criminal act.

In recognition of the fact that measurement (photometric, radar, or otherwise) data captured by our employee may comprise evidence, in the context of a possible future criminal investigation against me or my establishment, I, the undersigned, declare that my REQUEST FOR

Figure A-1: Front of Request For Deletion (RFD) form

DELETION is not for purposes of concealing criminal activity of myself or of others in my establishment.

I assert that my RFD is not intended to hide criminal activity of any kind occurring within my establishment, including, but not limited to fire exits chained or otherwise fastened shut illegally, or criminal activity of myself. I further assert that my RFD is not for purposes of concealing or destroying evidence of harassment of a representative of Personal Safety Devices (PSD), or to conceal discrimination against a PSD employee.

Name: _____

Social Security Number: _____

Signature: _____

Right Thumb: +---------------+
| |
| |
| |
| |
| |
| |
| |
| |
| |
| |
| |
+---------------+
(PSD employee to assist in centering thumb print in box)

A NOTE ON PUBLIC SAFETY:

You should understand that public safety (including that of our employees) is for YOUR benefit. For example, by eliminating credit card fraud, the credit card companies are able to continue to charge lower profit margins, which increases your store profits. As you are well aware, your use of surveillance cameras has reduced costs and benefits the customer (e.g. you often use the words "for YOUR protection you are being videotaped"). We appreciate your concern for your customers (us), and wish to return the gift of public safety to your establishment, in a manner in which we can all benefit, through the creation of a utopian world order of zero crime.

----------for PSD use only----------
Image index and type (e.g. v000100.jpg to v000199.jpg, or v123.pic), to be completed by PSD employee prior to submission to Company Headquarters

_____.jpg to _____.jpg _____.pic _____.mpg _____.dat

_____.jpg to _____.jpg _____.pic _____.mpg _____.dat

_____.jpg to _____.jpg _____.pic _____.mpg _____.dat
(use additional page if necessary)

Figure A-2: Back of Request For Deletion (RFD) form

Appendix B

Glossary of new terminology

This thesis contains new concepts and ideas, for which I have found it convenient to introduce new terminology[1]. I have attempted to select the terminology such that it is as self-explanatory as possible. However, to assist in summarizing, I provide this glossary of new terminology.

Groups of entries are alphabetized when practical, but, more importantly, are arranged together (in the sense that they refer to each other).

B.1 Science (mathematical/conceptual)

- 'lightspace': The tensor (outer) product of the plenoptic function with itself, used in the context of a plenoptic camera and a plenoptic light source.
- 'VideoOrbits' (or 'video orbits'): A general methodology for processing video or other sequences of pictures of the same scene or object, where images are processed as though they are related through a projective (homographic) coordinate transformation, possibly in conjunction with a homomorphic gain transformation process.
- 'Wyckoff set': A collection of real-valued functions, f, of one or more real variables, x, in which the functions are related to one another by only a change in scalar, k, in $f(kq(x))$. The most notable example is that of a collection of differently exposed images. The term is named in honor of Charles Wyckoff who formulated an exposure-bracketing film.
- 'Chirplet transform [55]': A signal representation based on analysis primitives that are windowed portions of chirp functions. There are two variations of the 'chirplet transform'. (1) The projective chirplet transform (p-chirplet) in which analysis functions are transitive under the projective group of coordinate transformations. (2) The quadratic chirplet transform (q-chirplet) in which analysis functions are metaplectomorphisms of one another (e.g. related to one another by affine coordinate transformations in the time-frequency plane). (1) was developed for image processing, and (2) was developed for VibraVest, etc..
- 'projective-Wyckoff set': A collection of real-valued functions, f, of one or more real variables, x, in which the functions are related to one another by a projective coordinate transformation on x, together with a change in scalar, k, in $f(kq(x))$.
- 'projectivity+gain group': A group of coordinate transformations in which any members of a particular 'projective-Wyckoff set' are transitive.
- 'Homomorphic homography': A relation among members of a 'projective-Wyckoff set.
- 'spatiotonal': Relating to both the domain and range of a function. Typical examples include spatial processing, combined with tone-scale adjustment.

[1] Single quotes denote terminology introduced by the author here or elsewhere in the literature.

B.2 Technology (WearComp/WearCam, etc.)

- 'Ephemeral criterion': That which requires any appreciable form of delay in the responsiveness of the apparatus to effectively disappear, to the extent that the user experiences a constancy of user interface that is both constant in operation and constant in interaction. Although the apparatus may have 'sleep modes' (e.g. advanced power management), it should never 'die' altogether (as when a laptop computer or PDA is switched off and carried in a pocket or briefcase).
- 'Eudaemonic criterion': That which situates the apparatus within the user's *prosthetic territory* [5] with at least the control point of the apparatus in close proximity to the user irrespective of the user's activity (e.g. while walking around, etc.). For example, the device may be worn in or on clothing to meet this criterion.
- 'Existential criterion': That which provides the user with control over an apparatus, such that the user exerts the principle of self-determination and mastery over the apparatus.
- 'existential technology': Technology which affords the user with increased self-determination or mastery over his/her own destiny.
- 'existech': short form for 'existential technology'
- 'Humanistic Intelligence' ('HI'): A framework for personal connected (networked) intelligence amplification that embraces humanistic ideals. Most notably, it is a technological framework that borrows from humanistic psychology the notion of "self-actualization" [7], and from existentialist philosophy the notion of self-determination and mastery over our own destiny. Loosely speaking, 'humanistic intelligence' is that which is facilitated through the WearComp apparatus as a tool of self-empowerment. Humanistic intelligence is characterized by three facets called the 'ephemeral criterion', the 'eudaemonic criterion', and the 'existential criterion.
- 'WearCam': A personal imaging apparatus in which the camera is affixed to the user in the 'eudaemonic' sense (e.g. worn).
- 'WearComp': A specific embodiment of computational apparatus which is (or at least the control point is) situated in close physical proximity to the user during many of the user's day-to-day activities, over which the user exerts substantial control (e.g. it's not just a monitoring device), and which itself is both interactionally and operationally constant. (These three criteria are referred to as the 'eudaemonic', 'existential', and 'ephemeral' criteria respectively.)
- 'WearComp0': The first series of "wearable computers" built from electromechanical components (such as steppers, motors, relays, etc.), and used for such tasks as sequencing electronic flash and wearable or portable battery-powered photographic lighting apparatus, often in conjunction with wireless communications.
- 'WearComp1': The first series of solid-state "wearable computers" (e.g. those with no moving parts except for cooling fans and the like).
- 'WearComp2': A 6502-based wearable computer system built into a metal frame pack, and used by the author in the early 1980s. WearComp2 comprised analog to digital and digital to analog converters (used for speech input, audio output, radar, etc.), as well as multichannel binary I/O (typically used to take input from a one-handed chording "keyboard" of sorts, and give output to a plurality of light sources). WearComp2 was completely self-contained (did not require a docking station) and "booted" almost instantly. Programs and data were saved on audio cassettes. Display was hybrid (analog NTSC or text with 12 rows of 40 characters each, alphanumeric but uppercase only).

- 'WearComp3': A matched pair of 8085-based wearable computers built for experiments in collaborative photographic and lighting efforts during the 1980s. Since they had no local storage, these computers required a mainframe computer which was a "docking station" of sorts, for startup, but once started, could be used with batteries and wireless communications until the batteries ran down. They could not be turned off and then on again without revisiting the docking station (hence the need to keep them powered up at all times while away from the locale of the mainframe). Separate displays were used for analog video (from a computer-controlled camera), and for programming. The programming display comprised a row of 7-segment display units each capable of displaying a decimal number, or a letter from A through F. Radio communications comprised a pair of two-way DSB-AM radios.

- 'WearComp4': An 80286-based wearable computer. Completely self-supporting (e.g. no need for docking station or the like). Comprised serial and parallel ports as well as floppy disk drive. Input comprised a collection of pushbutton switches (e.g. belt-mounted, mounted on the handle of an electronic flash, etc.).

- 'WearComp5': A 33MHz 80486-based wearable computer. WearComp5 took many different forms, and included serial, parallel, and PCMCIA ports, as well as floppy disk. Display was 480 by 640 pixels (image) or 24 rows of 80 characters. Input device was a belt-mounted collection of pushbutton switches later replaced with a "twiddler" keyboard made by HandyKey corp. Radio communications comprised either a 1987 WA4DSY rig with modified terminal-node controller, a system built around a voice transceiver, using F2D or G2D emission, or a modern G3RUH rig. Antenna typically comprised a resonant whip driven, along RG174 feedline, through a ground plane made from copper mesh. Analog to digital converter comprised an 8 channel unit manufactured by Thought Technologies Limited.

- 'WearComp6': A series of wearable computers based on the PC-104 standard. Input device was typically a "twiddler" keyboard made by HandyKey corp. Radio communications comprised either a 1987 WA4DSY rig with modified terminal-node controller, a system built around a voice transceiver, using F2D or G2D emission, or a modern G3RUH rig. Antenna typically comprised a resonant whip driven, along RG174 feedline, through a ground plane made from copper mesh. These units also typically featured special-purpose video processing hardware.

- 'WearComp7': The current series of covert (e.g. undetectable by others at close conversational distances, etc.) wearable computers, built into undergarments (also known as the 'underwearable'), and used in conjunction with display and imaging apparatus concealed in ordinary dark sunglasses or "mirrorshades". These units typically comprised special-purpose video processing hardware together with TMS320C3x/4x series processors, programmed through a 586 host processor. Programming is typically done through RTTY link to a SUN3 system running a cross-compiler. The close proximity of the 'underwearable' to the body, together with its tight (custom-tailored) fit (e.g. exact fit to my body), permitted various versions of it to be equipped with vibrotactile devices. Other accessories include radar, etc., but if too many features are built into the 'underwearable', it begins to show evidence of bulging, and ceases to be a covert system.

- 'WearComp8': A future proposed (currently under development) generation of covert wearable computers characterized by sleek and slender thin-frame eyeglasses. WearComp8 will be almost completely undetectable even under close scrutiny from all angles. It will use a large number of parallel processors, running from eight separate power channels.

- 'Personal Imaging': That which is facilitated through a WearComp which has visual output modality (comprising a screen over one or both eyes), together with one or more cameras.

- 'underwearable': A system built into an undergarment, typically a tank top of some kind, although other versions have been built into tanksuits and bodysuits. The garments are typically custom-made from materials having high tensile strength.

- 'ThinkTank': A version of the 'underwearable' that contains a computer system, where the undergarment is specifically a tank top.
- 'VibraVest': A vest containing a radar system and any one of a number of vibrotactile devices actuated by the radar system.
- 'Electric feel sensing': Electric field sensing coupled to vibrotactile output devices. Typically the electric field sensing was of the far-field variety (e.g. radar, typically operating at 10.525GHz, 10.250GHz, or 24.360GHz).
- 'Vibrotach': Tactile device that conveys, through the sense of feeling, a quantity of speed, with the sensation of a cyclic motion. Early 1980s embodiments of vibrotach were typically implemented using synchronous motors driven directly from the plates (i.e. with output transformer removed) of a body-worn amplifier, fed with a variable frequency square wave signal. Late 1980s embodiments were typically implemented using D.C. motors driven by solid state amplifiers.
- 'Saltach': Variation of Vibrotach that uses multiple tactile devices to create a sense of cyclic motion. Early embodiments of saltach comprised ten transducers which were activated sequentially, one-at-a-time.
- cyborg: short form for cybernetic organism.
- cybernetic organism: (as defined by Manfred Clynes [94]): A combination of human and machine, in which the human does not need to expend conscious thought or effort in order to sustain the interaction with the machine. (For example, a person riding a bicycle [93] is a cybernetic organism.)

B.3 Art (philosophical/conceptual/critical)

- 'surveillance superhighway' ('SS'): An infrastructure with the potential to be used for omniscient surveillance. Notable examples include cameras instaled by governments throughout entire cities, for purposes of monitoring the general activities of citizens, as in Liverpool or Baltimore. Other examples include consumer electronics devices, such as television sets containing cameras to watch the viewers, which are networked and run proprietary (e.g., "intellectually encrypted") operating systems.
- 'reflectionism': A form of cultural criticism, art, performance, or the like, in which one attempts to appropriate the methodology or characteristics of a problem, and use these as tools, resituating them in a disturbing and disorienting fashion, to allow those responsible or those associated with the problem to understand the problem.
- 'diffusionism': A form of cultural criticism, art, performance, or the like, in which one attempts to appropriate the methodology or characteristics of a problem, in particular a problem involving imbalance of power or the like, and diffuse the very source of the problem throughout society, so that the "problem", by the very virtue of its ubiquity, ceases to be a problem of imbalance. The 'diffusionist' philosophy applies to situations where the problem can be "equalized" through a restoration of "balance".
- 'personal verité': An attempt at defining a new genre of personal documentary, based on the principles of cinema verité, and characterized by the use of personal imaging. Most notably, personal verité attempts to modify the behaviour of the videographer, through a process of long-term adaptation (typically involving some kind of coordinate transformation on visual reality), to the extent that the camera begins to function as a true extension of the mind and body — quite literally as a prosthetic necessary for seeing. Loosly speaking, this is attained by virtue of the fact that the videographer lives "life through the screen", undergoing all manner of real-world visual experience upon a computer screen or the like.

Appendix C

About the preparation of this document

Much of this document was typed, originally in the form of stream-of-consciousness notes, in all lowercase, while walking around, waiting in line at the bank, or the like. This original material was typed on a variety of small battery operated systems as described in Chapter 1 (systems I called WearComp5, WearComp6, and WearComp7), using an editor I wrote myself (as well as some with VI and some with EMACS. These systems were running DOS/NOS and Linux (e.g. configured as dual boot systems), and running Linux/XFree86 most of the time. Partly out of paranoia, however, having had a good deal of hard-drive carnage, most of the documents were edited on the MS-DOS partition of my hard drive, so that on the off chance that Linux would not boot (as often was the case), I could still access the documents. Thus all the source file names were 8+3 characters or less.

It was perhaps a strange and trying experience to be using a hobbyist system I put together myself as a tool to get some real work done. Because of the experimental nature of the very tools I was using, I developed various precautions, such as using the MS-DOS partition as described above, where I could still send through the network wirelessly to other computers (using KA9Q NOS) to back the documents up. Again, out of healthy paranoia (which turned out to be very justified in many cases), I typically kept at least three copies of all documents backed up on three different computers.

This very rough material, observations, personal anecdotes, and the like, combined with other research papers I'd written were further edited, much of the time using a VT320 terminal sitting on my kitchen table, literally plugged into an undergarment — my 'underwearable' computer — hanging up in my closet. Ironically, I originally set up this terminal simply for "rescue" purposes (e.g., for when the device driver for the input system died, or when XFree86 froze up completely and failed to respond even to CTRL-ALT-DEL). However, increased use of the VT320 became necessary, in the final stretches of this effort, as the wearable apparatus, despite many years of development, is still crude enough that a change is as good as a rest[1] (e.g. sitting down in the traditional mode of work, and using a fullsize two-handed keyboard). There's still something to be said for typing with both hands! Not to mention with both feet, as I had also rigged up foot pedals to the keyboard, for shift (left foot), and ctrl (right foot), which I found spared me from ruining my left hand as so many emacs users have done.

Thus this thesis was typed in a variety of locations, from banks, department stores, and airport lounges, to my own kitchen table. As such, it was originally quite disorganized, but fortunately, with some helpful suggestions from my advisor, and various others (see acknowledgements) it eventually took shape.

[1] Despite the efforts to make it comfortable and usable, I found that, in the end, many hours of using WearComp did result in a certain amount of eyestrain, neck strain, etc., and that much remains to be done in order to make WearComp totally comfortable and totally usable for 16 hour-long time stretches that are typical of writing a PhD thesis.

The document itself, typed in three different text editors (my own that I customized for typing while walking or jogging, VI which I prefer for navigating in an existing document when changes are minimal, and EMACS which I preferred while doing serious re-arrangement of text via keyboard macros and the like), was then compiled in LaTeX, using the standard PhD dissertation style file retrieved from MIT's Athena network. The thesis is typeset in the Computer Modern font, at 11 points.

When it was time to print the resulting thesis (PostScript file approximately 50 megabytes), I found it preferable to transmit the source text files to one of my base stations, and then recompile, issuing commands remotely to the base station. Indeed, sending 50 megabytes over the radio is a slow process.

C.1 Figures from a personal imaging perspective

When I remarked I was wearing my thesis, Vaughan R. Pratt once joked that it should be called a prosthesis. In some sense, perhaps he was right — the thesis, written about WearComp was also written using WearComp. This self-referentiality also applies to the pictures in the thesis, which were mostly captured and processed using the methodology presented. Of course the figures that illustrate the techniques were of this variety, but so too were many of the illustrations of the apparatus itself. These pictures were often taken by another version of the apparatus that I was wearing, and often used to capture a collection of differently exposed/illuminated images, upon which the lightspace technique was applied (e.g. optimally combining information from hundreds of different exposures). Thus, in some way, I have demonstrated the utility of personal imaging as a photographic/documentary tool for use in everyday life.

Most of the figures/images were rendered at extremely high resolution and definition for the archives (e.g. for a future in which there is a practical means to display such resolution), and then reduced to far lower resolution, in greyscale (e.g. removing color information) for the final printed version (to keep the final PostScript file size under 100 megabytes or so). However, some day when display and print media exceed the standards of today, I will be able to simply recompile the thesis at a higher image resolution.

C.2 Bibliography

A decision for ordering of the entries in the bibliography, in the order in which they were cited, was based partly the fact that a good number of the references do not have any author associated with them (and therefore appeared strangely when alphabetized), and also that because of the interdisciplinary nature of this work, an alphabetized bibliography looked strange and disjoint, when considered by itself.

