Context and Configuration Based Scene Classification

by

Pamela R. Lipson

A.B. Computer Science
Harvard University, 1989

M.S. Computer Science
Massachusetts Institute of Technology, 1993

Submitted to the
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER, 1996

Signature of the Author____
Department of Electrical Engineering and Computer Science
September 1, 1996

Certified by:____
Eric Grimson
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by:____
Frederic R. Morgenthaler
Chairman, Committee on Graduate Students

Context and Configuration Based Scene Classification

by

Pamela R. Lipson

Submitted to the Department of Electrical Engineering and Computer Science
on September 1, 1996 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Computer Science

ABSTRACT

The problem of scene classification is one of the significant open challenges in the field of machine vision. During the past few years, there has been a resurgence of interest in this area due to the potential applications in content-based digital image database indexing. Most proposed solutions have either skirted the problem by using textual annotation or have employed image statistics such as color histograms or local textural measures. While adequate for some tasks, these approaches are unable to capture the global configuration of a scene, which seems to be of critical significance in perceptual judgments of scene similarity. The key question this thesis addresses is how to encode a scene so as to incorporate its overall structure in a manner that would allow subsequent generalization to other members of the scene class. We present a novel approach, called "configural recognition", as a partial solution to this problem. The main features of this approach are its use of qualitative spatial and photometric relationships within and across regions in low resolution images.The emphasis on qualitative measures endows the approach with an impressive generalization ability and the use of low-resolution images renders it computationally efficient. We present results of testing this approach on a large database of natural scenes. We also describe how qualitative scene concepts may be automatically learned from exampies. The applicability of the configural recognition approach is not limited to natural scenes; we conclude by describing some other domains for which the approach seems well suited.

Thesis Supervisor: Eric Grimson
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

There are many people whom I would thank for help and encouragement during this Ph.D. process.

Foremost, I would like to thank my advisor Eric Grimson for his continual guidance and support. I have found his advice invaluable. I also am very grateful for the freedom he gave me to explore a variety of interesting topics over the past several years. It has been a pleasure working with him.

I would also like to thank my thesis committee members, Tommy Poggio and Patrick Winston. I have benefited greatly from my interactions with them. They both provided novel insights and much appreciated advice on many topics.

I would also like to thank Shimon Ullman who has continued to provide advice and support throughout the years. It is true that one meeting with him is worth a year's full of help.

My officemates Aparna Lakshmi Ratan, Mike Leventon, and Greg Klanderman deserve a great deal of thanks, the least of which for putting up with wandering Coke cans and growing stacks of papers. Their help on both research, technical issues, and other matters has been invaluable.

Many thanks to Jay Thornton for his help and supervision over the past several years. I learned a great deal from my work at Polaroid and my interactions with him.

Greg Galperin, J.P. Mellor, and Robert Thau, and Bruce Walton have been overwhelmingly helpful with computer and web issues. Their tolerance for questions is remarkable. Without their help, much of this work could not have been done. I would also like to thank Paul Viola for his advice and especially for his enthusiasm for getting projects started.

Thanks also to Sajit Rao, Mona Singh, Bonnie Berger, and Margrit Betke for their support, advice and friendship.

Pawan Sinha has provided unwavering help and support. Much of this work has been in collaboration with him. He is an outstanding colleague and special friend.

Thanks also to those not-so-newcomers L.L., H.L., S.L. (and W.L.). They always had something to say.

Finally, I thank my family for their love and support. This thesis is dedicated to them.

This thesis describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the A.I. Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense, the Office of Naval Research, and the National Science Foundation.

Table of Contents

CHAPTER 1 Introduction

Object recognition is widely considered to be one of the most challenging problems in the domain of machine vision. A robust object recognition system needs to be capable of recognizing or classifying a single object under varying poses, over changes in illumination, in pre-segmented or cluttered environments, under small non-rigid deformations, despite occlusions, in both greyscale and color images, and through differing imaging equipment. Much work in machine vision has been devoted to developing efficient techniques that can tolerate such variations. However, these efforts have met with limited success so far. For instance, even the seemingly simple task of recognizing the American flags in Figure 1.1 will likely pose a challenge for most existing recognition systems. The American flag on the left is oriented horizontally, illuminated from above and to the right, and undergoing small non-rigid deformations. The flag on the right is oriented vertically, illuminated from above and to the right with a yellowish light, and due to non-rigid deformations is self-occluding. These differences in image conditions are probably of significant magnitude to defeat most existing recognition systems.

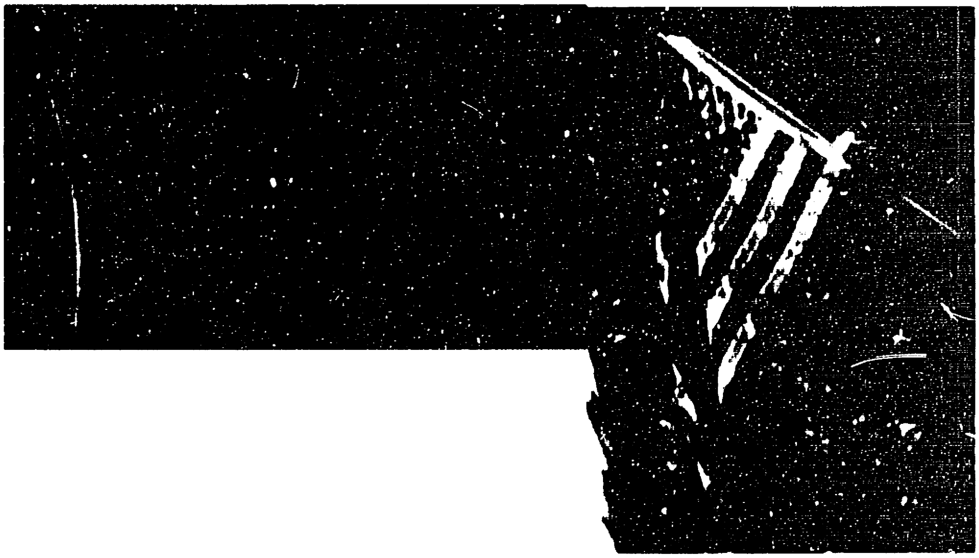


FIGURE 1.1. *Images of one object: an American Flag*

Scene recognition appears to be an even more complex problem than single object recognition. Scenes can encompass many objects which are subject to a variety of changes in imaging conditions and can be arranged in a vast number of compositions. Figure 1.2 shows a bustling city sidewalk at mid day. This image contains many objects, such as pedestrians, trees, street lamps, sign posts, cars, store fronts, and a flag. The problem of recognizing or classifying images such as this 'city sidewalk' scene is rendered difficult because of the high degree of image variability. The interplay between the large number of objects and the resulting object occlusions, shadows, illuminations, and reflections presents serious challenges to even the most sophisticated recognition strategy.

An important question to ask is whether it is necessary to recognize the individual objects in the image as a precursor to recognizing or classifying the scene. If object recognition is always required, then scene recognition inherits all the significant problems associated with object recognition compounded by the difficulties of recognizing multiple objects in complex configurations. Given the number of possible objects, their composition, and the variable viewing conditions, scene classification via individual object recognition becomes an almost intractable problem.

This thesis addresses the question of whether scene recognition or classification can be performed without the need for comprehensive object recognition. If one could eliminate the need for object recognition at the start of the computation queue, scene classification could potentially become a more tractable and computationally simpler problem. In this thesis, we will describe and demonstrate one approach. We will also address whether the technique for scene classification developed in this thesis can be used directly for, or at least to facilitate, object recognition.

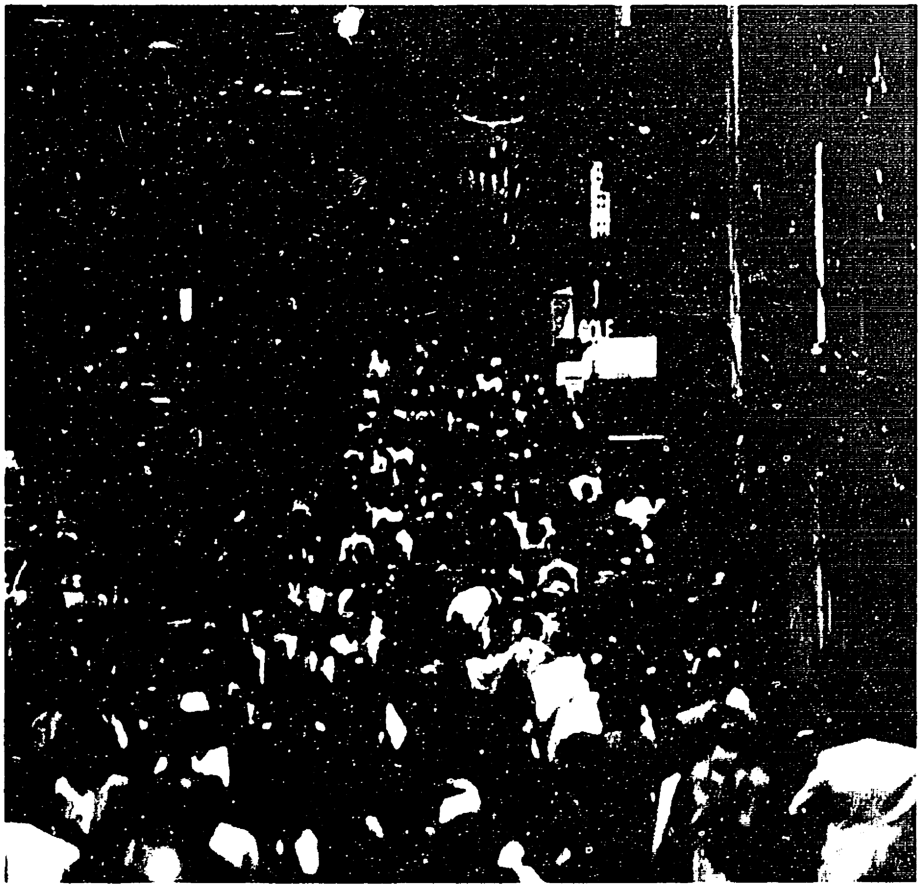


FIGURE 1.2. *A bustling city scene containing many objects.*

1.1 Scene classification

1.1.1 Definition of the problem

The goal of the work described in this thesis in the context of scenes is quite different from that of previous researchers who have addressed the problem of "scene understanding". The focus of much of conventional scene understanding work is to recover the geometry of the scene; the three-dimensional surface coordinates from a set of two-dimensional images. The geometry of the scene may be the desired solution to the problem, i.e. for use in surface relief map construction or a 3-D CAD model, or the information may be used as a precursor for object recognition [15][29][71].

In this thesis, however, scenes are viewed simply as compositions of two-dimensional features such as colored regions. The scene classification problem which is addressed is to build a concept of a class based on these image features from a set of examples and to identify members of the class from novel images. The input to a scene classification system is either a predefined scene model or a set of one or more example images. The desired output is a set of images which perceptually belong to the same class as described by the model or the set of examples. Figure 1.3 shows some input/output examples for the class of snowy mountain images.

Scene classification can thus be subdivided into two problems. The first is the problem of capturing or describing a scene concept in the form of a scene model. The model may be provided from an *a priori* description or from a description extracted automatically from a set of example images. The second problem is to classify or to detect novel images of the class given the scene model. The overall goal of an automated scene classification technique is to robustly and efficiently determine the similarities of images within a class and to determine the features that distinguish between different classes for both class model generation and detection.

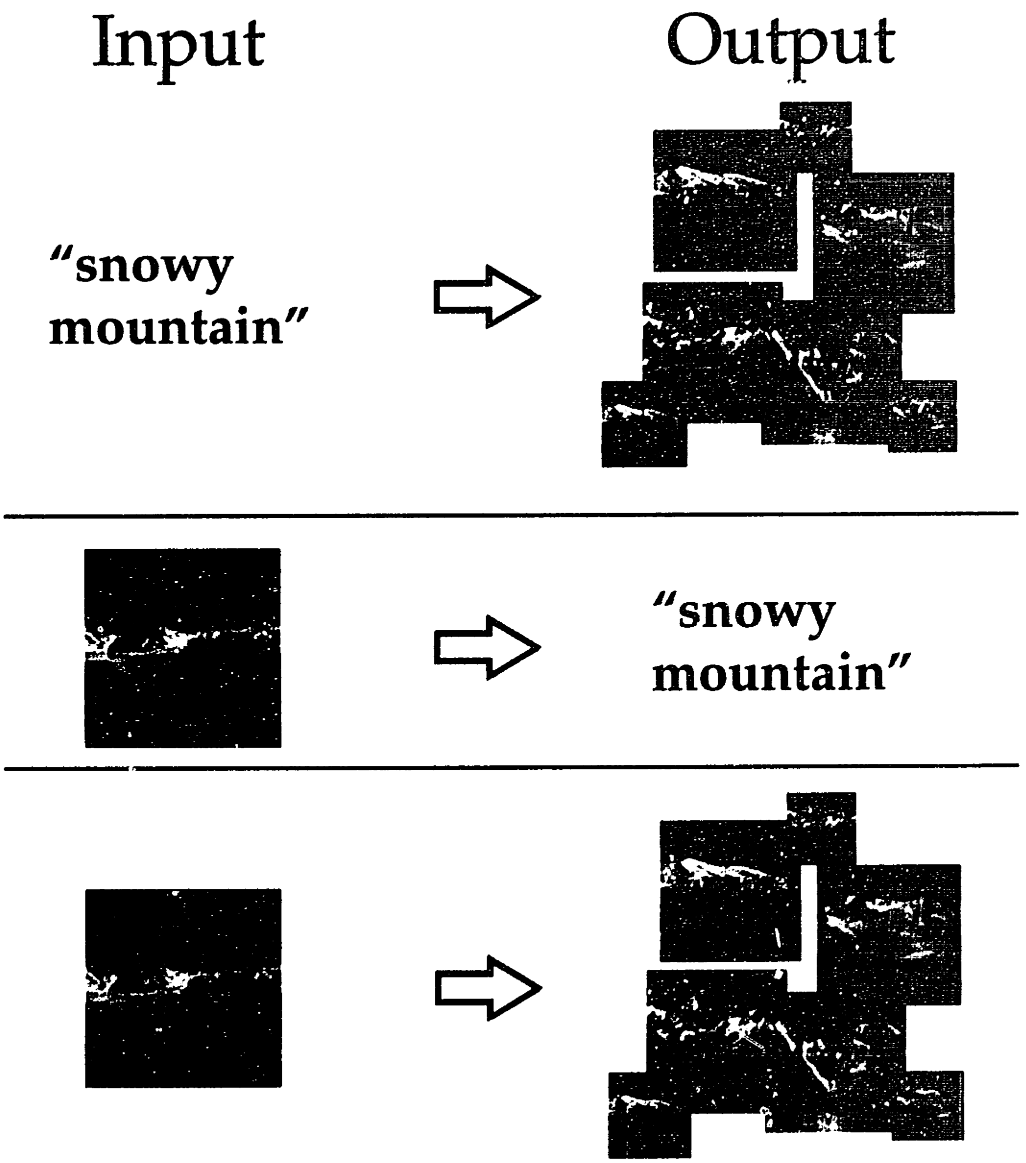


FIGURE 1.3. ***Example of inputs to and desirable corresponding outputs from a scene classification system.***

1.1.2 Difficulty of the problem

In this thesis, we deal mainly with the classification of natural images. Figure 1.4 shows several different classes of natural images, including coastal scenes, field panoramas, underwater scenes, and snowy mountain scapes. The problem of natural scene classification presents some formidable difficulties due to the wide variety of possible stimuli and the diverse conditions under which the stimuli may be observed. For instance, even within a single class, such as the coastal images shown, scenes may differ greatly in their distribution of color, absolute position of regions, illumination, viewing position and even content.

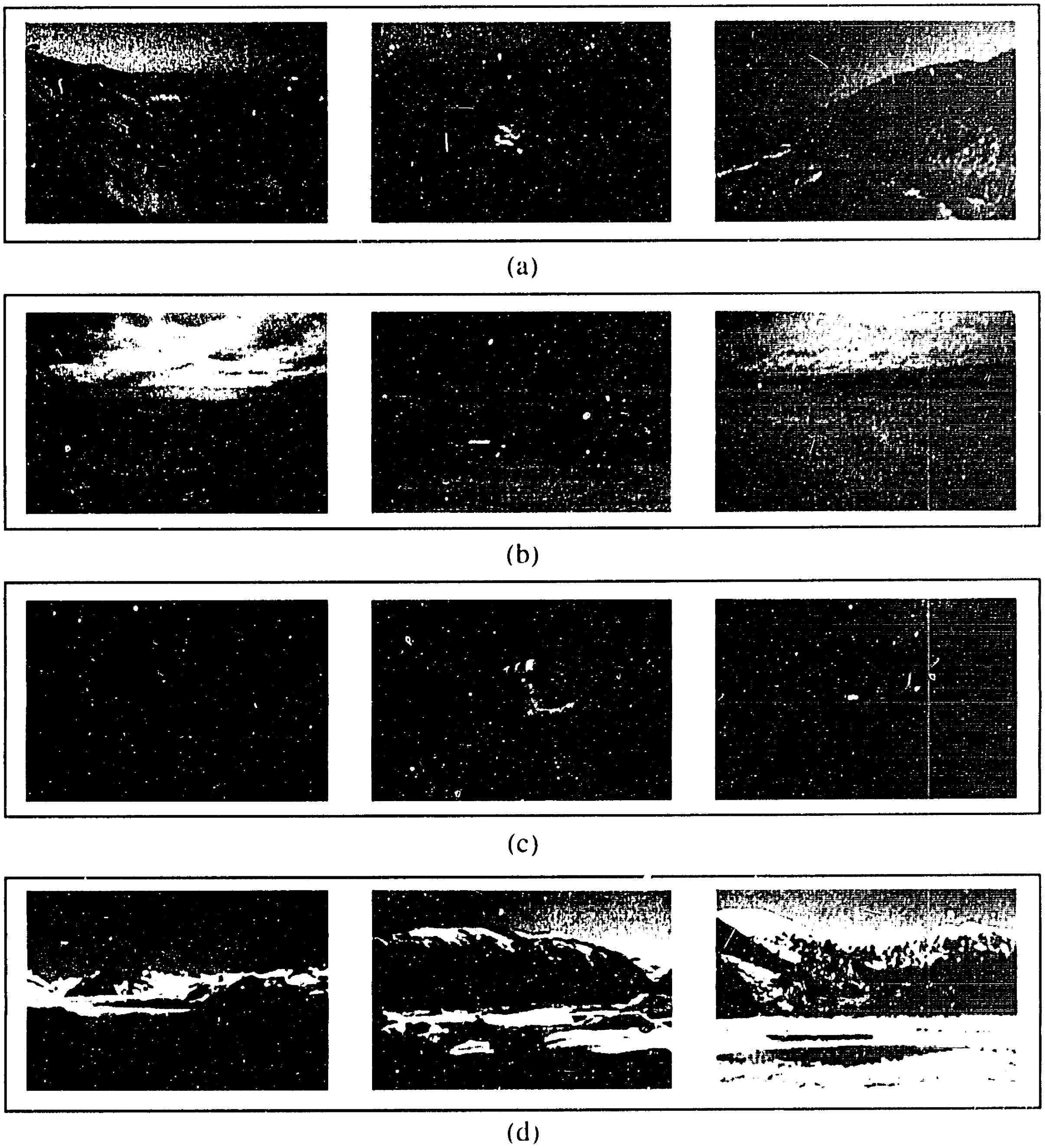
More generally, the difficulty of the problem of scene classification can be broken down into three areas:

1) The first is the dimensionality of the problem. Looking at a digitized image, every pixel in that image may be considered as a distinct piece of information. For instance, in a typical 512x512 pixel image, there are over 260,000 potential pieces of information. One way to reduce the complexity of the problem is to require the computer vision algorithm to extract only relevant or salient pieces of information from the image and use those as a basis for classification. We will refer to this reduced data as a set of primitives.

2) The second significant problem is to determine which are the relevant primitives and what are the relationships between these primitives that capture the "essence" of a scene. A scene may be characterized by, among other things, color, texture, geometrical shapes, and salient objects. A useful model needs to be able to capture the important aspects of the scene such that it is distinctive enough to distinguish between different stimuli and permissive enough to recognize the same stimuli under a variety of lighting, pose, and landscape changes.

3) The third significant problem is how to efficiently and robustly identify scenes from extracted image primitives. To identify a scene, we must relate the primitives found in the image to the primitives encoded in our model. In the worst case, the number of possible pairwise comparisons between image and model primitives is exponential. The complexity of the problem grows with increasing variations within a class and with the addition of sensor noise. It is desirable to have a method which can quickly identify relevant image primitives and decide whether those primitives together constitute a particular type of scene.

In this thesis, we will examine all of these problems.



(a)

(b)

(c)

(d)

FIGURE 1.4. *Example of several classes of images. (a) coastal images. (b) field panoramas. (c) underwater views. (d) snowy mountain scapes.*

1.2 Practical uses for scene classification

Scene classification is not just a problem of academic interest. It has practical uses in many types of domains, including but not limited to, a robot trying to identify its surroundings, a surveillance system that can automatically detect suspicious activity outside a high security area, a geological system that can identify and track different classes of weather patterns, a geographic monitoring system that can classify terrains both on Earth and on other planets. The most topical application for scene classification is in the management of large libraries of digital images.

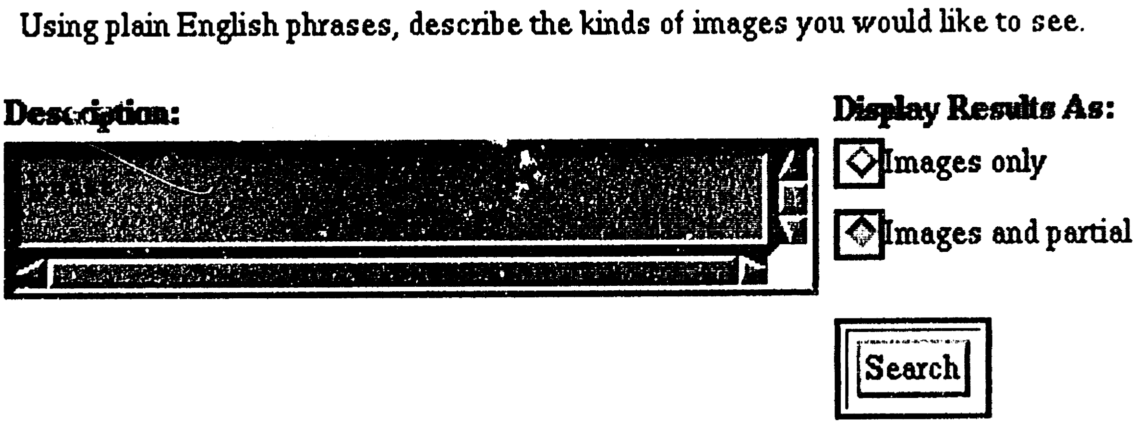
1.2.1 Image Database Indexing

Recently there has been a dramatic growth in digital image libraries. For example, collections from the National Geographic and the Louvre have been digitized. The Library of Congress intends to spend $60 million to digitize 5 million images [26]. In addition, the expansion of the Internet to a wide range of people coupled with user friendly web browsers have greatly increased the availability of and the demand to access on-line digital image stock from a wide range of providers.

With the increase in the number of available digital pictures and the ease of access to these libraries, the need has arisen for more complete and efficient annotation (attaching classification labels to images) and indexing (accessing specific images from the database via a query) systems. Image database indexing systems can be used in a wide variety of applications from a casual user browsing for images, to advertisers, news agencies, and magazine publishers seeking stock photography. Such systems may also aid in other areas including the management of art galleries, museums, scientific databases, and retail stock. Further applications are found in fashion and interior design. Database indexing is not limited to static image storage and retrieval. A complimentary stock of video sequences are being digitized and archived.

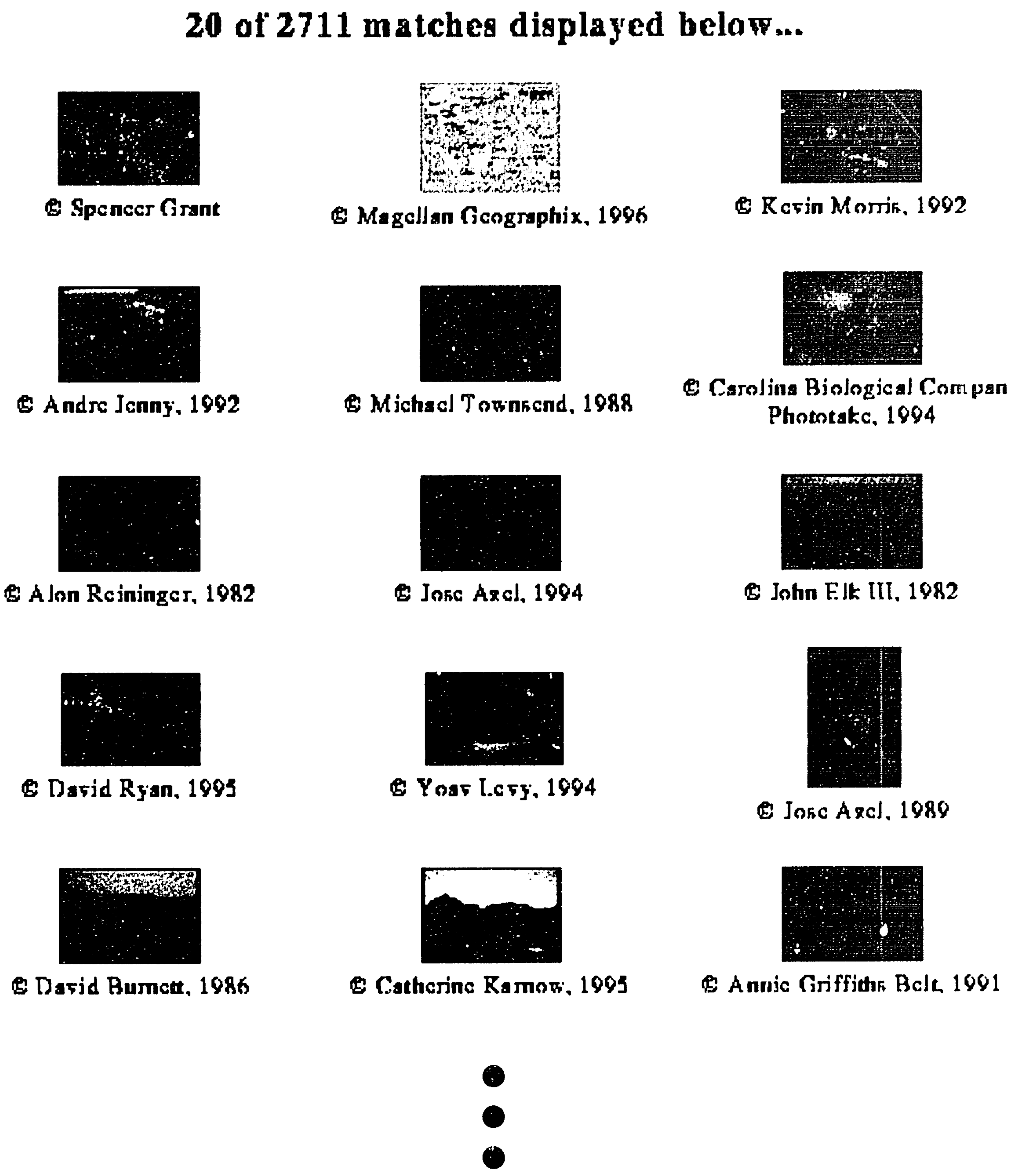
Traditionally, image/video database indexing systems have annotated the database with *key words*. A user retrieves desired images by submitting text-based queries to the system using one or a combination of these key words. Some of the textual query systems are so complexly designed that a middle-man familiar with a particular annotation system must be employed to carry out the third-party request. In a currenι state of the art system, the middle-men may actually perform the search by accessing the images from memory.

Figure 1.5 shows a simple textual query system from a stock photography vendor. The word "coast" was used as the input search string. Figure 1.6 shows a few of the 2700 images found by the query system based on the input. These two figures illustrate the main problems of trying to index into an image database using text based keys. The returned images can be quite varied and, in many cases, their content is often only indirectly tied to the original query. The latter point is best illustrated by the gull and map pictures. In such text based systems, the content of the image, the level of detail of description, and the vocabulary used is decided by the annotator. The annotator's biases and preferences in most cases will not match the biases and preferences of the user, resulting in many "false" positives, selected images that do not fit the user's expectations, and many "false" negatives, desirable images bypassed by the system.



Using plain English phrases, describe the kinds of images you would like to see.

FIGURE 1.5. *A textual query interface to an image database indexing system.*



20 of 2711 matches displayed below...

© Spencer Grant

© Magellan Geographix, 1996

© Kevin Morris, 1992

© Andre Jenny, 1992

© Michael Townsend, 1988

© Carolina Biological Compan
Phototake, 1994

© Alon Reininger, 1982

© Jose Azel, 1994

© John Elk III, 1982

© David Ryan, 1995

© Yoav Levy, 1994

© Jose Azel, 1989

© David Burnett, 1986

© Catherine Karnow, 1995

© Annie Griffiths Belt, 1991

FIGURE 1.6. *Images selected as a result of the text query "coast". Many of the selected images are quite varied and some do not match human intuition of what should comprise a coastal image.*

Such systems could be greatly enhanced if the queries could be performed automatically on the image content and if the query could be formulated and refined at the time of access. A more comprehensive system could tolerate queries of the form, "Given these example images that represent a class, return other images from the database that are also instances of this class". Similarly, a user could ask the system to return images that belong to a predefined class, such as "coastal". An interactive system would allow the user to refine the query by rating the returned images and resubmitting an improved class concept. A general automated image classification strategy could enhance the performance, flexibility, reliability, and ease of use of existing image indexing systems.

1.3 Aim and scope of the thesis

The aim of this thesis is to provide a solution for the problem of scene classification. This thesis will develop a novel approach called 'configural recognition' that is eminently suited to encode and utilize the global structure of scenes. Configural recognition utilizes qualitative relationships between colored regions as primitives and encodes such relationships in a global deformable template. The scheme will be shown to be computationally efficient - both theoretically and empirically. We will demonstrate our approach to classification on the domain of natural scenes. We will show results of the system tested on a database of 700 images. We will also suggest an interactive learning technique to develop the scene models.

The scheme developed here, although versatile in many situations, has some significant limitations. In particular, the solution presented in this thesis will not encompass situations where classification is based on fine quantitative discriminations. Additionally, the solution will not apply to classification of functionally or emotively defined scenes/objects.

1.4 Structure of the thesis

In this chapter we provided a brief introduction to the problem of scene classification and the important application of image database indexing. In the next chapter, we investigate how spatial configuration and global organization in scenes may encode the semantic content of a scene and, therefore, may be a key component for scene classification. Chapter 3 discusses the benefits of the use of qualitativc measures in scene models. In Chapter 4, we show how natural scenes can be classified using templates which encode a global configuration of qualitative relationships between scene patches. We demonstrate our approach on a 700 image dataset. Chapter 5 suggests strategies to automatically learn a class template from a set of example images. In chapter 6, we analyze the strengths and weaknesses of the configural recognition approach to scene classification. Finally, in chapter 7 we briefly discuss how the configural recognition approach might be applicable for other tasks such as object detection.

CHAPTER 2 The Importance of Global Organization

A key question that needs to be addressed by any approach seeking to perform scene classification is: what is the image information that would allow reliable and robust classification? More succinctly, what aspects of scene content are relevant for the purposes of classification?

2.1 The role of global organization or scene structure

Scenes may be described in a variety of ways. For instance, two popular proposals in this regard suggest describing a scene either as a collection of objects or as having some set of particular image statistics, such as color or texture measurements. However, it seems that in either proposal the description of the image pieces by themselves, either as labeled objects or as image statistics, may not fully capture a scene's content.

Figure 2.1 shows two collages both of which contain the same seven distinct and recognizable objects; four seashells, one piece of bone, and two pieces of rope. Perceptually, Figure 2.1(a) is interpreted as a largely random collection of these parts, while Figure 2.1(b) is seen as a person. The large disparity between the two perceptions can be attributed to how the objects are spatially arranged.

Let us consider another example to highlight the importance of an image's global spatial arrangement. Figure 2.2(a) shows several images of snowy mountains. They can be described as having the same histograms of colors; blue, white and grey-green. Figure 2.2(b) shows a waterfall image, a rocky coastline, and a scrambled snowy mountain image. All three of these images contain approximately the same amount of blue, grey-green, and white regions as the images in Figure 2.2(a), however, perceptually we would not characterize them as snow-capped mountains. The images in Figure 2.2(b) have the correct chromatic components, but they are arranged in the wrong overall configuration.

The main point of Figure 2.1 and Figure 2.2 is that, irrespective of the representation used for the image parts, object descriptions or image statistics, it is usually the overall configuration of those parts that is most critical for classification. These observations which have been presented anecdotally here derive strong support from several psychological studies. It has been shown that a stimulus in correct spatial configuration allows for more accurate and rapid detection or recognition of itself or its parts than the same stimulus with incorrect spatial relations [5][7][12]. The conclusion we arrive at is that the overall organization of a scene's parts or scene structure strongly influences its interpretation. This idea will be one of the central themes of this thesis.

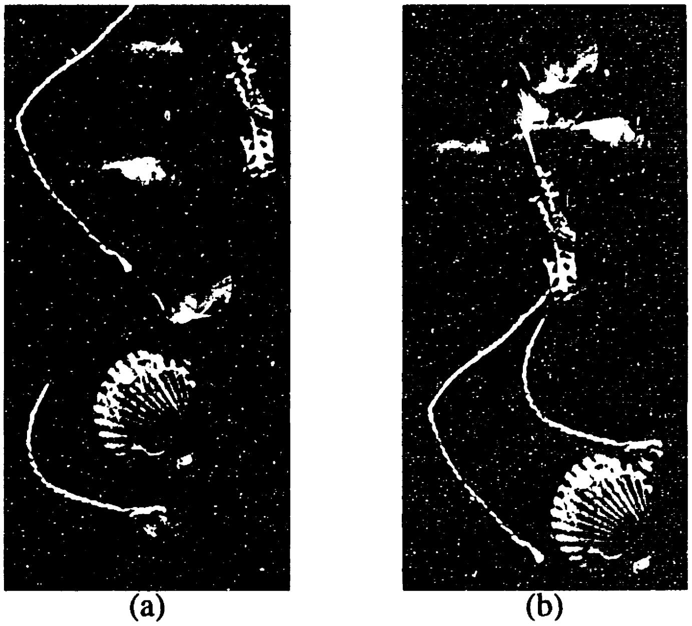


FIGURE 2.1. *Two versions of Andre Masson's "Ludion: Bottle-Imp", containing shells, rope, and bone; (a) contains a scrambled version, (b) shows the original.*

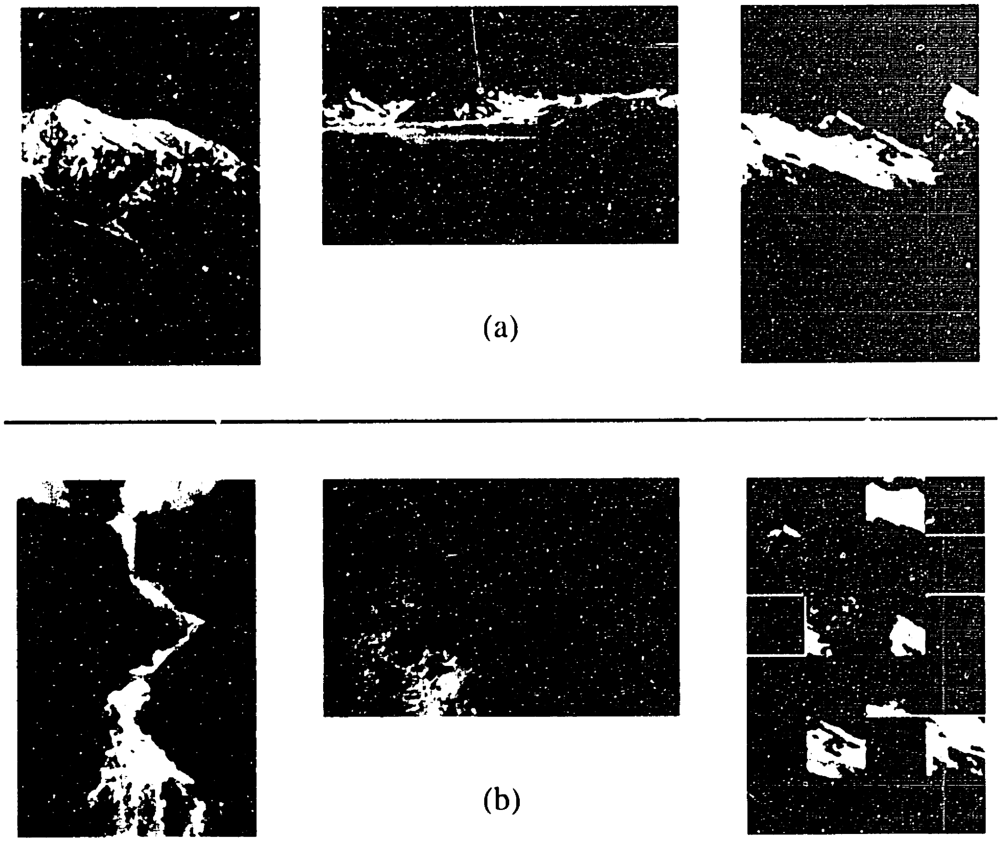


FIGURE 2.2. *(a) Three pictures of snow capped mountains. (b) Scenes of a waterfall, a coastline, and a scrambled mountain image.*

2.2 Representations for encoding scene structure

If scene organization is an important component for classification, the next question that we must address is precisely what information from the scene is used in that organization. In the previous section we described a scene in two ways, as a global organization of recognized objects, such as shell, bone, and rope, or as a structured configuration of image statistics, such as regions of blue, white and grey colors. Let us now examine these two strategies in a little more depth.

2.2.1 How important is image parcellation into distinct objects?

Whether it is necessary to recognize objects in an image before performing the classification is a fundamental question for scene classification. We addressed this issue briefly in the introductory chapter. In this section, we present a mcre comprehensive discussion.

Clearly, object recognition is important in some situations. Figure 2.3 shows Picasso's painting "Woman Dressing Her Hair". This is an example of a structure which is recognizable even though the spatial relationships between the constituent objects are jumbled. This is probably because of the recognizability of the individual objects. This is best illustrated in the face region. The face is recognizable, because its subparts, the eyes, nose and mouth are evident, even though they are incorrectly organized.

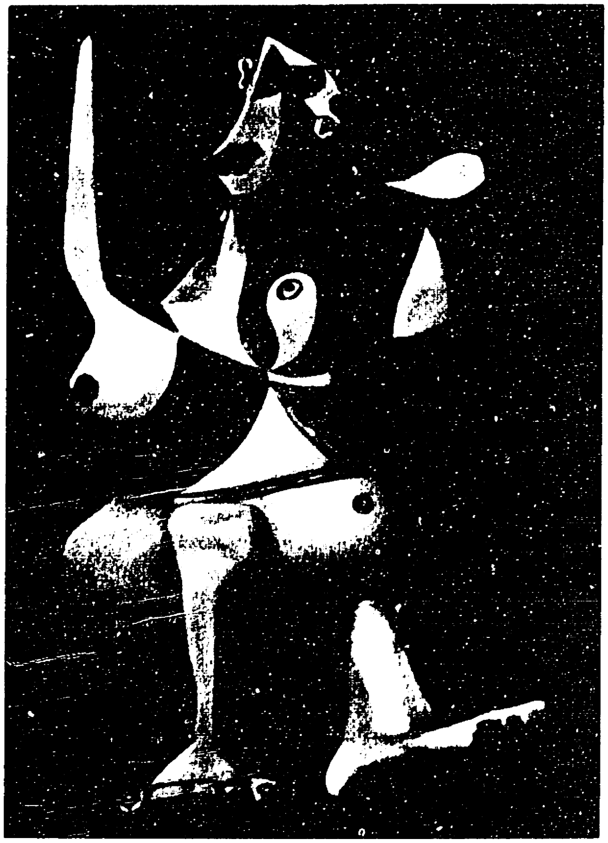


FIGURE 2.3. *"Woman Dressing Her Hair" by Picasso. This figure is recognized as a person because the individual parts such as the eyes, nose, and mouth are recognizable, even though these facial features are in the wrong spatial organization.*

On the other extreme are situations where in some scenes are recognizable even though their individual parts are not. Figure 2.4 shows four parts of an image. In isolation, these four parts are unidentifiable. However, in the context of the whole image, they become evident respectively as a portion of a tree, part of the sky, a person, and a swatch of grass. The intact painting is shown in Figure 2.5.

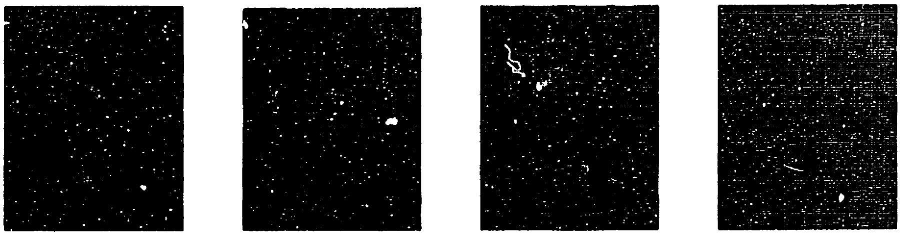


FIGURE 2.4. *Four parts of an image. Alone they are unrecognizable*

A similar and perhaps more convincing demonstration is provided by low frequency counterparts of images. Figure 2.6 shows the low frequency components of several images. The images contain only a global organization of color regions. These images are easily recognizable by human observers, even though none of their constituent parts by themselves are.

The last two points suggest that scene classification can proceed without the need for recognition of the individual parts. This suggestion is 'good news' for an automated scene classification technique, because object recognition is a difficult problem in its own right. A scene classification strategy which attempts to perform individual object recognition as an initial step may be thwarted by the difficulty of the recognition component.

There is a large body of work in computer vision which addresses the problem of object recognition. Some examples include [24][25][30][44][51][67][69]. Most of these strategies use geometric models and are aimed at recognizing static objects. The successes in this area have usually been when the objects have well-defined boundaries, are largely unoccluded and viewed under constrained lighting conditions. These strategies are not well suited for complex scenes consisting of multiple objects which may be viewed under varying viewing and lighting conditions. The arrangement of the objects and the resulting occlusions, shadows, and reflections greatly increases the complexity of object recognition (see Figure 2.7).

Some strategies have been developed to try to focus the attention of the recognition algorithm on parts of the scene that belong to one object [10][39][45]. Although there has been some success using geometry and color information to solve the problem of localizing unoccluded objects in a cluttered scene, focusing attention and recognizing most or all objects, irrespective of the arrangement of objects in a scene, is extremely difficult and currently computationally prohibitive.

