Eulerian Video Processing and Medical Applications

by

Hao-Yu Wu

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author .....
Department of Electrical Engineering and Computer Science
May 25, 2012

Certified by .....
Frédo Durand
Associate Professor of Computer Science
Thesis Supervisor

Certified by .....
William T. Freeman
Professor of Computer Science
Thesis Supervisor

Accepted by .....
Prof. Dennis M. Freeman
Chairman, Masters of Engineering Thesis Committee

Eulerian Video Processing and Medical Applications

by

Hao-Yu Wu

Submitted to the Department of Electrical Engineering and Computer Science
on May 25, 2012, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science

Abstract

Our goal is to reveal subtle yet informative signals in videos that are difficult or impossible to see with the naked eye. We can either display them in an indicative manner, or analyse them to extract important measurements, such as vital signs. Our method, which we call Eulerian Video Processing, takes a standard video sequence as input, and applies spatial decomposition, followed by temporal filtering to the frames.

The resulting signals can be visually amplified to reveal hidden information, the process we called Eulerian Video Magnification. Using Eulerian Video Magnification, we are able to visualize the flow of blood as it fills the face and to amplify and reveal small motions. Our technique can be run in real time to instantly show phenomena occurring at the temporal frequencies selected by the user.

Those signals can also be used to extract vital signs contactlessly. We presented a heart rate extraction system that is able to estimate heart rate of newborns from videos recorded in the real nursery environment. Our system can produce heart rate measurement that has clinical accuracy when newborns only have mild motions, and when the videos are acquired in brightly lit environments.

Thesis Supervisor: Frédo Durand
Title: Associate Professor of Computer Science

Thesis Supervisor: William T. Freeman
Title: Professor of Computer Science

Acknowledgments

This thesis would not have been possible without the the invaluable contributions of many individuals. I am very grateful to have great mentors, colleagues, friends and family members around who have always provided tremendous support during the entire process.

First, I would not be here without Frédo Durand, William T. Freeman, John Guttag if they hadn't given me the opportunity to join this exciting project. I have grown more intellectually in my time under their supervisions then any other periods in my life. They have always provided invaluable and insightful advices from many different perspectives at every meeting, which stimulated me and shaped my research approach and thinking.

Second, huge thanks to Eugene Shih and Michael Rubinstein. Without their helps, this work would not have been published to SIGGRAPH conference and get accepted. Having discussion with them is always provocative and open up many possible directions for the project. Their helps on the implementations and writings of the project are indispensable factors for the completion of this thesis.

This work would also not have been possible without helpful feedbacks from Guha Balakrishnan, Steve Lewin-Berlin, Neal Wadhwa, and the SIGGRAPH reviewers. Thanks to Ce Liu and Deqing Sun for helpful discussions on the Eulerian vs. Lagrangian analysis for our project. Thanks to Dr. Donna Brezinski, Dr. Karen McAlmon, and the Winchester Hospital staff for helping me collect videos of newborn babies dataset. Also thanks to the financial support for our project provided by DARPA SCENICC program, NSF CGV-1111415, and Quanta Computer.

Last but not least, I would like to thank my family members and my friends for constantly supporting me. They inspired and motivated me to pursue my dreams, gave me all the necessary help that I need, and cheered for my success. This thesis is more meaningful because of them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human visual system has limited spatio-temporal sensitivity, but many signals that fall below this capacity can be informative. For example, human skin color varies slightly with blood circulation. This variation, while invisible to the naked eye, can be exploited to extract pulse rate [14, 12, 11]. Similarly, motion with low spatial amplitude, while hard or impossible for humans to see, can be magnified to reveal interesting mechanical behavior [9]. The success of these tools motivates the development of new techniques to extract invisible signals in videos. The extracted signals can be used to either estimate interesting measurement, such as pulse rate, or to redisplay on the videos in an indicative manner. In this thesis, we show that a combination of spatial and temporal processing of videos can (a) amplify subtle variations that reveal important aspects of the world around us, and (b) extract vital signs contactlessly.

Our basic approach is to consider the time series of color values at any spatial location (pixel) and amplify variation in a given temporal frequency band of interest. For example, in Figure 1-1 we automatically select, and then amplify, a band of temporal frequencies that includes plausible human heart rates. The amplification reveals the variation of redness as blood flows through the face. For this application, temporal filtering needs to be applied to lower spatial frequencies (spatial pooling) to allow such a subtle input signal to rise above the camera sensor and quantization noise.

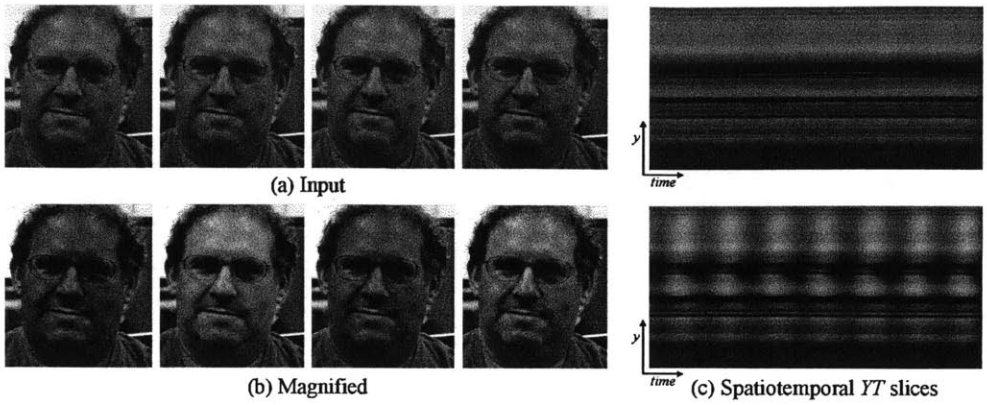


Figure 1-1: An example of using our Eulerian Video Magnification framework for visualizing the human pulse. (a) Four frames from the original video sequence (*face*). (b) The same four frames with the subject's pulse signal amplified. (c) A vertical scan line from the input (top) and output (bottom) videos plotted over time shows how our method amplifies the periodic color variation. In the input sequence the signal is imperceptible, but in the magnified sequence the variation is clear. The complete sequence is available in the supplemental video.

Our temporal filtering approach not only amplifies color variation, but can also reveal low-amplitude motion. For example, in the supplemental video, we show that we can enhance the subtle motions around the chest of a breathing baby. We provide a mathematical analysis that explains how temporal filtering interplays with spatial motion in videos. Our analysis relies on a linear approximation related to the brightness constancy assumption used in optical flow formulations. We also derive the conditions under which this approximation holds. This leads to a multiscale approach to magnify motion without feature tracking or motion estimation.

Previous attempts have been made to unveil imperceptible motions in videos. [9] analyze and amplify subtle motions and visualize deformations that would otherwise be invisible. [15] propose using the Cartoon Animation Filter to create perceptually appealing motion exaggeration. These approaches follow a *Lagrangian* perspective, in reference to fluid dynamics where the trajectory of particles is tracked over time. As such, they rely on accurate motion estimation, which is computationally expensive and difficult to make artifact-free, especially at regions of occlusion boundaries and complicated motions. Moreover, Liu et al. [9] have shown that additional techniques,

including motion segmentation and image in-painting, are required to produce good quality synthesis. This increases the complexity of the algorithm further.

In contrast, we are inspired by the *Eulerian* perspective, where properties of a voxel of fluid, such as pressure and velocity, evolve over time. In our case, we study and amplify the variation of pixel values over time, in a spatially-multiscale manner. In our Eulerian approach to motion magnification, we do not explicitly estimate motion, but rather exaggerate motion by amplifying temporal color changes at fixed positions. We rely on the same differential approximations that form the basis of optical flow algorithms [10, 5].

Temporal processing has been used previously to extract invisible signals [12] and to smooth motions [4]. For example, Poh et al. [12] extract a heart rate from a video of a face based on the temporal variation of the skin color, which is normally invisible to the human eye. We uses temporal processing similarly to select signal of interest, but in addition, we extends it to translate color variation to spatial motion when amplifying motion. Fuchs et al. [4] use per-pixel temporal filters to dampen temporal aliasing of motion in videos. They also discuss the high-pass filtering of motion, but mostly for non-photorealistic effects and for large motions (Figure 11 in their paper). In contrast, our method strives to make imperceptible motions visible using a multiscale approach. We analyze our method theoretically and show that it applies only for small motions.

Spatial filtering has been used to rise signal-to-noise ratio (SNR) of pulse signal [14, 12]. Poh et al. [12] use spatial pooling of the whole face region to extract pulse signal and rely on Independent Component Analysis(ICA) to eliminate noise induced by motions for higher SNR. Whereas, we use localized spatial pooling and bandpass filtering to extract and reveal visually the signal corresponding to the pulse. This primal domain analysis allows us to amplify and visualize the pulse signal at each location on the face. This has important potential monitoring and diagnostic applications to medicine, where, for example, the asymmetry in facial blood flow can be a symptom of arterial problems. The localized spatial pooling feature of Eulerian approach also provides us the local SNR and onset time difference of pulse signal at each

location on the face. Given the information, we can treat each localized extracted signal differently to achieve more accurate pulse rate estimation.

Contributions

Our project consists two parts. First part focuses on visual magnification of subtle signals in the video and second part focuses on numerical extraction of vital signs.

(1) Visual magnification

(a) Nearly invisible changes in a dynamic environment can be revealed through *Eulerian* spatio-temporal processing of standard monocular video sequences. Moreover, for a range of amplification values that is suitable for various applications, explicit motion estimation is not required to amplify motion in natural videos. Our approach is robust and runs in real time

(b) We provide mathematical analysis of the link between temporal filtering and spatial motion and show that our method is best suited to small displacements and lower spatial frequencies

(c) An unified framework is presented to amplify both spatial motion and purely temporal changes, e.g., the heart pulse, and it can be adjusted to amplify particular temporal frequencies—a feature which is not supported by Lagrangian methods

(d) We analytically and empirically compare Eulerian and Lagrangian motion magnification approaches under different noisy conditions

To demonstrate our approach, we present several examples where our method makes subtle variations in a scene visible

(2) Numerical extraction of vital signs

(a) We collected a dataset consisting of recordings of 11 newborn baby subjects in Special Care Nursery of Winchester Hospital. Each subject has several

recordings with different activity levels, lighting conditions and recording angles. Recordings of vital sign readings generated by state-of-art vital sign measurement devices in hospital were captured at the same time with synchronization clues. A stereo camera set and a near infra-red camera were used to acquire recordings to get more visual information that may be useful for future study.

(b) We prototype a heart rate measurement system based on *Eulerian* video processing. Localized time-series of color values are temporally aligned and weighted averaged. Temporal alignments and weights are estimated from data according to the local SNR and onset time difference. The combined time-series can generate heart rate measurement which is more resistant to noise and motion.

To evaluate our system, we extract heart rate of newborns and compare with ground truth from the dataset we built. We demonstrate that our system can produce reliable heart rate estimation under bright lighting condition and mild motions.

Chapter 2

Space-time video processing

Our approach combines spatial and temporal processing to emphasize subtle temporal changes in a video. The process is illustrated in Figure 2-1. We first decompose the video sequence into different spatial frequency bands. These bands might be magnified differently because (a) they might exhibit different signal-to-noise ratios; or (b) they might contain spatial frequencies for which the linear approximation used in our motion magnification does not hold (Sect. 3). In the latter case, we reduce the amplification for these bands to suppress artifacts. When the goal of spatial processing is simply to increase temporal signal-to-noise ratio by pooling multiple pixels, we spatially low-pass filter the frames of the video and downsample them for computational efficiency. In the general case, however, we compute a full Laplacian pyramid [2].

We then perform temporal processing on each spatial band. We consider the time series corresponding to the value of a pixel in a frequency band and apply a bandpass filter to extract the frequency bands of interest. For example, we might select frequencies within 0.4-4Hz, corresponding to 24-240 beats per minute, if we wish to magnify a pulse. If we are able to extract the pulse rate, we can use a narrow band around that value. The temporal processing is uniform for all spatial levels, and for all pixels within each level. We then multiply the extracted bandpassed signal by a magnification factor α. This factor can be specified by the user, and may be attenuated automatically according to guidelines in Sect. 3.2. Possible temporal filters

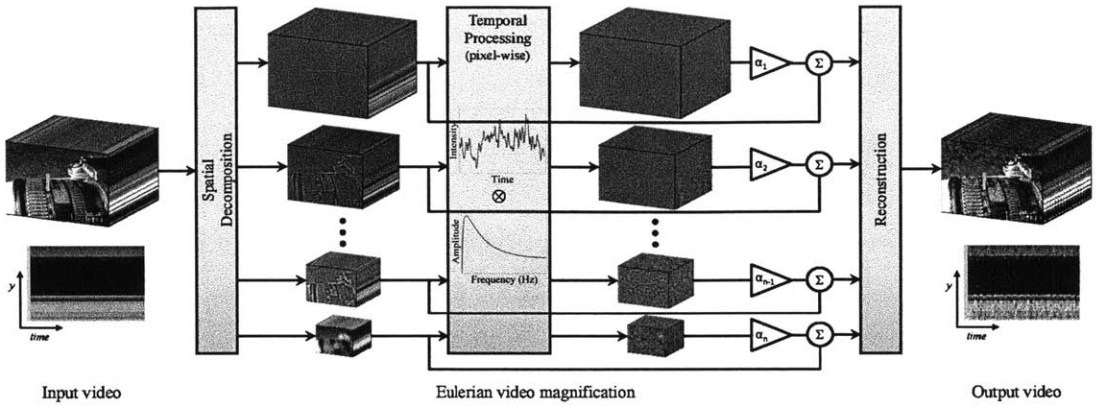


Figure 2-1: Overview of the Eulerian video magnification framework. The system first decomposes the input video sequence into different spatial frequency bands, and applies the same temporal filter to all bands. The filtered spatial bands are then amplified by a given factor α, added back to the original signal, and collapsed to generate the output video. The choice of temporal filter and amplification factors can be tuned to support different applications. For example, we use the system to reveal unseen motions of a Digital SLR camera, caused by the flipping mirror during a photo burst (*camera*; full sequences are available in the supplemental video).

are discussed in Sect. 4. We then add the magnified signal to the original and collapse the spatial pyramid to obtain the final output. Since natural videos are spatially and temporally smooth, and since our filtering is performed uniformly over the pixels, our method implicitly maintains spatiotemporal coherency of the results.

