Pergamon

Pattern Recognition, Vol. 31, No. 12, pp. 1921–1935, 1998

0031-3203/98 $19.00 + 0.00

PII: S0031-3203(98)00079-X

ON IMAGE CLASSIFICATION: CITY IMAGES VS. LANDSCAPES

ADITYA VAILAYA,* ANIL JAIN*,† and HONG JIANG ZHANG‡

*Department of Computer Science, Michigan State University, East Lansing, MI 48824, U.S.A.
‡Broadband Informations Systems Lab, HP Labs, Palo Alto, CA 94304, U.S.A.

(*Received* 11 *March* 1998; *in revised form* 19 *May* 1998)

Abstract—Grouping images into semantically meaningful categories using low-level visual features is a challenging and important problem in content-based image retrieval. Based on these groupings, effective indices can be built for an image database. In this paper, we show how a specific high-level classification problem (city images vs landscapes) can be solved from relatively simple low-level features geared for the particular classes. We have developed a procedure to qualitatively measure the saliency of a feature towards a classification problem based on the plot of the intra-class and inter-class distance distributions. We use this approach to determine the discriminative power of the following features: color histogram, color coherence vector, DCT coefficient, edge direction histogram, and edge direction coherence vector. We determine that the edge direction-based features have the most discriminative power for the classification problem of interest here. A weighted k-NN classifier is used for the classification which results in an accuracy of 93.9% when evaluated on an image database of 2716 images using the leave-one-out method. This approach has been extended to further classify 528 landscape images into forests, mountains, and sunset/sunrise classes. First, the input images are classified as sunset/sunrise images vs forest & mountain images (94.5% accuracy) and then the forest & mountain images are classified as forest images or mountain images (91.7% accuracy). We are currently identifying further semantic classes to assign to images as well as extracting low level features which are salient for these classes. Our final goal is to combine multiple 2-class classifiers into a single hierarchical classifier.

Image classification Clustering Salient features Similarity Image database Content-based retrieval

1. INTRODUCTION

Content-based image organization and retrieval has emerged as an important area in computer vision and multimedia computing, due to the rapid development in digital imaging, storage and networking technologies. Various systems have been proposed in the recent literature for content-based image retrieval, such as QBIC,[1] Photobook,[2] SWIM,[3] Virage,[4] Visualseek,[5] Netra,[6] and MARS.[7] These systems follow the paradigm of representing images via a set of feature attributes, such as color, texture, shape, and layout. These features are archived along with the images in the database. A retrieval is performed by matching the feature attributes of the query image with those of the database images.

For large databases with over tens of thousands of images, effective indexing becomes an important issue in content-based image retrieval. This problem has not been solved very successfully in current image database systems. What is even more challenging is how to group images into semantically meaningful categories or index images in a database based on low-level visual features of the images. A successful categorization of images will greatly enhance the performance of content-based image retrieval systems by filtering out images from irrelevant classes during matching. One attempt to solve this image indexing problem is the hierarchical indexing scheme proposed by Zhang and Zhong,[8,9] which uses a self-organization map (SOM) to cluster images into groups of visually similar images based on color and texture features. This indexing scheme was further applied in reference[10] to create a texture thesaurus for indexing a database of large aerial photographs. However, the success of such clustering-based indexing schemes is often limited, largely due to the low-level feature-based representation of image content. To achieve the goal of automatic categorization and indexing of images in a large database, we need to develop robust schemes to identify salient features of images that capture a certain aspect of semantic content of these images. This is a well-known and important problem in pattern recognition and computer vision.

In this paper, we show how a high-level concept can be learned from images using relatively simple low-level features. Understanding the entire content of an image may not be possible with the state-of-the-art feature extraction and matching algorithms in

†Author to whom correspondence should be addressed. Tel.: 001517 353 6484; Fax: 001 517 432 1061; E-mail: jain@cps.msu.edu.

computer vision, but we believe that it is feasible to attempt rather specific classification problems using low-level features geared towards the particular classes. In order to identify meaningful image categories that are of interest to users and that can be automatically identified by simple and efficient pattern recognition techniques, we conducted a simple experiment where 8 human subjects were asked to group 171 outdoor images into a number of clusters. The subjects were also asked to name these clusters. The experiment yielded the following 11 categories: forests and farmlands, natural scenery and mountains, beach and water scenes, pathways, sunset/sunrise scenes, city scenes, bridges and city scenes with water, monuments, scenes of Washington DC, a mixed class of city and natural scenes, and face images. While this experiment was useful in identifying meaningful categories for retrieval purposes, we feel that instead of generating a multi-class classification, it may be more feasible to attempt multiple two-class classifications based on features which have high discriminability for the particular two classes. Based on this observation, we focus here on two user-defined categories: city and landscape images.

The image classification problem of interest here can be defined as follows: Given an image, classify it into either a city or a landscape image. We assume an existing set of labeled images (training set) to which the input image is compared. The labels for the training set are assigned by a human subject. There obviously exist images that can be placed into both these categories (e.g. a city skyline at sunset). We assign these "ambiguous" images to the closest category (based on an independent subject's classification). We assume that the test image belongs to one of the two categories. The proposed system does not yet have the capability to reject images belonging to neither of the two classes. The input to the classification system is an image and the output is the confidence with which the system assigns the input to the city and landscape classes. Figures 1(a) and (b) shows typical images from the city and landscape classes, respectively. City scenes can be characterized by the presence of man-made objects and structures such as buildings, cars, roads, etc. Natural scenes, on the other hand, lack these structures.

The proposed classification scheme can be used to hierarchically organize images, making browsing and retrieval of images more efficient. Schemes for indoor–outdoor image classification[11,12] have already been reported in the literature. Outdoor images can further be dichotomized into city-landscape classes at the next level of classification hierarchy using our scheme. A query for a beach scene can, thus, reduce the search to outdoor *natural* scenes only, instead of searching the entire database.

The rest of the paper is organized as follows. Section 2 briefly discusses some of the recent work in the literature on high-level image classification. We provide details of our experiments with human subjects on image classification in Section 3. Our image database is described in Section 4. Section 5 describes the proposed approach and the classification results, and we outline future directions of our work in Section 6.

2. PREVIOUS WORK

A number of attempts have been made to understand high-level semantics from images using low-level features. Yiu[12] and Szummer and Picard[11] propose algorithms for indoor-outdoor scene classification. Yiu reports results on a database of 500 images and uses color and dominant directions to do the classification. Szummer and Picard show results on a larger database of 1343 images. They propose a combination of color and texture (MSAR—multiresolution, simultaneous autoregressive model[13]) features on 4×4 blocks of images to do the classification. These systems report classification accuracies of approximately 90%.

Forsyth *et al.*[14] use specialized grouping heuristics to classify coherent regions in an image under increasingly stringent conditions to recognize objects in the image. They demonstrate recognition of trees by fusing texture and geometric properties, and learn

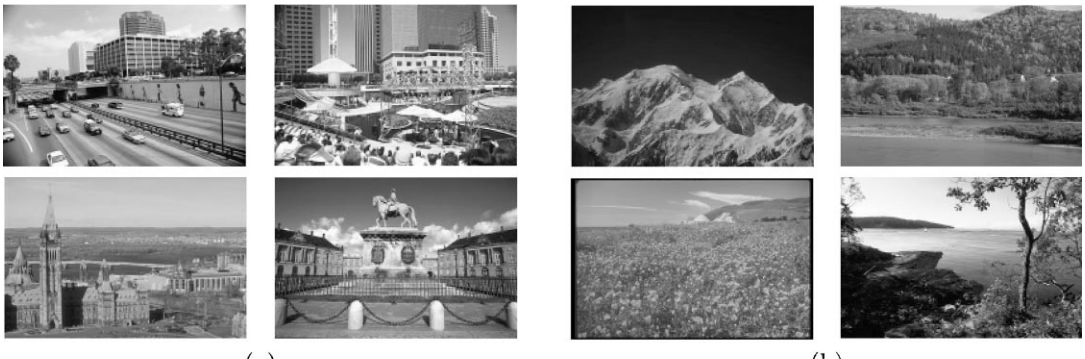


Fig. 1. Typical (a) city and (b) landscape images.

blob-like landscape concepts using grouped features. However, they do not report the efficiency of classification or the accuracy of the retrievals on a large database based on classification of regions into blobs.

Yu and Wolf[15] use one-dimensional hidden Markov models along horizontal and vertical blocks of images to do scene classification. Quantized color histogram vectors are used as features for the image sub-blocks and the hidden Markov models are used to learn statistical templates from examples. The scheme suffers from the drawback that one-dimensional model does not capture spatial relationships well.