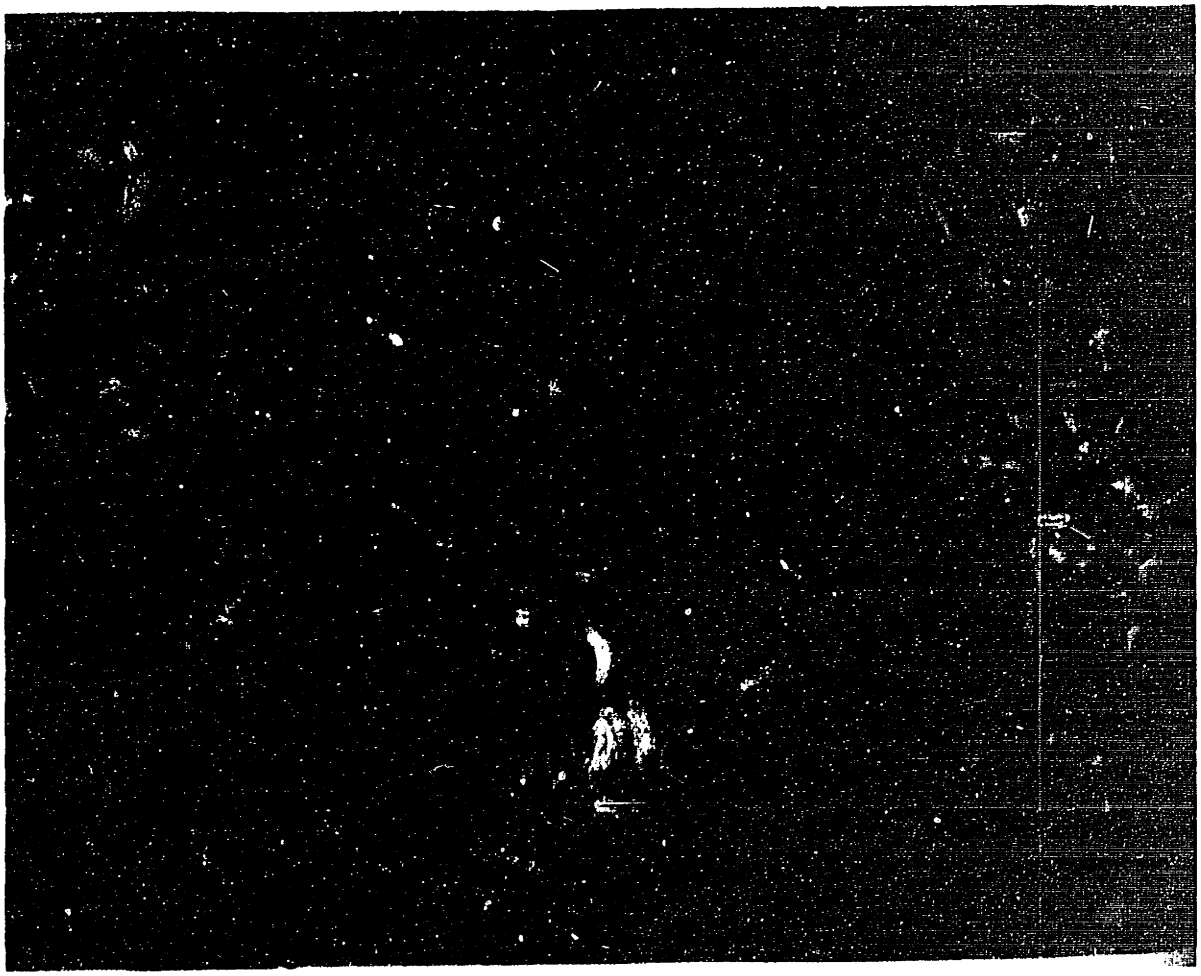


FIGURE 2.5. *Painting by Maurice Prendergast, "Summer in the Park". An example of an image that can be meaningfully interpreted due to the spatial arrangement of the different subparts, even though the latter are not identifiable by themselves (see Figure 2.4).*

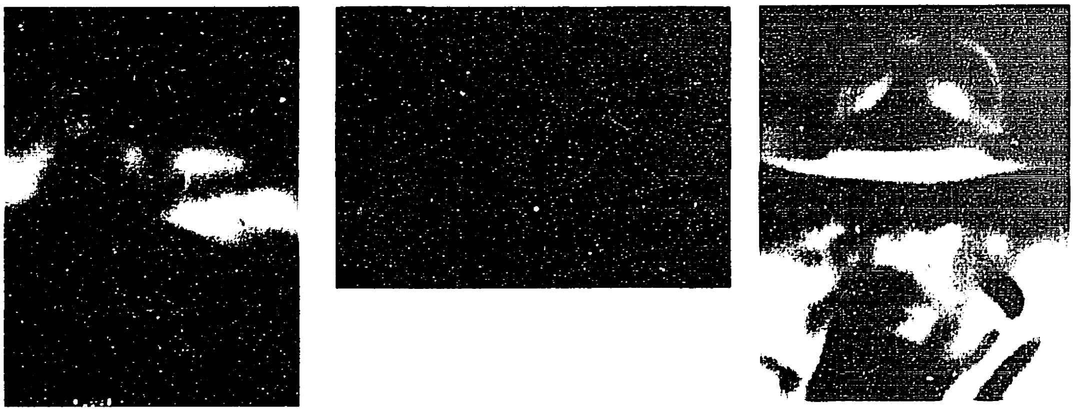


FIGURE 2.6. *Low frequency components of three images; a snowy mountain scene, a car, and a face. All are recognizable even though their individual parts may not be.*

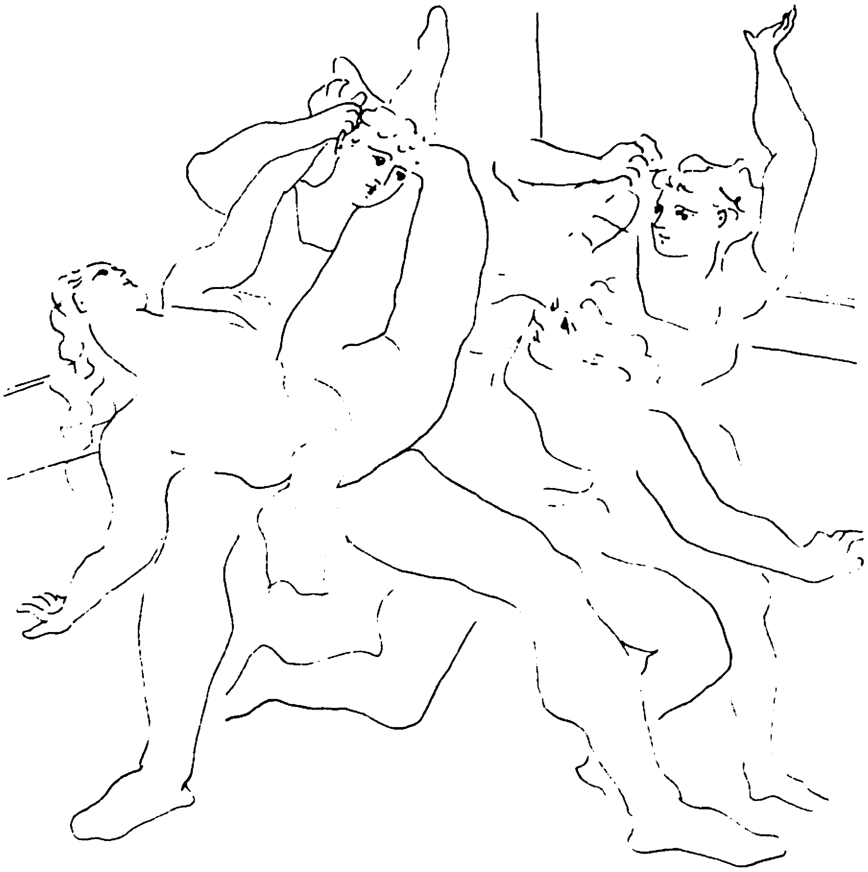


FIGURE 2.7. *Partioning objects and recognizing them in complex scenes is sometimes perceptually difficult. This drawing by Picasso has no complicating colors, textures, shadows, or reflections., however it is quite difficult even for humans to segment out the individual women due to their intricate arrangement. An automated system would be greatly challenged by such an image.*

An additional drawback that conventional object recognition schemes suffer from is that their largely geometric strategies are ill equipped for recognition of natural objects such as trees, rocks, water, coastlines, and clouds which constitute a major portion of all outdoor scenes (see Figure 2.8). There have been attempts to use color and texture information to segment natural objects in such scenes [27][50]. However, these techniques are often time consuming and may be prone to error due to the great color and textural variabilities of natural objects in the same image and across different images (see Figure 2.9).

Considering all the problems that plague individual object recognition schemes, the idea of recognizing the scene as a whole without first labeling its individual parts seems very attractive. This idea has some support in the area of psychology, especially in the area of face recognition/detection. Tanaka *et. al.* show that parts of faces, such as the eyes, nose, and mouth, of identifiable people are more quickly recognized when presented in an upright face, or in context, than those parts presented alone [66]. The results suggest that faces in the correct configuration are recognized in a more holistic manner, i.e. as a pattern, than via the identification or labeling of individual salient features. Results favoring a more holistic pattern matching strategy have been found with other objects. For instance, Cave and Kosslyn found that the particular type of division of an object into parts, e.g. unnatural vs. natural, where the parts were in a correct spatial arrangement, had little effect on speed or accuracy of recognizing the object [12]. Cave and Kosslyn suggest that representations of parts may be extracted *after*, rather than before, object identification.

Even though this thesis emphasizes the idea of recognizing scenes without first explicitly recognizing the individual scene parts, it is important to note that the approach we propose is not mutually exclusive with one based on specific object detection/recognition. In fact, it is likely that in a practical instantiation, such as an image database indexing system, both techniques may be fruitfully combined.

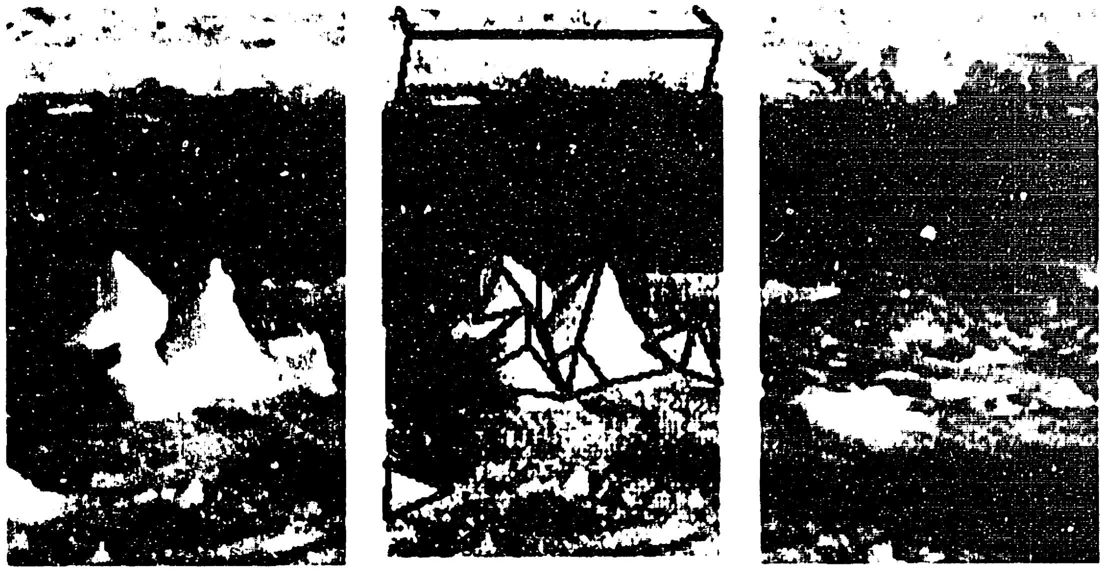


FIGURE 2.8. ***Illustration that geometric models are ill-suited for natural objects. For instance, although the ice chunks and ice slabs in the left image may be somewhat described via compact geometric forms such as polyhedral shapes, describing the ice in the right image via these shapes is much more difficult.***

Throughout this discussion, we have assumed that the boundary between what constitutes an object and what constitutes a scene is defined (or similarly what constitutes an object and an object subpart). In fact, this distinction is not always so clear. We will discuss in Chapter 7, how the approach developed for scene classification can effectively be used for individual object recognition/classification. So, the question of whether scene recognition is fundamentally distinct from individual object recognition might not, in the final analysis, be too meaningful.

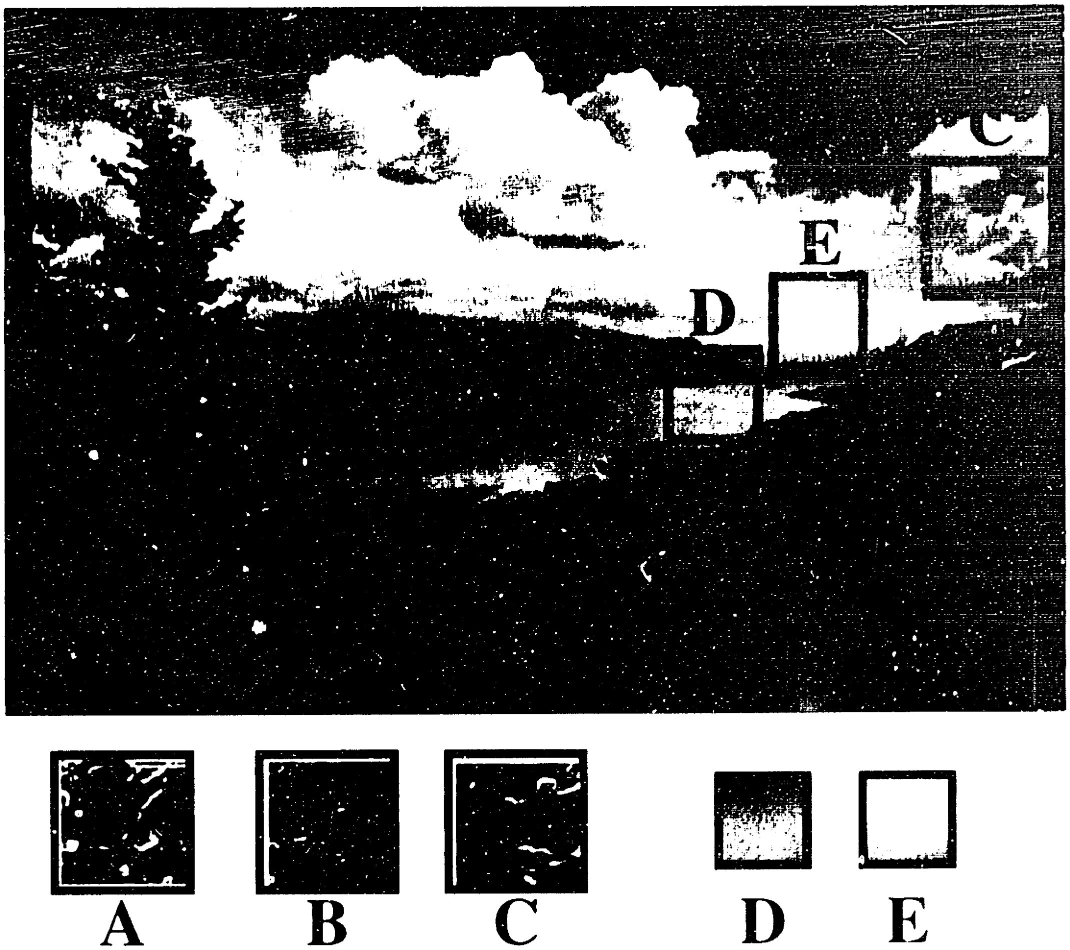


FIGURE 2.9. *Segmentation of natural images using color and texture properties is a difficult problem. This figure shows a scene with five highlighted regions: A and B are tree regions, D is a water region, E and C are sky regions. The textural properties of A, B, and C and the average color properties of D and E are shown below the image. The texture of A and C seem more similar than A and B even though A and B are both tree regions, while C is a sky region. The average color properties of D and E seem very similar even though they are from a water and sky region respectively. Segmentation based on these cues may lead to an incorrect partioning of the image.*

2.2.2 *Encoding image structure as an organization of colored pixels*

Instead of thinking of an image as a collection of distinct objects, let us consider it to be merely a set of colored pixels. In this framework, we have to handle the important question of how to represent the structure of this pixel set. There are multiple possibilities that can be placed on a continuum whose two ends are defined by 1) conventional template matching where image structure is represented exactly as the absolute color and absolute position information of every pixel and 2) cumulative statistics, such as color and luminance histograms and Fourier amplitude spectrum signatures, where no positioning information may be encoded.

In the first case, the actual image or subpart of an image is used as a model, often referred to as a template. New images are classified or recognized if they contain the model template or some constrained distortion of the template. Current applications that utilize templates include systems for face recognition [11] and sign post detection [6]. There are several benefits to using templates. The template is easy to store as an image. No costly preprocessing is necessary to generate the template or to prepare a novel input image for matching. Matching the template to an image involves differencing or correlational operations, which can be implemented quite efficiently.

On the other extreme, cumulative statistics have also been used successfully in some limited recognition applications. For instance, Swain and Ballard implemented a recognition technique using color histograms as object models [63]. Cumulative statistics have several advantages. Most statistics are easy to compute. Models in the form of these statistics are compact. Matching between models and the novel processed images involves comparing a small set of numbers which is computationally attractive.

Neither extreme, however, has been successfully used for the classification of scene content. Figure 2.10 and Figure 2.11 illustrate why neither extreme is appropriate for scene classification. Figure 2.10 shows two ice scenes and a picture of their difference. Although both belong to the same class, the absolute colors and absolute positions of the image regions are quite different, rendering the template matching scheme ineffective. In general, two instances of the same class may differ greatly in their absolute measurements. In addition, there is most likely no simple distortion of one to fit the other to produce a good match. Figure 2.11 addresses the other extreme. The figure shows a snowy mountain image and its scrambled counterpart. The color histograms for each of the images are exactly the same, even though the images do not belong to the same class. This suggests that scene content often may not be well described by global measurements.

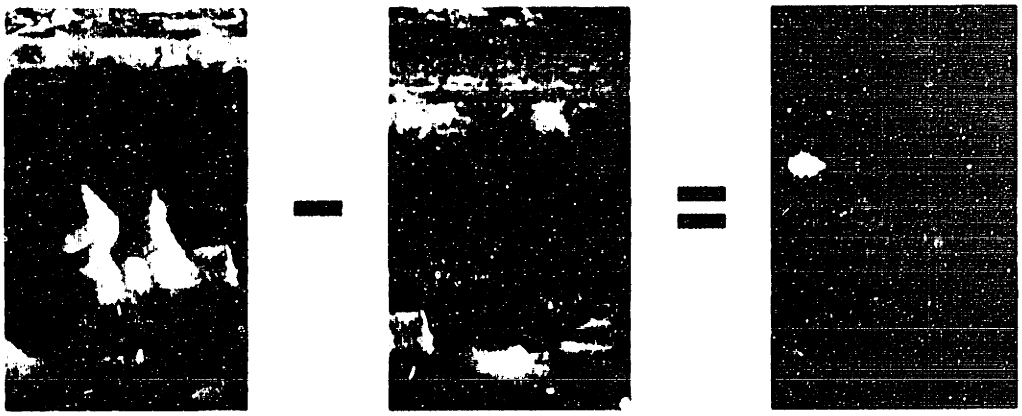


FIGURE 2.10. *A measure of similarity between a template (first ice scene) and a novel image (second ice scene) can be defined as function of their absolute difference. Although the two ice images belong to the same class, their absolute difference (third picture) is large, suggesting that template matching may not be an effective scene classification technique.*

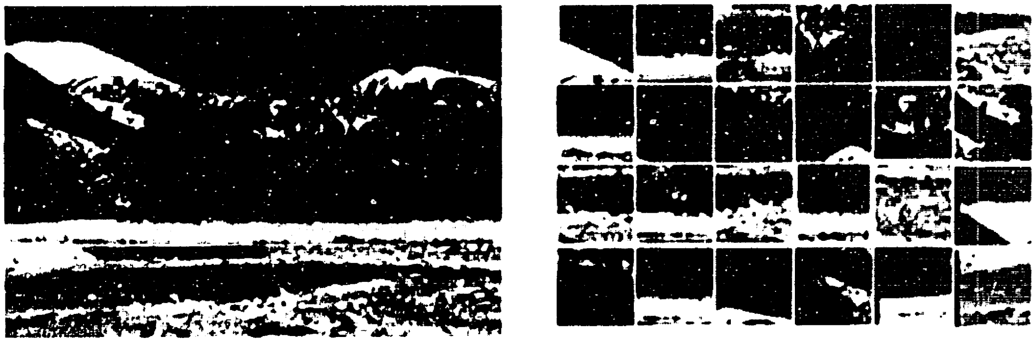
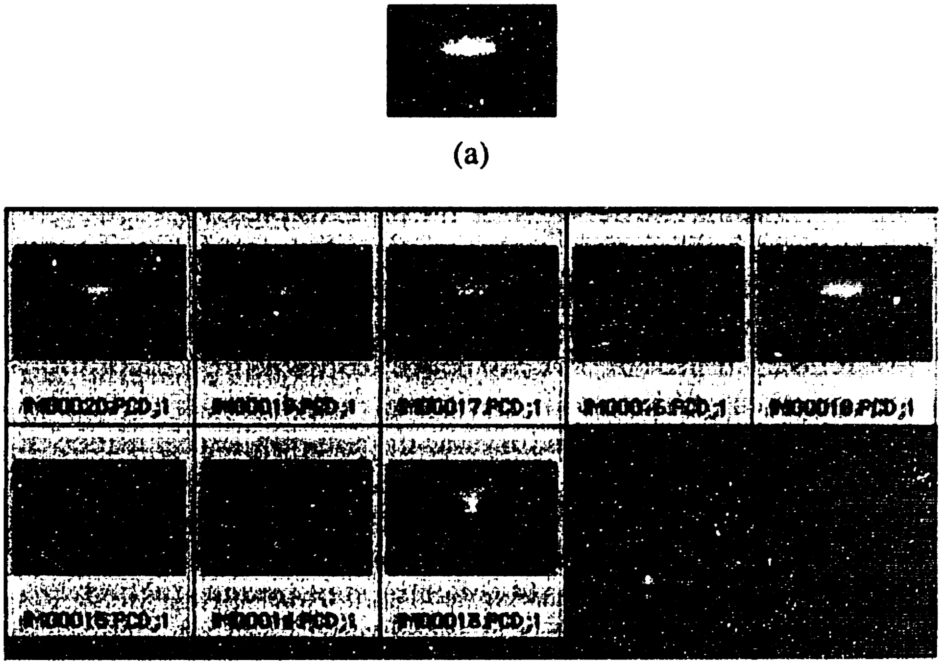


FIGURE 2.11. *A measure of dissimilarity for a cumulative statistic technique can be defined as function of the difference between the color histograms of two images. The two images shown in this figure (a snowy mountain image and its scrambled counterpart) have exactly the same color histograms even though they do not belong to the same class. This suggests that comparing cumulative statistics may not be an effective technique for image classification.*

Because of the problems outlined above, real applications of these two extremes in image database applications have not been too fruitful for retrieving images based on their perceptual content. For instance, Equitz uses a form of template matching to find images in the database that are most similar to an input image [18]. The result of the queries, shown in Figure 2.12(b), are images that are almost identical in configuration and luminance to the input image, shown in (a).

No evidence of class generalizabilty was demonstrated. Qubic, another indexing system, uses color histograms as one component of their search strategy [4]. When a color histogram of the input image is the main basis for the query, the resulting images are similar in overall color but can differ greatly in their perceptual content. Figure 2.13 shows the results of one query of this type. The top left image was the query image, the other seven are the ordered closest matches. This image of a boat at sunset was is found to match most closely with an image of money, a sand dune scene, a image of molten liquid and a picture of a woman eating a slice of watermelon. The two images that perceptually match most closely in content with the input image did not even have the highest color similarity scores.



(a)

(b)

FIGURE 2.12. ***A query based on low resolution template matching. The input image is shown in (a). (b) contains the retrieved images. Only images which are an exact match or a slight deviation from the input image are returned.***

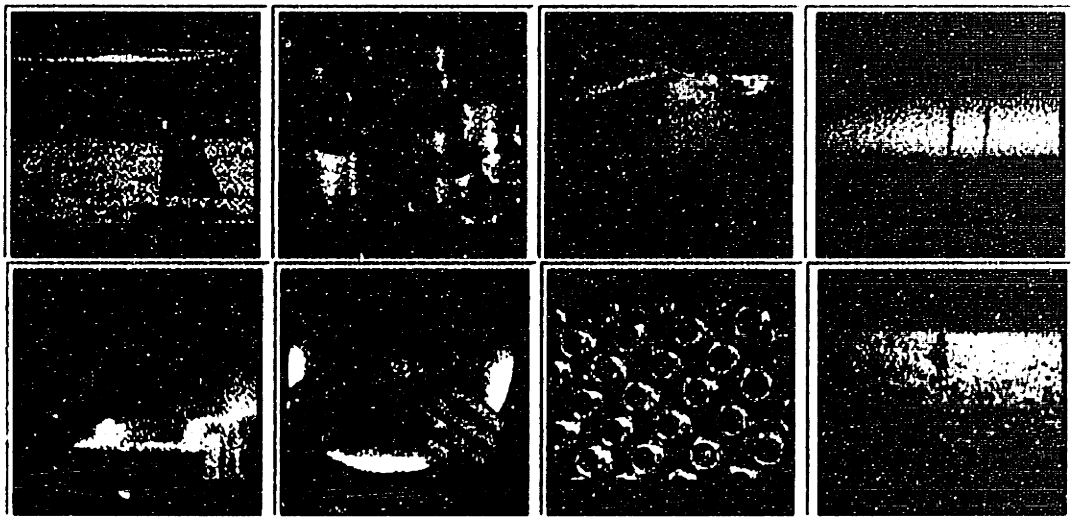


FIGURE 2.13. *A query based on color percentages. The top left image of the sailboat at sunset was the query image. The other seven are the closest matches from the database in order of computed similarity. 5 of the returned images come from such diverse classes as stacked money, desert scenes, molten liquid, and a woman eating watermelon.*

Ideally, one would like to use a strategy positioned between these two extremes. Such a strategy would encode the 'general' perceptually salient structure of the scene. General salient structure may be expressed as relative spatial relations between scene subparts. In the next section, we suggest a framework for encoding 'general' structure via qualitative relationships.

2.3 Qualitative encoding of scene structure

Qualitative measurements coarsely encode the *relative* relationships between entities, such as spatial position between two regions in an image. This is in contrast to quantitative measurements which express the absolute value of those entities, e.g. *x-y* image coordinates of those regions, or cumulative measures which express no information of the value of the individual entities. Qualitative relationships allow us to capture a flexible representation of the structure of a scene, while retaining some information about the individual components of that scene.

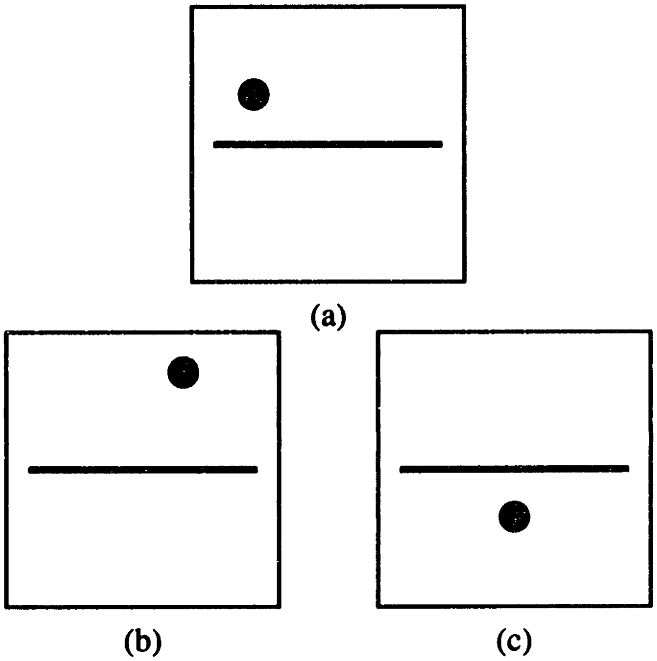


FIGURE 2.14. *(a) Original figure (b) dot displaced by distance d but still above the line. (c) original dot moved by distance d to a point below the line. Stimulus (a) is usually rated perceptually more similar to (b) than (c) suggesting the idea of perceptual grouping based on qualitative spatial relationships.*

Qualitative or relative relationships have been used effectively in the psychology community as a model of how humans make categorical judgments. Figure 2.14 shows one example of how relative relationships can be used as a measure of perceived similarity. Figure 2.14(a) shows a dot above a line. Figure 2.14(b) shows the dot above the line but displaced in the *x-y* plane by some amount *d*. Figure 2.14(c) shows the dot displaced from the original position in Figure 2.14(a) by the same amount *d*, but in a direction to put it below the line. Most observers when asked to rate whether (b) or (c) is more similar to (a) report that (b) is perceptually closer in nature to (a) than (c) to (a). This suggests that observers might be using some qualitative notion of whether the dot is above or below the line to make class judgments.

Above/below is one type of qualitative spatial relationship. The qualitative horizontal analog of this relation is left/right. Various combinations of these two types of spatial relationships provide a general language to describe a scene structure and a way to determine scene similarly or to perform scene classification. Although we have discussed the use of qualitative relationships for describing synthetic scenes, by analogy qualitative relationships might be important for real scenes as well.

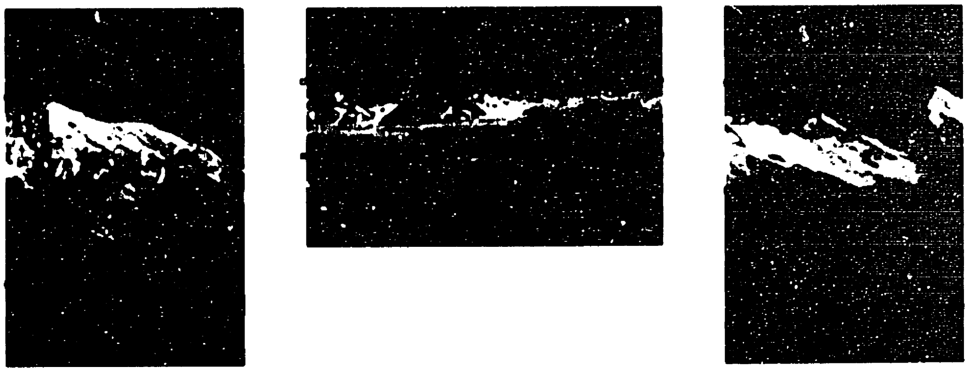


FIGURE 2.15. *The three snow-capped mountain scenes from Figure 2.2(a) are shown. Each is divided into three regions (A,B,C). Perceptually, the corresponding regions have similar content. Across all the images regions A, B, and C have the same relative spatial attributes (although they differ in their absolute sizes and positions).*

With respect to the spatial layout of real scenes, scene classes may also be described by image regions which relate to each other via these qualitative relationships. For example, Figure 2.15 shows the snow capped mountains from Figure 2.2(a). This class of images may be described as having three perceptually salient regions, a blue region (A), a white region (B), and a grey-green region (C). The corresponding regions have been annotated on the images. In all of the images region A is above region B which in turn is above region C. Therefore, even though the particular instances of the class exhibit these regions at diverse absolute locations (e.g. blue sky/white snow transition, measured in pixels from the top of the image, is at 50 in image 1, 38 in image 2, and 40 in image 3) and over different spatial extents (region B is 50, 27 and 33 percent respectively of images 1, 2 and 3), one constant is that the regions all have the same relative spatial layout.

2.4 *Summary*

In this chapter, we have discussed the importance of the global spatial organization or scene structure for the classification of images. There are several points we addressed:

•We illustrated that the spatial organization of a scene, whether the scene is described as a collection of objects or as having some set of particular image statistics such as color or texture, is critical for its classification.

• We addressed the question of what representations should be used for encoding scene structure and suggested that scenes may be represented as an organization of colored pixels rather than as an arrangement of objects. We also suggested that scene classification may in some cases precede object recognition or recognition of scene-subparts.

• We demonstrated two extremes for representing scenes as an organization of colored pixels. On one extreme the scene may be viewed as a template, where absolute spatial positions are encoded. On the other extreme the scene can be represented as a set of cumulative statistics, where no positioning information is encoded. We suggested that neither strategy is well suited for scene classification.

• We suggested that a more fruitful strategy may be to encode scene structure in terms of the relative spatial relationships between scene parts.

Relative spatial relationships between colored image regions provide a partial language to describe scene content. However, we still need to define how to represent other properties of the local regions such as color. For instance, loosely defined terms such as "blue", "white", or "grey-green" may have some perceptual meaning, but, they are difficult to encode in terms of digital image color spaces, such as the red, green, and blue color gamut (i.e. what values of red, green, and blue components combine to make "grey-green?). One idea to surmount this problem is to relate descriptions of the region properties to the structure of the scene. Therefore, just as we can define qualitative spatial relationships between image regions, we also can define qualitative relationships between other region attributes such as chromatic and luminance content.

We will show in following chapters that a combination of qualitative spatial relationships and qualitative region attributes provides a flexible but rich description of scene content and that this type of description can be used for reliable and efficient scene classification.

CHAPTER 3 Qualitative Models

The human visual system is remarkably adept at perceiving differences between the colors or luminances of image regions. However, it is very limited in its ability to estimate the absolute values of these attributes. Even in the detection of differences, the visual system seems to partition the relationships into coarse equivalence classes such as "brighter than" and "bluer than". This suggests that qualitative inter region relationships might be more important than quantitative absolute measurements for at least some visual tasks. This observation partly motivates our approach of encoding image structure in a effort to capture perceptually meaningful content in terms of qualitative relationships.

In many cases the relative or ordinal relationships between image regions are important for perceptual classification of scenes. The classification of a scene may remain valid long as the relative relationships between the image regions remain the same, even though the absolute region values may change. However, when the ordinal relationships are violated, often the percept and therefore the classification of that image is greatly altered. Figure 3.1(a) shows three images; a coastal view, a sunset panorama, and a picture of clouds. Figure 3.1(b) shows the three images where the contrast has been increased, however, the inter-region color and luminance relationships remain the same. Increasing the contrast is a linear function that stretches the difference between the high and low values in an image. Figure 3.1(c) shows the three images from Figure 3.1(a) inverted. In this case, the R, G, and B components of each image pixel have been inversely mapped, therefore, reversing the signs of most of the ordinal relationships within these three color bands. Observers report that the corresponding images to Figure 3.1(a) in Figure 3.1(b) do fall into the same class, although some perceive that the camera parameters or lighting conditions might have changed. On the other hand, subjects report that the corresponding images in Figure 3.1(a) and Figure 3.1(c) do not belong to the same class. For instance observers have reported that the inversion process on the coastal image produced an image of a glacier. The sunset panorama seems to have changed to a daytime water image with waves or clouds. Finally, the transformed cloud scene appears to be a relief map illuminated from below.

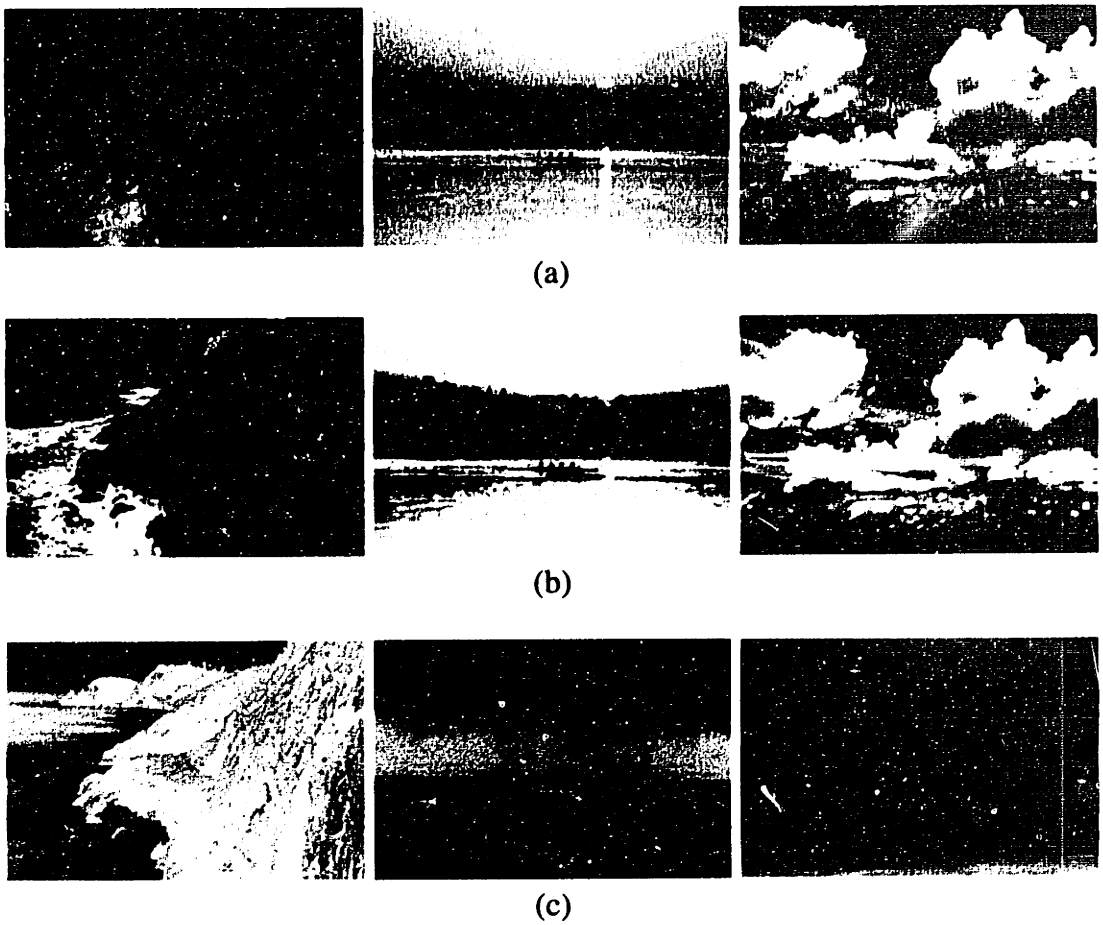


FIGURE 3.1. *Example of how relative relationships between image regions may be important for scene classification. (a) shows a coastal image, a sunset view, and a cloud picture. (b) shows the same images with increased contrast; the absolute values of the regions have changed, however, the ordinal relotions remain the same. (c) shows the images from (a) inverted; the ordinal relations between image regions are reversed. Perceptually, classification of the corresponding images in (a) and (b) are similar. However, classification of the corresponding images in (a) and (c) are quite different.*

Another motivation for using relative relationships in describing a scene class is that human perception of a visual stimulus is greatly influenced by the surroundings of the stimulus. For instance, the perceived luminance of an image patch can be altered by what surrounds that patch. Figure 3.2 is an example of the simultaneous contrast illusion [17]. The figure shows how a background gradient can make two equiluminant grey patches appear dissimilar. Figure 3.3, shows that a natural image patch may be described as "light" in one context, e.g. in a the coastal image, and "dark" in another, e.g. in a cloud scene, even though the average luminance of those patches are the same in both images. Analogous displays can be created to demonstrate a similar effect of context on perceived patch color. Thus, we suggest that an image region should be described in terms relative to the scene which contains it.

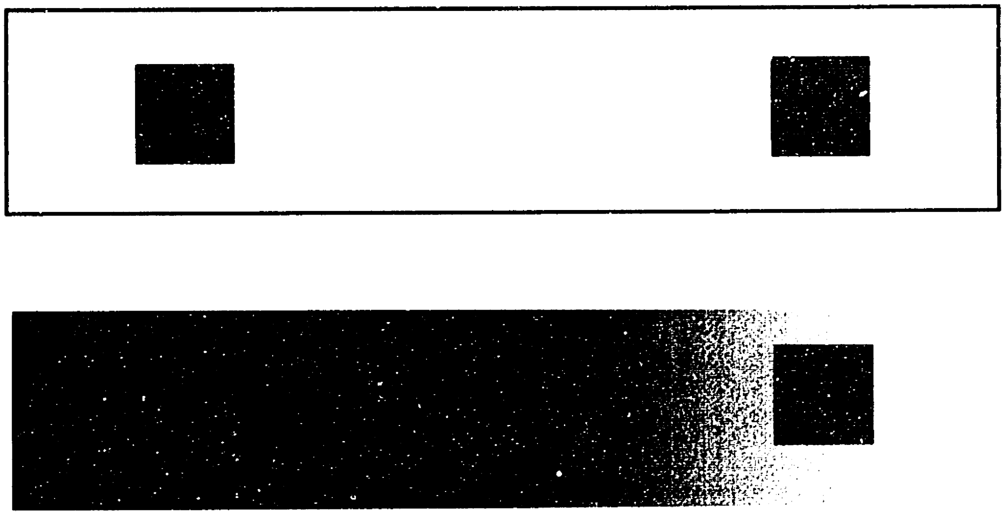


FIGURE 3.2. *Example of the perceptual effect of a background on two equiluminant grey patches. On a uniform background they appear to be the same intensity. However, when displayed on a background with a gradient the left patch appears to be brighter than the right.*

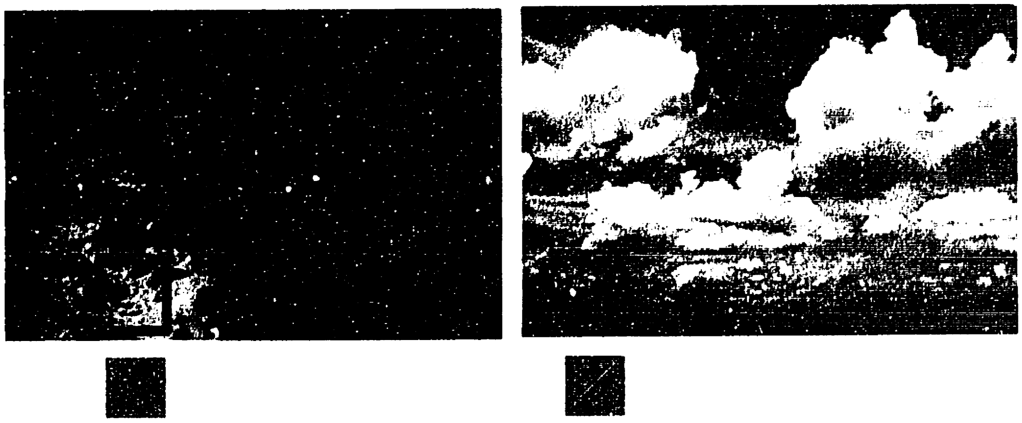


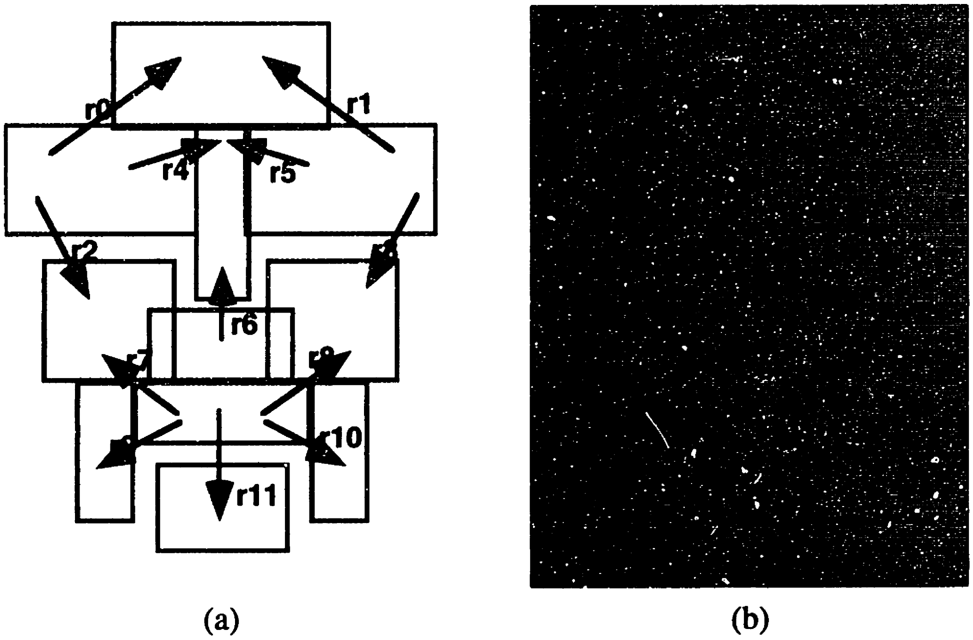
FIGURE 3.3. *Example of the effect of context on lightness perception. Within the context of the coastal image, the indicated patch is considered "light". Within the context of the cloud image, the highlighted patch seems "dark". However, both patches have the same average luminance. The average luminance of the indicated patches are shown below the images.*

In this chapter, we describe class models that use qualitative relationships between image regions. After reviewing prior work on qualitative object models, we give a concrete example of a language which can be used to describe a qualitative class model. We also describe the computational complexity of generating the class model and of matching the class model to novel images for the purposes of classification.