Appendix D

Technical details of 'WearComp' and other related inventions

In this appendix, I provide a brief summary of some of the technical details of the WearComp invention, and related inventions. These details are separated from the main body of the thesis, and will become part of the thesis after the filing of related patents.

This appendix will also serve as a "howto" guide for those interested in building WearComp. Most notably, WearComp6 can be easily built by most electronic hobbyists, from off-the-shelf components. THESE INSTRUCTIONS ARE PROVIDED AS GENERAL INFORMATION ONLY; USE AT OWN RISK. NEITHER I NOR MY EMPLOYER ASSUME ANY RESPONSIBILITY FOR DAMAGE TO EQUIPMENT, INJURY, OR DEATH THAT MAY ARISE FROM THIS TECHNOLOGY. Furthermore, I suggest that anyone practicing this art be well versed in possible hazards of faulty wiring, possible hazards of long-term exposure to radio frequency energy, possible eye damage from displays in close proximity to the eye, possible brain damage from long-term usage of the apparatus, possible hazards from reduced attention span, flashback effects, and any other possible hazards that may be related to this technology. This set of instructions is based on an earlier "howto" guide I wrote in 1995, which was, to the best of my knowledge, the first set of instructions on how to build a wearable computer. Originally, this guide was placed on the MIT wearables site[1] but has subsequently been appropriated by Starner who has added details pertaining to use with a Private Eye display[2].

Finally, I point the way to the future, by outlining details of the ongoing WearComp7 project. It is hoped that this vision will provide inspiration for a new era of a truly wearable apparatus that is comfortable and unobtrusive enough to wear in ordinary day-to-day situations.

D.1 Brief history of the WearComp effort

Name	When completed	Processor	Text,Graphics	Where on body
WearComp0	1970s	electromech.	---	back
WearComp1	1970s	SSI, MSI	ATV RS170	back+waist+shoulder
WearComp2	1981	6502	40x12,280x,NTSC	back+waist+shoulder
WearComp3	early 1980s	8085	7segment displays	waist+chest
WearComp4	1989,1990	80286	80x24,640x480	ordinary backpack
WearComp5	early 1990s	80486/33	80x24,640x480	large waistbag
WearComp6	early 1990s	PC104,80x86	80x24,640x480	medium waistbag

[1] At the time I put this on the http://n1nlf-1.media.mit.edu/computing site which was equivalent to the http://wearcam.org/computing site.

[2] See http://lcs.www.media.mit.edu/projects/wearables/lizzy/assembly.html.

WearComp7	mid/late 1990s	TMS320C3x/4x	RS170	underwearable

It is debatable whether WearComp0 or WearComp1 were really computers, as they were specifically designed for control of experimental body-worn lighting equipment and the like, and thus certainly not "general-purpose" computers. WearComp2 was the first system that could be regarded as a general-purpose computer, as it could execute a general instruction set, and even had a BASIC interpreter, making it easy to write programs to edit ASCII text files, or exchange messages (e.g. "email" of sorts), do floating-point calculations, and other things that one might regard as falling in the domain of general-purpose computing.

WearComp3 was much less capable than WearComp2, but at the same time, the WearComp3 effort emphasized small size and better integrating the unit into clothing to some degree. This was accomplished by an early attempt to make the unit more like clothing than like a backpack. Furthermore, WearComp3 marked the beginning of the use of the chest area as a display space that others could see. This design choice arose out of the fact that WearComp3 put more emphasis on computer-supported collaborative work than on the more individual spirit upon which WearComp2 was designed.

D.1.1 Smart clothing

Beginning in the era of WearComp2 and WearComp3, another parallel effort was directed toward 'Smart Clothing', which was characterized by conformal antennas, flexible conductive members, etc., as outlined briefly in Chapter 1. Although a complete general purpose wearable computer system with all of the photographic applications of the time was never realized fully in 'smart clothing', various steps were taken toward building electronic devices of various sorts into clothing. Thus some portions of the apparatus were integrated truly into clothing, while other portions (such as the main motherboard) remained "lumped". Some additional innovations toward 'smart clothing' are discussed in this section.

Shielding in 'Smart Clothing'

Shielding in 'smart clothing' is attained by running a conductive member between other grounded conductive members within BC1, IC1, or C0 into which conductive members are sewn (Fig D-1). Typically the "shield" (set of conductors on either side of the signal-carrying line) is grounded at one end only. Such methodology is useful for microphone signals, as well as video, and the shielding is useful both to protect from interference and to reduce spurious emissions.

Antenna design

The transition from WearComp2 to WearComp3 was characterized by a desire to use it indoors (WearComp2 was designed for outdoor use), where it was necessary to clear doorways, low ceilings, and the like. With the low frequencies used for radio communications at this time (current versions of WearComp use much higher frequencies), the antenna systems were such that the larger the antenna surface area, the better was the performance (e.g. the wavelengths used were much longer than the dimensions of the human body). Accordingly, conformal antennas facilitated the use of the whole body as a radiating surface. It was desired, also, that none of the radiation was used to heat the body. There is much debate as to whether such radiation is harmful, but whether or not it is harmful, it is certain that energy absorbed into the body as heat is wasted. Therefore, the use of BC2 was explored in the form of ground planes.

D.2 How to build a WearComp (WearComp6)

The purpose of this section is to provide anyone who has a moderate amount of skill in building electronic circuits with enough knowledge to build a version of WearComp6, the most recent version of WearComp that is solid and highly reliable, and that does not require any special non-standard devices.

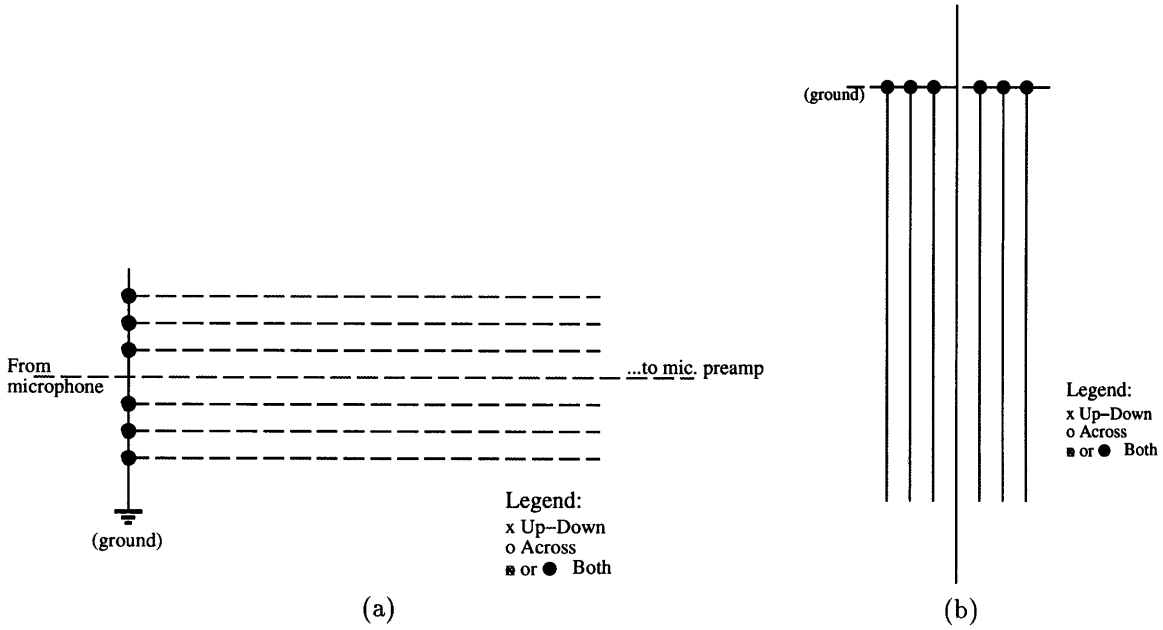


Figure D-1: Shielding methodology used in 'smart clothing'. Notation is the same as that described in Chapter 1 (e.g. solid lines for up-down and dashed lines for across conductors). (a) Low-level signal line for microphone preamp input, taken from LED shirt design. (See Fig 1-7 for rest of diagram but with this portion removed.) Signal-carrying line is flanked by grounded lines on either side. Here I have illustrated 3 ground lines on either side, without loss of generality. (often more are used.) (b) Situation when low-level signal needs to travel in the up-down direction.

D.2.1 Batteries for WearComp

Past/low cost: lead-acid batteries

Early versions of WearComp used lead-acid batteries. Later (Mid '80s) versions used NiCad batteries.

Lead-acid batteries are typically available surplus (e.g. taken out of used surplus equipment or the like) for around $10 each. For constant operation you will want to obtain at least two 12 volt batteries. These batteries typically have lugs that connect to crimp-on connectors. However, in wearable applications, the lugs are easily broken off or shorted (fire/explosion hazard) by stray materials such as keys or tools one might be carrying in a pocket with the batteries. Therefore, I generally soldered wires right to the lugs, and then insulated these very well.

Be sure to place a fuse right next to one of the lugs of the battery, not in the cord going to the battery. The reason for this is that if the fuse is in the cord, something can wear through the insulation on the cord upstream of the fuse, and cause a fire/explosion or the like.

The best fuses to use are the automotive type that have solder lugs. Place a fuse right near the positive lug, as close as possible. Typically one lug of the fuse can be soldered right to the positive lug of the battery. Now solder a red wire to the other end of the fuse, and solder a black wire to the negative lug of the battery. Wrap both lugs in several layers of fiberglass tape and epoxy. It is important to totally encase both the positive lug, and the fuse near it, wrapping all the way around the entire battery for strength, as general wear and tear on wearable apparatus is much higher than for other uses.

NiCad batteries

I do not recommend the purchase of surplus NiCad batteries as NiCad batteries are generally very susceptible to "memory" effects and other possible malfunction. Consequently, those found in salvage equipment are generally found in a state of malfunction already.

A new "battery vest" may be purchased for around $600 (see http://www.nrgresearch.com). This solution has the advantage of providing a ready-to-wear power supply without the need to devise

one's own solution. Furthermore, the vest provides plenty of pockets for placement of computational apparatus, etc., and provides a good means of physical placement of the additional components. These vests are designed for high-current output (e.g. video lights and large cameras), so it is advisable to include an additional fuse of lower current rating, consistent with the actual usage patterns expected.

Alternatively, one can purchase new NiCad packs for under $100 and sew them into a vest or the like. Again, make sure the batteries are fused properly and well insulated as there is an extreme fire hazard owing to their high short-circuit current capability, and the potential hazard is multiplied by the effect of close proximity to the body, and potential difficult of removing the apparatus or undressing quickly enough to avoid being trapped in burning material.

Present/high performance: Li-Ion batteries

In the early to mid 1990s, I began to use lithium ion (Li-Ion) batteries. Most notably, SONY had provided me with camcorder batteries before they were commercially available. However, at this time, in view of the lack of general availability of these batteries, I had recommended the use of lead-acid batteries or NiCad batteries.

However, now that Li-Ion camcorder batteries are commercially available, I recommend their use. You will need a minimum of four batteries (two sets of two in series) for a constant-running 12 volt supply. You can either purchase four SONY NP-F730 batteries (cost approx. 4* $140 = $560 at large department store such as Fry's Electronics where I purchased some recently), or four NP-F530 batteries (approx. 4* $80 = $320).

These camcorder batteries have built in female mini banana connectors. Therefore, to connect to WearComp, which has historically used banana connectors (all versions of WearComp since 1985 have used banana plugs), the following cables are useful (one set for each pair of batteries):

- One white cable, approx. 8 inches long, with a white mini banana plug on each end.
- One red cable, approx. 8 inches long, with a red mini male banana plug on one end, and a red regular-sized female banana socket on the other.
- One black cable, approx. 8 inches long, with a black mini male banana plug on one end, and a black regular-sized female banana socket on the other.

This facilitates connection of each battery in the pair in series (using the white wire), and adaptation to the standard banana connectors of the rig. Alternatively, the adaptor and the power bridge described in the next subsection, may be subsumed into a single entity.

While the choice of connectors is arbitrary, I have advocated banana connectors initially (among small groups of people) so that we can all share common batteries, chargers, etc., and also because they make field repairs simple (e.g. when wires break off while on long trips away from the workshop or lab). However, care is needed, as these connectors should be held together with gaffer's tape or the like, to prevent gradual separation in the clothing, resulting in exposed conductors. I suggest the purchase of three rolls of gaffer's tape in red, white, and black, and the use of appropriate colors to make sure that correct polarity is visible at all times.

In the next subsection, I will explain how the two pairs of batteries are connected together.

In the preferred embodiment of WearComp6 and WearComp7, presently, a large number of very small batteries are distributed throughout the clothing (e.g. more than two pairs of batteries).

WearComp8 will use a conformal polymer Li-Ion battery. These batteries can be totally flexible, and easily incorporated into clothing.

D.2.2 Bridging the power gap

Ordinarily, when a battery is removed from the computer to insert a new one, there is a brief power gap (a time gap between when the first battery is removed and the second is inserted). One or more large capacitors may be used to keep power to the circuit during this time period. However, I found a better alternative, which solved various problems:

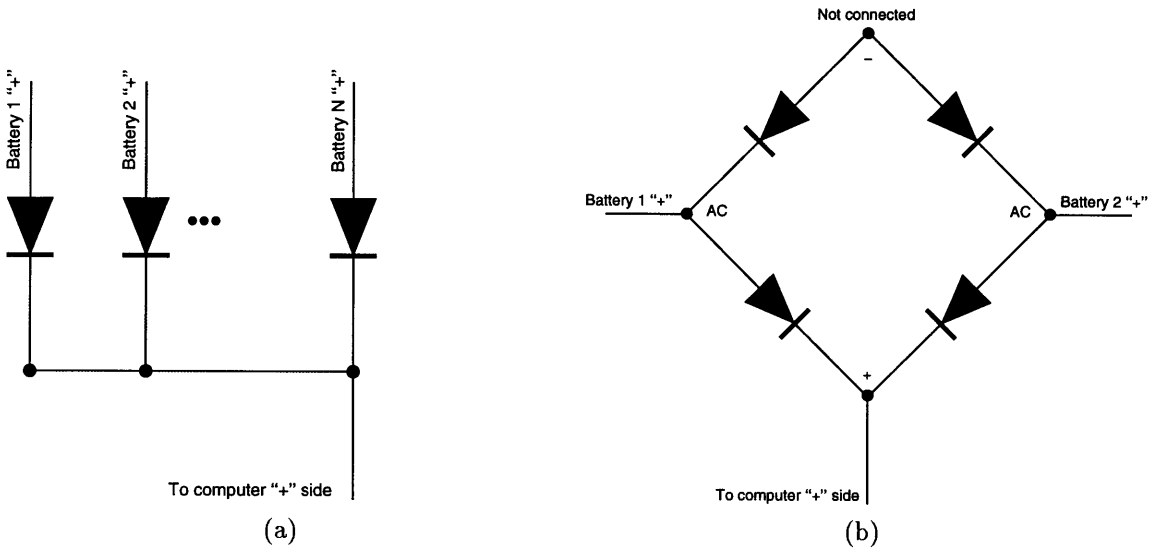


Figure D-2: Combining the output of multiple batteries of possibly different types. (a) A large number of batteries may be combined, and in this way, there is also protection from accidental polarity reversal. (b) Where only two batteries are needed, a commercial bridge rectifier may be used. In this case, only 2 of the 4 internal diodes are used.

1. Allows new battery to be inserted before old battery removed, so that the power gap is bridged.
2. Allows multiple batteries to be bridged together for increased power capacity.
3. Protects against possible damage if battery polarity is incorrect.
4. Allows mixed brands and capacities of batteries to be bridged together without damage resulting from one "charging" the other.

This alternative is implemented through the use of a unit comprised of diodes in series with each set of battery terminals, as depicted in Fig D-2. The diodes dissipate some heat, and must also carry the full current of the maximum anticipated load. Thus it was found that a bridge rectifier, by virtue of its larger surface area, etc., could dissipate the heat, and also be easily sewn into the clothing.

D.2.3 Building the bridge

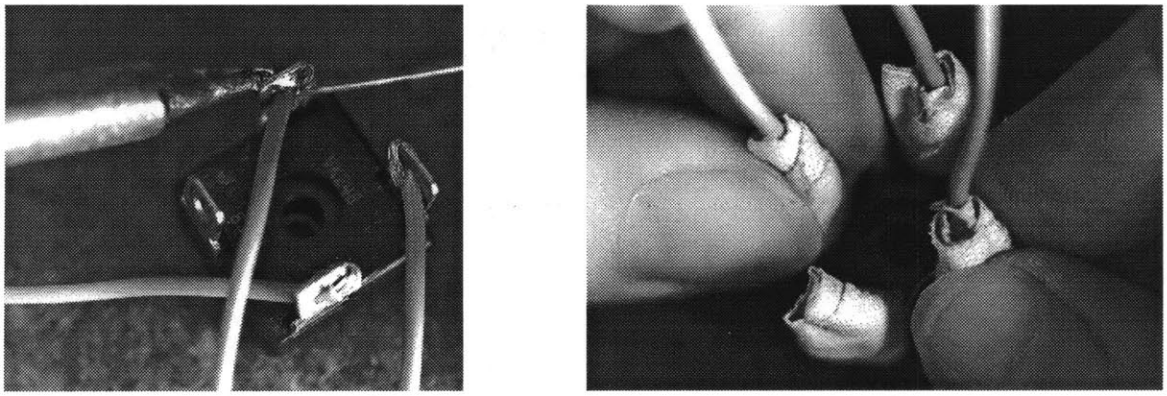
Cut three lengths of sufficiently thick red multistranded wire. Solder ends of these wires to the bridge rectifier, pins "+", "AC" and "AC" (e.g. the two "AC" pins are identical). See Fig D-3(a). Break off or insulate the fourth pin ("-").