Gorkani and Picard[16] have proposed the use of dominant orientations found using a multiscale steerable pyramid in 4×4 sub-blocks of 98 images to discriminate between city/suburb scenes from photos of landscape scenes. They classify an image as a city scene if a majority of sub-blocks have a dominant vertical orientation or a mix of vertical and horizontal orientations. Their database consisted of vacation photographs provided by British Telecom (BT). Their system misclassified 7 out of the 98 database images.

Our approach differs from[16] not only in the classification scheme but also on the features used. Our approach is based on weighted k-nearest neighbor classifier with leave-one-out option for error estimation. On the same 98 BT images used in reference,[16] our system misclassified only four images, two each from the city and landscape classes. We report more comprehensive results on a larger database of 2716 images in Section 5.3.

3. CLASSIFICATION BY HUMAN SUBJECTS

What do users typically want to search for in an image database? Users either have an *a priori* idea of what they are looking for (e.g. someone's face) or they have a rather vague abstract notion of the pictures (e.g., looking for scenic photographs for planning a vacation).[17] In the second type of queries, there is a need for classification of database images based on the abstract concepts. The user can then browse through the classes of the desired images.

The first step is to identify meaningful image categories that are of interest to users and which can be automatically identified by simple and efficient pattern recognition techniques. For this purpose, we conducted an experiment with eight human subjects to classify a set of 171 images. We asked eight human subjects to group these 171 color images (including the 98 images used by Gorkani and Picard[16]) into meaningful categories. The goal of this experiment was to identify semantic and abstract classes that humans assign to these images. The subjects were given no explicit criteria for judging the similarity and could create any number of categories with any number of images per category. The subjects were also asked to explain the semantic meanings of the groups that they formed and explain the reasons for placing each image into a particular group. The subjects were given unlimited time to complete the task, except that they had to do so in one sitting without contacting anyone else. In general, the subjects took between 1 and 2 hours to complete the task.

The experimental data indicated that the number of categories selected by the eight subjects varied from 7 to 17. Some of the categories formed by various subjects included buildings, streets, cities, bridges, monuments, people, natural scenes, mountains, farms, country side, forests, sunset/sunrise scenes, etc. Of these numerous categories, there were certain categories which were coherent in their semantic descriptions. For example, the class of buildings, streets, and monuments, used by some subjects can be grouped under the *city class* which was created by another subject. Similarly, the categories such as mountains, forests, farms, and country side, can be grouped under a broader category, called *natural scenes*. Figures 2(a) and (b) show 25 images each from the forests and farmlands, and city clusters identified by the human subjects, respectively. In order to generate a hierarchical grouping based on the categorization of the human subjects, we generated a dissimilarity measure between every pair of images.

A 171×171 dissimilarity matrix (this is a symmetric matrix, so only the upper diagonal entries need to be considered) was generated, where the dissimilarity value for every pair of images was assigned the number of subjects (out of 8) who *did not* group the pair into one category (regardless of what the actual category meant for each subject). Note that the integer entries in the dissimilarity matrix vary from 0 to 8. Based on the dissimilarity matrix, we performed a complete-link hierarchical clustering of the 171 images. The resulting dendrogram is shown in Fig. 3(a). Note that clusters can be formed by cutting this dendrogram at a specific level of dissimilarity. At the highest level of dissimilarity (by cutting the dendrogram at the dissimilarity level of 7; at level 8, all the images are in one cluster), 11 clusters were formed. The semantic labels that can be assigned to these clusters and the number of images in each cluster are as follows: (i) forests and farmland (37 images), (ii) natural scenery and mountains (19 images), (iii) beach and water scenes (6 images), (iv) pathways (roads and streams, 9 images), (v) sunset/sunrise shots (21 images), (vi) city shots (38 images), (vii) bridges and city scenes with water (6 images), (viii) monuments (9 images), (ix) pictures of Washington DC (5 images), (x) a mixed class of city and natural scenes (20 images), and (xi) a face image (1 image). We refer to the mixed category of city and landscape shots as the miscellaneous class.

We organized the 11 categories generated by the human subjects into a hierarchy as shown in Fig. 3(b). The first four classes of forests, natural scenery, beach scenes, and pathways can be grouped into a single class, labeled natural scenes (71 images). The clusters

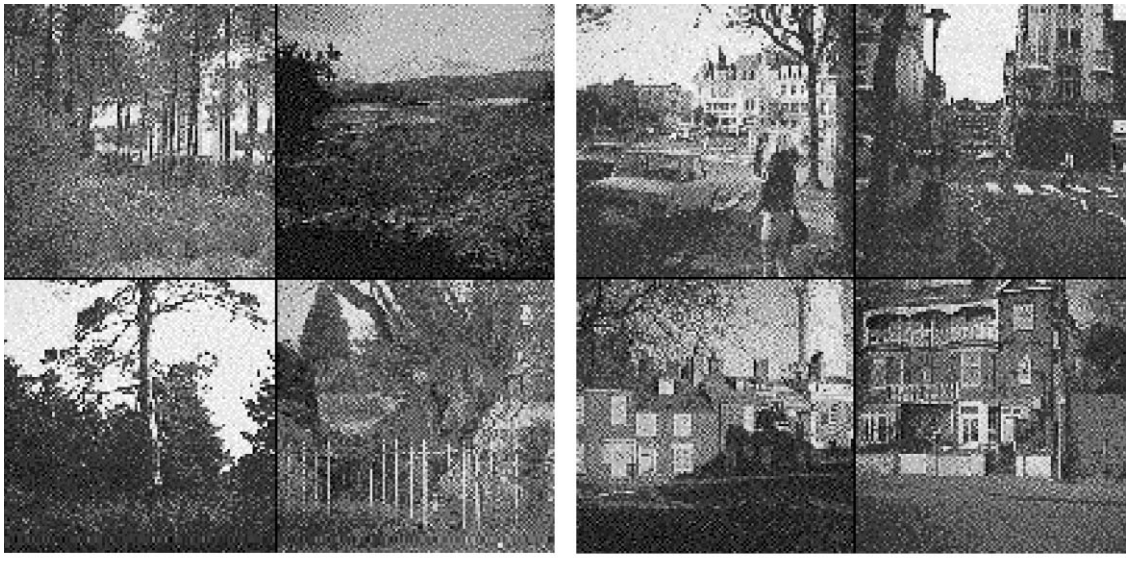


Fig. 2. Representative images of (a) forest and farmland and (b) city clusters identified by human subjects.

of city shots, monuments, and shots of Washington DC can be grouped into the category of city scenes (58 images). At the next level, the sunrise/sunset group can be classified into the category of landscapes (92 images). At the last stage, the landscapes, the city scenes, the miscellaneous class, and the lone face image are grouped together into one class (all 171 images). Hence, this rather small database can be considered to have two major classes (city shots and landscapes) at the highest level which account for 150 out of the 171 images. This also explains our emphasis on these two semantic categories.

4. IMAGE DATABASE

The image database used in this experiment consists of 2716 images of various city and landscape scenes collected from various sources. It includes 2145 images from Corel stock photo library, 170 city and landscape images used in the human experiment reported in Section 3 (the miscellaneous class of mixed city-landscape scenes was also included), around 250 key frames from digitized video of television serial (mostly pictures inside buildings), 30 images from VisTex database at MIT Media Lab, 20 scanned images of vacation photos, and around 100 photographs downloaded from the web. The images are of different sizes ranging from 150×150 to 750×750. The color images are represented by 24-bits per pixel. No pre-processing was done on the images. The ground truth for the 2716 images was assigned based on a single subject. Ambiguous images were assigned to the closest category. The final labeling assigned 1128 images to the city class and 1588 images to the landscape class.