3.1 Prior related work using qualitative models

The most closely related work to what we are about to describe is the ratio-template construct devised by Sinha [57]. Sinha encodes sets of ratios of luminance values between image regions as qualitative object models. He has discovered that such models may be used for object detection under varying conditions. To show this, Sinha developed an invariant for frontal face detection under varying illumination conditions. The invariant consists of a set of image regions in a fixed spatial position, corresponding to facial features, and relative luminance relationships between the image regions. For instance, the model encodes that the regions corresponding to the eyes should be darker than the regions corresponding to the forehead, cheeks, and nose. Sinha demonstrated that the relative relations remained valid for a majority of faces and over many changes in illumination. The template can be evaluated at different locations and over several spatial scales of an image to detect instances of faces. Figure 3.4(a) contains a schematic of the template. Figure 3.4(b) shows the template overlaid on a face. Sinha has also developed a correlational learning scheme to learn ratio-templates from a set of example images. While this scheme performs well for the task of object detection, it seems not directly suited for situations wherein the structural arrangement of image entities can under go changes from one instance to another.

Smith and Chang have developed a system called VisualSEEK for image database indexing which uses spatial relationships between image regions as one component of a metric of image similarity. The query mode is to have the user specify color, texture, size, and absolute position of several image regions on a grid. Based on the patches input by the user, the system extracts measures of color, texture, position, and relative position and then uses a weighted combination of these cues to retrieve similar images from a database [58][59]. The emphasis, however, is on the use of multiple absolute measures for determining image similarity.



(a)

(b)

FIGURE 3.4. ***Example of a qualitative model which describes the class of frontal faces. (a) shows the patches, their spatial positions, and the corresponding relative luminance relationships. (b) shows the template applied to an image. The face is correctly detected.***

There has been some other prior work using qualitative spatial relationships in the context of scene classification to describe the relationships between objects or object subparts in images. One goal of such scene classification systems is to compute queries of the form “give me all the images where object O_i and object O_j have relation r_{ij}”. Another goal of such systems is to measure the similarity between two images or an image and a sub-image based on how well the geometric attributes and relative relationships of the scene parts match. Most of these systems bypass the difficult object recognition bottleneck by assuming that the objects are already labeled or by defining objects as simple geometric entities such as closed curves, which can be extracted from images using simple image processing techniques. The main focus of most of this work is on how the relationships, such as “next to” are defined given a set of object properties, such as center of mass, how the symbolic queries are represented, and how such queries can be efficiently processed [1][13][14][28][49][65]. However, the assumption of already labeled or easily extracted objects makes discussions of these issues more academic than practical. In addition, these strategies ignore the critical step of determining or learning what relationships between which salient objects are important for a scene class definition.

3.2 Model parts- image patches

As described in the last chapter, we suggested that a scene class may be modeled in terms of relative relationships between spatial and photometric attributes of image regions. Precisely what is implied by 'image region' must still be defined. The example of the snow capped mountain images partitioned into three salient regions in Figure 2.15 may have suggested that we should use a "smart" segmentation process to partition the images into patches of blue, white and grey-green in order to recover the structure of the scene. However, as suggested in chapter 2, segmentation of a scene via color or texture can be computationally expensive and can lead to perceptually incorrect partitions.

To partition an image into regions or patches, we adopt a much simpler approach. We can simply break up the image into *n* blocks, irrespective of the contents of those blocks. The blocks may be the size of a pixel or extend over many image pixels. The blocks may be of any size or shape. In the practical demonstration described in the next chapter, the blocks are usually equally sized and square.

The goal is to partition the image finely enough so that some patches extend over part of one perceptually cohesive image region. It is desirable, but not necessary, that the partitioning is coarse enough such that only a few patches together cover a cohesive image region. It may be necessary to partition the image at different scales so that hopefully at least one scale has a somewhat optimal covering of the image by these partitions. Figure 3.5 shows a snowy mountain region partitioned at three different scales. The image in the middle is an example of a desirable partitioning.

There will be many cases where the arbitrarily chosen image patches cover parts of two or more perceptually cohesive image regions. These patches may in effect be ignored, as long as there are some patches which cover part of only one cohesive image region. Such regions may be ignored both in the example images used to build a class model (see chapter 5) and in novel images.

The attributes of the image regions can be described in a number of ways. For instance, the luminance of an image region may be computed as the mean, median, or mode of the luminances' of the pixels that comprise that region. Definitions of mean and median are provided below [52]. In the most trivial case, where the image region corresponds to one pixel, the luminance of the image region equals the luminance of the pixel it covers. The position of the image region may be described by the centroid or one of the bounding points on the region. Similarly, when the region covers only one image pixel, the position of the region is the x and y coordinates of the pixel.

mean of $x_1, \ldots x_N$:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

median of $x_1, \ldots x_N$:

if the values x_j $j = 1, \ldots, N$ are sorted in order, then

$$x_{med} = x_{(N+1)/2} \quad \text{for } N \text{ odd}$$

$$x_{med} = 1/2(x_{N/2} + x_{(N/2)+1}) \quad \text{for } N \text{ even}$$

There are several benefits of this partitioning strategy. First, it is computationally simple. There are no complicated preprocessing or abstraction steps. The majority of the preprocessing work consists of image smoothing and averaging of region attributes. Second, we do not presume that the partitioning has segmented the image into connected perceptually salient parts. Segmentation of an image is a difficult problem. The expectation that the image has been segmented "properly", in most cases, will cause problems for the subsequent stages which rely on this information. Third, the class model comprises of image regions which are simple to describe (e.g. square image regions) and also easy to compare to novel images which have been partitioned in a similar manner.

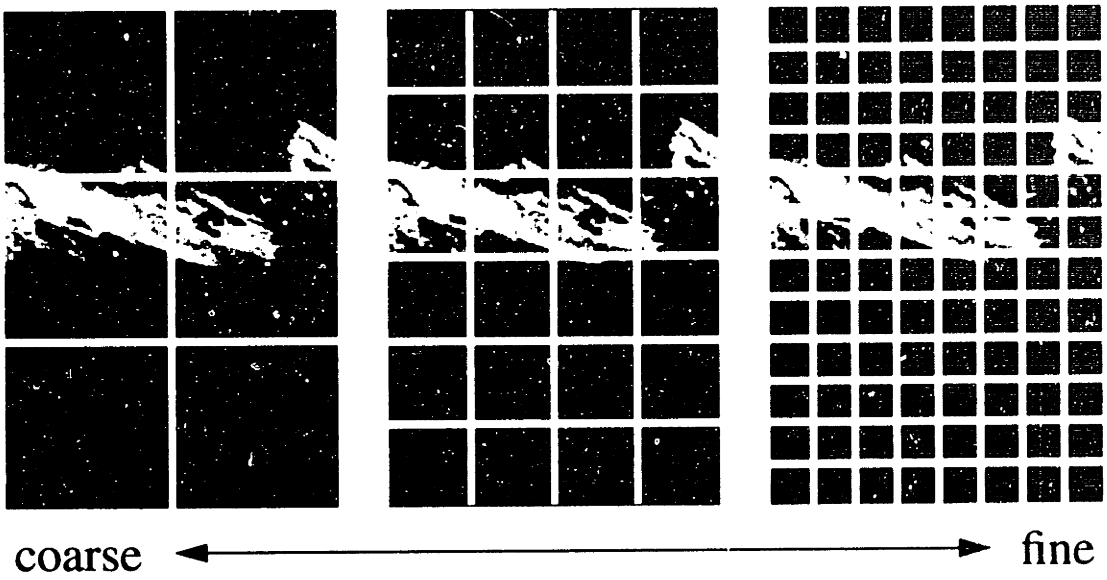


FIGURE 3.5. *Coarse to fine partitioning of the image into equal size square regions.*

3.3 Model parts - salient regions and their mutual relationships

A class model is a compact representation of a set of logically related images which share the same or similar characteristics. Thus, the class model should consist of only the salient regions and their relationships which embody those similar characteristics and which can be used to discriminate members of that class from non-members. In the most specific case, the model may represent only one example image. Therefore, all the image regions should be included in the model. This configuration is now similar to a template. In the most general case, the model covers all possible images and therefore need not contain any image regions or relationships. In between these extreme cases, only the salient or discriminatory details need be encoded in the model. For instance, many coastal scenes contain waves breaking on the shore, resulting in white image patches between the blue of the ocean and the brown or grey of the shore. However, the breaking waves are not a requirement for the general class of coastal images, and therefore, may be omitted from the model. The relationship between the ocean patches and shore patches will be the salient ones that distinguish the images that belong to the coastal class from images that are members of other image classes. However, if the desired class is "coasts with pounding surf", then the relationships between the breaking waves and the water, and the breaking waves and the shore become significant.

3.4 *Describing Qualitative relationships*

We now need a language which describes the qualitative relationships between model regions. Regions have many attributes. Some of the most basic ones are spatial position, color, luminance, and size. In the following four sub-sections we give examples of how to represent these relationships. We focus on this set in particular because they are the ones which we found were most useful in our applied task of classifying natural images, which is discussed in the following chapter. In subsection 3.4.4, we discuss how other region attributes such as shape, orientation, and density may be encoded in a qualitative fashion. These region attributes may become important for classifying other sets of images, such as synthetic images, space scenes, and fabric catalogues.

Qualitative relationships for each attribute may be computed between any subset of the patches in the model. For instance, if the relationships are pairwise, there are $O(n^2)$ total number of relationships to compute for any one attribute. Relations between subsets of triples or quadruples may be used. However, the increase in number of patches in the subset increases the exponent of the polynomial in the computation time. Experimentally, we found model descriptions of pairwise relationships to be sufficiently descriptive. In the following subsections, we give examples of relative pairwise relationships.

Figure 3.6 shows an image which consists of three salient patches (A, B, C). We can compute position, luminance, color, and size of each patch. Using these measurements, we can then find the qualitative pairwise relationships between the patches. The value of a qualitative measure is either greater than (>), less than (<), or equal to (=).

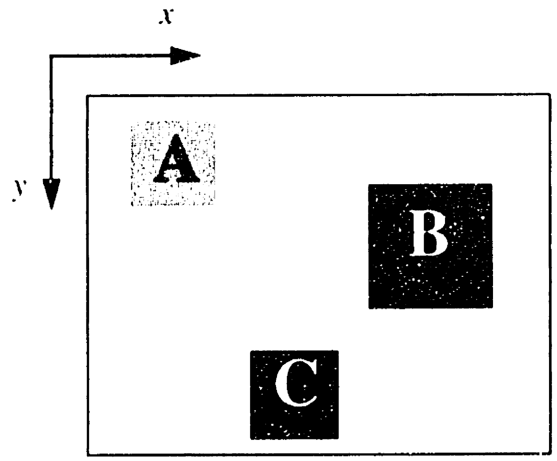


FIGURE 3.6. ***Example image with three salient patches. Patch A is light brown. Patch B is a purple blue. Patch C is dark green.***

3.4.1 Qualitative spatial relationships

Qualitative spatial relationships are based in terms of the image coordinates. Images are usually described as a two-dimensional array of basic elements or pixels. Image size therefore is denoted as the width times the height of the image in pixels. Figure 3.6 is a 167 by 127 sized image. We assume that the origin of the image (0,0) is in the upper left hand corner. If we use *x* and *y* coordinates of the centers of the image regions to denote their position in the image, region A is at position (26,22), region B is at position (121,51), and region C is at position (70,106). Let A_x and A_y denote the *x* and *y* positions of region A. Corresponding notation can be used for regions B and C. We can compute qualitative spatial relationships in both the horizontal direction (x) and vertical direction (y).

The six pairwise spatial relationships between the regions are the following:

$$\mathbf{A_x < B_x; A_y < B_y}$$

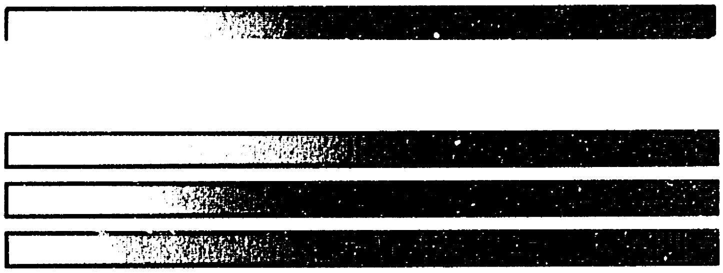
$$\mathbf{B_x > C_x; B_y < C_y}$$

$$\mathbf{A_x < C_x; A_y < C_y}$$

We can describe these relationships also in terms of above/below and right/left which may be more perceptually meaningful. In the vertical direction, "less than" corresponds to "above" and "greater than" corresponds to "below". In the horizontal direction, "less than" corresponds to "to the left of" and "greater than" corresponds "to the right of". Thus, region A is above and to the left of region B. Region B is to the right and above region C. Region A is to the left and above region C.

If we allow patches of varying size, it is possible for a larger patch to substantially overlap another patch in either the *x* or *y* direction. In such a situation, computing relative spatial relationships based on patch centers may not provide a perceptually valid or informative description. For instance, in the case where a larger patch overlaps a smaller patch in the *y* direction, the smaller patch may be described as "above" or "below" the larger patch even if its center is slightly above or below the center of the larger patch. In such cases, we would want to use information other than the middle of the regions to describe relative spatial positions. More sophisticated methods and terminology for describing the spatial relationships between two geometric regions or between two objects have been developed. References to some of this work are provided in section 3.1.

3.4.2 Qualitative photometric relationships



The photometric properties of an image region are its color and brightness distributions. Color and brightness of a region are dependent on the light received by an observer from that region. These photometric properties are a function of the lighting of the scene, reflectivity of the surface materials, and the three-dimensional geometry of the scene. An array of photosensitive devices in the image plane can be used to measure scene lightness at particular points or over small regions. Generally, the photometric properties of an image may be described by several sensors with different spectral sensitivities in the visible spectrum. The photometric values at one point in the array of sensors corresponds to a pixel's color and luminance[52].

One common way to describe the color of a pixel is via the measured red, green, and blue spectral intensities at a corresponding small patch in the real scene. This color scheme is commonly referred to as RGB. The R, G, and B values that are stored are based on a bounded and quantized partitioning of the spectral intensities. Often the scale for each of these color components ranges from 0 to 255. The color gamut described by this color model is produced by adding together all the different combinations of R, G, and B values. The luminance or brightness of a pixel is usually computed as a weighted sum of its R,G, and B. The weights commonly used for the R, G, and B values are respectively 0.3, 0.6 and 0.1. It is easy to visualize this color space as a unit cube oriented so that it is sitting on one of its points (see Figure 3.7). The three edges from that vertex correspond to the R, G, and B scales. The value at that point corresponds to black where R=G=B=0. The diagonal axis from that point through the cube, with equal amounts of each primary color, is the luminance axis. The other endpoint of this axis corresponds to white [20].

There are many other color schemes which can be used to describe the photometric properties of an image including hue, saturation, and brightness (HSB), cyan, magenta, yellow and black (CMYK), Luminosity and color axes a & b (LAB), and a retinal cone coordinate system (see [20] and [52] for full descriptions of these and other color models). Experimentally, we found the RGB coordinate system to be sufficiently descriptive of a scene's colors and well suited as a basis for qualitative measurements.

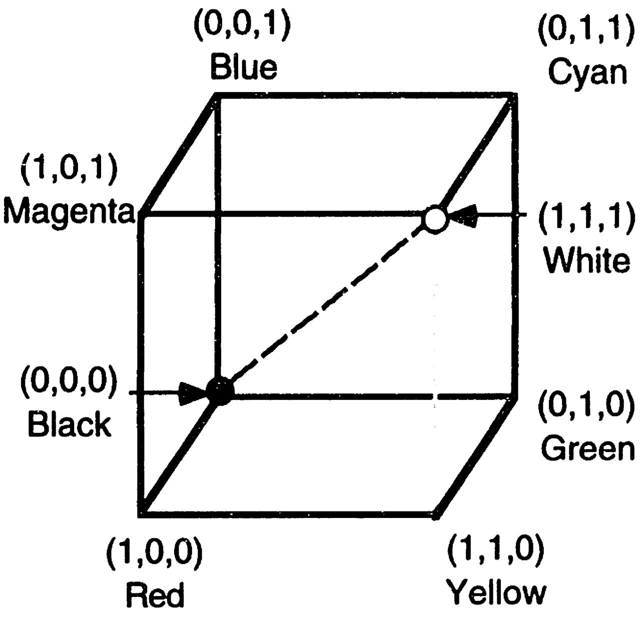


FIGURE 3.7. *The RGB cube.*

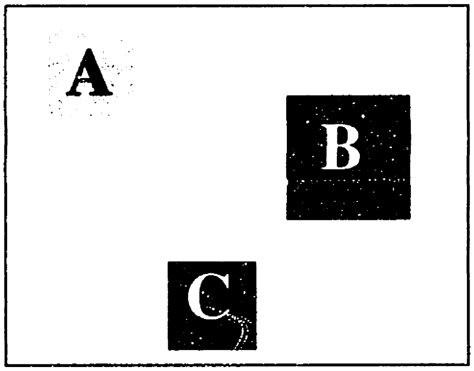


FIGURE 3.8. *The luminance component of Figure 3.6 is shown here.*

The luminance of each of the patches in Figure 3.6 is shown in Figure 3.8. The luminance values of region A,B, and C respectively are 206, 90, and 69. We can denote the luminance of a region in the form of A_l.

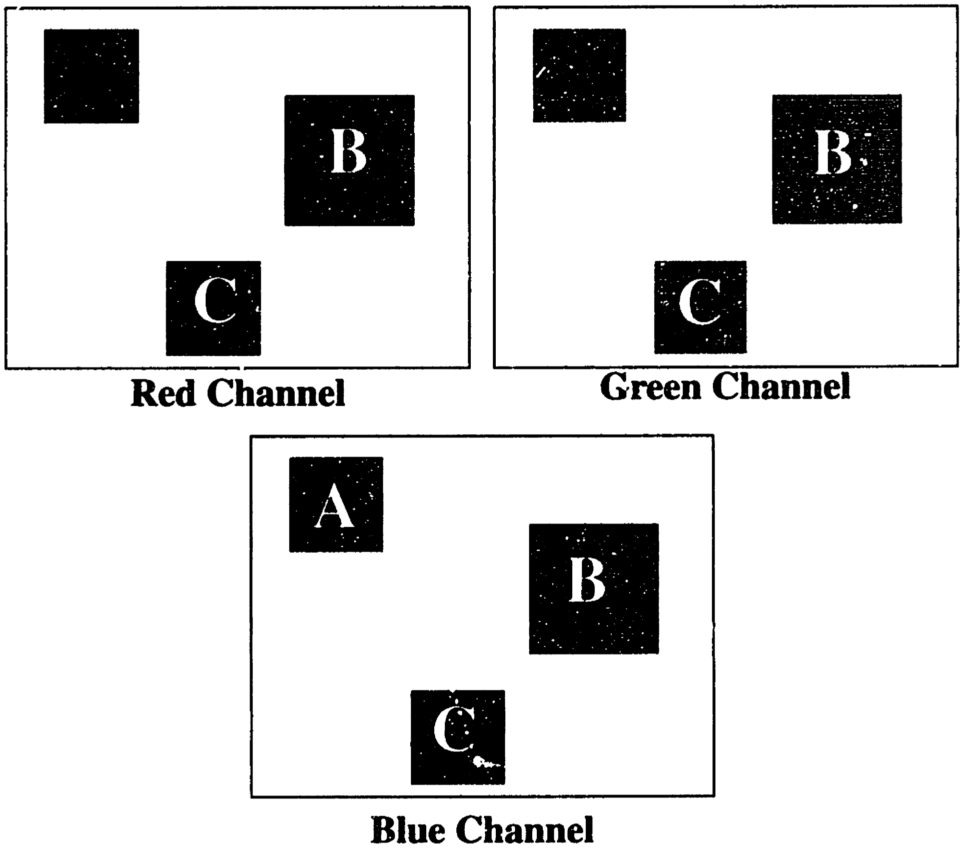
The three pairwise luminance relationships between the regions are the following:

$$A_l > B_l$$

$$B_l > C_l$$

$$A_l > C_l$$

We can describe these relationships in a language that is more perceptually meaningful where “greater than” corresponds to “brighter than” and “less than” corresponds to “darker than”. In this language, A is brighter than B, B is brighter than C, and A is brighter than C.



Red Channel

Green Channel

Blue Channel

FIGURE 3.9. *The red, green, and blue components of Figure 3.6 shown separately.*

Figure 3.9 shows the red, green, and blue components of Figure 3.6 separately. Brighter colors correspond to greater values in either R,G, or B. The (R, G, B) tristimulus color components of A, B, and C in Figure 3.6 are respectively (254, 200, 100), (75, 80, 200), and (50, 80, 60).

The color component of a patch is denoted in the following manner, e. g. for patch A as A_r or A_b or A_g. We can compute relative patch colors by separately computing the relative values in each color band.

The 9 pairwise color relationships between the regions are the following

In the red channel:

$$A_r > B_r$$

$$B_r > C_r$$

$$A_r > C_r$$

In the green channel:

$$A_g > B_g$$

$$B_g = C_g$$

$$A_g > C_g$$

In the blue channel:

$$A_b < B_b$$

$$B_b > C_b$$

$$A_b > C_b$$

Cross channel relative relationships within a patch are a way of encoding that patch's general color. For instance patch B is perceptually blue. This is reflected in the cross channel relationships. $B_b > B_g$ and $B_b > B_r$. Cross channel relative relationships *between different* patches may also be computed.

Relative intra-pixel cross-channel color measurements have been used in a system by Fishler to classify each pixel of an image into a set of predefined categories [19]. This pixel classification algorithm is part of a larger body of work to develop natural scene interpretation algorithms for autonomous robots. With respect to its color, a pixel is classified to be one of "water/rock", "cloud/snow/sky", "ground", "live vegetation", or "shadow-unknown" based on the absolute and relative relationships of the red, green, and blue color components of that pixel. The cross-channel color relationships are used effectively to quantize the color space into bins which represent these categories. The pixel classification system is hard coded. For instance, a pixel is categorized as ground if its red component is greater than its green component, its green component is greater than its blue component, and the ratio of its blue component to the sum of its red, green, and blue components is greater than 0.27.

3.4.3 Qualitative size relationships

Patches may have different sizes. For instance, in Figure 3.6, regions A and C are 32*x*32 pixels. Region B, on the other hand, is 45x45 pixels. We can, thus, encode that region B is greater in overall size than regions A and C. Region size is an easy way to emphasize the difference between fine and gross details in a model without performing texture calculations. For instance, in a snow capped mountain class, the white of the snow may consist of only a small portion in each of the images in that class. This is in contrast to the blue of the sky and grey of the mountain which may dominate most of the images.

3.4.4 Other qualitative relationships

There are many other relative region properties that can be used in the description of the model. Examples include relative patch shape and texture. Figure 3.10 illustrates some these properties. For instance region B is rounder than region A. Region C is more elongated than region B. The texture in region C is more dense than the textures in regions A and B. The texture in region A is more horizontally oriented than the texture in region B.

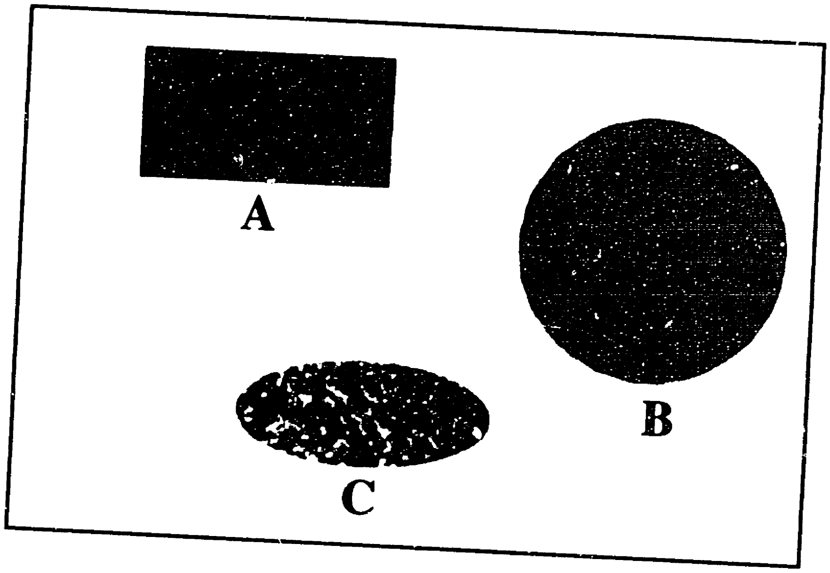


FIGURE 3.10. *Image with three salient patches that can be described by their relative shape, orientation, and density in addition to the attributes of relative spatial position, color, luminance, and size.*

3.5 Example of a qualitative model

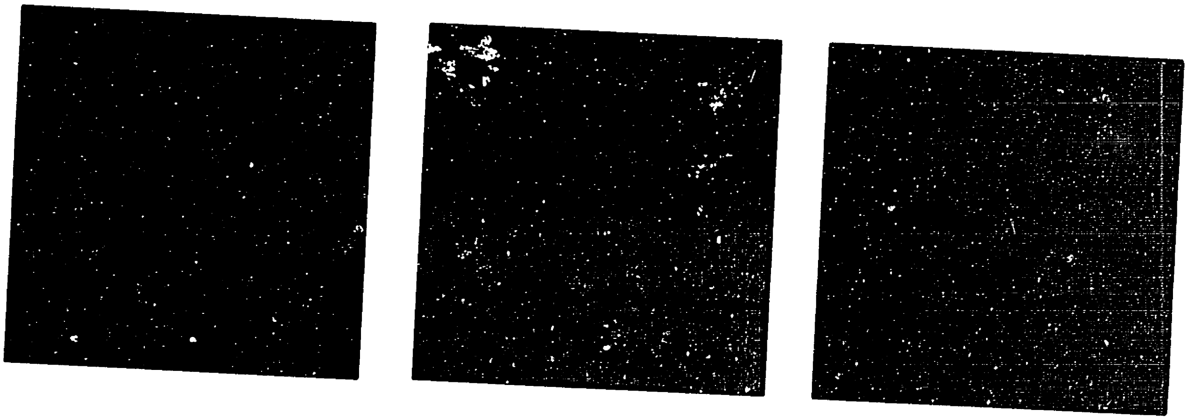


FIGURE 3.11. *Three synthetic field scenes*

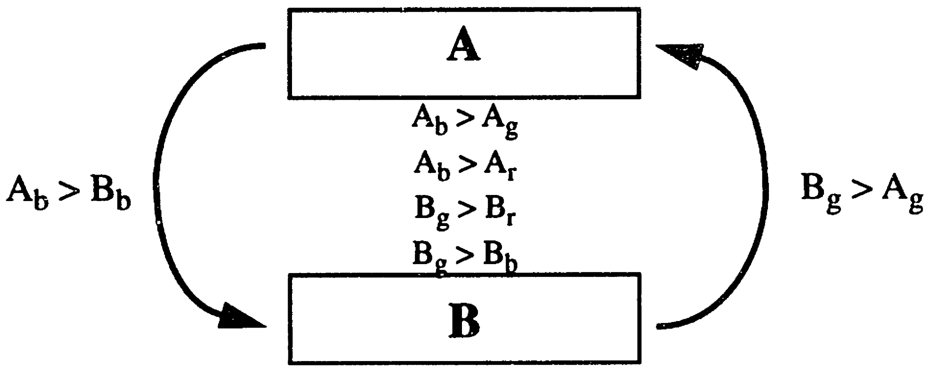


FIGURE 3.12. ***A model which captures the commonalities between the synthetic field scenes in Figure 3.11.***

Figure 3.11 shows three synthetic field scenes. A model which encompasses the similarities of scenes is shown in Figure 3.12. All of the scenes contain at least one pair of regions where a first region (A) is more blue, less green, and above a second region (B). In addition, to the relative relationships between the patches, there are some intrapatch relationships associated with A and B. The first region's blue component should be greater than both of its red and green components. The second region's green component should be greater than both of its red and blue components. This qualitative model captures the concept of a relatively blue "sky" over a relatively green "field", irrespective of the extents and absolute colors of the sky and field.

3.6 The role of quantitative information

Throughout this chapter, we have discussed the nature of qualitative models for scene classification. One important question is whether there is scope for any quantitative information in the scene models. Quantitative information may indeed be important in defining a particular scene class. Therefore, our qualitative models should have the ability to also encompass some quantitative information. For instance, it may be important to a scene class that there be a yellow stream of light in the exact middle of the image, that the sky region be a particular color of vibrant blue, or that the brightness difference between two regions be greater than some amount. We may want to incorporate information that is not expressible in terms of the attributes defined in this chapter. For instance, we may want to incorporate in the model that there should be a sign in the image with the word "restaurant" on it. In most cases, the quantitative information acts to restrict the size of the class described by the model.

3.7 Qualitative model as a directed graph

The qualitative model shown in Figure 3.12 looks very similar to a directed graph. In fact, we can think of both the model and the portioned image as graph structures. The patches represent the nodes in the graph while the pairwise relative relationships between the patches represent the edges in the graph. The pairwise relative relationships are explicit in the model and implicit in the image. The act of checking whether the image satisfies the constraints of the model, or classifying the image, is similar to trying to find a subgraph in the partitioned image that matches the model.

A directed graph G is defined as a collection of vertices V and a set of directed edges E, which connect those vertices. The direction of the edges specifies how one can traverse the graph, i.e. to go from vertex v_i to vertex v_j.

The qualitative model and the image can be represented with a different directed graphs, where a is the number of attributes used to describe the model and the image. For instance in Figure 3.12, the attributes used to describe the class model are vertical spatial position and the R,G, and B color components. The direction of the edges in each graph denotes the relative relationship between the image regions or graph vertices. Let us define that there is a directed edge in the direction from v_i to v_j, if with respect to attribute a, $v_i < v_j$. If, with respect to attribute a, $v_i = v_j$, then there is an undirected edge from v_i to v_j. If in the model graph there is no relation between v_i and v_j, then no edge is encoded between the two. This last condition is meaningful only for the model graph, as there may not be a relevant or consistent relationship between two model regions over one or more attributes.

Figure 3.13 and Figure 3.14 show respectively the model from Figure 3.12 and one of the images from Figure 3.11 expressed as four directed graphs (ignoring luminance and size). These graphs represent the interpatch relationships (the intrapatch relationships are not considered here). Note that in the model, there are no encoded relationships between the red components of the regions

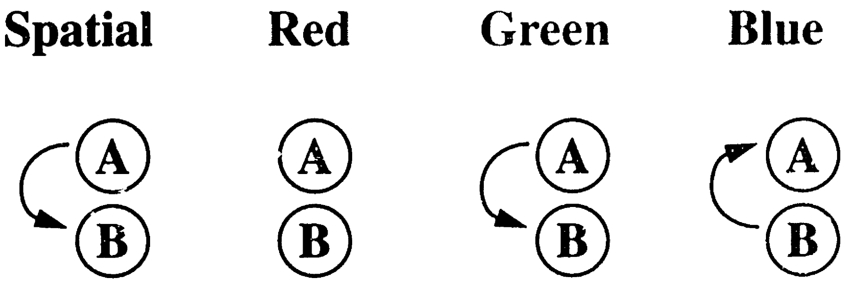
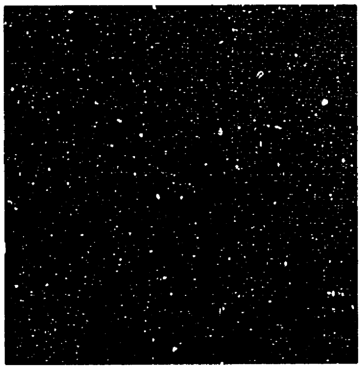


FIGURE 3.13. *The model in Figure 3.12 can be represented as four different graph structures, one for vertical spatial arrangement and one for each of three color channels. These graphs encode the relative relationships between the patches*



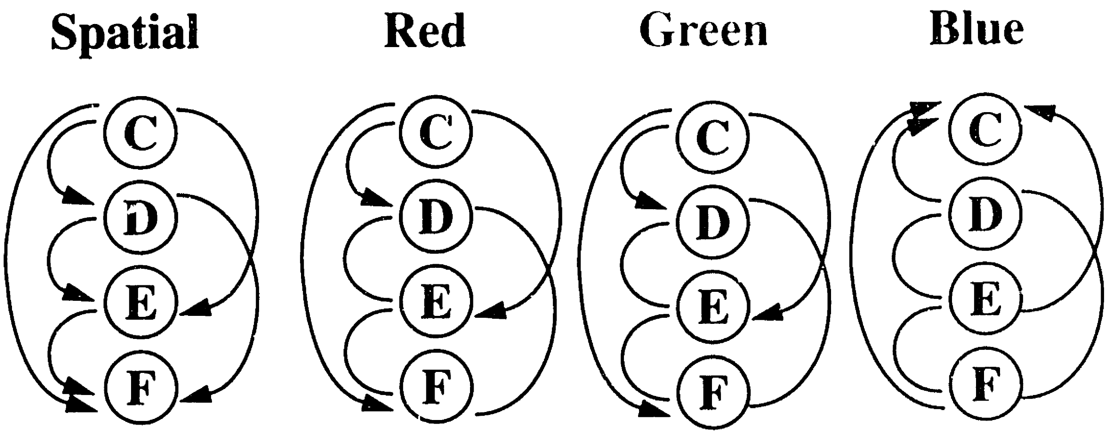


FIGURE 3.14. *The graph representation of one of the images from Figure 3.11 is shown here. The pairwise relationships between all the image patches have been computed and the values are represented as the direction of the edges.*

Using graph terminology, matching the model to the image is equivalent to computing whether the image graph contains a subgraph which is isomorphic to the model graph. Let I_a and M_a correspond to the image and model graphs for an attribute a. The formal definition of the graph isomorphism problem is as follows:

Given two graphs I_a=(V1, E1) and M_a=(V2, E2), does I_a contain a subset $V \subseteq V1$ and a subset $E \subseteq E1$ such that $|V| = |V2|$, $|E| = |E2|$, and there exists a one-to-one function f: $V2 \rightarrow V$ satisfying $\{u, v\} \in E2$ if and only if $\{f(u), f(v)\} \in E$? [20] We define that for a graph M_a that if E2 ={ }, as in the case of the red component graph in Figure 3.13, then all subgraphs of I_a, where $|V| = |V2|$ are isomorphic to M_a.

In the framework of our qualitative model, there are two conditions for M to match a subgraph in I:

1) There exists a vertex set $V \subseteq V1$ such that for all *a*, there must exist an image subgraph (V, E) in I_a that is isomorphic to graph M_a.

2) The intrapatch relationships of the vertices in M and corresponding subset of image vertices V must be the same.

For the model graphs in Figure 3.13 and the image graphs in Figure 3.14, the vertices of the subgraphs in the image that the match model graph are {C,D}, {C,E}, and {C,F}. The correspondence between the vertices in the model and the vertices in the image is that A matches C, and B matches D,E, or F.

The example shows that there may exist more than one subgraph in the image that matches the model. This gives us some important information regarding the grouping of regions. For instance if the vertex pairs {C,D}, {C,E}, and {C,F} match {A,B}, then we might group C into one region and D,E, and F into another region. Thus, the end result of the matching process may result in a perceptually consistent segmentation of the image. This segmentation may be used to recognize or label different regions of the image. A segmentation of two or more images would allow us to compute a registration between those images.

The example we have just described shows that there exists at least one perfect match between the model and the image. However, in the real applications, the model may not always exactly match part of the image. In such cases, we would want to find the best common subgraph in both the model and the image. The definition of "best" however is variable and may change according to the application. In some cases, the best subgraphs are the ones which contain the maximum number of image and model regions, where at least one of the relations over all the attributes are consistent. In other cases, it would be desirable to find the largest number of image and model regions, where all the relations over all the attributes are consistent. Some of the relations (or edges) may be more important than others in the model. We can represent this by putting weights on the model edges. Therefore, another definition of the best matching subgraph is the one which contains the edges with the highest total score. Figure 3.15 shows an example image which does not contain any subgraphs that match the model. Such an image may definitely be labeled as not part of the class described by the model.

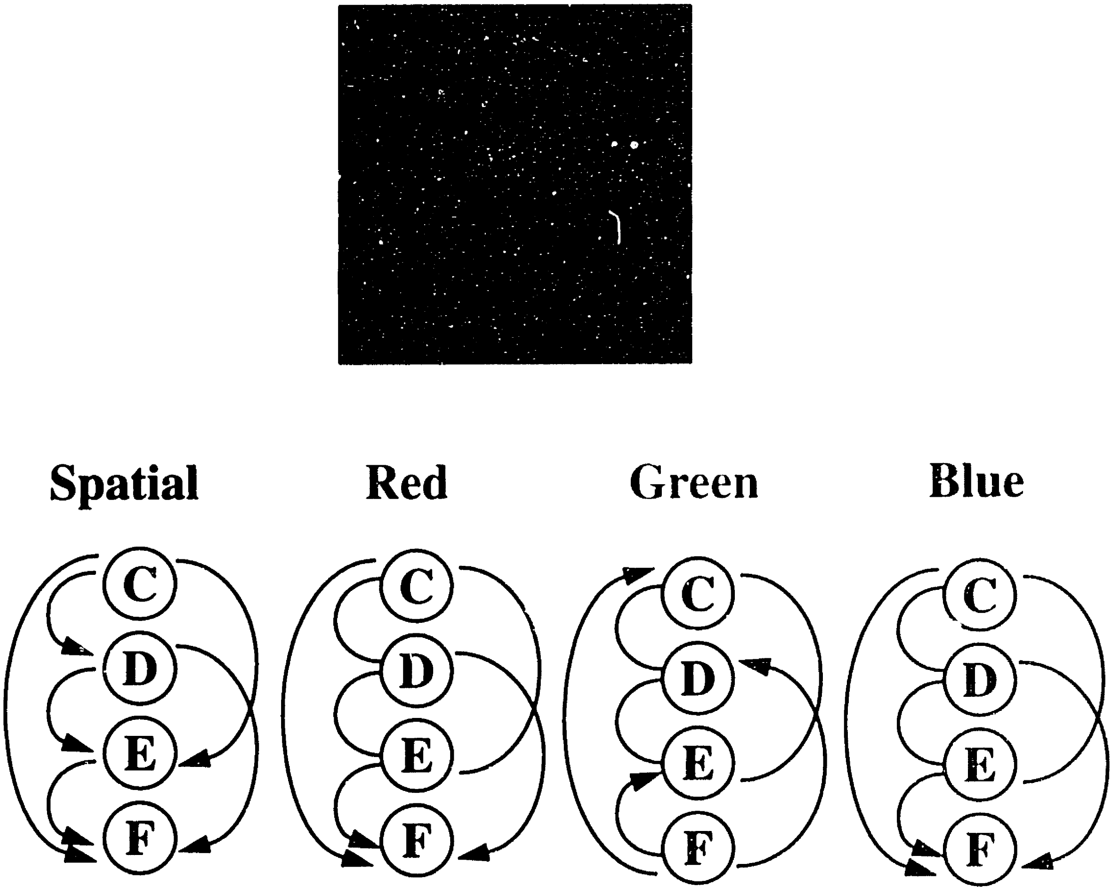


FIGURE 3.15. *Example of an image which does not match the model in Figure 3.12, and therefore would not be classified as part of the synthetic field class.*

3.7.1 Complexity of graph matching

The problem of computing subgraph isomorphism or the largest common subgraph is NP Complete (see [20] for a more detailed explanation). This means that to find a match between the model and the image, we would have to consider all possible one to one mappings of model regions to image regions. If the model contains m regions and the image contains i regions, where $m \le i$, the number of possible one to one mappings is $\binom{i}{m}m!$, which according to Stirling's approximation is exponential[16]. For instance, if there are 8 model regions and 64 image regions, then the number of pairings is $\frac{(64)!}{(64-8)!}$ = 1.7e+!4, which is approximately 64^8.

However, the model's qualitative relationships may provide some constraints which reduce the number of pairings. For example, not all model and image region pairings satisfy the relative spatial constraints encoded in the model. Let us assume because of the spatial constraints that each model patch can only match a bounded number of image regions. Let the number of image regions that can match each model patch equal $k \le i$. Using these constraints the number of matching image and model regions is k^m. Although this is still exponential, it is much less than $\binom{i}{m}m!$ for small values of k. Such a scenario is depicted in Figure 3.16, where each of 8 model regions can only match to a set of 4 independent image regions. The number of possible pairings is $4^8 = 65536$, which is a reduction from the previous number of model and image pairs on the order of 2.7e+9 times.

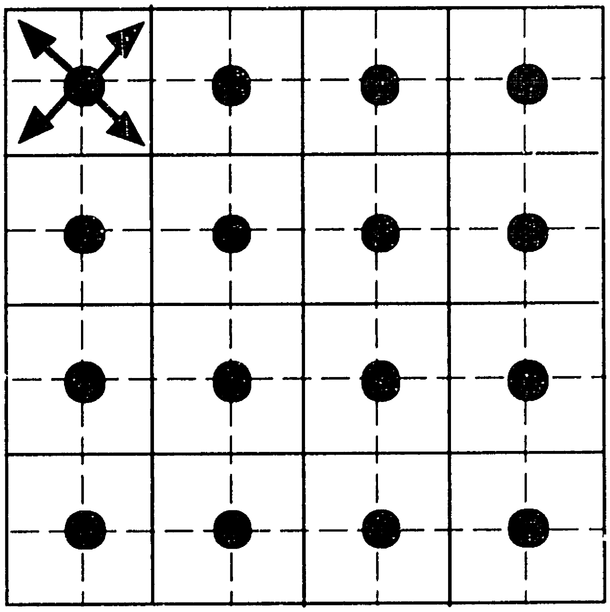


FIGURE 3.16. *Because of constraints from the model, the model regions may only validly match a subset of the of image regions. This figure shows a 64 pixel image where each model region, shown as circles, may only match within a local neighborhood 4 out of the 64 image regions.*

The values of 8 model regions and 64 image regions used in the previous calculations are not a drastic underestimate. We will show in the next chapter that only a few model regions can classify a large set of natural images. In addition, we will show that the match need only be performed on low resolution images which contain between 64 (8*x*8) and 1024 (32*x*32) image regions.

Verifying whether there is a match between the model and a sub-graph of the image is linear in the number of edges in the model and the number of attributes.

3.7.2 Probability of false positives

One potential argument against using qualitative models is that by discarding difference in magnitudes between regions, the model may be too general, therefore, resulting in many false positives. In this subsection, we analyze the probability of an image containing a false positive.

Let a model M contain a graphs, one for each attribute, where each graph contains the same v nodes and a different set and number of directed edges. For simplicity, let us assume that the number of edges e in each graph is the same. An edge has three possible values, two directed and one undirected (corresponding to the three types of ordinal relationships between two model regions). Thus, the number of model possible configurations is 3^{ea}. If the image graphs have the same topology as the model graphs, a first order approximation of the probability of randomly generating a specific model configuration from a uniform independent set a of attributes is $(1/3)^{ea}$.

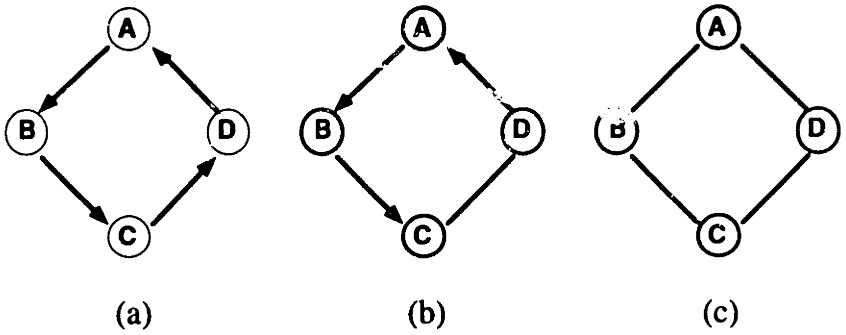


FIGURE 3.17. *All cycles expect for the one shown in (c) represent physically invalid graphs. (a) shows a invalid cycle with directional edges. (b) shows a non-valid cycle with 3 directional edges and one non-directional edge. (c) shows a valid cycle which consists of only non-directional edges.*

This first order approximation, however, is an lower bound on the probability of a false match. Some graphs are physically impossible. For instance, there cannot be any cycles in the graphs. Figure 3.17(a) depicts a cycle. If the edges represent luminance relationships, this suggests that node A is brighter than node B, which is brighter than node C, which in turn is brighter than node A. The resulting conclusion is that node A is brighter than itself, which is not possible. A cycle may contain a combination of uni-directional and non-directional edges. Figure 3.17(b) shows an invalid cy:le which contains 1 non-directional edge which still suggests that A is brighter than itself. The only cycle that is valid is one that is made of all non-directional edges as shown in Figure 3.17(c) (This validly suggests that A is as bright as itself.) If there are *c* edges in a cycle, then there are $2(2^c - 1)$ possible cycles with that number of edges. For a model which contains a graphs of e edges, if a cycle occurs in any one of the graphs the model is not physically realizable. Thus, the number of impossible configurations due to cycles are

$$\sum_{j=1}^{e} \langle 2(2^j - 1) \rangle^a - 1$$

Therefore, the number of possible models without cycles is

$$3^{ea} - \sum_{j=1}^{e} \langle 2(2^j - 1) \rangle^a - 1$$

There are some other invalid configurations. For instance, the attributes are not always independent of each other. For instance, luminance is a function of color. Therefore, there cannot exist a luminance relationship where A is brighter than B if the weighted average of the color components of B is greater the weighted average of the color components of A.

