

Using Structure in Content-based Image Retrieval *

Qasim Iqbal and J. K. Aggarwal
Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712, USA
aggarwaljk@mail.utexas.edu

Abstract

In this paper we present a study of the comparison of the performance of content-based image retrieval systems based on structure [1] with those based on histogram and texture analysis methods, where retrieval is concerned with locating images containing manmade objects. The advantage of using structure in such queries is demonstrated by analyzing an image database containing monocular grayscale outdoor images taken from a ground-level camera to retrieve images containing buildings.

Keywords – *Content-based image retrieval, histogram, texture, structure, nearest neighbor classifier, Bayesian formulation.*

1 Introduction

Advances in computing and signal processing have enabled researchers to extend the image retrieval paradigm for multimedia systems. This extension has brought advances in visual databases, spatial databases, object-oriented databases and the interlinkage of database management and artificial intelligence techniques [2].

Our overall motive is to extend the current stage of content-based image retrieval (CBIR), which is limited to the treatment of lower-level image descriptions, such as histograms of pixel values [3] and texture analysis [4]. The histogram and texture analysis techniques analyze an image at a lower level on a strictly quantitative basis and are unable to capture higher-level scene descriptions that relate different primitive image features with each other. These descriptions will help in image retrieval where queries are concerned with locating images containing manmade objects (such as buildings or architectural objects). Moreover, such descriptions are relatively less sensitive to illumination

changes as compared to lower-level histogram and texture analysis.

It is known that higher-level semantic knowledge, exhibited as the structural information in an image, may be used as effective domain knowledge to isolate potential regions of interest comprised of manmade objects [5]. In our previous work we have successfully developed a retrieval methodology based on structure that detects the presence of manmade objects in an image [1]. In this paper we have undertaken a study to compare the efficacy of using structure to the use of histogram and texture analysis for that purpose.

We extract structure by applying the general principles of perceptual grouping. Perceptual grouping refers to the human visual ability to extract significant image relations from lower-level primitive image features without any knowledge of the image content. It uses such concepts as grouping by proximity, similarity, continuation, closure, and symmetry [5] to group primitive image features into meaningful higher-level image relations.

For the comparison of the three paradigms, viz., histograms, texture and structure, we develop retrieval methodologies for each of them that query a database of monocular grayscale outdoor images taken from a ground-level camera in order to retrieve images that contain buildings. The organization of the rest of the paper is as follows: section 2 describes the framework underlying the formulation of the retrieval methodologies, section 3 outlines the results obtained, and finally, section 4 presents the conclusions.

2 Framework

We assume that the image space consists of three classes, building, non-building, and intermediate, which are denoted as Ω_1 , Ω_2 , and Ω_3 , respectively. The intermediate class is added to account for the fact that in natural outdoor images, some images may be ambiguous and, therefore, difficult to classify even for human operators, and it is convenient to treat them

*This research was supported in part by the Army Research Office under contracts DAAD19-99-1-0012 and DAAG55-98-1-0230, and by the Texas Higher Education Coordinating Board, Advanced Research Project 97-ARP-275.

as belonging to a third class.

Each of these three classes, $\Omega_1, \Omega_2, \Omega_3$, has an associated discriminant function, denoted as g_1, g_2 , and g_3 , respectively. Representing an image classifier in a canonical form through a set of these discriminant functions, the classifier assigns a feature vector \mathbf{X} , and hence, the image from which it is extracted, to class Ω_i if

$$g_i(\mathbf{X}) > g_j(\mathbf{X}), \quad j \neq i, \quad i, j \in \{1, 2, 3\} \quad (1)$$

where ties are resolved arbitrarily.

In the histogram and texture analysis methods, classification is performed using a nearest neighbor classifier, where the structural analysis is performed using a classifier based on Bayesian decision theory. The development of the methodologies outlining the formulation of the appropriate discriminant functions for the three paradigms is described in the next sections.

Our database consists of 150 images of size 640×480 , with 55 building images, 51 non-building images, and 44 intermediate images. For each of the three classes we have employed a total 30 images, with 10 images in each of the three classes, as training images for the individual classifier. The remaining 120 images are used for testing.