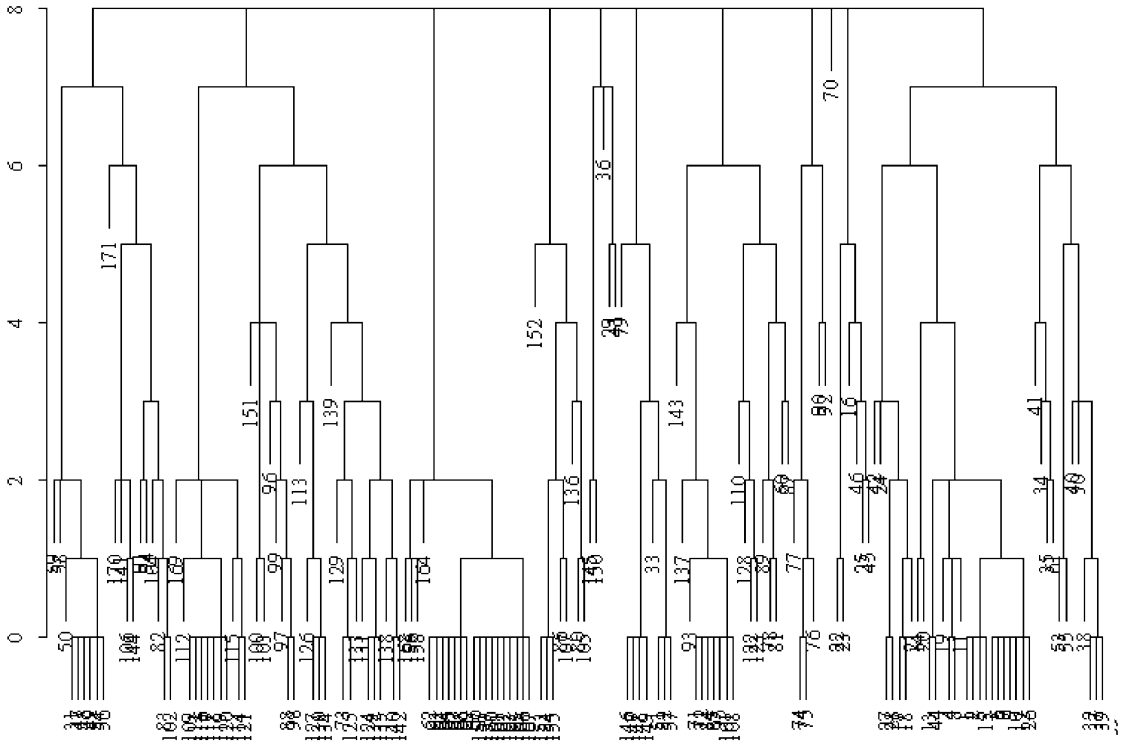
5. IMAGE FEATURES

It is a challenge to select a good feature set for image classification. A number of image features based on color and texture attributes have been reported in the literature for image retrieval. Yet, quantifying their discrimination ability to the given classification problem is difficult. We have used the following procedure to select the most salient feature from a given set of features for city-landscape classification.

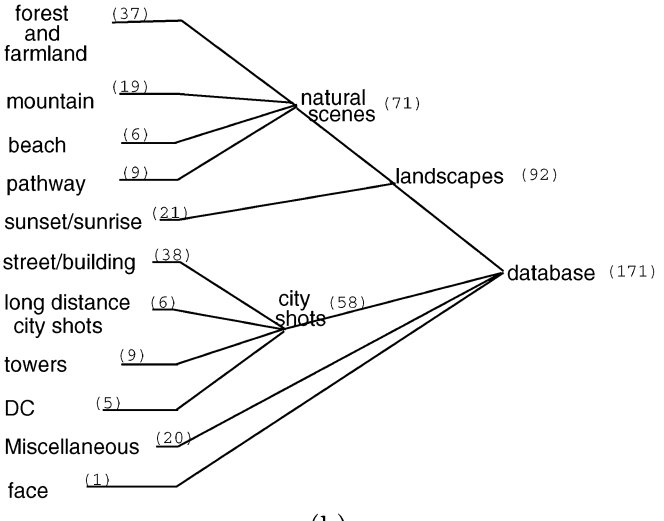
5.1. Feature evaluation

A given feature is said to have a large discrimination power if its intra-class (within-class) distances are small and the inter-class (between-class) distances are large. We thus empirically constructed distributions for the intra- and inter-class distances for the given set of features. The intra-class distribution was generated by computing the distances between every pair of images in the same class. The inter-class distribution was generated by computing the distance of every image in the city class to every image in the landscape class. Based on the overlap of the intra- and inter-class distance distributions, we can identify the most discriminative features.

We evaluated the following five features for their ability to discriminate between the city and landscape classes: color histograms, color coherence vectors, moments of the Discrete Cosine Transform (DCT) coefficients of the image, edge direction histograms,



(a)



(b)

Fig. 3. (a) Complete-link dendrogram of 171 images based on dissimilarities generated by groupings provided by human subjects; (b) a hierarchical organization of the 11 categories obtained from the dendrogram in (a). The number of images in each category is indicated in the parenthesis.

and edge direction coherence vectors. These features have been used earlier, so we only describe them briefly here.

- *Color histograms*: We define color features in terms of a histogram in the quantized *HSV* color space. The *RGB* space is uniformly sampled into $16 \times 16 \times 16$ bins which are represented by the center color and transformed to the *HSV* color space. The 4096 colors in the *HSV* space are then clustered into 64 colors using a K-means clustering algorithm.[18] A look-up table is generated to map

every *RGB* value to a bin based on the cluster membership. We extract 5 local color histograms for every image (top left quarter, top right quarter, bottom left quarter, bottom right quarter, and central quarter) to incorporate spatial information. A color similarity matrix, $A(i, j)$, (generated from the distances between the 64 color centers) is used to smooth the histograms as follows:

$$H'(i) = \sum_{j=1}^{64} H(j) * A(i, j),$$

where H is the original histogram and H' is the smoothed histogram. The smoothing spreads the value of a bin into bins of similar color.

- *Color coherence vector*: A color coherence vector is a color histogram refinement scheme that divides each bin into coherent and non-coherent pixels.[19] A pixel in a bin is said to be coherent if it is part of a large similarly colored region. An 8-neighbor connected component analysis is used to extract connected regions of the same color. Pixels in regions whose size exceeds a threshold (1% of image size) are counted as coherent pixels, and those from smaller regions count towards non-coherent regions. We extended the color coherency concept to our local histograms in the quantized *HSV* space. Figures 4(a)–(h) show a city and landscape image and their color histogram and color coherence vector features, respectively. Let $d_{CH}(X, Y)$ and $d_{CCV}(X, Y)$ represent the dissimilarity values between two images X and Y using the color histogram and color coherence feature vectors, respectively. For the given images, $d_{CH}(a, b) = 0.51$ and $d_{CCV}(a, b) = 0.52$. Long distance landscape (mountains, sunset/sunrise, etc) shots tend to have more coherent pixels (large regions of similar color, (see Figs 4(f) and (h)) whereas city scenes and some landscape scenes (forest images) tend to have an equal fraction of coherent and non-coherent pixels (see Figs 4(e) and (g)).
- *DCT coefficients*: Color and texture features are also described in terms of the Discrete Cosine Transform (DCT) coefficients of an image. Every image in the database was transformed to the JPEG compressed format. Central moments of the second- and third-order[20,21] of the DCT coefficients were used to represent an image. In particular, 4 moments for each of the 17 DCT coefficients (first 9 coefficients in the intensity field and the first 4 DCT coefficients in the two chrominance fields) were used as the features. The resulting 2716×68 pattern matrix was then normalized to zero mean and unit variance.
- *Edge direction histograms*: We define texture in an image in terms of an edge direction histogram.[22] The Canny edge detector is used to extract the edges from an image. We have modified the edge direction histograms defined in reference[22] by adding an extra bin to measure the frequency of non-edge pixels in the image. A total of 73 bins are used to represent the edge direction histogram of an image; the first 72 bins are used to represent edge directions quantized at 5° interval and the last bin represents a count of the number of pixels that didn't contribute to an edge. To compensate for different image sizes, we normalize the histograms as follows:

$$H(i) = H(i)/n_e,\ i \in [0, \ldots, 71]; \quad H(72) = H(72)/n_p,$$

where $H(i)$ is the count in bin i of the edge direction histogram, n_e is the total number of edge points detected in the image, and n_p is the total number of pixels in the image.
- *Edge direction coherence vector*: An edge direction coherence vector stores the number of coherent versus non-coherent edge pixels with the same directions (within a quantization of 5°). A threshold (0.1% of image size) on the size of every connected component of edges in a given direction is used to decide whether the region is coherent or not. This feature is thus geared towards discriminating structured edges from arbitrary edge distributions when the edge direction histograms matched. Figures 5(a)–(h) show a city and landscape image and their edge direction histogram and edge direction coherence vector features, respectively. As can be seen in Figs 5(e) and (f), the city images tend to produce coherent edge pixels in the vertical direction, where as most edges in landscape images are non-coherent. Let $d_{EDH}(X, Y)$ and $d_{EDCV}(X, Y)$ represent the dissimilarity values between two images X and Y using the edge direction histogram and edge direction coherence feature vectors, respectively. For the given images, $d_{EDH}(a, b) = 0.13$ and $d_{EDCV}(a, b) = 0.39$. The coherent edge pixels in the vertical direction in the city images (and the lack of these in landscape images) increase the discriminating power of the edge direction coherent vectors over the edge direction histograms for the two classes under consideration.

