Configuration based scene classification and image indexing

Article *in* Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition · January 1997

DOI: 10.1109/CVPR.1997.609453 · Source: DBLP

3 authors, including:

Pawan Sinha
Massachusetts Institute of Technology
203 PUBLICATIONS **7,292** CITATIONS

Configuration Based Scene Classification and Image Indexing

P. Lipson, E. Grimson, P. Sinha
MIT Artificial Intelligence Lab,
545 Technology Square, Cambridge, MA 02139
lipson@ai.mit.edu, welg@ai.mit.edu, sinha@ai.mit.edu

Abstract

Scene classification is a major open challenge in machine vision. Most solutions proposed so far such as those based on color histograms and local texture statistics cannot capture a scene's global configuration, which is critical in perceptual judgments of scene similarity. We present a novel approach, "configural recognition", for encoding scene class structure. The approach's main feature is its use of qualitative spatial and photometric relationships within and across regions in low resolution images. The emphasis on qualitative measures leads to enhanced generalization abilities and the use of low-resolution images renders the scheme computationally efficient. We present results on a large database of natural scenes. We also describe how qualitative scene concepts may be learned from examples.

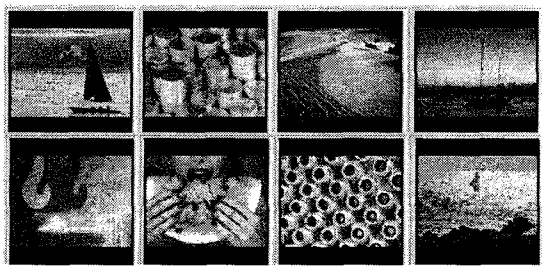


Figure 1. Using color histograms to find the most similar images to a water scene at sunset (upper left) returns pictures of money, molten liquid, and a woman eating watermelon. Although these images all have the same overall golden color, most differ greatly in semantic content.

1. The Problem

The goal of our work is to classify scenes based on their content. Scene classification has applications for the problem of image and video database indexing. With the increase in the number and sizes of digital libraries there is a need for automated, flexible, and reliable image search algorithms.

Recently several strategies have been developed for image classification. Most use aggregate measures of an image's color and texture as a signature for an image, then compare signatures to determine how similar one image is to another. Several image database indexing systems are based on this idea, such as QBIC [1], VIRAGE [3], and VisualSeek [10]. These similarity measures are adequate if the goal is to find images with similar distributions of color or other low level signal characteristics. However, if the goal is to find images from a given object/scene class, such as snowy mountains or waterfalls, the previously defined similarity measures often produce results incongruent with human expectations (Figure 1).

Figure 2 shows three images that perceptually belong to the same class, viz. coasts. However, the elements of which they are composed vary significantly in color distribution, texture, illumination, and spatial layout. Given such possible variations, how can we represent a class of images in order to detect instances of that class?

One possibility is to first recognize subpieces of an image and then to classify the image based on the identity of those subpieces. This strategy, however, presumes that a wide range of objects or scene-subparts can be reliably recognized in possibly complex configurations. There is a large body of work in computer vision which addresses the problem of object recognition. The successes in this area have usually been when the objects have well-defined boundaries, can be modeled by simple geometric primitives, are largely unoccluded and viewed under constrained lighting conditions. Such strategies are not well suited for complex scenes, especially those which consist mostly of natural objects.

Given these difficulties inherent in individual object recognition, our approach classifies scenes without first attempting to recognize their components. This strategy is supported by psychophysical evidence showing that humans may holistically classify visual stimuli before recognizing the individual parts [5][12].

1063-6919/97 $10.00

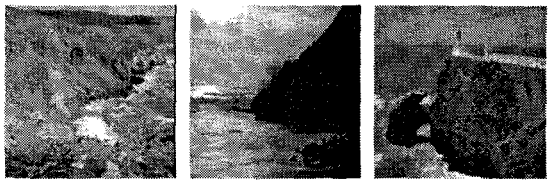


Figure 2. These images all belong to the coastal images class although colors, illumination, and layout vary widely.

In this paper, we suggest a novel representational strategy, "configural recognition", as a partial solution to the scene classification problem. The strategy encodes class models as sets of qualitative relationships between low resolution image regions. We also show how a flexible template matching strategy proves effective for detecting these qualitative models on novel images. Additionally, we suggest a technique for learning these models from a set of example images. Finally, we describe how other qualitative relations may be incorporated into the configural representation strategy.