Connect a red male banana connector (plug) to each of the wires going to the "AC" terminals, and connect a red female banana connector (socket) to the wire going to the "+" terminal. In this way the power bridge will be modular (e.g. easy to take out or insert at will, depending on usage requirements).

D.2.4 Voltage regulators

The weight, for a given energy level, is much less for Li-Ion batteries compared to lead-acid and NiCad batteries, but the output voltage of Li-Ion batteries varies widely, and drops significantly, with usage from a full charge. Lead acid batteries exhibit this nonconstancy of output voltage to some degree (compared to NiCads which are much more self-regulating), but Li-Ion batteries are far worse in this regard, and therefore, almost certainly, need a voltage regulator.

Another reason that a voltage regulator is needed is that various components of WearComp require different voltages. Typically the computational apparatus requires 5 volts while the analog



(a) (b)

Figure D-3: Combining the output of two batteries of possibly different types. (a) Solder three red wires to all but the "-" pin. (b) Wrap pins and wire in fiberglass tape and epoxy to protect and insulate them. Note that the fourth pin, with no wire connected, is still insulated.

video circuits, and the RF components require 12 volts. It is desirable that a single battery power the entire rig.

With the exception of WearComp0-3, all current versions of WearComp use 12 volt batteries[3]. The original reason for this voltage selection arose from the automotive battery voltage standard, so that WearComp could be operated from an automobile cigarette lighter or accessory outlet fitted with a long cord, either for testing, or for additional runtime when the batteries were low. Furthermore, because much of the peripheral radio equipment operated at 12 volts, this voltage was convenient.

Accordingly, a single "12 volt" battery is used to power most of the apparatus, together with a voltage regulator to bring the 12 volts down to 5 for powering the computational portion of the apparatus.

A linear voltage regulator is undesirable, due to the dissipation of excess heat, since much more efficient switching regulators are available.

Regulators may be compared by:

1. linear versus integrated switching regulators (ISRs);
2. isolated versus 3 terminal

The most efficient are the non-isolated stepdown ISRs.

Furthermore, the voltage variation of Li-Ion batteries is typically excessive for certain components, which require exactly 12 volts, so it is often desirable to have separate switching regulators, one to provide 5 volts, and another to provide 12 volts. I generally use a so-called "step down" regulator to provide 5 volts for the computational apparatus, and a 12v to 12v regulator to take in the varying battery voltage and provide a fixed 12v output for other devices (video, radio, etc.). Furthermore, it is often desirable to use separate regulators for individual components, so that they don't affect each other. (e.g, I often use more than one 12 volt to 12 volt regulator, so that, for example, when the radio transmitter keys up to transmit a packet of data, it doesn't affect other 12 volt components).

D.3 Specific details about how to build WearComp6

WearComp6 is built from standard PC104 modules, which may be purchased from Ampro (www.ampro.com), as well as a large number of other vendors. The PC104 modules are small-sized low-power-consumption

[3] WearComp2 used a 24 volt battery at one point, after which the design was changed to operate from a 12 volt battery. WearComp3 used a 4.8 volt battery comprised of four large NiCad cells connected in series and fixed to a belt.

components that stack together. To build WearComp6, you need to purchase the desired PC104 computer modules (which modules you purchase depends on desired functionality), desired hard drive(s) (again, depending on desired capacity, etc., you may wish to purchase one or two), case, etc.. Each of these items is described in the corresponding subsection below.

D.3.1 Power supply

Isolation is not needed, therefore I have chosen to use a nonisolated (e.g., "3 terminal") integrated switching regulator. In particular, I selected the PowerTrends PT6302 (3 amp ISR) which is much more efficient than the isolated regulators (e.g. Datel, etc.). Not only does this result in extended battery life, but also much less heat is produced by it.

WearComp6 is generally built from the Ampro CoreModule, together with various other modules. Most of the other modules do not have a power connector; power is connected only to the CoreModule, and the other boards derive their power through the interconnecting pins. The CoreModule has a 10 pin (or on some older versions, an 8 pin) power connector. The power connector provides both 5 volt and 12 volt connection terminals. However, most modern boards do not require the 12 volt connection, so you generally only need to connect 5 volts to the core module.

It is generally worth the extra money to get the CoreModule development system (e.g. the version that comes with all the connectors), especially if this is the first unit you build. Subsequently this gives you time to track down the sources for the various connectors, yet still lets you make sure you have a "stock" reference system to compare against cables you make up yourself.

Included in the CoreModule development system, you will generally find the power connector (e.g. MX40 or the like), with a 10 (or 8) pin female connector — 2 rows of 5 (or 4) to mate with the header pins on the CoreModule. See Fig D-4.

You can cut off the 12 volt wires. You might also be inclined to think that some of the 5 volt wires are redundant (e.g. there are 3 pairs of wires for 5 volts). However, it is important to use all 3 pairs; I found using all 3 pairs gave rise to greater system reliability. Furthermore, position the ISR in such a way as to minimize the lead length going to the power connector. The lead-length and actual layout will depend on the specific enclosure you build or purchase.

Originally, I built my own enclosures using sheet metal and a metal bending machine[4]. If you have access to a metal bending machine, this is quite easy to do; first draw the spread-out design on paper, then glue the paper to sheet metal (typically aluminum), and cut with the machine, then bend appropriately.

Here, however, I will illustrate putting together a system using a commercial off-the-shelf enclosure, for the benefit of those who do not have a metal bending machine or the like. The most suitable enclosure is the so-called "half cube enclosure" which can be obtained from Enclosure Technologies Inc (ETI), distributed by Tri-M (http://www.tri-m.com/). Tri-M also sell many other PC104-related products.

With the "half cube enclosure", you can easily keep the power cables 2 inches or less in length. (I found, for example, that the original 6 inch power cable was unreliable due to this excessive length.) The connection from the power cable to the ISR is shown in Fig D-5.

The next step is to mount the ISR inside the enclosure. The reason for mounting it solidly inside the enclosure is twofold:

1. This prevents it from being jostled around where it may touch and short other components. Even if wrapped in insulating material, it could move around and obstruct airflow. It is important when building the PC104 system to keep the insides as neat and tidy as possible so that there can be good ventilation.

[4] In our lab, this is located in the basement machine shop. Subsequent to my building a PC104 enclosure from sheet metal, others in the lab have also been successful in also building similar enclosures, e.g. Jeremy Levitan (see acknowledgements) has built a couple of such enclosures for Ken Russell. Levitan is the "local expert" on the use of the metal bending machine (and on the use of the machine shop in general). If you have never worked with a metal bending machine, it is a good idea to find a similar "local expert" who has the patience to teach you this art, and first practice on some scrap metal to become proficient in the use of the machine. This is a simple skill to learn, and will prove quite valuable.

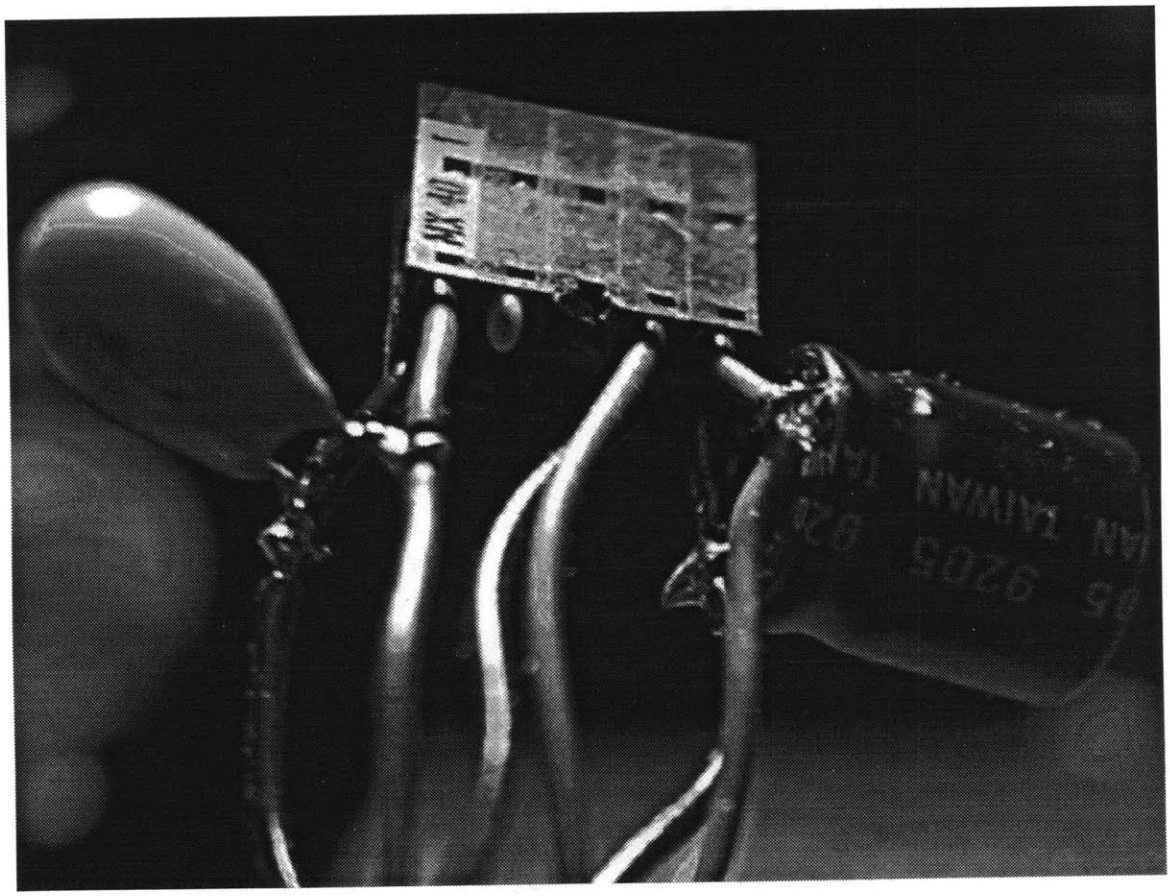


Figure D-4: The power connector mates with 10 pins (some versions are made to mate with only 8 pins). Here I have used all three pairs of 5 volt wires. Note the key pin (right next to the "MX 40" designation) which is marked by a triangular "arrow". Note also the two additional capacitors (one electrolytic, and one tantalum which has lower effective series resistance than electrolytic) which I have added as close as possible to the connector. These are optional, and arise from my healthy level of power spike paranoia.

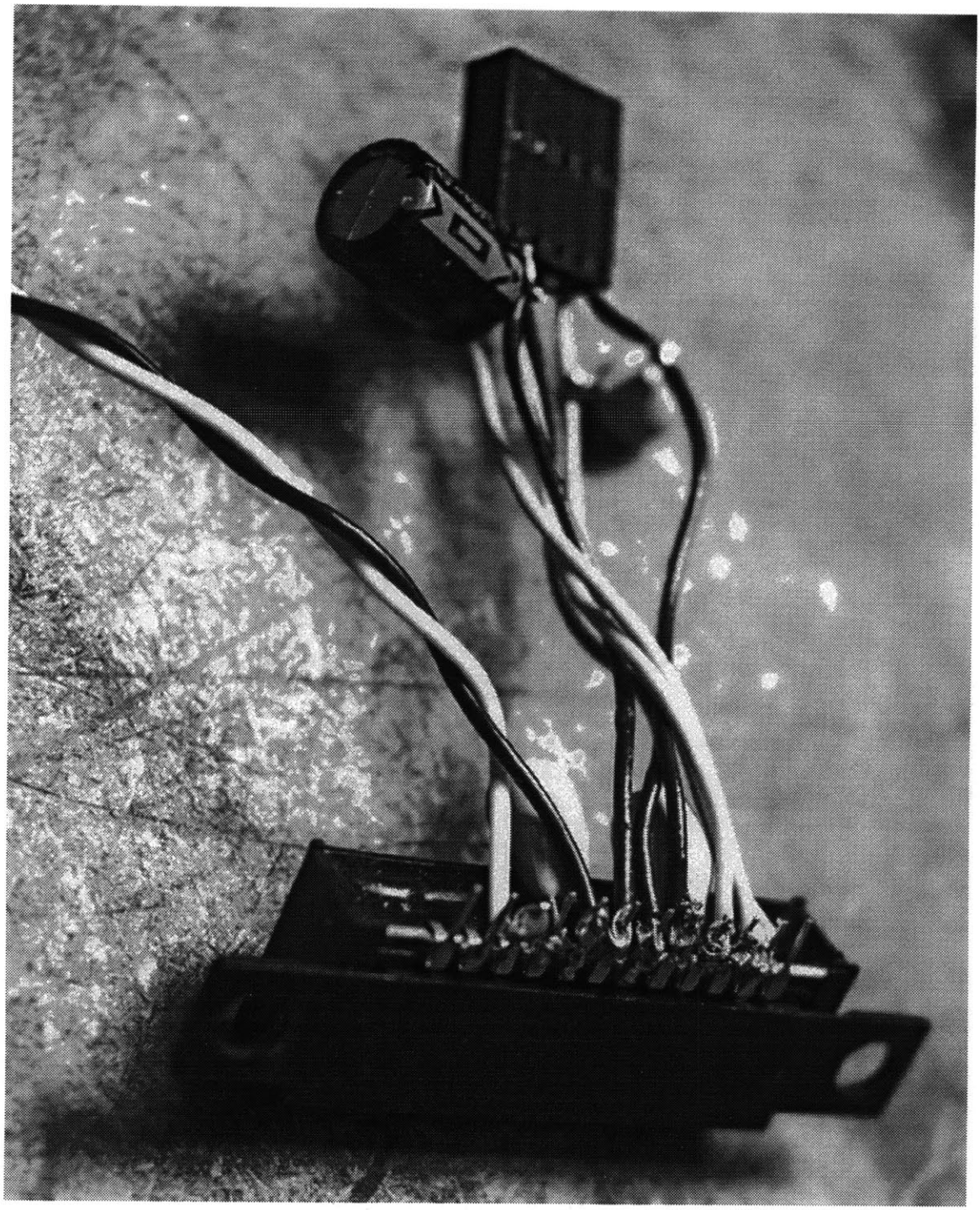


Figure D-5: All three pairs of 5 volt wires may be connected to the various parallel pins of the pt6302 ISR. In particular the ISR also has three "redundant" +5 volt pins. Connect one of each red wire from the power cable to each of these. The ISR has four "redundant" ground pins. Connect the three black wires from the power cable to three of these. That leaves one ground connection for the 12 volt input to the ISR (higher voltage and correspondingly less current). Connect a single twisted pair of wires to the input (conductors do not need to be so thick owing to the lesser current, as well as the fact that the ISR will make up for line losses). Make sure that the twisted pair of wires has tough insulation as this will be outside the enclosure and subject to wear and tear. Here I used a 100μf output capacitor and a 47μf input capacitor with leads soldered to the appropriate pins of the ISR for additional filtering. It is desirable to select an input capacitor which has a high enough voltage rating to match the range of input voltage that the pt6302 can handle, since this will allow you to run the rig on a wider range of input voltages.

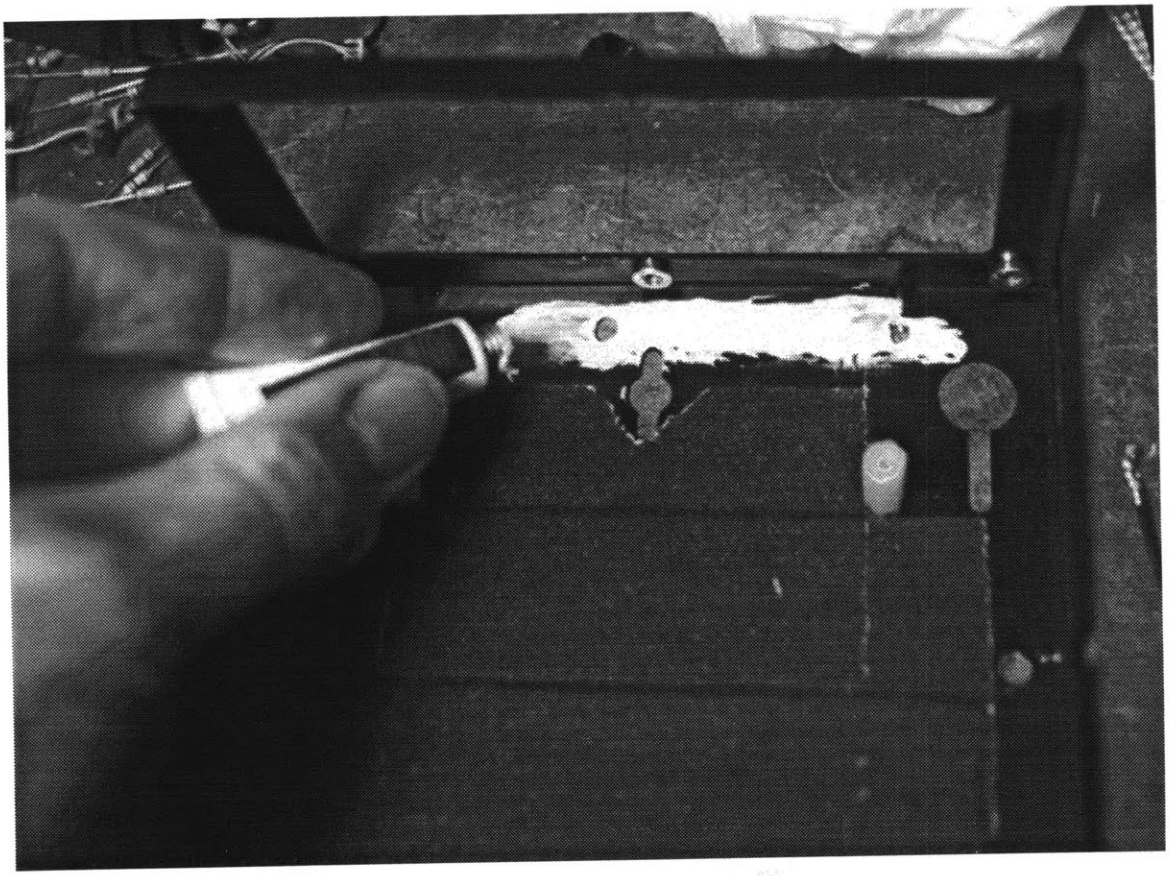


Figure D-6: Once the bottom of the "half cube enclosure" is lined with cloth tape, heatsink compound is applied where the ISR will go. Note that I have removed the front of the enclosure (held on with six screws) for easier access later on when it comes time to insert the ISR.