Using the Euclidean distance as a distance measure for each of the features, we generated the intra-class and the inter-class distance distributions for the database images. Figures 6(a)–(d) show the intra- and inter-class distributions for color histogram, DCT moment, and edge direction histogram and coherence vector features. The color coherence vectors yield similar inter-class and intra-class distributions as the color histograms. The large overlap between the two distributions based on color histograms and DCT coefficients shows that these features are not suited for the city-landscape classification problem. On the other hand, the distributions based on the edge direction coherence vectors have a substantially smaller overlap (especially, at lower distances). These results suggest that the use of a small set of near

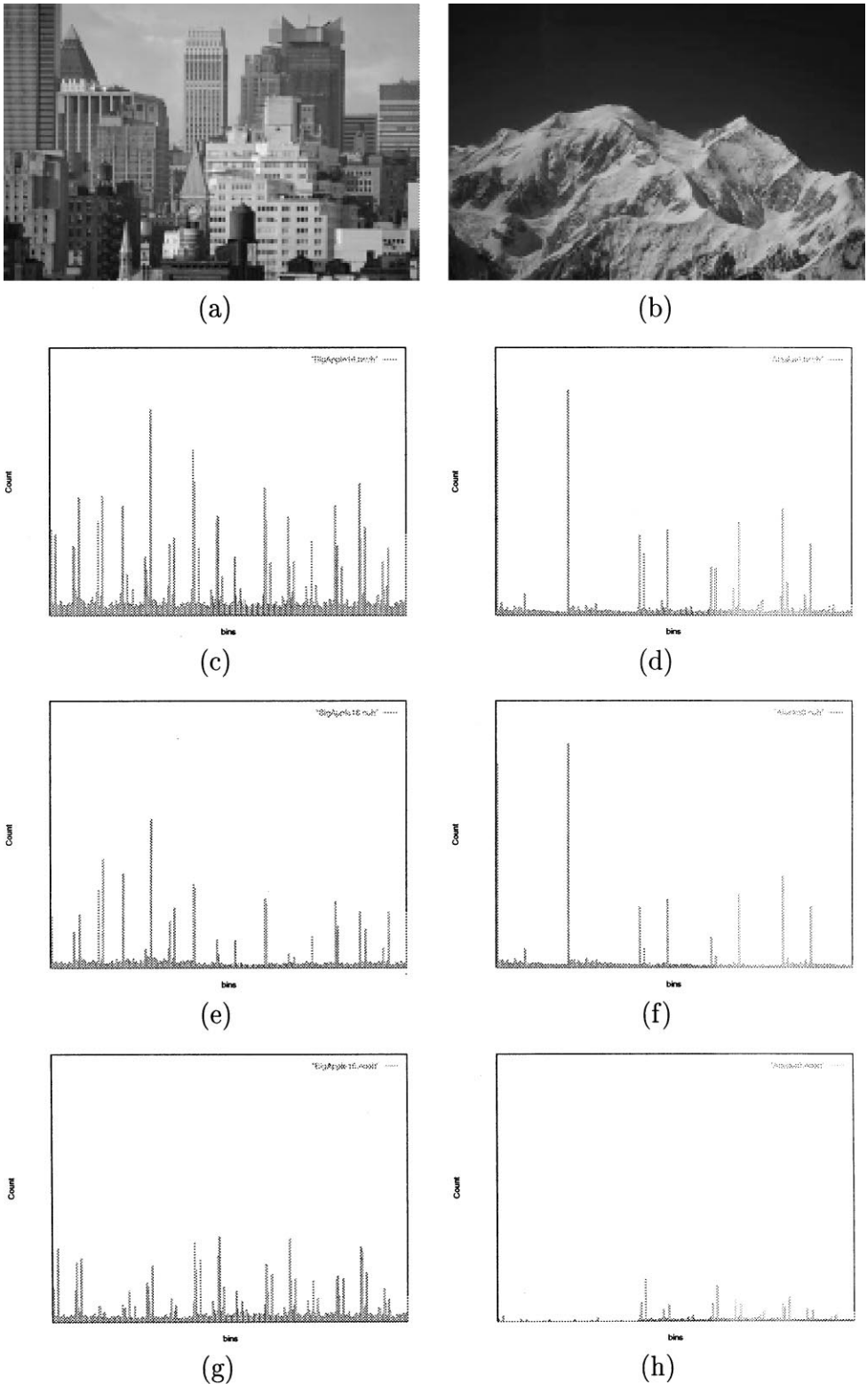


Fig. 4. Color-based features for (a) city and (b) landscape image; (c) & (d) show the color histogram features for (a) & (b); (e) & (f) show the coherent color bins for (a) & (b); (g) & (h) show the non-coherent color bins for (a) & (b); $d_{CH}(a, b) = 0.51$, $d_{CCV}(a, b) = 0.52$.

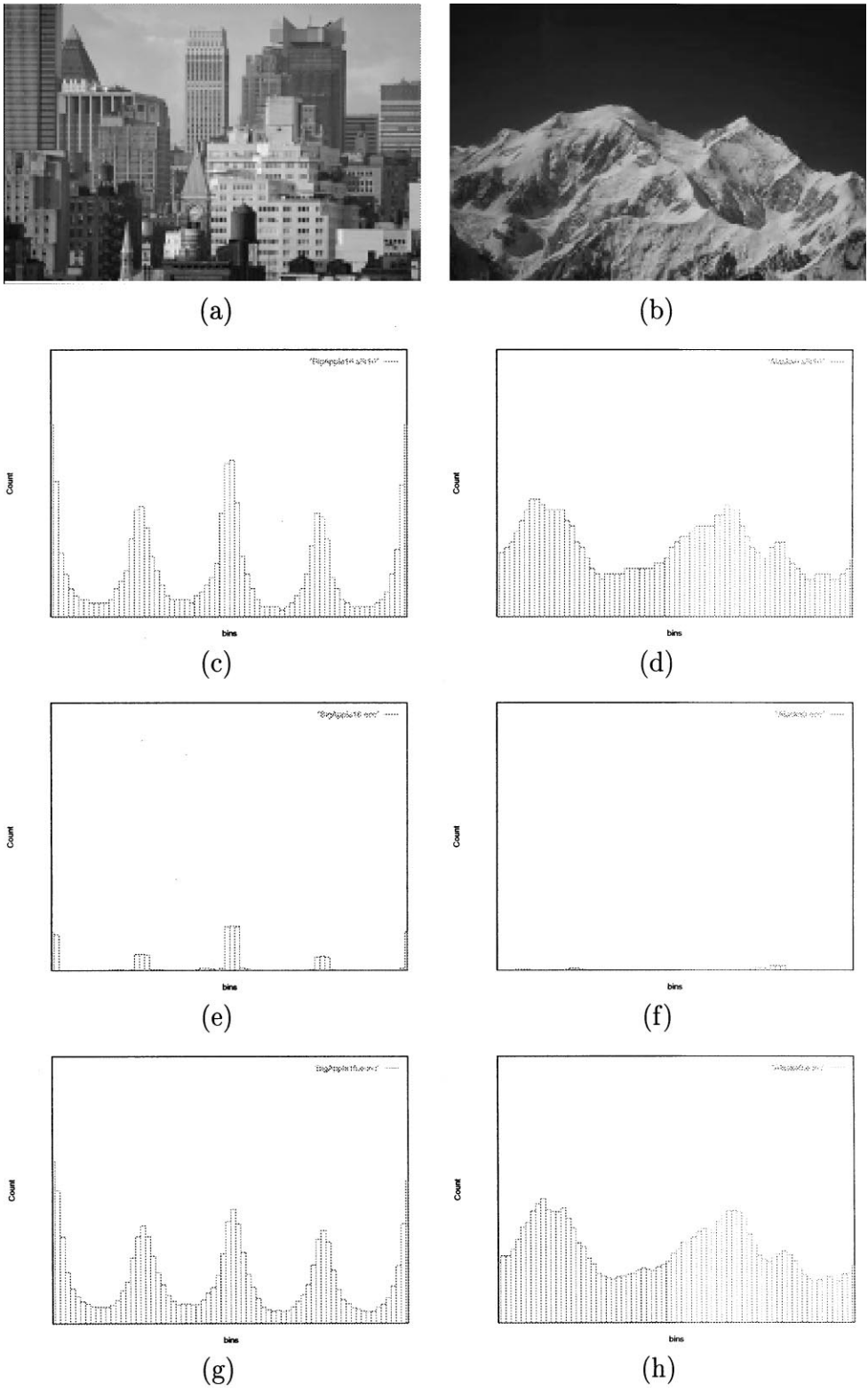


Fig. 5. Edge direction-based features for (a) city and (b) landscape image; (c) & (d) show the edge direction histogram features for (a) & (b); (e) & (f) show the coherent edge direction bins for (a) & (b); (g) & (h) show the non-coherent edge direction bins for (a) & (b); $d_{EDH}(a, b) = 0.13$, $d_{EDCV}(a, b) = 0.39$.