Chapter 3

Eulerian motion magnification

Our processing can amplify small motion even though we do not track motion as in Lagrangian methods [9, 15]. In this section, we show how temporal processing produces motion magnification using an analysis that relies on the first-order Taylor series expansions common in optical flow analyses [10, 5].

3.1 First-order motion

To explain the relationship between temporal processing and motion magnification, we consider the simple case of a 1D signal undergoing translational motion. This analysis generalizes directly to locally-translational motion in 2D.

Let $I(x,t)$ denote the image intensity at position x and time t. Since the image undergoes translational motion, we can express the observed intensities with respect to a displacement function $\delta(t)$, such that $I(x,t) = f(x+\delta(t))$ and $I(x,0) = f(x)$. The goal of motion magnification is to synthesize the signal

$$\hat{I}(x,t) = f(x+(1+\alpha)\delta(t)) \tag{3.1}$$

for some amplification factor α.

Assuming the image can be approximated by a first-order Taylor series expansion,

we write the image at time t, $f(x+\delta(t))$ in a first-order Taylor expansion about x, as

$$I(x,t) \approx f(x) + \delta(t)\frac{\partial f(x)}{\partial x} \tag{3.2}$$

Let $B(x,t)$ be the result of applying a broadband temporal bandpass filter to $I(x,t)$ at every position x (picking out everything except $f(x)$ in Eq. 3.2). For now, let us assume the motion signal, $\delta(t)$, is within the passband of the temporal bandpass filter (we will relax that assumption later). Then we have

$$B(x,t) = \delta(t)\frac{\partial f(x)}{\partial x} \tag{3.3}$$

In our process, we then amplify that bandpass signal by α and add it back to $I(x,t)$, resulting in the processed signal

$$\tilde{I}(x,t) = I(x,t) + \alpha B(x,t) \tag{3.4}$$

Combining Eqs. 3.2, 3.3, and 3.4, we have

$$\tilde{I}(x,t) \approx f(x) + (1+\alpha)\delta(t)\frac{\partial f(x)}{\partial x}. \tag{3.5}$$

Assuming the first-order Taylor expansion holds for the amplified larger perturbation, $(1+\alpha)\delta(t)$, we can relate the amplification of the temporally bandpassed signal to motion magnification. The processed output is simply

$$\tilde{I}(x,t) \approx f(x + (1+\alpha)\delta(t)) \tag{3.6}$$

This shows that the processing magnifies motions—the spatial displacement $\delta(t)$ of the local image $f(x)$ at time t, has been amplified to a magnitude of $(1+\alpha)$.

This process is illustrated for a single sinusoid in Figure 3-1. For a low frequency cosine wave and a relatively small displacement, $\delta(t)$, the first-order Taylor series expansion serves as a good approximation for the translated signal at time $t+1$. When boosting the temporal signal by α and adding it back to $I(x,t)$, we approximate that

wave translated by $(1+\alpha)\delta$.

For completeness, let us return to the more general case where $\delta(t)$ is not entirely within the passband of the temporal filter. In this case, let $\delta_k(t)$, indexed by k, represent the different temporal spectral components of $\delta(t)$. Each $\delta_k(t)$ will be attenuated by the temporal filtering by a factor γ_k. This results in a bandpassed signal,

$$B(x,t) = \sum_k \gamma_k \delta_k(t) \frac{\partial f(x)}{\partial x} \tag{3.7}$$

(compare with Eq. 3.3). Because of the multiplication in Eq. 3.4, this temporal frequency dependent attenuation can equivalently be interpreted as a frequency-dependent motion magnification factor, $\alpha_k = \gamma_k \alpha$, resulting in a motion magnified output,

$$\tilde{I}(x,t) \approx f(x + \sum_k (1+\alpha_k)\delta_k(t)) \tag{3.8}$$

The result is as would be expected for a linear analysis: the modulation of the spectral components of the motion signal becomes the modulation factor in the motion amplification factor, α_k, for each temporal subband, δ_k, of the motion signal.

3.2 Bounds

In practice, the assumptions in Sect. 3.1 hold for smooth images and small motions. For quickly changing image functions (i.e., high spatial frequencies), $f(x)$, the first-order Taylor series approximations becomes inaccurate for large values of the perturbation, $1 + \alpha\delta(t)$, which increases both with larger magnification α and motion $\delta(t)$. Figures 3-2 and 3-3 demonstrate the effect of higher frequencies, larger amplification factors and larger motions on the motion-amplified signal of a sinusoid.

As a function of spatial frequency, ω, we can derive a guide for how large the motion amplification factor, α, can be, given the observed motion $\delta(t)$. For the processed signal, $\tilde{I}(x,t)$ to be approximately equal to the true magnified motion,

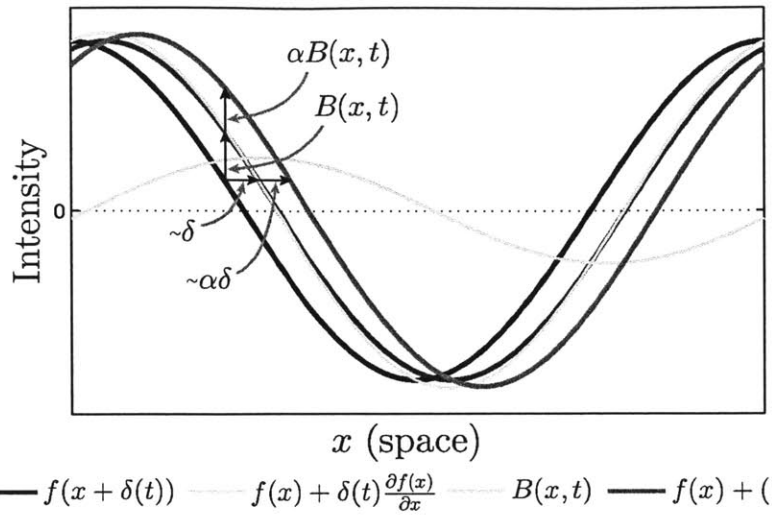


Figure 3-1: Temporal filtering can approximate spatial translation. This effect is demonstrated here on a 1D signal, but equally applies to 2D. The input signal is shown at two time instants: $I(x,t) = f(x)$ at time t and $I(x,t+1) = f(x+\delta)$ at time $t+1$. The first-order Taylor series expansion of $I(x,t+1)$ about x approximates well the translated signal. The temporal bandpass is amplified and added to the original signal to generate a larger translation. In this example $\alpha = 1$, magnifying the motion by 100%, and the temporal filter is a finite difference filter, subtracting the two curves.

$\hat{I}(x,t)$, we seek the conditions under which

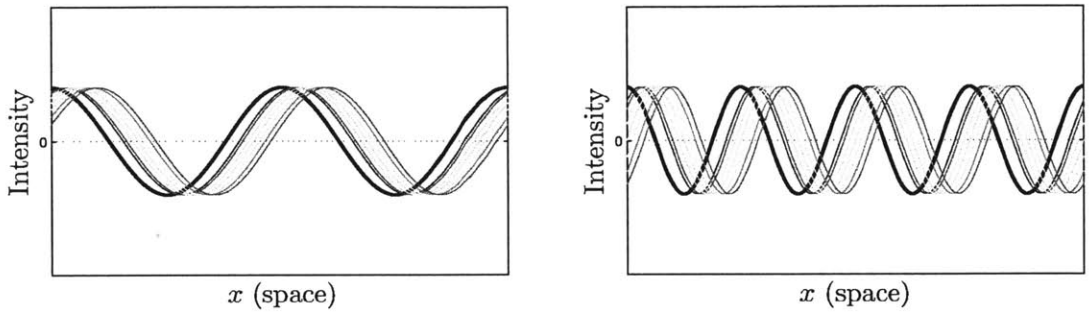
$$\begin{aligned} \tilde{I}(x,t) &\approx \hat{I}(x,t) \\ \Rightarrow f(x) + (1+\alpha)\delta(t)\frac{\partial f(x)}{\partial x} &\approx f(x + (1+\alpha)\delta(t)) \end{aligned} \tag{3.9}$$

Let $f(x) = \cos(\omega x)$ for spatial frequency ω, and denote $\beta = 1 + \alpha$. We require that

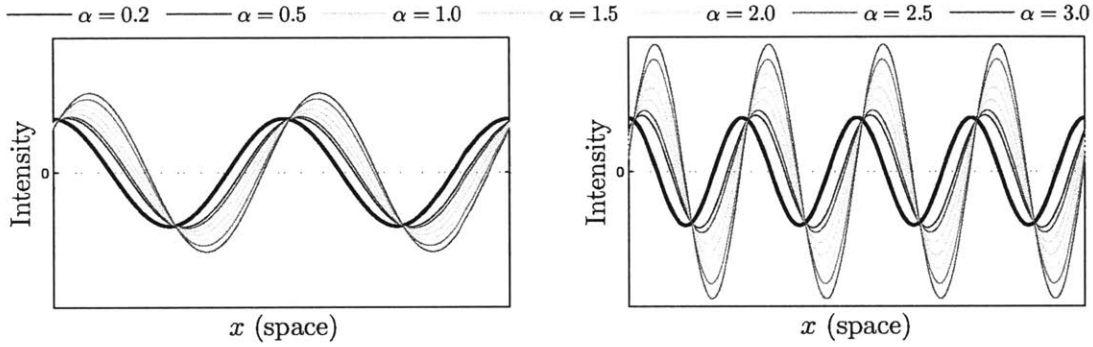
$$\cos(\omega x) - \beta\omega\delta(t)\sin(\omega x) \approx \cos(\omega x + \beta\omega\delta(t)) \tag{3.10}$$

Using the addition law for cosines, we have

$$\begin{aligned} &\cos(\omega x) - \beta\omega\delta(t)\sin(\omega x) = \\ &\qquad \cos(\omega x)\cos(\beta\omega\delta(t)) - \sin(\omega x)\sin(\beta\omega\delta(t)) \end{aligned} \tag{3.11}$$



(a) True motion amplification: $\hat{I}(x,t) = f(x + \alpha\delta(t))$.



(b) Motion amplification via temporal filtering: $\tilde{I}(x,t) = I(x,t) + \alpha B(x,t)$.

Figure 3-2: Illustration of motion amplification on a 1D signal for different spatial frequencies and α values. For the images on the left side, $\lambda = 2\pi$ and $\delta(1) = \frac{\pi}{8}$ is the true translation. For the images on the right side, $\lambda = \pi$ and $\delta(1) = \frac{\pi}{8}$. (a) The true displacement of $I(x,0)$ by $(1 + \alpha)\delta(t)$ at time $t = 1$, colored from blue (small amplification factor) to red (high amplification factor). (b) The amplified displacement produced by our filter, with colors corresponding to the correctly shifted signals in (a). Referencing Eq. 3.14, the red (far right) curves of each plot correspond to $(1+\alpha)\delta(t) = \frac{\lambda}{4}$ for the left plot, and $(1+\alpha)\delta(t) = \frac{\lambda}{2}$ for the right plot, showing the mild, then severe, artifacts introduced in the motion magnification from exceeding the bound on $(1 + \alpha)$ by factors of 2 and 4, respectively.

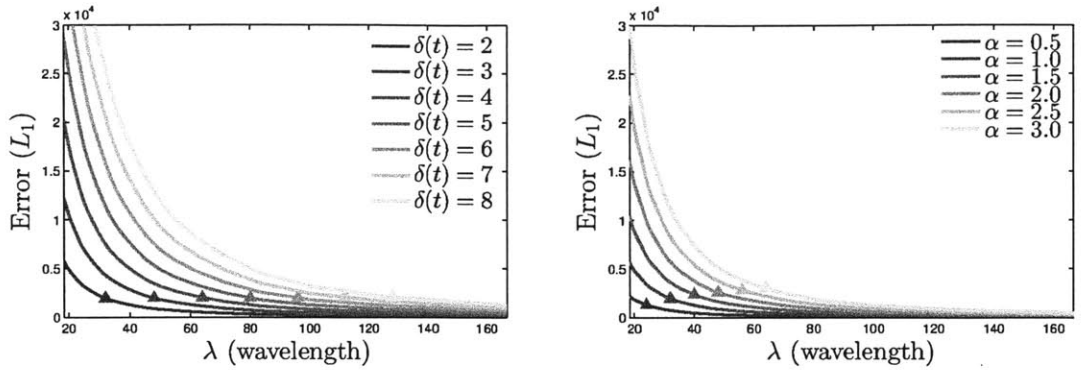


Figure 3-3: Motion magnification error, computed as the L_1-norm between the true motion-amplified signal (Figure 3-2(a)) and the temporally-filtered result (Figure 3-2(b)), as function of wavelength, for different values of $\delta(t)$ (a) and α (b). In (a), we fix $\alpha = 1$, and in (b), $\delta(t) = 2$. The markers on each curve represent the derived cutoff point $(1 + \alpha)\delta(t) = \frac{\lambda}{8}$ (Eq. 3.14).

Hence, the following should approximately hold

$$\cos(\beta\omega\delta(t)) \approx 1 \tag{3.12}$$

$$\sin(\beta\omega\delta(t)) \approx \beta\delta(t)\omega \tag{3.13}$$

The small angle approximations of Eqs. (3.12) and (3.13) will hold to within 10% for $\beta\omega\delta(t) \leq \frac{\pi}{4}$ (the sine term is the leading approximation and we have $\sin(\frac{\pi}{4}) = 0.9\frac{\pi}{4}$). In terms of the spatial wavelength, $\lambda = \frac{2\pi}{\omega}$, of the moving signal, this gives

$$(1 + \alpha)\delta(t) < \frac{\lambda}{8}. \tag{3.14}$$

Eq. 3.14 above provides the guideline we seek, giving the largest motion amplification factor, α, compatible with accurate motion magnification of a given video motion $\delta(t)$ and image structure spatial wavelength, λ. Figure 3-2 (b) shows the motion magnification errors for a sinusoid when we boost α beyond the limit in Eq. 3.14. In some videos, violating the approximation limit can be perceptually preferred and we leave the λ cutoff as a user-modifiable parameter in the multiscale processing.