2.1 Grayscale histogram

The normalized grayscale histogram extracted from an image is a 256 dimensional vector that is contained in the histogram space $\mathcal{S}_{\mathcal{H}}$ (represented by a unit hypercube), i.e.,

$$\mathcal{H} = (h_0, h_1, \dots, h_{255})^t, \quad h_i \geq 0, \quad 0 \leq i \leq 255 \quad (2)$$

where \mathcal{H} represents the histogram, $h_i = \frac{B_i}{B_T}$ represents the normalized value of the i^{th} grayscale, B_i represents the number of pixels corresponding to the i^{th} grayscale, B_T is the total number of pixels in an image, and $\sum_{i=0}^{255} h_i = 1$. A histogram is extracted from each image of the database. We treat \mathcal{H} as the feature vector \mathbf{X} , i.e., $\mathbf{X} = \mathcal{H}$.

Images are assigned to one of the three classes using the nearest neighbor classifier. The nearest neighbor classifier assigns a pattern histogram (feature vector), \mathbf{X} , to the same class Ω_i , $i \in \{1, 2, 3\}$, as the training feature vector nearest in the histogram space, i.e., \mathbf{X} is assigned to that class which has the highest discriminant value given by equation 1:

$$\delta_{NN}(\mathbf{X}) = \Omega_i \quad \text{if} \quad g_i(\mathbf{X}) > g_j(\mathbf{X}), \quad j \neq i \quad (3)$$

where $\delta_{NN}(\mathbf{X})$ assigns a class label to the test feature vector \mathbf{X} , and $g_k(\mathbf{X}) = -d(\mathbf{X}, \mathbf{X}_k)$, $k \in \{1, 2, 3\}$, is a measure of the distance between \mathbf{X} and the feature

vector $\mathbf{X}_k \in \Omega_k$ which is nearest to \mathbf{X} for the class Ω_k . The distance function $d(\mathbf{X}, \mathbf{X}_k)$ may be selected as the L_2 norm:

$$d(\mathbf{X}, \mathbf{X}_k) = \|\mathbf{X} - \mathbf{X}_k\| = \sqrt{(\mathbf{X} - \mathbf{X}_k)^t(\mathbf{X} - \mathbf{X}_k)} \quad (4)$$

2.2 Texture

Gabor filters have been utilized for texture analysis because they have optimal joint localization (resolution) in both the spatial and the spatial frequency domains. For performing texture analysis the size of each image was adjusted to 512×512 to achieve computational efficiency by the use of fast Fourier transform. The impulse response of an even-symmetric two-dimensional Gabor filter is expressed as:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \cos(2\pi u_0 x) \quad (5)$$

where $f(x, y)$ represents the response at spatial locations x and y , u_0 is the frequency of a sinusoidal plane wave along the \mathcal{X} -axis (i.e., the 0^{th} orientation), and σ_x and σ_y are the spreads of the Gaussian envelope along the \mathcal{X} and \mathcal{Y} axes, respectively. In the spatial frequency domain the above mentioned Gabor filter is given as:

$$F(u, v) = \frac{1}{2} \left[e^{-\frac{1}{2}\left(\frac{(u-u_0)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right)} + e^{-\frac{1}{2}\left(\frac{(u+u_0)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right)} \right] \quad (6)$$

where $F(u, v)$ represents the response at spatial frequency locations u and v , $\sigma_u = \frac{1}{2\pi\sigma_x}$ and $\sigma_v = \frac{1}{2\pi\sigma_y}$ are the spreads of the Gaussian envelopes along the \mathcal{U} and \mathcal{V} axes, respectively. The set of self-similar Gabor filters is obtained by appropriate rotations and scalings of $f(x, y)$ through the generating function:

$$\hat{f}_{mn}(x, y) = a^{-m} f(\hat{x}, \hat{y}), \quad a \geq 1 \quad (7)$$

where $\hat{f}_{mn}(x, y)$ is the rotated and scaled version of the original filter, a is the scale factor, $n = 0, 1, \dots, K-1$ is the current orientation index, K is the total number of orientations, $m = 0, 1, \dots, S-1$ is the current scale index, S is the total number of scales, and \hat{x} and \hat{y} are the rotated and scaled coordinates:

$$\hat{x} = a^{-m}(x \cos \theta + y \sin \theta), \quad \hat{y} = a^{-m}(-x \sin \theta + y \cos \theta) \quad (8)$$

where $\theta = \frac{n\pi}{K}$ is the orientation. The scale factor a^{-m} in equation 7 ensures that the filter energy is independent of m . Representing the upper and lower center frequencies of interest as \mathcal{U}_h and \mathcal{U}_l respectively, the following design ensures that the half-peak magnitude support of the filter responses in the spatial frequency domain touch each other [4]:

$$a = \left(\frac{\mathcal{U}_h}{\mathcal{U}_l}\right)^{\frac{1}{S-1}} \quad (9)$$

$$\sigma_u = \frac{(a-1)\mathcal{U}_h}{(a+1)\sqrt{2\ln 2}} \quad (10)$$

$$\sigma_v = \tan\left(\frac{\pi}{2K}\right) \frac{\mathcal{U}_h - (2\ln 2)(\sigma_u^2/\mathcal{U}_h)}{\sqrt{2\ln 2 - (2\ln 2)^2(\sigma_u^2/\mathcal{U}_h^2)}} \quad (11)$$