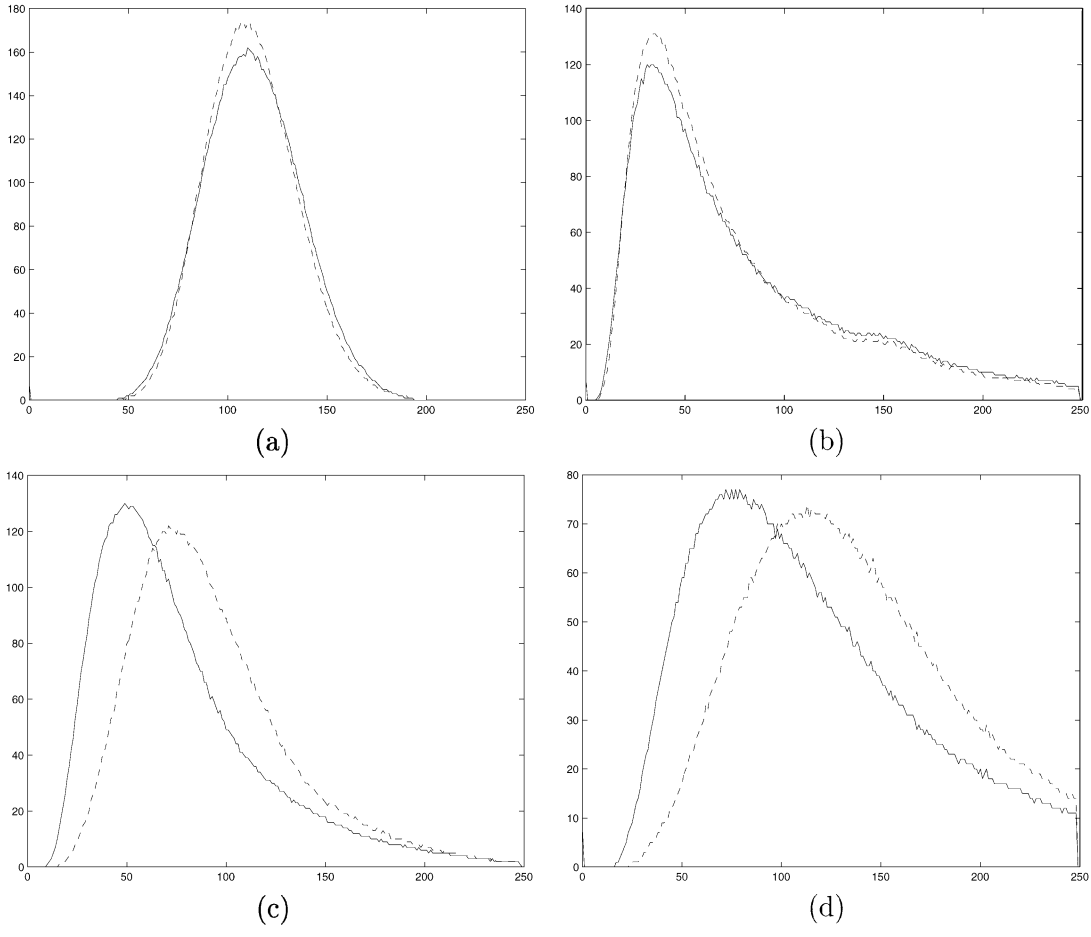


Fig. 6. Intra-class and inter-class distance distributions using (a) color histograms; (b) DCT moments; (c) edge direction histograms; (d) edge direction coherence vectors; solid-line represents the intra-class distance, while the dotted-line represents the inter-class distance; x-axis represents inter-image distance; y-axis represents the frequency.

neighbors based on edge direction features are better in discriminating between the two classes than the color histograms and DCT coefficient features. In order to verify the claim, we used Sammon's non-linear projection algorithm to plot the high-dimensional patterns in two dimensions. As can be seen in Figs 7(a) and (b), the edge direction coherence vectors do a better job in discriminating between the given two classes.

Man-made objects in the city scenes usually have strong vertical and horizontal edges, whereas non-city scenes tend to have edges randomly distributed in various directions. A feature based on the distribution of edge directions can discriminate between the two categories of images. On the other hand, color features would not have sufficient discriminatory power as man-made objects have arbitrary color distributions (two buildings need not have the same color). We feel that color features may be better suited for further classification of landscapes (natural scenes), where colors are relatively constant (grass has a yellowish-green hue, sky has a blue hue, etc).

5.2. *Classification*

A weighted k-nearest-neighbor classifier with leave-one-out option is used to classify an image into either city or landscape class.[18] Dissimilarity is based on the Euclidean distance between features of an image (edge direction histograms/coherence vectors, color histograms/coherence vectors, or moments of DCT coefficients). Let the true class of the ith nearest neighbor of a test image be c_i ($i = 1, \ldots, k$), $c_i \in [city, landscape]$. Let the Euclidean distance between the feature vector of the test image and its ith nearest neighbor be d_i ($i = 1, \ldots, k$). If the distances are normalized to lie in the range $[0, 1]$, the confidence, p_j, that the test image belongs to class j can be defined as:

$$p_j = \frac{\sum_{i:c_i=j}(1 - d_i)}{\sum_{i<=k}(1 - d_i)}, \quad j = 1, 2.$$

The test image is assigned to class j if $p_j > 0.5$.

5.3. *Classification results*

Table 1 shows the classification results for the k-NN classifier with different values of k. As indicated

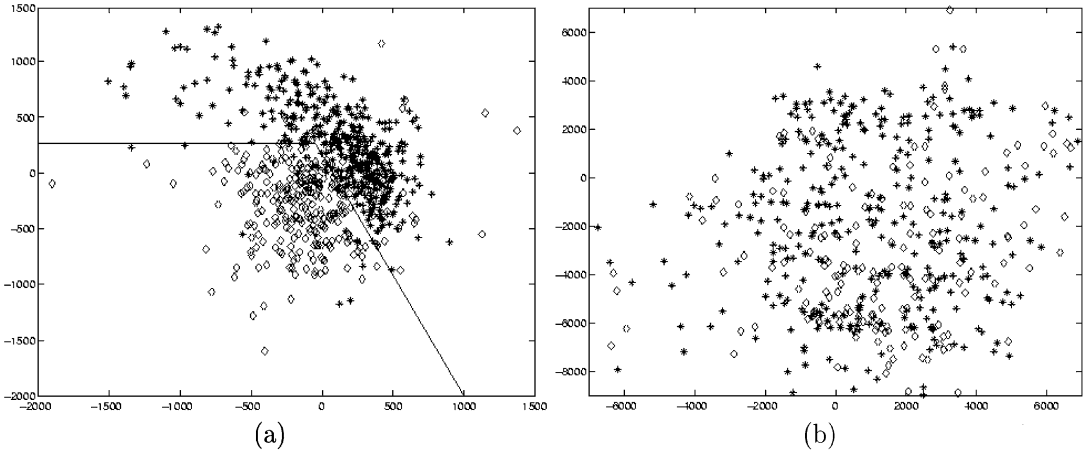


Fig. 7. 2*D* feature space plots showing (a) edge direction coherence vector and (b) color coherence vectors; ∗ represents the landscape patterns and ◇ represents the city patterns; only a subset of 2716 patterns have been plotted here for clarity of display.