2. Attaching it to the inside of the case will help with heat dissipation. Depending on what components you are using, this may or may not be an important issue, but in any case, a cool ISR will operate more efficiently.

The optimal place to mount it in the 'half cube' is on the bottom of the enclosure, near the front, and toward the left. This location was selected for three reasons:

- Proximity to power entry point on board stack (keeping leads as short as possible).
- The bottom is the thickest and largest piece of metal, and therefore the best heatsink.
- Best choice of location for space, e.g. to leave open access to all other connections.

In all three regards, the selected location was optimal (e.g. it was not necessary to make a compromise).

Begin by marking and drilling holes for the ISR. Once these holes are drilled, and once all other holes that you think you will want in the case are drilled, clean off all debris (metal flakes, etc.) and proceed to put the nylon standoffs into the case (for anchoring the board stack). Other holes you may wish to drill are wire tie holes for mounting the hard drive (read ahead to next section). Line the bottom of the case with heavy cloth tape, leaving space for the ISR (this is more healthy paranoia — just to make sure nothing could short to it). See Fig D-6. Once you have proceeded this far (lining the bottom) you should not drill any more holes in the case, or debris (metal flakes, etc.) may become stuck to the cloth tape. Next apply heatsink compound to install the ISR.

The pt6302 ISR comes in six variations, with and without mounting tabs (select the one with mounting tabs), and each of these comes in three variations (horizontal mount, surface mount, and

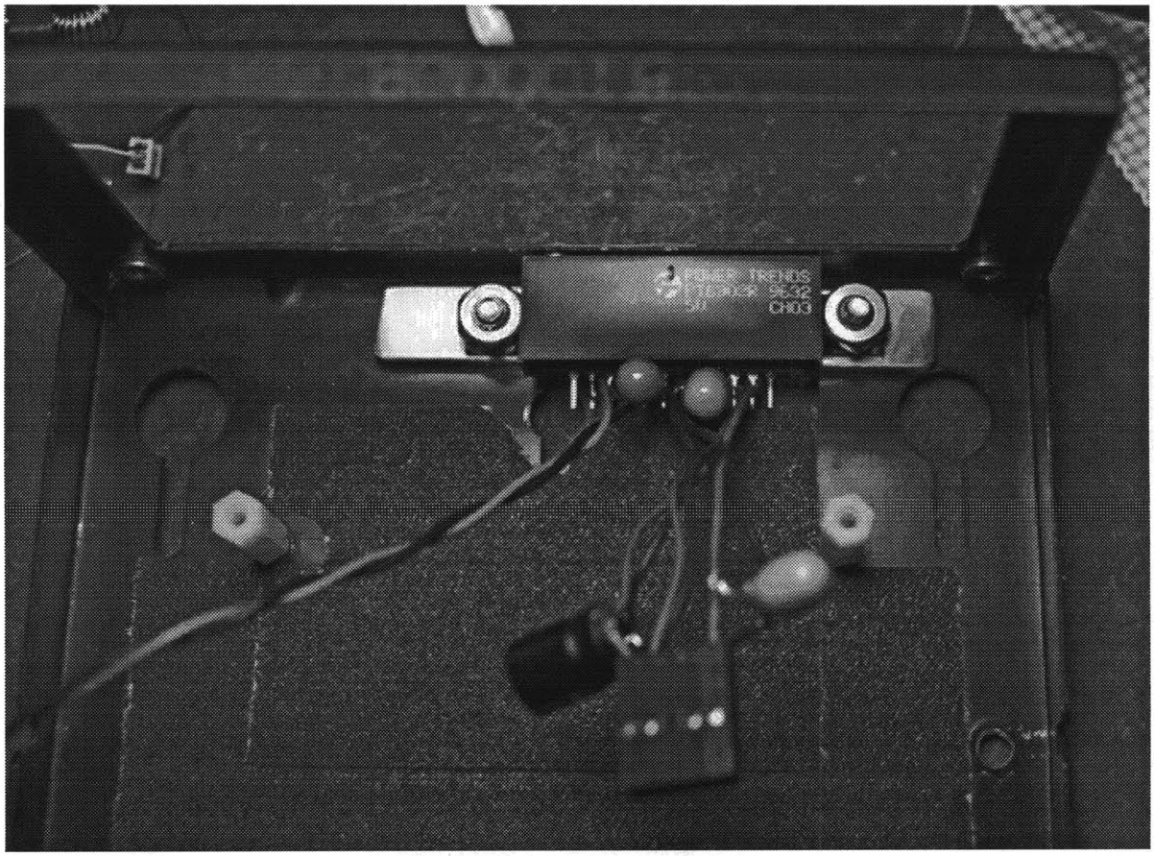


Figure D-7: The pt6302 ISR is installed near the front of the enclosure, facing inwards. Locate it so that the power cable emanates from directly below where the power connector is located on the PC104 CoreModule. Note the aluminum shim I have bolted underneath the ISR to keep the pins from touching the case. Be careful not to locate objects near pin 12 (the sense pin) of the ISR. For example, if the disk cable comes too close to pin 12, stray emissions will affect the computer and make it unreliable. Touching pin 12 when the computer is running will generally cause a spike of sufficient strength to reboot the computer. If you don't need it, you might consider breaking it off or cutting it short so it doesn't act like a receive antenna.

vertical mount). Vertical mount is preferable, but often out of stock. The most readily available is surface mount, and this would otherwise create a problem as the pins would touch the case, but a small aluminum shim will fix this problem and keep the pins sufficiently far away from the case. Fig D-7 shows the ISR installed with a shim I made from 1/8 inch aluminum.

Bring the 12 volt power leads out of the enclosure, thread through a ferrite bead if you like (more healthy paranoia), and then solder on banana plugs (red and black) for connection to power later.

D.3.2 Hard drive

Assuming you are using the Ampro 100MHz 486 CoreModule, the best place to put the hard drive is on the bottom of the case, assuming you have 3 boards or less in the stack. If you have four boards, which is the maximum you can fit in the case, then put the hard drive on its side to the left of the board stack, but then you will not be able to get the case closed all the way, and this puts the hard drive in possible jeopardy if extreme forces are applied to the case. It is far better to have a 3 board stack and get the case properly closed (I've even sat on top of my case with my full body weight, and not had trouble in this regard). It is important to have the hard drive inside the case. Otherwise it can easily be damaged (e.g. if the rig is in a lumbar pack and you sit down on it, or in a backpack and you lean back on it, the hard drive can be damaged if it is outside the case).

Cover the circuit board side of the hard drive with more cloth tape (thick gaffer's tape works best). Assuming you are placing the hard drive on the bottom of the case, put it upside-down in the

case, and use a piece of insulated stiff (single-stranded) wire to "wiretie" it down. (See Fig D-8.)

With the hard drive underneath, you can make a straight run to the header on the CoreModule. Therefore, you can shorten the ribbon cable appreciably (shorter cables all-around make the insides of the rig much neater, and result in greater reliability and improved air circulation). I left the second hard drive connector accessible. The second connector may also be left protruding outside the case if desired. This makes it quick and easy to do backups or copies (e.g. to help someone else get a system up and running) onto a second hard drive.

D.3.3 Assembling the computer

I found that the Ampro VGA board did not properly support 24 bit true color. (Even though it purported to, in hardware, it lacked the appropriate device drivers to do so.) Therefore, I have generally used a VGA board from another vendor, most commonly, Advantech, located at www.advantek.com — note the difference in spelling between their company name and their domain name. This board uses the Tseng4000 chip which is fully supported in linux. It works well with both SVGA lib and in XF86. I prefer standard VGA displays over esoteric displays such as the Private Eye, because this makes debugging and testing easier, and allows for a greater degree of interoperability. Specifically, I find that the Private Eye gives me a headache over extended usage. (I tend to wear my rig sometimes more than 16 hours a day, over several weeks.) Since the Private Eye is a binary red-only display, it is not well suited to personal imaging applications (and the color is part of the reason it gives me a headache). The Private Eye is also difficult to obtain, owing to its esoteric nature (e.g. it is not manufactured in large volumes).

Over the last 20 years of WearComp, various display standards have come and gone, and one standard that has remained has been NTSC. Modern versions of WearComp are leaning toward use of NTSC displays, which tend to have very good color rendition, so most of the recent designs use full 24 bit color. I will discuss NTSC versus VGA later. In any case, both NTSC and VGA are likely to remain for some time, and are good choices as display formats.

Once you have decided which boards to assemble, lay these out on a clean surface. Be careful not to get small blobs of solder, metal flakes, or the like, on the boards, since the very fine traces are quite susceptible to short circuits. Also, if you have not had experience pulling the boards apart without bending the pins, you may wish to plan ahead to minimize wear and tear. For example with the 100MHz 486 CoreModule, connect the hard drive cable (and set appropriate jumpers) prior to assembling the stack together, as it is inaccessible once in between boards.

The boards are easy to assemble. The fact that I have a visual record of this assembly is yet another example of the utility of personal imaging — much of my work in this area is documented from the first-person perspective of the apparatus I wear. For example, the assembly of the first version of WearComp6 was documented by WearComp5 which I was wearing at the time. This provided a video sequence showing the assembly procedure. In Fig D-9, I have selected six frames from this video sequence in order to illustrate the assembly of WearComp6.

D.3.4 Installing the computer in the case

The computer is now ready to be inserted into the case. The easiest way to do this is to first unscrew the metal plate at the back of the case (two screws, facing the outside of the case, are removed) which has the slots for the boards (otherwise it is very difficult to get the boards in), take it out, put it on back of the board stack, and then insert the entire stack in. The plate is then screwed back on. The board stack may be held in place by using other screws to screw into the nylon standoffs (or two more standoffs themselves may be used as screws).

The rig is now (as depicted in Fig D-10) ready for attachement of the rest of the connectors, speaker, power indicator LED, etc., and then put the front plate on. Shorten serial and parallel cables when possible.

The beeping speaker is annoying to others (e.g. in meetings, etc.), so consider using an earphone jack instead. Alternatively, I use a step-up transformer (e.g. to generate a mild electric shock to enable me to "feel" the beep), or a vibrotactile device.

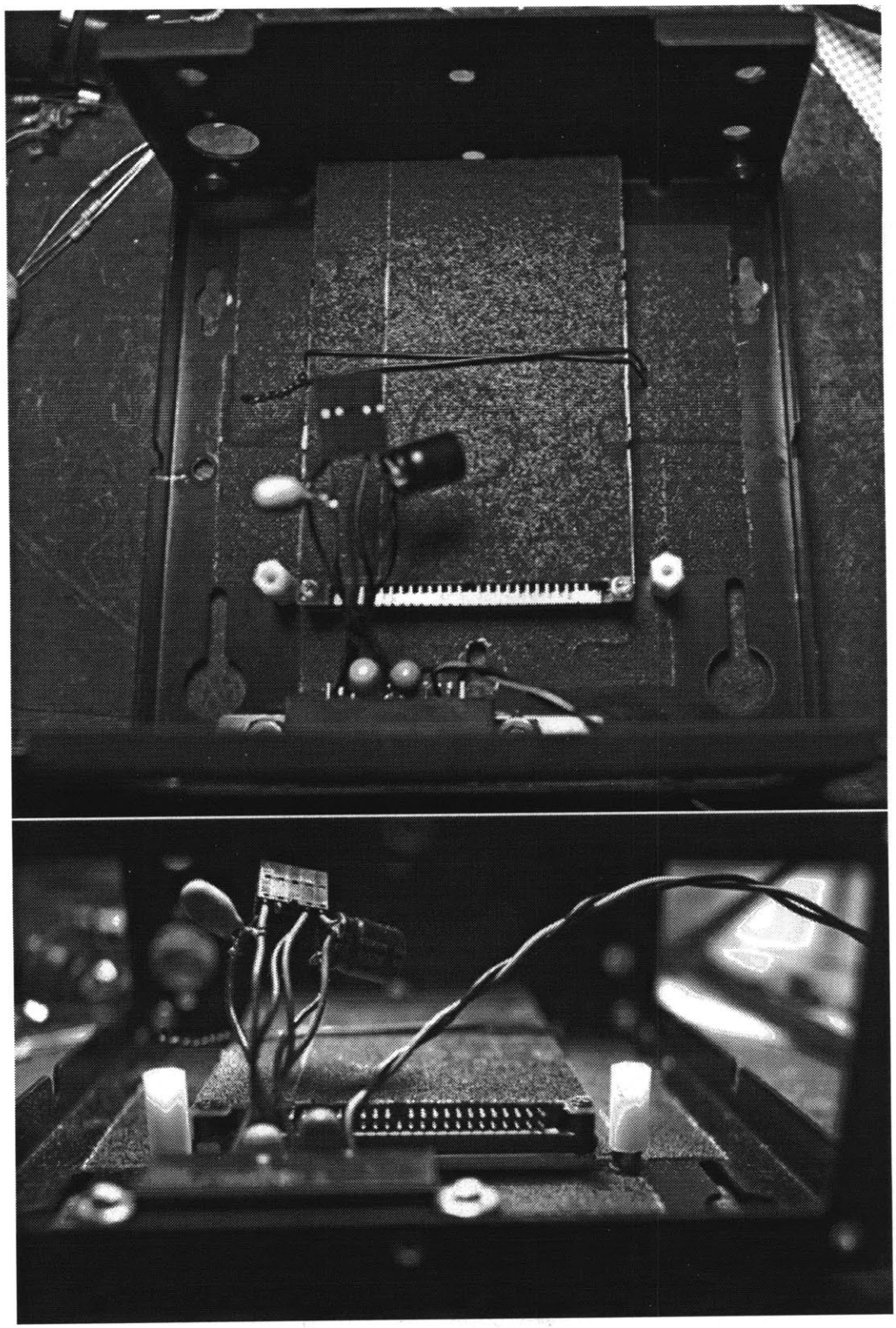


Figure D-8: Next, the bottom of the hard drive is covered with cloth tape prior to mounting it upside-down in the enclosure. It is preferable to "wiretie" it to the bottom to keep it from moving around. Alternatively, you may wish to use angle brackets and the appropriate mounting hardware. Placement is such that it fits under all the boards in the stack.

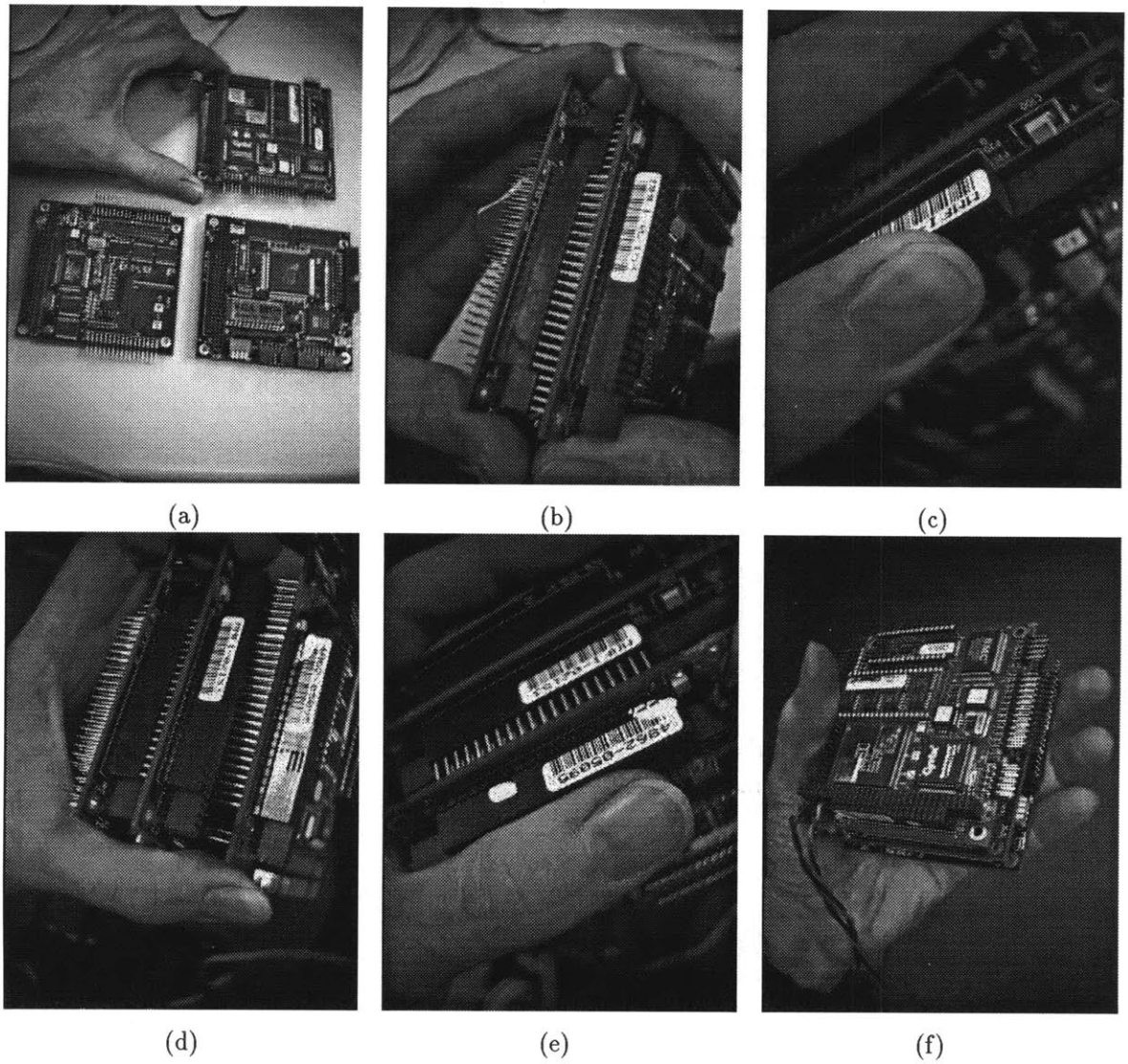


Figure D-9: Decide which PC104 boards you wish to use. Generally the Ampro CoreModule is selected, together with the Advantech VGA board, and perhaps a third board (such as video capture board, sound board, analog to digital converter, etc.) for some additional functionality. (a) Pictured here are boards from an early version of WearComp6 that required a separate floppy disk and IDE controller. Newer core modules include this functionality, so that you may only need two boards (CoreModule and VGA), unless you wish to have extra functionality. (b) Put the first two boards together; carefully insert pins from one, into the other. It is a lot easier to insert (put together) than to remove (take apart) without bending the pins. Therefore, prior to insertion, decide on the odering of the boards (e.g. which should go on top). With experience, you will learn how to pull apart boards without bending the pins, but you should either practice on old boards, or plan carefully so pulling apart is not necessary. When I am using a video capture board, I put that on top because it generates most heat compared to its level of sensitivety to heat (tendency to overheat). Consider also which board you will want easiest access to (e.g. in case you need to change jumpers). The video capture board is the most troublesome in this regard, hence another reason for putting it on top of the stack. If you have only a 2-board stack, consider putting the CPU on top. I usually put the VGA board on the bottom of the stack because it is the cheapest board, and the one for which I am most willing to cut off the bottom pins. You can save space in the whole stack by cutting off all the pins on the bottommost board. Be sure to think carefully and test carefully prior to this committment, as this commits you to making that board the bottom board from then on. (c) Once pins are aligned, press the first two boards together. (d) If a third board is going on your stack, align it next. It is easier to add one board at-a-time than it is to press together all three. (e) Press the new board together onto the rest of the stack. (f) You now have a battery-operatable multimedia computer in the palm of your hand. Test it thoroughly for functionality in your selected board-ordering before cutting off the bottom pins.