However, if the attributes are independent, for an image with the same topology of the model, the probability that the image matches the model is

$$1/\langle 3^{ea} - \langle \sum_{j=1}^{e} \langle 2(2^j - 1) \rangle^a - 1 \rangle \rangle$$

In general the image will contain more vertices and edges than the model. The image implicitly describes a fully connected graphs of $v1$ nodes and $e1$ edges. If the model contains v nodes, the number of mappings of model nodes to image nodes is $\binom{v1}{v} v!$. We can describe this in terms of edges. Because the image graph is a complete graph $v1 = \sqrt{e1}$. A model with v nodes can contain at most e edges. An upper bound on the probability that a model configuration exists in a randomly created larger image is:

$$\left(\binom{\sqrt{el}}{e}e!\right)/\langle 3^{ea}-\langle \sum_{j=1}^{e} \langle 2(2^{j}-1)\rangle^{a}-1\rangle\rangle$$

which is equivalent to:

$$((\sqrt{el})!/(\sqrt{el}-e)!)/\langle 3^{ea}-\langle \sum_{j=1}^{e} \langle 2(2^{j}-1)\rangle^{a}-1\rangle\rangle$$

For 6 attributes, 8 model edges, and 64 image edges, an upper bound of the probability of the model matching an image subgraph from a randomly generated image is approximately 2.2e-9. Table 1 shows this probability calculated for different combinations of model and image edges. The probability of a false positive is 1 when there are 2 model edges and 100 image edges. In general, the probability of a false positive under these conditions is quite low.

TABLE 1. Probability of a false positive over *e* model edges (columns) and *el* image edges (rows)

el \ e	1	2	3	4
1	0.001372	0.000000	0.000000	0.000000
4	0.001372	0.000002	0.000000	0.000000
9	0.001372	0.000006	0.000000	0.000000
16	0.001372	0.000023	0.000000	0.000000
25	0.001372	0.000113	0.000000	0.000000
36	0.001372	0.000677	0.000000	0.000000
49	0.001372	0.004742	0.000000	0.000000
64	0.001372	0.037935	0.000000	0.000000
81	0.001372	0.341411	0.000000	0.000000
100	0.001372	3.414114	0.000001	0.000000

3.8 Qualitative model as a deformable template

There is an alternative way of considering the qualitative model and the matching process between the model and the image. We can think of the model as initially having a fixed spatial arrangement. This fixed spatial arrangement may represent the most likely or average position of the model regions in images classified by the model. (There is some evidence that humans use averages of examples to create class models [68].) When the model is compared to the image, the model can be deformed by moving the patches around so that the model best matches the image in terms of relative luminance and photometric attributes without violating the encoded relative spatial arrangements. The model in this sense is acting as a deformable template. The image regions themselves may grow or shrink in order to better fit to the image. The resulting position and size of the model regions provides some information on how the image may be segmented into perceptually meaningful regions and how regions between two different images may be registered.

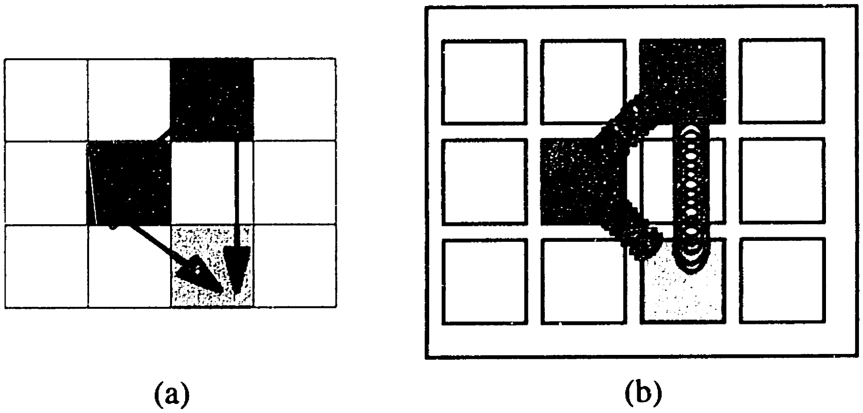
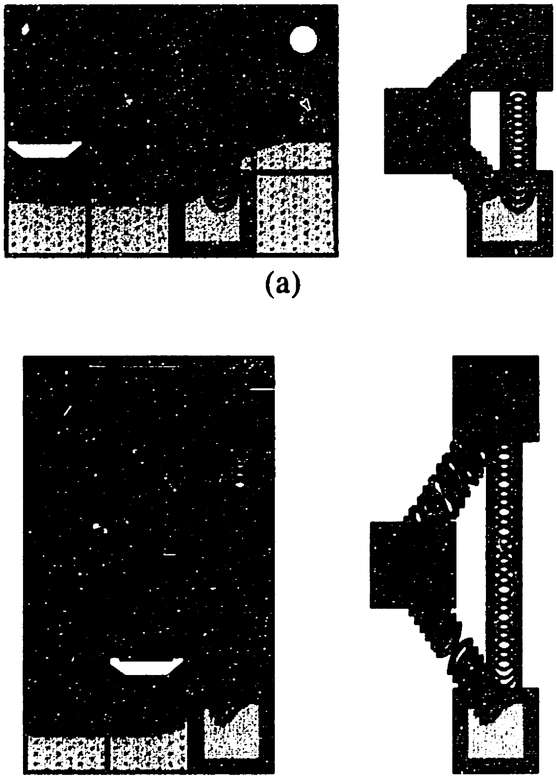


FIGURE 3.18. *Example of how a qualitative model can be formulated as a deformable template. The arrows in (a) denote the relative relationships between the image patches. In figure (b) springs replace the arrows to illustrate that the model can be deformed to match the image. The deformation is limited by the relative spatial relationships and any quantitative information regarding the positioning of the patches. A measure of goodness of fit is a combination how well the other relative relationships such as color and luminance are satisfied and how little the model needed to be deformed to detect that match.*

Figure 3.18(a) shows a model which contains three salient image regions. The lines between the image regions denote the relative region relationships, including relative luminance and color. We can replace the lines with springs to show that the model can be deformed by moving the patches around, as shown in Figure 3.18(b). A match between the model and a subset of the image can be defined by how well the deformed model matches the image subset and how little deformation was required to find that match.

Deformable models have been used for other computer vision tasks such as object recognition and image segmentation. However, the formulation for these models are mostly as deformable contours or snakes which try to maximize their fit to the boundary of an object or a region [35]. Yuille *et al* utilized deformable outlines of mouths and eyes to recognize to extract these features from images [71]. Kapur *et al* used snakes as part of a solution to cleanly segment brain tissue from MRI scans [34]. The idea of a deformable model that utilizes image regions and relative relationships between those regions differs greatly from the previous work.



(b)

FIGURE 3.19. ***Detecting both a distal and close-up view of a beach by deforming the qualitative model.***

Figure 3.19(a) and (b) shows how our formulation of a deformable qualitative model may be used to classify both a distal view and close up of the beach scene. The deformed model is overlaid on the images and is also shown beside the images. In both cases, the relative spatial, color, and luminance relationships of the selected image regions satisfy the constraints of the qualitative model. The amount of deformation from the original model (shown in Figure 3.18(b)) reflects the change in the viewing position of the observer.

3.9 Summary

In this chapter, we have described many aspects of qualitative models. The main points we wish to highlight are:

- Models which encode qualitative relationships between image regions may be effectively used to describe scene content.
- We can encode many types of qualitative comparisons in a model, such as relative spatial position, luminance, color, and size.
- Some quantitative information can be included in a model.
- Information in the model can be weighted according to its importance.
- Finding all instances of the model in the image provides information for how the image may be segmented in perceptually meaningful regions.
- Detection of the model in two or more images provides information for image to image registration.
- Computing whether a match exists between a model and an image is computationally tractable.
- The probability of a false positive in a randomly generated image whose attributes have uniform distributions is low.
- Qualitative models can be thought of as a novel type of deformable template.

In the next chapter, we describe a system for using qualitative models for scene classification

CHAPTER 4 Natural Scene Classification

4.1 Introduction to the problem

In this chapter and the next, we describe a system which is an implementation of the ideas described in the previous chapters for the specific task of natural scene classification. The work is motivated in part by the steadily increasing need for image indexing systems which could annotate and retrieve images based on their content.

The system is intended to be capable of discriminating between different classes of natural scenes. The goal of the system is to develop models which can concisely represent scene classes and can be used to retrieve other members of the classes from the image database. The user may then post queries to the system in the form "find all coastal images", where a coastal model has been predefined, or given a set of example images, return other pictures that belong to the class described by the examples. Users may also refine the models based on the results of the queries. In this chapter, we discuss the nature of the class models and present some examples of their generalization and discrimination ability. In the next chapter, we suggest strategies for automatically generating such models from a set of examples.

Although the system can 'tolerate' man-made objects in natural scenes, its current implementation, as specified in this chapter and the next, has not been specifically designed to classify other types of stimuli such as artificial objects, man-made scenes, indoor views, or images defined by texture patterns. In chapter 7, we suggest how some of the techniques employed to classify natural images may be used to classify or recognize individual objects.

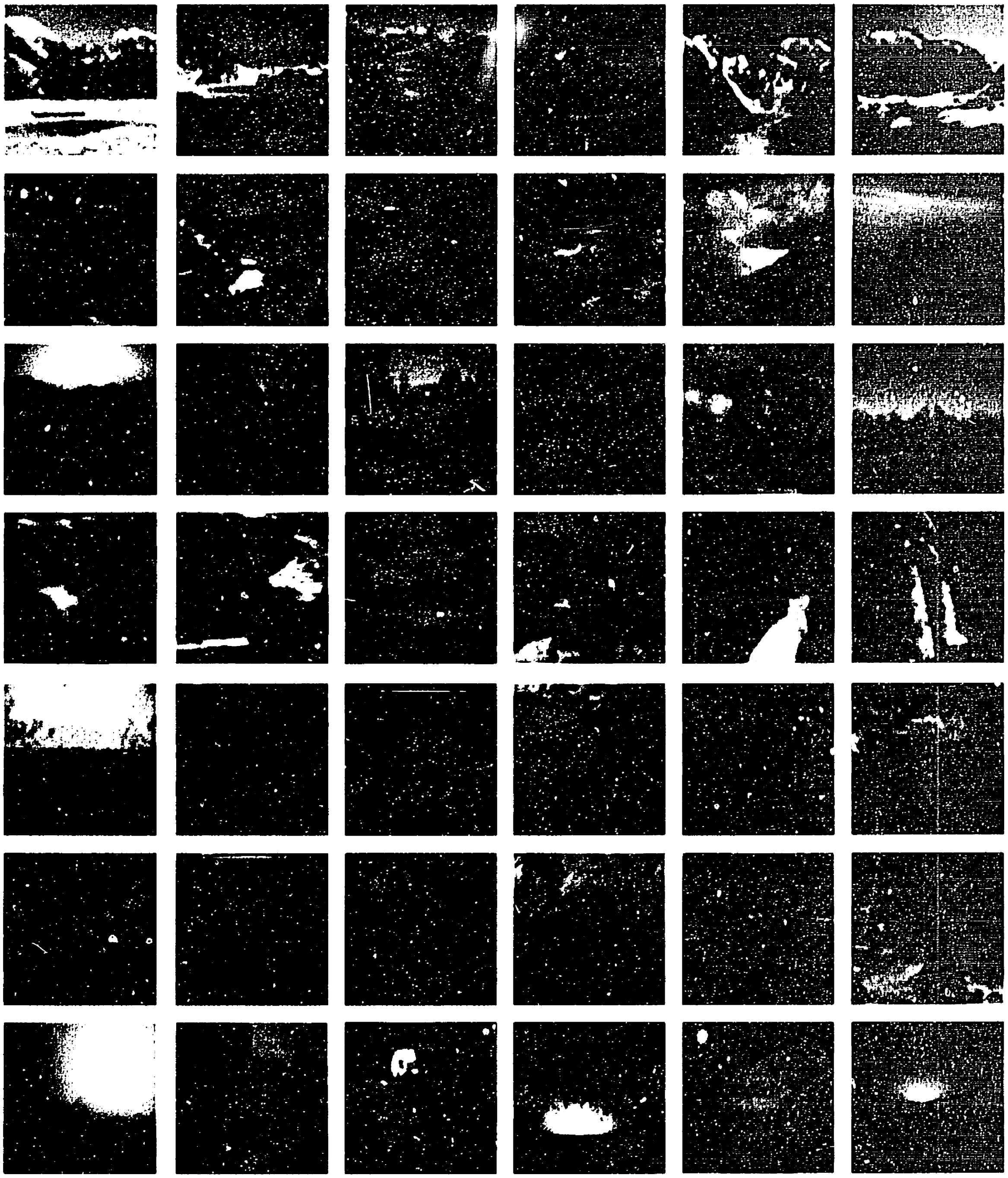


FIGURE 4.1. *Examples of natural images.*

4.2 Prior work on scene classification

Most of the approaches for scene analysis developed so far, classify or identify scenes based on a collection of statistical or local properties, including color, texture, and some shape information. The goal of a majority of color and texture based techniques, and a few shape based algorithms, is to perform scene classification based on non-semantic image attributes. The general philosophy underlying examples of this type of approach is that a combination of these image statistics produces a good measure of image similarity and, therefore, indirectly captures the image's semantic meaning. In contrast, the goal of other shape analysis techniques and a few color/texture segmentation algorithms is to delineate the image into meaningful objects or parts of objects, such as cups, buildings, sky regions, and trees. The reasoning behind this type of approach is that the classification of a scene can be derived from the semantic classification of its parts.

The majority of the scene analysis work is directed toward the applied problem of image database indexing. Queries to such an indexing system include a real picture, a user annotated image, or a user sketched image. These systems attempt to retrieve the images in the database that are most similar to the input image. Most of these systems are optimized for efficient image retrieval from a database by representing both the query image and the images in the database as feature vectors. Image retrieval is, therefore, reduced to comparing vectors. In this section, we describe a few distinct approaches to scene classification or image database indexing. We have classified the approaches as using either a combination of cues, using the cues in isolation, such as texture, shape, and color, and also matching of templates.

QBIC (Query By Image Content), developed at IBM, was one of the first such commercial database indexing systems to use the image itself as a query instead of a text string [4]. This system, which has evolved considerably over the past several years, uses a combination of many "low-level" cues to describe an image and is highly optimized in terms of speed of retrieving images from a database based on a visual query. The system uses color measurements in the form of a three element vector and a color histogram. Image texture is described in terms of coarseness, contrast, and directionality. Some shape features such as moments and parametric curves are also used. Properties of the images in the database are precomputed and stored as a vector. A query involves comparing features of the new image with features of the stored images using a vector match metric with relative weights for each of the properties. The strong points of the system are its use of simple cues and efficient indexing. The system is very good at retrieving images with similar color, textural, and shape distributions. The difficulty, however, is that these measures alone or in combination do not capture the meaning of the scene (see Figure 2.13).

There are several other systems that use the combination of cues approach. These include VisualSEEk [58][59], CANDID [36], JACOB[38], a commercial system from VIRAGE (see http://virage.com for more details). Some of these systems, including QBIC, have been extended to index into video databases using direction of motion and dramatic scene changes as dynamic cues.

In contrast to the combination of cues technique, Jacobs *et al.* use a multi-resolution wavelet approach to describing scene content [31]. Images are encoded as "signatures" of the highest *m* coefficients of the wavelet decomposition. This particular implementation uses the Haar basis set. Image similarity is based on the number of matching wavelet coefficients between an input image and a target image. This strategy is computationally attractive. The image "signatures" are small, therefore, the storage requirements, processing times, and matching times are low. The difficulty of this approach is that it is very sensitive to rotation, translation, non-uniform occlusion, and clutter in the input image [40]. The system seems most adapted for finding a particular image in a database which the user is able to roughly sketch.

A number of different systems use multiresolution wavelet or filter based approaches to indexing. Most of these systems use the outputs of the filters to derive shape from local areas of the image rather than a global image signature. Examples of such systems include [54] which takes in user delineated regions of interest from an image and processes these regions with derivatives of Gaussians at several different scales. The goal is to retrieve images with similar objects in roughly the same pose as the query image by matching the filtered patches in the query image to filtered patches in target images.

Picard and Minka use a combination of different texture models to describe the contents of images based on user labeled example image patches [50]. The system uses the texture model (or a combination of models) that best explains similarly labeled example patches. The system automatically propagates these labels on unclassified image patches. Scenes are then classified based on these annotations. Queries of the form "Give me all images with trees and grass", where trees and grass are defined as textural patterns, are also supported. This system suffers from two problems. The first is a class of natural objects such as trees will have varying textures and colors, for instance an oak tree in the fall vs. an evergreen tree. Secondly, there is no inherent concept of a scene class. The user must specify which objects belong to a class such as "city scenes".

There is a large body of work using shape cues to index images, used mostly to delineate objects in images or given an image of an object to retrieve pictures of similar objects. Photobook is an image database indexing system developed at the MIT Media Lab that can classify several different categories of images [48]. The overall goal of the system is to reduce the images to a small set of perceptually-significant coefficients. One of the classification modes uses the texture work of Picard and Minka. Another of the classification modes computes parameterized 2-D shapes of objects via eigenvectors. The eigenvectors encode how a 2-D shape has been deformed with respect to a base shape. Two shapes may be compared by looking at the amplitudes of their eigenvectors. The more similar the amplitudes, the more similar the shapes. The difficulty of this system is that there must be a sufficient number of training examples to build the parameterized model. In addition, objects in an image must have well defined unoccluded edges. Finally, the technique requires that 2-D query shape must be aligned or brought into correct correspondence with the base shape, which is a difficult problem in its own right.

A number of other systems use shape based cues to compare the similarity of two images. For instance, Gallant and Johnston extract oriented edges and compute angles between pairwise sets of edges [21]. Ang *et al* compute object shape attributes as region compactness, boundary eccentricity, region moment, and region convexity [3].

There are several proposals for using color histograms as either image or object signatures [61][62][64]. This body of work focuses mainly on how histograms may be matched. For instance, they may be matched by using a sum of squares difference, comparing the dominant features, and by incremental intersection. The difficulty with these approaches is that the spatial relationships between subparts in an image or subparts in an object are not preserved.

In contrast, Equitz [18] uses a low resolution template approach to classifying images which uses the absolute positions of the image elements. Equitz partitions the query image into rectangles, usually by division of the width and height of the image into 8 sections each, and computes the average color of each rectangle. The resulting grid of colors is then matched to processed images in the database. Color similarity is computed as a nonlinear function of distance in LAB color space. Rectangles from the query image may seek out the best match in a limited neighborhood of the target image. Image similarity is, thus, the sum of the color similarities for each rectangle. As shown in Figure 2.12, however, the images retrieved by this type of approach are very similar to the input image in terms of spatial layout and color.

In the next section, we describe our approach entitled "configural recognition" that uses novel qualitative metrics of photometric and spatial similarities to classify natural scenes. This approach overcomes several of the problems faced by image classification systems reviewed here.

4.3 The "Configural Recognition" approach and its benefits

As mentioned in chapter 1, the goal of this thesis is to develop a computational strategy for scene classification. This entails two related questions: 1) How should scene concepts be represented? and 2) What is an appropriate metric of similarity between scenes? We describe here a novel technique, entitled "configural recognition", that uses global organization with relative relationships over low-frequency photometric image regions to represent scenes. This representation also suggests computationally simple strategies for assessing similarity between different scene instances.

As described in chapters 2 and 3, a global organization of relative scene measures seems to encode much of an images's semantic or category information. By using this type of approach, we are able capture more salient descriptions of a class rather than merely its overall color, textural, and shape properties or descriptions based on local disconnected image features. In addition, the combination of the use of low frequency images and qualitative relationships between image patches solves some complexity problems inherent in scene classification.

In the introductory chapter, we suggested that the difficulty of scene classification can be broken down into three basic areas. We briefly recapitulate them here as a prelude to suggesting how the configural recognition approach ameliorates each of them.

• The first is the complexity of the problem. Images contain a wealth of information, where each pixel may be considered as an important piece contributing to the scene's classification. In addition, there exist a large variety of algorithms which produce different descriptions of images, such as texture measurements, statistical descriptions, geometry features, shading and reflectance information, and color information, depending on the technique. A combination of the amount of information and the great variety of ways to describe an image, leads to almost infinite possibilities for processing a potentially large amount of data.

• The second problem is evident from the description of the first difficulty. Given that we have all of these computational methods to describe an image, which if any, can capture the content or essence of a scene or scene class. The demands on the choice of representation are especially high with natural scene classification, because scenes which may be perceptually grouped into one class can differ greatly in their geometry, illumination, absolute colors, textures, and even content (for instance not all water scenes contain breaking waves). Scenes may also be viewed under different weather conditions, which can produce great image variabilities in scenes that belong to the same class. It is desirable to have a vocabulary of scene descriptors which is rich enough to describe may different types of classes, but is also easy to compute and represent.

• The third problem is one that is inherent in all computer recognition/classification problems. Given a model and an image, how may they be efficiently compared to determine if the image is an instance of the model. In the worst case, every pixel in the image may be compared to every basic element in the model.

In the next several sub-sections, we will detail the basics of the configural recognition approach to scene classification and also suggest how such an approach can be used to effectively classify scenes while solving many of the problems listed above.

4.3.1 The use of low frequency/low resolution information

Surprisingly, humans need little detailed information to recognize many objects and scenes. Figure 4.2 contains 12 reduced images ranging in source from man-made objects, natural scenes, animals, and human faces. From these thumbnail low resolution images, we can make coarse interpretations such as 'flower scene' to fine identifications of familiar faces such as 'Indiana Jones with a smirk'. Figure 4.3 illustrates that the only information retained in these small images is an arrangement of low frequency photometric regions. Informal observations of this sort suggest that we can conceivably base our classification algorithm on an image's low frequency information.

FIGURE 4.2. *Much of the work in recognition has been done using high-resolution images. However, several psychophysical studies have shown that low resolution images are sufficient for several recognition tasks. The images shown here are identifiable even though they contain only low-frequency information. For instance, the images in the top row are only 30x30 pixels each.*

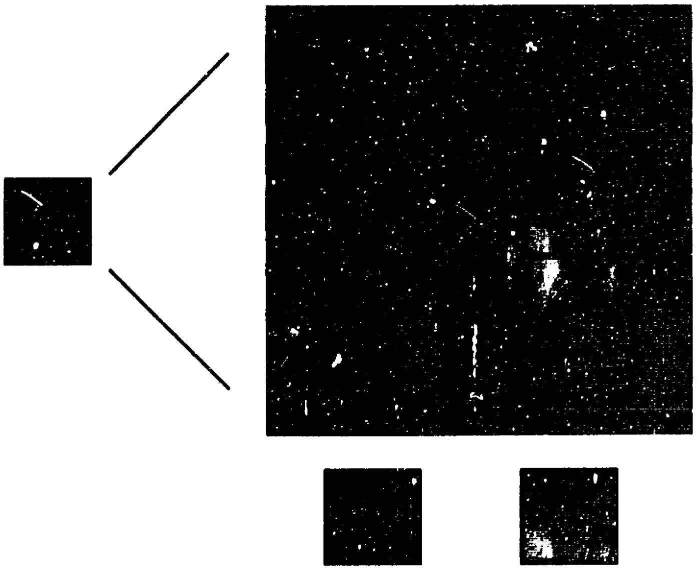


FIGURE 4.3. *This figure demonstrates that the only information retained in the low frequency subsampled images is the arrangement of color regions. The low resolution lion image has been enlarged and the color removed. The patches marked in red on the large greyscale image are shown side by side below the image for comparison. No discernible local texture regions or other local features are evident in this image.*

Figure 4.2 and Figure 4.3 have suggested that the lack of highly detailed information does not hamper the recognition of a large subset of scenes and objects. Figure 4.4 is a complimentary illustration which shows that the general organization of the scene or object may be perceptually evident even in the face of conflicting texture information or high resolution information. Figure 4.4 shows an outline of a fish. The fish is on a highly textured newspaper background. The interior texture of the fish consists of people and paraphernalia, such as umbrellas, in a beach setting. The image is easily classified as a fish even though its exterior and interior texture are possibly confusing. The fish is identifiable in full resolution, when the conflicting texture information can be resolved, and also in its low resolution version (shown in Figure 4.5). This demonstration suggests that high resolution texture information may, even when available, not participate in strongly in the recognition process.

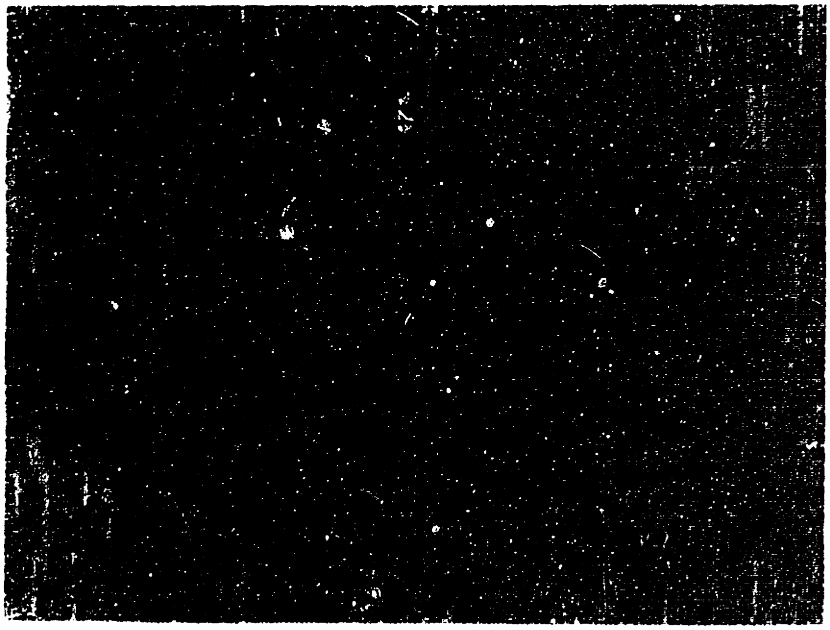


FIGURE 4.4. *A picture of a fish with a highly textured background and interior. The fish is identifiable even though the texture in the background (newspaper) and the texture in the interior (people at a beach) are possibly confusing.*

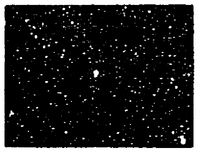


FIGURE 4.5. *In this low resolution version of the previous figure, the fish is identifiable even though the original interior and exterior textures cannot be resolved.*

There are several benefits of using low-frequency/low resolution images. The first is that the dimensionality of the problem is greatly reduced. The size of the images in Figure 4.2 are roughly thirty times smaller than their original counterparts. The dimensionality of the problem is thus lowered by almost an order of magnitude. Using the low-frequency image content also confers immunity to high-frequency sensor noise and leads to a shift in focus from potentially confusing detail to the use of relationships between image regions. Thus, the use of low frequency information is compatible with the idea that relative relationships between photometric image regions captures the perceptually salient content of a scene.

4.3.2 The use of qualitative relationships between image regions.

The configural recognition approach uses relative relationships between photometric regions from the low-frequency images as image and model primitives. There are several benefits of using low-frequency image regions and their relative relations as image and model primitives.

The first benefit is that no complex abstractions are performed on the low frequency pictures to partition the image into regions. The current implementation of the system subdivides an image into equally sized regions. The advantages of this type of approach are that complex time consuming image operations are not required. In addition, the possibility of incorrect region segmentations or identifications are eliminated.

The second benefit is that the use of relative relationships over these low frequency patches allows the system to describe class similarities even though the exemplars may differ in appearance due to various lighting conditions, viewing positions, and other scene parameters. Thus, while *absolute* color, luminance, and spatial position of the image regions, especially in the high frequency bands, cannot be used as reliable indicators of a scene's identity, their overall *rel-*

ative luminance, color and spatial positions can be expected to remain largely constant over these various conditions. In the base version of the system, three kinds of inter-region relationships are used in the scene model description. These are relative color, relative luminance, and relative spatial position of the regions. Figure 4.6 shows a synthetic example of a beach scene concept constructed of three image regions and their relative relationships. Figure 4.6(c)-(f) illustrate how these types of relative relationships remain valid over different but very commonplace image distortions. The beach concept includes a first patch which is bluer and above the other two, a second patch which is greener and below the first and above the third, and a third patch which is browner and below the first two. These relationships remain valid even though the beach images may undergo changes in scale, illumination, viewing parameters, and geometry.

Descriptions of scene classes using qualitative relationships based on low frequency information provide a rich vocabulary to differentiate between many classes of images. For example, consider a concept based on six attributes; relative color in terms of R, G, and B chromatic components, relative luminance, and relative spatial position in *x*-*y* plane. Each attribute may have one of three values; less than, greater than, or equal to. Assuming the relationships are independent, there are 3^6 possible relative relationships (see section 3.7.2 for a discussion of when the attributes are not independent). In a model of *m* image regions, there are $\binom{m}{2}$ possible pairwise relationships between those regions. Thus, there are $\langle 3^6 \rangle^{\binom{m}{2}}$ possible scene concepts which can be described by those *m* patches. Therefore, for a model of eight patches there are over 1.4e+80 possible concepts. An important point, which has already been addressed in chapter 3, is that not all relationships between all model regions are necessary to define a concept of a scene class. In both hand-crafted concepts and automatically learned concepts, only the salient or discriminatory relationships need to be encoded. Thus, although multiple patches provide a rich description, we will show that only a few patches are necessary to classify images within one perceptual group and to discriminate others not in that group.

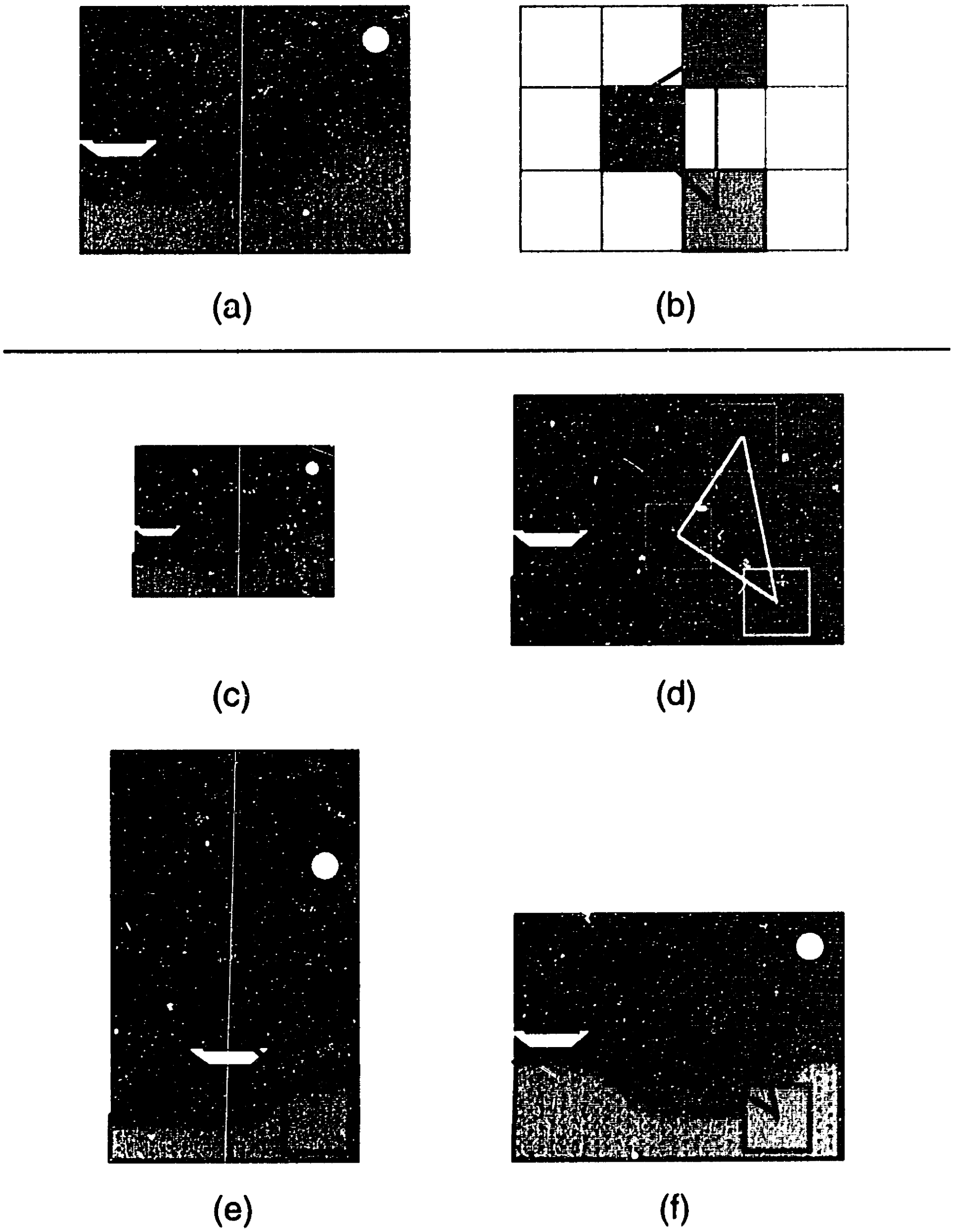


FIGURE 4.6. *(a) Example beach images. (b) Qualitative beach concept consisting of relative spatial and photometric relationships between three image regions (capturing the relationships between the sky, water, and sand). For instance the model includes a region that is more blue than two other regions below it. The model remains valid over many scene variations, including (c) scale changes, (d) illumination variations (the colors have changed but the relationships remain the same), (e) differing viewing parameters (distal vs. close up view), and (f) geometry changes (in this case a different coast line).*

4.3.3 Using model information to detect the concept in novel images.

A central theme of the configural recognition approach is that a scene should be characterized as a whole of parts, rather than as a collection of parts which comprise a whole. The model, which encodes the global organization of the parts, can be used to roughly guide the technique as to where to look for or where to expect each piece of the scene concept in the image. The technique, therefore, only considers image features that are consistent with the general framework of the model. The model driven approach greatly reduces the number of image features that need to detected and evaluated. Model-guided approaches have been used successfully in other applications to reduce the complexity of computation and to increase the robustness of model-based recognition and classification [32][42]. There is also psychophysical evidence that when humans are given information regarding what type of stimuli they will be presented with, their time for recognition or classification is more rapid than when they must interpret the visual stimuli without such *a priori* information [37].

4.4 Model descriptions

4.4.1 Qualitative information

In the current implementation of the system, there are seven types of qualitative relationships that are used to encode relationships between regions in the model. Each of these relationships can have the following values: less than, greater than, or equal to. The relative color between image regions is described in terms of their red, green and blue components. We may compute the relative relationships between each of the bands independently or include cross-band relationships, such as red to green comparisons between model patches. We can also encode relative luminance relationships. (Other relative chromatic relationships, such as relative hue and saturation, can be easily incorporated into the system.) The spatial relationships used are relative horizontal and vertical descriptions with respect to the upper left corner of the image and the cardinal axes. We also encode relative size. The size of a patch is described by how many square image regions it covers relative to the scale of the image. Three types of intra patch relative chromatic relationships can also be used in the model description. Thus, for one patch we can compute the relationship between its values for R and G, G and B, and B and R.

Let a model *m* consist of two regions A and B. Regions A and B can be described in terms of the *x* and *y* values of their centers. They can also be described in terms of their R, G, and B chromatic values. For each patch we can compute the luminance from the tri-stimulus color components. A and B can also be described in terms of their size. The possible comparisons between pairs of attributes are:

Relative Spatial Positions:

(A_x, B_x) (A_y, B_y)

Single Channel Color relationships:

(A_r, B_r) (A_g, B_g) (A_b, B_b)

Cross Channel Color relationships:

(A_r, B_g) (A_r, B_b)

(A_g, B_r) (A_g, B_b)

(A_b, B_r) (A_b, B_g)

Relative luminance:

(A_l, B_l)

Relative Size:

(A_s, B_s)

Intrapatch color relationships:

(A_r, A_g) (A_g, A_b) (A_b, A_r)

(B_r, B_g) (B_g, B_b) (B_b, B_r)

Together, these qualitative relationships provide a language to describe class models.

4.4.2 Quantitative information

We employed several types of quantitative information in our class models to refine the concept and reduce the possibility of false positives. The quantitative information used includes bounds on the difference between two measurements, bounds on the relative magnitude of two measurements, bounds on the absolute values of those measurements, and bounds on the sizes of the model patches with respect to the scale of the image. An example of this type of information is, for instance, that the red component of patch A should be greater than the red component of patch B by a difference of at least 10 units. The model can incorporate information, such as, the blue component of patch A should be 1.2 times greater than the blue component of patch B. We may specify that the luminance of patch A should be greater than a specific value such as 60 or that the *y* component of the center of patch A should exceed 2. In addition, we can specify that at a particular image scale, patch A should encompass two image regions. These constraints can be written more concisely as:

$$\mathbf{A_r - B_r > 10 \quad A_b > 1.2{*}\,B_b \quad A_l > 60 \quad A_y > 2 \quad A_s = 2}$$

Quantitative information can be used to limit the acceptable values of the relative relationships. Often, this is necessary to discriminate between changes in values which reflect a true difference between two distinct image regions versus insignificant changes which most likely come from within one or similar image regions. Cases where values change by one unit from one region to the next are probably not too significant, while changes of greater values or greater magnitudes are important to recognize. The limit as to what constitutes a "valid" change may be set as a parameter. This limit may differ depending on the type of relative attribute measured.

Quantitative information may also be used to encode a color, luminance, or spatial position for some of the model patches. For instance, it may be important for night scenes that many patches have the color black measured as r=0, g=0, and b=0.

4.5 Multiple templates and hierarchical scene concepts

We recognize that one template may not fully describe a scene class. Several templates may be needed to account for varied scene complexity of dramatically different viewing conditions within one class of natural images. Figure 4.7 shows three scenes which show a single tree or a pair of trees under different weather and lighting conditions. The color relationships between the tree, ground, and sky patches is quite different in each of the images. For instance, in the first picture the sky is more blue and less green than the trees. In the second picture, the sky is more white than the trees. In the third picture, the sky is more orange and lighter than the tree. To capture the concept of scenes with a few trees, it may be necessary to encode several templates such as those that describe "trees in summer", "trees in winter", and "trees at sunset".

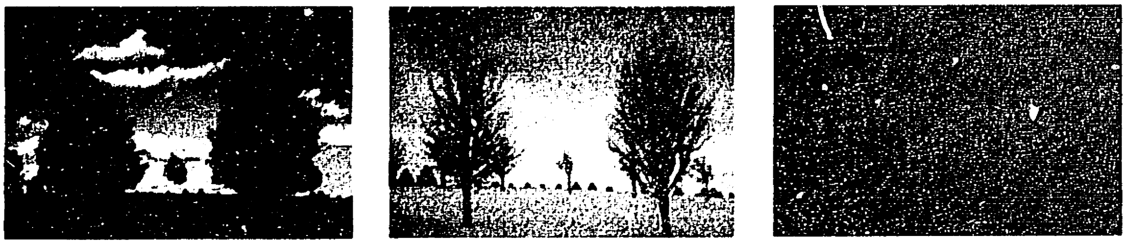


FIGURE 4.7. *Separate templates may be needed to describe a broad class of natural images. For instance, this figure shows three tree pictures under different conditions.; in summer, in winter, and at sunset. To describe the class of tree scenes, it may be necessary to encode templates for each of these variations.*

In addition, we may need to encode multiple templates which describe different levels of details of a class or different aspects of the class. Figure 4.8 shows three mountain scenes. One template which captures the common relationships between the sky, snow, and mountain regions may be used to categorize all these scenes as part of the "snowy mountain" class. However, more detailed templates may be needed to discriminate between these three scenes or to describe subclasses of the more general mountain class. For instance, the middle picture contains a snowy mountain with a lake. The picture on the right contains a snowy mountain with a lake and two sets of trees. Thus, we can create more specific templates which encode the relationships between the sky, snow, mountain *and lake regions* or a template which expresses the relationships between the sky, snow, mountain, *lake, and tree regions*.

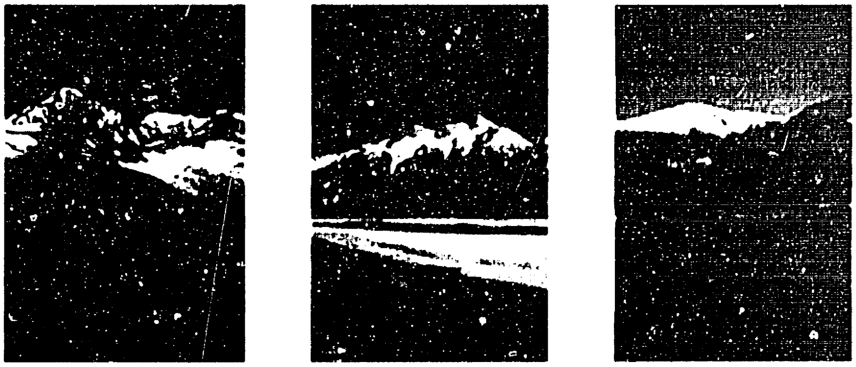


FIGURE 4.8. ***More specific templates can be generated to describe variations on a more general concept. Three snowy mountain images are shown. The middle image may be described as a snowy mountain with a lake. The right most image may be described as a snowy mountain with a lake and trees.***

The templates may be organized into a hierarchical or intertwined data structure that expresses the relationships between them. Figure 4.9 shows such a data structure based on the templates used to describe the images in Figure 4.7 and Figure 4.8. The templates are shown as boxes. Abstract labels, such as trees or lakes, are denoted as upper case text. Organizing templates and labels in this manner allows for a more comprehensive annotation system. For instance, if an image satisfies the "snowy mountain with lake and trees" template, it automatically inherits the labels of mountain, lake, and tree scenes. The data structure may also suggest a more efficient strategy to classify novel images. For instance, in a coarse to fine strategy more general templates may be applied before more specific ones. Based on the responses of the general templates, the system can be directed to try a subset of the more specific templates. Using the same example, if the image is a "snowy mountain with lakes and trees", it will satisfy the broad snowy mountain template. The template hierarchy implies that the more specific mountain templates should be applied, while other templates under the trees and lakes categories may be ignored.

