Perceptual Shape-Based Natural Image Representation and Retrieval

Xiaofen Zheng
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
xzheng@cs.dal.ca

Scott A. Sherrill-Mix
Department of Biological Sciences
Dalhousie University
Halifax, Nova Scotia, Canada
sherrill@mathstat.dal.ca

Qigang Gao
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
qggao@cs.dal.ca

Abstract

Human visual recognition is based largely on shape, yet effectively using shapes in natural image retrieval is a challenging task. Most existing methods are based on the geometric equations of curves computed from processing an entire image. These processes are computationally intensive, lack flexibility and do not take advantage or with minimum use of the Gestalt rules of human vision. By applying certain mechanisms based on the human visual perception process, our methods extract generic shape features from real world images. We extract and group perceptually significant segments and use their properties to create a Euclidean distance matrix for image retrieval. As all the computing is based on simple calculation and one pixel width edges instead of the whole image, this method provides a novel and efficient image feature representation. Testing on standard benchmark datasets and comparison with other well-known methods show this shape analysis method using only compact feature vectors performs well and robustly for real world images.

1 Introduction

Content-based image retrieval (CBIR) is the task of automatically finding images relevant to a query image from the Web or large image datasets using the inherent characteristics of image itself. To successfully achieve this goal, there are two main issues: how to find suitable features to encode image content and how to quantify these features to support efficient similarity measurements.

Finding effective image features is the first step in CBIR. Most existing general-purpose CBIR systems use primitive features, i.e. color, texture and shape.

Color information is relatively easy to extract and calculate and, therefore, is popularly used in CBIR systems. Color histogram and color moments are often used color features. A color histogram represents the distribution of colors in an image, derived by quantifying the pixels in each given set of color ranges. Color histogram matching techniques are discussed in [16]. In [14], a set of moments are extracted based on the chromaticity diagram to represent the frequency and distribution of colors in the image. Compared with the full chromaticity histogram methods, this representation is compact and constant but has a high computational cost.

Texture features provide more spatial or regional information than color features. Tamura et al [17] proposed one of most popular sets that contains six features selected by psychological experiments: coarseness, contrast, directionality, line-likeness, regularity and roughness. The disadvantages of texture-based methods are that they cannot be applied to different classes of texture with reasonable success and some methods involve high computation costs and implementational complexity [1]. The local binary pattern (LBP) [11, 12] is a texture analysis operator which is related to many well-known texture analysis methods.

Some research results suggest that using both color features and spatial relations is a better solution [15]. The SIMPLIcity [20] system classifies images into semantic categories, such as textured-nontextured and graph-photograph, before the retrieval, extracts features according to the semantic class and uses a region-matching scheme based on K-means algorithm that integrates properties of all the re-

0-7695-2997-6/07 $25.00 © 2007 IEEE
DOI 10.1109/ICSC.2007.85

gions in the images. The number of k is adaptively selected by gradually increasing k until a stopping criteria is met. However since the initial cluster assignment is random, different runs of the K-means clustering algorithm may not give the same final clustering solution.

Human visual recognition is largely based on shape. Shape is also important in image sets lacking large color differences, such as medical images [10]. Most real world objects have clear contours which are important clues for recognition. Most shape-based image retrieval techniques rely on Fourier descriptor and moment invariants unrelated to human perception [15, 22].

Some shape-based methods attempt to apply the rules of human perception observed by Gestalt psychologists. Iqbal et al [9] applied perceptual grouping rules to retrieve large manmade objects. Manmade objects generally have sharp edges and straight boundaries which exhibit a large number of significant edges, junctions, parallel lines and polygons. In [2], a retrieval method based on local perceptual shape descriptor and similarity indexing is proposed. Each shape is partitioned into several tokens in correspondence to a set of perceptually salient attributes and each token is represented by its orientation in 2D space. Perceptual shape features have shown some potential with limited data and only global properties [21] but no study using standard natural image datasets and local features has been undertaken.

Shape-based systems usually focus on images with isolated objects in uniform backgrounds. Effectively using edges in natural image retrieval is especially difficult. Most edge feature-based methods emphasize contour simplification in removing noise and other irrelevant features for shape matching [4]. This filtering often results in an inability to utilize texture information and handle the retrieval of real world images.