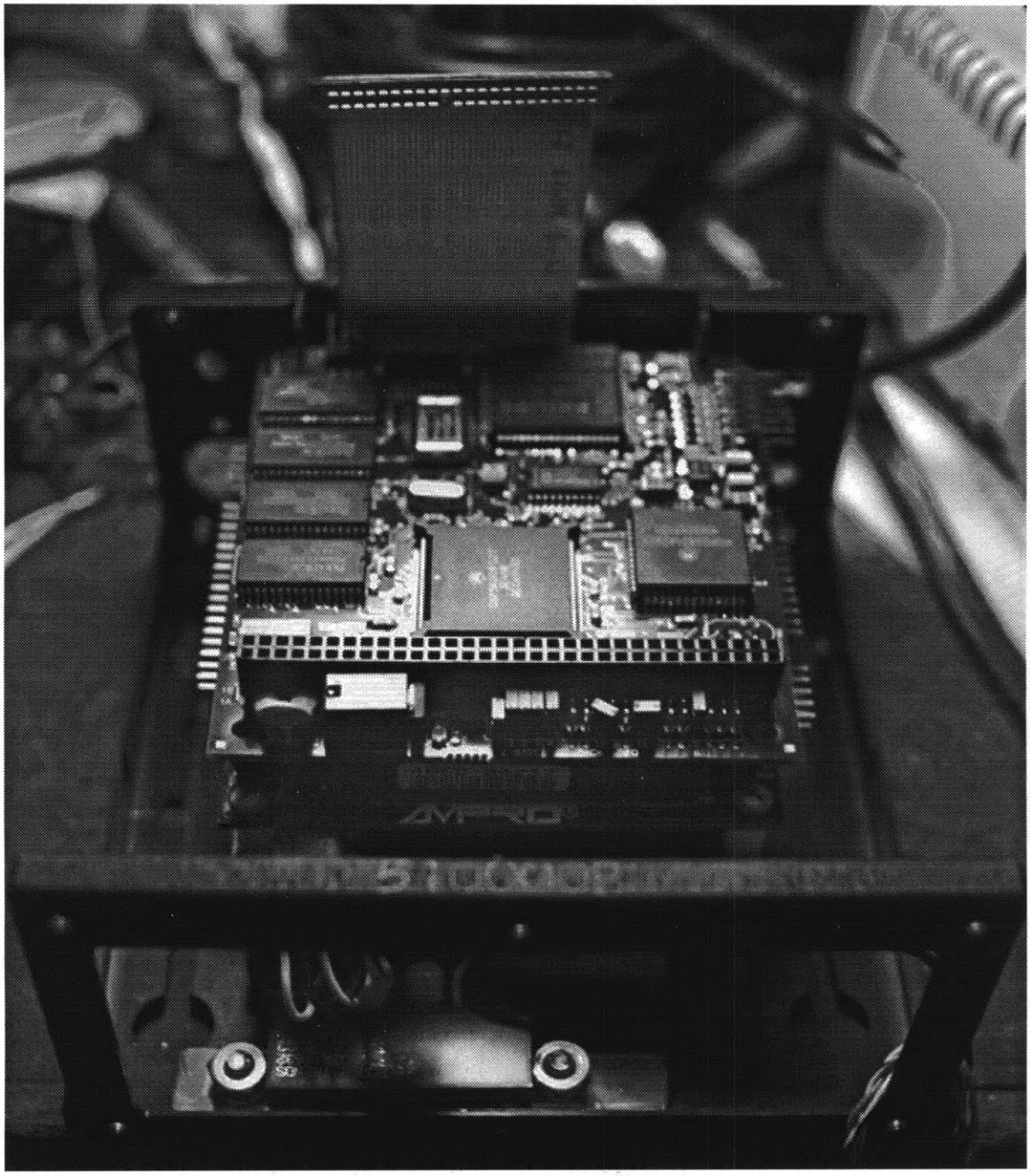


Figure D-10: Connect hard drive cable to hard drive on bottom of case, connect other end in board stack (to CoreModule), and insert the board stack into the enclosure. Plug in the power connector. You are now ready to connect the serial cables, parallel cable, keyboard, speaker, etc..

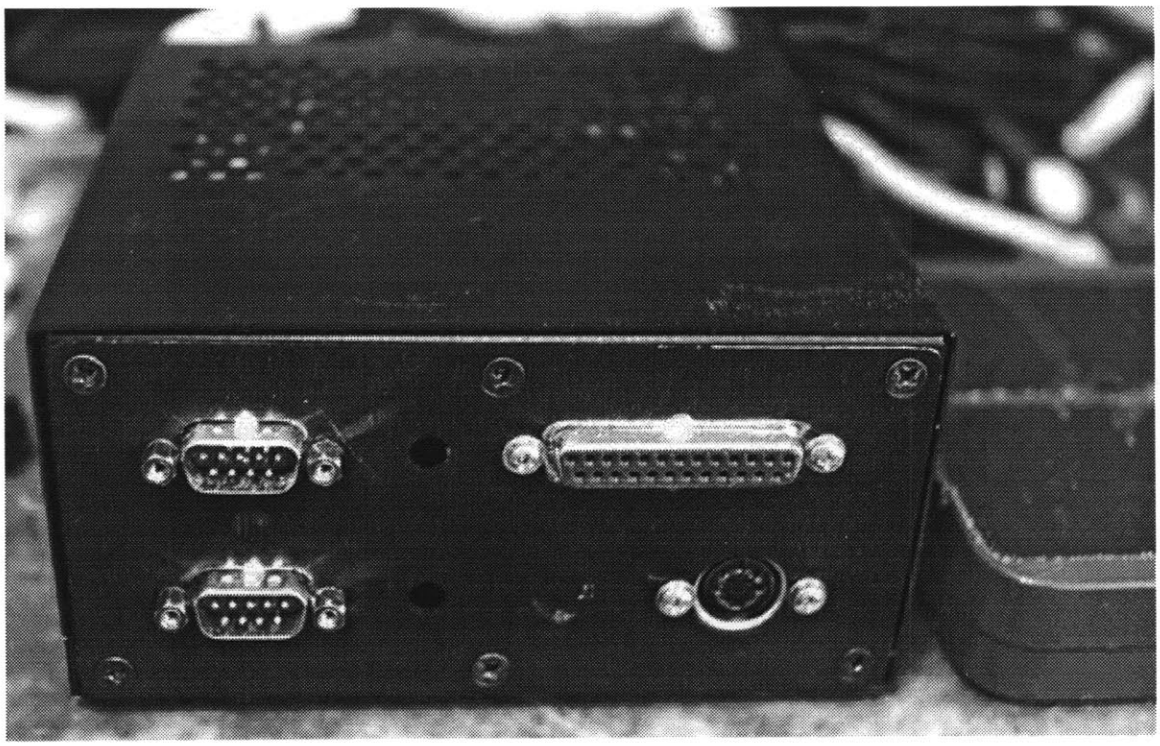


Figure D-11: Completed WearComp6 on workbench next to VGA to NTSC scan converter.

D.3.5 Case closed!

You are now ready to close up the case. First replace the front (held on with six screws), while routing the cables to the appropriate connectors. Then put the lid on. You are now ready to plug into a VGA desktop monitor or VGA head-mounted display. The finished computer may be operated with either a standard keyboard, or with a hand-held keyboard. In wearable operation, I typically wear the computer in a waist bag or lumbar pack. (The Mountainsmith daypack or tourpack is appropriate for the computer and a good collection of other materials.) By connecting it to a head-mounted display, and plugging in a hand-held keyboard (such as the twiddler described in Chapter 1 — see http://www.handykey.com), you have a computational environment that you can carry with you and use while walking around in ordinary day-to-day situations.

Fig D-11 shows the completed WearComp6 on my workbench, next to a VGA to NTSC scan converter (described in the next section).

D.4 Video for your head

The limited availability of VGA head mounted displays at reasonable cost suggests NTSC as a possible alternative. Indeed, early versions of WearComp have used NTSC, and there is a long history of availability of NTSC displays. Most notably, camcorder viewfinders may often be salvaged from broken camcorders and built into eyeglasses. I commonly obtain these units for under $20, so this is clearly the lowest cost solution. Larger tubes (such as some that I still have from 15 or 20 years ago) often last for many years and provide good resolution. There is a common misconception that NTSC resolution is significantly less than VGA. This misconception arises from cheap game displays and consumer television both of which have poor resolution. However, good camera viewfinders can have as much as 1000 vertical lines of resolution, and can therefore adequately display VGA resolution images or text. Some experimentation is needed because text modes in VGA are often not 60Hz, but many camcorder viewfinders will sync at 60 or 72Hz.

D.4.1 Transition from WearComp6 to WearComp7

A VGA to NTSC converter is useful in making the transition from VGA computers to NTSC computers, because it will allow you to move toward NTSC displays, yet still use these with the older generation of VGA computers. Accordingly, I describe how a low cost converter can be adapted to use with WearComp6.

Begin by purchasing a "Pocket Scan Converter" from AiTech (cost approx. $129). This unit consumes considerable power, owing to a very inefficient regulator inside. However, the efficiency can be roughly doubled by removing this regulator, and replacing it with a PowerTrends ST105VC integrated switching regulator (ISR).

The "pocket scan converter" is held shut with a single screw, which is hidden under one of the labels on the bottom. Peel back and unscrew (Fig D-12). Once the screw is removed, pry open and take out the circuitboard. Locate the offending 7805 regulator (Fig D-13) and remove the screw holding it in.

Now desolder the 7805 and install the ISR in its place. (See Fig D-14.) There is plenty of room inside the scan converter case for the larger ISR, even though this is a very small-sized scan converter.

Connect red and black wires for 12 volt input, route through case, and re-assemble. Add red and black banana plugs, and you are ready to use the scan converter together with the computer. This combination may be used with a low cost wearable television set (such as VirtualVision), or with a high quality camcorder viewfinder. If you are using the lower resolution TV set (like the standard VirtualVision unit), then you may want to run XF86 with increased font size (e.g. 30x12).

D.5 WearComp7: getting the fit right

WearComp7 is characterized by NTSC output and a small wearable television set built into a pair of sunglasses. This system is not suggested for the typical hobbyist, as building it is quite involved. WearComp7 is an ongoing (as yet incomplete) effort.

The goal of WearComp7 is to make an unobtrusive (e.g., covert) version of WearComp (while also designing it to be comfortable enough to wear in most day-to-day situations). The secret to success of WearComp7 is in getting all the components to fit within the space of ordinary everyday familiar objects (like ordinary sunglasses and ordinary clothes). The underwearable has already been described, so the rest of this section pertains to the unobtrusive eyeglass-based television set.

D.5.1 Imaging of the head

The first step in fitting all the components into the sunglasses is to determine exactly how much space there is between the glasses and the head. This is done through making first a "negative" 3-D copy of the head, and then making the "positive" from that. Without loss of generality (e.g. there are many possible ways to make the "positive"), I describe one means of making a positive in plaster.

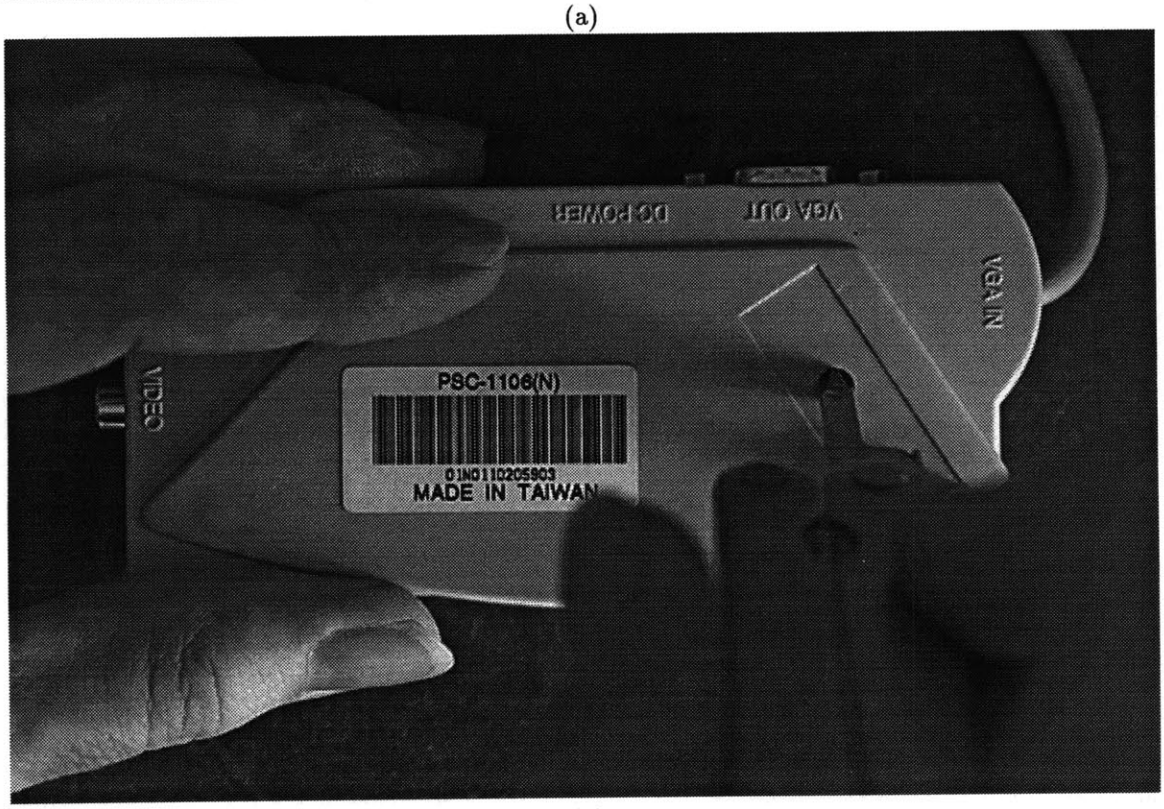
Direct application of plaster to the body is unwise as it will stick to hairs and other features, and could be difficult to remove. A blue-green algae powder, or similar material forms a buffer between plaster and the head.

It is difficult to image the head by yourself, as you can't see from your own perspective (eyes will be covered, etc.), thus it is better to have at least one other person work with you. Ideally, 2 other people do this (one to hold, the other to mix the materials, as timing is critical).

Generally all hair is shaved off first (although one can use a bathing cap, to avoid the need to shave the head, if one is careful to later account for this dimensional error through filing down the positive).

Assuming that the glasses would be worn and used while standing or sitting, it is essential that the "positive" be made while standing or sitting (e.g. so that the image of the head is consistent with the shape of the head during normal orientation, as the shape of the head changes as a function

(a)



(b)

Figure D-12: The "pocket scan converter" is held shut with one screw, which is hidden under one of the labels on the bottom. (a) Peel back. (b) Unscrew.