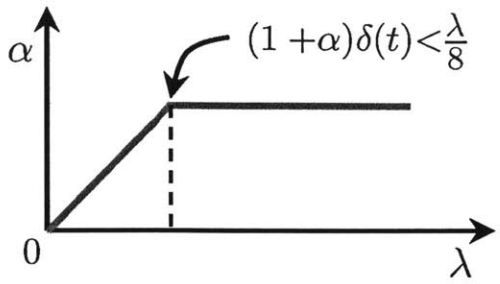


Figure 3-4: Amplification factor, α, as function of spatial wavelength λ, for amplifying motion. The amplification factor is fixed to α for spatial bands that are within our derived bound (Eq. 3.14), and is attenuated linearly for higher spatial frequencies.

3.3 Multiscale analysis

The analysis in Sect. 3.2 suggests a *scale-varying* process: use a specified α magnification factor over some desired band of spatial frequencies, then scale back for the high spatial frequencies (found from Eq. 3.14 or specified by the user) where amplification would give undesirable artifacts. Figure 3-4 shows such a modulation scheme for α. Although areas of high spatial frequencies (sharp edges) will be generally amplified less than lower frequencies, we found the resulting videos to contain perceptually appealing magnified motion. Such effect was also exploited in the earlier work of Freeman et al. [3] to create the illusion of motion from still images.

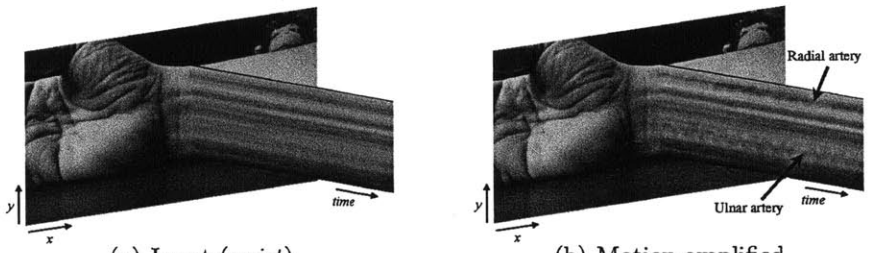
Chapter 4

Results

The results were generated using non-optimized MATLAB code on a machine with a six-core processor and 32 GB RAM. The computation time per video was on the order of a few minutes. We used a separable *binomial filter* of size five to construct the video pyramids. We also built a prototype application that allows users to reveal subtle changes in real-time from live video feeds, essentially serving as a microscope for temporal variations. It is implemented in C++, is entirely CPU-based, and processes 640×480 videos at 45 frames per second on a standard laptop. It can be sped up further by utilizing GPUs. A demo of the application is available in the accompanying video. The code is available on the project webpage.

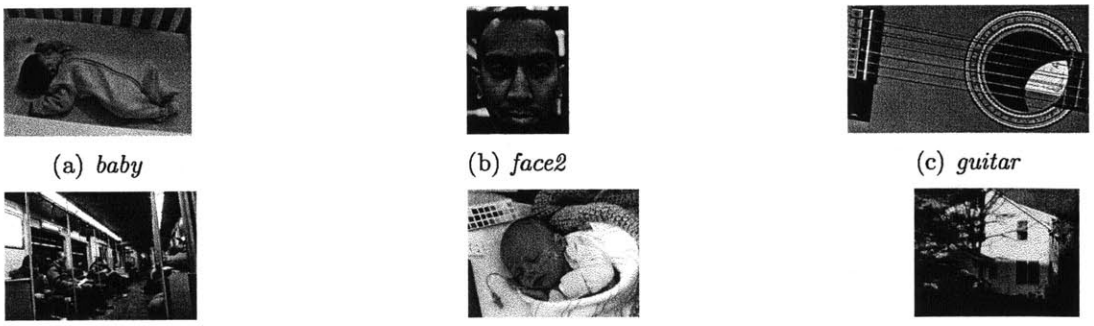
Given an input video to process by Eulerian video magnification, there are four steps the user needs to take: (1) select a temporal bandpass filter; (2) select an amplification factor, α; (3) select a spatial frequency cutoff (specified by spatial wavelength, λ_c) beyond which an attenuated version of α is used; and (4) select the form of the attenuation for α—either force α to zero for all $\lambda < \lambda_c$, or linearly scale α down to zero. The frequency band of interest can be chosen automatically in some cases, but it is often important for users to be able to control the frequency band corresponding to their application. In our real-time application, the amplification factor and cutoff frequencies are all customizable by the user.

We first select the temporal bandpass filter to pull out the motions or the signal desired to be amplified (step 1 above). The choice of filter is generally application



(a) Input (*wrist*) (b) Motion-amplified

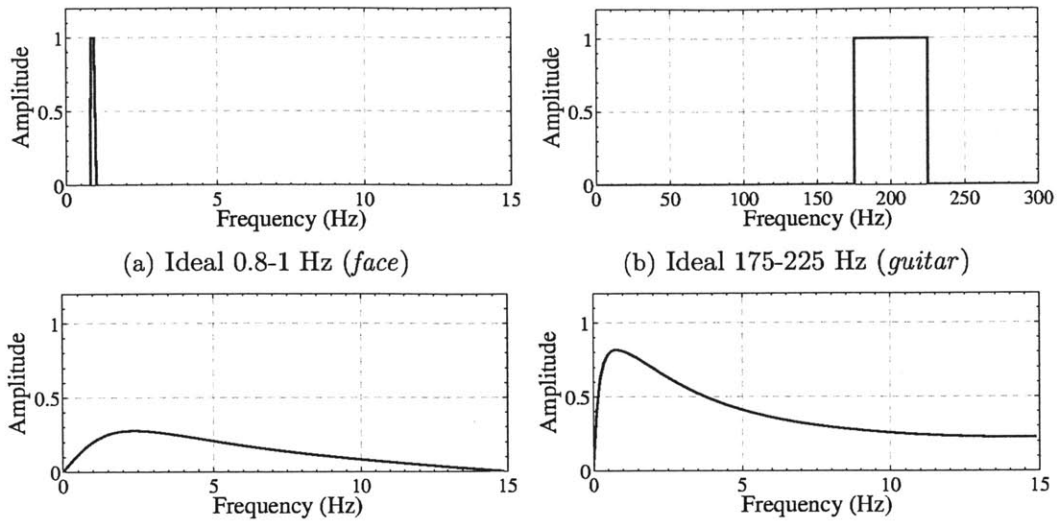
Figure 4-1: Eulerian video magnification used to amplify subtle motions of blood vessels arising from blood flow. For this video we tuned the temporal filter to a frequency band that includes the heart rate—0.88 Hz (53 bpm)—and set the amplification factor to $\alpha = 10$. To reduce motion magnification of irrelevant objects, we applied a user-given mask to amplify the area near the wrist only. Movement of the radial and ulnar arteries can barely be seen in the input video (a) taken with a standard point-and-shoot camera, but is significantly more noticeable in the motion-magnified output (b). The motion of the pulsing arteries is more visible when observing a spatio-temporal YT slice of the wrist (a) and (b). The full *wrist* sequence can be found in the supplemental video.



(a) *baby* (b) *face2* (c) *guitar*

(d) *subway* (e) *baby2* (f) *shadow*

Figure 4-2: Representative frames from additional videos demonstrating our technique, which can be found in the accompanying video and webpage.



(a) Ideal 0.8-1 Hz (*face*)

(b) Ideal 175-225 Hz (*guitar*)

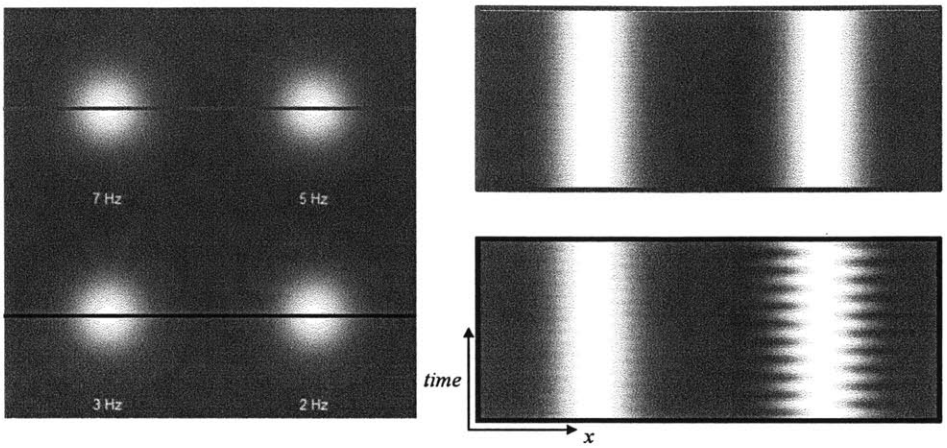
(c) Butterworth 3.6-6.2 Hz (*subway*)

(d) Second-order IIR (pulse detection)

Figure 4-3: Temporal filters used in the paper. The ideal filters (a) and (b) are implemented using DCT. The Butterworth filter (c) is used to convert a user-specified frequency band to a second-order IIR structure and is used in our real-time application. The second-order IIR filter (d) also allows user input. These second-order filters have a broader passband than an ideal filter.

dependent. For motion magnification, a filter with a broad passband is preferred; for color amplification of blood flow, a narrow passband produces a more noise-free result. Figure 4-3 shows the frequency responses of some of the temporal filters used in this paper. We use ideal bandpass filters for color amplification, since they have passbands with sharp cutoff frequencies. Low-order IIR filters can be useful for both color amplification and motion magnification and are convenient for a real-time implementation. In general, we used two first-order lowpass IIR filters with cutoff frequencies ω_l and ω_h to construct an IIR bandpass filter.

Next, we select the desired magnification value, α, and spatial frequency cutoff, λ_c (steps 2 and 3). While Eq. 3.14 can be used as a guide, in practice, we may try various α and λ_c values to achieve a desired result. Users can select a higher α that violates the bound to exaggerate specific motions or color changes at the cost of increasing noise or introducing more artifacts. In some cases, one can account for color clipping artifacts by attenuating the chrominance components of each frame. Our approach achieves this by doing all the processing in the YIQ space. Users can attenuate the chrominance components, I and Q, before conversion to the original color space.



(a) Input (*sim4*) (b) Motion-amplified spatiotemporal slices

Figure 4-4: Selective motion amplification on a synthetic sequence (*sim4* on left). The video sequence contains blobs oscillating at different temporal frequencies as shown on the input frame. We apply our method using an ideal temporal bandpass filter of 1-3 Hz to amplify only the motions occurring within the specified passband. In (b), we show the spatio-temporal slices from the resulting video which show the different temporal frequencies and the amplified motion of the blob oscillating at 2 Hz. We note that the space-time processing is applied uniformly to all the pixels. The full sequence and result can be found in the supplemental video.

For human pulse color amplification, where we seek to emphasize low spatial frequency changes, we may force $\alpha = 0$ for spatial wavelengths above λ_c. For motion magnification videos, we can choose to use a linear ramp transition for α (step 4).

We evaluated our method for color amplification using a few videos: two videos of adults with different skin colors and one of a newborn baby. An adult subject with lighter complexion is shown in *face* (Figure 1-1), while an individual with darker complexion is shown in *face2* (Figure 4-2). In both videos, our objective was to amplify the color change as the blood flows through the face. In both *face* and *face2*, we applied a Laplacian pyramid and set α for the finest two levels to 0. Essentially, we downsampled and applied a spatial lowpass filter to each frame to reduce both quantization and noise and to boost the subtle pulse signal that we are interested in. For each video, we then passed each sequence of frames through an ideal bandpass filter with a passband of 0.83 Hz to 1 Hz (50 bpm to 60 bpm). Finally, a large value of $\alpha \approx 100$ and $\lambda_c \approx 1000$ was applied to the resulting spatially lowpass signal to emphasize the color change as much as possible. The final video was formed by adding

this signal back to the original. We see periodic green to red variations at the heart rate and how blood perfuses the face.

baby2 is a video of a newborn recorded *in situ* at the Nursery Department at Winchester Hospital in Massachusetts. In addition to the video, we obtained ground truth vital signs from a hospital-grade monitor. We used this information to confirm the accuracy of our heart rate estimate and to verify that the color amplification signal extracted from our method matches the photoplethysmogram, an optically obtained measurement of the perfusion of blood to the skin, as measured by the monitor.

To evaluate our method for motion magnification, we used several different videos: *face* (Figure 1-1), *sim4* (Figure 4-4), *wrist* (Figure 4-1), *camera* (Figure 2-1), *face2*, *guitar*, *baby*, *subway*, *shadow*, and *baby2* (Figure 4-2). For all videos, we used standard Laplacian pyramid for spatial filtering. For videos where we wanted to emphasize motions at specific temporal frequencies (e.g., in *sim4* and *guitar*), we used ideal bandpass filters. In *sim4* and *guitar*, we were able to selectively amplify the motion of a specific blob or guitar string by using a bandpass filter tuned to the oscillation frequency of the object of interest. These effects can be observed in the supplemental video. The values used for α and λ_c are shown in Table 4.1.

For videos where we were interested in revealing broad, but subtle motion, we used temporal filters with a broader passband. For example, for the *face2* video, we used a second-order IIR filter with slow roll-off regions. By changing the temporal filter, we can magnify the motion of the head instead of amplifying skin color change. Accordingly, $\alpha = 20, \lambda_c = 80$ were chosen to magnify the motion.

By using broadband temporal filters and setting α and λ_c according to the bound, our method is able to reveal invisible motions, as in the *camera* and *wrist* videos. For the *camera* video, we used a high-speed camera with a sampling rate of 300 Hz to record a Digital SLR camera capturing photos at about 1 exposure per second. The vibration, caused by the moving mirror in the SLR, though invisible to the naked eye, is revealed by our Eulerian video magnification approach. To verify that the vibrations amplified are indeed caused by the flipping mirror in the SLR, we secured a laser pointer to the camera and recorded a video of the laser light, appearing at a

Table 4.1: Table of $\alpha, \lambda_c, \omega_l, \omega_h$ values used to produce the various output videos. For *face2*, two different sets of parameters are used—one for amplifying pulse, another for amplifying motion. For *guitar*, different cutoff frequencies and values for (α, λ_c) are used to "select" the different oscillating guitar strings. f_s is the frame rate of the camera.