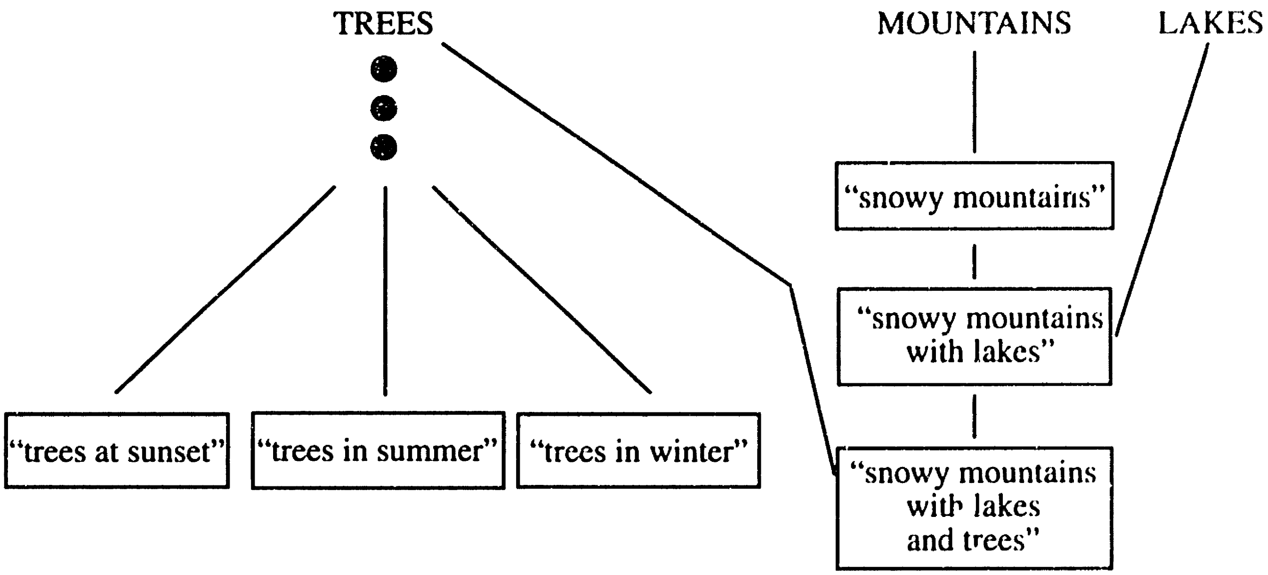


FIGURE 4.9. *Hierarchical encoding of templates and labels. The templates are denoted by the boxes. The abstrac: labels are shown as upper case text.*

4.6 The model to image matching process

The current implementation of the model to image matching process is very straightforward. All spatially valid configurations of model regions are evaluated. If the model is thought of as a flexible template, the matching process essentially stretches and compresses each spring like connection between each of the model patches until all valid deformations of the model are evaluated.

For some classes of images, qualitative relationships may only be important across the vertical or horizontal axis of the image. The templates described in section 4.8.3 utilize this information to decrease the complexity of the matching problem. Thus, for a model of p patches and an image of dimensions *mxm*, the complexity of the matching process is reduced from m^{2p} to $m^{(p+1)}$, which is a significant reduction. However, because we are already using such low-frequency images (ranging in size from *32x32* to *8x8*), this optimization is not critical for the matching process.

It is possible to further optimize the matching process by scanning the image for a first model region pair that satisfy the constraints of the model. This reduces the number of possible mappings of the model to the image. The number can be further reduced by iteratively adding in the constraints from additional model regions. The model regions used first should be the most significant to the description of the class. The regions added in at later stages may not be as important to the class concept. This strategy eliminates the need to exhaustively enumerate and test all possible model to image mappings and is similar in spirit to pruning the potentially large interpretation tree [23].

For each valid spatial configuration of the model, we compute a measure of how well the corresponding image patches and their relative relationships match the model. Currently, the match metric reflects how many model interpatch and intrapatch relationships are satisfied by the image. The image is classified as an instance of the model, if the number of interpatch and intrapatch relationships are greater than an *a priori* set of thresholds. The thresholds may be different for each relationship, based on how well we expect the relationships to be satisfied in a real image. For instance, in the class of snowy mountain images, the snow component is often scattered across the mountain tops, rather than providing a full covering. The sky and bare mountain regions, however, are usually visible across the image and have somewhat uniform colors. Assuming that the model is correctly positioned over the sky, snow, and mountain regions of the image, the sky to mountain relationship should hold across the image, while the sky to snow and snow to mountain relationships may occur with only a 50% frequency across the image. In the current implementation of the system, if one configuration of the model matches the image, then the image is categorized as a member of the model class.

This measure of fit, however, does not take into account several other important attributes. For instance, it is possible to add in a quantitative color component to the measure. In addition, the relationships may be weighted according to how important they are to the concept. We also may increase the measure of fit by how many deformations of the model match the image. It is also possible to decrease the measure according to how much the model needed to be warped to find a match. The percentage of the image covered by the warped model may also be added to the measure of fit function.

4.7 *Refining the scene class model*

The scene class model may be tailored to fit the preferences of a particular user. In the first iteration, the pre-existing model is used to classify a portion of the database. These images are returned to a user, who may rate them. Based on this rating, the weight of the existing qualitative relationships between the patches may be altered and new relationships may be extracted. Refinement of the scene class may be used to make the class more general, to narrow the scope of the class, or to encode in the class other important salient relationships. We discuss the refinement process in more detail in the next chapter on learning class templates from a set of exampl· images.

4.8 *Testing the approach*

The configural recognition approach to scene classification was tested by generating four class templates and by using these models to classify a large database of natural images. In this section, we describe the content of the image database, how the images were processed to extract the low-frequency, low resolution content, the details of the class templates, and the results of applying the templates to the images. For each template, the automated classification was reported as a binary decision of either a member or non-member of the class. We compared the results of the template classification to perceptual class judgments. The possible perceptual ratings of each image with respect to a particular class consisted of "a member of the class" or "not a member of the class". We compared the output of the templates with the perceptual judgments of the observers. For each template, we report the "true positives", "false positives", "false negatives". The experiments were run using a C program on a Pentium P.C. with a linux operating system.

4.8.1 *The test database*

The test database consists of 700 images from prepackaged CD-Rom collections which contained 100 images each. Each 100 image collection consists of images that the vendor, Corel, classified into one theme. The titles of the CD-Roms were "Fields", "Sunsets and Sunrises", "Glaciers and Mountains", "Coasts", "California Coasts", "Waterfalls", and "Lakes and Rivers".

The total collection contains pictures which have a wide range of content, colors, textures. The pictures have been taken from a variety of viewing positions (i.e. close-up vs. panorama) and under many different types of viewing conditions. Although the images in these compilations were mostly of natural images, many contain people, animals, and man-made structures such as fences, houses, and boats. The images in these collections contain a variety of sub-classes of images. For instance, the "California Coasts" compilation contains scenes of lighthouses, animals, bridges, and sand formations, the "Lakes and River" collection encompasses mountains with lakes, trees with lakes, ice images, waterfalls, and even a person fishing, the "Waterfall" CD has waterfalls through barren rocks and also through dense vegetation, the "Field" images consist of fields of many different types of grass, flowers, and trees. The collections also contain overlapping classes, for instance both "California Coasts" and "Sunsets and Sunrises" contain sunset images and both the "Glaciers and Mountains" and "Lakes and Rivers" contain snowy mountain images.

4.8.2 Processing the database

Each image in the database was iteratively smoothed and subsampled to create a pyramid of low resolution images (see [2] for more details on image pyramids in image processing). Each pyramid consists of three images of sizes respectively 32x32, 16x16, and 8x8 pixels. Figure 4.10 shows the result of the pyramid generation process on one image. In this figure, the computed low-resolution images have been scaled to the same size to illustrate the decreasing frequency content as a result of each iteration. Pixels in the low resolution images represent averages of larger regions in the higher frequency images.

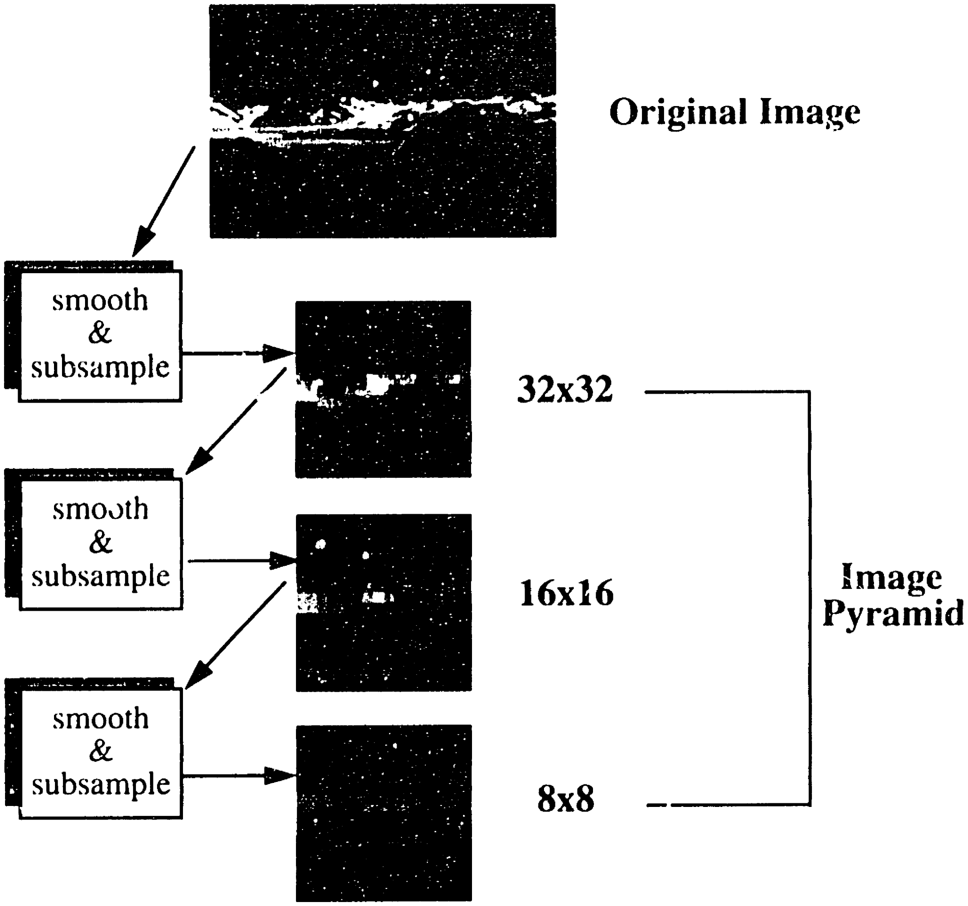


FIGURE 4.10. *Creation of a low-resolution image pyramid. The original image is iteratively smoothed and subsampled to create three images of sizes 32x32, 16x16, and 8x8. The three images in this illustration have been scaled to the same size to show the decrease in the level of detail from the first image in the pyramid to the last.*

4.8.3 Construction of qualitative templates for several scene classes

We constructed class templates for snowy mountains, snowy mountains with lakes, fields, and waterfalls. For each class model, we describe the details of template, the resolution of the images to which the template was matched, how the template was matched to the low resolution images, and the complexity of the matching process. All the templates described in this section were hand crafted after visual inspection of a few sample images. The templates are intended to serve as proofs of concept; i.e. they demonstrate that it is indeed possible to design simple qualitative concepts that may be used to correctly classify a reasonably diverse set of input images. More systematic general ways of deriving such templates will be discussed in chapter 5.

4.8.3.1 Snowy Mountain Template

The snowy mountain template represents the class of images which contain frontal views of snowy mountains. The template consists of three regions (A,B,C). Region A corresponds to the sky. Region B corresponds to the snow. Region C corresponds to the mountain. The qualitative relationships between the regions, the quantitative information, and the intraregion relationships encoded in the model are shown in Table 2. These relationships express the photometric and spatial relationships which are often valid in snowy mountain images. The spatial relationships express that the sky should be above the snow or low-lying clouds, which in turn should be above the mountain. The photometric relationships express that the sky is usually bluer than the snow and the mountain, the snow is generally lighter than both the sky and the mountain, and, finally, the mountain most likely is darker than both the sky and the snow. The intraregion relationships for the sky suggest that its blue color component is usually greater than its red or green components. The size relationships express that the snow usually comprises a smaller part of the image than the sky or mountain. Figure 4.11 illustrates the general structure of the snowy mountain template. The colors shown in the figure are examples of colors that satisfy the specified photometric relationships between the regions.

TABLE 2. Relationships encoded in the snowy mountain template.

mountain template	spatial	luminance	color	size
A to B	$A_x = B_x$	$A_l < B_l$	$A_r < B_r$	$A_s > B_s$
	$A_y < B_y$		$A_g < B_g$	
			$A_b < B_b$	
A to C	$A_x = C_x$	$A_l > 1.2*C_l$	$A_b > 1.2*C_b$	$A_s < C_s$
	$A_y < C_y$			
B to C	$B_x = C_x$	$B_l > C_l$	$B_r > C_r$	$B_s < C_s$
	$B_y < C_y$		$B_g > C_g$	
			$B_b > C_b$	
OTHER			$A_b > A_r$	$A_s = 1x2$
			$A_b > A_g$	$B_s = 1x1$
				$C_s = 1x4$

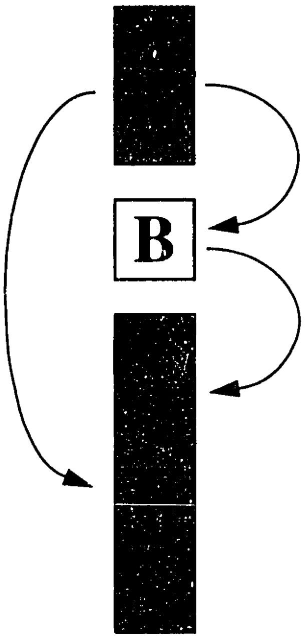


FIGURE 4.11. *The snowy mountain template consists of three regions A, B, and C. The figure shows their relative spatial relationships and sizes. The colors of the regions satisfy the photometric relationships of the model.*

The model is compared to low resolution images at the first level of the pyramid (*32x32* pixels). Figure 4.12 shows examples of mountain images at this resolution. At this resolution, the sky, snow, and mountain are still distinct.

Generally, the difference in snowy mountain regions exist along the vertical image axis. The regions are largely uniform horizontally across the image. We use this information to optimize the detection process. For a given vertical configuration of the model, where region A is above B which is above C, the technique starts at the left side of the image ($x = 0$) and sweeps the configuration horizontally across the image to try to detect triples that satisfy the pairwise relationships and other attributes shown in the table 2.

The model regions are specified to contain one or more low resolution image pixels. For instance, region A is hard coded to contain 2 images pixels, one on top of the other. Similarly, region C is specified to contain 4 image pixels in a vertical configuration. For reasons of efficiency, the image pixels in each of the regions are compared to the pixels in other regions that share the same column (or have the same x value). Figure 4.13 shows how the image pixels in each region A (A1, A2) are compared to the image pixels in region C (C1, C2, C3, C4). There are no cross column comparisons of regions as the template moves across the image.

The relations are encoded as sets of pairwise relationships. The pairwise relationships are defined as valid if all the pertinent constraints shown in the table hold for a given pixel in a first region and a given pixel in the second region. As the template is swept across the image, the number of configurations of valid (A,B), (B,C) and (A,C) pixel pairs are counted. The image is considered to be in the class described by the model if 85% of the image pixels in regions A and B satisfy the specified relative relationships, 60% of image pixels in regions B and C satisfy the specified relative relationships, and 90% of image pixels in regions A and C satisfy the specified relative relationships. These percentages reflect the fact that the sky (A) and mountain (C) regions usually span the image, while the snow distribution (B) is variable.

All spatially valid vertical orderings of the three regions are compared to the image in the manner described above. For a column size of m pixels and three model regions, there is an upper bound of m^3 configurations. If the model regions contain r image pixels, the number of pairwise comparisons for one image column is $3*r^2$. For an image of width n, the number of computations is $3*nr^2 *m^3$. In the case of the mountain template, the upper bound on the number of comparisons made is $48*32^4$ or $O(32^4)$, which is not very large. This is comparable to an operation which accesses every pixel once in a *1024x1024* image.

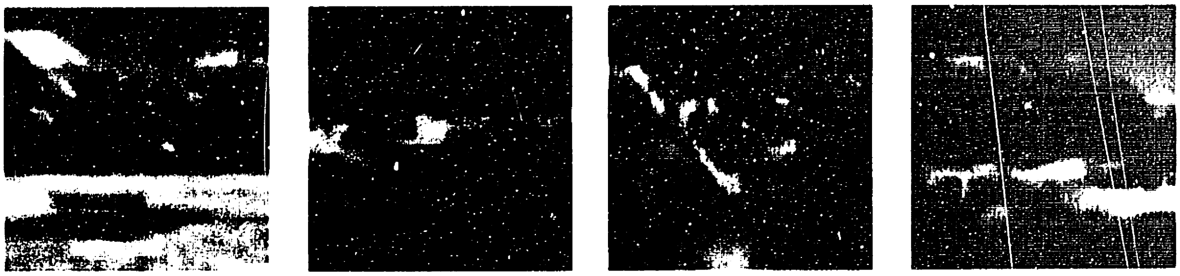


FIGURE 4.12. *Four low resolution snowy mountain images of size (32x32) scaled by 300% for viewing purposes.*

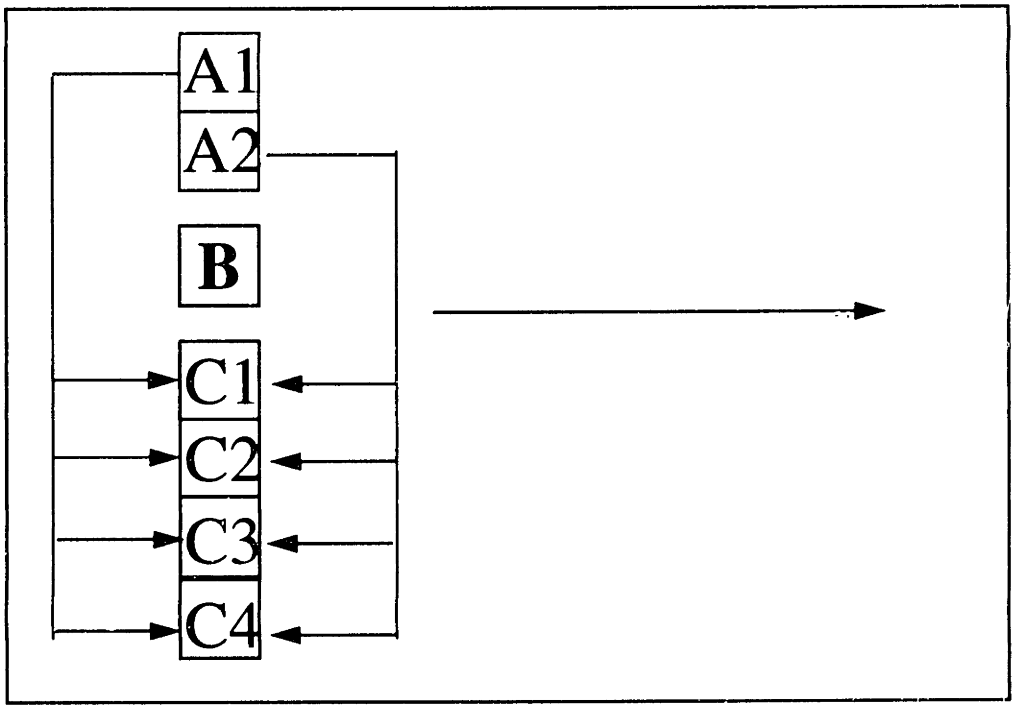


FIGURE 4.13. *At a specific location of the template, the image pixels in each of the regions are compared to each other. This example shows how image pixels in region A are compared to image pixels in region C. As the template is moved, the interregion comparisons are made at each column in the image. No cross column relationships are computed.*

4.8.3.2 *Snowy Mountain with lake template*

The snowy mountain with lake template describes a subset of the images in the class represented by the previous template. One extra region (D) which corresponds to the lake is added to the template. The template describes all views of mountains with lakes in the foreground of the image. The template does not distinguish between frozen or liquid lakes. This template encompasses all the previous relationships from the snowy mountain template and adds the spatial and photometric relationships between the mountain and lake regions. These relationships express that the lake should be below the mountain in the image plane. In general, the lake should also be lighter <u>or</u> lighter and bluer than the mountain. We express this OR relationship by encoding two sets of relationships between regions C and D in Table 3. Figure 4.14 shows the revised mountain template. The colors shown in the figure are examples of colors that satisfy the specified photometric relationships between the regions.

TABLE 3. The extra relationships encoded in the snowy mountain template with lake

	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	$C_x = D_x$	$C_l < D_l$	$C_b < D_b$	$C_s = D_s$
	$C_y < D_y$			
[illegible]	$C_x = D_x$	$C_l < D_l$		$C_s = D_s$
	$C_y < D_y$			
[illegible]				$C_s = \mathit{1x4}$
				$D_s = \mathit{1x4}$

The matching of this model to the image uses the same process as the previous template. In addition the same size resolution images are used (*32x32* pixels). The image is considered an instance of the class, if the mountain concept is detected and if the relationships between the pixels in region C and D are valid 90% of the time. The computation complexity of matching this model to the image versus the previous model is on the order $O(32^5)$.

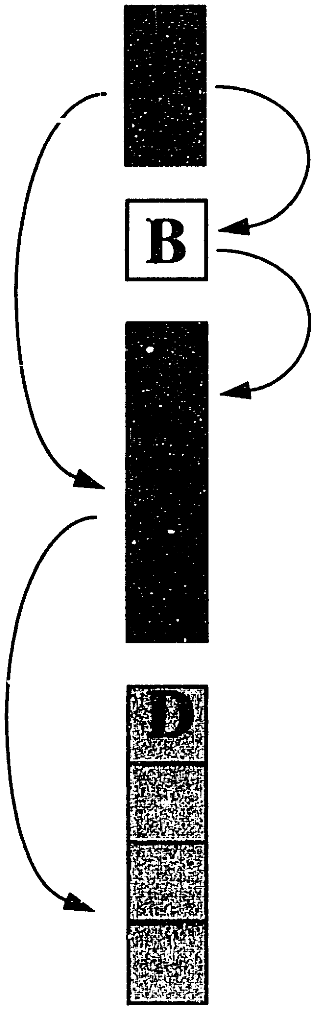


FIGURE 4.14. ***The revised snowy mountain template which includes a lake region. The template consists of four regions A, B, C, and D. The figure shows their relative spatial relationships and sizes. The colors of the regions satisfy the photometric relationships of the model.***

4.8.3.3 Field Template

The field template characterizes the class of images that contain sky and field regions. The template consists of two regions (A,B). Region A corresponds to the sky. Region B corresponds to the field. The specifics of the model are shown in Table 4. These relationships express the photometric and spatial relationships which are often valid in field scenes. The spatial relationships express that the sky is above the field in the image plane. The photometric relationships express that the sky is generally bluer than the field. The color components of the field should have the relationships that the red and green chromatic values should be greater than the blue chromatic value. Figure 4.15 illustrates the general structure of the field template. The colors shown in the figure are examples of colors that satisfy the specified photometric relationships between the regions.

The model is compared to the processed images at the lowest level of the pyramid, which contain the fewest details. The size of the images are *8x8* pixels. Figure 4.16 shows examples of field images at this resolution.

TABLE 4. The relationships encoded in the field template.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	$A_x = B_x$		$A_b > B_b$	$A_s = B_s$
[illegible]	$A_y < B_y$		$A_b > B_r$	
[illegible]			$A_b > B_g$	
[illegible]			$B_b < B_r$	
[illegible]			$B_b > B_g$	
[illegible]				$A_s = 1x1$
[illegible]				$B_s = 1x1$

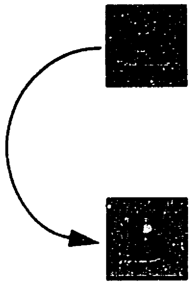


FIGURE 4.15. *The field template consists of two regions A and B. The figure shows their relative spatial relationships and sizes. The colors of the regions satisfy the photometric relationships of the model.*

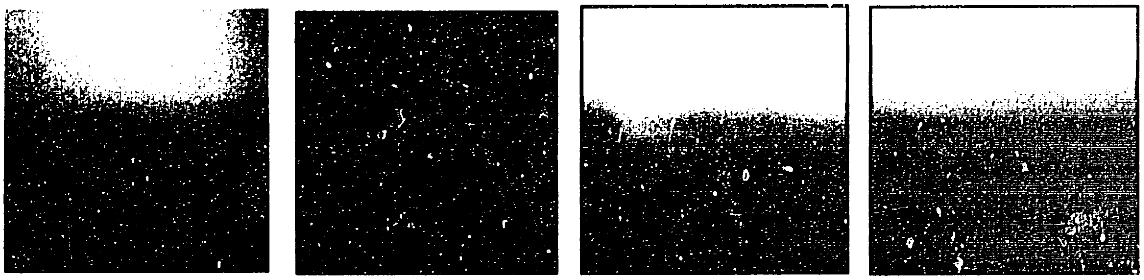


FIGURE 4.16. *Four low resolution field images of size (8x8) scaled by 1200% for viewing purposes.*

Similar to the case with mountains, the perceptual difference in regions of field images usually exists along the vertical image axis. The regions generally are largely uniform along the horizontal axis. This information is used to optimize the detection process. For a given configuration of the model, where region A is above region B, the technique starts at the left side of the image ($x = 0$) and sweeps the configuration horizontally across the image to try to detect pairs that satisfy the relationships and other attributes shown in Table 4.

The model patches are specified to contain only one image region. For reasons of efficiency, the image regions in A and B are only compared if they are in the same image column. The number of valid sets cf pairwise relationships are counted as the template is swept horizontally across the image. The image is considered to be a member of the class described by the model if the relationships between regions A and B are satisfied 80% of the time.

All spatially valid vertical orderings of the two regions are compared to the image in the manner described above. The upper bound on the number of comparisons to detect the field concept is $O(8^3)$.

4.8.3.4 Waterfall Template

The waterfall template represents the class of images which contain waterfalls in a somewhat vertical orientation. The waterfall must be at least as long as 40% of the vertical size of the image. The waterfall template consists of three regions (A, B, C). Region B corresponds to the waterfall. Region A and C respectively correspond to the areas on the left and right side of the fall. The qualitative relationships between the regions, the quantitative information, and the intraregion relationships encoded in the model are shown in Table 5. These relationships express the photometric and spatial relationships which are often invariant in waterfall images. The spatial relationships express that the waterfall should be in between two regions. The photometric relationships express that the waterfall is usually lighter than the surrounding regions and that the waterfall is more blue and green than red. Figure 4.17 illustrates the general structure of the waterfall template. The colors shown in the figure are examples of colors that satisfy the specified photometric relationships between the regions.

The model is compared to the processed image at the second level of the pyramid which is of size *16x16* pixels. Figure 4.18 show examples of waterfalls at this resolution

TABLE 5. Relationships encoded in the waterfall template.

	$A_x < B_x$	$1.6A_l < B_l$	$A_b < B_b$	$A_s > B_s$
	$A_y = B_y$		$A_r < B_r$	
			$A_g < B_g$	
	$B_x = B_x$	$B_l > 80$	$B_r > 1.2*B_b$	$B_s < B_s$
	$B_y = B_y$		$B_r > 1.2*B_g$	
	$B_x < C_x$	$1.6C_l < B_l$	$C_b < B_b$	$B_s < C_s$
	$B_y = C_y$		$C_r < B_r$	
			$C_g < B_g$	

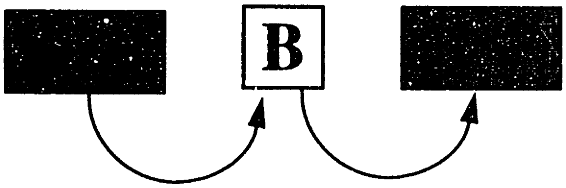


FIGURE 4.17. *The waterfall template consists of three regions A, B, and C. The figure shows their relative spatial relationships and sizes. The colors of the regions satisfy the photometric relationships of the model.*

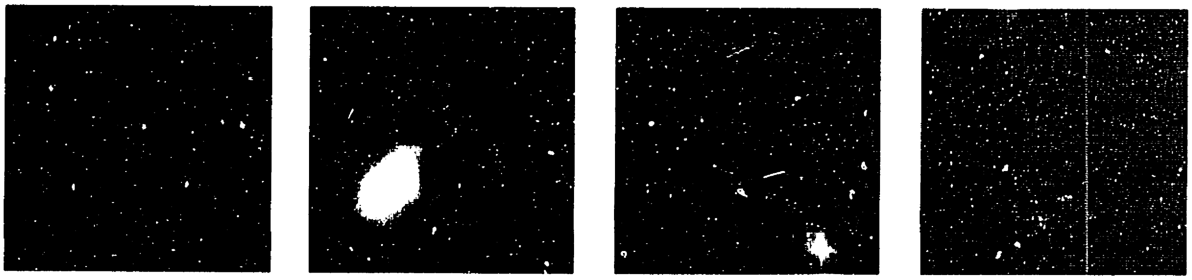


FIGURE 4.18. *Four low resolution waterfall images of size (16x16) scaled by 600% for viewing purposes.*

In comparison to the previous classes, the differences in waterfall regions usually exist along the horizontal image axis and are largely uniform along the vertical axis. This information is used to optimize the detection process. For a given horizontal configuration of the model, where region A is to the left of region B which is to the left of region C, the technique starts at the top of the image (y = 0) and sweeps the configuration vertically down the image to try to detect triples that satisfy the pairwise relationships and intrapatch relationships shown in Table 5.

The model regions are specified to contain one or more low resolution pixels. For instance, regions A and B are designated to contain 2 image pixels, one next to the other. For reasons of efficiency, the image pixels in each of the regions are only compared to the pixels in other regions that share the same row (or have the same y value).

The relations are encoded as sets of pairwise relationships. The pairwise relationships are defined as valid if all the pertinent constraints shown in the table hold for a given pixel in a first region and a given pixel in the second region. As the template is swept down the image, the number of configurations of valid (A,B) (B,C) and (B,B) pixel pairs are counted. In addition, the longest string of pixels that satisfy the (B,B) relationships is remembered. The image is considered to be in the class described by the model if 60% of the (A,B) and (B,C) relationships are satisfied, 50% of the (B,B) relationships are satisfied, and the longest string of contiguous B regions is at least 40% of the image. These percentages reflect that the waterfall should comprise at least 40% of the image, allowing for possible sky regions.

All spatially valid horizontal orderings of the three regions are compared to the image in the manner described above. The complexity of matching the model to the image is $O(16^4)$.

4.9 Results

In this section, we report the results of the automated classification using the described templates. Subsections 4.9.1 - 4.9.4 show examples of the "true positives", "false positives", and "false negatives" for each of the templates. The "ground truth" for the results is based on the perceptual judgments of two observers. Table 6 summarizes the results in terms of percentages over the 700 image database broken into two groups for each scene class; perceptually true positives and negatives. (True positives and false negatives are described with respect to the total number of positives. False positives and true negatives are described with respect to the total number of negatives.)

TABLE 6. Results described as percentages over the 700 image database

RESULTS	"true" positives	"false" negatives	"false" positives	"true" negatives
Snowy mountain template	75%	25%	12%	88%
Snowy mountain with lake template	67%	33%	1%	99%
Field template	80%	20%	7%	93%
Waterfall template	33%	67%	2%	98%

4.9.1 Results for the snowy mountain template

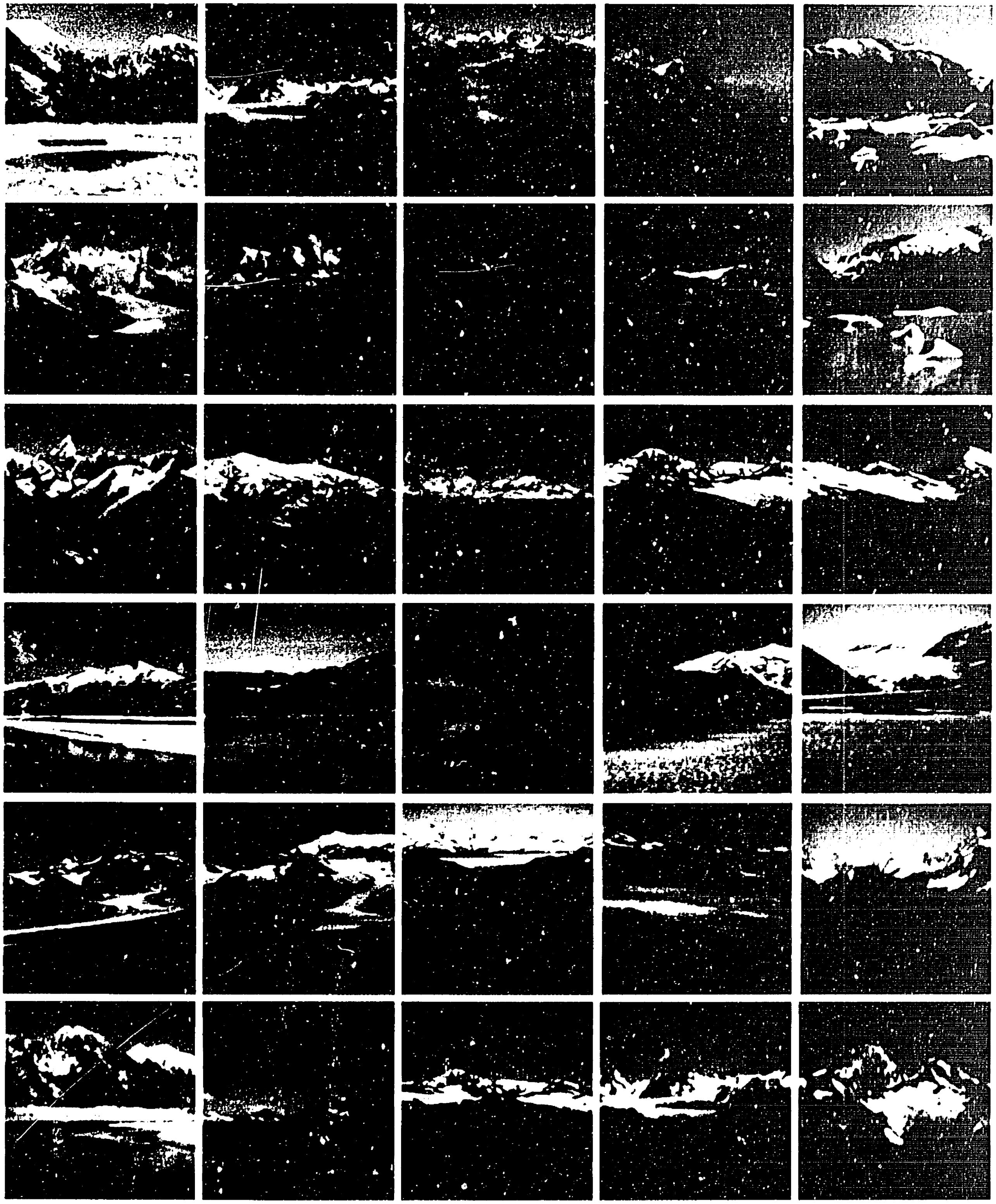


FIGURE 4.19. *Some true positives detected by the snowy mountain template.*

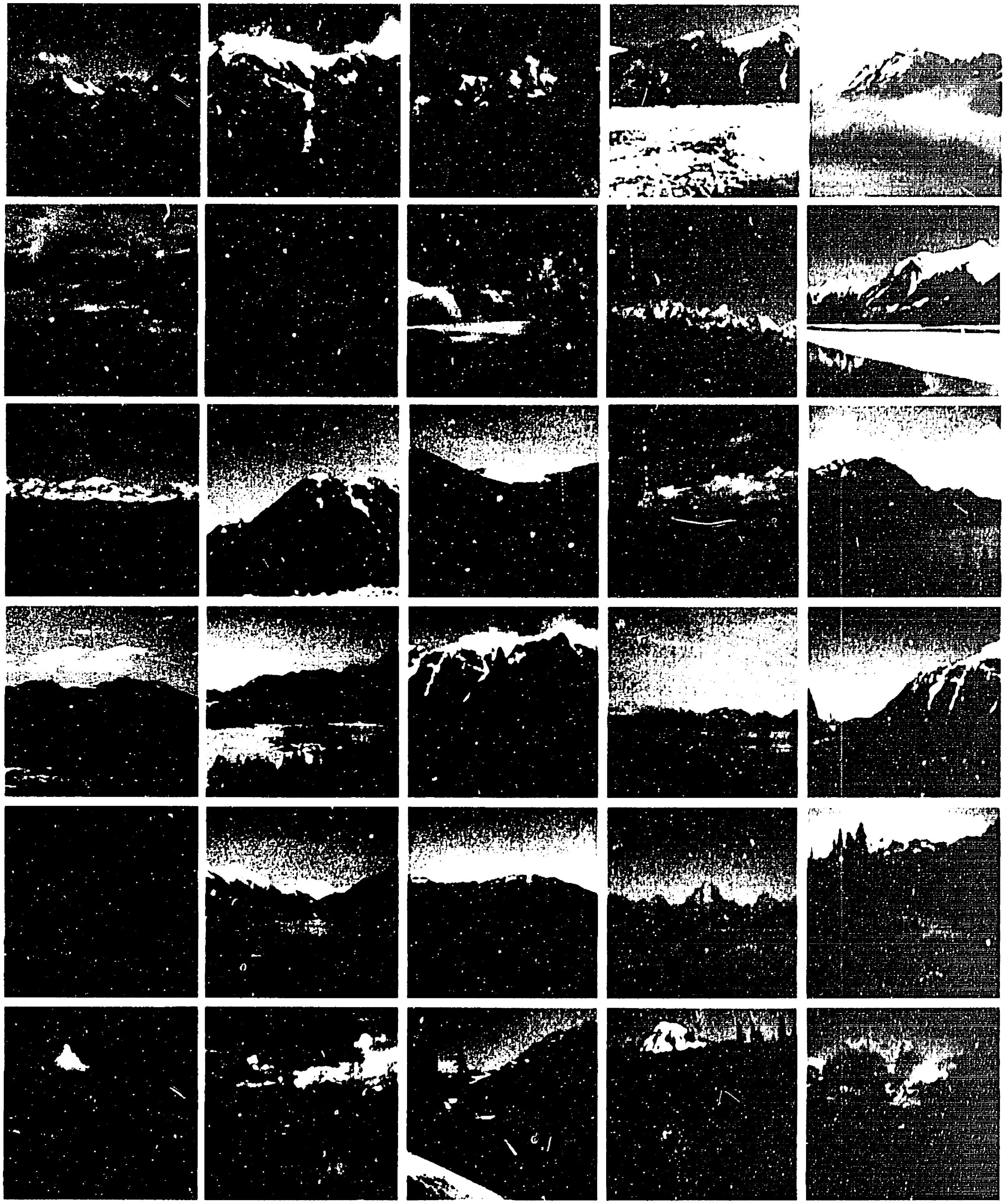


FIGURE 4.20. *More true positives detected by the snowy mountain template*

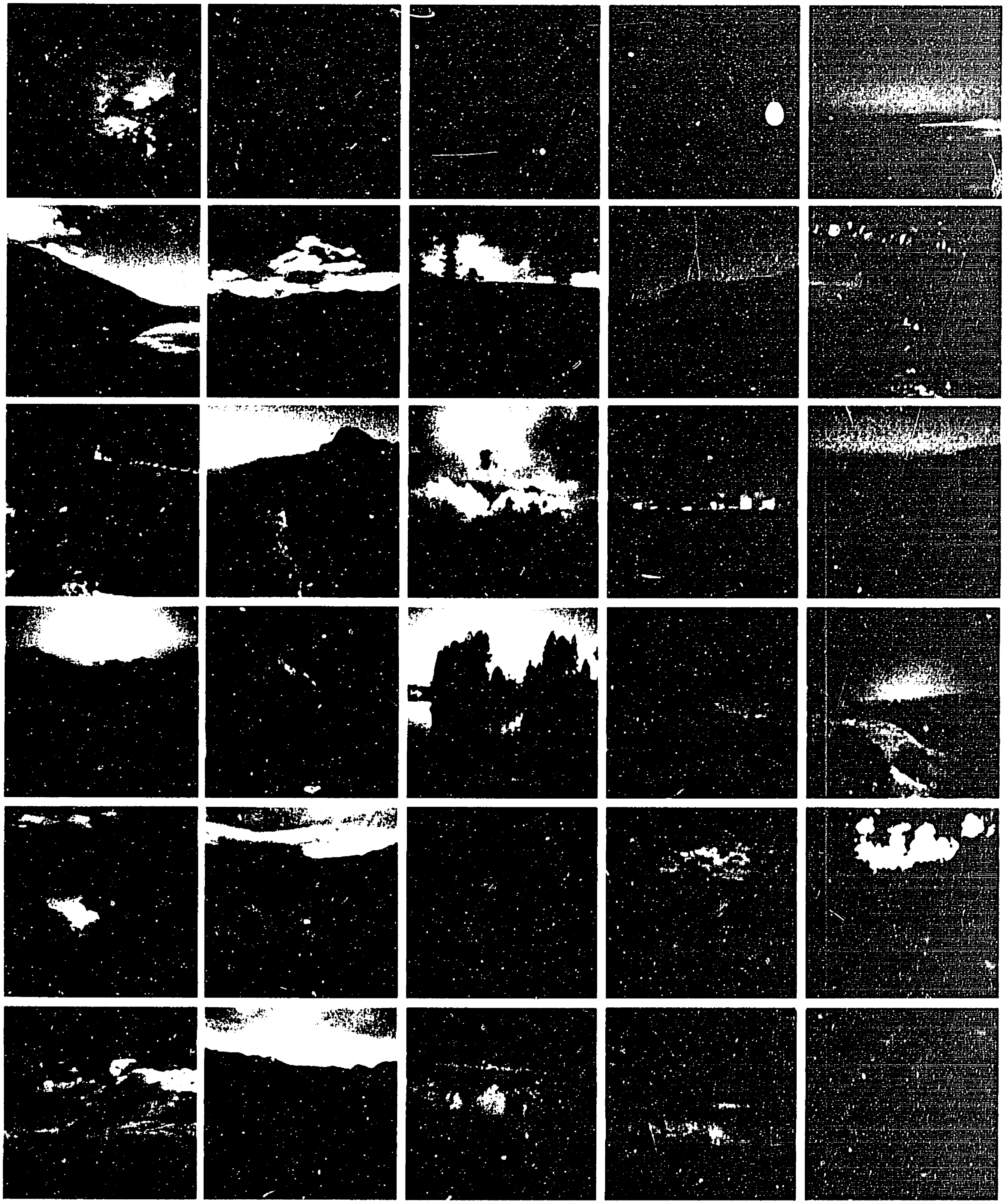


FIGURE 4.21. *Some false positives detected by the snowy mountain template*

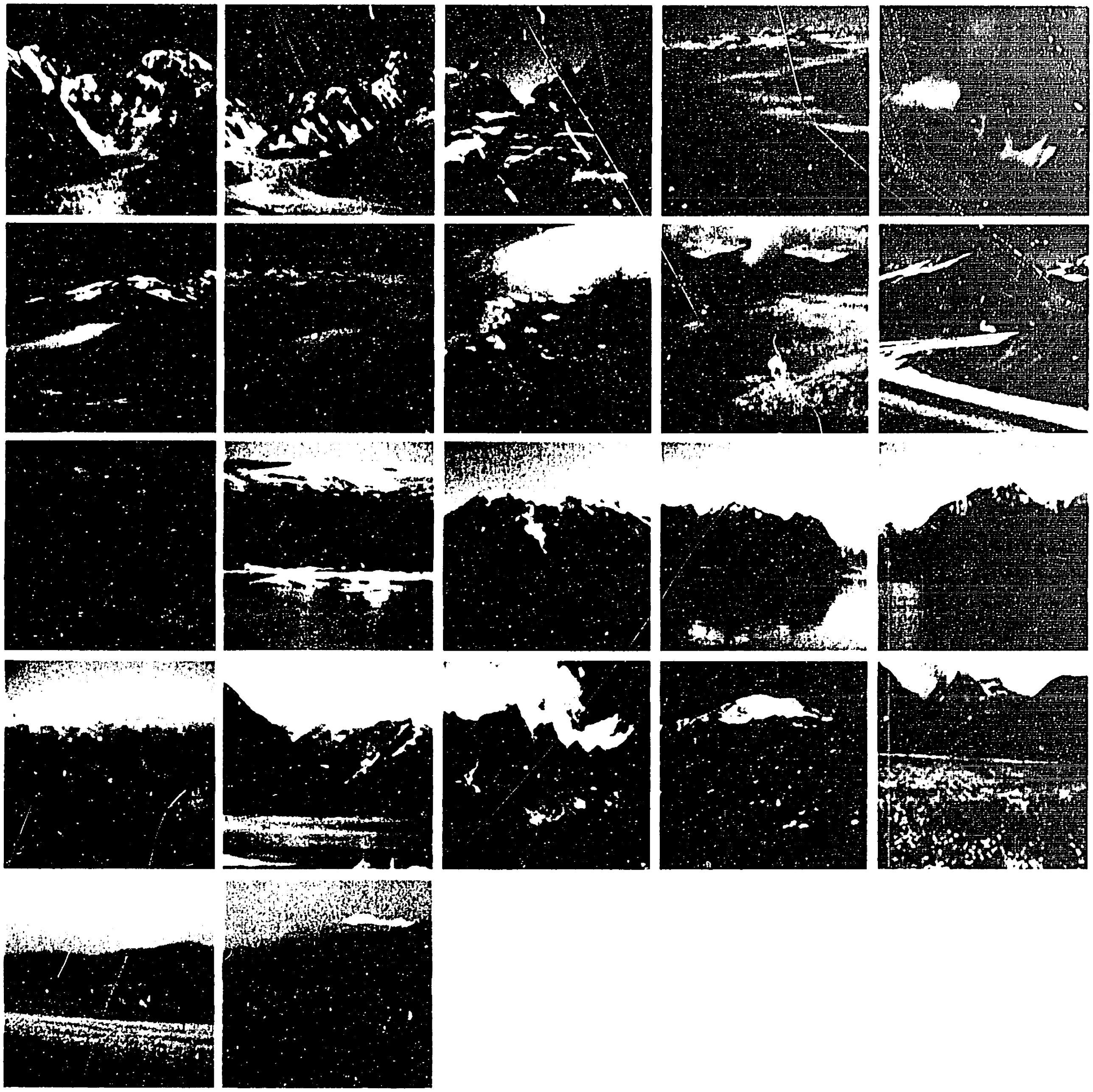


FIGURE 4.22. *All the false negatives that were not detected by the snowy mountain template.*