In this paper, our strategy is to use the Gestalt laws of human vision to develop feature extraction and content representation that can easily be used as the basic elements for qualitative image analysis and image retrieval. We propose a perceptual shape-based image content representation for image retrieval. This method uses both global and local shape features and incorporates textural information from often ignored noise segments. We apply these perceptual features to image retrieval on standard natural image datasets of and compare the results with several other common methods.

2 A Shape-based CBIR Model

In our previous research, a Perceptual Shape Language (PSL) [23, 24] was established which provides tools used to extract the features and forms that users and researchers desire to support their different vision applications, such as motion object tracking [8] and medical image registration [18]. The features consist of lists of generic edge segments and partitioning points. This paper applies these features to represent images and support similarity measuring for retrieval purpose.

The query technique in this CBIR system is query by example, i.e, the user specifies an example image and the image database is searched and compared against this query image. There are two options of providing example image. It can be a normal image provided by the user, or the user can draw a rough approximation using graphical painting tools. The first option is chosen for this system.

Figure 1 shows the PSL solution for CBIR application in this research. The rounded rectangles are the processes and the rectangles show data. The user specifies a query image to search for its best matchings in image dataset. In the offline preprocessing, all the images in database are represented by image feature vectors to be matched against user's query image. Retrieval becomes a matter of measuring the similarity between the feature vectors of the query image and images in the dataset. To represent an image by a feature vector involves two processes. First, PSL functions extract the perceptual edge tokens from the image. This feature partitioning and extraction process is an edge-based parsing. Second, these edge tokens are encoded into a feature vector to describe the image. Image matching compares the vector of the user's query image with vectors from the database images. Image matching includes two processes: similarity computing and ranking. Similarity is computed between the query image and each database image by measuring the distance between their image feature vectors. The distance functions are specified in the system. Usually, the bigger the distance value is, the less similar the two images are. Ranking arranges the database images according to their similarities with the query image and the most similar images are the retrieved image results.

In an information retrieval system, precision is the number of correct images divided by the number of retrieved images and recall is the number of correct images divided by the total number of possible correct images. Mean average precision (MAP) is the average of the precision at each successful retrieval averaged over a query for each image in the database. MAP is an effective measure of retrieval success [19]. The formula of mean average precision is

$$MAP = \frac{1}{N}\sum_{i=1}^{N} AveP_i$$

where N is the number of images in the dataset and $AveP_i$ is the average precision using the image i as the query image. Average precision is the average of the precision after each relevant image is retrieved, which is defined as

$$AveP = \frac{\sum_{r=1}^{M}(P(r) \times rel(r))}{\#\ of\ relevant\ images}$$

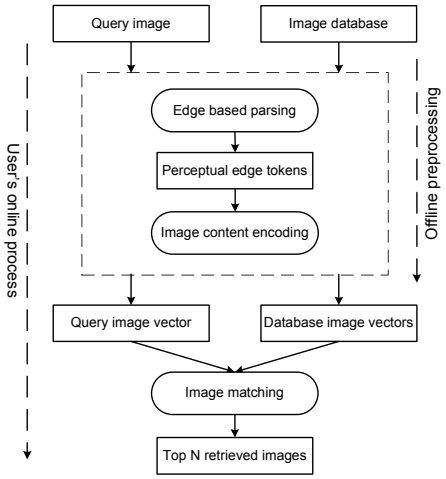


Figure 1. The PSL solution for CBIR application.

where r is the rank, M is the number retrieved, $rel()$ is a binary function on the relevance of a given rank and $P()$ is the precision at a given cut-off rank. Mean average precision provides an index of retrieval performance over all images and retrievals with higher value indicating better retrieval results.

3 Perceptual Shape Parsing and Grouping

Gao and Wong [7] presented a generic curve grouping and partitioning model which allows machines to perform curve segmentations following rules based on the observations of Gestalt psychologists. Generic segments (GSs) are perceptual tokens perceived by humans as atomic features. As shown in Figure 2, GS is a set of points satisfying certain properties which are classified into eight categories according to the tangent functions of GS $y = f(x)$ and its inverse function $x = \varphi(y)$. The computative definitions for these generic segments and the detailed description of this model can be found in [23].