Video	α	λ_c	ω_l (Hz)	ω_h (Hz)	f_s (Hz)
baby	10	16	0.4	3	30
baby2	150	600	2.33	2.67	30
camera	120	20	45	100	300
face	100	1000	0.83	1	30
face2 motion	20	80	0.83	1	30
face2 pulse	120	960	0.83	1	30
guitar Low E	50	40	72	92	600
guitar A	100	40	100	120	600
shadow	5	48	0.5	10	30
subway	60	90	3.6	6.2	30
wrist	10	80	0.4	3	30

distance of 3 to 4 meters from the source. At that distance, the laser light visibly oscillates with each exposure, and the oscillations were in sync with the magnified motions.

Our method is also able to exaggerate visible, yet subtle motion, as seen in the *baby*, *face2*, and *subway* videos. In the subway example we deliberately amplified the motion beyond the derived bounds of where the 1st order approximation holds in order to increase the effect and to demonstrate the algorithm's artifacts. We note that most of the examples in our paper contain oscillatory movements because such motion generally has longer duration and smaller amplitudes. However, our method can be used to amplify non-periodic motions as well, as long as they are within the passband of the temporal bandpass filter. In *shadow*, for example, we process a video of the sun's shadow moving linearly yet imperceptibly over 15 seconds. The magnified version makes it possible to see the change even within this short time period.

Finally, some videos may contain regions of temporal signals that do not need amplification, or that, when amplified, are perceptually unappealing. Due to our Eulerian processing, we can easly allow the user to manually restrict magnification to particular areas by marking them on the video (this was used for *face* and *wrist*).

Chapter 5

Discussion

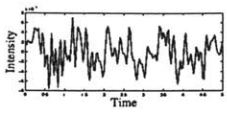
5.1 Sensitivity to Noise.

The amplitude variation of the signal of interest is often much smaller than the noise inherent in the video. In such cases direct enhancement of the pixel values will not reveal the desired signal. Spatial filtering can be used to enhance these subtle signals. However, if the spatial filter applied is not large enough, the signal of interest will not be revealed (Figure 5-1).

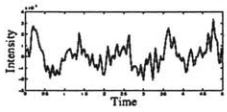
Assuming that the noise is zero-mean white and wide-sense stationary with respect to space, it can be shown that spatial low pass filtering reduces the variance of the noise according to the area of the low pass filter. In order to boost the power of a specific signal, e.g., the pulse signal in the face, we can use the spatial characteristics of the signal to estimate the spatial filter size.

Let the noise power level be σ^2, and our prior on signal power over spatial frequencies be $S(\lambda)$. We want to find a spatial low pass filter with radius r such that the signal power is greater than the noise in the filtered frequency region. The wavelength cut off of such a filter is proportional to its radius, r, so the signal prior can be represented as $S(r)$. The noise power σ^2 can be estimated by examining pixel values in a stable region of the scene, from a gray card, or by using a technique as in [8]. Since the filtered noise power level, σ'^2, is inversely proportional to r^2, we can solve

(a) Input noisy frame



(b) Insufficient spatial pooling



(c) Sufficient spatial pooling

Figure 5-1: Proper spatial pooling is imperative for revealing the signal of interest. (a) A frame from the *face* video (Figure 1-1) with white Gaussian noise ($\sigma = 0.1$ pixel) added. On the right are intensity traces over time for the pixel marked blue on the input frame, where (b) shows the trace obtained when the (noisy) sequence is processed with the same spatial filter used to process the original *face* sequence, a separable binomial filter of size 20, and (c) shows the trace when using a filter tuned according to the estimated radius in Eq. 5.1, a binomial filter of size 80. The pulse signal is not visible in (b), as the noise level is higher than the power of the signal, while in (c) the pulse is clearly visible (the periodic peaks about one second apart in the trace).

the following equation for r,

$$S(r) = \sigma'^2 = k\frac{\sigma^2}{r^2} \tag{5.1}$$

where k is a constant that depends on the shape of the low pass filter. This equation gives an estimate for the size of the spatial filter needed to reveal the signal at a certain noise power level.

5.2 Eulerian vs. Lagrangian Processing.

Because the two methods take different approaches to motion—Lagrangian approaches explicitly track motions, while our Eulerian approach does not—they can be used for complementary motion domains. Lagrangian approaches, e.g. [9], work better to enhance motions of fine point features and support larger amplification factors, while our Eulerian method is better suited to smoother structures and small amplifications. We note that our technique does not assume particular types of motions. The first-order Taylor series analysis can hold for general small 2D motions along general paths.

In Sect. 5.3, we further derive estimates of the accuracy of the two approaches

with respect to noise. Comparing the Lagrangian error, ε_L (Eq. 5.15), and the Eulerian error, ε_E (Eq. 5.17), we see that both methods are equally sensitive to the temporal characteristics of the noise, n_t, while the Lagrangian process has additional error terms proportional to the spatial characteristics of the noise, n_x, due to the explicit estimation of motion (Eq. 5.13). The Eulerian error, on the other hand, grows quadratically with α, and is more sensitive to large spatial frequencies (I_{xx}). In general, this means that Eulerian magnification would be preferable over Lagrangian magnification for small amplifications and larger noise levels.

We validated this analysis on a synthetic sequence of a 2D cosine oscillating at 2 Hz temporally and 0.1 pixels spatially with additive white spatiotemporal Gaussian noise of zero mean and standard deviation σ (Figure 5-2). The results match the relationship of error-to-noise and error-to-amplification predicted by the derivation (Figure 5-2(b)), as well as the expected region where the Eulerian approach outpeforms the Lagrangian results (Figure5-2(a)-left). The Lagrangian method is indeed more sensitive to increases in spatial noise, while the Eulerian error is hardly affected by it (Figure 5-2(c)). While different regularization schemes used for motion estimation (that are harder to analyze theoretically) may alleviate the Lagrangian error, they did not change the result significantly (Figure 5-2(a)-right). In general, our experiments show that for small amplifications the Eulerian approach strikes a better balance between performance and efficiency. Comparisons between the methods on natural videos are available on the project webpage.

5.3 Eulerian and Lagrangian Motion Magnification Error

In this section we derive estimates of the error in the Eulerian and Lagrangian motion magnification results with respect to spatial and temporal noise. The derivation is done again for the 1D case for simplicity, and can be generalized to 2D. We use the same setup as in Sect. 3.1.

Both methods only approximate the true motion-amplified sequence, $\hat{I}(x,t)$ (Eq. 3.1). Let us first analyze the error in those approximations on the clean signal, $I(x,t)$.

Without noise. In the Lagrangian approach, the motion-amplified sequence, $\tilde{I}_L(x,t)$, is achieved by directly amplifying the estimated motion, $\tilde{\delta}(t)$, with respect to the reference frame, $I(x,0)$

$$\tilde{I}_L(x,t) = I(x + (1+\alpha)\tilde{\delta}(t), 0) \tag{5.2}$$

In its simplest form, we can estimate $\delta(t)$ in a point-wise manner (See Sect. 5 for discussion on spatial regularization)

$$\tilde{\delta}(t) = \frac{I_t(x,t)}{I_x(x,t)} \tag{5.3}$$

where $I_x(x,t) = \partial I(x,t)/\partial x$ and $I_t(x,t) = I(x,t) - I(x,0)$. From now on, we will omit the space (x) and time (t) indices when possible for brevity.

The error in in the Lagrangian solution is directly determined by the error in the estimated motion, which we take to be second-order term in the brightness constancy equation (usually not paid in optical flow formulations),

$$\begin{aligned} I(x,t) &\approx I(x,0) + \delta(t)I_x + \frac{1}{2}\delta^2(t)I_{xx} \\ \Rightarrow \frac{I_t}{I_x} &\approx \delta(t) + \frac{1}{2}\delta^2(t)I_{xx} \end{aligned} \tag{5.4}$$

The estimated motion, $\tilde{\delta}(t)$, is thus related to the true motion, $\delta(t)$, by

$$\tilde{\delta}(t) \approx \delta(t) + \frac{1}{2}\delta^2(t)I_{xx} \tag{5.5}$$

Plugging (5.5) in (5.2) and using a Taylor expansion of I about $x + (1+\alpha)\delta(t)$, we have

$$\tilde{I}_L(x,t) \approx I(x + (1+\alpha)\delta(t), 0) + \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x \tag{5.6}$$

Subtracting (3.1) from (5.6), the error in the Lagrangian motion-magnified sequence,

ε_L, is

$$\varepsilon_L \approx \left|\frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x\right| \tag{5.7}$$

In our Eulerian approach, the magnified sequence, $\hat{I}_E(x,t)$, is computed as

$$\tilde{I}_E(x,t) = I(x,t) + \alpha I_t(x,t) = I(x,0) + (1+\alpha)I_t(x,t) \tag{5.8}$$

similar to Eq. 3.4, using a two-tap temporal filter to compute I_t.

Using a Taylor expansion of the true motion-magnified sequence, $\hat{I}$ (Eq. 3.1), about x, we have

$$\hat{I}(x,t) \approx I(x,0) + (1+\alpha)\delta(t)I_x + \frac{1}{2}(1+\alpha)^2\delta^2(t)I_{xx} \tag{5.9}$$

Using (5.4) and subtracting (3.1) from (5.9), the error in the Eulerian motion-magnified sequence, ε_E, is

$$\varepsilon_E \approx \left|\frac{1}{2}(1+\alpha)^2\delta^2(t)I_{xx} - \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x\right| \tag{5.10}$$

With noise. Let $I'(x,t)$ be the noisy signal, such that

$$I'(x,t) = I(x,t) + n(x,t) \tag{5.11}$$

for additive noise $n(x,t)$.

The estimated motion in the Lagrangian approach becomes

$$\tilde{\delta}(t) = \frac{I'_t}{I'_x} = \frac{I_t + n_t}{I_x + n_x} \tag{5.12}$$

where $n_x = \partial n/\partial x$ and $n_t = n(x,t) - n(x,0)$.

Using a Taylor Expansion on (n_t, n_x) about $(0,0)$ (zero noise), and using (5.4), we have

$$\tilde{\delta}(t) \approx \delta(t) + \frac{n_t}{I_x} - n_x\frac{I_t}{I_x^2} + \frac{1}{2}\delta^2(t)I_{xx} \tag{5.13}$$

Plugging (5.13) into (5.2), and using a Taylor expansion of I about $x + (1+\alpha)\delta(t)$,

we get

$$\tilde{I}'_L(x,t) \approx I(x+(1+\alpha)\delta(t),0)+ \\ (1+\alpha)I_x(\frac{n_t}{I_x} - n_x\frac{I_t}{I_x^2} + \frac{1}{2}\delta^2(t)I_{xx})) + n \tag{5.14}$$

Using (5.5) again and subtracting (3.1), the Lagrangian error as function of noise, $\varepsilon_L(n)$, is

$$\varepsilon_L(n) \approx \Big|(1+\alpha)n_t - (1+\alpha)n_x\delta(t) \\ - \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}n_x + \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x + n\Big| \tag{5.15}$$

In the Eulerian approach, the noisy motion-magnified sequence becomes

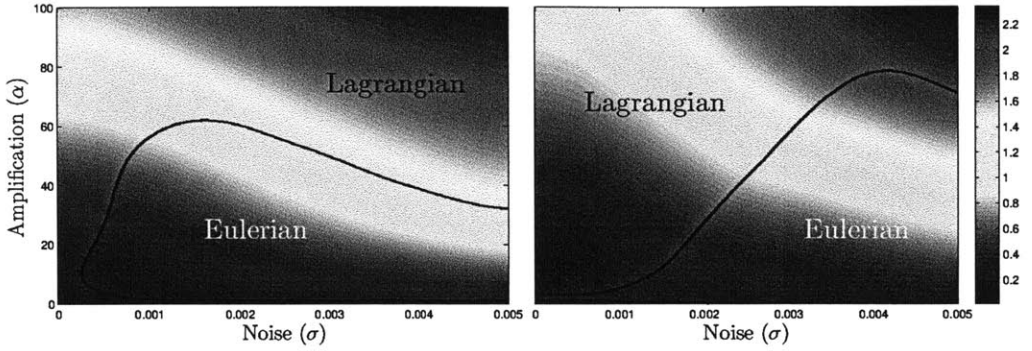
$$\tilde{I}'_E(x,t) = I'(x,0) + (1+\alpha)I'_t = I(x,0) + (1+\alpha)(I_t + n_t) + n \tag{5.16}$$

Using (5.10) and subtracting (3.1), the Eulerian error as function of noise, $\varepsilon_E(n)$, is

$$\varepsilon_E(n) \approx \Big|(1+\alpha)n_t + \frac{1}{2}(1+\alpha)^2\delta^2(t)I_{xx} - \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x + n\Big| \tag{5.17}$$

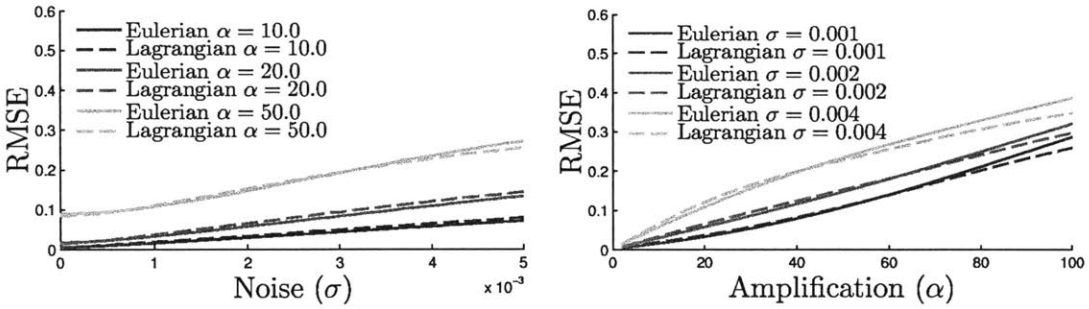
Notice that setting zero noise in (5.15) and (5.17), we get the corresponding errors derived for the non-noisy signal in (5.7) and (5.10).

The detailed steps of derivation can be found in Appendix A.