A total of 16 Gabor filters are selected with 4 filters in different orientations at 4 different scales, i.e., $K = 4$, and $S = 4$. We select $\mathcal{U}_h = 128\sqrt{2}$ cycles per image width, and $\mathcal{U}_l = 16\sqrt{2}$ cycles per image width, resulting in $a = 2$. Given an image $I(x, y)$, we compute

$$\hat{I}_{mn}(x, y) = \mathcal{F}^{-1}[\mathcal{I}(u, v) F_{mn}(u, v)] \quad (12)$$

where $\mathcal{F}^{-1}[\cdot]$ represents inverse Fourier transform, $\hat{I}_{mn}(x, y)$ represents the filtered spatial image associated with the spatial frequency channel $F_{mn}(u, v)$, $F_{mn}(u, v)$ represents the Fourier transform of $f_{mn}(x, y)$, and $\mathcal{I}(u, v)$ represents the Fourier transform of the image $I(x, y)$. In order to eliminate the sensitivity of the filters to absolute intensity values we set $F_{mn}(0, 0) = 0$. The 16 dimensional feature vector \mathbf{X} is constructed using the fractional energy in each of the 16 spatial channels, i.e.,

$$\mathbf{X} = (\tilde{\mathbf{x}}_{00}, \tilde{\mathbf{x}}_{01}, \tilde{\mathbf{x}}_{02}, \tilde{\mathbf{x}}_{03}, \tilde{\mathbf{x}}_{10}, \tilde{\mathbf{x}}_{11}, \dots, \tilde{\mathbf{x}}_{33})^t \quad (13)$$

where $\tilde{\mathbf{x}}_{mn}$ is given as:

$$\tilde{\mathbf{x}}_{mn} = \frac{\sum_{y=0}^{M-1} \sum_{x=0}^{M-1} \hat{I}_{mn}^2(x, y)}{\sum_{m=0}^{S-1} \sum_{n=0}^{K-1} \sum_{y=0}^{M-1} \sum_{x=0}^{M-1} \hat{I}_{mn}^2(x, y)} \quad (14)$$

where $M = 512$, and $\sum_{m=0}^{S-1} \sum_{n=0}^{K-1} \tilde{\mathbf{x}}_{mn} = 1$. The feature space is represented by a unit hypercube. The feature vectors extracted from the images are classified using the nearest neighbor classifier as described in the previous section, viz., $g_k(\mathbf{X}) = -d(\mathbf{X}, \mathbf{X}_k)$, $k \in \{1, 2, 3\}$.

2.3 Structure

Buildings are manmade objects with sharp edges and straight boundaries. Searching for the highest level features representing the peripheral shape of a building may give inaccurate results because of the large search space. However, the presence of a building in an image will generate a large number of significant edges, junctions, parallel lines and groups, in comparison with an image with predominantly non-building objects. These structures are generated by the presence of corners, windows, doors, boundaries of the building, etc. These intermediate-level features exhibit regularity and relationships, and are strong evidence of structure present in an image.

Straight lines extracted from non-building images are generally randomly distributed. The presence of the distinguishing intermediate-level features mentioned above follow the ‘‘principle of non-accidentalness’’ [6] and, therefore, are more likely to be

generated by buildings. Hence, these features can discriminate between a building image and a non-building image.

We detect the presence of buildings in an unconstrained environment, i.e., with no constraints on the viewing angle and depth, by extracting these intermediate-level features using the principles of perceptual grouping. The following features are extracted hierarchically from an image: *line segments*, *longer linear lines*, *‘‘L’’ junctions*, *‘‘U’’ junctions*, *parallel lines*, *parallel groups*, *‘‘significant’’ parallel groups*. Perceptual grouping rules of similarity, continuity, and parallelism have been used to extract these features. The details of the extraction of these descriptions may be found in [1].