2. Motivation for the approach

Our approach to scene classification is motivated by three considerations derived from studies of human perception, which we describe below.

1) The importance of global scene configuration. The image on the right in Figure 3 has been derived from the one on the left by dividing the latter into pieces and permuting their positions. Clearly, both images have identical chromatic and (to a large extent) textural statistics. Yet, perceptually, they do not belong to the same class since they have different overall configurations. This observation has been replicated in several systematic psychological studies which demonstrate that a stimulus in correct spatial configuration allows for more accurate and rapid detection of itself or its parts than the same stimulus with incorrect spatial relationships [2][4][5]. Our conclusion is that the overall organization of a scene's parts strongly influences its interpretation.

2) The use of qualitative measurements. Figure 4 shows three snow-capped mountain scenes. This class of images may be described as having three perceptually salient regions: a blue region (A), a white region (B), and a grey region (C). In all cases region A is above region B which is above region C. Therefore, even though the particular instances of the class exhibit these regions at diverse absolute locations and over different spatial extents, one constant is that all the regions across the images have the same *relative* spatial layout.

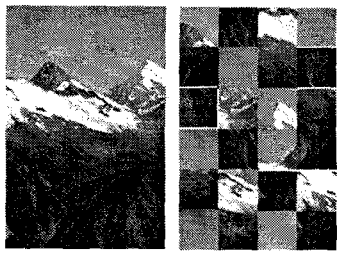


Figure 3. A mountain picture and its scrambled counterpart. Although both images contain the same color and textural characteristics, perceptually we would not classify the second image in the same category as the first.

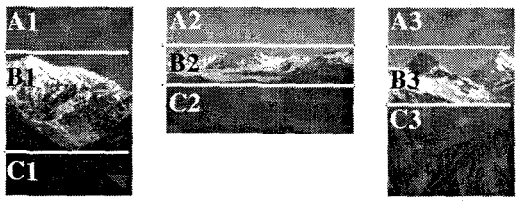


Figure 4. Three snow-capped mountains are shown. Each is divided into three regions (A, B, C). Even though the corresponding regions in the three images differ in their absolute sizes, positions, and colors, their relative spatial and photometric relationships are largely the same.

Just as relative spatial relationships may be used to encode the overall configuration of scene content, relative photometric relationships between image regions may also be important for perceptual classification of scenes. The corresponding image regions in figure 4 may not have the same absolute color, however, the relative photometric relationships between the regions are the same. For instance in all the images region A is bluer and brighter than region C.

This suggests that the classification of a scene may remain valid as 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. The difficulty observers experience in recognizing photographic negatives is a case in point.

3) The sufficiency of low spatial frequency information for scene classification. Humans need little detailed information to recognize many objects and scenes. Figure 5 shows several readily recognizable thumbnails. The only information retained in these small images is an arrangement of low frequency photometric regions. Such observations suggest that we can base our classification algorithm on an

Figure 5. Low resolution images may be sufficient for classification. These images are identifiable despite their extremely poor resolutions.

image's low frequency information.

These three observations are motivating factors for the configural recognition approach which is based on a qualitative encoding of a scene's photometric and spatial structure in low-resolution images.

3. Qualitative encoding of scene structure

The configural recognition scheme encodes class models as a set of salient image regions and salient qualitative relationships between those regions. The most closely related work to what we are about to describe is the ratio-template construct devised by Sinha to detect faces under varying illumination conditions. The construct consists of relative luminance relationships between image regions with fixed spatial positions [9]. Some researchers have previously considered using qualitative spatial relationships in the context of scene classification to describe the relationships between objects or object subparts in images [6][8][11]. They have typically assumed, however, that the objects or object subparts are pre-labeled.

The configural recognition system differs from these approaches by constructing class models for scenes from a wide vocabulary of relative relationships, including both spatial and photometric, between image regions that have not been preclassified. In the current system, class models are described using seven types of relative relationships between image patches. Each of these relationships can have the following values: less than, greater than, or equal to. The first three relations encode the relative color between image regions in terms of their red, green, and blue components. The fourth relationship used is relative luminance between the patches. (Other relative chromatic relationships, such as relative hue and saturation, can be easily incorporated into the system. The system is not dependent on the use of the RGB color space.) The spatial relationships used are relative horizontal and vertical descriptions. We also encode relative size, where the size of the patch is described by how many pixels it covers. Intra patch relative chromatic relationships may also be used in the model description.