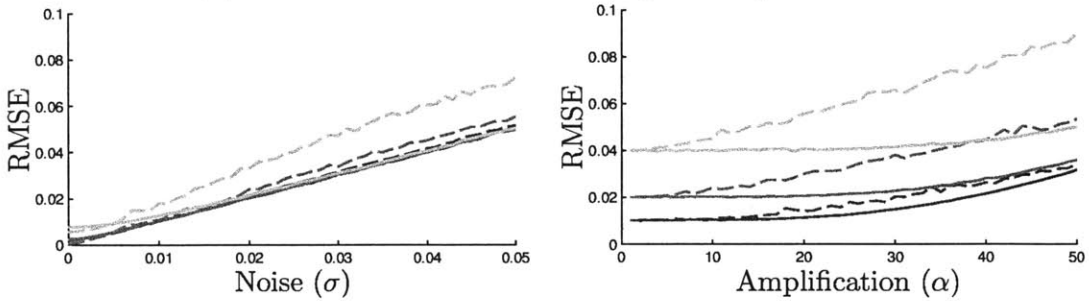


No regularization With regularization

(a) $\min(\varepsilon_E, \varepsilon_L)$, spatiotemporal noise



(b) Error as function of σ and α, Spatiotemporal noise



(c) Error as function of σ and α, Spatial noise only

Figure 5-2: Comparison between Eulerian and Lagrangian motion magnification on a synthetic sequence with additive noise. (a) The minimal error, $\min(\varepsilon_E, \varepsilon_L)$, computed as the (frame-wise) RMSE between each method's result and the true motion-magnified sequence, as function of noise and amplification, colored from blue (small error) to red (large error), with (left) and without (right) spatial regularization in the Lagrangian method. The black curves mark the intersection between the error surfaces, and the overlayed text indicate the best performing method in each region. (b) RMSE of the two approaches as function of noise (left) and amplification (right). (d) Same as (c), using spatial noise only.

Chapter 6

Vital signs extraction

6.1 Introduction

By showing how Eulerian video magnification can be used to reveal color changes caused by blood circulation, we have demonstrated the potential for extracting medical information from videos instead of contact sensors. Given the earlier success on measuring heart rate and respiration rate from videos [12, 14], we are motivated to explore how to use *Eulerian*-based spatio-temporal processing to achieve more accurate heart rate estimation under noise and motion.

In this chapter, we first describe some human physiological factors that suggest the feasibility of vital signs extraction from videos. Secondly, we describe the kind of data we acquired for evaluating the accuracy of vital signs estimation and how we collected them. Lastly, we present an early attempt for heart rate extraction that uses *Eulerian*-based spatio-temporal processing. Our method combines the localized time series of color values produced by our Eulerian processing according to their local characteristics to achieve an accurate heart rate estimate.

6.1.1 Background

Modern vital signs monitoring systems typically require physical contact with patients to obtain vital signs measurements. For example, the state-of-the-art technique for

measuring cardiac pulse uses the electrocardiogram (ECG), which requires patients to wear adhesive patches on their chest. However, physical contact can be undesirable because it may cause discomfort and inconvenience for the patient. For some patients, such as burn patients and premature babies, contact sensing is infeasible. For these patients, a contact-free vital signs extraction system with comparable accuracy is not only preferred but also necessary.

Contactless detection of the cardiovascular pulse wave has been explored using optical means since the 1930s [1, 6, 14]. This optically obtained cardiovascular measurement is called the photoplethysmogram (PPG). Because blood absorbs light more than surrounding tissues, we can illuminate the thin parts of the body, such as fingers or earlobes, and measure the amount of light transmitted or reflected to obtain the PPG. Since the cardiac pulse wave alters blood vessel volume, the measured light intensities vary accordingly. It has been shown that the PPG can be used to extract pulse rate, respiration rate, blood pressure, and oxygen saturation level [13]. Commercial devices for physiological assessment, such as the pulse oximeter, uses the PPG to accurately measure pulse rate and oxygen saturation.

The PPG signal can be affected by many sources of noise, such as changes in ambient lighting and patient motion, that compromise the accuracy of vital sign measurement. Dedicated lighting is used by most sensors, e.g., the pulse oximeter, that use the PPG signal. Typically, light sources with the wavelengths in the red and infrared bands are used because those wavelengths penetrate human body tissues better and can be measured more accurately upon transmission. Recent studies show that using visible light and measuring reflected light intensity can also produce clear PPGs for heart rate measurements [6, 14, 11]. However, in all cases, either contact probes have to be used or the subjects have to remain still to remove the effect of motion. With the presence of motion, none of the current techniques can produce PPG signals suitable for medical purposes.

Many signal processing techniques have been used in previous works to reduce the noise of the raw observed data. Verkruysse et al.[14] extract the PPG signal from videos of still subjects recorded by off-the-shelf commercial cameras in ambiently lit

environments. The PPG is extracted by spatially averaging the pixel values of the green channel in the face region. The signal-to-noise ratio is high enough such that pulse rate and respiration rate can be detected up to the fourth harmonic. Philips [11] applied the same idea and implemented a real-time heart rate measurement application using a mobile device and embedded camera. Poh et al.[12] used face detection to locate the face region of the subjects and performed independent component analysis (ICA) on the spatially averaged color values to eliminate the effect of motion. Their method is able to compensate for the noise induced by mild translational motions and produce an accurate heart rate measurement.

These previous works show that pulse rate can be extracted accurately from frontal face video in ambiently lit environments. However, they are still vulnerable to noise induced by general motions, such as talking and head shaking. In addition, since Poh et al.[12] uses temporal filtering to stabilize the output, their method may hide fatal symptoms and be unsuitable for clinical applications. A better method to minimize the effect of motion artifact is needed.

6.1.2 Problem Statement

We aim to provide a contactless vital signs monitoring system for patients, such as premature babies, where conventional physiological monitoring devices cannot apply. Our goal is to estimate vital signs with comparable accuracy to those produced by state-of-the art devices in the hospitals. To obtain accurate and robust vital signs measurements with the presence of noise, we need to use more rigorous physiological and mathematical models.

Physiologically, each location on the face can emit different amounts of visual clues because blood flow to the skin is modulated by a variety of physiological phenomena. For example, regions with a lower density of peripheral blood vessels should generate a PPG with less amplitude. In addition, the cardiovascular pulse wave may arrive at each location of the face at different times. In previous works, every point was considered equally informative with the blood arriving at each point with no time differences. As such, simple averaging was used. To address these physiological

phenomena, we consider these differences when combining the time series of color values at each location (pixel). Specifically, we associate with each pixel (a) a weight that indicates the importance of the pixel, and (b) an onset time which indicates the time difference. We time shift each time series of color values by its onset time and multiply by its weight before taking an average. This more flexible framework allows us to statistically optimize how we combine the observed signal from each pixel.

In determining the weights, we want to find a set of weights that (a) gives the minimum estimation error variance compared with ground truth measurements and (b) can be physiologically explained. Our system computes the weights and onset time of each point automatically from data.

6.2 Experiment Data Collection

We collected a set of videos for 11 different newborns. This data was used to help us evaluate our system. The video was recorded using typical lighting conditions as would be found in the Special Care Nursery at Winchester Hospital. While we hoped to record videos of newborns with different skin colors, most of the subjects had a fair complexion.

For each subject, we recorded four to eight videos (hereafter known as the "recording session") using digital video recording equipment. Multiple recordings were necessary to capture the patient during sleeping and waking phases and to vary the lighting conditions in the environment. Each recording lasted approximately five to fifteen minutes. Recordings were captured during periods when the act of recording did not interfere with the care provided by the medical personnel in the nursery or with parental visitation. For each video that we recorded, we varied the brightness of the lights in the immediate area around the newborn. Note that we only enrolled patients that were not affected negatively by changes in the brightness level and who did not require specialized lighting such as phototherapy.

We recorded video of the baby using three medium-sized consumer digital camera and one compact digital camera. The cameras were attached to a portable IV pole

that we positioned next to the baby using the camera's tripod mount. The cameras were not mounted directly above the newborn and we ensured beforehand that the cameras did not occlude or hide medical equipment. We started and stopped the recording manually.

The three cameras consisted of one near infrared (NIR) camera (SONY NEX-3) and two standard visible light spectrum cameras (SONY NEX-5K). We used the NIR spectrum camera because that oxygenated and deoxygenated hemoglobin has very different absorption characteristics in the NIR spectrum [16]. Two standard cameras with different recording angles allowed us to do 3D reconstruction of the subjects. When we were actively capturing video, we placed a small gray card and color reference paper card—approximately 3 inch by 4 inch in size — in the field of view. These cards allowed us to compensate for any changes in the ambient light and to estimate the noise level when we analyzed the video's contents.

In addition to recording video, we also recorded the gestational-age and corrected gestational age of the patient. This information was written on the study number card that was placed in front of the video camera at the start of the recording session. We simultaneously acquired real-time vital signs for each patient from the vital signs monitor that is routinely used on all infants in the Special Care Nursery by recording the display of the monitor using a compact digital camera (SONY DSC-T110). Note that we properly de-identified the data collected by the monitors to protect the privacy of the patient. However, the face of the newborn, if captured, is not de-identified in the video, since facial features are an essential input for our algorithms.

The detailed protocol description is included in Appendix B, which we submitted to Institutional Review Board of Winchester Hospital for approval before our data collection process.

6.3 Heart rate extraction system

6.3.1 Overview

Our system decomposes input video into localized time series of color values (hereafter known as the "color series"), computes the local importance and onset time for each pixel, and combines the color series for heart rate estimation in a way that minimizes the error variance. The process is illustrated in Figure 6-1. We first spatially low-pass filter (spatial pooling) every frame of the video and downsample them to form a bank of localized color series. The localized color series is then temporally band-pass filtered for denoising purposes. These two steps are the same as the first two steps of the Eulerian video magnification shown in Figure 2-1, and hereafter we call these two steps "Eulerian preprocessing. Since we know our target signal is a pulse signal, we spatially filter the frames using Gaussian pyramid downsampling and temporally filter the color series using a bandpass filter that selects frequencies within 0.4-4Hz.

We then take the denoised color series generated by Eulerian preprocessing and estimate a weight and a onset time for each pixel. The set of weights/onsets is referred to as the weight/onset map. Next, we shift each color series by its onset time in time, multiply it by the weight, and average each color series. The weights are computed in a way such that the linear combination of temporally aligned color series gives rise to the smallest error variance when compared with the ground truth heart rate measurements. The weight/onset map is also visualized to check if they coincide with physiological explanations. An example of a visualized weight/onset map is shown in Figure 2-1. Details on how we estimate the weight/onset map are described in Sec 6.3.3.

Lastly, we convert the combined color series into a heart rate measurement. The temporal characteristics of the combined color series is analysed to produce the final heart rate estimation. We have investigated two methods to estimate heart rate from color series, one using peak position in frequency band of interest, and one looking at peak-to-peak distance in the primal domain.

The weight/onset map can be estimated iteratively if we estimate the heart rate

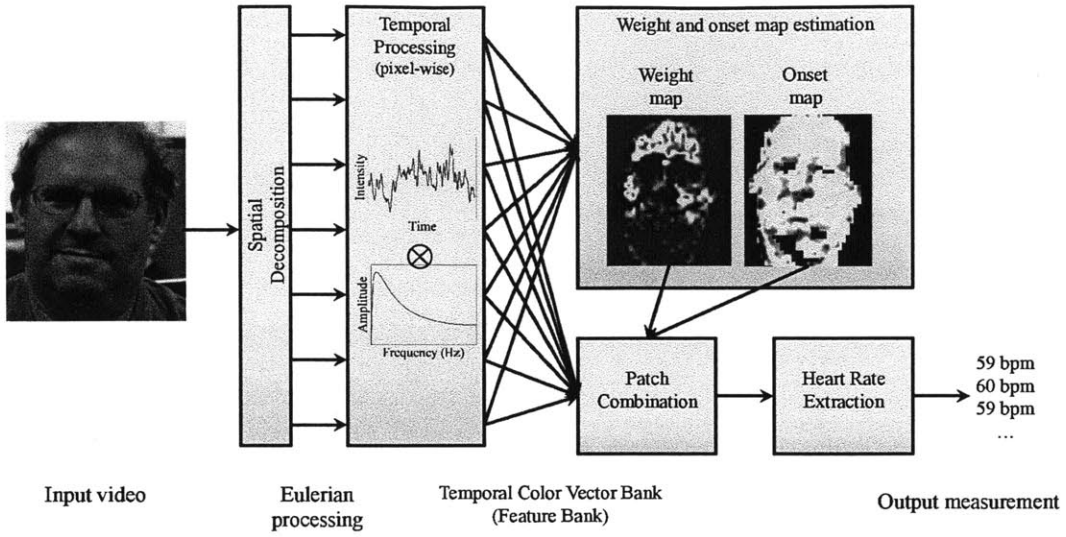


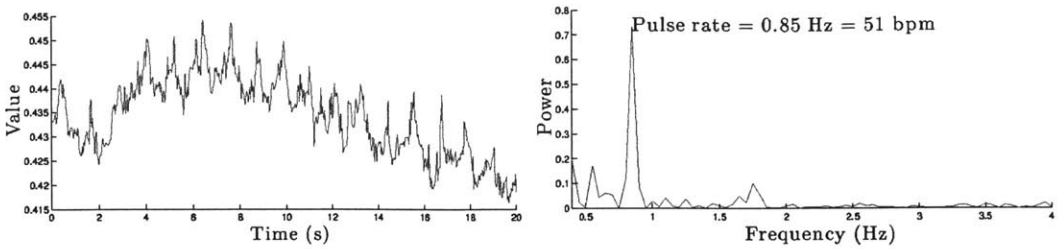
Figure 6-1: Overview of the vital sign extraction framework. The system first uses Eulerian preprocessing to generate a bank of denoised temporal color series from input video. Weight/onset map is computed from these denoised color series. These color series then is combined according to the weight/onset map and the combined series will be used for generating better heart rate estimation. A partial history of all color series are kept for updating the weight/onset map.

offline. An initial weight map is selected by the user and an initial offset map is set to zero for all pixels. The system can combine the color series according to the weight/onset map from a previous iteration and uses the resulting color series as a reference to estimate the new weight/onset map. The system continues to iterate until a user designated number of iterations.