Table 1. Classification rates using k-NN classifier for edge direction histogram (EDH), edge direction coherence vector (EDCV), color histogram (CH), color coherence vector (CCV), and DCT moment (DCT) features; the values in **bold** are the best accuracy rates for the given features

Feature	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$	$k=11$	$k=13$
EDH	91.7	92.4	92.7	92.9	**93.0**	92.9	92.7
EDCV	92.3	93.5	**93.9**	93.7	93.8	93.5	93.4
CH	72.1	74.9	75.6	75.3	**75.8**	75.8	75.6
CCV	73.0	74.6	75.5	75.5	**75.8**	74.8	74.8
DCT	66.4	67.9	68.6	69.7	69.7	69.7	**70.3**

earlier by analyzing the intra- and inter-class distance distributions, the edge direction features outperform the color and DCT coefficient features. The best accuracy of 93.9% (164 images were misclassified) was obtained with a 5-NN classifier using edge direction coherence vectors. Of the 164 misclassified images, 87 were city scenes and 77 were landscape scenes. The confusion matrix is given in Table 2. The edge direction features, in general, resulted in classification accuracies of over 92% for different values of k ($1 \leq k \leq 20$).

Table 2. Confusion matrix for 5-NN classifier using edge direction coherence vector feature

True class	Classification	
	City	Landscape
City	1,041	87
Landscape	77	1,511

Figures 8 and 9 show some typical misclassified city and landscape images, respectively, for 5-NN classifier using edge direction coherence vectors. The misclassified city images mainly lacked strong vertical edges and can be placed into three categories: (i) long distance city shots at night (Figs 8(a)–(e)); (ii) scenes of buildings obstructed by trees (Figs 8(f)–(j)); and (iii) scenes of highly textured buildings (Figs 8(k)–(o)). While the inability to extract strong edges from the dark pictures was the reason for misclassification of city images in the first category (city scenes at night), the presence of textured regions resulted in misclassification of the city images in the second and third case (textured buildings and buildings covered by trees). The landscape images were mostly misclassified due to the presence of strong vertical edges of tree trunks, fences, or close-ups of plant stems. Table 3 shows the distribution of confidence values associated with the true class for every image in the database. A total of 2144 images (79%) were classified with 100% confidence to their true class and 2429 images (89%) were classified with a confidence of over 75% to their true class. These results demonstrate that the edge direction features have sufficient discriminatory power for the classification problem considered here.

5.4. *Combining multiple features*

We have also conducted experiments with integration of multiple feature attributes with the hope of

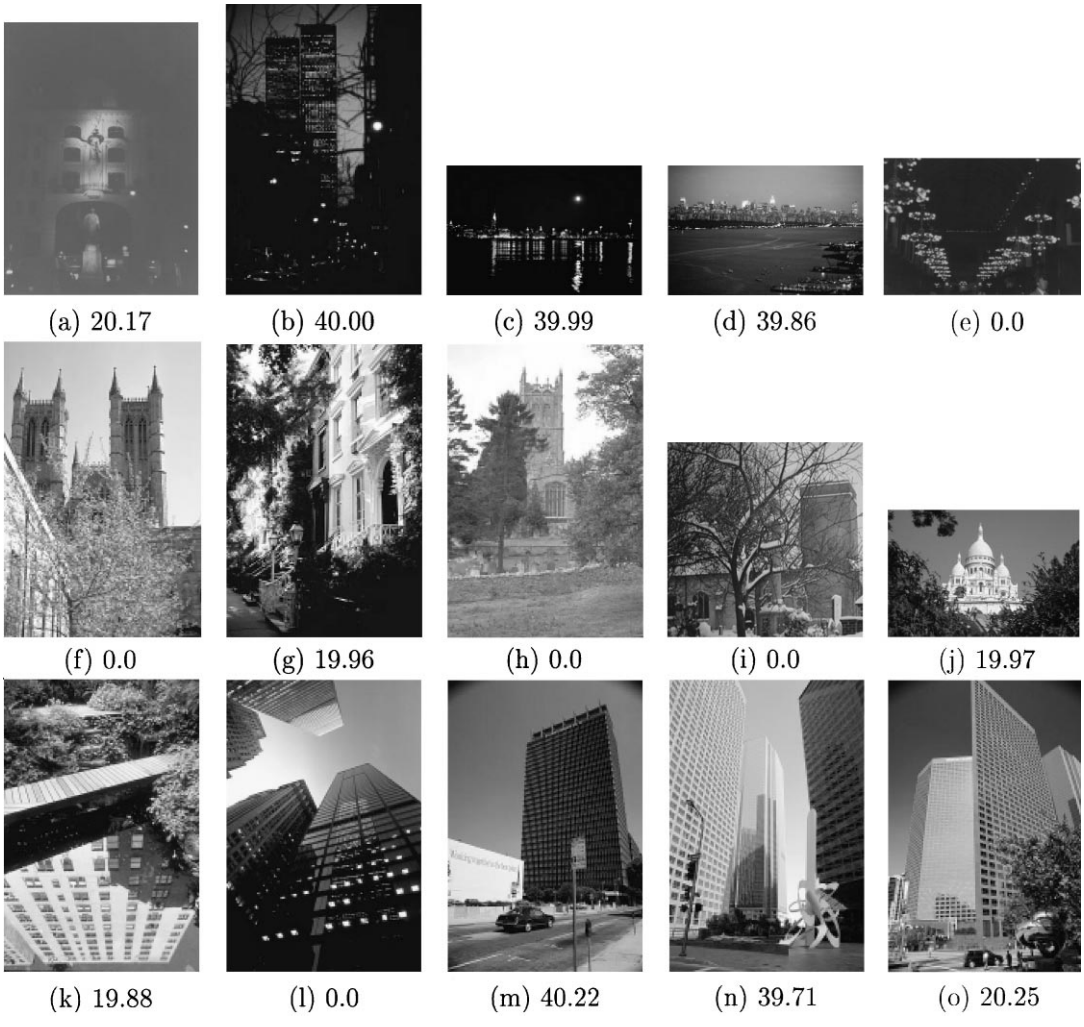


Fig. 8. A subset of misclassified city images and the corresponding confidence values (in%) associated with the true class using edge direction coherence vector-based 5-NN classifier.

improving the classification results.[22] Let X be a query image and Y be a database image. Let $d_1(X, Y)$ and $d_2(X, Y)$ be the dissimilarity indices between X and Y on the basis of two different feature vectors (e.g. color histograms and edge direction histograms). We define an integrated dissimilarity index $d_I(X, Y)$ between X and Y as

$$d_I(X, Y) = \frac{w_1 * d_1(X, Y) + w_2 * d_2(X, Y)}{w_1 + w_2}, \quad (1)$$

where w_1 and w_2 are the weights assigned to the two feature dissimilarity values. In our experiments we set the weights to 1. Since, the edge direction features performed much better than the color and DCT coefficient features, we integrated results of the edge direction features (histograms/coherence vectors) with either the color or DCT coefficient features. We found that the classifiers using multiple feature vectors did much better than the classifiers using individual color, DCT coefficient, and the edge direction histogram features with classification accuracies of approximately 93%. Among the various combinations of feature vectors, edge direction coherence vector combined with color coherence vector gave the best accuracy of 94%. But, this performance is essentially the same as the performance when edge direction coherence vectors (93.9%) alone are used. These results again demonstrate that the edge direction coherence vector has good discriminatory power for city vs. landscape classification.

5.5. *Classifying landscape images*

We have met with some success in extending the above classification scheme to further classify landscape images. A subset (528 images) of the landscape images were classified into the following three classes: forests, mountains, and sunset/sunrise classes. Of the 528 images, a human subject labeled 177, 196, and 155