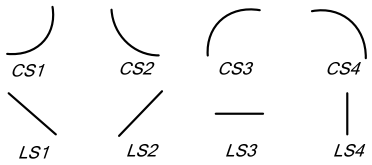


Figure 2. The eight categories of generic segments [7].

A curve partition point (CPP) is a perceptually significant point at which a transition of monotonicity takes place [24]. CPPs best divide edge traces into perceptually atomic tokens, which are useful and meaningful for observation and representation. All these generic perceptual tokens are perceptually distinguishable and can be defined qualitatively. They can easily be used as the basic elements of perceptual organization for qualitative image analysis and image retrieval. For example, vertical lines and horizontal lines commonly occur in the structures of manmade objects. This important clue for the image retrieval is included into the perceptual edge tokens. As shown in Figure 2, clusters of $LS3$ and $LS4$ are the groups of vertical lines and horizontal lines in an image and the segments in each cluster of $LS3$ and $LS4$ are parallel to each other. Figure 3(a) shows an example of parsing an image into perceptual edge tokens and their clusters. This parsing process divided images into perceptual edge tokens.

Due to the complexities of images, such as natural objects or scenes, contour-based methods often over-segment these images. Regular shape contours are mixed with texture edge fragments and background noise. GSs with continuous gradient change are often the true boundaries of objects. Usually noise has rapid random changes of gradient and direction along the edge and exhibits severe discontinuity on the strength and smoothness of the edge. A classification method to distinguish segments into noise and GSs are proposed in [24]. In most cases, noise edges are unwanted in contour-based object recognition methods, but groups of noise edges may capture texture patterns in images which would be useful features. Therefore, GSs and noise are used to represent the image content.

Image feature vectors contain the values that encode the image content. In this paper, the feature vectors are called perceptual shape features which are built upon GSs and noise edges provided by PSL functions.

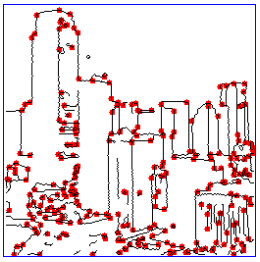
4 Perceptual Shape Feature-based Image Content Representation

Perceptual shape features are built in a generic way and both the global properties and the local properties of the features are used to represent an image. There are three considerations when building the perceptual shape features: the types of segments to consider, how to measure the significance of each type of edges in the image and the spatial groupings of the features.

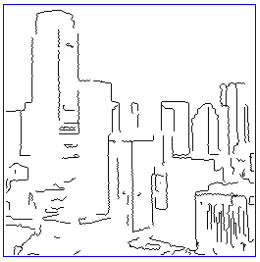
We propose a short and a long feature vector to represent the images. The short vector includes the significance measures of the segments but omits the spatial grouping for the sake of conciseness. The long vector contains both the significance information and the spatial distribution of the segments.

For this study, we used all eight GSs and noise segments.

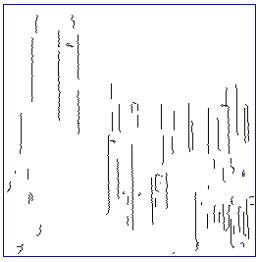
(a) The graylevel original image.



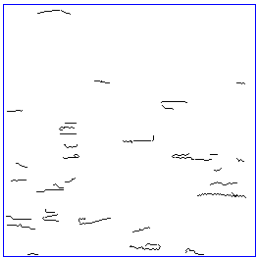
(b) The edge segments and CPPs.



(c) GSs.



(d) Cluster of LS3 segments.

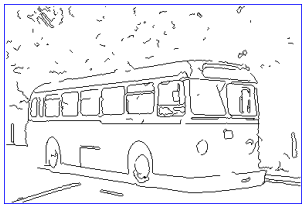


(e) Cluster of LS4 segments.

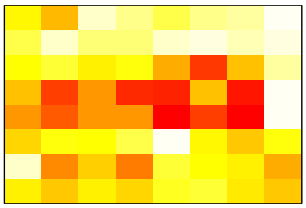
Figure 3. Illustration of parsing an image into perceptual edge tokens and clusters of segments.

(a) Graylevel original image.