Some of these features are self-explanatory, others are explained in the following. Longer linear lines are obtained by the extension of approximately collinear fragmented line segments that either overlap or are close to each other. The lines obtained are further pruned to eliminate those lines which are very small. All other features are extracted using these longer linear lines. Parallel groups are obtained by putting constraints on the amount of the overlaps of orthogonal projections of parallel lines and projections along the \mathcal{X} and \mathcal{Y} axes, while incorporating differences in local and intrinsic orientation of the lines. ‘‘Significant’’ parallel groups are extracted by further constraining the search to only those parallel groups in which at least one member line is enclosed by an ‘‘L’’ or ‘‘U’’ junction, while accommodating the obliqueness of the viewing angle.

The feature vector \mathbf{X} extracted from each of the images is expressed as:

$$\mathbf{X} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \tilde{\mathbf{x}}_3)^t \quad (15)$$

where

$$\tilde{\mathbf{x}}_1 = \frac{\text{Lines in ‘‘L’’ junctions}}{\text{Total \# of longer linear lines}} \quad (16)$$

$$\tilde{\mathbf{x}}_2 = \frac{\text{Lines in ‘‘U’’ junctions}}{\text{Total \# of longer linear lines}} \quad (17)$$

$$\tilde{\mathbf{x}}_3 = \frac{\text{Lines in ‘‘significant’’ parallel groups}}{\text{Total \# of longer linear lines}} \quad (18)$$

where $\tilde{\mathbf{x}}_i \in [0, 1]$, $i \in \{1, 2, 3\}$, i.e., an image is mapped into a feature space bounded by a unit cube. For minimum-error-rate classification [7] we may set

$$g_i(\mathbf{X}) = \ln[P(\Omega_i|\mathbf{X})] \quad (19)$$

where the *a posteriori* probability of observing the class Ω_i given \mathbf{X} , $P(\Omega_i|\mathbf{X})$, is evaluated by using the

Class	T	R	C	Recall	Precision	
				(C/T)	(C/R)	
Building	45	51	20	44.44%	39.22%	Histogram
Non-building	41	51	18	43.90%	35.29%	
Intermediate	34	18	7	20.59%	38.89%	
Building	45	27	15	33.33%	55.56%	Texture
Non-building	41	42	19	46.34%	45.24%	
Intermediate	34	51	11	32.35%	21.57%	
Building	45	43	36	80.00%	83.72%	Structure
Non-building	41	32	25	60.98%	78.13%	
Intermediate	34	45	21	61.76%	46.67%	

Table 1: Recall and precision for the three paradigms.

Bayes’ rule, which is fundamental in decision theory. Mathematically it is expressed as [7]:

$$P(\Omega_i|\mathbf{X}) = \frac{p(\mathbf{X}|\Omega_i)P(\Omega_i)}{p(\mathbf{X})} \quad (20)$$

where $p(\mathbf{X}|\Omega_i)$ is the class conditional probability density function of \mathbf{X} w.r.t. the class Ω_i , $P(\Omega_i)$ is the *a priori* probability of observing Ω_i , $p(\mathbf{X}) = \sum_{j=1}^N p(\mathbf{X}|\Omega_j)P(\Omega_j)$ is the probability density function of observing \mathbf{X} , and N is the number of classes.

We have assumed that $p(\mathbf{X}|\Omega_i)$ is multivariate Gaussian:

$$p(\mathbf{X}|\Omega_i) = \frac{1}{(2\pi)^{3/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}[(\mathbf{X}-\mu_i)^t \Sigma_i^{-1}(\mathbf{X}-\mu_i)]} \quad (21)$$

where $\mu_i = (\mu_{i_1}, \mu_{i_2}, \mu_{i_3})^t$ is the 3-component mean vector and Σ_i is the 3×3 covariance matrix for the class Ω_i . Parameters μ_i and Σ_i are estimated using maximum likelihood estimation [7] from the feature vectors extracted from the training images. Therefore,

$$g_i(\mathbf{X}) = -\frac{1}{2} \ln |\Sigma_i| - \frac{1}{2}[(\mathbf{X}-\mu_i)^t \Sigma_i^{-1}(\mathbf{X}-\mu_i)] + \ln[P(\Omega_i)] - \ln[p(\mathbf{X})] \quad (22)$$

We assume that the *a priori* probabilities of the three classes are equal. The images present in the database were specifically selected to represent an approximately equal *a priori* distribution.