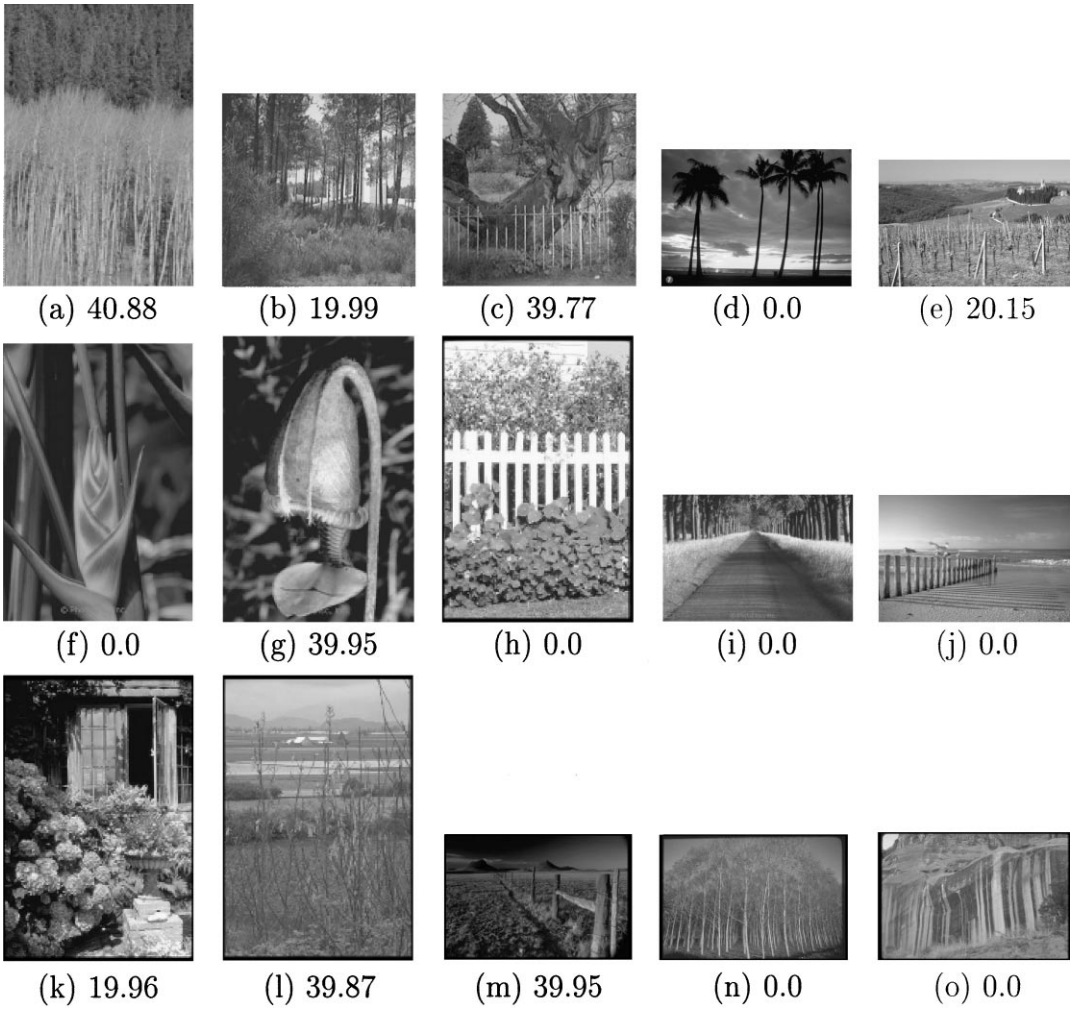


Fig. 9. Fifteen misclassified landscape images and the corresponding confidence values (in %) associated with the true class using an edge direction coherence vector-based 5-NN classifier.

Table 3. Confidence value associated with the true class for every image in the database using edge direction coherence vector-based 5-NN classifier; 89% of the 2716 images were correctly classified with a confidence of 75% and above

Confidence (%)	100	80	75	60	55	40	35	20	15	0
Number of Images	2,144	185	100	75	48	29	20	28	31	56

images as belonging to the forest, mountain, and sunset classes, respectively. Figures 10(a)–(c) show typical images from the forest, mountain, and sunset categories. These three categories do not represent all the landscape images, and hence, we considered only the subset of landscape images (528 images) that could be assigned to one of these three categories. Our future goal is to identify other semantically meaningful classes into which most of the landscape images can be classified.

Plotting the intra- and inter-class distance distributions for the 528 images, we found that color-based features had the most discriminating power. Natural objects tend to have uniform color distributions; trees have brown and green colors, sky is usually blue and white during the day, while sunset scenes usually have dark red or orange colors. Therefore, color can be an important cue in discriminating natural images. Edge direction features showed a high discrimination power for forest vs. sunset images, but lacked any discrimination power for forest vs. mountain images. This is mostly due to the fact that some mountain images also have pictures of trees in them. We find that color features are able to discriminate between these two classes since, forest scenes have little or no sky in them, whereas mountain scenes have

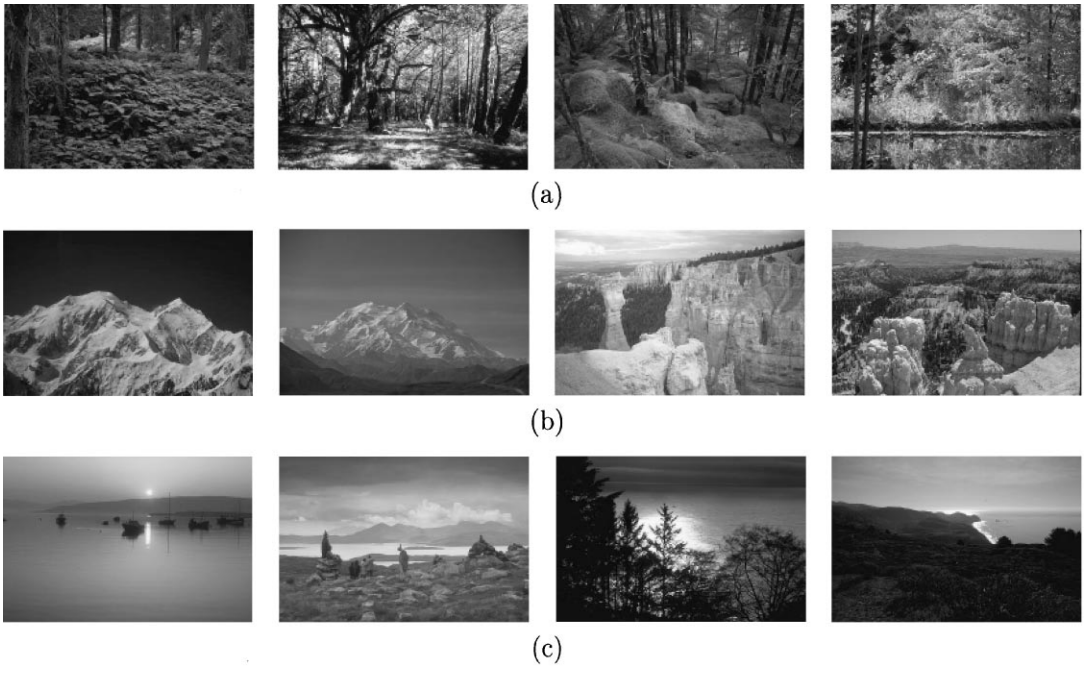


Fig. 10. Typical (a) forest, (b) mountain, and (c) sunset/sunrise images.

a considerable region of sky. Sunset/sunrise images yield very few edge points and hence edge direction features can easily discriminate them from the mountain and forest scenes.

Instead of implementing a three-class classifier, we first approach the problem of classifying an image into either sunset/sunrise class or the forest & mountain class (intra- and inter-class distance distributions show that both color and edge direction features can differentiate between these two classes and hence, it is a simpler problem to tackle). We next approach the problem of forest vs mountain classification if the input image was not classified as a sunset/sunrise image. Table 4 shows the classification accuracies for the given classification problems using a k-NN classifier with leave-one-out method. The results reported here are the best accuracies for $1 \leq k \leq 9$. As the intra- and inter-class distance distribution plots suggested, color-based features performed much better at the classification tasks, whereas the edge direction histogram-based features were good for discriminating between sunset/sunrise images from mountain and forest images. The best accuracy for forest & mountain vs. sunset/sunrise classification was 91.9% using color coherence vectors. The best accuracy for forest vs. mountain classification was 90.9% using color histograms. Combining the color and edge direction features improves the classification accuracy as shown in Table 5. The best accuracies of 94.5% and 91.7% for the two classification problems are obtained by combining the color and edge direction histogram features.

Table 4. Classification accuracies using k-NN classifier with leave-one-out method for the various features; MF/S denotes the problem of classifying an input image as a mountain & forest scene vs a sunset/sunrise scene; F/M denotes the problem of classifying an image as a forest scene vs a mountain scene; CH, CCV, EDH, EDCV, and DCT represent color histogram, color coherence vector, edge direction histogram, edge direction coherence vector, and DCT features, respectively

Classification problem	CH (%)	CCV (%)	EDH (%)	EDCV (%)	DCT (%)
MF/S	89.0	91.9	89.0	89.0	79.2
F/M	90.9	90.4	82.6	76.1	81.2

Table 5. Classification accuracies using k-NN classifier with leave-one-out method for combination of color and edge direction features; CH, CCV, and EDH represent color histogram, color coherence vector, and edge direction histogram features, respectively

Classification problem	CH & EDH (%)	CCV & EDH (%)
MF/S	94.58	93.8
F/M	91.7	90.4

6. CONCLUSIONS AND FUTURE WORK

Our experiments with human subjects show that users assign abstract and semantic labels to images while classifying them. Hence, it would be desirable to automatically attach semantic labels to a collection of

images if the goal is to generate semantic content-based indices for retrieval and browsing. Rather than learn all the concepts in an image, we show how constrained classification problems can be solved. We propose that instead of generating a hierarchical clustering of images based on a single feature, it is more feasible to perform pairwise classifications based on features that are appropriate for distinguishing between the 2 classes. Specifically, we approach this problem by classifying images into city or landscape category. We show that edge direction features (histograms and coherence vectors) have sufficient discrimination power for city-landscape classification. Our experiments on a database of 2716 images yield good results with an accuracy of about 93%. The proposed classifier assumes that the database consists of only those images that belong to the two classes. Our classification scheme classifies every test image as a city or landscape scene and does not have a reject option. We are currently investigating means to reject an input image that does not belong to any of these two classes (e.g. fingerprints, face images, etc.).