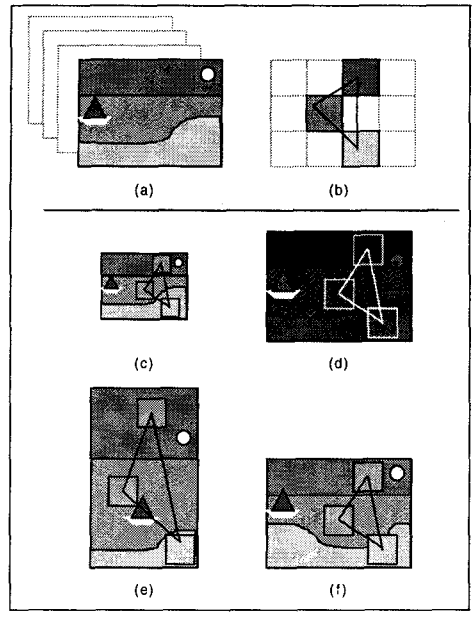


Figure 6. (a) Example beach images. (b) One encoding of a qualitative beach concept. The model remains valid over many scene variations including (c) scale changes, (d) illumination variations (the colors have changed but the relationships between the patches remain the same), (e) differing viewing parameters (distal vs. close up view), and (f) geometry changes.

Figure 6(a) denotes example beach scenes. (The example beach scene shown when rendered in color has blue sky green water, and tannish colored sand.) Figure 6(b) shows three highlighted patches, from an image grid of large equally sized patches, and their relative relationships, denoted by lines. This constitutes one possible model for beach scenes. The relationships in the model are that there is a bluer region, which is above and to the right of a greener region, both of which are above a more tan and lighter region.

3.1. Benefits of qualitative encoding

There are at least four significant benefits to using low-frequency image regions and their relative relations to encode scene classes. *1. Invariance to scene transformations.* The prime benefit is that the use of relative relationships over 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. Figure 6(c-f) illustrate how the relative relationships encoded in the model remain

valid over different but very commonplace image distortions such as changes in scale, illumination, viewing parameters and geometry. *2. Immunity to high-frequency sensor noise. 3. Dimensionality reduction.* Instead of having to use high-resolution images, 32x32 thumbnails sufficed for the classification task. *4. Simple image partitioning requirements.* Partitioning the image with a uniform grid suffices.

4. Model to image matching

We can think of the model as a prototype of a class. 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 regions themselves may grow or shrink in order to better fit to the image. 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.

5. Implementation and testing

The configural approach to scene classification was tested by generating several class templates and subsequently using these models to classify a large database of natural 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 made by human observers.

The test database consists of 700 images from prepackaged CD-ROM collections from Corel which contained 100 images each. Each 100 image collection consists of images that the vendor classified into one theme, e.g., "Fields", "Glaciers and Mountains", and "Waterfalls". The total collection contains pictures which have a wide range of content, colors, and textures. The pictures have been taken from a variety of viewing positions (close-up vs. panorama) and under different types of weather 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.

Each image was iteratively smoothed and subsampled to create a Gaussian pyramid of low resolution images. Each pyramid consists of three image sizes: 32x32, 16x16, and 8x8 pixels.

We manually constructed class templates for snowy mountains, snowy mountains with lakes, fields, and waterfalls. In figure 9 we describe the snowy mountain template and the waterfall template and show pictorially the database retrieval results using these class models (figures 10- 15).

RESULTS	"true pos."	"false pos."
Snowy mount.	75%	12%
S. mnt. w/lake	67%	1%
Field	80%	7%
Waterfall	33%	2%

Table 1. The classification results from four hand crafted templates for snowy mountains, snowy mountains with lakes, fields, and waterfalls. The results are reported in terms of the "true positives" and "false positives" with respect to human perceptual classification of the 700 image database.

In table 1 we describe the results of all four templates on the database in terms of "true positives", "false positives". (See [7] for more details.)

One template may not cover a whole class. For instance, two templates which are mirror images of each other may be needed to cover the examples of coastal images in figure 2. The waterfall template is a very narrow detector, selecting only 33% of the "true" waterfall images. We found experimentally that a combination of a small number of narrow detectors was a good strategy to cover most of the members of the class and very few instances of non-members.

6. Learning the scene concept

We have demonstrated that models consisting of qualitative relationships between low frequency image regions can be used effectively to classify images. It would be desirable if instead of hand-crafting the models, an automated process could take a set of example images and generate a set of templates which describe the relevant consistencies between the pictures in the example set.