The weight/onset map can also be updated in real-time. We keep a partial history of all the color series as training data to compute the new weight/onset map. The previous weight/onset map is used to obtain the reference color series and this reference series is treated as the ground truth for computing the new weight/onset map. The weight/onset map at time index i is updated according to

$$Map[i] = (1-\gamma)Map[i-1] + \gamma Map_{training}[i] \tag{6.1}$$

where the choice of γ value depends on how much we want our map to be adaptive to scene changes. Large γ gives us more adaptivity but the map will be less stable



(a) Temporal color series (b) Power spectrum

Figure 6-2: Power spectrum shows peak at pulse rate frequency. (a) A color series generated from the *face* video using Eulerian preprocessing. (b) shows the power spectrum in the frequency band of interest (0.4-4Hz) and the estimated pulse rate.

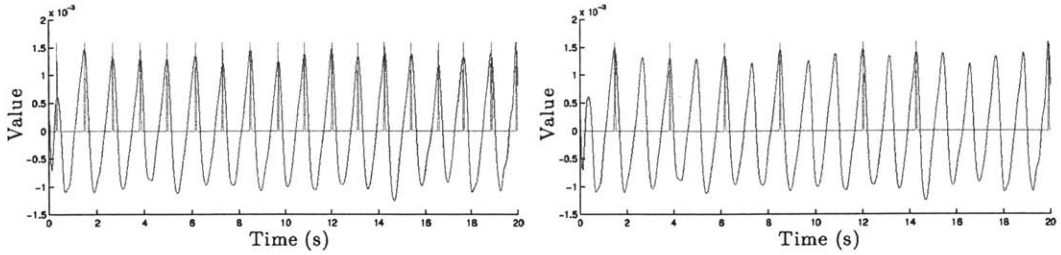
because of the noise.

6.3.2 Heart Rate Extraction

To extract the frequency of a periodic pulse signal from a noisy signal, we investigated two different approaches: Fourier frequency analysis and peak detection. In this section, we will describe specifically how we implement heart rate extraction and the pros and cons of each method. We choose peak detection as our final implementation because it provides the ability to extract the individual heart beat, which is potentially useful for clinical applications.

Frequency Analysis. Fourier frequency analysis is a standard approach to examine signal strength at each frequency. The Fast Fourier Transform(FFT) can be performed on temporal signals to produce a power spectrum. If the pulse signal strength rises above noise strength level, we can simply detect the peak in our frequency band of interest and read off the frequency position of the peak as our measurement of pulse rate, as shown in Figure 6-2. Some previous works used similar methods but with different noise reduction techniques. [12][14].

In order to capture temporal variations of the pulse, we need to choose a size for the local temporal window to perform the FFT. This choice is a trade off between frequency resolution and time resolution. A larger time window will give us more frequency resolution but less time resolution. Specifically, our video sequences are



(a) Good temporal window (b) Temporal window with size too large

Figure 6-3: Peak detection results of different temporal windows. Blue traces show the band passed signal and red traces show the detected peak positions.

recorded at frame rate of 30 fps (e.q. sampling rate $F_s = 30Hz$). If we want our frequency resolution Δf to be 5 beats per minute, we should choose the length of the time window to be

$$L = \frac{F_s}{\Delta f} = \frac{30}{\frac{5}{60}} = 360 frames \tag{6.2}$$

Beat-to-beat interval of human heart fluctuates constantly from beat to beat. This phenomenon causes the power of pulse signal to spread out in frequency domain, so the FFT-based method may has less accurate measurement. The beat-to-beat variation is measured as heart rate variability (HRV). FFT-based method is unable to extract beat by beat segments to measure HRV, which is an important diagnostic measurement for cardiac diseases such as myocardial infarction [7].

Peak Detection. Peak detection is commonly used for heart rate extraction using the electrocardiogram (ECG). In a typical ECG signal, there is a high-amplitude R wave in every heart beat. The time interval between successive R waves (RR interval) is used to estimate instantaneous heart rate. We can also average over several RR intervals to estimate average heart rate. The same methodology can be applied to the temporal color series which we obtained from Eulerian preprocessing.

To get the peak positions, we select the maximum point within a local temporal window. The size of the temporal window should be chosen carefully because if the window size is larger than the true peak interval, then a peak with smaller amplitude will not be detected. Results of peak detection using different sizes of temporal windows are shown in Figure 6-3.

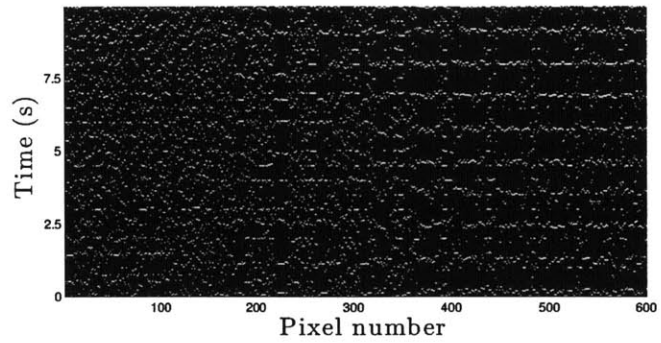


Figure 6-4: Peak positions detected of all pixels are shown as white dot in the figure. The right part(pixel number 400 - 600) of the figure contains clear periodic stripe pattern. The pixels with this pattern are considered informative and should be assigned more weight. The left part(pixel number 1 - 150) of the figure has peaks detected randomly, and these pixels are considered as noise.

Peak detection gives us the segments of the heart beat waveform. We can use them to estimate the instantaneous heart rate, average heart rate, and HRV. We can also normalize these segments to the same length and average them to get an average PPG profile waveform. Peak detection allows us to extract heart rate with beat-by-beat accuracy, and we have more control on how to temporally smooth the measurements without missing potentially abnormal heart beats. Given these advantages, we prefer to use the peak detection method to extract heart rate.

6.3.3 Weight/onset map estimation

Not every pixel contains the pulse signal, and we would like to use a data-oriented way to determine which pixel is informative and how much weight we should assign to each pixel. The peak detection results of all pixels are shown in Figure 6-4. From the figure, we observe periodic peak patterns in some pixels with different onset times. In this section, we will explain how to estimate the weight and onset map from peak detection results given a temporal color series.

Linear combination of estimators. Our goal is to find the best set of weights to linearly combine the color series which are temporally aligned according to their onset times. We want the combination to give us the most accurate estimation of peak

positions, which can be formulated as a linear programming optimization problem.

First of all, we consider temporal color series which contains the true pulse signal. The peak detection algorithm is a peak position estimator, which takes a color series v_i at position i ,and produces a peak position estimation p_i. We denote the onset time difference at position i as o_i, the expected value and variance of the estimator p_i are

$$E[p_i] = p_{true} + o_i \tag{6.3}$$

$$Var[p_i] = \sigma_i^2 \tag{6.4}$$

where p_{true} is the ground truth peak position, and σ_i^2 is the estimation error variance of p_i.

Assuming p_i's are uncorrelated, our goal is to find the set of weight w_i's

$$\begin{aligned} \min_{w_i} \quad & Var[\sum_i w_i(p_i - o_i)] \\ \text{subject to} \quad & E[\sum_i w_i(p_i - o_i)] = p_{true} \\ & \sum_i w_i = 1 \end{aligned}$$

Solving this optimization problem, we get the weight and onset time for each temporal color series v_i

$$o_i = E[p_i] - p_{true} \tag{6.5}$$

$$w_i = \frac{\sigma_i^{-2}}{\sum_i \sigma_i^{-2}} \tag{6.6}$$

We can get estimates of o_i and w_i by estimating $E[p_i]$ and $Var[p_i]$ from data. The following paragraph explains explicitly how we estimate $E[p_i]$ and $Var[p_i]$ and use

them to compute o_i and w_i according to eq.(6.5) and eq.(6.6).

Estimating weight and onset time. Given the ground truth peak impulse train (one at peak positions and zero otherwise), we can estimate the weight and onset for each pixel as follows:

(1) For each color series v_i, detect its peaks and denote their position as $p_i[k]$ for $k = 1 \ldots K$, where K is the number of peaks detected in the color series

(2) For each $p_i[k]$, find the nearest ground truth peak $p_{true}[j]$ using L1 norm.

(3) Compute $d_i[k] = p_i[k] - p_{true}[j]$.

(4) Estimate onset time $\hat{o}_i$ using the empirical mean of $d_i[k]$'s

$$\hat{o}_i = \frac{\sum_k d_i[k]}{K} \tag{6.7}$$

(5) Estimate weight $\hat{w}_i'$ using is empirical variance of $d_i[k]$'s, which is inversely proportional to the weight

$$\hat{w}_i' = \left(\frac{\sum_k (d_i[k] - \hat{o}_i)^2}{K} \right)^{-1} \tag{6.8}$$

(6) Estimate weight $\hat{w}_i = \hat{w}_i' / \left(\sum_i \hat{w}_i' \right)$

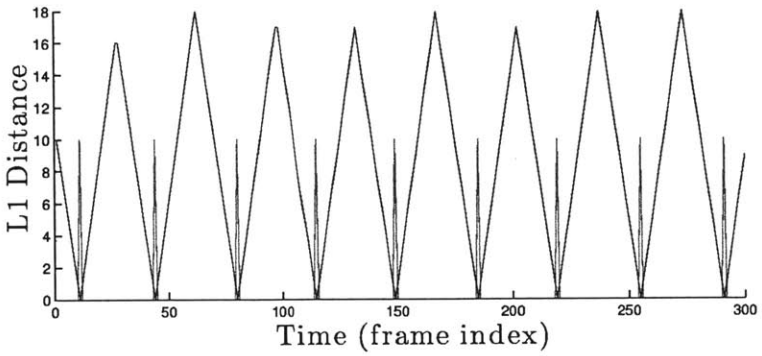
Steps (1)-(3) are illustrated in Figure 6-5.

For those color series containing only noise, we assume that a peak will be detected randomly in the interval Δ of two ground true peaks. The empirical variance of $d_i[k]$ can be approximated as

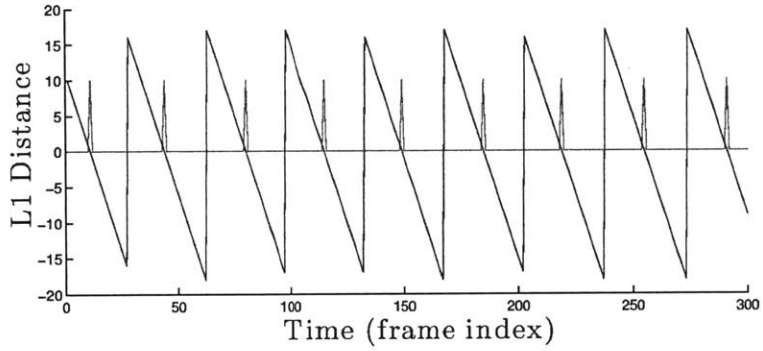
$$Var[d_i] \approx \frac{1}{\Delta} \int_{\frac{-\Delta}{2}}^{\frac{\Delta}{2}} x^2 \, \mathrm{d}x = \frac{\Delta^2}{24}. \tag{6.9}$$

We assign a variance threshold th_n of noise such that the color series with empirical variance larger than the noise threshold will be considered as noise and its weight is assigned zero. Currently, we empirically set th_n to $\frac{\Delta^2}{24}$, where Δ is the period corresponding to the average heart rate.

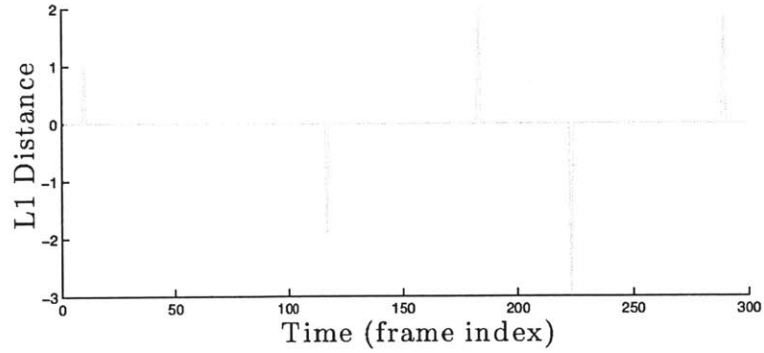
An example estimated weight and onset map for the sequence *face* is shown in Figure 6-6. The result shows that the smooth skin region has the highest weight.



(a) Compute L1 distance transform of reference peak impulse train

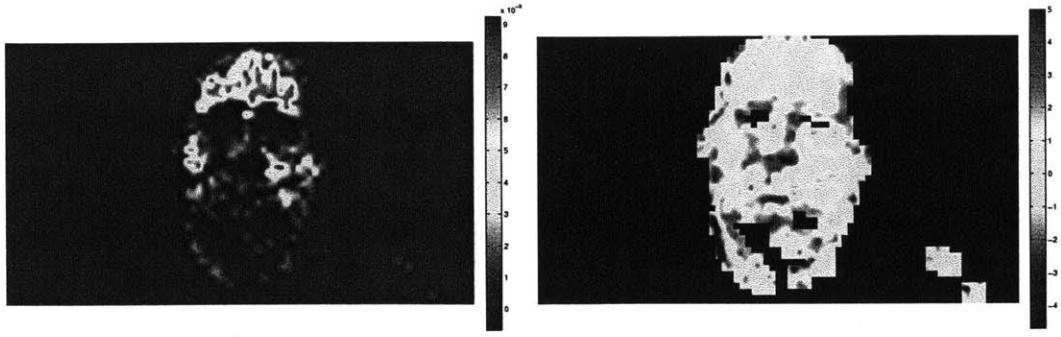


(b) Add sign information



(c) Distance to the nearest reference peak at each detected peak position of v_i

Figure 6-5: The process to get $d_i[k]$ for every peak $p_i[k]$ detected in color series v_i



(a) Weight map (b) Onset map

Figure 6-6: The estimated weight and onset map. The onset map is shown only for those regions with weights > 0.

The smooth region has a smaller image gradient, and therefore, is more resistant to motion-induced noise. The fact that the weight map coincides with the smooth skin region is consistent with our physiological understanding that blood circulation causing skin color changes.