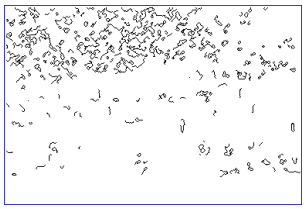
(b) Edges and CPPs.



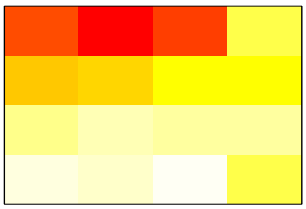
(c) GS edge map.



(d) 8×8 GS regions.



(e) Texture map.



(f) 4×4 noise regions.

Figure 4. An example image and its perceptual shape representation illustration. In (d) and (f), darker color indicate more significance of segments in the region.

Their significance was represented by the number of noise segments, the total length of noise segments and the total length of each GS type.

The length of a curve segment or noise is counted as the number of the pixels on that segment. Since the images are digitized as a mesh grid, the extracted straight lines are zigzagged slightly. Therefore the length of a line segment is calculated as the distance between its two end point pixels. For an image, short vector SV can be expressed as:

$$SV = \{\#(NS), L(NS), L(LS_i), L(CS_i) | i = 1, ..., 4\}$$

where $\#(NS)$ is the number of noise segments, $L(NS)$ is the total length of noise segments, $L(LS_i)$ is the total length of straight line segments of category LS_i, $L(CS_i)$ is the total length of curve segments CS_i and i denotes the type of the generic segment. The resulting short vector has 10 elements.

The long vector combines the global properties of the short vector with local spatial properties of the edges. We created a GS edge map, shown in Figure 4(c), by dividing the image into 8×8 regions. The noise features were divided into 4×4 regions to create a texture map, shown in Figure 4(e). For an image, the long vector LV can be expressed as:

$$LV = \{SV, R_i(GS), R_j(NS) | i = 1, ..., 64, j = 1, ..., 16\}$$

where SV is the short vector of this image, i and j are the indexes of the regions, $R_i(GS)$ is the significance value of the segments in region i in GS edge map, $R_j(NS)$ is the significance value of the noise in region j in texture map. For this study, the significance value of the segments in a region is the number of pixels of the segments falling in the region. The spatial information is therefore encoded in 64 GS spatial bins, shown in Figure 4(d) and 16 texture spatial bins, shown in Figure 4(f). The resulting long vector has 90 elements. These regions could be adjusted to best represent the images in particular datasets.

Images are compactly represented for retrieval purposes using these perceptual feature vectors. To distinguish from feature vectors of other methods, for the remainder of the paper we refer to the short vector as Perceptual Shape Features (10) and the long vector as Perceptual Shape Features (90).

5 Experiments

To investigate the potential applications of perceptual shape features, we conducted image retrieval experiments on two datasets. For comparison, we also calculated Fourier transform based features[3], LBP [11] and intensity histogram representing the basic features of shape-based methods, texture-based methods and color-based methods. In all the experiments, the similarity of images was measured by Euclidean distance. The Euclidean distance d between two image feature vectors $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ is

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

where n is the number of the elements in an image feature vector, p_i and q_i are the elements in feature vectors.

5.1 Experimental Setup

The Wang dataset [20], available from `http://wang.ist.psu.edu/docs/related/`, has 1000 images. There are 10 categories: Africa people and village, Beach, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and glaciers, Food. Each category has 100 images. The image size is 256×384 pixels or 384×256 pixels.

The Oliva dataset [13], available from `http://cvcl.mit.edu/database.htm`, has 2688 urban and natural scene images. This includes 8 categories: Coast and beach, Open country, Forest, Mountain, Highway, Street, City center, Tall building. The image size is 256×256.

Figure 5 and Figure 6 show a few example query results from the two datasets. The query image and the corresponding ten most relevant images are shown in each retrieval. The average precision of the query image is also given.

5.2 Comparison Study

The Fourier feature is suggested as a low-level shape feature for content-based image retrieval in [3]. The Fourier transform is performed on the 256×256 normalized image. The Fourier spectrum is low pass filtered and then decimated by factors of 16 or 32 resulting in 128 and 32 element feature vectors.

(a) Image query of 671.

(b) Image query of 13.