We have developed an algorithm that computes the consistent relationships between regions across a set of example images. The algorithm first computes all pairwise qualitative relationships between each low resolution image region. For each region, the algorithm also computes a rough estimate of its color from a coarsely quantized color space as a measure of perceptual color. The next step, for each region, groups the image into directional equivalence classes, such as "above" and "below", with respect to that region. Redundant relationships in each equivalence class for each region are eliminated. The next step is to compute the consistent set of region relationships across the set of examples. There is, however, a problem in that the correspondence of regions across the images is not known. A reasonable assumption is that corresponding regions in each image are likely to occur in similar positions. To determine the set of consistent relationships, we only compare the set of relationships/colors

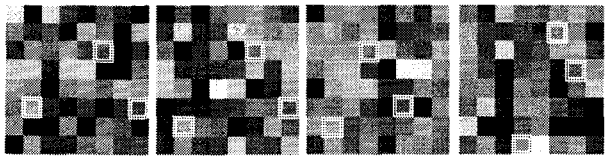


Figure 7. Four example input images to the learning algorithm. The patches which correspond to the embedded qualitative concept are highlighted in white in each image.

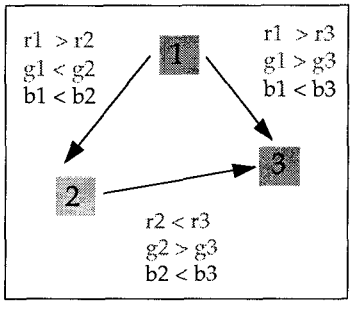


Figure 8. Resulting qualitative concept determined by the learning algorithm. The learned concept matches the concept embedded in each of the images.

in a neighborhood surrounding each region location across all the example images.

We tested the approach by generating randomly colored synthetic images. A three patch qualitative concept 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. Figures 7 and 8 respectively show the inputs to and output from the learning system. The extracted concept *matched* the original randomly generated concept. We are currently testing this approach on real imagery. We are also extending our algorithm to allow the user to delineate particularly salient regions in the images.

7. Summary and Conclusions

We have presented a novel approach to classifying scenes in terms of qualitative relationships between low frequency image regions that provides a computationally efficient way to encode overall scene structure. We demonstrated the effectiveness of our approach by creating four natural scene class models and testing each of these on a large database of natural images. 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 dissimilar configurations, resulting in low false positive rates. We described a method to learn qualitative scene classes from a set of examples.

Although the configural recognition approach appears to be a promising strategy for scene classification, it also has limitations. For instance, the technique is not suited to make fine quantitative discriminations, such as between different types of mountains. The technique is not designed 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.

We are experimenting with an expanded repertoire of qualitative and quantitative information for classification of a broader class of images. For instance, we have created a template which includes relative texture measurements in order to classify cityscapes (see figure 16).

This paper is available on-line (with color images) at the URL http://www.ai.mit.edu/people/lipson/

Acknowledgments

This work was sponsored in part by ARPA under ONR contract N00014-95-1-060.

References

[1] J. Ashley, M. Flickner, D. Lee, W. Niblack, and D. Petkovic, "Query by image content and its applications," IBM Research Report, RJ 9947 (87906), Computer Science/Mathematics, March, 1995.

[2] M. Bar and S. Ullman, "Spatial context in recognition," *Perception*, vol. 25, pp. 342-352, 1996..

[3] J.R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R.C. Jain, and C. Shu, "Virage image search engine: an open framework for image management," *SPIE Storage and Retrieval of Image and Video Databases*, Vol. 4, pp. 76-87, 1996.

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

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

[6] C. Chang, S. Lee, "Retrieval of similar pictures on pictorial databases," *Pattern Recognition*, Vol. 23, No. 7, pp. 675-680, 1991.

[7] P. Lipson, "Context and Configuration Based Scene Classification", *Ph.D. Thesis*, MIT, Sept. 1996.

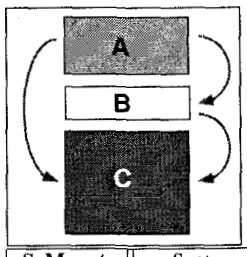
[8] E. Petrakis and C. Faloutsos, "Similarity searching in large image databases," CS Tech. Report 3388, U. Maryland, College Park, Dec. 1994.

[9] P. Sinha, "Image invariants for object recognition," *Invest. Opth. and Vis. Science*, 34/6, 1994.

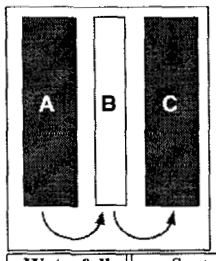
[10] J.R. Smith and S. Chang, "Local color and texture extraction and spatial query," *IEEE Int. Conf. on Image Processing*, 1996.