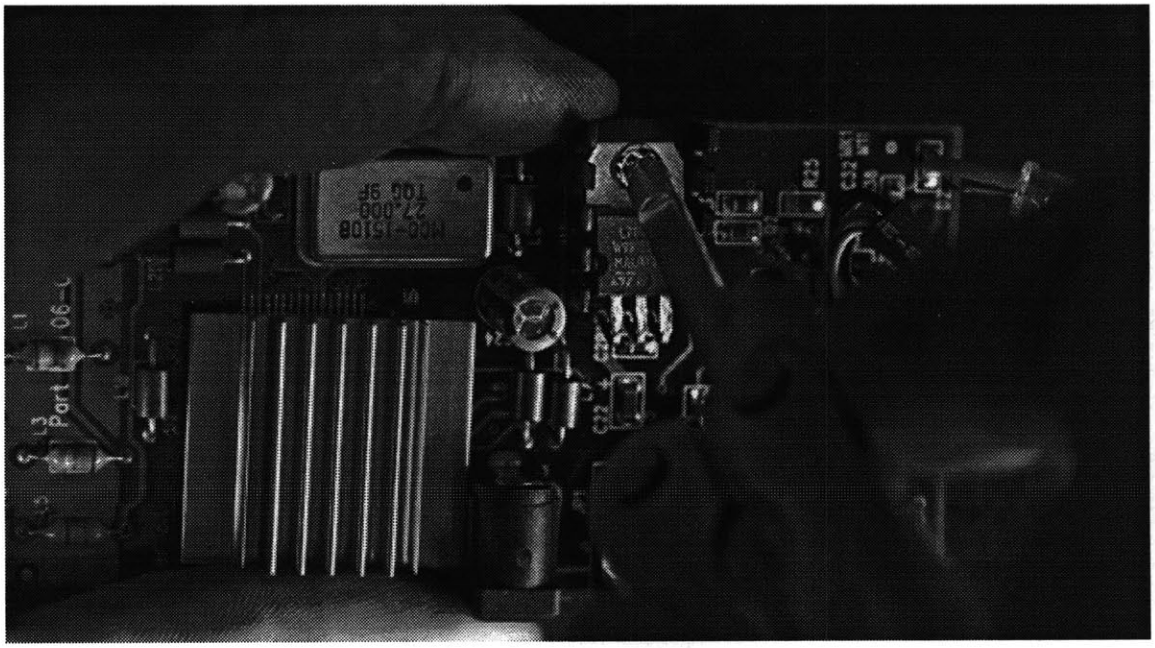


Figure D-13: Take the circuit board out of the "pocket scan converter". Remove the screw holding it in.

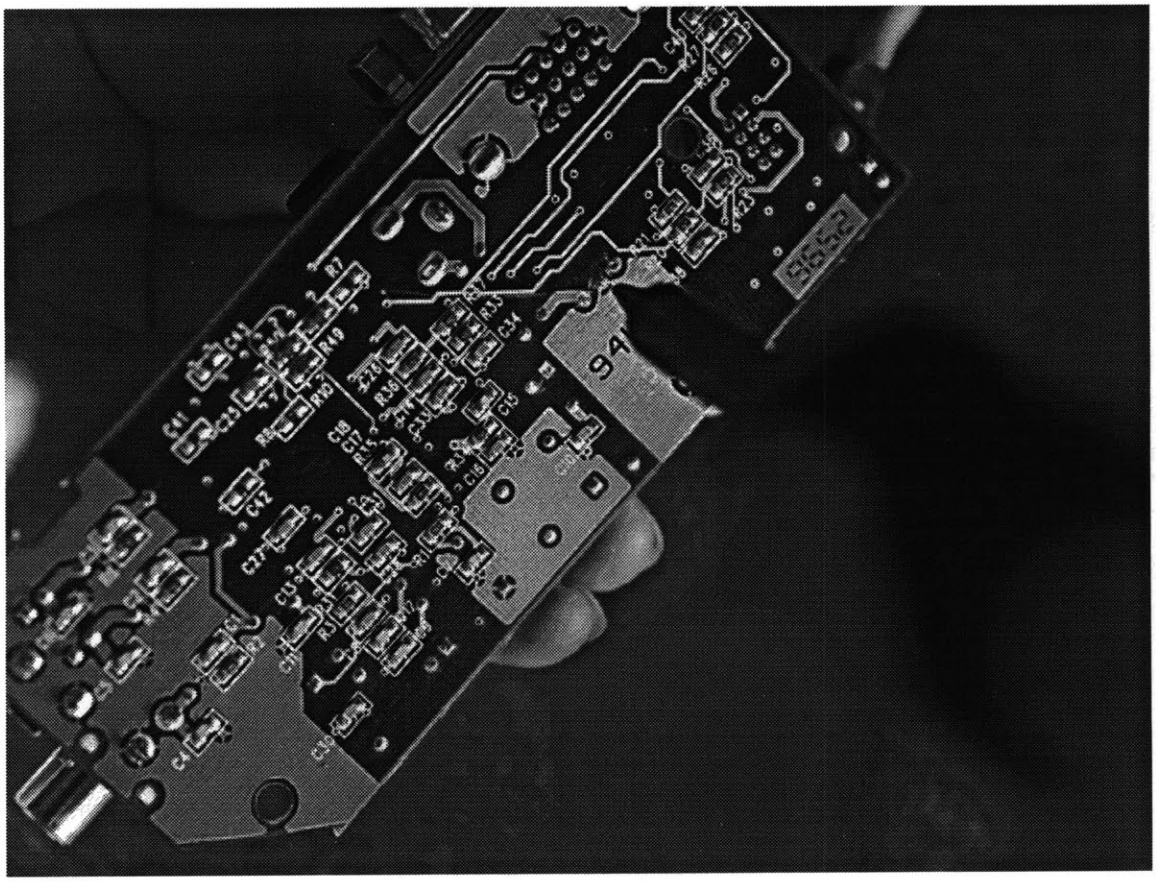


Figure D-14: Desolder the 7805 linear regulator and replace with ISR.

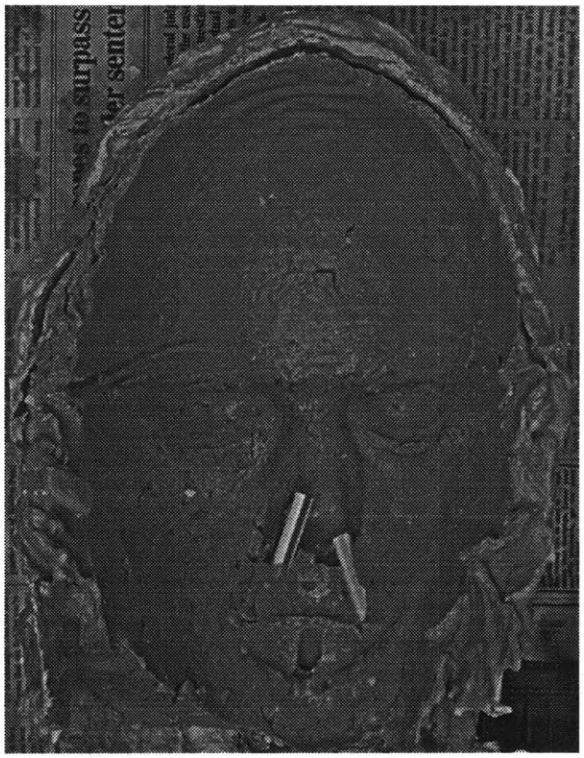


Figure D-15: The "negative" of my head, with nostril straws still in place.

of the direction of gravity[5]).

While building WearComp7, having WearComp6 plugged into a large screen television, together with a Twiddler, is beneficial, in order to communicate with others while the mouth is covered in the process of forming the "negative". (A straw in the mouth, and in each nostril is used to breathe through but otherwise it is necessary to be still for a couple of hours or so while the plaster, etc., sets). A notepad could be used, but the Twiddler is preferable as this minimizes movement, e.g., I had problems with the X-windows driver getting killed, as I was typing without being able to see what I was typing). A large plastic garbage bag, with hole cut out for head, is placed over head, to cover the body. It is desired to keep the excess waste plaster from the rest of the body; while seated, it tends to accumulate in the crotch area, so it should be cleared away as appropriate.

Old clothes are worn (clothes often get soaked in plaster despite care in covering — this is difficult to remove; I expect to throw away the clothing when done).

It is also preferable to plug the ears to avoid getting plaster in.

The "negative" is made in 2 parts, so that it can be separated from the head. One half (the front half) of the negative is depicted in Fig D-15. The two halves are brought together, and propped up ready to pour in the plaster for the positive. (Fig D-16) This must be done carefully because the negative will be destroyed in the process. The plaster is poured (Fig D-17) At the correct time, when the plaster is just hard enough to be solid, yet not too strong that it may bond the halves of the negative together, break it apart and take out the "positive" (Fig D-18(a)). Clear debris and remove image of nostril straws with knife if desired, while plaster is still soft. Before plaster hardens, correct any dimensional errors (Fig D-18(b)).

Shower well afterwards to remove any plaster from body before it hardens.

Once the head is imaged, this may be used as a substrate for building the sunglasses and laying out the components.

[5] On the other hand, if you were making the glasses for use by a medical patient while confined to a bed, laying down, you would want to make the "positive" of the head as it appears while facing up, with gravity acting front-to-back.

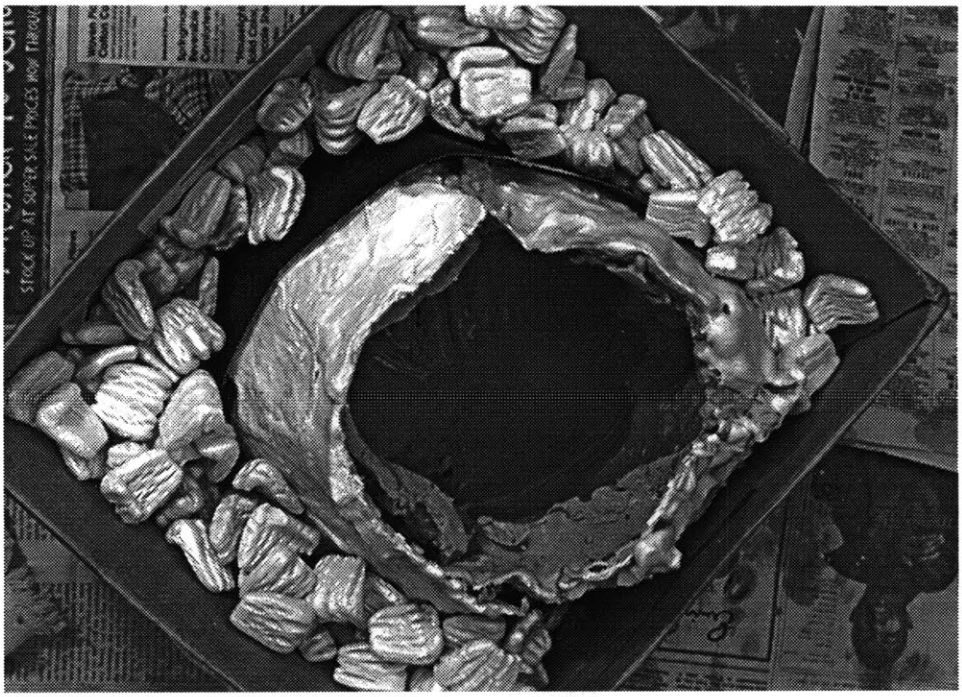


Figure D-16: The "negative" is propped up in a box with styrofoam "peanuts", ready for pouring in the plaster that will form the "positive".

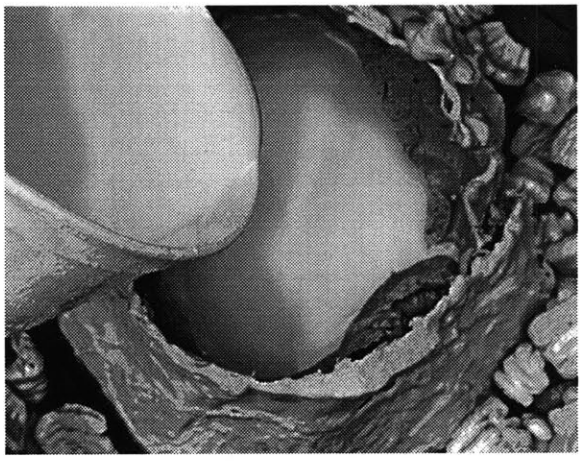
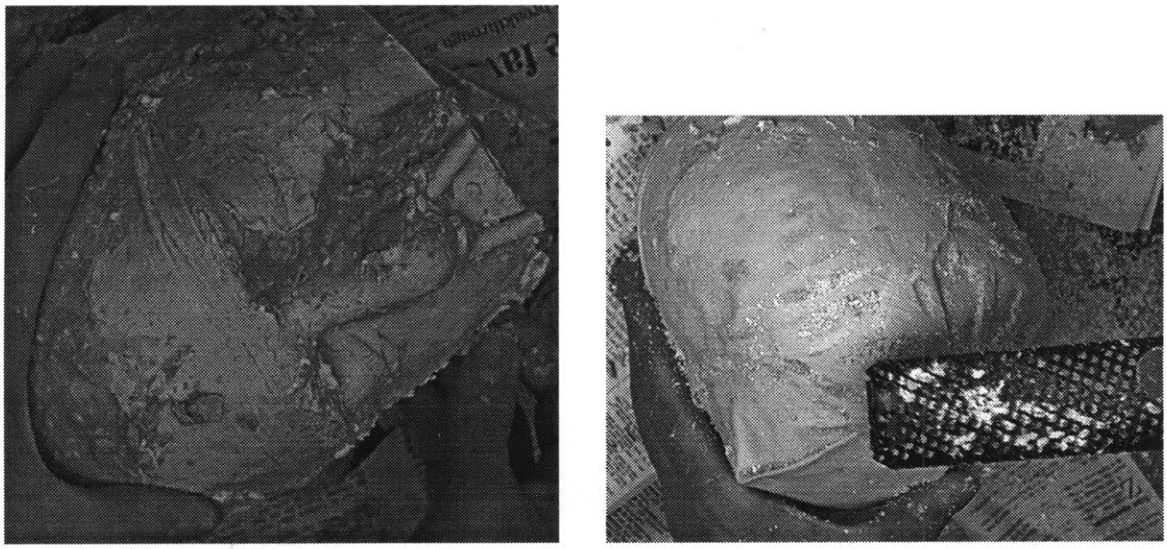


Figure D-17: Pour plaster into the "negative" slowly, and allow it to become somewhat consistent, before pouring in more. Otherwise the "negative" may begin leaking.



(a) (b)

Figure D-18: "Positive" of my head, removed just when plaster solidifies but before it becomes too hard to work with. (a) Image of nostril straws may be cut away with a sharp knife. (b) Errors in dimensional accuracy may be corrected before plaster hardens.

D.6 Layout for WearComp7

An LCD television screen as may be found in modern camcorder viewfinders (Fig D-19) is assembled with the appropriate drive electronics.

This is loosely gathered together with gaffer's tape into the sunglasses (Fig D-20) for fitting onto the head. I emphasize here that the device is put on the real head to adjust for optical path, and the "positive" copy of the head to adjust for fit, as it is necessary to work the fit with adhesive materials. Thus it is frequently necessary to move the rig back and forth between the real head and the "positive" copy. I should also emphasize that there is a considerable difference between the glasses made for one person and those made for another (e.g. those I made for myself versus those I made for my wife). Just as you would not try on someone else's mouthguard, you are not likely to want to try on someone else's glasses (e.g. many people ask me if they can try on my glasses, but I have found that they are easily broken or damaged by others trying them on). A useful metaphor to instill in people's minds is that of underwear and mouthguards (e.g. items that most people don't like to share).

D.6.1 Optics

The optical path needs to be lengthened appropriately (See Fig D-21). This may be done with a series of relay mirrors hidden in the frames as depicted in the figure. Front silvered mirrors are generally used[6] and aligned for the individual user. This process is quite complicated and often takes many hours. It may be helpful to make an alignment jig (e.g. "positive" head with camera or other optics inside), or to work with an assistant skilled in optical alignment. If very dark glass or mirrored glass is used, others may not notice the beam splitter and lens. If moderately dark glasses are used (as might be desirable for mixed indoor/outdoor use), another lens to cancel the effect of the first, and a lesser silvered beam splitter combined with increased screen brightness are used. Indoors, the screen brightness may be sufficient with a moderately silvered beamsplitter. Other embodiments of WearComp7 use flip-up welder's style goggles so that the view to the outside world

[6] Total internal refraction is another possibility, but then customization is is complicated owing to reduced number of degrees of freedom if manufacturability is considered.