6.4 Results and discussion

We evaluated our system by comparing against ground truth using the dataset we obtained from Winchester hospital. In all the results, we used a 2D separable binomial filter of size 5 to spatially blur and downsample the video sequence six times. We used an ideal band pass with a passband from 1 Hz to 4 Hz to temporally filter the sequence. The initial weight map is set to be a rectangular region containing mostly the face region of the subject in the video. The weight and onset map is refined iteratively five times and the resulting maps are used to combine the temporal color series for the final heart rate estimation. After performing peak detection on the combined temporal color series, we generated the heart rate estimation by averaging the peak-to-peak intervals in five second time windows. In Figure 6-7, 6-8 and 6-9, we show the heart rate estimate for three different subjects with different activity levels (still and active) or different lighting conditions (dim and bright). We will refer to the babies in the videos by the subject ID numbers that were assigned during the

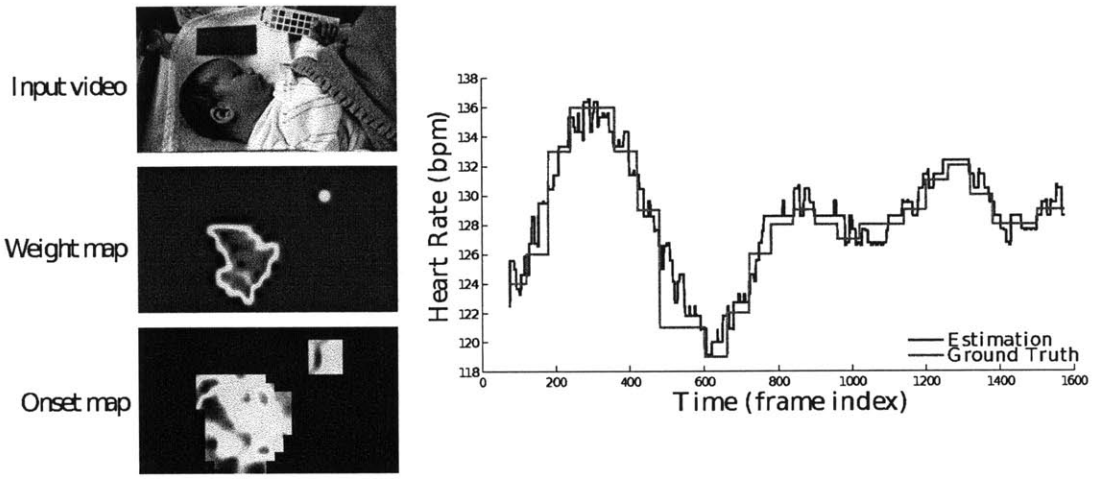
data collection process.

Heart rate estimation is performed on subject 3 and subject 4 with different activity levels under bright lights. When the subjects are still, we can see that the weight map has weights distributed over most skin regions. The accuracy of heart rate measurements in these conditions are comparable to the heart rate measurement obtained from the state-of-the-art monitor. (The root mean square error per unit time is 1.47 beats per minutes for subject 3 and 6.36 for subject 4.) When subjects are active, the weight map has weights only centralized at the smoothest skin regions because these regions are least affected by motions. The results show that our system can still generate reasonable heart rate measurements when motion is fairly small. (The root mean square error per unit time is 14.9 beats per minutes for subject 3 and 12.3 for subject 4.) We also estimated that the heart rate for subject 7. Subject 7 was recorded during periods of low activity, but under varying lighting conditions. Under bright lighting conditions, our method produces a good heart rate estimation except during the interval from frame index 500 to frame index 1000, when the subject has a periodic chewing motion that our system treats it as pulse signal. On the other hand, under dim lighting conditions, although our method seems to find reasonable weight map, the noise level is too high for our heart rate estimation to generate accurate measurements. (Excluding frames 500 to 1000, the root mean square error per unit time is 8.28 beats per minutes in bright lighting and 20.7 in dim lighting.)

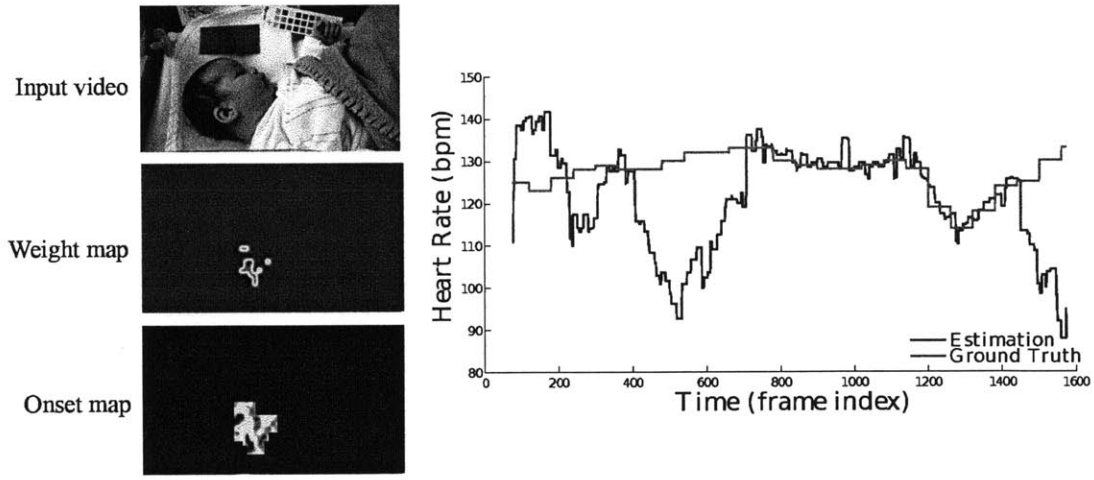
In Figure 6-8(a), the heart rate estimate between between frame index 1200 and 1500 varies considerably from the ground truth measurement because there is a false-detected peak. That peak negatively affects the heart rate measurements which use a temporal averaging window that contains it. Assuming there are K peaks in the temporal averaging window, a false-detected peak will make the heart rate $(K+1)/K$ times larger than it should be ,and an undetected peak will make the heart rate $(K-1)/K$ times smaller than it should be. Our system needs to better handle false-detected/undetected peaks to get more accurate measurements.

In bright lighting conditions, if subjects only have mild motions, our system can produce heart rate measurements with clinical accuracy. We have also shown that if

our system can handle false-detected/undetected peaks properly, we can achieve even higher accuracy. However, our system still has limitations in dim lighting conditions and when subjects have big motions.

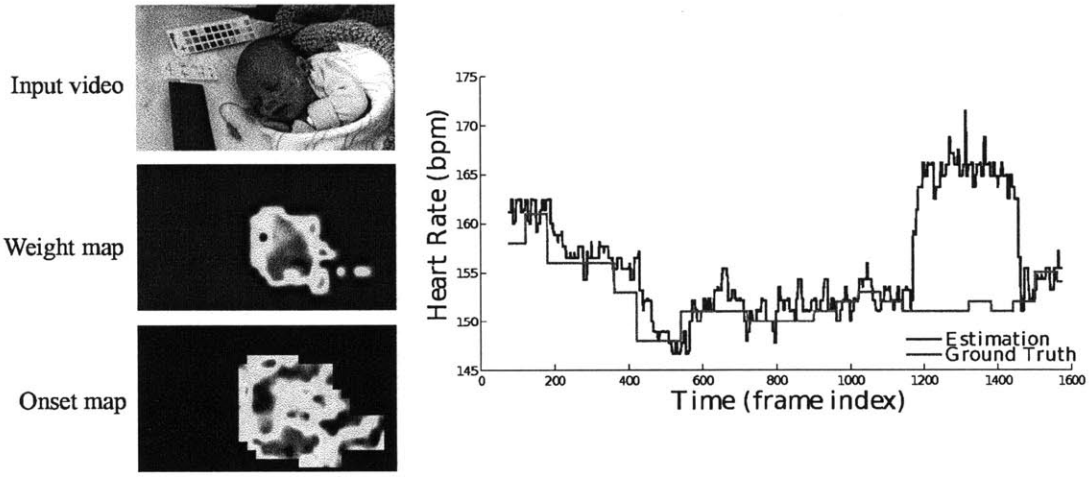


(a) Heart rate estimation of subject 3 in still phase

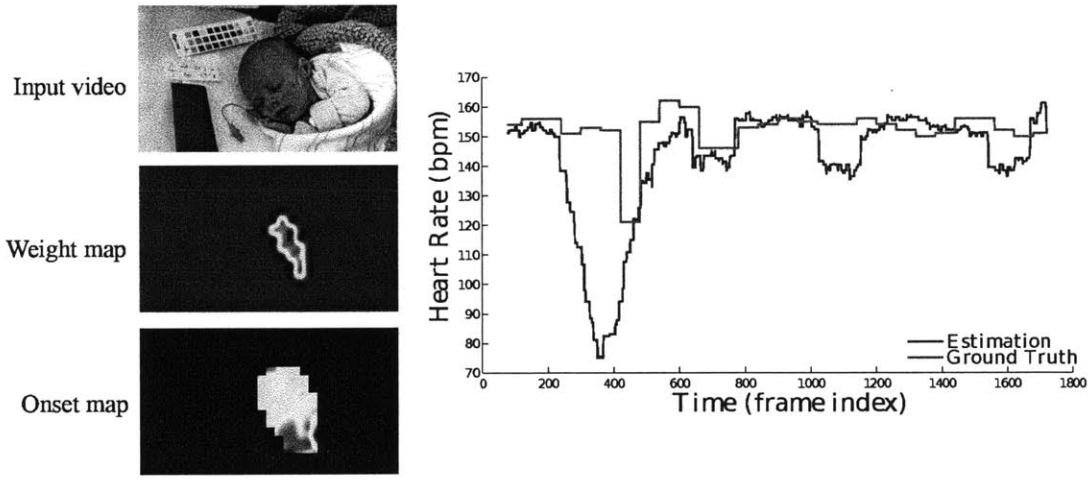


(b) Heart rate estimation of subject 3 in active phase

Figure 6-7: Heart rate estimation results, weight and onset map of subject 3 with different activity levels.

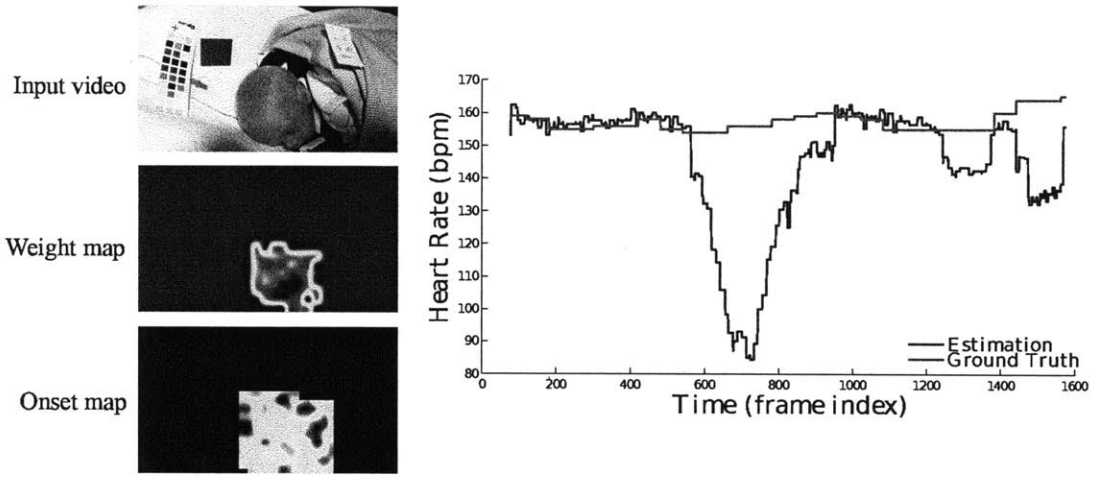


(a) Heart rate estimation of subject 4 in still phase

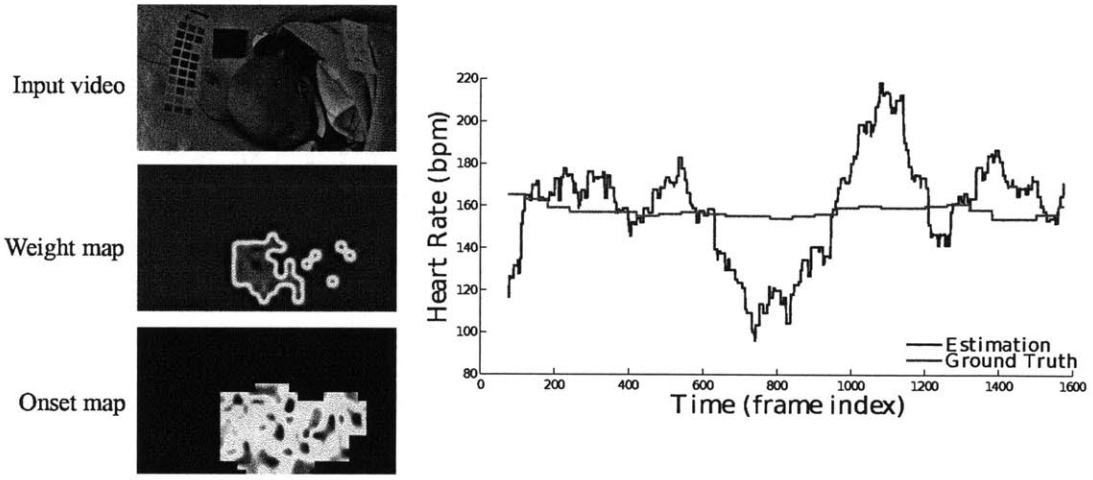


(b) Heart rate estimation of subject 4 in active phase

Figure 6-8: Heart rate estimation results, weight and onset map of subject 4 with different activity levels.



(a) Heart rate estimation of subject 7 in bright lighting condition



(b) Heart rate estimation of subject 7 in dim lighting condition

Figure 6-9: Heart rate estimation results, weight and onset map of subject 7 with different lighting conditions.

Chapter 7

Conclusion

We presented a simple method that takes a video as input and reveals the information of subtle signals that are invisible in the video. The *Eulerian*-based spatio-temporal processing, which temporally processes pixels in a fixed spatial region, successfully extracts informative signals. In this project, these signals are revealed in two different ways. One is visual magnification in the videos for general signals and the other is numerical extraction for vital sign signals.