4.9.2 Results for the snowy mountain with lake template

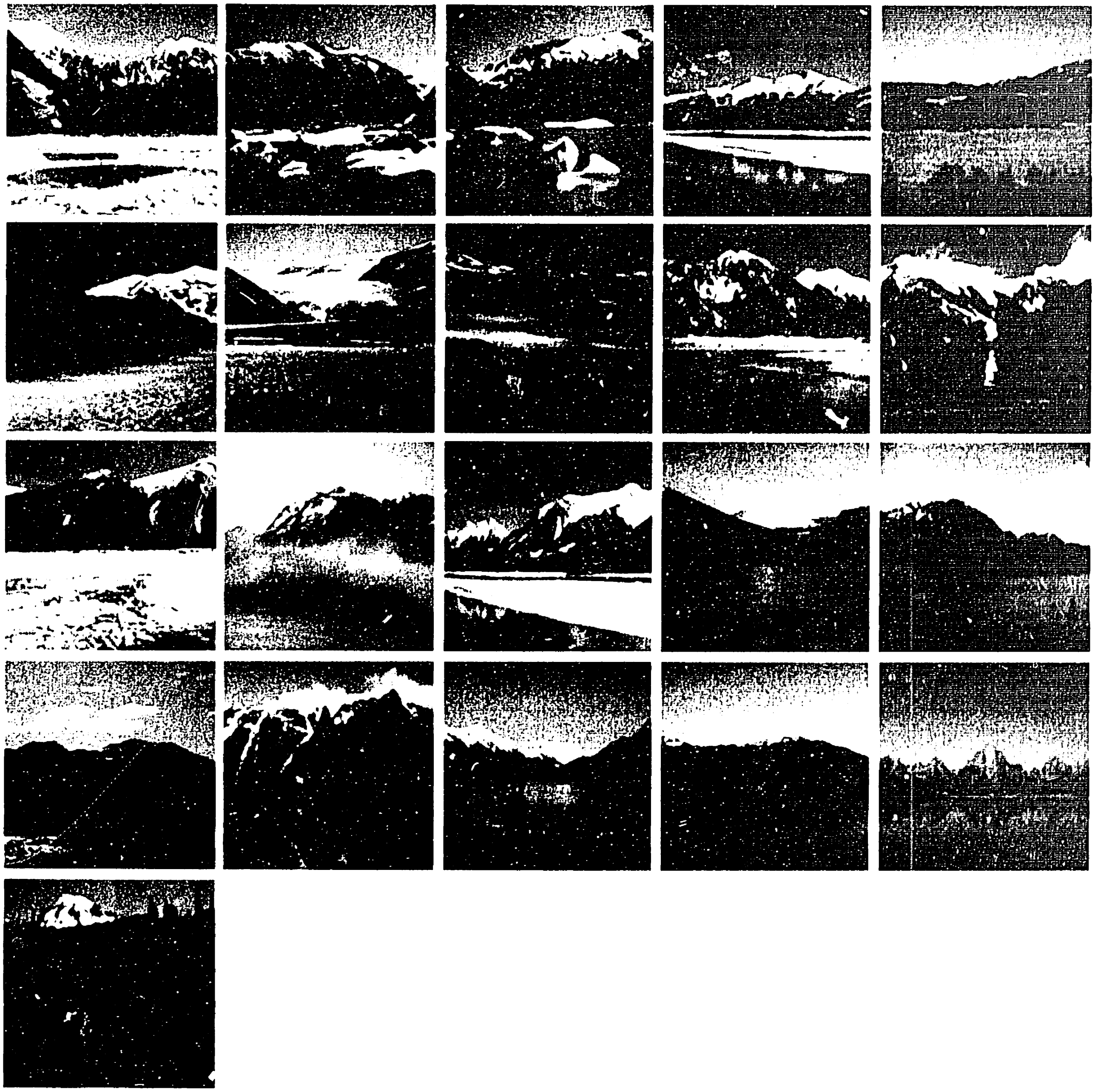


FIGURE 4.23. *All the true positives detected by the snowy mountain with lake template.*

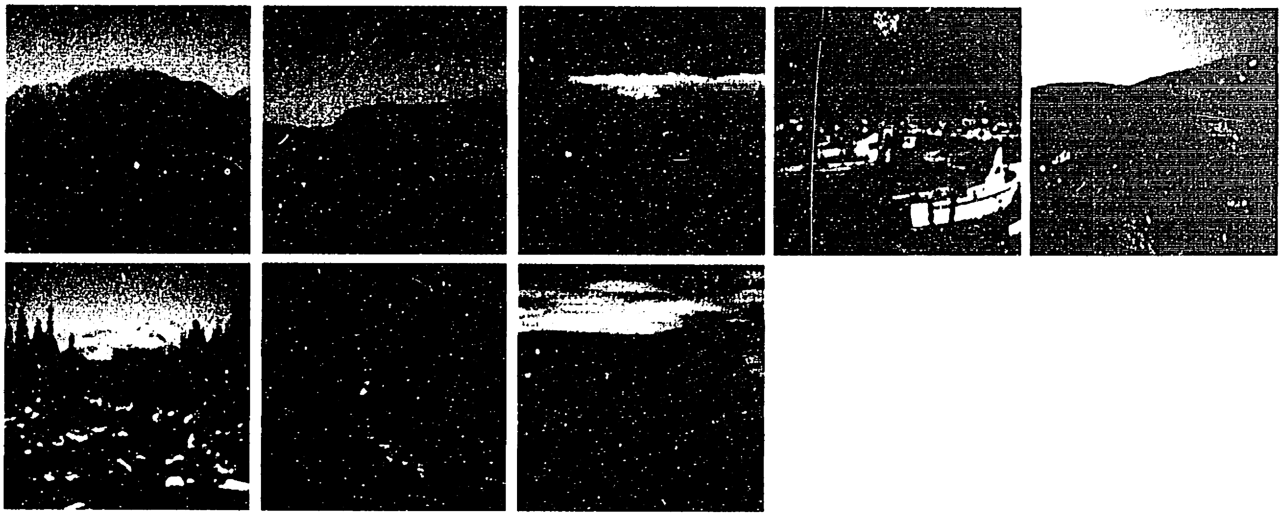


FIGURE 4.24. *All the false positives detected by the snowy mountain with lake template.*

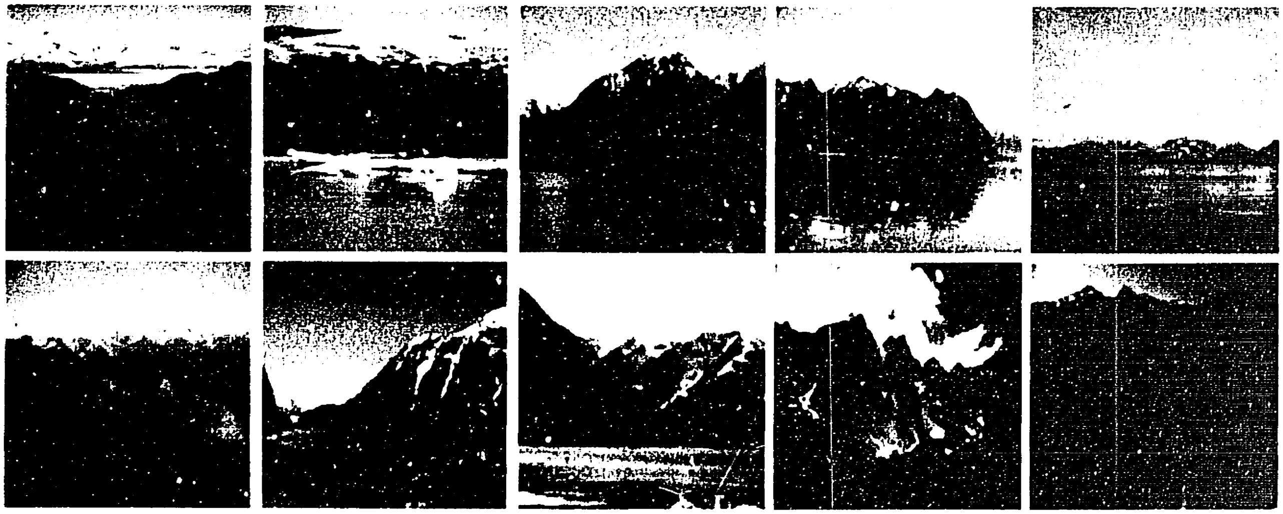


FIGURE 4.25. *False negatives not detected by the snowy mountain with lake template.*

4.9.3 Results for the field template

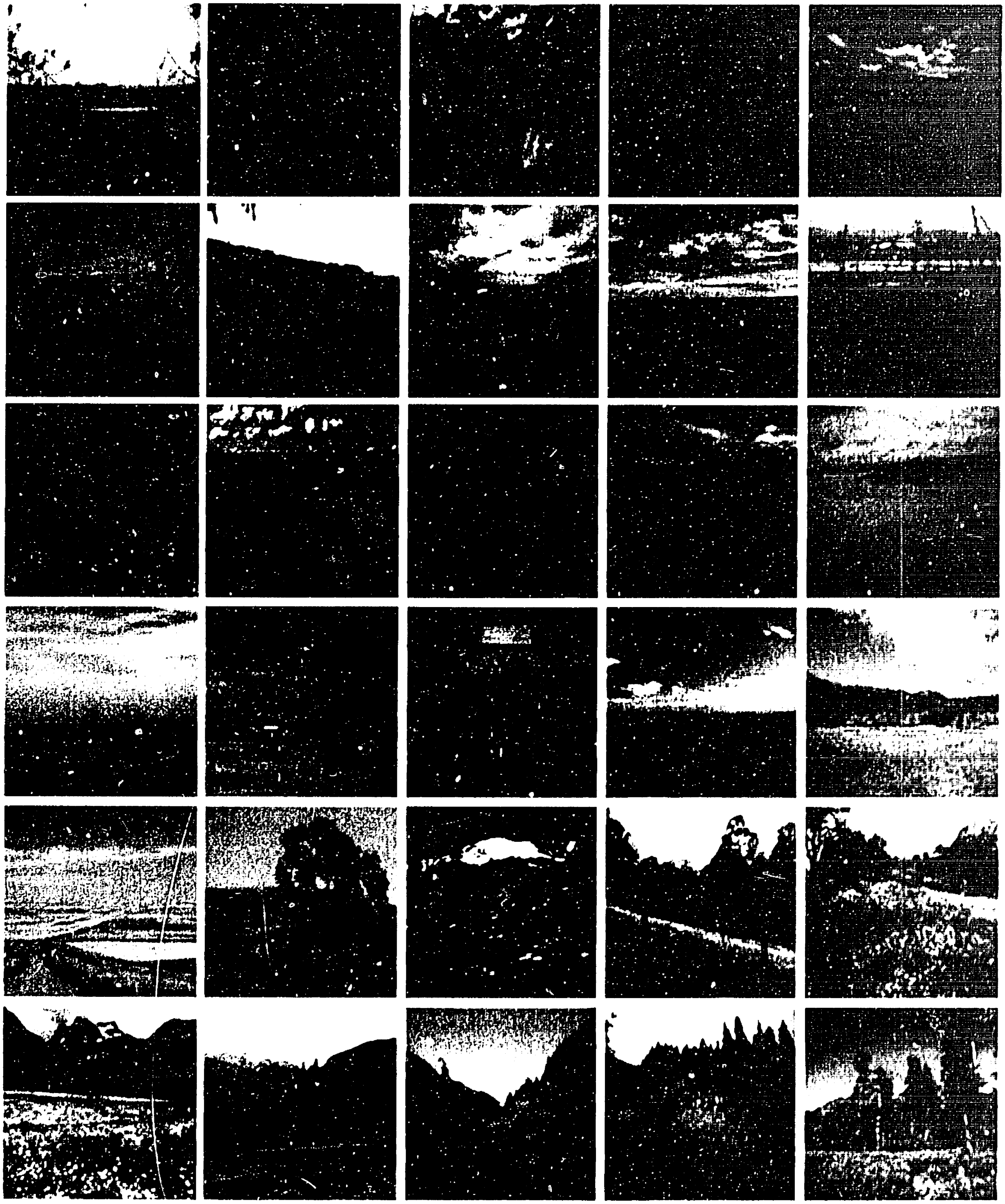


FIGURE 4.26. *Some true positives detected by the field template*

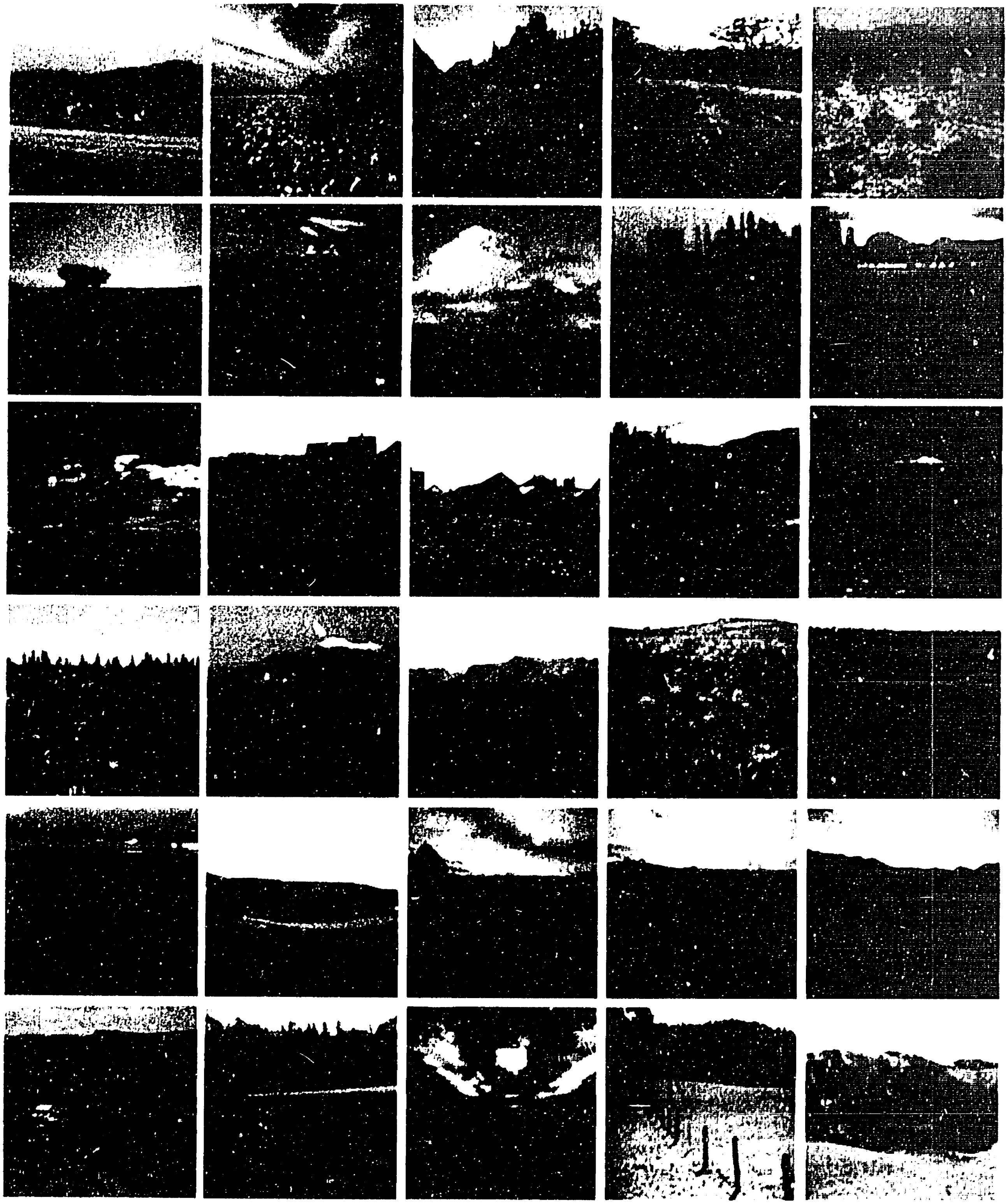


FIGURE 4.27. *Additional true positives detected by the field template.*

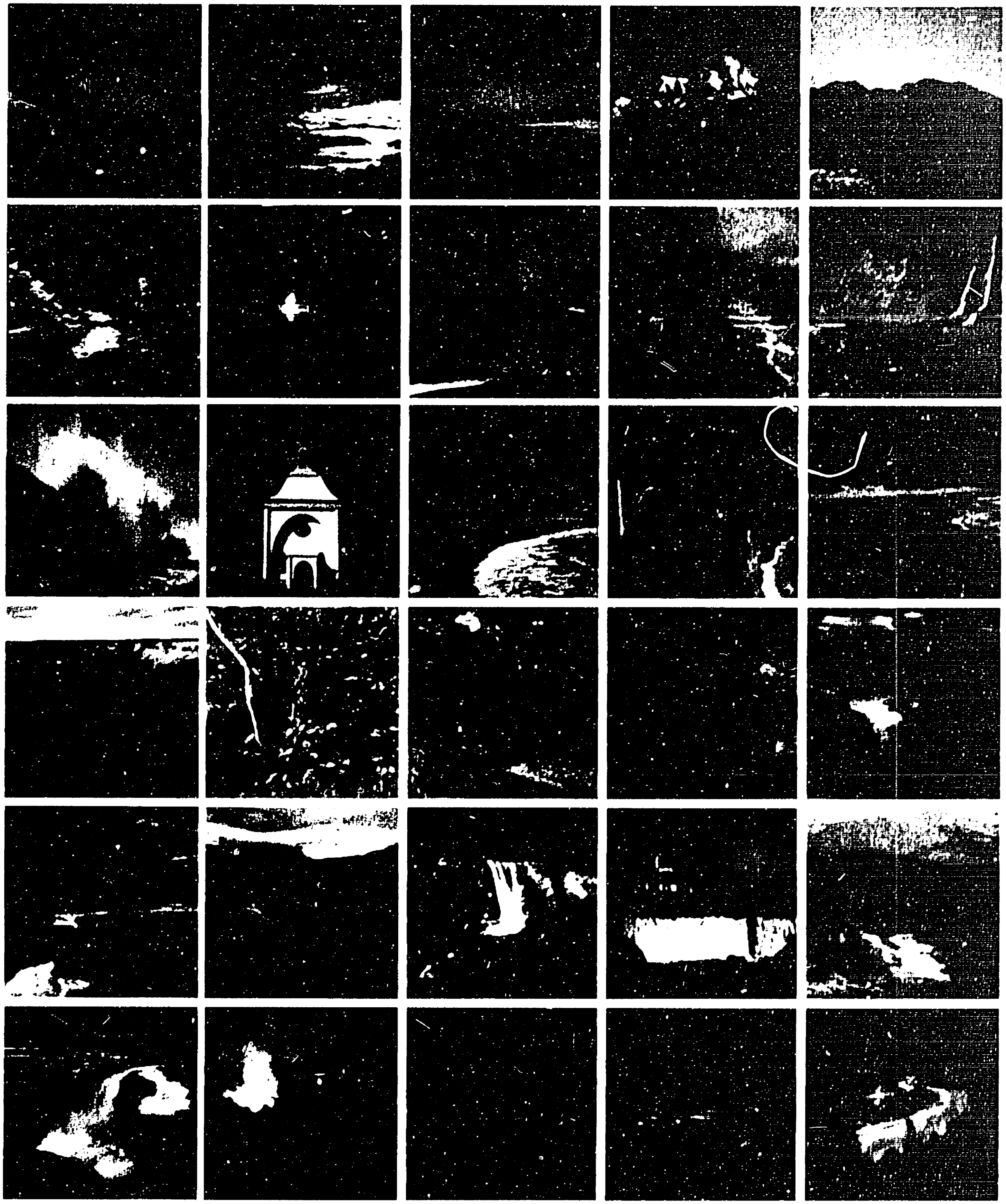


FIGURE 4.28. *Some false positives detected by the field template.*

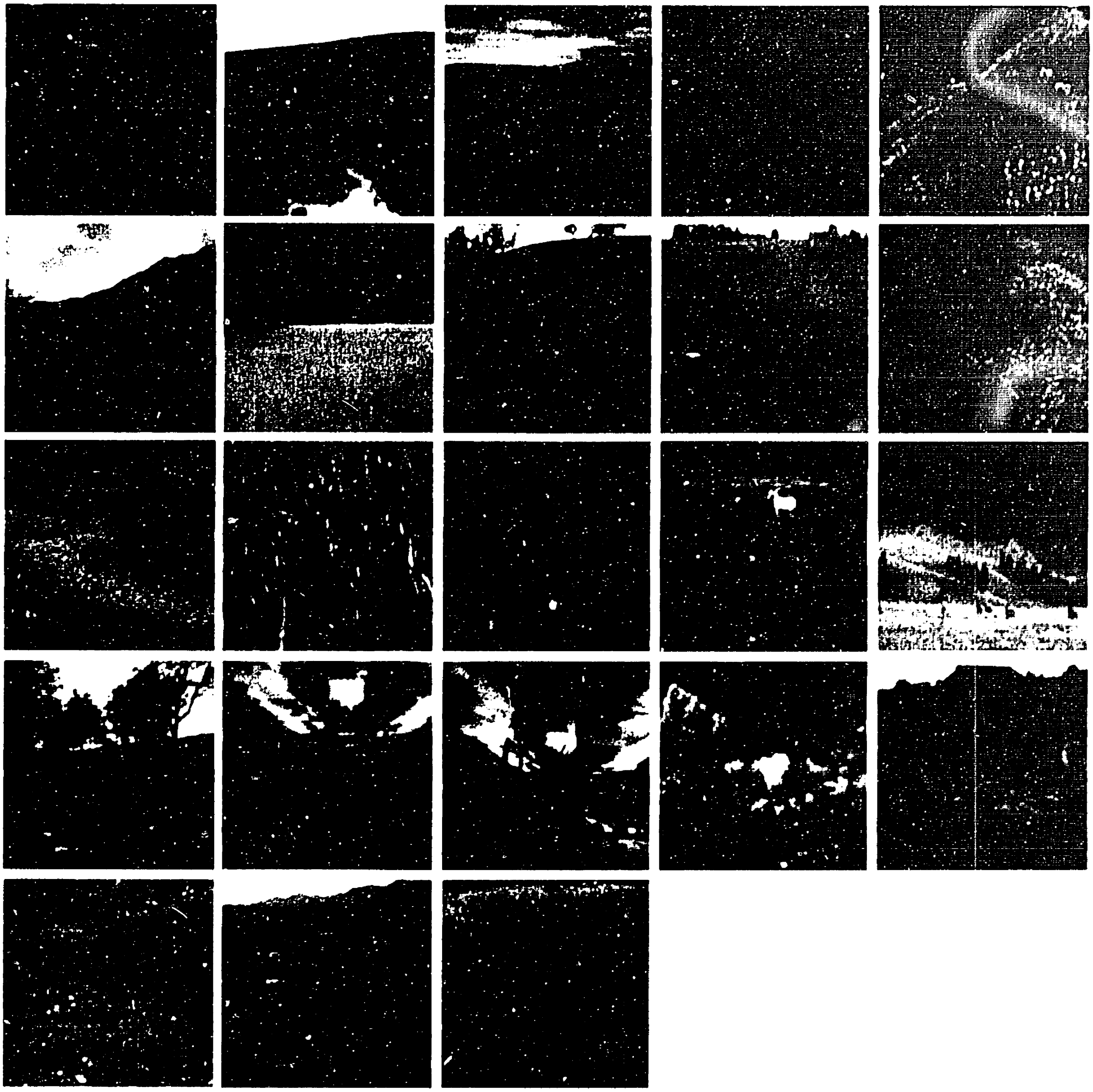


FIGURE 4.29. *False negatives not detected by the field template*

4.9.4 Results for the waterfall template.

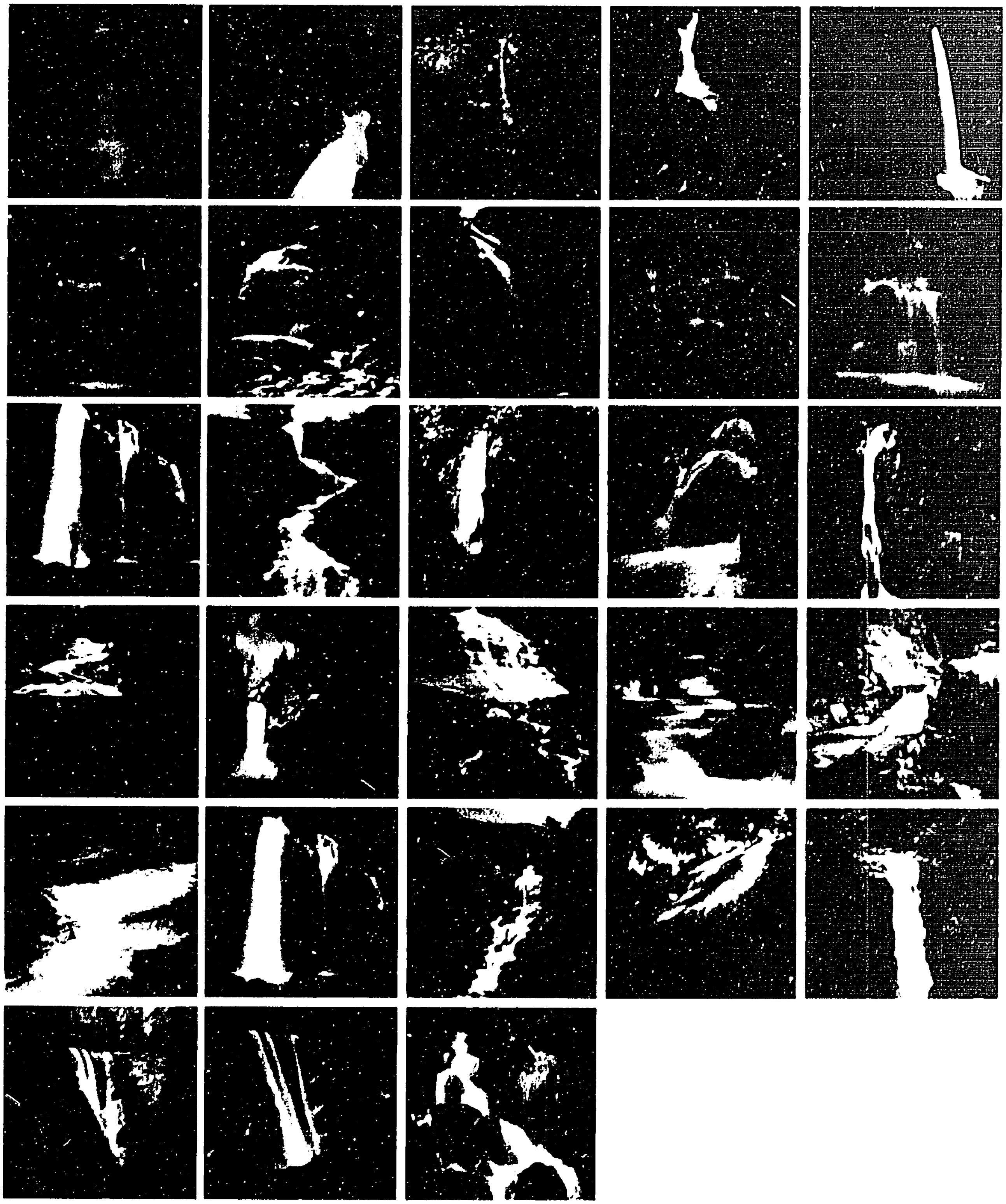


FIGURE 4.30. *True positives detected by the waterfall template*

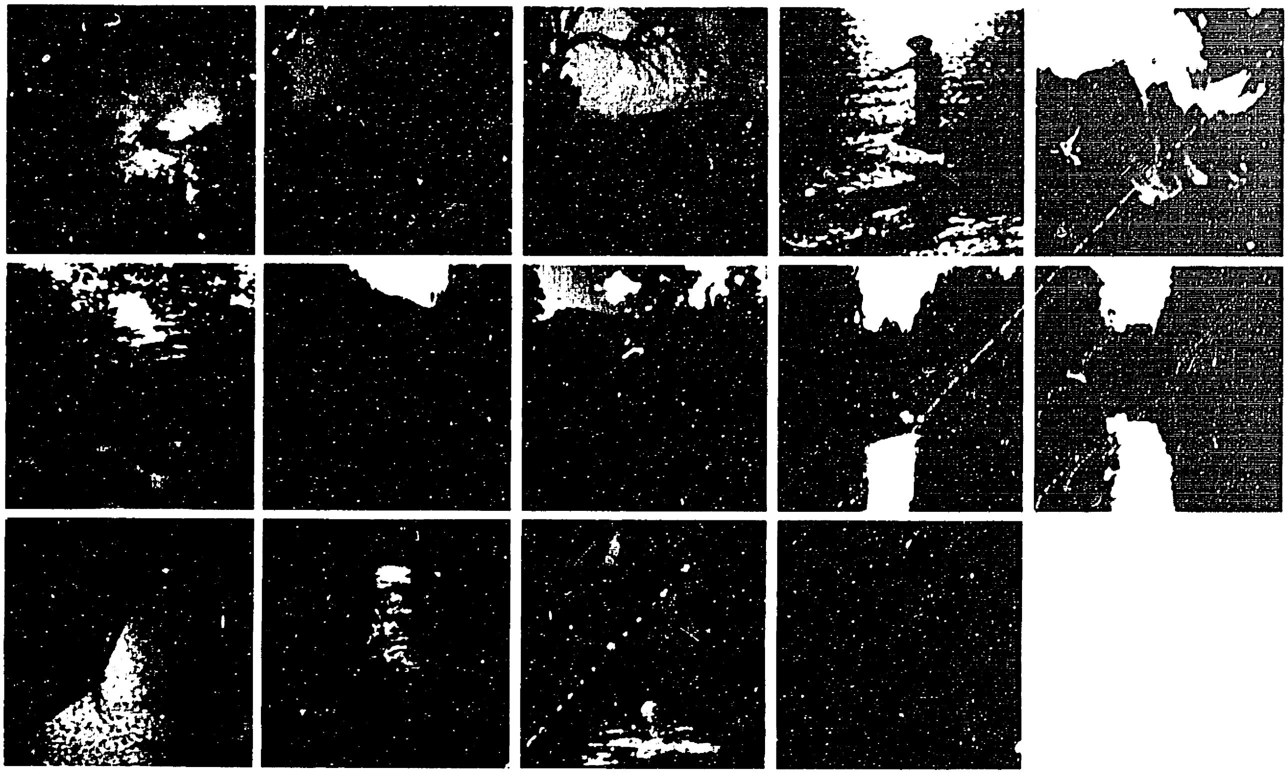


FIGURE 4.31. *All the false positives detected by the waterfall template*

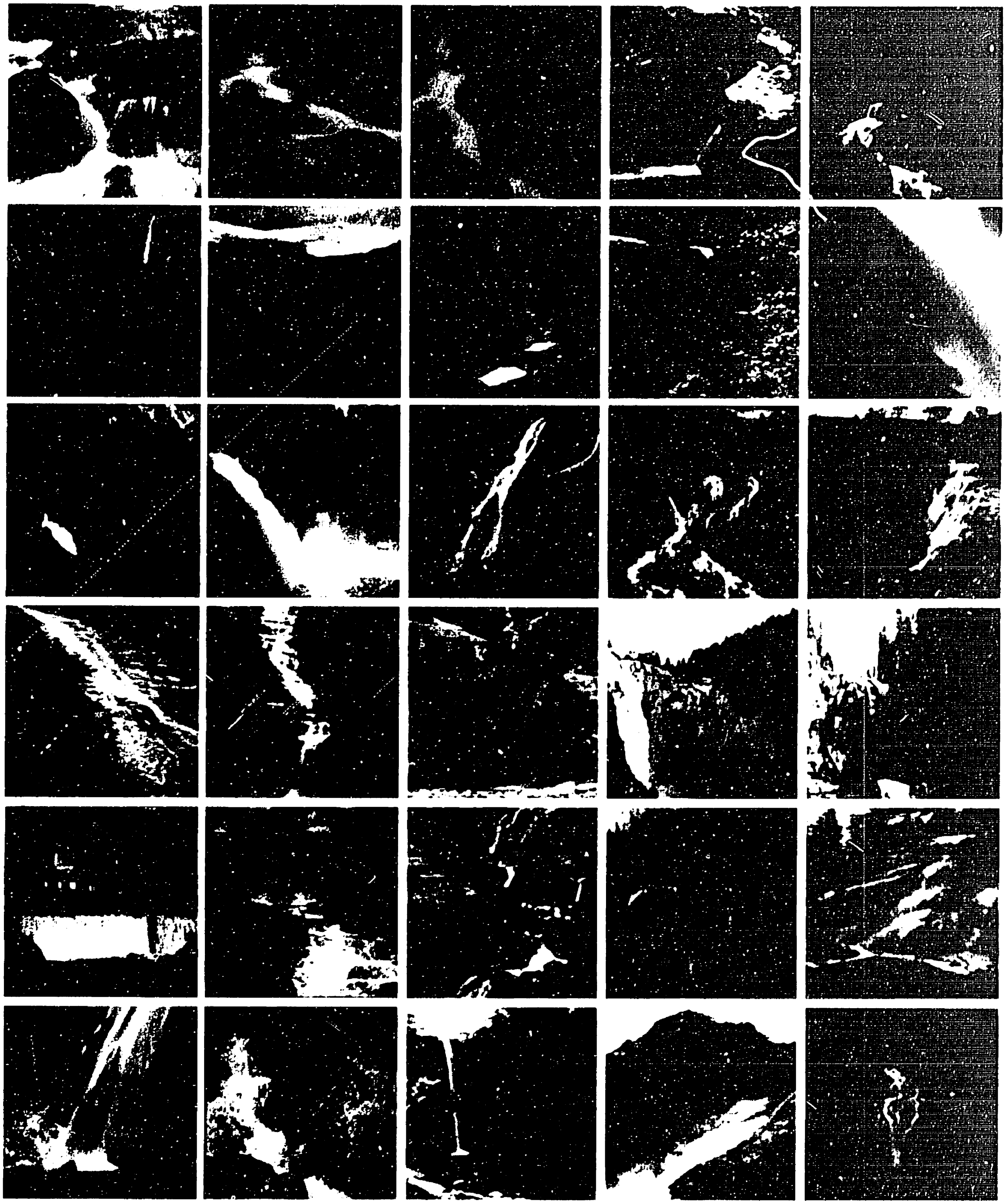


FIGURE 4.32. *Some of the false negatives not detected by the waterfall template*

4.10 Testing the approach on other types of scenes

In this section, we present results which demonstrate that the configural recognition approach is extendable to scenes that include man-made structures such as buildings. This example will also serve to demonstrate the ability of our approach to incorporate other image attributes such as texture for image classification purposes.

We developed a cityscape template that characterizes the class of images that contain panoramic views of sky and large buildings. The salient aspects of many cityscape scenes are the color of both the sky and the buildings and also the texture of the buildings (and lack of texture in the sky). In most of our examples of city panoramas, the sky was generally blue and without texture, the buildings had either a golden brown, greenish, or blue hue and together created a texture with significant vertical orientation.

The template constructed to characterize this class consists of two regions (A,B). Region A corresponds to the sky. Region B corresponds to the buildings. The specifics of the model are shown in Table 7. These relationships express the photometric, spatial, *and texture* relationships which are often valid in cityscape scenes. The spatial relationships express that the sky is above the buildings in the image plane. The intraregion photometric relationships express that the sky is generally blue (written as one intraregion relationship for region A) and that the buildings are predominantly brown, green or blue (written as three possible intraregion relationships for region B). The texture relationship between the two patches (denoted as A_t and B_t) encodes that region B should have almost an order of magnitude more vertical texture than region A. Figure 4.33 illustrates the general structure of the cityscape template. The colors and textures shown in the figure are examples of colors and textures that satisfy the specified photometric and texture relationships.

The model is compared to both a low resolution color image and a filtered higher resolution greyscale image. The color images used contained very little detail and were of size *8x8* pixels. The greyscale images were of size *128x128* pixels and were filtered with a simple vertically oriented edge detector. Figure 4.34(a) shows examples of color low resolution cityscape images. Figure 4.34(b) shows examples of the filtered higher resolution greyscale images and Figure 4.34(c) shows the filter used to generate those images.

TABLE 7. The relationships encoded in the cityscape template.

	[illegible]	[illegible]
	$1.2*A_b > A_r$	
	$1.2*A_b > A_g$	
	$B_b < 1.2*B_r$	
	$B_b > 1.2*B_g$	
	$1.3*B_b > B_r$	
	$1.3*B_b > B_g$	
		$9*A_t < B_t$

Each of the model regions encompasses two contiguous rows of pixels in the low resolution color images and two regions of size *128x16* pixels at the corresponding positions in the filtered higher resolution images. For instance, if region A encompasses the first and second rows in the color image, it is also encompasses the first 32 rows in the filtered image. All non-overlapping configurations of model region A and B, where A is above B, are evaluated to determine if the criteria in Table 7 hold for a given image. The criteria hold if all the color pixels encompassed by region A and region B satisfy the color criteria and if the sum of the luminances in the texture regions encompassed by region A and B satisfy the texture criteria.

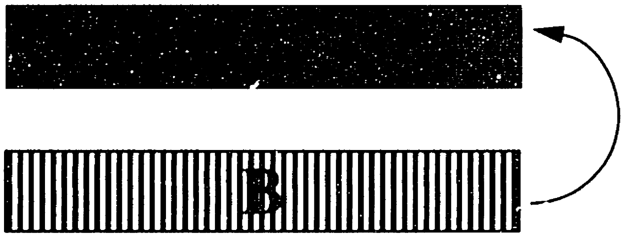
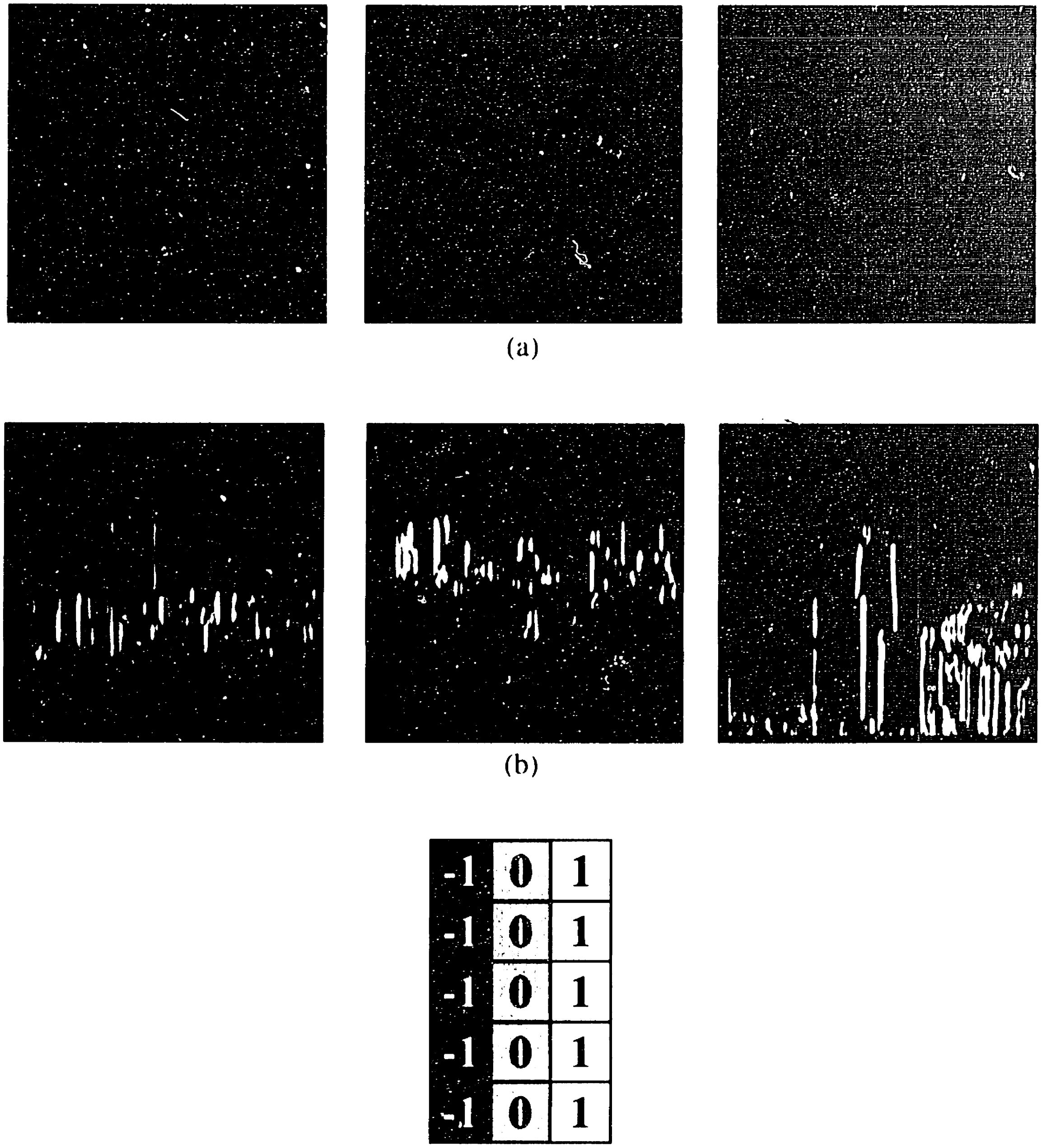


FIGURE 4.33. *The cityscape template consists of two regions A and B. The figure shows their relative spatial relationships. The colors and textures of the regions satisfy the relationships encoded in the model.*



(a)

(b)

-1	0	1
-1	0	1
-1	0	1
-1	0	1
-1	0	1

(c)

FIGURE 4.34. *cityscape images (a) low resolution color images (magnified by 1600% for viewing purposes) (b) higher resolution texture images (c) the filter to used to generate the texture images.*

4.10.1 Results

We tested the cityscape template on a database of 32 city images and the original database of 700 natural images. The template correctly identified 71% of the cityscapes. The "true" positives are shown in Figure 4.36. The template incorrectly identified 14% of the natural images as cityscapes. The false positives were mainly from the coastal and field images. The misidentified images had the correct relative colors. They also had a great deal of texture in some regions. However, the texture of the images existed at all orientations, not just in the vertical direction. Additionally, the texture was not arranged in long vertical lines, but appeared fragmented. Figure 4.35 shows one of the misidentified coastal images and its filtered counterpart. The cityscapes template can easily be adjusted to be more discriminating via the use a more sophisticated texture filter or by encoding that region B should not have dominant texture in bands other than the vertical and horizontal ones.