[11] H. Tagare, F.M. Vos, C.C. Jaffe, and J.S. Duncan, "Arrangement: A spatial relation between parts for evaluating similarity of tomographic sections," *PAMI*, Vol. 17, No. 9, pp. 225-245, Sept. 1995.

[12] J.W. Tatanka and M. Farah, "Parts and wholse in face recognition," *Quarterly J. of Exp. Psych.*, 46A (2), pp. 225-245, 1993.



S. Mount.	Spat.	Lum.	Color
A to B	$A_y < B_y$	$A_l < B_l$	$A_r < B_r$
			$A_g < B_g$
			$A_b < B_b$
A to C	$A_y < C_y$	$A_l > 1.2C_l$	$A_b > C_b$
B to C	$B_y < C_y$	$B_l > C_l$	$B_r > C_r$
			$B_g > C_g$
			$B_b > C_b$
A to A			$A_b > A_r$
			$A_b > A_g$



Waterfall	Spat.	Lum.	Color
A to B	$A_x < B_x$	$1.6A_l < B_l$	$A_r < B_r$
			$A_g < B_g$
			$A_b < B_b$
B to B		$B_l > 80$	$B_r < 1.2B_b$
			$B_r < 1.2B_g$
B to C	$B_x < C_x$	$B_l > C_l$	$B_r > C_r$
			$B_g > C_g$
			$B_b > C_b$

Figure 9. Qualitative models for two natural-scene classes, viz. snowy mountain scenes and waterfalls. The figures on top show the schematic layouts of the models while the tables list the qualitative constraints that inter- and intra-region relationships in a given image have to satisfy for the image to be accepted as a member of the scene class.

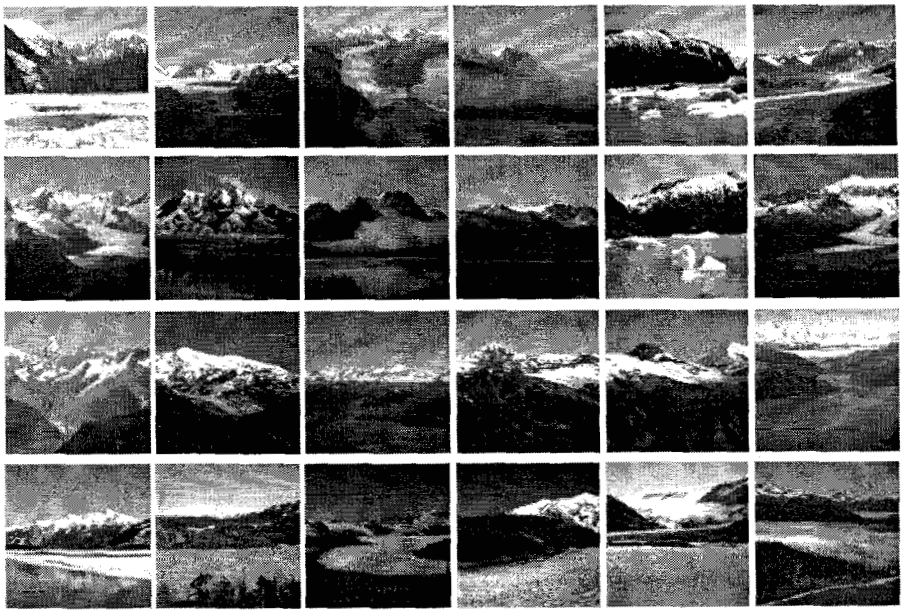


Figure 10. "True positives" detected by the snowy mountain template. Notice the diversity in these scenes captured by a single qualitative model.

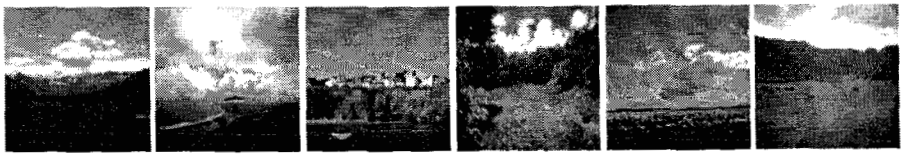


Figure 11. "False positives" detected by the snowy mountain template. Since the qualitative model does not encode fine textural details, it sometimes fails to distinguish between snowy mountains and white clouds.