We introduced a framework called Eulerian video magnification that takes a video as input and exaggerates subtle color changes and imperceptible motions to make them visible. To amplify motion, our method does not perform feature tracking or optical flow computation, but merely magnifies temporal color changes using spatio-temporal processing. This *Eulerian*-based method successfully reveals subtle color variations and amplifies small motions in real-world videos. The Eulerian video magnification is also computationally efficient enough to process the video in real-time.

We built a heart rate extraction system that takes localized time series of color values that generated by the *Eulerian*-based spatio-temporal processing, and combined them based on their local SNR and onset time differences to estimate heart rate. The estimations were evaluated on the dataset with ground truth that we acquired from Winchester Hospital. We showed that our system can produce accurate heart rate measurement from real-world videos when subjects have mild motions in bright lighting condition.

Appendix A

Eulerian vs. Lagrangian Sensitivity to Noise Detailed Derivation

Here we give the derivation in Appendix A in the paper in more detail.

In this section we derive the error in the Eulerian and Lagrangian motion magnification results with respect to spatial and temporal noise. The derivation is done again for the 1D case for simplicity, and can be generalized to 2D. We use the same setup as in Sect. 3.1 in the paper, where the true motion-magnified sequence is

$$\begin{aligned} \hat{I}(x,t) &= f(x + (1+\alpha)\delta(t)) \\ &= I(x + (1+\alpha)\delta(t), 0) \end{aligned} \tag{A.1}$$

A.0.1 Inherent Errors

Regardless of noise, the two approaches only approximate the true motion magnified sequence (Eq. A.1). Let us first examine the errors in the approximation on the clean signal.

Lagrangian. In the Lagrangian approach, the motion-amplified sequence, $\tilde{I}_L(x,t)$, is achieved by directly amplifying the estimated motion, $\tilde{\delta}(t)$, with respect to the

reference frame $I(x,0)$

$$\tilde{I}_L(x,t) = I(x + (1+\alpha)\tilde{\delta}(t), 0) \tag{A.2}$$

In its simplest form, we can estimate $\delta(t)$ using point-wise brightness constancy (See the paper for discussion on spatial regularization)

$$\tilde{\delta}(t) = \frac{I_t(x,t)}{I_x(x,t)} \tag{A.3}$$

where $I_x(x,t) = \partial I(x,t)/\partial x$ and $I_t(x,t) = I(x,t) - I(x,0)$. From now on, we will omit the space (x) and time (t) indices when possible for brevity.

The error in in the Lagrangian solution is directly determined by the error in the estimated motion, which we take to be second-order term in the brightness constancy equation

$$\begin{aligned} I(x,t) &= I(x + \delta(t), 0) \\ &\approx I(x,0) + \delta(t) I_x + \frac{1}{2}\delta^2(t) I_{xx} \\ \frac{I_t}{I_x} &\approx \delta(t) + \frac{1}{2}\delta^2(t) I_{xx} \end{aligned} \tag{A.4}$$

So that the estimated motion $\tilde{\delta}(t)$ is related to the true motion, $\delta(t)$, as

$$\tilde{\delta}(t) \approx \delta(t) + \frac{1}{2}\delta^2(t) I_{xx} \tag{A.5}$$

Plugging (A.5) in (A.2),

$$\begin{aligned} \tilde{I}_L(x,t) &\approx I\left(x + (1+\alpha)\left(\delta(t) + \frac{1}{2}\delta^2(t) I_{xx}\right), 0\right) \\ &\approx I\left(x + (1+\alpha)\delta(t) + \frac{1}{2}(1+\alpha)\delta^2(t) I_{xx}, 0\right) \end{aligned} \tag{A.6}$$

Using first-order Taylor expansion of I about $x + (1+\alpha)\delta(t)$,

$$\tilde{I}_L(x,t) \approx I(x + (1+\alpha)\delta(t), 0) + \frac{1}{2}(1+\alpha)\delta^2(t) I_{xx} I_x \tag{A.7}$$

Subtracting (A.1) from (A.7), the error in the Lagrangian motion-magnified sequence, ε_L, is

$$\varepsilon_L \approx \left| \frac{1}{2}(1+\alpha)\delta^2(t) I_{xx} I_x \right| \tag{A.8}$$

Eulerian. In our Eulerian approach, the magnified sequence, $\hat{I}_E(x,t)$, is computed as

$$\begin{aligned} \tilde{I}_E(x,t) &= I(x,t) + \alpha I_t(x,t) \\ &= I(x,0) + (1+\alpha) I_t(x,t) \end{aligned} \tag{A.9}$$

similar to Eq. 4 in the paper, using a two-tap temporal filter to compute I_t.

Using Taylor expansion of the true motion-magnified sequence, $\hat{I}$ (Eq. A.1), about x, we have

$$\hat{I}(x,t) \approx I(x,0) + (1+\alpha)\delta(t) I_x + \frac{1}{2}(1+\alpha)^2 \delta^2(t) I_{xx} \tag{A.10}$$

Plugging (A.4) into (A.10)

$$\begin{aligned} \hat{I}(x,t) &\approx I(x,0) + (1+\alpha)(I_t - \frac{1}{2}\delta^2(t) I_{xx} I_x) + \frac{1}{2}(1+\alpha)^2 \delta^2(t) I_{xx} \\ &\approx I(x,0) + (1+\alpha) I_t - \frac{1}{2}(1+\alpha)\delta^2(t) I_{xx} I_x + \frac{1}{2}(1+\alpha)^2 \delta^2(t) I_{xx} \end{aligned} \tag{A.11}$$

Subtracting (A.9) from (A.11) gives the error in the the Eulerian solution

$$\varepsilon_E \left| \approx \frac{1}{2}(1+\alpha)^2 \delta^2(t) I_{xx} - \frac{1}{2}(1+\alpha)\delta^2(t) I_{xx} I_x \right| \tag{A.12}$$

A.0.2 Errors as Function of Noise

The errors derived above are inherent to the two methods based on their first-order Taylor approximations. Let us now examine these approximations on the noisy signal

$I'(x,t)$ defined as

$$I'(x,t) = I(x,t) + n(x,t) \tag{A.13}$$

for additive noise $n(x,t)$.

Lagrangian. The estimated motion becomes

$$\tilde{\delta}(t) = \frac{I'_t}{I'_x} = \frac{I_t + n_t}{I_x + n_x} \tag{A.14}$$

where $n_x = \partial n/\partial x$ and $n_t = n(x,t) - n(x,0)$.

Using Taylor Expansion on (n_t, n_x) about $(0,0)$ (zero noise), and using (A.4), we have

$$\begin{aligned}\tilde{\delta}(t) &\approx \frac{I_t}{I_x} + n_t\frac{1}{I_x + n_x} + n_x\frac{I_t + n_t}{(I_x + n_x)^2} \\ &\approx \delta(t) + \frac{n_t}{I_x} - n_x\frac{I_t}{I_x^2} + \frac{1}{2}\delta^2(t)I_{xx}\end{aligned} \tag{A.15}$$

where we ignored the terms involving products of the noise components.

Plugging into Eq. (A.2), and using Taylor expansion of I about $x + (1+\alpha)\delta(t)$, we get

$$\tilde{I}'_L(x,t) \approx I(x + (1+\alpha)\delta(t), 0) + (1+\alpha)I_x(\frac{n_t}{I_x} - n_x\frac{I_t}{I_x^2} + \frac{1}{2}I_{xx}\delta^2(t))) + n \tag{A.16}$$

Arranging terms, and Substituting (A.4) in (A.16),

$$\begin{aligned}\tilde{I}'_L(x,t) &\approx I(x + (1+\alpha)\delta(t), 0) + (1+\alpha)\left(n_t - n_x\left(\delta(t) + \frac{1}{2}\delta^2(t)I_{xx}\right) + \frac{1}{2}\delta^2(t)I_{xx}I_x\right) + n \\ &= I(x + (1+\alpha)\delta(t), 0) + (1+\alpha)n_t - (1+\alpha)n_x\delta(t) - \frac{1}{2}(1+\alpha)n_x\delta^2(t)I_{xx} + \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x\end{aligned} \tag{A.17}$$

Subtracting (A.1) from (A.17), we get the Lagrangian error as function of noise

$$\varepsilon_L(n) \approx (1+\alpha)n_t - (1+\alpha)n_x\delta(t) - \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}n_x + \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x + n \tag{A.18}$$

Eulerian. The noisy motion-magnified sequence becomes

$$\begin{aligned}\tilde{I}'_E(x,t) &= I'(x,0) + (1+\alpha)I'_t \\ &= I(x,0) + (1+\alpha)(I_t + n_t) + n \\ &= I_E(x,t) + (1+\alpha)n_t + n \end{aligned} \tag{A.19}$$

Using (A.12), the Eulerian error as function of noise is

$$\varepsilon_E(n) \approx (1+\alpha)n_t + \frac{1}{2}(1+\alpha)^2\delta^2(t)I_{xx} - \frac{1}{2}(1+\alpha)\delta^2(t)I_{xx}I_x + n \tag{A.20}$$

Notice that setting the noise to be zero in (A.18) and (A.20), we get the corresponding errors derived for the non-noisy signal in (A.8) and (A.12).

Appendix B

Data Collection Detailed Protocol

Our goal is to capture a set of videos for 10-15 different newborns. We plan to record videos of newborns with different skin colors and under different lighting conditions as would be found in the Special Care Nursery. We estimate that the entire study will take approximately eighteen months to complete. To gather this video, a member of our research team will carry out the protocol outlined below on subjects who are enrolled in the study.

For each subject, we will record approximately eight separate short videos (hereafter known as the recording session) using digital video recording equipment. Multiple recordings are necessary to capture the patient during sleeping and waking phases and to vary the lighting conditions in the environment. Each recording will last approximately five to fifteen minutes. Recordings will be captured during periods when the act of recording will not interfere with the care provided by the medical personnel in the nursery or with parental visitation. For each video that we record, we will vary the existing nursery lighting within the normal spectrum of use in immediate area around the study subject to simulate the changes in lighting that normally occur in the nursery. Non-participating infants would not be exposed to any lighting variations other than those that occur routinely. Note that we will only enroll patients that will not be affected negatively by changes in the brightness level and who do not require specialized lighting such as phototherapy.

Each baby will be assigned a study number. A master list will be kept with the

babys name and study number. This list will be secured in a password-protected electronic file. Each time the baby is recorded his or her unique study number will be used as an identifier by displaying a card in front of the video camera with the subjects study number at the onset of the recording session. Once the study is completed, the master list will be destroyed.

Before each recording, we may ask the nurse or physician responsible for the care of the patient to position the patient within the crib or warmer to allow us to record the infants face and, if already exposed, chest, abdomen and extremities. The patient will be subject only to movements that are used by the nurses or physicians during typical care for the patient (e.g., diaper changes). However, repositioning will be performed only if the physician believes it will not be harmful to the baby.

We will record video of the baby using two medium-sized consumer digital cameras (each weighing less than one kilogram). We will position the cameras to the right or left side of the crib from above. Both cameras will be securely attached to either a portable IV pole that we can position next to the baby or to nearby fixed mounting rails already present at the bed space. Hospital grade equipment similar to what is used for attaching monitors or IV equipment will be used for attaching the cameras. The position of the camera will not occlude or hide medical equipment. Recording will be started and stopped manually by the research scientist. The research scientist will not reposition or contact the subject. Sound will not be recorded.

When we are actively capturing video, we will place a smallapproximately 3 inch by 4 inch in sizegray color reference paper card in the field of view. This card will allow us to compensate for any changes in the ambient light when we analyze the videos contents.

In addition to recording video, we will also record the gestational-age and corrected gestational age of the patient. This information will also be written on the study number card that is placed in front of the video camera at the start of the recording session. We will simultaneously acquire real-time vital signs for each patient from the vital signs monitor that is routinely used on all infants in the Special Care Nursery by recording the display of the monitor or by obtaining the information electronically

from the monitor. The research scientist conducting the study will set up a small laptop computer near the monitor for collecting the vitals signs data. Data collected by the monitors will not contain the study subject name to protect the privacy of the subject. However, the face of the newborn, if captured, will not be de-identified in the video, since facial features are an essential input for our algorithms.

Bibliography

[1] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1, 2007.

[2] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *Transactions on Communications*, 31(4):532–540, 1983.

[3] William T. Freeman, Edward H. Adelson, and David J. Heeger. Motion without movement. *SIGGRAPH Comput. Graph.*, 25:27–30, Jul 1991.

[4] Martin Fuchs, Tongbo Chen, Oliver Wang, Ramesh Raskar, Hans-Peter Seidel, and Hendrik P.A. Lensch. Real-time temporal shaping of high-speed video streams. *Computers & Graphics*, 34(5):575–584, June 2010.

[5] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[6] Sijung Hu, Jia Zheng, V. Chouliaras, and R. Summers. Feasibility of imaging photoplethysmography. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, volume 2, pages 72 –75, may 2008.

[7] Robert E. Kleiger, J.Philip Miller, J.Thomas Bigger Jr., and Arthur J. Moss. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *The American Journal of Cardiology*, 59(4):256 – 262, 1987.

[8] Ce Liu, W.T. Freeman, R. Szeliski, and Sing Bing Kang. Noise estimation from a single image. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 901 – 908, june 2006.

[9] Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand, and Edward H. Adelson. Motion magnification. *ACM Trans. Graph.*, 24:519–526, Jul 2005.

[10] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 674–679, Apr 1981.

[11] Philips. Philips Vitals Signs Camera. `http://www.vitalsignscamera.com`, 2011.

[12] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, May 2010.

[13] Andrew Reisner, Phillip A. Shaltis, Devin McCombie, and Harry H. Asada. Utility of the photoplethysmogram in circulatory monitoring. *The Journal of the American Society of Anesthesiologists*, 108(5):950 – 958, 2008.

[14] Wim Verkruysse, Lars O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, Dec 2008.

[15] Jue Wang, Steven M. Drucker, Maneesh Agrawala, and Michael F. Cohen. The cartoon animation filter. *ACM Trans. Graph.*, 25:1169–1173, Jul 2006.

[16] Susan Wray, Mark Cope, David T. Delpy, John S. Wyatt, and E.Osmund R. Reynolds. Characterization of the near infrared absorption spectra of cytochrome aa3 and haemoglobin for the non-invasive monitoring of cerebral oxygenation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 933(1):184 – 192, 1988.