3 Results obtained

We performed two experiments on the database for each of the three paradigms. The first experiment measured recall and precision. Recall is defined as the fraction of the total number of images in a particular class that are retrieved correctly by the system for that class. Precision is defined as the fraction of images retrieved that actually belong to that class. Images are retrieved by classifying them into one of the

three classes by utilizing their respective discriminant functions, $g_i(\mathbf{X})$'s.

Recall and precision are shown in Table 1 for the histogram, texture, and structural analysis, respectively. The first column shows the three classes. The second, third and fourth columns show the number of images (T) in each of the three classes, the number of images retrieved (R) in the respective classes, and the number of correct images (C) in the set of images retrieved, respectively. As is evident from the table, the recall and precision obtained by histogram and texture analysis are lower than those obtained by structural analysis.

The second experiment retrieved the best matches for each of the three classes and measured the efficiency of the system. The best matches were obtained by sorting the corresponding values of $g_i(\mathbf{X})$ in descending order. The number of images that actually belong to a particular class within the best matches are shown in ranges of 20 images in columns 2 - 7 of Table 2. Efficiency is defined as the number of images (O) that actually belong to a particular class that are obtained in the first T best matches for that class, expressed as a fraction. These values are shown in columns 8 - 10 of the table. It is observed again that the distribution of best matches and efficiency obtained by histogram and texture analysis are inferior to that obtained by using structure.

A grayscale histogram is a global description of an image and lacks the ability to directly relate to spatial locations in the image. Texture deals with the analysis of an image at local scales, but a wide variety of images in all of the three classes may have both smooth textures or rapidly varying textures (e.g. close-up images of a uniform surface and images of vegetation would have smooth and rapidly varying textures, respectively, although both of them belong to the non-building class). Both histogram and texture analysis techniques provide a lower-level quan-

1	2	3	4	5	6	7	8	9	10	
Class	1-20	21-40	41-60	61-80	81-100	101-120	T	O	Efficiency (O/T)	
Building	5	9	8	8	6	9	45	17	37.78%	Histogram
Non-Building	10	8	5	6	7	5	41	18	43.90%	
Intermediate	5	4	7	5	8	5	34	6	17.65%	
Building	13	10	8	9	5	-	45	25	55.56%	Texture
Non-Building	8	11	6	10	4	2	41	20	48.78%	
Intermediate	1	6	6	7	8	6	34	3	8.82%	
Building	19	15	8	2	1	-	45	36	80.00%	Structure
Non-building	17	12	8	3	1	-	41	29	70.73%	
Intermediate	10	9	8	3	3	1	34	17	50.00%	

Table 2: Distribution of correct images in the best matches, and the efficiency of the system for the three paradigms.

titative description of an image. However, buildings have well-defined higher-level spatial relationships exhibited by the lines, corners, junctions and parallel groups. Therefore, structure readily discriminates a building image from a non-building image.

Finally, Figure 1 shows some of the images retrieved by the system that are classified as building images using the structural analysis. As seen from the figure, the results encompass retrieved building images in a wide variety of viewing angles and depths.

4 Conclusions

With recent advances in computing technology content-based image retrieval systems are becoming increasingly useful and desirable. Current techniques in image retrieval do not analyze an image for the extraction of higher-level semantic features that describe the structural content of an image. We have shown that the extraction of semantic features describing the structural content of an image provides an advantage over histogram and texture analysis in methodologies where retrieval is based upon the presence of manmade objects in an image. Structure was extracted by applying the principles of perceptual grouping. We analyzed an image database consisting of monocular grayscale outdoor images taken from a ground-level camera to retrieve building images and compared the results obtained by histogram, texture and structural analysis. The judicious use of domain knowledge exhibited in the form of structure has helped us to achieve higher retrieval performance.

References

[1] Q. Iqbal and J. K. Aggarwal, "Applying perceptual grouping to content-based image retrieval: Building images," in *IEEE Int. Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 42–48, June 1999.



Figure 1: Some of the building images retrieved using structure.

[2] W. I. Grosky and R. Mehrotra, "Image database management," *Computer*, vol. 22, no. 12, pp. 7–8, 1989.

[3] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[4] W. Y. Ma and B. S. Manjunath, "Texture features and learning similarity," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 425–430, 1996.

[5] H. Q. Lu and J. K. Aggarwal, "Applying perceptual organization to the detection of man-made objects in non-urban scenes," *Pattern Recognition*, vol. 25, no. 8, pp. 835–853, 1992.

[6] D. G. Lowe, *Perceptual organization and visual recognition*. Kluwer Academic publishers, 1985.

[7] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.