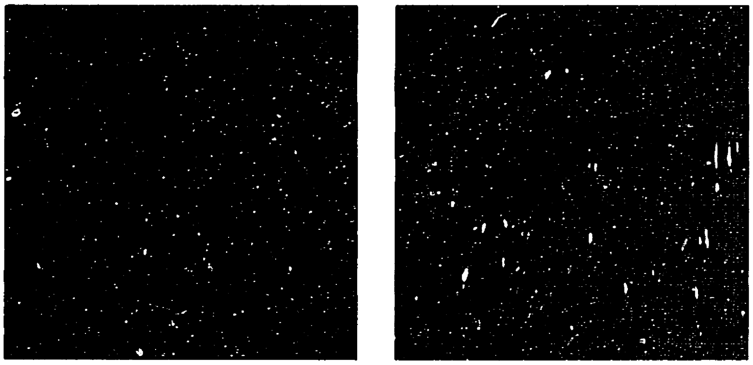


FIGURE 4.35. ***One of the coastal "false positives" and its filtered version.***

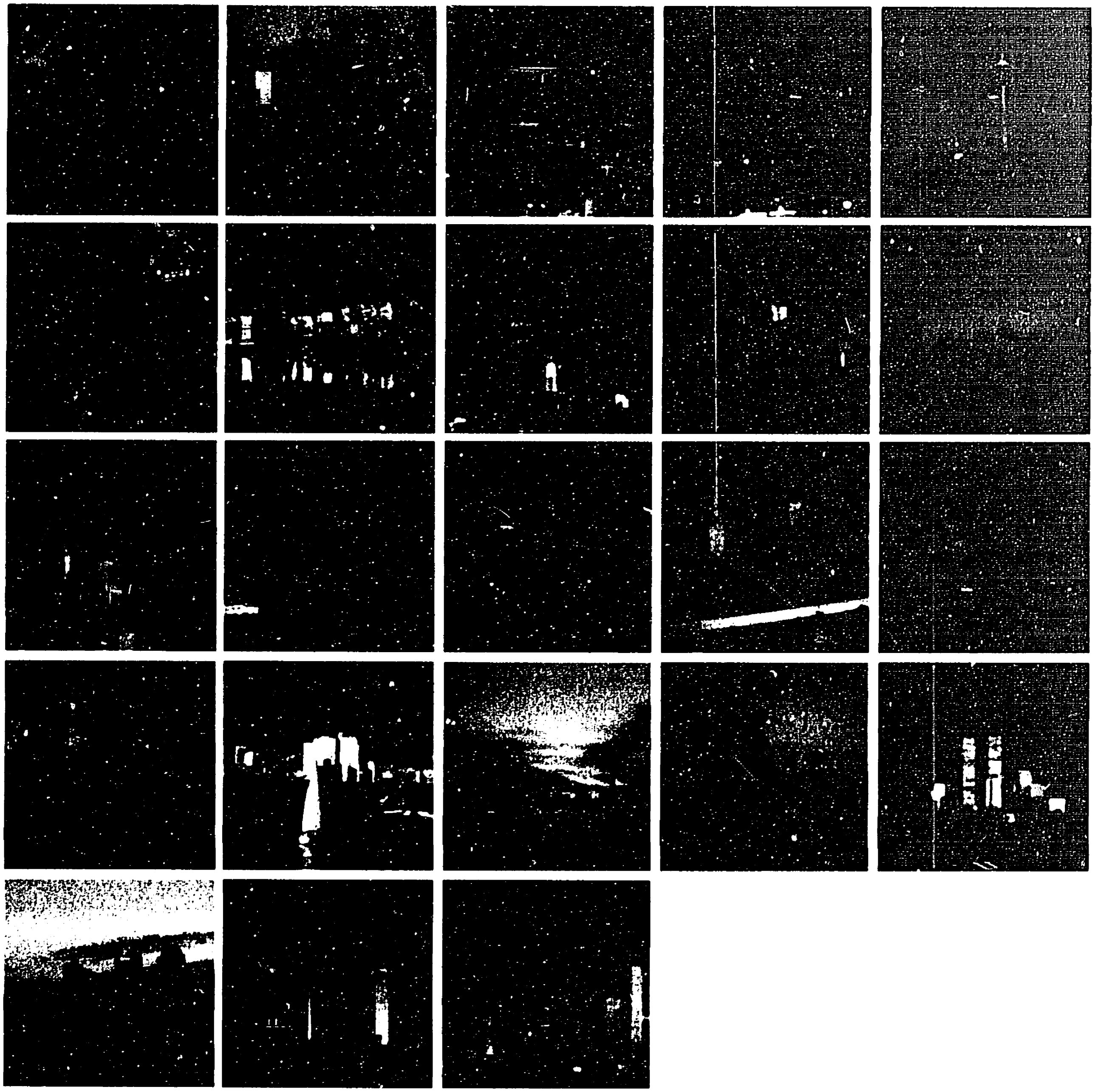


FIGURE 4.36. *True positives detected by the cityscapes template*

4.11 Discussion

In this chapter, we introduced a novel approach to image classification called "configural recognition". The approach employs novel qualitative metrics over low resolution image regions to classify images or to compute image similarity. This is in contrast to the majority of prior work on scene classification that uses quantitative measures of an image's color, texture, and shape characteristics to compute image similarity. We found that qualitative, rather than quantitative, measures are effective to describe scene content. We presented the configural recognition approach to classification in the form of a system that can be used for image database indexing. Configural recognition has several advantages, but also some significant shortcomings. In this section, we will discuss the benefits and the limitations in the context of the results we presented in sections 4.9 and 4.10.

Generalization across scene class:

A significant benefit of the configural recognition approach to scene classification is that by using qualitative models we are able to effectively generalize over different attributes of a class. For instance, the field class model is able to detect field scenes even though the scenes differ in their colors (green vs. red or brown grass/flowers), textures (solid field vs. tall grass or rows of crops), illumination (bright sunny day vs. cloudy and overcast), in the viewing parameters (close up vs. panorama), and in some content (i.e. the presence or absence of trucks, trees, barns or fences). Figure 4.37 shows some of the field scenes detected by the field template that differ in these parameters. A classification strategy based on quantitative color and texture, template matching, or object recognition would most likely not be able to generalize as effectively.

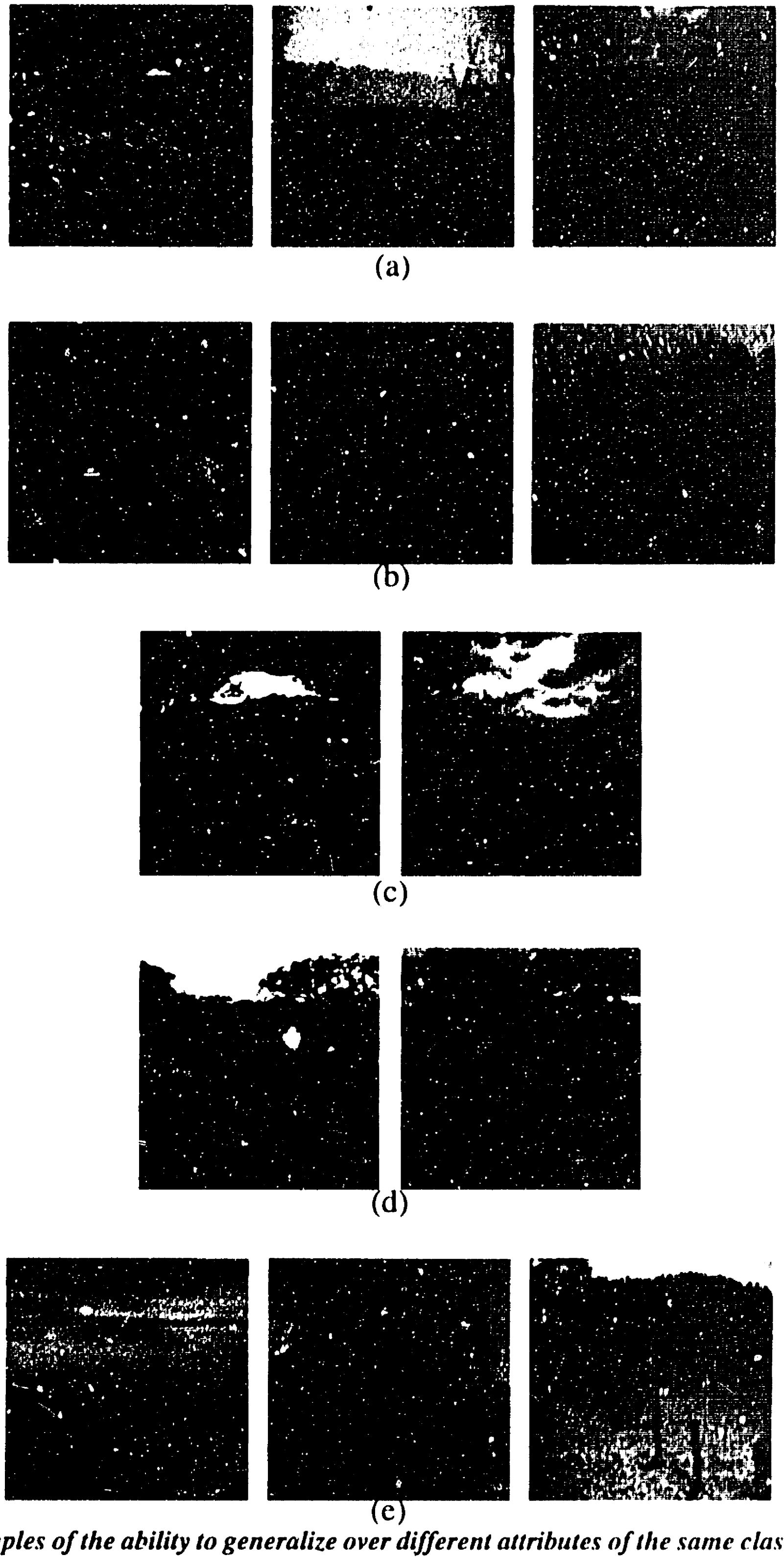


FIGURE 4.37. ***Examples of the ability to generalize over different attributes of the same class. The images are all field scenes even though they differ in the attributes of (a) color, (b) texture, (c) illumination, (d) viewpoint, (e) content.***

Discrimination between classes:

Another significant benefit of the configural recognition approach to scene classification is that we are able to effectively discriminate between different classes. For instance, the snowy mountain template correctly did not detect some glaciers, coastal images, ice pictures, and a city scene with a gushing river which share the same color attributes and some of the same texture attributes as snowy mountains scenes, but have incorrect spatial configurations. In addition, the mountain template did not classify a scrambled snowy mountain image as part of the class. Figure 4.38 shows a few of these true negatives. A strategy based solely on color or texture measurements would not be able to make some of these distinctions.

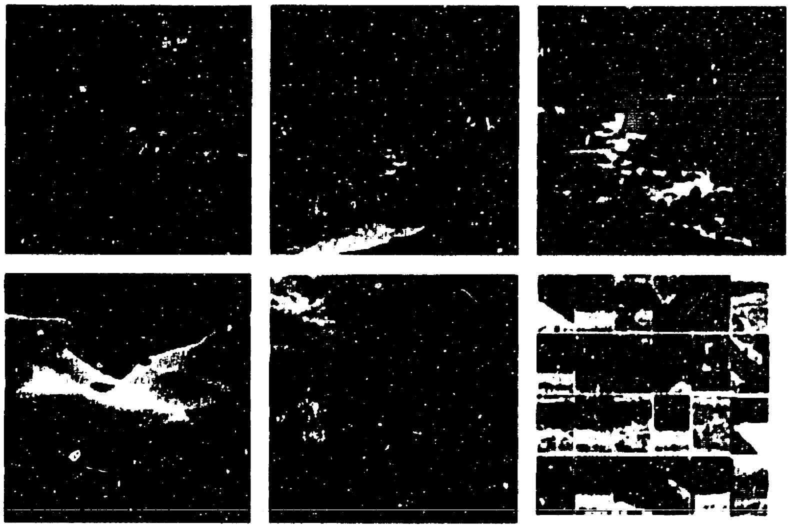


FIGURE 4.38. *Some of the true negatives not detected by the snowy mountain template.*

Computational efficiency:

The system is also computationally efficient. The models are matched to very low resolution imagery containing at most *32x32* pixels. In addition, the models used to demonstrate the approach contained very few regions (between 2 and 4). To test four templates, of varying complexity, on 100 images in our database took under 10 minutes on a pentium p.c. with an external file server.

The advantages of significant generalization ability and computational efficiency, however, come with a price. The impressive ability to generalize sometimes leads to overgeneralization. For instance, the snowy mountain template detected such false positives as a sunset scene, a coastal scene, a tree scens with clouds, a group of white houses on a cliff, an animal standing on a hill overlooking water, and a picture of pounding surf, which to a human observer clearly do not belong to the class. The snowy mountain template was also not able to finely discriminate cloudy mountain scenes from snowy mountain scenes. In addition, the snowy mountain template, as presented, would not be able to discriminate a real snowy mountain from a synthetic image which consists of three horizontal bands of blue, white, and grey in the correct spatial configuration. Figure 4.39 shows some of these false positives.

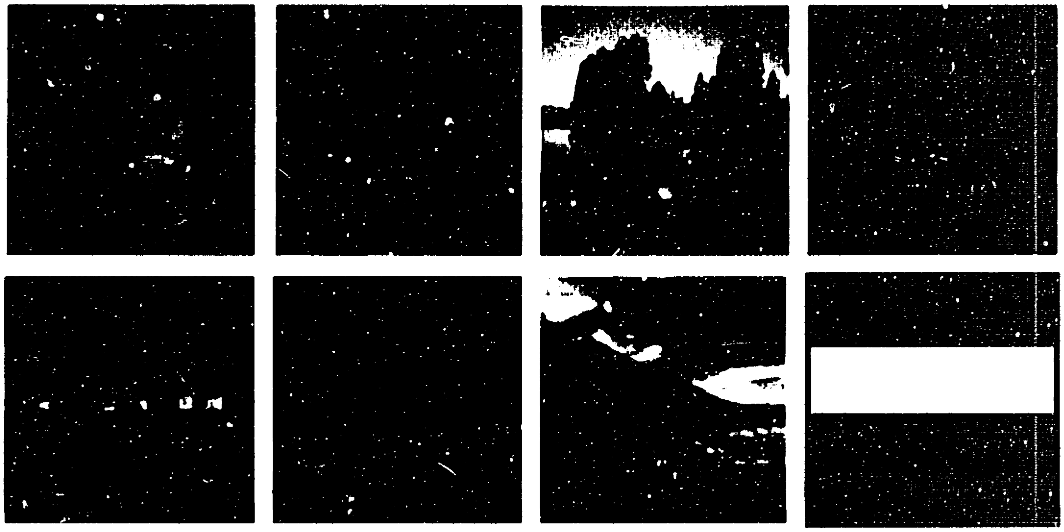


FIGURE 4.39. ***Some of the false negatives detected by the snowy mountain class.***

There are several additions, however, to the configural recognition approach that may allow us to prevent overgeneralization. Some of the false positives are incorrect because they have the wrong quantitative colors, despite having the correct qualitative colors. Examples include the sunset, coastal, tree, and animal scenes. This may be corrected by using a perceptually based quantitative color system (we discuss one possibility for such a system in the next chapter). Other false positives such as the houses on the coast, the pounding surf, the cloudy mountain, and the color patch image are detected because the system does not have any fine discrimination ability in small details. It is possible that using quantitative texture measurements in conjunction with the existing qualitative measures could mitigate this problem.

In some cases, the qualitative models may be very restrictive. For instance, the waterfall template only detected 33% of all the waterfalls in the database. However, the concept of any class is unlikely to be captured by a single qualitative template, which is one reason why a single template does not detect all the instances of one class in the database. However, a small set of such templates can be expected to capture a majority of the instances of the scene class. For instance, many of the waterfalls not detected by the existing template had green or brown surrounds which were almost as bright as the waterfall itself. Other waterfalls not detected were not vertically oriented. Provisions for these conditions were not encoded in the original waterfall template. By adding two or three more templates to capture these variations, it is possible to cover most of the instances of the class.

It is desirable to design a set of templates which together capture most of the true positives and have few false positives. From the data in Table 6, the templates with the fewest false positive rates were the waterfall and snowy mountain with lake templates. The waterfall template, as just described, is an example of a narrow detector due to the qualitative constraints. The snowy mountain with lake template is also a narrow but accurate detector most likely due to the fact that it had four model regions, the greatest number of regions amongst all the templates. This suggests that a small set of narrow templates, due to the qualitative constraints or the number of model patches, might span a large class and result in a low false positive rate.

This chapter has provided essentially a proof of existence that qualitative concepts that can describe significant subsets of a scene class. All the templates included here were handcrafted. In the next chapter, we discuss strategies to automate the template construction process.

CHAPTER 5 *Learning the Scene Class*

In the previous chapter, we demonstrated that models consisting of qualitative relationships between low frequency image regions could be used effectively to classify images. The models described there were hand-crafted after visually inspecting salient regions in example images and extracting the consistent relationships between those regions. It would be desirable if, instead of having to hand-craft the models, an automated process could take a set of example images and generate a template or a set of templates which describe the relevant consistencies between the pictures in the example set.

The automatic learning of scene classes is very important for a practical system. The system must be flexible enough to identify the particular class a user is interested in. Two users of the system may partition the images into very different categories depending on the class of images they wish to retrieve. An image retrieval system must be able to discern the biases and preferences of the user with respect to the class of images he is trying to access.

There are several ways for a user to indicate his preferences without handcrafting a template himself. The user may show several examples of images which fall into the desired class. The user may also indicate the salient regions which should contribute to the class concept in each of the example images.

The learning process does not have to terminate in only one pass. The system may generate an initial template based on the information provided by the user and can then retrieve images based on this rough estimate of the class. The user in turn can rate the returned images using a range from "very desirable" to "completely undesirable". Based on these responses the system can refine the class concept so that the images it retrieves are more compatible with the expectations of the user. The refinement process may be stopped at any time, when one or more acceptable images have been retrieved.

In this chapter, we discuss three techniques to learn a class concept in the form of a qualitative template. The techniques are variants on a correlational type of learning algorithm to extract the relevant consistent attributes over a set of examples. These algorithms are similar in part to those described by Sinha in [57]. First, however, we briefly discuss why the problem of learning in this context is computationally difficult.

5.1 Difficulty of the problem

To generate a model which encodes the consistent relationships between salient regions, an automated system must identify the salient regions in each of the images, compute a correspondence between these regions over the example set, and determine the consistent relationships.

The problem of learning the model from a set of examples is difficult in part due to the amount of variation between the example images, the amount of variation within a single image, the amount of noise or irrelevant attributes, and also the size of the images. These problems manifest themselves in both the partitioning of the regions and computing inter-image region correspondence. For instance, image regions that are perceptually grouped by humans may not have uniform quantitative color and textural properties throughout the region. It is, thus, difficult to determine when the variations between pixels are significant such that the set of pixels should be divided into several image regions rather than grouped into one image region. In addition, the corresponding salient regions in one image may have very different quantitative characteristics in another image. For instance, the sky in field scenes may be cloudy and textured or blue and without texture. Even if the sky regions in each of the field images were correctly partitioned, it would be difficult to decide if they have the same label or, in other words, if they correspond to each other. In general, there are an exponential number of pairings between regions in one image and regions in another image. Finally, the segmentation and the correspondence are dependent on the class concept. For instance, if the desired class concept is a field with sky and clouds, then the image should be partitioned into regions corresponding to these labels. However, if the desired concept is field with sky, then any white cloud regions and blue sky regions should be grouped together into one larger region. This problem can not be addressed in a purely bottom up manner or from the image information alone. The segmentation and resulting correspondence may have to be altered based on feedback from the user.

In order to learn a scene concept from a set of examples, the issues of partitioning an image into regions and computing region correspondences between a group of images must be addressed. In addition, for these strategies to be used in a real system, they must be computationally efficient. In the next section, we describe some techniques which address the problems of region partitioning and region correspondence to derive qualitative models of scene classes from a set of examples.

5.2 Learning the class templates from a set of example images

The task of learning a class model is as follows: given a set of example images, to extract the relative relationships and quantitative information that is consistent amongst the example images and is relevant to the class concept. The goal is to create a qualitative model which is derived from the low frequency components of the example images. We discuss three approaches to solving this problem.

5.2.1 Learning without region grouping

The first approach to solving the problem of learning a class concept from a set of examples ignores the problems of region grouping and, therefore, treats each pixel in the low resolution images as a distinct region. The algorithm computes all pairwise qualitative relationships between each image pixel and based on this information attempts to compute a correspondence between regions with consistent relationships across the set of example images. The goal of this technique is to determine the consistent relationships in a robust and computationally efficient manner while allowing for some positional variance between the corresponding regions.

There are several steps to the learning algorithm:

The first step creates low resolution versions of the full resolution images by extracting the low frequency components. As described in section 4.8.2, this can be achieved by blurring and subsampling the set of example images. The pixels in the resulting smaller images represent space averaged image regions from the full resolution images.

The second step, for each of the low resolution images, computes all the pixel pairwise relationships. Each of these relationships is encoded as an n dimensional vector, where n is equal to the number of attributes. The attributes may, for instance, include relative color and luminance. At this point it is not necessary to compute the spatial relationships due to the fact that the positions of both pixels in a pair are known.

For each pixel, we also compute a rough estimate of its color from a coarsely quantized color space, as a measure of perceptual color. In this implementation, we map the full R,G,B color space to 27 equally sized partitions. This is achieved by mapping the range of each color component (0-255) to three values (0,1,2). The estimate of color is "attached" to the high dimensional relationship vector.

At this point in the algorithm, for an image of size m, there are on the order of m^2 high dimensional vectors, m for each pixels. Figure 5.1(a) shows for one pixel its pairwise relationships encoded as vectors and quantized color. It is not necessary, however, to retain all the m relationships for that pixel. Instead, the image can be grouped into directional equivalence classes with respect to that pixel (notice that an equivalence class does not imply a space averaging of the underlying regions). In Figure 5.1(b) we show one example of this grouping. This step has two beneficial effects. The first is that relative spatial relationships, such as "above" and "below" have been introduced. The second is that the we can eliminate the redundant relationships in each equivalence class, thereby reducing the number of relationship vectors associated with each pixel. This is shown pictorially in Figure 5.1(c). The steps of grouping the image into equivalence classes and computing the non-redundant relationships in each equivalence class may be done for each pixel in the image.

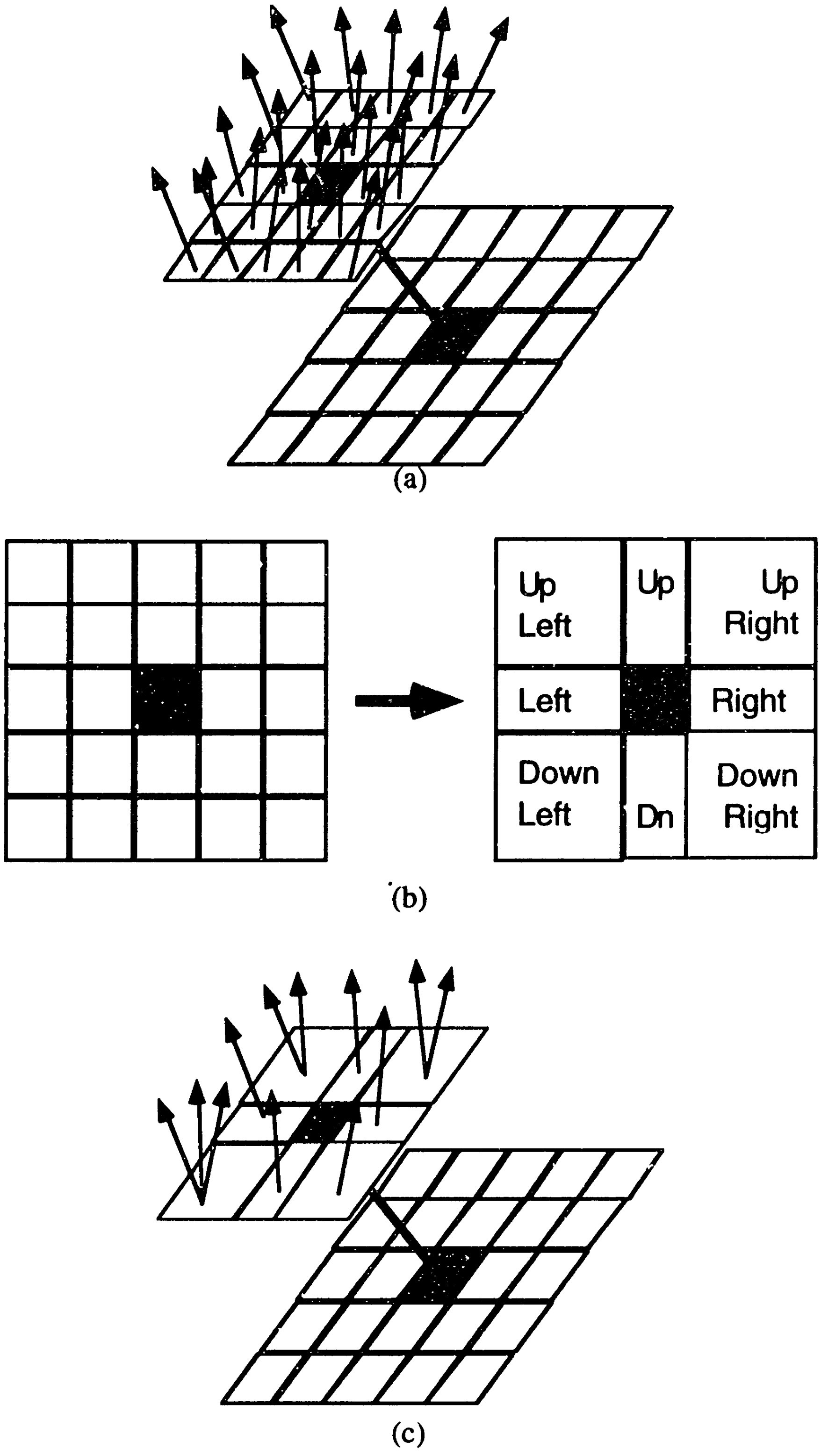


FIGURE 5.1. *(a) Relative relationships to all other pixels (described as vectors) and quantized color shown for one pixeL (b) Example of directional equivalence classes. (c) Illustration of the reduction in redundant relationships via the use of directional equivalence classes.*

Each image in the example set may be processed in the manner described above. The next step is to compute the consistent set of relationships for each region in each image. There is, however, a problem in that the correspondence of pixels across images is not known. To compute the consistent relationships, it would not be meaningful to compare relationships between different regions such as a sky pixel in image 1 and a grass pixel in image 2. It is possible to look at all pairings of pixels across images and try to determine the set of pairings that have the greatest number of consistent relationships and colors. Hopefully, the set of consistent pairings will be between corresponding image region. This strategy, however, is computationally expensive because there are an exponential number of such pairings.

A seemingly reasonable assumption that overcomes this problem is that corresponding pixels in each image are likely to occur in approximately similar positions. Thus, we assume that a pixel in image 1 located at position (x,y) will most likely correspond to a pixel in image 2 at position $(x \pm i,\ y \pm j)$, where i and j are parameters.

We can incorporate this assumption in our processed images by having each pixel inherit the relative relations/color from its neighbors. The extent of the neighborhood is variable. It also is possible to weight the importance of the attributes from the neighbors of a pixel as an inverse function of their distance from that pixel. This step is somewhat analogous to smoothing the processed images.

To determine the set of consistent relationships, we now need only compare the set of relationships/colors at each pixel location across all the example images. The example images can be summarized by one image which contains all the relationships/colors for each pixel location and the frequency with which they occurred. The frequency of occurrence of the relationships and associated colors is a measure of their consistency.

The summary image may be used intact as a class model or it may be translated into a "graph" structure similar to the models shown in chapters 3 and 4 by looking for pairs of pixels in the summary image that have inverse relative relationships. To convert the summary image into a graph structure, it is first necessary to choose a consistency threshold to eliminate relationships and nodes that are not important to the model. Nodes that remain after the thresholding steps become vertices in the graph structure. Nodes with inverse relationships can be connected via directional edges. For instance, the graph structure should contain an edge between node *i* and node *j*, if node *i* specifies that it is more blue, less red, and less green, than a node which is below it and node *j* specifies that it is less blue, more red and more green than a node which is above it.

It may be desirable to keep the summary image for further refinement of the model. The initial model may be used to retrieve images. The user may select a subset of the returned images and ask the system to update the model by increasing or decreasing the consistency of already encoded relationships or by adding new relationships.

The measures of consistency of regions encoded in the processed model are equivalent to the thresholds used in the prior chapter to determine if the relationships between two image regions were "sufficiently" satisfied. This algorithm computes what the tolerable thresholds are for making that determination.

The algorithm as described is computationally efficient. The preprocessing steps of smoothing and subsampling the image, computing the pairwise relationships between low resolution image pixels, reducing the number of pairwise relationships by computing directional equivalence classes for each pixel, and allowing each pixel to inherit its neighbor's attributes are all steps which can be computed in $O(m^2)$ time, where m is the number of pixels in the image. These steps may performed off-line for each image in the database. The step of computing the consistent relationships, which must be done at the time of the query, is only $O(m)$, since each pixel in each image is considered once in creating the class model. m, for our purposes, is generally small, since we use greatly subsampled versions of the original images, ranging from *(8x8)* to *(32x32)* pixels.

5.2.1.1 Testing the approach on synthetic images

We first tested the approach by generating 25 synthetic images of size *(8x8)* pixels. As described in the previous chapter *(8x8)* is not an unrealistic image size for natural scene classification. Each pixel in the synthetic images was given a random color. A three patch qualitative concept, also generated randomly, was embedded in each image. The absolute colors and positions of the patches in the concept were allowed to vary as long as the qualitative color and spatial relationships were not violated. A different instance of the qualitative concept was embedded in each image. Figure 5.2(a) shows four example images from the synthetic example set of 25 images. The concept patches are indicated by white outlines. The goal was to test the learning approach to see if the concept could be recovered.

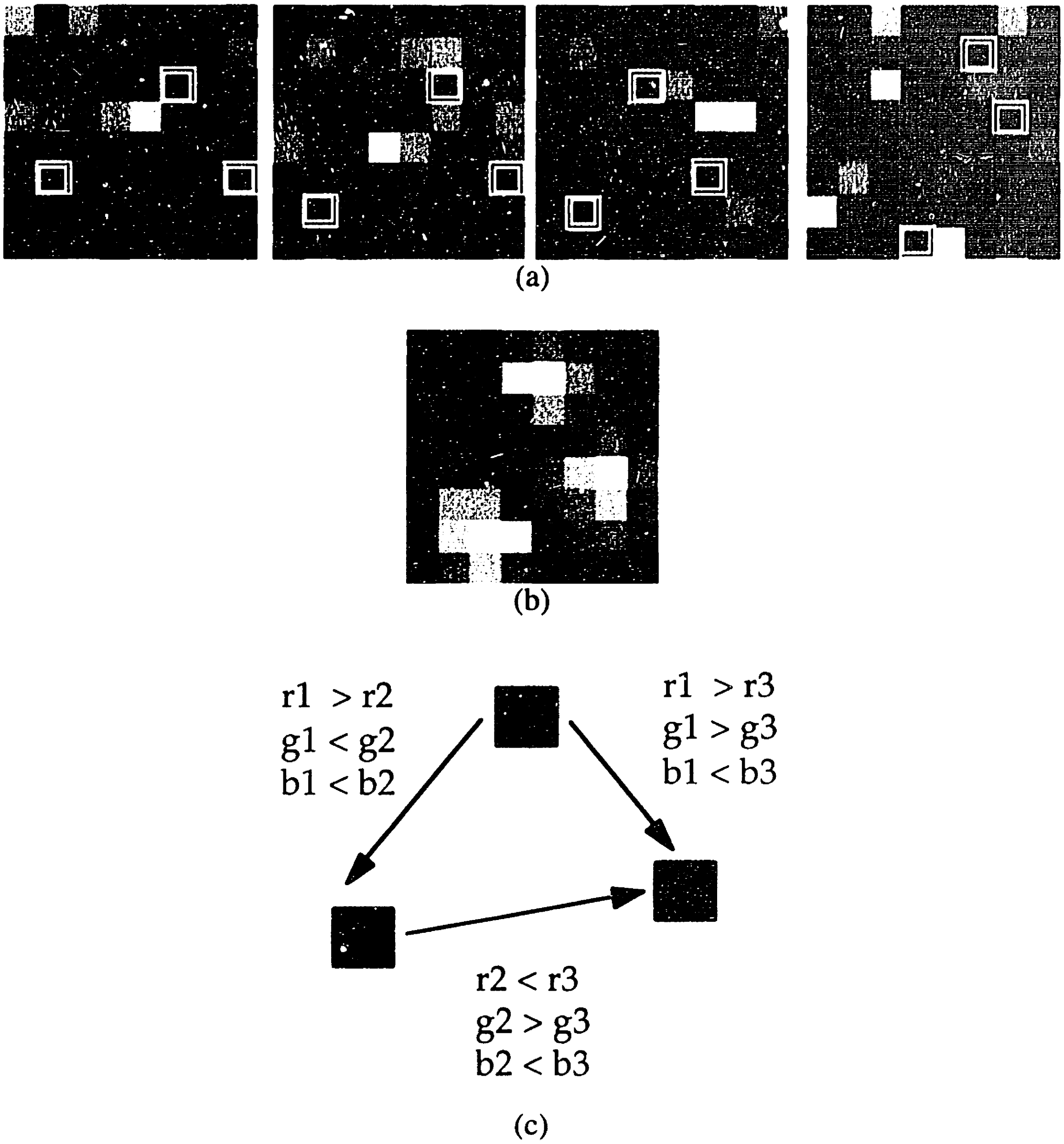


FIGURE 5.2. *Demonstration of the learning algorithm on synthetic images. (a) Four sample synthetic images which consist of randomly colored patches are shown. The three patches of the qualitative concept embedded in the images are highlighted in white. The spatial locations and colors of the patches were allowed to vary in each image as long as they satisfied the relative constraints. (b) The summary image shows the consistency of the patches across all the sample images. Black denotes no consistency. White denotes 100% consistency. (c) The relationships between the three 100% consistent patches from the summary image are shown as a qualitative model. The model matches the embedded concept.*

Figure 5.2(b) shows the summary image generated from the implementation of the algorithm. The patches in the image range from black to white, where black denotes no consistency and white corresponds to total consistency. The image indicates that there were three patches in each example image which had 100% consistent relationships to other patches. Note that the three consistent patches in the summary image are not in exactly the same locations as the qualitative concept patches in each image. The locations of the patches in the summary image represent the average location of the concept patches across the example image. Figure 5.2(c) shows the graph model derived from the three patches and their consistent relationships. The graph is identical to the original concept.

5.2.1.2 Testing the approach on real images

The results of testing the approach on real images suggests that major modifications are required to attain the level of performance obtained with the synthetic image database. There are several reasons for this. The first is that each image pixel should not be considered as a candidate for a node in the model graph. In addition, our assumption of corresponding regions occurring in localized areas of the example images is somewhat unrealistic for real images.

The algorithm works well on the synthetic images because each image pixel does represent a distinct image region. Therefore, it is realistic that each image pixel initially should be considered equally important to the class concept. In real images, such as field scenes, image pixels will not cover an entire perceptually coherent region. For instance, even in an *8x8* pixel field image, the entire top row or the first 8 pixels is likely to correspond to the sky. The algorithm as stated will give equal importance to each pixel in the sky region. These pixels will likely have the relevant attributes that they are bluer than the majority of the pixels below them, both directly below and below and to the right and below and to the left, and each sky pixel may have the relationship that it is just as blue as many of the pixels around it. These relationships are accurate for field models. However, they are also accurate for the majority of natural images that contain sky. Figure 5.3 shows a low frequency and full resolution version of a field and a coastal image. Many of the relative relationships between the low frequency regions in the field image will also be valid in the low frequency coastal image.

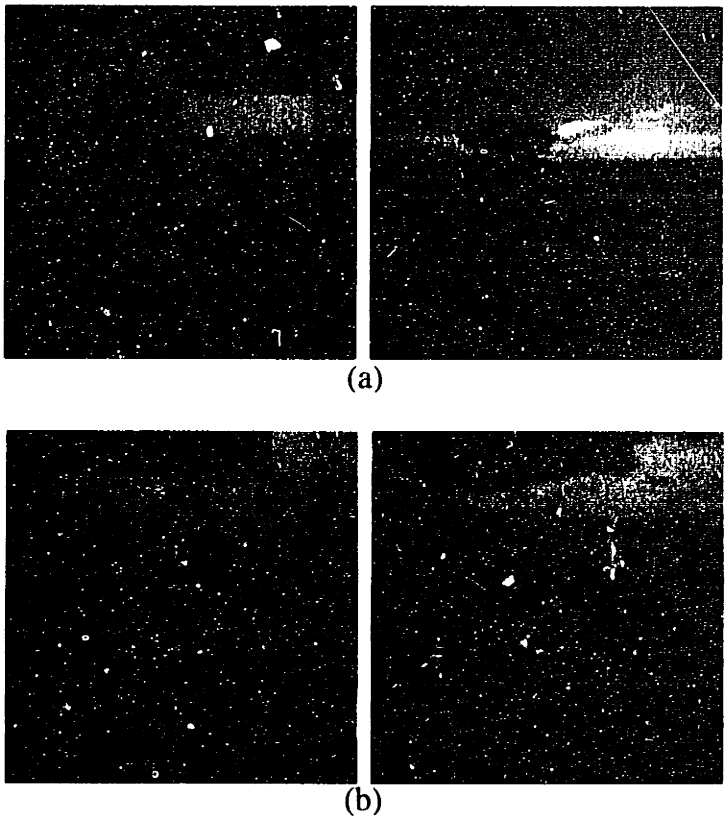


FIGURE 5.3. *(a) A low resolution and full resolution field image. (b) A low resolution and a full resolution coastal image. Many of the relative color relationships between low resolution image regions in the field scene are also satisfied in the coastal scene.*

In general, the discriminating features, such as field regions which specify that they should be more green or red that the regions above them, may be drowned out by the overwhelming consistency of the sky regions with respect to the other image pixels. Even the relationships which do encode that there should be regions which are more green than those above them, may exist to some extent in images from other classes. In general, we found models generated via the algorithm had a very slight discrimination ability between different classes of natural images. The difference in discrimination was not sufficient to accurately classify novel images as examples or non-examples of the class covered by the models.

The second problem is that our assumption that image features will be found in a localized area in every example image is not valid for some classes of natural images. For instance, the regions corresponding to the stream of water in a waterfall scene may be in any column of the image and comprise of a variable subset of the column's pixels. To fully generate a waterfall template using the framework of the algorithm it may be necessary to compare more than just localized regions for consistency. Figure 5.4 shows three waterfalls with varying positions and orientations.

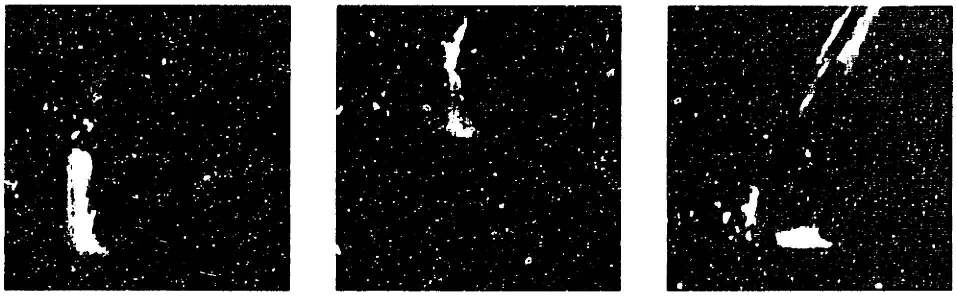


FIGURE 5.4. ***Three waterfalls with significant variations in location and geometry. The assumption that image features will be found in a localized area across example images is not valid for this set.***

Our results suggest that the algorithm may be greatly augmented by a region grouping step and a better correspondence algorithm in order to build more concise and accurate scene models. This idea is supported by the observation that humans tend to establish relationships between extended regions that have homogeneous properties. In the next section, we suggest a novel strategy for region grouping. We also suggest how this strategy may aid the correspondence problem.

5.2.2 Learning with region grouping

Region grouping may help us eliminate some of the problems that the previous algorithm faces. In general, region grouping allows us to specify that pixels in one region should be considered together and that the group of pixels should have some common relationships to each other and to other pixels outside the group. For matching purposes, most of the shared relationships associated with the grouped pixels should apply in order for the model region to match a subpart of a novel image. Region grouping also specifies that regions that belong to one group should generally occur together within a contiguous area. Models generated by the previous algorithm would allow that two sky regions be matched to very different spatial positions in a novel image. For instance one sky region may be matched to a valid sky region at the top of a novel image, while a second sky region could be matched to a lake in that novel image, as long as there was something less blue below the lake.

Many algorithms have been developed to segment distinct regions in an image based on their quantitative color, textural, and shape attributes. These algorithms applied to natural scene classification are generally not robust due to the fact that regions often differ greatly in their quantitative attributes [27][43][50][55]. However, such segmentation, even if imperfect may be suitable for our purposes.

We suggest that region grouping via quantitative attributes may be augmented by the use of qualitative intrapatch relationships and qualitative interpatch relationships. Such measures may provide some invariance to differences in quantitative attributes. In addition, grouping image regions which have consistent relationships among themselves and to other regions is directly compatible with the goal of generating a qualitative model which encodes sets of image regions which have consistent relationships to other sets of image regions.

We have developed a novel perceptually motivated technique of partitioning up the RGB color space. Each partition of the color space represents groups of colors which may be perceptually equivalent. For each image region, we compute a comparison between the red and green, the green and blue, and the red and blue color components. This is similar in spirit to the color opponency found in the early stages of the primate visual system [33]. For viewing purposes, we remapped the red channel to 0, 128, or 255 if the red color component respectively was less then, equal to or greater than the green color component of the patch. The green and blue channels can be remapped in a similar manner, where the green channel reflects the comparison between a patch's red and green components and the blue channel reflects the difference between a patch's blue and red components. Note that there exist other ways to remap the red, green, and blue channels which may better reflect the intrapatch relative color. In addition, we may quantize the color space in a finer manner. For instance, we found that colors especially which lie along the luminance axis should be more finely represented in the color quantized space. Currently, the RGB color of value (0,0,0) which corresponds to black and the value (255,255,255) are mapped to the same value (128,128, 128). Perceptually, however, these two colors appear very different. In terms of scene classification, would not want to determine that a dark island in the middle of a blue sea and a white iceberg in the middle of blue sea are similar in terms of their color attributes.

Figure 5.5(a) shows nine full frequency field images. Figure 5.5(b) shows the low resolution versions of the images remapped into a relative color space. Even though the sky and field regions vary in terms of their quantitative color both within the images and across the images, each of these regions appear to have distinct homogeneous quantized colors. The colors for each of the regions are similar across the example images. We have found similar results for other classes of natural images such as snowy mountains and sunset scenes.

The results suggest that region grouping may be significantly easier using measures of intrapatch color rather than quantitative color. We are currently investigating the idea of grouping regions which have similar relative relationships to other image patches.

Grouping the image regions may significantly decrease the complexity of computing the correspondence between regions in example images. First, the number of regions to be considered is be reduced from the total number of pixels in the image to a much smaller number. Additionally, from our initial experiments, corresponding regions across images of the same class may have the same or similar relative intrapatch attributes. These issues need to be explored in a more detailed manner. Once we have computed the corresponding regions across the set of example images, we can determine which relationships are consistent between those regions in each image. We can then generate a qualitative model which consists of the common regions in each image and their consistent relative attributes.