Figure 5. Two sample queries and the corresponding top ten retrieved images in the 1000 image Wang dataset. Top image: user's query image. Bottom ten images: the retrieved images.

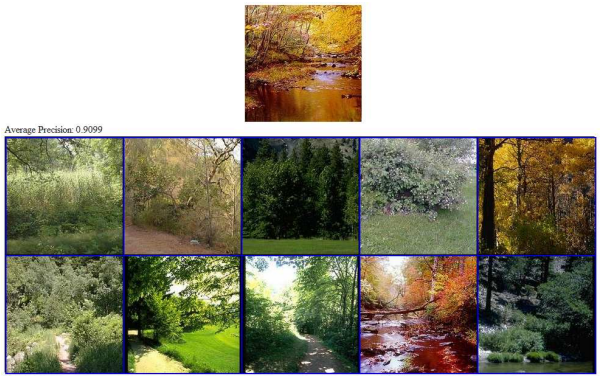
The local binary pattern (LBP) [11, 12] is a texture analysis operator which is invariant to monotonic changes in gray scale. Two-dimensional distributions of the LBP and local contrast measures are used as features. A binary code that describes the local texture pattern is built by thresholding a neighborhood by the gray value of its center. The LBP operator is related to many well known texture analysis methods.

Grayscale histogram comparisons convert the image from the true color RGB images to grayscale intensity images by eliminating the hue and saturation information while retaining the luminance. Images are then represented by 256 element vectors where each element corresponds to the number of the pixels with that gradient value.

Figure 7 and Figure 8 are the precision vs. recall results with perceptual shape features, Fourier features, LBP texture features and gray histogram features implemented on the Oliva dataset and on the Wang dataset.

Table 1 shows the number of elements of each feature and the mean average precision of using perceptual shape

(a) Image query of a804068.



(b) Image query of natu169.

Figure 6. Two sample queries and the corresponding top ten retrieved images from the 2688 image Oliva dataset. Top image: user's query image. Bottom ten images: the retrieved images.

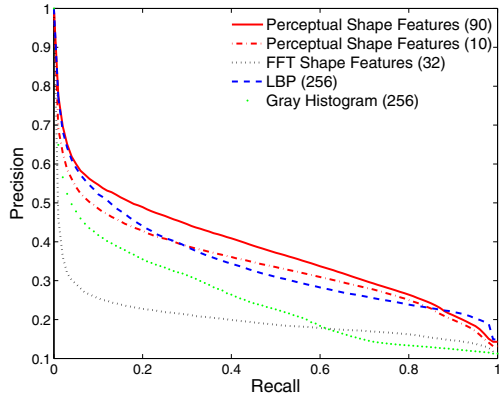


Figure 7. Average precision-recall of 1000 retrievals on the Wang dataset.

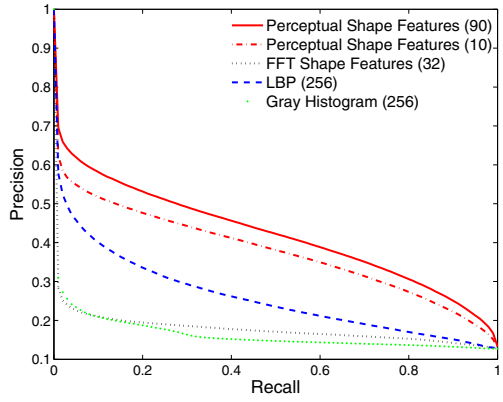


Figure 8. Average precision-recall of 2688 retrievals on the Oliva dataset.

features, Fourier features, LBP and gray histogram for image retrieval on the two datasets.

Perceptual shape features with 90 elements have the best retrieval results on these two datasets. Perceptual shape features with only 10 elements also performed well. In the Wang dataset, perceptual shape features (10) and LBP performed equivalently, but in the Oliva dataset perceptual shape features (10) outperformed although LBP has about 25 times more elements in its features than perceptual shape features (10).