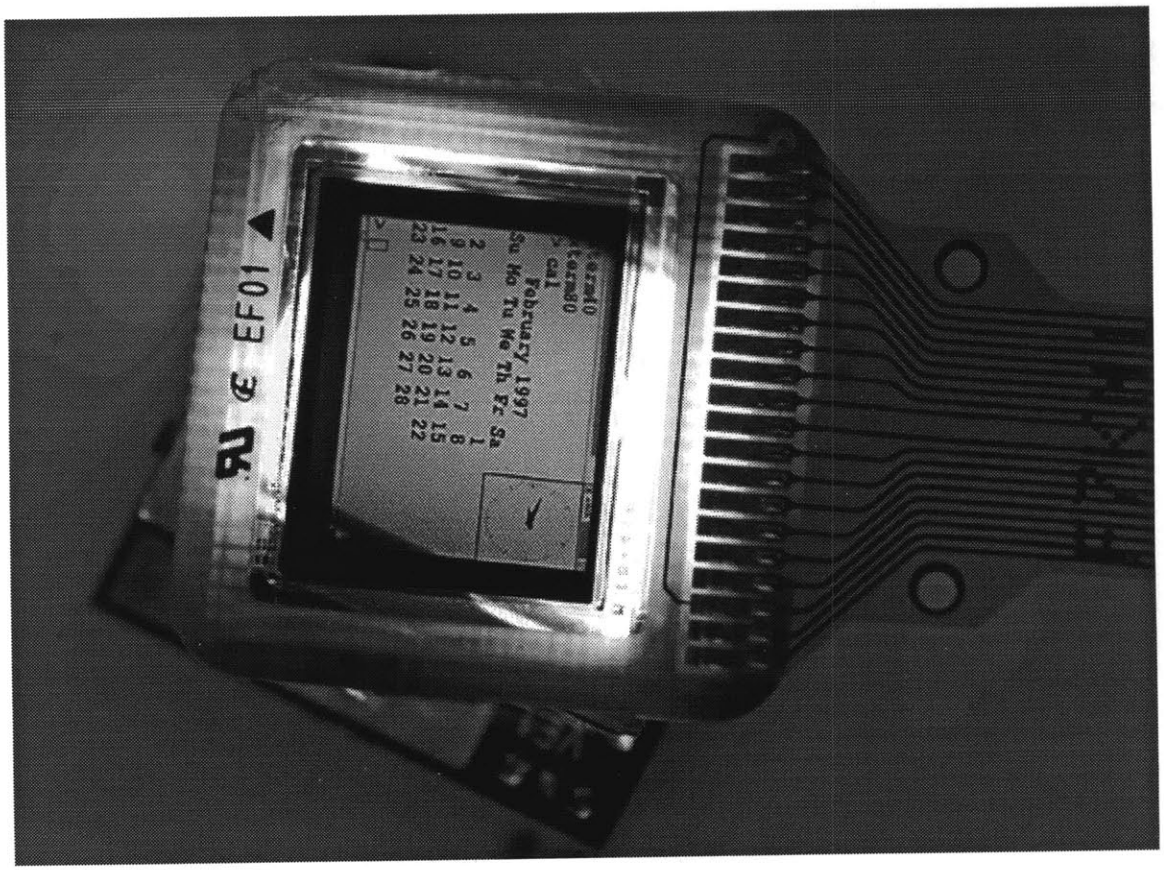


Figure D-19: LCD television screen. Test prior to assembly into eyeglasses.

Figure D-20: Materials loosely held in place with gaffer's tape. All wiring is concealed in the "croakies" eyeglass safety strap. If earphones are desired, these are inserted now, and may later be stripped down for concealment into the glasses, or may be left as they are for less unobtrusive more conventional usage. Subsequently, the display unit, sensor array, and wiring is sealed in epoxy, consistent with the amount of space available between the head and the glasses. The unit pictured here is a right-eyed unit, while the original embodiments of WearComp7 were identical but left-eyed. Some variations were also two-eyed systems.

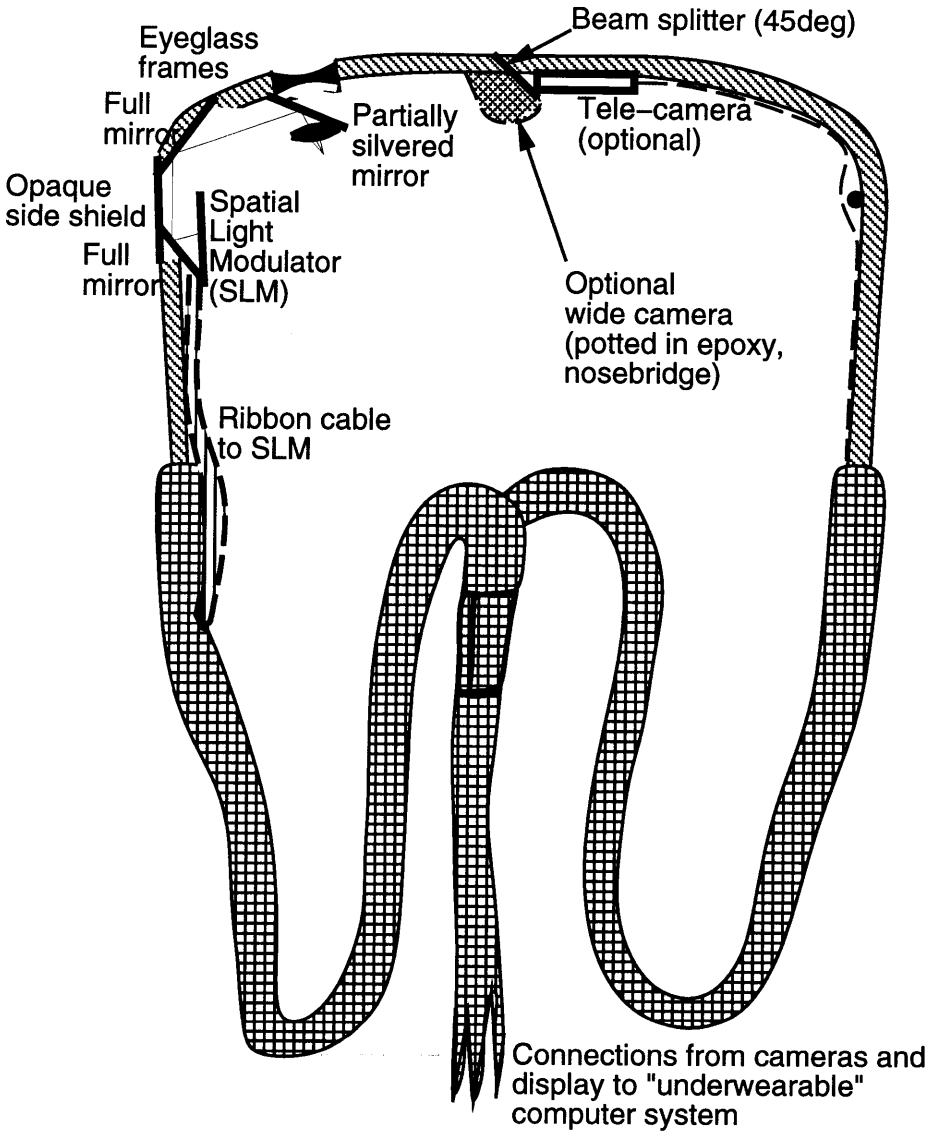


Figure D-21: In a preferred embodiment, optics for WearComp7 are based on a path-lengthening principle. In this manner, the path length can be increased sufficiently, while remaining hidden in the frames of the glasses. The negative lens is to conceal the effect of the positive lens to the outside viewer. Alternatively, extremely dark or mirrored glasses can be used to make this addition less visible. A secondary (but less important) reason for this addition is so that the wearer sees normally in that portion of the visual field of view. (The primary concern of WearComp7 is to make the rig unobtrusive, while making it unobstructive to the wearer was given less importance in the design.)

.

may be darkened enough to make reading the screen easy outdoors as well.

Getting the fit right may require compromising the design and alignment of the optics (Fig D-21). Here a tradeoff is made, that allows tighter profiling of the face. Note that the primary weakness of WearComp7 is a view from above or below. Thus an "eye in the sky" surveillance camera may be able to detect the unusual nature of the glasses because of the view afforded from directly above, even though the glasses look normal in face-to-face interaction.

Hopefully this deficiency will be corrected in future developments of WearComp.

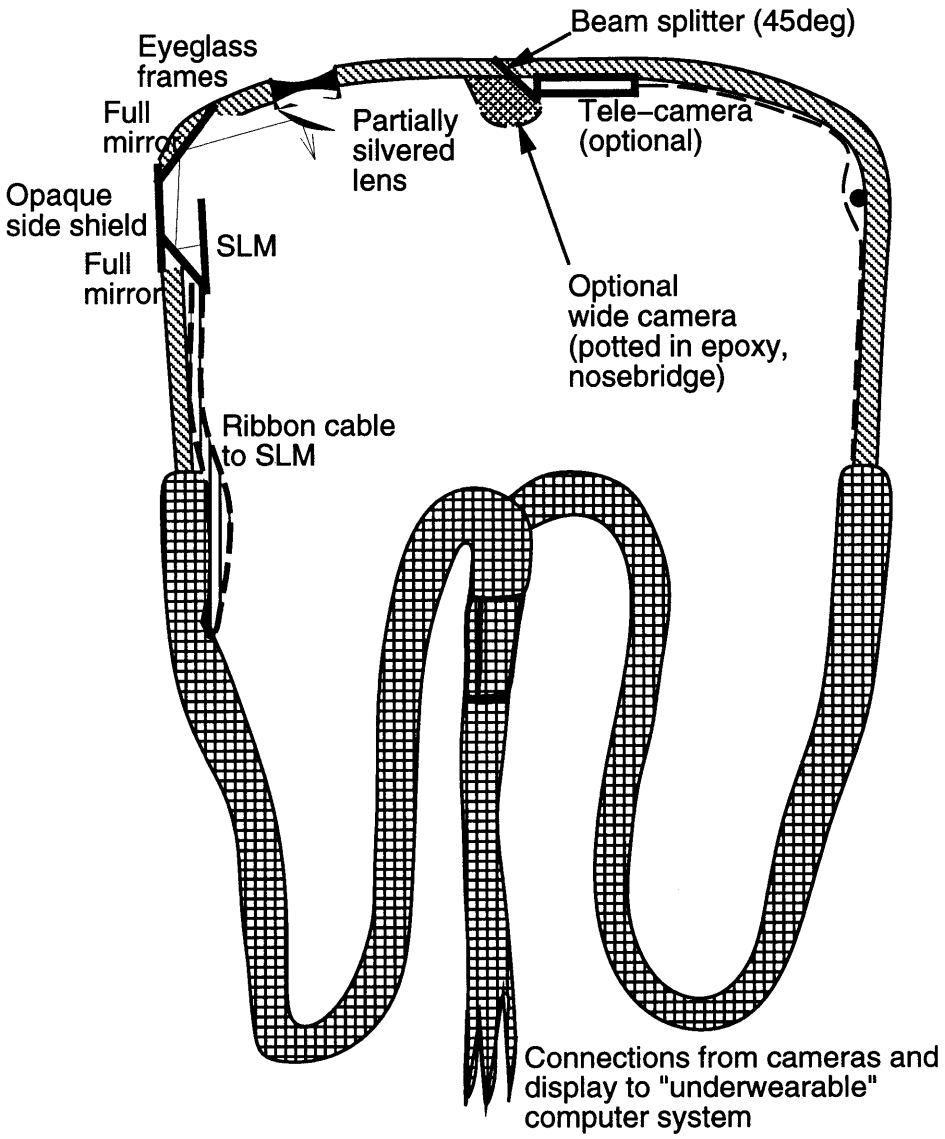


Figure D-22: In an alternative embodiment of WearComp7, optics are angled to match the profile of the eye surface, so that the unit can be built to tighter tolerance. Slight cancellation error results, so special lenses are required, or slightly darker glass is needed.

.

Appendix E

Video Orbits v1.0, or, how to "paint with a video camera"

This chapter is a brief tutorial on how to use Video Orbits version 1.0, while further information can be obtained from **http://wearcam.org/orbits**

Video Orbits version 1.0 partially implements the methodology of Chapter 4. (The work of Chapters 2, 3, and 5 will be incorporated into Video Orbits version 2.0.)

Video Orbits 1.0 allows you to "stitch together" multiple pictures of the same scene or object, taken from a fixed center of projection or of a flat surface. These assumptions may also be violated to some extent with successful results (see some of the examples in **http://wearcam.org/orbits**). The Video Orbits 1.0 can reassemble overlapping pictures of the same scene into one large "image composite", given the assumption that they lie in the same orbit of the projective group of coordinate transformations (e.g. static scene, constant lighting, and zero parallax, which happens if either the camera is at a fixed point in space but still free to pan, tilt, rotate about its optical axis, or zoom, or if the subject matter is flat). Given enough overlap in the input images, and a reasonable approximation to images being in the same orbit, the software can function without user intervention.

One can regard the process as embodying a "painting" metaphor, where the camera "paints" out images onto an empty "image canvas". If the camera "paints" back on itself (e.g. path loops around in a self-intersecting way so that it covers an area already "painted"), some cumulative error may become visible, because release 1.0 does not re-adjust to distribute the error over various possible frame pairs (this will be fixed in release 1.1).

The four main programs you need to use to assemble such image sets are estpchirp2m, pintegrate, pchirp2nocrop, and cement (Computer Enhanced Multiple Exposure Numerical Technique).

The programs use the "isatty" feature of the C programming language to provide documentation which is accessed by running them with no command line arguments (e.g. from a TTY) to get a help screen. The sections for each program give usage hints where appropriate. Future versions will support the "pipe" construct (e.g. some programs may be used without command line arguments but will still do the right thing in this case rather than just printing a help message).

The first program you need to run is estpchirp2m, which estimates coordinate transformation parameters between pairs of images. These "chirp" parameters are sets of eight real-valued quantities which indicate a projective (i.e., affine plus chirp) coordinate transformation on an image.

The images are generally numbered sequentially, for example, v000.ppm, v001.ppm, ... v116.ppm (e.g. for an image sequence with 117 pictures in it).

After you run estpchirp2m on all successive pairs of input images in the sequence, the result will be a set of sets of eight numbers, in ASCII text, one set of numbers per row of text (the numbers separated by white space). The number of lines in the output ASCII text file will be one less than the total number of frames in the input image sequence. For example, if you have a 117-frame sequence (e.g. image files numbered v000.ppm to v116.ppm), there will be 116 lines of ASCII text in the output file from estpchirp2m.

The first row of the text file (e.g. the first set of numbers) indicates the coordinate transformation

between frame 0 and frame 1; the second row, the coordinate transformation between frame 1 and frame 2, ... A typical filename for these parameters is "parameters_pairwise.txt"

These pairwise *relative* parameter estimates are then to be converted into "integrated" *absolute* coordinate transformation parameters (e.g. coordinate transformations with respect to some selected 'reference frame'). This conversion is done by a program called pintegrate.

This program takes as input the filename of the file containing parameters from the ASCII text file produced by estpchirp2m (e.g. "parameters_pairwise.txt" and a 'reference frame' (specified by the user), and calculates the coordinate transformation parameters from each frame in the image sequence to this specified 'reference frame'.

The output of pintegrate is another ASCII text file which lists the set of chirp parameters (again, 8 numbers per chirp parameter, each set of 8 numbers in ASCII, on a new row of text), this time one parameter per frame, designed to be used in order. That is, the first row of the output text file (first set of 8 real numbers) provides the coordinate transformation from frame 0 to the reference frame, the second from frame 1 to the reference frame...

The program called pchirp2nocrop takes the ppm or pgm image for each input frame, together with the chirp parameter for this frame and 'dechirps' it (applies the coordinate transformation to bring it into the same coordinates as the reference frame). Generally the parameters passed to pchirp2nocrop are those which come from pintegrate (e.g. *absolute* parameters, not relative parameters). The output of pchirp2nocrop is another ppm or pgm file.

The program called cement[1] assembles the dechirped images (which have been processed by pchirp2nocrop) and assembles them onto one large image 'canvas'.

E.1 Additional notes

The estimation program, estpchirp2m, can take optional initial parameter "guess" inputs. This is useful when incoming images do not have a lot of overlap (i.e., taken with a still camera rather than a video camera).

Initial estimation parameters can be generated using the Abelian pre-processors (still to be ported to C), or idr.c (interactive dechirp/rechirp) may be used to experiment with manual specification of the starting "guess".

E.2 The history of Video Orbits

The collection of tools comprising the Video Orbits software was originally developed over a period of several years by Steve Mann while at McMaster University, University of Toronto, HP Labs, and MIT. Others at MIT, including Shawn Becker, Nassir Navab, Chris Ggraczyk, Jeffrey Levine, and Kenneth Russell have recently contributed extensively to this project (see acknowledgements section).

[1] CEMENT is an acronym for Computer Enhanced Multiple Exposure Numerical Technique.

Bibliography

[1] Hubert Dolezal. *Living in a world transformed.* Academic press series in cognition and perception. Academic press, Chicago, Illinois, 1982.

[2] Ivo Kohler. *The formation and transformation of the perceptual world*, volume 3 of *Pshchological issues.* International university press, 227 West 13 Street, 1964. monograph 12.

[3] Stanley Milgram. *The Individual in a Social World*, chapter 25, pages 337–345. McGraw-Hill, Inc., New York, second edition.

[4] *Cold Iron and Lady Godiva.* University of Toronto press, 1973.

[5] Jennifer González. *Prosthetic Territories*, chapter 9, Autotopographies, pages 133–150. Westview Press, Boulder.

[6] Joseph Hoshen, Jim Sennott, and Max Winkler. Keeping tabs on criminals. *IEEE SPECTRUM*, pages 26–32, February 1995.

[7] A. H. Maslow. *Toward a psychology of being.* Van Nostrand, Princeton, NJ, 2nd edition, 1968.

[8] A. Chakra and Kundalini workbook Dr. John Mumford. Self-actualization, self-understanding and personal transformation, ISBN 1-56718-473-1. http://www.llewellyn.com/xselfact.htm.

[9] Phillip Laplante editor Marvin Minsky. Steps toward artificial intelligence. In *Great papers on computer science*, West Publishing Company, Minneapolis/St. Paul, 1996 (paper in IRE 1960).

[10] Compiled and edited from the original manuscripts by Jean Paul Richter. *The Notebooks of Leonardo Da Vinci*, volume 1. Dover Publications, Inc., 1452-1519; 1970.

[11] S. Mann. Wearable Wireless Webcam, 1994. http://wearcam.org.

[12] Steve Mann. **Wearable, tetherless computer–mediated reality**: WearCam as a wearable face–recognizer, and other applications for the disabled. TR 361, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Ma, February 2 1996. Also appears in **AAAI Fall Symposium on Developing Assistive Technology for People with Disabilities**, 9-11 November 1996, MIT.

[13] S. Mann. 'mediated reality'. TR 260, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Ma, 1994.

[14] Harold E. Edgerton. *Electronic flash, strobe.* MIT Press, Cambridge, Massachusetts, 1979.

[15] Steve Mann. 'smart clothing': Wearable multimedia and 'personal imaging' to restore the balance between people and their intelligent environments. Proceedings, ACM Multimedia 96, Nov. 18-22 1996.