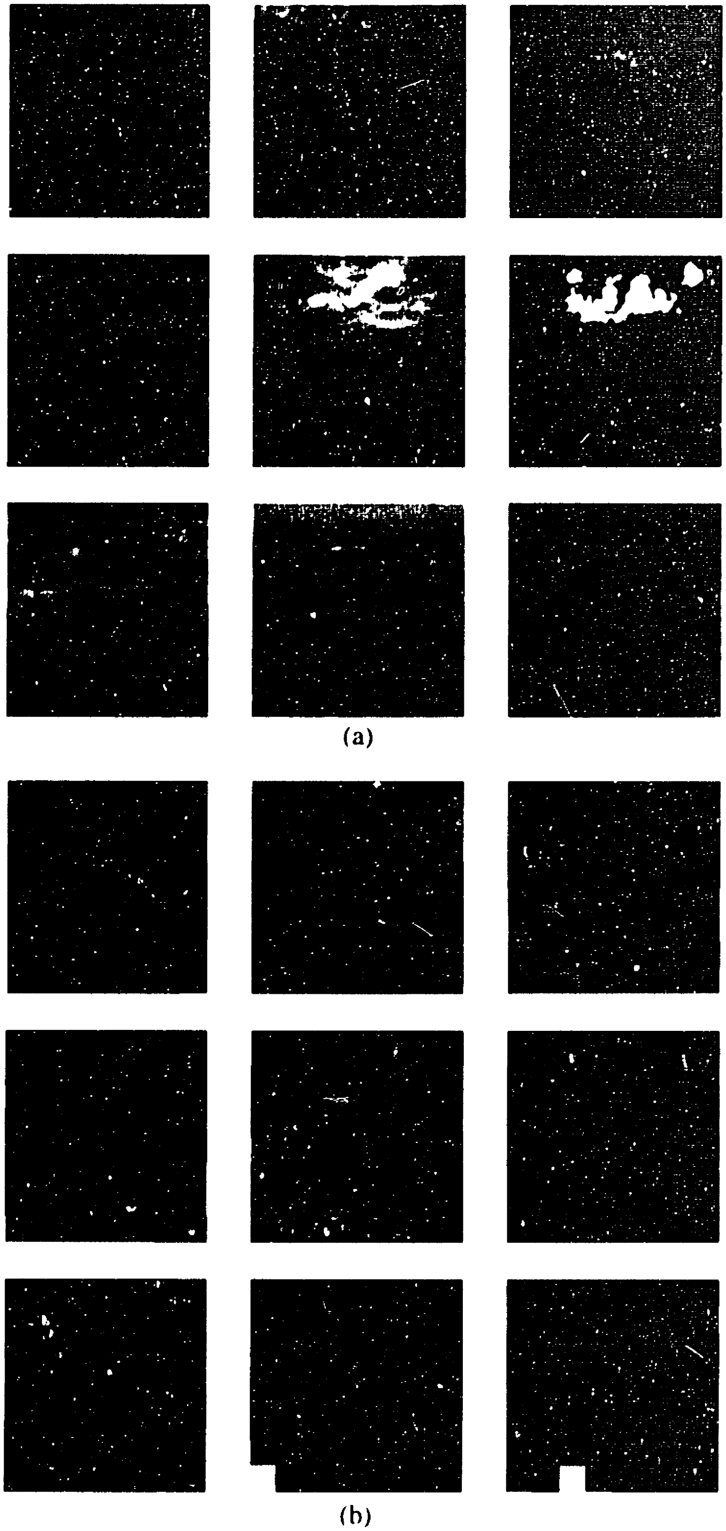


FIGURE 5.5. *This figure shows 9 field images (a) in full resolution and (b) in the low resolution color quantized representation.*

5.2.3 interactive region grouping and region correspondence

Instead of totally automating the process of region grouping and region correspondence, the user may provide some of this information. We are developing an interactive system in which the user selects a few salient regions in one or more example images. The system, using coarse similarity criteria, retrieves some images consistent with the user's initial specification. The regions are implicitly in correspondence across the images. The user can score the retrieved images. A correlational learning process determines the consistent relationships between the salient regions across images by taking into account the user's rating of the images.

The learning process may proceed in a systematic manner to try to extract the relevant relationships between the selected regions. On each pass of the algorithm, the system may try to determine the relevant relationships across one attribute class by allowing that attribute to vary in the selected images. For instance, on a first pass, the system may provide images that have regions with similar spatial relationships and slightly varying colors. Based on the response of the user, some of the valid color relationships may be determined. In a second pass, the system may return regions with similar colors to the original input regions, but in different spatial arrangements. As the process continues, the consistent relationships over the set of attributes may be used to generate a qualitative class model. The qualitative model may then be used to probe into a large database.

Figure 5.6 shows an example of the interactive learning process for a snowy mountain image. The goal was to try to automatically develop a template that could classify the set of snowy mountain images from this example. Figure 5.6(a) shows the example input image with three salient regions highlighted in red. Figure 5.6(b) shows the low resolution version of the image (8x8 pixels). Figure 5.6(c) shows the low resolution color quantized version of the image. The color quantized image was created by comparing the intrapatch color components. The quantization was computed in a slightly finer manner than described in the previous section. For instance, the luminance axis has been quantized into several distinct colors, rather than just one.

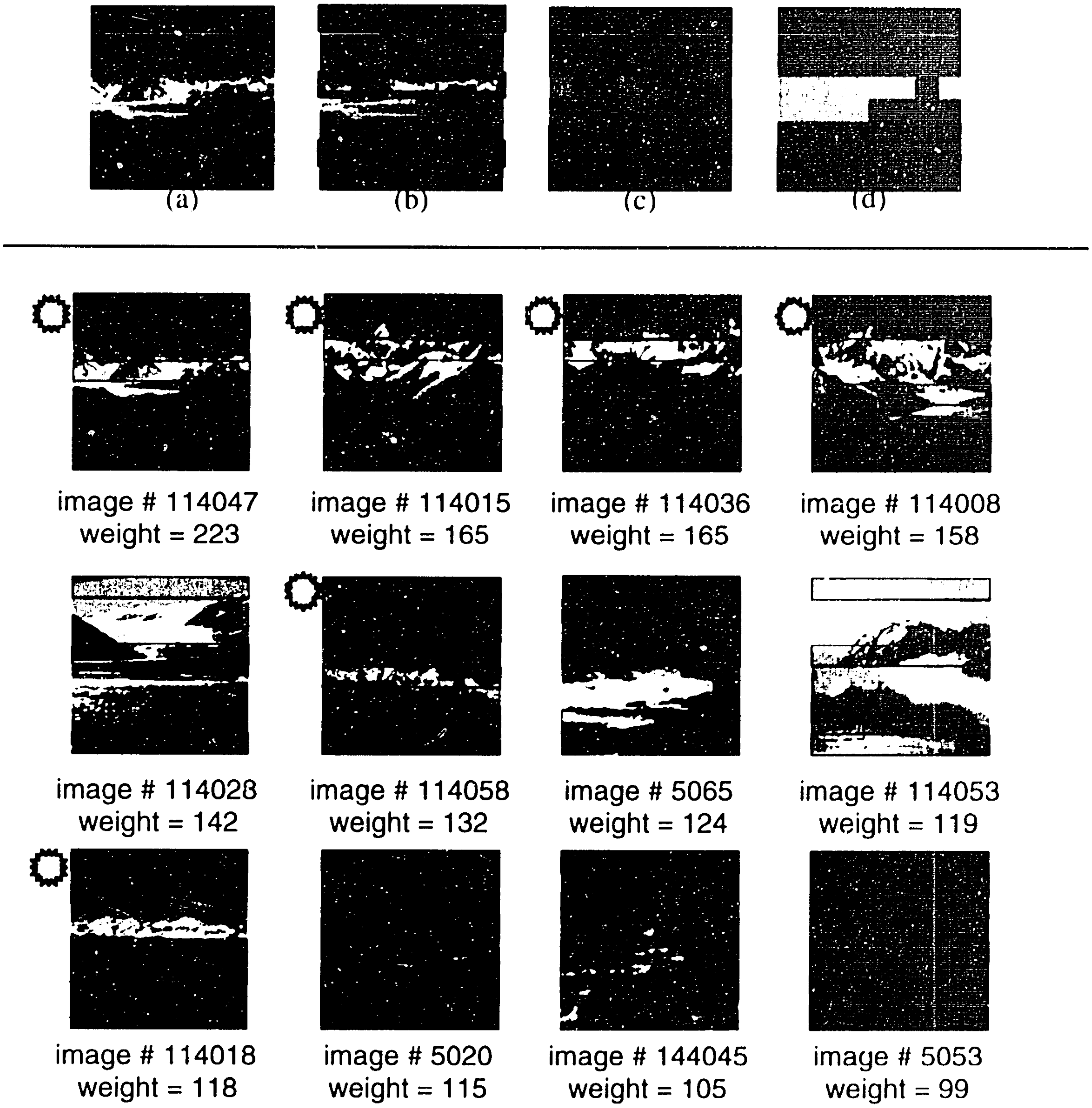


image # 114047
weight = 223

image # 114015
weight = 165

image # 114036
weight = 165

image # 114008
weight = 158

image # 114028
weight = 142

image # 114058
weight = 132

image # 5065
weight = 124

image # 114053
weight = 119

image # 114018
weight = 118

image # 5020
weight = 115

image # 144045
weight = 105

image # 5053
weight = 99

FIGURE 5.6. *Example of a first iteration of the interactive learning process. (a) shows the query image. (b) shows the salient regions marked by the user. (c) shows the low resolution version of the query image. (d) shows the color quantized version of the low resolution image. The images below the line are the result of a first query based on the salient regions in the input image. The corresponding regions found by the algorithm are denoted on each image in red. For each image, its number image and its similarity score to the input image are reported. The seals denote the images which were chosen by the user for use in refining the class model.*

The algorithm uses a weighted combination of the color quantized values of the salient regions and the color quantized values of the background to select a few images from the natural scene database of 700 images which may be similar to the input image. In this first iteration, the positions of the patches were allowed to vary slightly as long as they maintained their relative spatial relationships. The goal was to retrieve images that were similar in spatial arrangement to the input image but had a varying range of absolute colors. Figure 5.6 shows the twelve images with the highest similarity score to the input image. The image numbers and similarity scores are shown below the images.

At this point, the user may rate the input images. The returned images which were snowy mountains and had the salient patches in roughly the right locations are indicated with a seal. This information can be fed back into the system in order to extract the consistent relationships between the regions in each of the chosen images.Table 8 shows the initial qualitative model based on the input image in Figure 5.6. The consistency measures in terms of percentages for the relationship pairs over the color attributes are shown. The patches 0, 1, and 2 correspond to the three patches in Figure 5.6(b) starting from the top. The possible values over one attribute are patch *i* < patch *j*, patch *i* = patch *j*, patch *i* > patch *j*. Given this initial model and the input image, the system may again probe the database, return some example images, and based on the user's response revise the model.

TABLE 8. Initial qualitative model based on the regions in the query image in Figure 5.6

[illegible]	[illegible]	[illegible]	Green	Relation-ships
[illegible]	<	100%	83%	44%
	=	0%	17%	33%
	>	0%	0%	23%
[illegible]	<	0%	0%	0%
	=	8%	0%	0%
	>	92%	100%	100%
[illegible]	<	0%	0%	0%
	=	0%	0%	0%
	>	100%	100%	100%

5.3 Implementing learning in a practical system

Figure 5.7 shows a database search/template learning protocol for a practical indexing system, irrespective of how the model is generated (e.g. no segmentation, automatic segmentation, or user specified segmentation of regions). The basic approach is to probe a small number of images on the first iteration of learning the class model. As the model is further refined, the system may probe deeper into the large database of images. At the end of the learning process, the refined model may be used to retrieve images from the rest of the database or other databases.

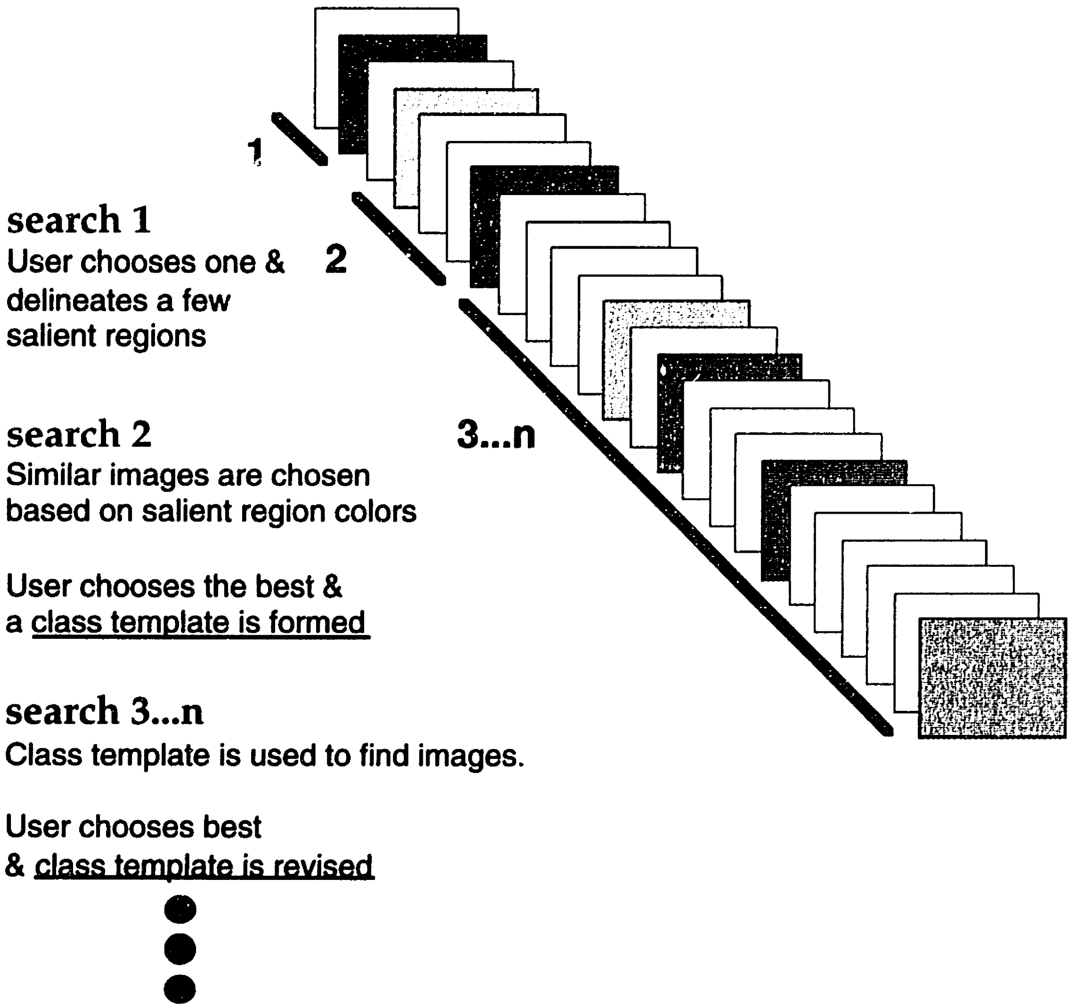

1
2
3...n
search 1
User chooses one & delineates a few salient regions
search 2
Similar images are chosen based on salient region colors
User chooses the best & a class template is formed
search 3...n
Class template is used to find images.
User chooses best & class template is revised

FIGURE 5.7. *Database search/template learning protocol for a practical indexing system*

5.4 Discussion

In this chapter, we have discussed several techniques for learning a scene class from a set of examples. From our experiments, we found that an initial region grouping step was highly beneficial for determining the relevant consistent relationships over a set of images. We discussed how region grouping may be performed via quantitative and qualitative information. In an interactive learning approach, we demonstrated that a qualitative model can be generated based on corresponding regions in a set of example images. We suggested how this model may be refined through feedback from the user.

There are several other issues that need to be addressed in learning a scene class from a set of examples. Our learning technique builds up a union of the most likely relationships between salient image regions. We need to learn the appropriate quantitative bounds on those relationships to limit the scope of the model. We also need to expand our techniques to learn disjunctions of relationships (e.g. the field region should be browner or greener than the sky region). All of our learning techniques incorporate information from positive examples. However, we may fruitfully use negative examples in the learning process. They may be used to better delineate the boundaries between classes. They may also be used generate negative nodes or negative relationships in the class model which explicitly encode that these nodes/relationships should not be found in images which are instances of the class.

CHAPTER 6 Conclusion

In this thesis, we presented a novel approach to classifying scenes. The goal of scene classification is the following: given a specification for a scene class, either as a symbolic query or as a set of example images, to retrieve other images which would perceptually be categorized as instances of the class. Automated scene classification techniques are eminently suited for the applied problem of image database indexing. With the increase in the number of digital libraries, there is a need for automated comprehensive annotation and robust indexing systems.

One possible method for classifying scenes is to try to recognize each individual object in the scene and then based on some inference procedures evaluate the type of scene the image represents. This solution is not attractive because object recognition is a difficult problem, especially when the objects are embedded in a complex scene. One of the main questions we attempted to address was whether scene classification could proceed without a prior step of individual object recognition.

In the first chapter we discussed the problem of scene classification. Scene classification is difficult because members of one class may differ greatly in their color, texture, illumination, viewing position, and geometry. Our goal was to try to establish what characteristics of a scene best describe the content of the scene and whether these characteristics could be used to capture the essence of a scene class.

In chapter 2, we discussed how global spatial configuration is important for scene classification. We motivated our approach by showing several examples whereby changing the spatial configuration of the scene changed the perception of the scene. In this chapter, we suggested that scene classification could proceed before object recognition. Therefore, scene classification could be based on image regions, which are viewed as two-dimensional patches with spatial, photometric and texture attributes, rather than semantically meaningful image parts. We argued that neither absolute spatial configurations or cumulative statistics of these regions was necessary or sufficient for scene classification. Instead, we suggested that relative spatial organization may be critical for models of scene classes. Additionally, we suggested that image details were not crucial for image classification and therefore may be unnecessary for most class models.

In chapter 3, we discussed that other relative measures of image patches might be important to describe scene content. For instance, we suggested that the absolute photometric values of image regions were not crucial to the description of a scene class, but rather the direction of their contrast was important. In the same chapter, we presented a detailed description of qualitative models.

In chapter 4, we presented a novel approach to scene classification, entitled "configural recognition". This approach draws heavily from the inferences made in the prior two chapters. The approach to classification uses global scene organization, proceeds directly on the image without the need for complex abstractions, and represents class models as sets of qualitative relationships between low frequency image regions. We described how such models can tolerate many commonplace scene distortions, such as those from changes in scale, illumination, viewing position, and geometry.

Our approach differs significantly from the prior work in the area of scene classification. Most of the prior work uses quantitative image statistics such as color, texture and shape to compute image similarity. Most of these statistics are based on high frequency image detail. We showed several demonstrations that the quantitative nature of these approaches could not overcome the problems of general scene classification and therefore in most cases were not good measures of scene content.

We discussed how the configural recognition technique could be applied effectively to the problem of natural scene classification. We demonstrated our approach by hand-crafting four natural scene class models, one for snowy mountains, snowy mountains with lakes, fields, and waterfalls. We tested each of these class models on a database of 700 natural images and compared the results to human perceptual judgements. We found that the templates had an impressive ability to generalize over a large perceptual class. The templates were able to discriminate between many images of different classes, some of which had the same color and textural characteristics but in incorrect configurations, resulting in low false positive rates. We showed that the template matching process was computationally efficient due to the use of low resolution information and simple models. Additionally, we suggested how other attributes such as texture could be incorporated into the models to classify other types of scenes containing man-made objects.

In chapter 5, we discussed how such class models could be learned automatically from a set of examples. We presented three approaches for learning a class model. We also presented a novel technique for region grouping based on relative intrapatch relationships.

In summary, the thesis makes four contributions.

• We introduced qualitative image representation strategies that allow us to capture meaningful scene content, which has not been demonstrated in the prior work. The representation strategy also provides a robust metric for inter-image similarity.

• The approach presented is able to incorporate global scene configuration in a manner that allows subsequent generalization to other members of a scene class. In addition, the approach is computationally efficient.

• We demonstrated the approach in a practical system for natural scene classification.

• We also described methods to learn qualitative scene classes from a set of examples.

Although configural recognition appears to be a promising strategy for the difficult problem of scene classification, we need also to be aware of its limitations. For instance, the technique is not suited to make fine quantitative discriminations, such as between different types of mountains or various varieties of grass. The technique is not well suited to describe classes of functionally defined objects. Additionally, the technique is not able to classify scenes which depend on object recognition, such as office scenes or living rooms.

In the next chapter, we discuss how the configural recognition approach may be applied to other domains.

CHAPTER 7 Looking Ahead

The success of the configural recognition approach for the problem of natural scene classification suggest some interesting ideas for future research. In this chapter, we discuss how scene classification via configural recognition can aid object recognition in a complex environment. We also suggest that the configural approach may be adapted for object classification.

7.1 Object recognition in scenes

The goal of much computer vision research is to identify objects in scenes. The scenes may contain multiple objects in a variety of configurations and under varying viewing parameters. Traditional approaches to the problem of recognizing objects in a scene search the whole image for salient regions which might contain one object. This action is often described as "focusing the attention" of the visual system. The focus of attention mechanisms usually involve looking for regions in the image that contain colors or geometrical features that might belong to a single object or a specified target object.

Many researchers have suggested that if the context of the scene is known, the complexity of the object recognition problem is greatly reduced in at least two ways [47][60]. First, the type of scene dictates the kind of objects that are usually found in such scenes. For instance, beach scenes may contain beach balls, beach chairs, and people, but rarely do they contain objects such as planes, sofas, or phone handles. Second, the scene also dictates where the objects associated with that scene might be positioned in the image. For instance, beach chairs are usually placed on the beach. They are sometimes found near the water. But they are rarely in the sky. Thus, if the type of scene is known, the range of objects and number of possible locations for the objects is constrained, thereby reducing the complexity of the object recognition problem (see Figure 7.1).

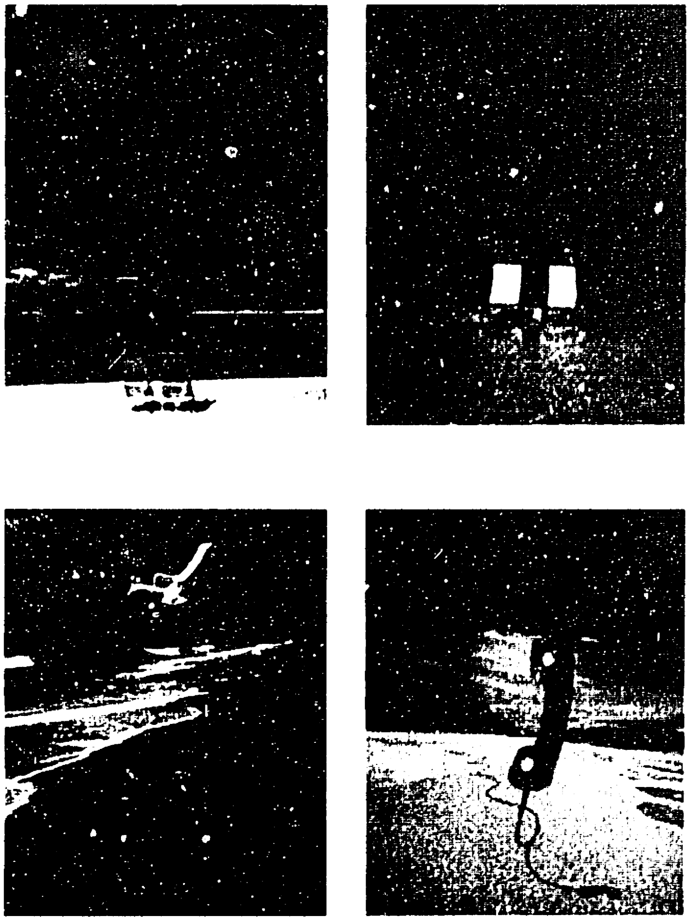


FIGURE 7.1. *The four pictures in this figure all beach scenes. One would expect to find beach chairs in the scene, but not telephones. Additionally, the beach chairs are most likely to be on the sand or in the water rather than in the sky. Studies such as those performed by Biederman have found that observer's recognition performance in flashed presentations is impaired by these incorrect spatial contexts[8][9]. Even though these pictures are not too complex in terms of their perceptual content, together they make the point that knowing scene context aids object recognition by reducing the set of objects likely to be found in the scene and also by reducing where those objects may be located in the scene.*

The idea of establishing scene context before object recognition is one of the themes of this thesis. The configural recognition approach to scene classification, therefore, might be used as a precursor to and also an aid to object recognition. Using the configural technique we can first classify a given scene, thus, providing us with the scene context. Secondly, the template which detects the scene can be used to label the regions of the scene, such as sky water and sand. The configural recognition approach to scene classification, thus, provides the information regarding the scene context *and* the approximate position of the some of the scene parts. With this information, the complexity of the problem of objecf recognition may be greatly reduced.

7.2 *Static object classification*

Scene classification and object detection may not be fundamentally different types of operations that require separate types of processing. Our approach to scene classification is to detect a qualitative pattern in low frequency regions over the majority of the image. Object detection strategies may use the same approach, but look for qualitative patterns in sub portions of the image.

Sinha has already demonstrated the use of qualitative models for detecting frontal views of faces under varying illumination conditions [57]. We suggest that qualitative view based models can be designed for other objects such as cars, people, planes, and steam trains.

Figure 7.2 shows an example of a qualitative model for side views of vehicles[1]. The qualitative model is coupled with geometry information specifying the shape of the wheels. The model is matched to subportions of low frequency images (*15x15* pixels). The wheel detectors are used as verification devices after the qualitative model has found a potential match. (The wheel detectors are applied to high frequency image data at the locations specified by the template).

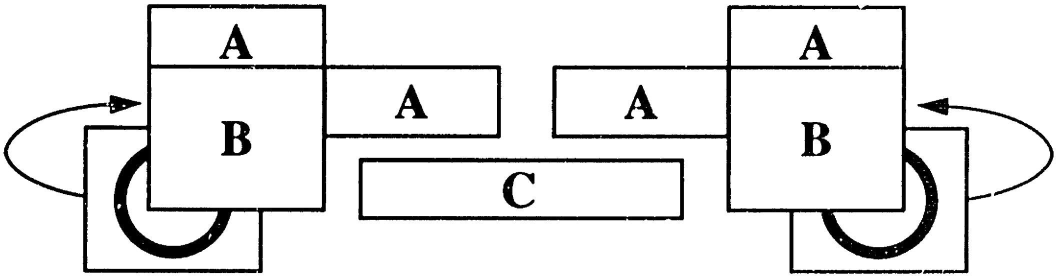


FIGURE 7.2. *Qualitative model of side views of vehicles. The model consists of 7 regions, classified into three types. The model specifies that the regions labeled A should be similar and different from the regions labeled B and C. In addition, the regions labeled B should contain circular regions.*

[1] The vehicle detector was designed and tested in collaboration with A. Lakshmi Ratan of the MIT Artificial Intelligence Laboratory and P. Sinha of the MIT Department of Brain and Cognitive Science.

The model was tested on over 100 images. Figures 7.3, 7.4, and 7.5 show some of the results.

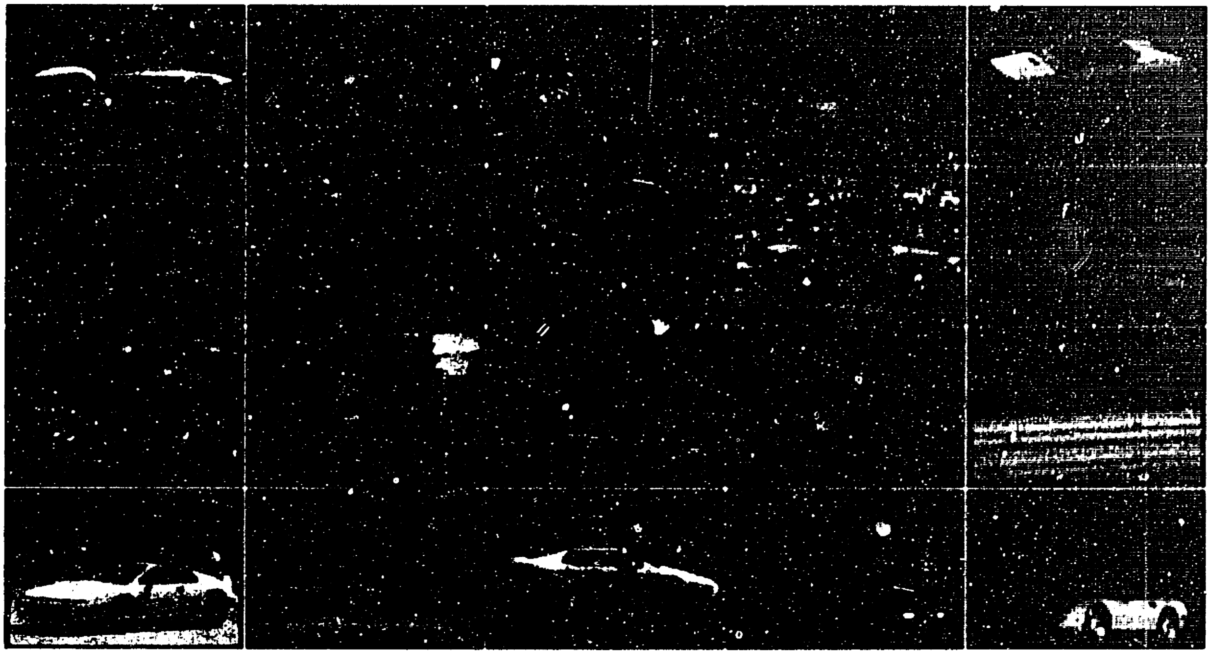


FIGURE 7.3. *Some true positives detected by the vehicle template.*

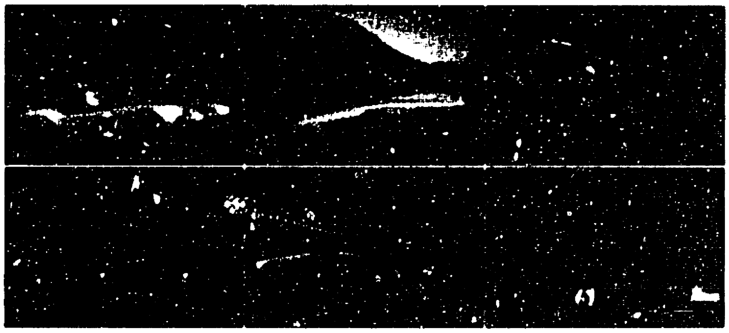


FIGURE 7.4. *Some false negatives not detected by the vehicle template.*

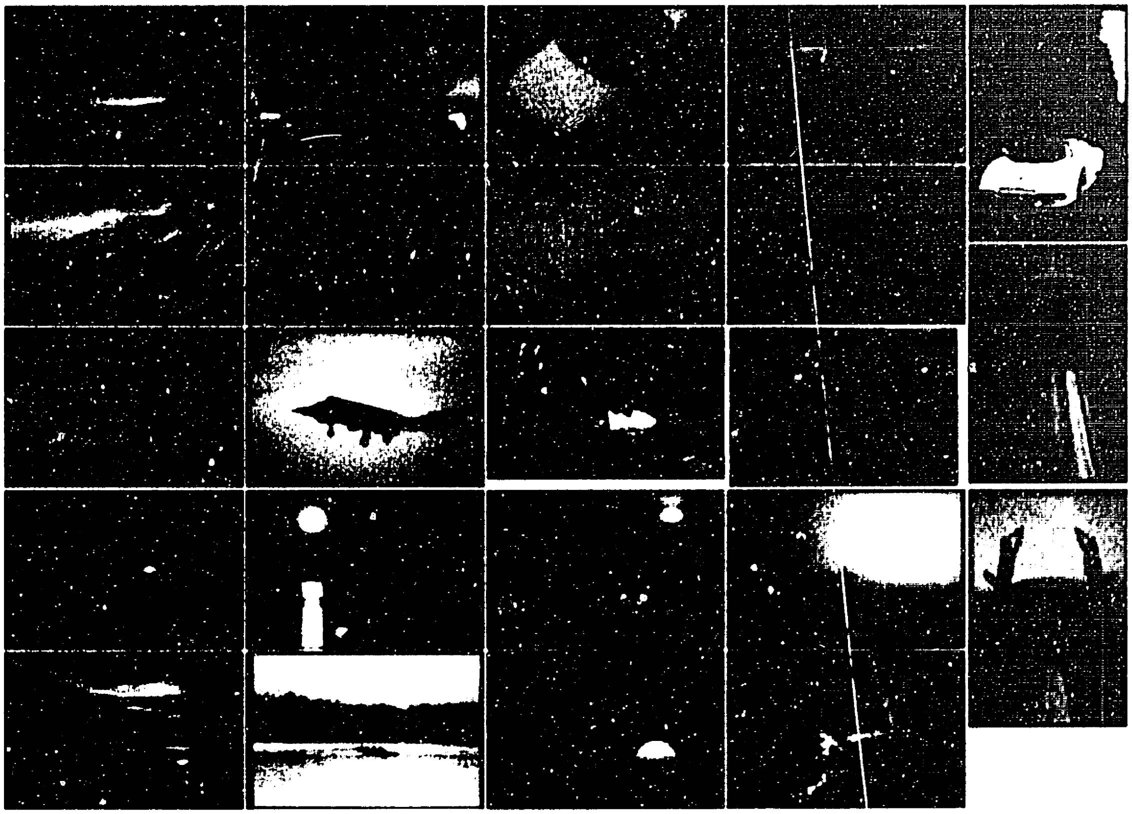


FIGURE 7.5. *Some true negatives not detected by the vehicle template. Non side views of cars are considered to be true negatives.*

Bibliography

[1] Abella, A., and Kender, J.R., "Qualitatively describing objects using spatial propositions". *IEEE Workshop on Qualitative Vision.* New York, June, 1993, pgs. 33-38.

[2] Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., and Ogden, J.M., "Pyramid methods in image processing". *RCA Engineer*, 29-6, Nov/Dec, 1984.

[3] Ang, Y.H., Zhao, L., and Ong. S.H., "Image retrieval based on mulitdimensional feature properties", *SPIE Storage and Retrieval for Image and Video Databases III.* San Jose, Feb., 1995. pgs. 4757.

[4] Ashley, J, Flickner, M., Lee, D., Niblack, W. and Petkovic, D, "Query By Image Content and Its Applications". Research Report, RJ 9947 (87906) , Computer Science/Mathematics, March, 1995.

[5] Barr, M. and Ullman, S., "Spatial context in recognition". *Perception*, vol. 25, pgs. 342-352, 1996.

[6] Betke, M. and Makris, N. Fast Object Recognition in Noisy Images Using Simulated Annealing. A.I. Memo No. 1510, MIT, December, 1994.

[7] Biederman, I., "Perceiving real-world scenes". *Science*, Vol. 177, pgs 77-80, 1972.

[8] Biederman, I., "On the semantics of a glance at a scene". In M. Kubovy & J.R. Pomerantz (Eds.), *Perceptual Organization.* Erlbaum, Hillsdale, NJ, 1981.

[9] Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C., "Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 1982, pgs. 143-177.

[10] Bober, M., Hoad, P., Mataas., J., Remangnino, P., Kittler, J. and Illingworth, J., "Control of perception in an active vision system: sensing and interpretation". *IROS*, 1993

[11] Brunelli, R. and Poggio, T., "Face recognition: Features versus templates." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042-1052, October, 1993.

[12] Cave, C. and Kosslyn, S., "The role of parts and spatial relations in object identification". *Perception*, Vol. 22, 1993, pgs. 229-248.

[13] Chang, C. and Lee, S., "Retrieval of Similar Pictures on Pictorial Databases", *Pattern Recognition.* Vol. 23, No. 7, pgs. 675-680, 1991.

[14] Chang, S.K., Lee, C.M., and Dow, C.R., "A 2-D string matching algorithm for conceptual pictorial queries". *SPIE Image Storage and Retrieval Systems*, Vol. 1662, 1992, pgs. 47-58.

[15] Cohen P. and Feigenbaum, E., *The Handbook of Artificial Intelligence*, Vol. 3. William Kaufmann, Inc., Los Altos, CA, 1982.

[16] Corman, T., Leiserson, C., and Rivest, R., *Introduction to Algorithms.* MIT Press, Cambridge, 1991.

[17] Cornsweet, T. N. *Visual Perception.* Academic Press, New York, 1970.

[18] Equitz, W., "Image searching in a shipping product". *SPIE Storage and Retrieval for Image and Video Databases III.* San Jose, Feb., 1995. pgs. 186-196.

[19] Fischler, M.A., "Robotic vision: sketching natural scenes". *Image Understanding Workshop*, Palm Springs, Feb., 1996. pgs. 979-890.

[20] Foley, J., van Dam, A., Feiner, S., and Hughes, J., *Computer Graphics.* Addison Wesley, New York, 1993.

[21] Gallant, S. and Johnston, M., "Image retrieval using context vectors". *SPIE Storage and Retrieval for Image and Video Databases III.* San Jose, Feb., 1995. pgs. 82-94.

[22] Garey, M. and Johnson, D., *Computers and Intractibility.* W.H. Freeman and Co., New York, 1979.

[23] Grimson, W.E.L., and Lozano-Perez, T., "Model-based recognition and localization from sparse range or tactile data". *The International Journal of Robotics*, Vol. 3, No. 3, Fall 1984.

[24] Grimson, W.E.L, "On the recognition of curved objects". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):632-643, 1989.

[25] Grimson,W.E.L. and Huttenlocher,D.P. , "On the verification of hypothesized matches in model-based recognition". A.I. Memo 11110, The Artificial Intelligence Lab., M.I.T., 1989.

[26] Haftner, K., "Picture This". *Newsweek.* June 24, 1996.

[27] Hild, M. and Shirai, Y., "Interpretation of natural scenes using multi-parameter default models and qualitative constraints". *Fourth International Conference on Computer Vision*, Berlin, May, 1993, pgs. 497-501.

[28] Hou, T.-Y., Hsu, A., Liu, P., and Chiu, M.-Y. , A content-based indexing technique using relative geometry features, *Image and Storage Retrieval Systems*, SPIE vol. 1662, 1992.

[29] Hsieh, Y. , SiteCity: A Semi-Automated Site Modelling System, *CVPR*, San Francisco, June, 1996.

[30] Huttenlocher, D.P. and Ullman, S., "Object recognition using alignment". *International Conference on Computer Vision*, pages 102-111,1987.

[31] Jacob, C., Finkelstein, A., and Salesin, D., "Fast Multiresolution Image Querying". *SIGGRAPH*, Los Angeles, 1995.

[32] Jones, J. and Poggio, T., "Model-Based Matching of Line Drawings by Linear Combinations of Prototypes". *Fifth International Conference on Computer Vision*, pgs. 531-536, June, 1995.

[33] Kandel, E. and Schwartz, J (Eds.), *Principles of Neural Science*. Elsevier, New York, 1994.

[34] Kapur, T., Grimson, W.E.L., Wells, W.M., and Kikinis, R., "Segmentation of brain tissue from MRI Images". *Medical Image Analysis*, Vol 1, No 2, 1996.

[35] Kass, M., Witkin, A., and Terzopoulos, D., "Snakes: Active Contour Models". *International Journal of Computer Vision*, pgs. 321-331, 1988.

[36] Kelly, P.M., Cannon, T.M., and Hush, D.R., "Query by image example: the CANDID approach". *SPIE Storage and Retrieval for Image and Video Databases III.* San Jose, Feb., 1995. pgs 238-248.

[37] Kosslyn, S.M. *Image and Brain.* MIT Press, Cambridge, MA, 1994.

[38] La Cascia, M. and Ardizzone, E., "JACOB: Just a content-based query system for video databases". To appearn in *Proc. ICASSP*, Atlanta, May, 1996.

[39] Lakshmi Ratan, A., "The role of fixation and visual attention in object recognition", MIT AI-TR 1529, 1995.

[40] Lakshmi Ratan, A. and Phi Bang, D., "Wavelets for indexing into databases: A study", *unpublished*, 1996.

[41] Lipson P., Yuille A.L., O'Keefe D., Cavanaugh J., Taaffe J., and Rosenthal D. "Deformable Templates for Feature Extraction from Medical Images." ECCV. France, April, 1990. pages 413-417.

[42] Lipson, P., "Model-based Correspondence". M.S. Thesis. Massachussetts Institute of Technology, 1993.

[43] Liu, J. and Yang Y-H., “Multiresolution color image segmentation”. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 7, July 1994, pgs. 689-700.

[44] Lowe,D.G, “Perceptual organization and visual recognition”. Technical Report STAN-CS-84-1020, Stanford University, 1984.

[45] Mahmood, T.F. Syeda, “Data and model driven selection using color regions”. *Proceedings of the European Converence on Computer Vision,* 1992, pgs. 321-327.

[46] Minka, T.P. and Picard, R.W., "Interactive learning using a "society of models". MIT Media Laboratory Perceptual Computing Section Technical Report No. 349, 1996.

[47] Mundy, J., “Object recognition: the search for representations”.

[48] Pentland, A., Picard. R, and Sclaroff, S., "Photobook: Content-Based Manipulation of Image Databases", *SPIE Storage and Retrieval Image and Video Databases II,* No. 2185, Feb 6-10, 1994, San Jose.

[49] Petrakis, E. and Faloutsos, C., "Similarity Searching in Large Image DataBases", Computer Science Technical Report 3388, Univeristy of Maryland, College Park, Dec. 1994.

[50] Picard, R.W. and Minka, T.P., "Vision Texture for Annotation", MIT Media Laboratory Perceptual Computing Section Technical Report No. 302, 1994.

[51] Pollard, S.B. , Porrill, J., Mayhew, J.E.W., and Frisby, J.P., “Matching geometrical descriptions in three-space”. *Image and Vision Computing* ,5(2):73-78, May 1987.

[52] Pratt, W., *Digital Image Processing.* John Wiley & Sons, Inc., New York, 1991.

[53] Press, W., Flannery, B., Teukolsky, S., and Vetterling, W., *Numerical Recipies in C.* Cambridge University Press, Cambridge, 1990.

[54] Ravela, S., Manmatha, R., and Riseman, E., “Scale-space matching and image retrieval”. *IU Workshop,* Palm Springs, 1996. pgs 119-1207.

[55] Rubner, Y. and Tomasi, C., “Coalescing texture descriptors”. *IU Workshop,* Palm Springs, 1996. pgs 927-935.

[56] Rosenfeld, A. and Kak, A.C., *Digital Picture Processing.* Vol. 2. Academic Press, San Diego, 1982.

[57] Sinha, P., "Image Invariants for Object Recognition". *Investigative Ophthalmology and Visual Science,* 34/6, 1994.

[58] Smith, J.R. and Chang, S., “Single color extraction and image query”. *International Conference on Image Processing*, Washington, Oct., 1995.

[59] Smith, J.R. and Change, S., “Local color and texture extraction and spatial query”. *IEEE International Converence on Image Processing*, 1996.

[60] Strat, T., “Employing contextual information in computer vision”. Image Understanding Workshop, April, 1993, pgs. 217-229.

[61] Stricker, M., “Color and geometry as cues for indexing”. U.Chicago Technical Report CS92-22, Nov, 1992.

[62] Stricker, M., “Similarity of color images”. *Storage and Retrieval for Image and Video Databases*. Feb., 1993, pgs. 381-392.

[63] Swain, M. and Ballard, D, “Indexing via color histograms”. *International Conference on Computer Vision*, 1990, pgs. 390-393.

[64] Swain, M., "Interactive Indexing into Image Databases". *Storage and Retrieval for Image and Video Databases*. Feb., 1993, pgs. 95-103.

[65] Tagare, H., Vos, F.M, Jaffe, C.C., and Duncan, J.S., “Arrangement: A spatial relation between parts for evaluating similarity of tomographic section”. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 9, Sept. 1995, pgs 880-893.

[66] Tatanka, J.W. and Farah, M., “Parts and wholes in face recognition”, *Quarterly Journal of Experimental Psychology*, 46A (2), 1993, pgs. 225-245.

[67] Ullman, S. and Basri, R., “Recognition by linear combinations of models”. A.I. Memo 1152, The Artificial Intelligence Lab., M.I.T., 1989.

[68] Ullman, S., *High-level Vision: Object Recognition and Visual Cognition*. MIT Press, Cambridge, 1996.

[69] Viola, P. and Wells,W., “Alignment by Maximization of Mutual Information”. *Fifth Internation Conference on Computer Vision*, pages 16-23, June, 1995.

[70] Winston, P., *The Psychology of Computer Vision*, McGraw-Hill, Inc., New York, 1975.

[71] Yuille, A., Cohen D., and Hallinan, P., "Feature Extraction from faces using deformable templates", *CVPR*, June, 1989, pgs. 104-109.