5.3 Further Discussion

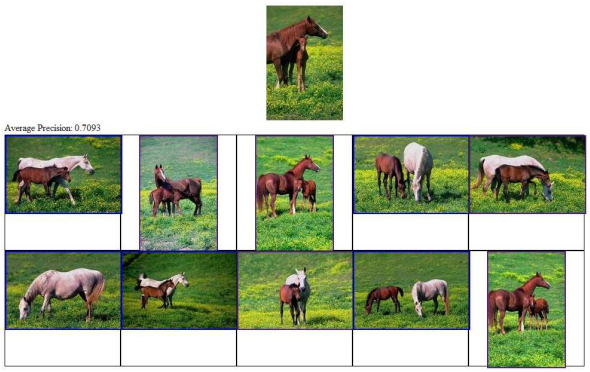
Perceptual shape features (10) provides a concise representation of image, requires only a small amount of storage and may allow faster matching when a dataset is especially large. To develop a powerful image retrieval system, adding more information, like perceptual shape features (90), improves performance. Additional features built up on perceptual edge tokens, such as joints, closure and structure forms, could be added.

Perceptual shape features focusing on contour and texture are based on the edges extracted in the image. Figure 9 gives two image queries. One query is good, while another one shows the limitation of only using edges in CBIR. In this example, the contours of the overlapped horses and elephants are not clearly distinguishable although they belong to different categories in the Wang dataset. This similarity unavoidably influences the retrieval precision of using perceptual shape features. One of the possible solutions is to define the common colors of elephants and horses and therefore to distinguish them. The intention of this paper was to explore the usefulness of perceptual shape features. How to combine them with other useful information is open

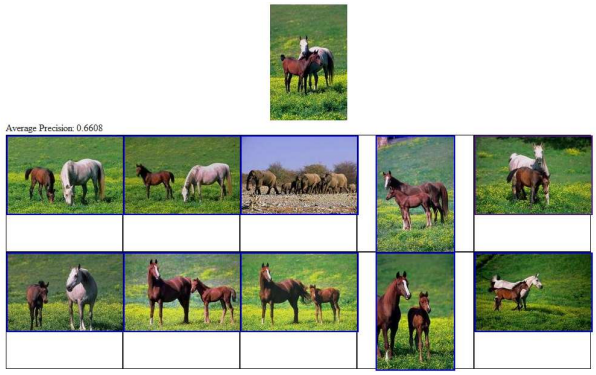
Features	# of Elements	MAP of the Wang Dataset	MAP of the Oliva Dataset
Perceptual Shape Features	10	0.342	0.377
Perceptual Shape Features	90	0.379	0.420
Fourier Features	32	0.198	0.176
Fourier Features	128	0.194	0.170
LBP	256	0.343	0.259
Gray Histogram	256	0.250	0.162

Table 1. Features and their mean average precisions (MAP) on two datasets.

for future research.



(a) Image query of 777.



(b) Image query of 778.

Figure 9. Two sample queries and the corresponding ten retrieved images in the Wang dataset. Top image: user's query image. Bottom ten images: the retrieved images.

The Wang dataset is a widely used benchmark dataset for CBIR. Deselaers et al. [6] compared nine methods of image retrieval, including Wang's SIMPLIcity [20] method, on the Wang dataset (Table 2). They used an alternative comparison of image retrieval, the error rate, which correlates well with other measures of recall and precision [5]. Error rate is $1-P(1)$ where $P(1)$ is the precision for the first successful retrieval averaged over the entire dataset [5].

Feature	ER (%)
inv. feat. histogram	15.9
color histogram	17.9
pixel values (IDM)	22.3
Tamura histogram	31.0
local feature histogram	32.5
Gabor histogram	48.2
regions (SIMPLIcity)	54.3
pixel values (Euclidean)	55.1
local features	62.5
Perceptual Shape Features (90)	22.4
Perceptual Shape Features (10)	30.1

Table 2. Error rates [%] of different retrieval methods (See [6] for details) for the Wang dataset.

Although our method uses fewer features and no color information, this perceptual shape feature method compares well with the texture and color-based methods reported by [6]. No other shape feature method has been reported on the Wang dataset.

The Oliva dataset consists of urban and natural scenes categories. Originally it was built to study human and computational abilities at real world scene understanding. At the time of this study, there appears to be no image retrieval experiments conducted on this dataset. To illustrate the robustness and efficiency of using perceptual shape feature in natural scenery image retrieval, the experiments were conducted on this dataset as well. Results were also positive on this dataset (Error Rate for 90 elements: 20.4% and 10 elements: 26.8%). These experiments demonstrate that shape-based methods have potential in real world image retrieval applications.