We have also shown that a classification based on *ad hoc* low-level image features does not often have enough discriminatory power towards the classification problem. It is necessary to identify salient low-level features for a given classification problem. We have shown that color features such as histograms and coherence vectors and texture features such as DCT coefficients are not suited for city-landscape classification. Plots of intra- and inter-class distance distributions can be used to qualitatively determine the discrimination ability of a feature towards a classification problem.

We are currently identifying additional semantically meaningful classes that can be easily assigned to images. As a first step in this direction, we have further classified a subset of landscape images into forest, mountain, and sunset/sunrise classes. A combination of color- and edge direction-based cues can be used to hierarchically classify natural scenes into these three classes with over 91% classification accuracy. Our future efforts are directed towards identifying further semantically meaningful categories to assign to images as well as extracting features geared for the particular classes. Our final goal is to combine multiple 2-class classifiers into a single hierarchical classifier.

REFERENCES

1. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic and W. Equitz, Efficient and effective querying by image content, *J. Intell. Inform. Systems* **3,** 231–262 (1994).
2. A. Pentland, R. W. Picard and S. Sclaroff, Photobook: content-based manipulation of image databases, *SPIE* Vol. 2185: *Storage and Retrieval for Image and Video Databases II*, pp. 34–47 (1994).
3. H. J. Zhang, C. Y. Low, S. W. Smoliar and J. H. Wu, Video parsing, retrieval and browsing: an integrated and content-based solution, *Proc. ACM Multimedia '95*, pp. 15–24, San Francisco, CA (1995).
4. A. Hampapur, A. Gupta, B. Horowitz, C. F. Shu, C. Fuller, J. Bach, M. Gorkani and R. Jain, Virage video engine, in *Proc. SPIE*: *Storage and Retrieval for Image and Video Databases V*, pp. 188–197. San Jose (1997).
5. J. R. Smith and S. F. Chang, Visualseek: A fully automated content-based image query system, in *Proc. ACM Multimedia*, pp. 87–98 (1996).
6. W. Y. Ma and B. S. Manjunath, Netra: A toolbox for navigating large image databases, *Proc. Int. Conf. on Image Proc*., Vol. 1, pp. 568–571. Santa Barbara, CA (1997).
7. S. Mehrotra, Y. Rui, M. Ortega and T. S. Huang, Supporting content-based queries over images in MARS, *Proc. IEEE Int. Conf. on Multimedia Computing and Systems* (1997).
8. H. J. Zhang and D. Zhong, A scheme for visual feature based image indexing, *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases*, pp. 36–46. San Jose, CA (1995).
9. D. Zhong, H. J. Zhang and S.-F. Chang, Clustering methods for video browsing and annotation, *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases*, San Jose, CA (1995).
10. W. Y. Ma and B. S. Manjunath, Image indexing using a texture dictionary, *Proc. SPIE Conf. on Image Storage and Archiving System*, Vol. 2606, pp. 288–298. Philadelphia, PA (1995).
11. M. Szummer and R. W. Picard, Indoor-outdoor image classification, *IEEE Int. Workshop on Content-based Access of Image and Video Databases*, *in conjunction with ICCV'98*, Bombay, India, (1998). http://www white.media.mit.edu/people/szummer/profile.html.
12. E. C. Yiu, Image classification using color cues and texture orientation, Master's Thesis, Department of EECS, MIT (1996).
13. J. Mao and A. K. Jain, Texture classification and segmentation using multiresolution simultaneous autoregressive models, *Pattern Recognition* **25**(2), 173–188 (1992).
14. D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson and C. Bregler, Finding pictures of objects in large collections of images, *Int. Workshop on Object Recognition for Computer Vision*, Cambridge, England, (1996). http://www.cs.berkeley.edu/projects/vision/publications.html.
15. H.-H. Yu and W. Wolf, Scenic classification methods for image and video databases, *Proc. SPIE*, *Digital Image Storage and Archiving Systems*, pp. 363–371 (1995).
16. M. M. Gorkani and R. W. Picard, Texture orientation for sorting photos "at a glance", *12th Int. Conf. on Pattern Recognition*, pp. 459–464. Jerusalem (1994).
17. R. M. Bolle, B.-L. Yeo and M. Yeung, Video query: Beyond the keywords, *IBM J. Res. Development* (1998) (to appear) (see also IBM Res. Rep. RC20586, 1996).
18. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ (1988).
19. G. Pass, R. Zabih and J. Miller, Comparing images using color coherence vectors,, *Proc. 4th ACM Conf. on Multimedia*, Boston, MA (1996). http://simon.cs.cornell.edu/Info/People/rdz/rdz.html.
20. M. K. Hu, Visual pattern recognition by moment invariants, *IRE Trans. Inform. Theory* **IT-8,** 179–187 (1962).
21. A. Vailaya, Y. Zhong and A. K. Jain, A hierarchical system for efficient image retrieval, *13th Int. Conf. on Pattern Recognition*, pp. C356–360. Vienna (1996).
22. A. K. Jain and A. Vailaya, Image retrieval using color and shape, *Pattern Recognition* **29,** 1233–1244 (1996).

About the Author—ADITYA VAILAYA received his B. Tech from Indian Institute of Technology, Delhi in 1994 and his M.S. degree from Michigan State University, East Lansing in 1996. He is currently working for his Ph.D. at Michigan State University. His research interests include image and video databases, image understanding, pattern recognition and classification, computer vision, and multimedia.

About the Author—ANIL JAIN is a University Distinguished Professor and Chair of the Department of Computer Science and Engineering at Michigan State University. His research interests include statistical pattern recognition, Markov random fields, texture analysis, neural networks, document image analysis, fingerprint matching and 3D object recognition. He received the best paper awards in 1987 and 1991 and certificates for outstanding contributions in 1976, 1979, 1992, and 1997 from the Pattern Recognition Society. He also received the 1996 IEEE Trans. Neural Networks Outstanding Paper Award. He was the Editor-in-Chief of the IEEE Trans. on Pattern Analysis and Machine Intelligence (1990–94). He is the co-author of Algorithms for Clustering Data, Prentice-Hall, 1988, has edited the book Real-Time Object Measurement and Classification, Springer-Verlag, 1988, and co-edited the books, Analysis and Interpretation of Range Images, Springer-Verlag, 1989, Markov Random Fields, Academic Press, 1992, Artificial Neural Networks and Pattern Recognition, Elsevier, 1993, 3D Object Recognition, Elsevier, 1993, and BIOMETRICS: Personal Identification in Networked Society to be published by Kluwer in 1998. He is a Fellow of the IEEE and IAPR, and has received a Fulbright research award.

About the Author—HONG JIANG ZHANG obtained his Ph.D in 1991 from the Technical University of Denmark and his BS in 1982 from Zhengzhou University, China, both in Electrical Engineering. From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, leading a project on video/image content analysis, representation, indexing and retrieval, which pioneered the field. He joined Hewlett-Packard Labs in October 1995, where he is currently managing the Computation Video Research. His research interests are in video/image content analysis and retrieval, interactive and internet video, image processing and computer vision. He has published about 50 papers and book chapters in these areas, and is a co-author of the book, "Image and Video Processing in Multimedia Systems", published by Kluwer Academic Publishers in 1995. He serves on program committees of many international conferences on multimedia and image processing, and has been an Associate Chair of Program Committee of ACM Multimedia Conferences, 96, 97, and 98. He is a member of the Editorial Board of the international journal "Multimedia Tools and Applications". He also serves as an Associate Editor of the Visual Communication and Image Representation journal. Dr. Zhang is a Senior Member of IEEE and IEEE Computer Society and a Member of ACM.