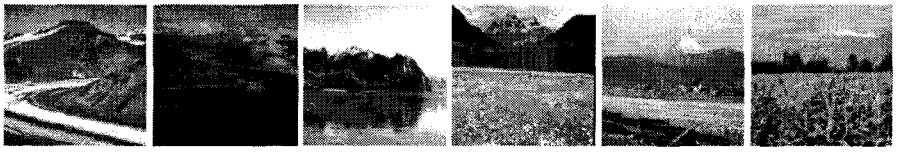


Figure 12. Mountain scenes not detected by the snowy mountain template. These failures are often due to significant differences between image configurations and the general scene structure encoded in the model. Also, sometimes the scene entities such as the snowy mountains are too small to be picked up by the qualitative model in the low frequency images.

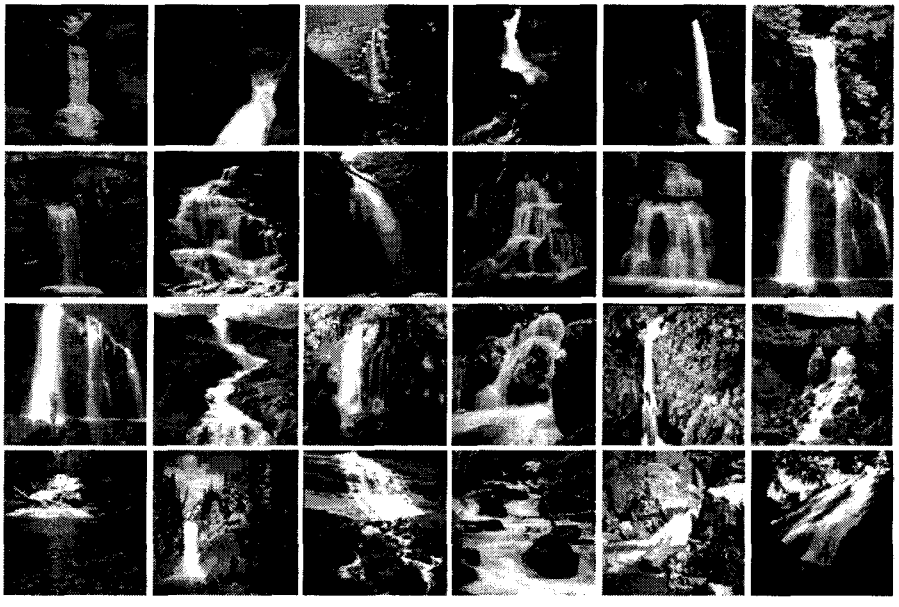


Figure 13. "True positives" detected by the waterfall template.

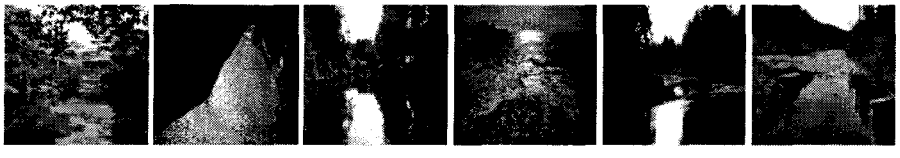


Figure 14. "False positives" detected by the waterfall template.

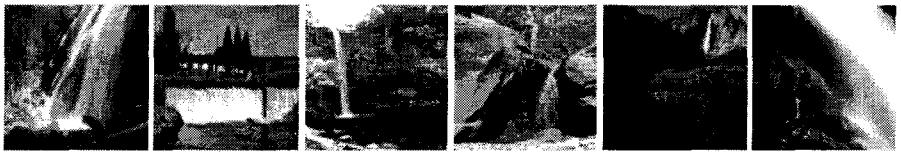


Figure 15. Waterfall scenes not detected by the waterfall template.

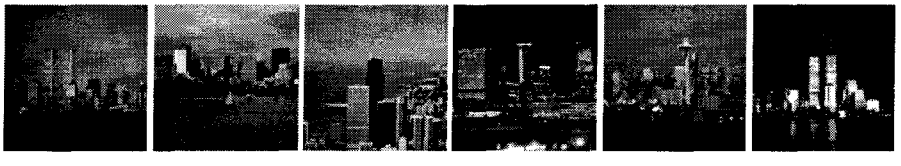


Figure 16. A demonstration of the use of relative textural statistics in the configural recognition framework. These cityscape scenes were detected by a qualitative mode that encoded not only qualitative chromatic and spatial relationships but also ordinal relations between the orientation energies in different image regions.