6 Conclusions

This work is based on perceptual edge features and can be used in the lowest level of query. The perceptual shape

features are simple characterizations of images but can handle complex images retrieval, such as natural scenes. Retrieval performances on two datasets show the features are concise and robust for supporting general purpose content-based image retrieval. This study used only a simple similarity measure and performed well without optimization of weighting for each element in the feature vector. For specific tasks, it may be possible to determine appropriate weights with a training dataset. As these methods are shape-based, they are robust to color change or grayscale images and may have potential use in other tasks, such as medical image or satellite image analysis when color information is limited. In addition, these shape-based methods may have potential synergies with color-based methods.

References

[1] M. Amadasun and R. King. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man and Cybernetics*, 19(5):1264–1274, October 1989.

[2] S. Berretti, A. D. Bimbo, and P. Pala. Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Transactions on Multimedia*, 2(4):225–239, Dec 2000.

[3] S. Brandt, J. Laaksonen, and E. Oja. Statistical shape features for content-based image retrieval. *Journal of Mathematical Imaging and Vision*, 17:183–194, 2002.

[4] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval - approaches and trends of the new age. In *ACM International Workshop on Multimedia Information Retrieval, ACM Multimedia*, Singapore, 2005.

[5] T. Deselaers, D. Keysers, and H. Ney. Classification error rate for quantitative evaluation of content-based image retrieval systems. In *17th International Conference on Pattern Recognition*, pages 505–508, Washington, DC, USA, 2004. IEEE Computer Society.

[6] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval - a quantitative comparison. In *DAGM'04: 26th Pattern Recognition Symposium*, LNCS, Tbingen, Germany, Sep 2004.

[7] Q. Gao and A. Wong. Curve detection based on perceptual organization. *Pattern Recognition*, 26(1):1039–1046, 1993.

[8] Q. Gao, Y. Zhang, and A. Parslow. Motion stream analysis based on perceptual feature partitioning and grouping. In *7th International IEEE Conference on Intelligent Transportation Systems*, pages 575–579, 2004.

[9] Q. Iqbal and J. K. Aggarwal. Retrieval by classification of images containing large manmade objects using perceptual grouping. *Pattern Recognition*, 35:1463–1479, 2002.

[10] H. Mller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, Feb 2004.

[11] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29:51–59, 1996.

[12] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, Jul 2002.

[13] A. Oliva and A. B. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[14] G. Paschos, I. Radev, and N. Prabakar. Image content-based retrieval using chromaticity moments. In *IEEE Transactions on Knowledge and Data Engineering*, volume 15, pages 1069–1072, Piscataway, NJ, USA, 2003. IEEE Educational Activities Department.

[15] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, April 1999.

[16] M. Stricker and M. Swain. The capacity of color histogram indexing. In *Computer Vision and Pattern Recognition*, pages 704–708, 1994.

[17] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. In *IEEE Transactions on Systems, Man and Cybernetics.*, volume SMC-8, pages 460–473, 1978.

[18] Y. Tao and Q. Gao. Vessel junction detection from retinal images. In *16th International Conference on Vision Interface*, pages 388–394, Halifax, Canada, June 2003.

[19] E. Voorhees. Information retrieval: Roots and future directions. In *IS&T/SPIE 11th Annual Symposium on Electronic Imaging*, 1999.

[20] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[21] M. Wu and Q. Gao. Content-based image retrieval using perceptual shape features. In *Proceedings of International Conference on Image Analysis and Recognition*, pages 567–574, Toronto, Canada, 2005.

[22] D. Zhang and G. Lu. Study and evaluation of different Fourier methods for image retrieval. *Image and Vision Computing*, 23(1):33–49, 2005.

[23] X. Zheng and Q. Gao. Generic Edge Tokens: Representation, Segmentation and Grouping. In *16th International Conference on Vision Interface*, pages 423–430, Halifax, Canada, June 2003.

[24] X. Zheng and Q. Gao. Efficient edge noise removal and perceptual feature classification. *ICGST International Journal on Graphics, Vision and Image Processing - Special Issue on Edge Detection and Tracking*, pages 1–8, 2006.