[16] S. Mann and R. W. Picard. Video orbits of the projective group; a simple approach to featureless estimation of parameters. TR 338, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Ma, 1995. IEEE Trans. Image Proc., Sept 1997.

[17] T. S. Huang and A.N. Netravali. Motion and structure from feature correspondences: a review. *Proc. IEEE*, Feb 1984.

[18] R. W. Picard and J. Healey. Affective wearables. In *Proceedings of the First International Symposium on Wearable Computers*, Cambridge, MA, Oct. 1997. To appear.

[19] William J. Mitchell. *The Reconfigured Eye.* The MIT Press, 1992.

[20] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. *M. Landy and J. A. Movshon, (eds) Computational Models of Visual Processing*, 1991.

[21] Graham Saxby. *practical holography.* Prentice hall, second edition, 1994.

[22] B. R. Alexander, P. M. Burnett, J.-M.R. Fournier, and S. E. Stamper. Accurate color reproduction by Lippman photography. In *SPIE Proceedings #3011-34 Practical holography and holographic materials*, Bellingham WA 98227 USA, Tues. Feb. 11, 1997. Photonics West 97 SPIE, Cosponsored by IS&T. Chair T. John Trout, DuPont.

[23] Steve Mann. Lightspace. Submitted to SIGGRAPH, 1992 (Paper available from author. Also see example images in http://wearcam.org/lightspace), July 1992.

[24] T. G. Stockham, Jr. Image processing in the context of a visual model. *Proc. IEEE*, 60(7):828–842, July 1972.

[25] S. R. Curtis and A. V. Oppenheim. Signal reconstruction from Fourier transform sign information. Technical Report No. 500, MIT Research Laboratory of Electronics, May 1984.

[26] Charles W. Wyckoff. An experimental extended response film. Technical Report NO. B-321, Edgerton, Germeshausen & Grier, Inc., Boston, Massachusetts, MARCH 1961.

[27] Charles W. Wyckoff. An experimental extended response film. *S.P.I.E. NEWSLETTER*, JUNE-JULY 1962.

[28] S. Mann. Compositing multiple pictures of the same scene. In *Proceedings of the 46th Annual IS&T Conference*, Cambridge, Massachusetts, May 9-14 1993. The Society of Imaging Science and Technology.

[29] Martin Bichsel and Krystyna W. Ohnesorge. How to measure a camera's response function from scratch .

[30] A.M. Tekalp, M.K. Ozkan, and M.I. Sezan. High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration. In *Proc. of the Int. Conf. on Acoust., Speech and Sig. Proc.*, pages III–169, San Francisco, CA, Mar. 23-26, 1992. IEEE.

[31] M. Irani and S. Peleg. Improving Resolution by Image Registration. *CVGIP*, 53:231–239, May 1991.

[32] S. Mann and R. W. Picard. Virtual bellows: constructing high-quality images from video. In *Proceedings of the IEEE first international conference on image processing*, Austin, Texas, Nov. 13-16 1994.

[33] S. Mann. Recording 'lightspace' so shadows and highlights vary with varying viewing illumination. Technical Report 348, MIT Media Lab, Cambridge, Massachusetts, December 1994. MAS854 final course project, also appears, OPTICS LETTERS/Vol. 20 No. 24 December 15, 1995.

[34] Cynthia Ryals. Lightspace: A new language of imaging. *PHOTO Electronic Imaging*, 38(2):14–16, 1995. http://www.novalink.com/pei/mann2.html.

[35] S.S. Beauchemin J.L. Barron, D.J. Fleet. Systems and experiment performance of optical flow techniques. *International journal of computer vision*, pages 43–77, 1994.

[36] Qinfen Zheng and Rama Chellappa. A Computational Vision Approach to Image Registration. *IEEE Transactions Image Processing*, July 1993. pages 311-325.

[37] L. Teodosio and W. Bender. Salient video stills: Content and context preserved. *Proc. ACM Multimedia Conf.*, August 1993.

[38] R. Szeliski and J. Coughlan. Hierarchical spline-based image registration. *CVPR*, pages 194–201, 1994.

[39] George Wolberg. *Digital Image Warping.* IEEE Computer Society Press, 10662 Los Vaqueros Circle, Los Alamitos, CA, 1990. IEEE Computer Society Press Monograph.

[40] G. Adiv. Determining 3D Motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Machine Intell.*, pages 304–401, July 1985.

[41] Nassir Navab and Steve Mann. Recovery of relative affine structure using the motion flow field of a rigid planar patch. *Mustererkennung 1994, Tagungsband.*, 1994.

[42] R. Y. Tsai and T. S. Huang. Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch. *Trans. Accoust., Speech, and Sig. Proc.*, 1981.

[43] O. D. Faugeras and F. Lustman. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(3):485–508, 1988.

[44] Amnon Shashua and Nassir Navab. Relative Affine: Theory and Application to 3D Reconstruction From Perspective Views. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 1994. 1994.

[45] H.S. Sawhney. Simplifying motion and structure analysis using planar parallax and image warping. *ICPR*, 1, October 1994. 12th IAPR.

[46] R. Kumar, P. Anandan, and K. Hanna. Shape recovery from multiple views: a parallax based approach. *ARPA image understanding workshop*, 10 Nov 1994.

[47] Lee Campbell and Aaron Bobick. Correcting for radial lens distortion: A simple implementation. TR 322, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Ma, Apr 1995.

[48] S. Mann and R.W. Picard. Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. Technical Report 323, M.I.T. Media Lab Perceptual Computing Section, Boston, Massachusetts, 1994. Also appears, IS&T's 46th annual conference, pages 422-428, May 1995.

[49] M. Artin. *Algebra.* Prentice Hall, 1991.

[50] S. Mann. Wavelets and chirplets: Time–frequency perspectives, with applications. In Petriu Archibald, editor, *Advances in Machine Vision, Strategies and Applications.* World Scientific, Singapore . New Jersey . London . Hong Kong, world scientific series in computer science - vol. 32 edition, 1992.

[51] L.V. Ahlfors. *Complex Analysis.* International Series in Pure and Applied Mathematics. McGraw Hill, Inc., 3rd edition, 1979.

[52] R. Y. Tsai and T. S. Huang. Multiframe image restoration and registration. *ACM*, 1984.

[53] Nassir Navab and Amnon Shashua. Algebraic Description of Relative Affine Structure: Connections to Euclidean, Affine and Projective Structure. *MIT Media Lab Memo No. 270*, 1994.

[54] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory (Part I).* John Wiley and Sons, 1968.

[55] Steve Mann and Simon Haykin. The chirplet transform: Physical considerations. *IEEE Trans. Signal Processing*, 43(11), November 1995.

[56] A. Berthon. Operator Groups and Ambiguity Functions in Signal Processing. In J.M. Combes, editor, *Wavelets: Time-Frequency Methods and Phase Space*. Springer Verlag, 1989.

[57] A. Grossmann and T. Paul. Wave functions on subgroups of the group of affine cannonical tranformations. Lecture notes in physics, No. 211: Resonances — Models and Phenomena, pages 128–138. Springer-Verlag, 1984.

[58] R. K. Young. Wavelet theory and its applications. 1993.

[59] Lora G. Weiss. Wavelets and wideband correlation processing. *IEEE Signal Processing Magazine*, pages 13–32, 1993.

[60] Steve Mann and Simon Haykin. Adaptive "Chirplet" Transform: an adaptive generalization of the wavelet transform. *Optical Engineering*, 31(6):1243–1256, June 1992.

[61] B. Horn and B. Schunk. Determining Optical Flow. *Artificial Intelligence*, 1981.

[62] John Y.A. Wang and Edward H. Adelson. Spatio-Temopral Segmentation of Video Data . In *SPIE Image and Video Processing II*, pages 120–128, San Jose, California, February 7-9 1994.

[63] J. Bergen, P. Burt, R. Hingorini, and S. Peleg. Computing two motions from three frames. In *Proc. Third Int'l Conf. Comput. Vision*, pages 27–32, Osaka, Japan, December 1990.

[64] Lucas and T. Kanade. An iterative image-registration technique with an application to stereo vision . In *Image Understanding Workshop*, pages 121–130, 1981.

[65] J. Y. A. Wang and Edward H. Adelson. Representing moving images with layers. *Image Processing Spec. Iss: Image Seq. Compression*, 12(1), September 1994.

[66] Roland Wilson and Goesta H. Granlund. The Uncertainty Principle in Image Processing . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1984.

[67] J. Segman, J. Rubinstein, and Y. Y. Zeevi. The canonical coordinates method for pattern deformation: Theoretical and computational considerations. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 14(12):1171–1183, Dec. 1992.

[68] J. Segman. Fourier cross correlation and invariance transformations for an optimal recognition of functions deformed by affine groups. *Journal of the Optical Society of America, A*, 9(6):895–902, June 1992.

[69] J. Segman and W. Schempp. *Two methods of incorporating scale in the Heisenberg group.* 1993. JMIV special issue on wavelets.

[70] Bernd Girod and David Kuo. Direct estimation of displacement histograms. *OSA Meeting on IMAGE UNDERSTANDING AND MACHINE VISION*, June 1989.

[71] Yunlong Sheng, Claude Lejeune, and Henri H. Arsenault. Frequency-domain Fourier-Mellin descriptors for invariant pattern recognition. *Optical Engineering*, May 1988.

[72] Peter J. Burt and P. Anandan. Image stabilization by registration to a reference mosaic. *ARPA image understanding workshop*, 10 Nov 1994.

[73] M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt. Real-time scene stabilization and mosaic construction. *ARPA image understanding workshop*, 10 Nov 1994.

[74] S. Intille. Computers watching football, 1995. http://www-white.media.mit.edu/vismod/demos/football/football.html.

[75] R. Wilson, A. D. Calway, E. R. S. Pearson, and A. R. Davies. An introduction to the multiresolution Fourier transform. Technical report, Department of Computer Science, University of Warwick, Coventry CV4 7AL UK., 1992. ftp://ftp.dcs.warwick.ac.uk/reports/rr-204/.

[76] A D Calway, H Knutsson, and R Wilson. Multiresolution estimation of 2-d disparity using a frequency domain approach. pages 227–236. Springer-Verlag, September 1992.

[77] R. A. Earnshaw, M. A. Gigante, and H Jones. *Virtual reality systems.* Academic press, 1993.

[78] I. Sutherland. A head-mounted three dimensional display. In *Proc. Fall Joint Computer Conference*, pages 757–764, 1968.

[79] S. Feiner, B. MacIntyre, and D. Seligmann. Knowledge-based augmented reality, Jul 1993. Communications of the ACM, 36(7).

[80] S. Feiner, B. MacIntyre, and D. Seligmann. Karma (knowledge-based augmented reality for maintenance assistance), 1993. http://www.cs.columbia.edu/graphics/projects/karma/karma.html.

[81] Henry Fuchs, Mike Bajura, and Ryutarou Ohbuchi. Teaming ultrasound data with virtual reality in obstetrics. http://www.ncsa.uiuc.edu/Pubs/MetaCenter/SciHi93/ 1c.Highlights-BiologyC.html.

[82] David Drascic. David drascic's papers and presentations, 1993. http://vered.rose.utoronto.ca/people/david_dir/ Bibliography.html.

[83] P. St Hilaire, S.A. Benton, and M. Lucente. Electronic display system for computational holography. In *SPIE Proceedings #1212 "Practical holography IV"*, pages 174–182, 1990.

[84] M. Lucente, P. St Hilaire, and S.A. Benton. A new approach to holographic video. *SPIE Proceedings #1732 "Holography '92"*, 1992.

[85] Ronald Azuma. Registration Errors in Augmented Reality: NSF/ARPA Science and Technology Center for Computer Graphics and Scientific Visualization , 1994. http://www.cs.unc.edu/~azuma/azuma_AR.html.

[86] George M. Stratton. Some preliminary experiements on vision. *Psychological Review*, 1896.

[87] Simon Haykin. *Communication Systems.* Wiley, second edition, 1983.

[88] G. Arfken. *Mathematical Methods for Physicists.* Academic Press, Orlando, Florida, third edition, 1985.

[89] K. Nagao. Ubiquitous talker: Spoken language interaction with real world objects., 1995. http://www.csl.sony.co.jp/person/nagao.html.

[90] Michael W. McGreevy. The presence of field geologists in mars-like terrain. *PRESENCE*, 1(4):375–403, FALL 1992. MIT Press.

[91] Stuart Anstis. Visual adaptation to a negative, brightness-reversed world: some preliminary observations. In Gail Carpenter and Stephen Grossberg, editors, *Neural Networks for Vision and Image Processing*, pages 1–15. MIT Press, 1992.

[92] Lions vision research and rehabilitation center, 1995. http://www.wilmer.jhu.edu/low_vis/low_vis.htm.

[93] Manfred Clynes. personal communication.

[94] M. Clynes and N.S. Kline. Cyborgs and space. *Astronautics*, 14(9):26–27,and 74–75, September September 1960.

[95] Bob Shaw. *Light of Other Days.* Analog, August 1966.

[96] Steve Mann. The eudaemonic eye. *CHI-97.*

[97] Maes, Darrell, Blumberg, and Pentland. The alive system: Full-body interaction with animated autonomous agents. TR 257, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Ma, 1994.

[98] R. W. Picard. Affective computing. Media Laboratory, Perceptual Computing TR 321, MIT Media Lab, 1995.

[99] T. Starner. The remembrance agent, 1993. Class project for intelligent software agents class of Patti Maes.

[100] S. Feiner, Webster, Krueger, B. MacIntyre, and Keller. Architectural anatomy, 1995. Presence, 4(3), 318-325.

[101] R. E. Cytowic. *Synesthesia: A union of the senses.* Springer-Verlag, New York, 1989.

[102] R. E. Cytowic. *The Man Who Tasted Shapes.* G. P. Putnam's Sons, New York, NY, 1993.

[103] Don Norman. *Turn signals are the facial expressions of automobiles.* Addison Wesley, 1992.

[104] Steve Mann. Wearable computing: A first step toward personal imaging. *IEEE Computer,* 30(2), Feb 1997. http://computer.org/pubs/computer/1997/0297toc.htm.

[105] Vannevar Bush. As we may think. *Atlantic Monthly,* July 1945. http://www2.theatlantic.com/atlantic/atlweb/flashbks/ computer/bushf.htm.

[106] P. J. O'Connell. *Robert Drew and the Development of Cinema Verite in America.* Southern Illinois University Press, 1992.

[107] Mick Hans. Cameras catch red-light runners cities install photo-enforcement systems at problem intersections. *WIRED,* January/February 1997.

[108] Erving Goffman. *The Presentation of Self in Everyday Life.* Doubleday, Garden City, New York, 1959.

[109] Walter Benjamin. The work of art in the age of mechanical reproduction. In Hannah Arendt, editor, *Illuminations.* Schocken, New York, 1968.

[110] Phil Patton. Caught. *WIRED,* January 1995.

[111] Sr. Consumer Correspondent Hattie Kauffman. CBS good morning, July 3, 1996.

[112] LynNell Hancock, Claudia Kalb, and William Underhill. You don't have to smile. *Newsweek,* July 17 July 17, 1995.

[113] Andrew Wilson, Aaron Bobick, Lee Campbell, Elvis the Monster, Jim Davis, Freedom Baird, Stephen Intille, Arjan Schutte, Claudio Pinhanez, and Yuri Ivanov. Kids room, 1987. http://www-white.media.mit.edu/vismod/demos/kidsroom/.

[114] ADVANCED IMAGING. Imaging in the internet. *Solutions for the Electronic Imaging Professional,* February 1994.

[115] Bradley corporation. The quick guide to electronic plumbing control, 9101 fountain boulevard menomonee falls, WI 53051.

[116] Joe Constance. Nowhere to hide. holographic imaging radar may soon be uncovering hidden dangers at u.s. airports. http://www.ingersoll-rand.com/compair/octnov96/radar.htm.

[117] Moderated by Lauren Weinstein. Privacy forum digest, Monday, 28 October 1996. http://wearcam.org/privacy_forum_digest_nakedradar.html.

[118] Paul Virilio. *The Vision Machine.* British Film Institute and INDIANA UNIVERSITY PRESS, Bloomington and Indianapolis, 1994.

[119] Allan Sekula. The body and the archive. *October*, pages 3–64, 39.

[120] Simon Davies. Privacy international. http://www.privacy.org/pi/activities/idcard/campaigns.html.

[121] Michel Foucault. *Discipline and Punish.* 1977. Translated from "Surveiller et punir".

[122] Adam Oranchak. personal communication. http://wearcam.org/previous_experiences/adam_oranchak/.

[123] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. Picard, and A. Pentland. Augmented reality through wearable computing. *PRESENCE*, 6(4), 1997. MIT Press.

[124] S. Mann. "smart clothing". TR 366, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Ma, February 2 1996.

[125] Michael Schneider. In baltimore, big brother moves in. *The Detroit News Home Page*, January 20 1996. http://www.detnews.com/menu/stories/32681.htm.

[126] Elisabeth Sussman. on the passage of a few people through a rather brief moment in time: The situationist international. 1957-1972.

[127] Tom Ward. The situationists reconsidered. In Douglas Kahn and Diane Neumaier, editors, *Cultures in Contention.* The real comet press.

[128] Steve Mann. Shootingback: Personal imaging in personal documentary, 1996. Submitted to American Cinematographer; see also http://wearcam.org/shootingback.html or http://18.85.20.100/shootingback.html.

[129] Stelarc official web site - Australia, 1997. http://www.merlin.com.au/stelarc/.

[130] Arthur elsenaar, 1997. http://www.desk.nl/~ acsi/WS/artists/elsenaar.htm and http://wearcam.org/previous_experiences/arthur_elsenaar/.