published in the Proceedings of ICPR2000, Barcelona, Spain, September 2000
available at <URL:http://www.cis.hut.fi/picsom/publications.html>

Statistical Shape Features in Content-Based Image Retrieval

Sami Brandt
Helsinki University of Technology
Laboratory of Computational Engineering
P.O. BOX 9400, FIN-02015 HUT, Finland
Sami.Brandt@hut.fi

Jorma Laaksonen and Erkki Oja
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. BOX 5400, FIN-02015 HUT, Finland
Jorma.Laaksonen@hut.fi, Erkki.Oja@hut.fi

Abstract

In this article the use of shape features in content-based image retrieval is studied. The emphasis is on such techniques which do not demand object segmentation. PicSOM, the image retrieval system used in the experiments, requires that features are represented by constant-sized feature vectors for which the Euclidean distance can be used as a similarity measure. The shape features suggested here are edge histograms and Fourier-transform-based features computed for an edge image in Cartesian and polar coordinate planes. The results show that both local and global shape features are important clues of shapes in an image.

1. Introduction

Content-based image retrieval has been an active research area since the early 1990's. Many image retrieval systems, both commercial and research, have been built. The best known are Query By Image Content (QBIC) [3] and Photobook [10] and its new version FourEyes. Other well-known systems are the search engine family VisualSEEk, MetaSEEk and WebSEEk [1, 11], NETRA [9], Multimedia Analysis and Retrieval System (MARS) [4].

In the above systems, image content is stored in visual features which can be divided into four classes according to the properties they describe. The classes are color, texture, shape, and structure. Color and texture contain important information but, for instance, two images with similar color histograms can represent very different things. Therefore the use of shape-describing features is essential in an efficient content-based image retrieval system. Although shape description has been intensively researched, there exists no direct answer as to which kind of shape features should be incorporated into such a system.

A major problem in automatic feature extraction is segmentation. Even if it were known that there is a single object in the image, it is in general a non-trivial problem to locate it. On the other hand, when there are no specific objects the result of segmentation is probably an irrelevant part of the original image. For the use of a general database of images, such as the World Wide Web, it might then be reasonable to use some statistical shape features for the whole image instead. However, in the current literature these kinds of shape features are an exception (see [2] for a thorough review). Indeed, all the above systems having shape features rely on manual segmentation. Only NETRA has a fully automated segmentation algorithm, whose results indicate the extreme difficulty of the problem.

Another basic concept to be considered in selecting an appropriate shape description technique is whether some invariant properties such as transformation, rotation, and scaling invariances are needed. The use of these is not always beneficial because they reduce the discrimination power of the features.

2. PicSOM

The aim of this work was to design shape features for a content-based image retrieval system PicSOM [7, 8] which is going to be used as an image search engine for large-scale databases like the World Wide Web. The system is based on the Tree Structured Self Organizing Maps [6] which are used as the indexing structure of the images. Ideally, images that are similar to each other with respect to a particular feature extraction method, should cluster together on the corresponding map. The responses on the different maps are combined in such a way that the most relevant features are automatically weighted as the query proceeds. The incorporated relevance feedback mechanism thus adapts the system to the user's preferences until he/she finds the preferred images from the database. An on-line version of PicSOM can be found at `http://www.cis.hut.fi/picsom/`.

In this paper, due to the general characteristics of shape features, the feature extraction methods were tested directly in the feature space, rather than experimenting on them in the PicSOM system.

3. Statistical Shape Features

The experiments study a few kinds of statistical features which do not require segmentation but are computed from the shape properties of the whole image. The question whether the invariance to the affine transformations is beneficial or not is also addressed.

3.1. Histogram of Edge Directions

The first experiments on shape features were made by using the histogram of edge directions. The edge histogram is translation invariant and it captures the general shape information in the image. Because the feature is local, it is robust to partial occlusion and local disturbance in the image. The major disadvantage is that two perceptually very different images may have similar edge histograms.

The edge histogram is computed as follows. At first, the color image is transformed to the HSI space from which the hue channel is neglected. The other two channels are convolved with the eight Sobel operators. The resulting gradient images are next thresholded to binary images by a proper threshold value for each channel. The threshold values are manually fixed to certain levels which are the same for all images. The thresholded intensity and saturation gradient images are combined by the logical OR operation. The threshold value for the intensity gradient image was manually set to 15% of the maximum gradient value and for the saturation image to 35%. In the OR operation, the direction of the larger gradient value is chosen. Finally the 8-dimensional edge histograms are calculated by counting the edge pixels in each direction. Still, it is necessary to normalize the histograms somehow. Our experiment showed that it is better to normalize the histograms by the number of pixels in each image rather than by number of edge pixels as was done in [5].

The effect of smoothing proposed in [5] was also studied. The smoothing should make the histograms more robust to rotation. It is performed as follows:

$$\mathcal{H}^{\mathrm{s}}(i) = \frac{\sum_{l=i-k}^{i+k} \mathcal{H}(l)}{2k+1}, \tag{1}$$

where $\mathcal{H}(l)$ stands for the original edge histogram and the parameter k determines the degree of smoothing.

3.2. Co-occurrence Matrix of Edge Directions

The edge histogram can yet be generalized. By taking every neighboring edge pixel pair and enumerating them based on their directions a two-dimensional histogram or co-occurrence matrix is obtained. The resulting 64-dimensional histogram is normalized by the number of pixels in the image and is defined as

$$\mathcal{H}_{\mathrm{co}}(i,j) = \frac{1}{NM} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \mathcal{I}_i(x,y) \sum_{(\hat{x},\hat{y}) \in U(x,y)} \mathcal{I}_j(\hat{x},\hat{y}), \tag{2}$$

where $\mathcal{I}_i$ is the binary edge image of direction i, $i = 1,\ldots,8$, $j = 1,\ldots,8$, and $U(x,y)$ is the causal neighborhood set of the pixel (x,y). In the above expression, the last sum is the number of edge pixels in the direction j in the neighborhood of each pixel (x,y). Hence, the resulting value after all the summations indicates the number of neighboring edge pixel pairs which are positioned in the directions i and j.

3.3. Fourier Features

The edge image contains the most relevant shape information and the discrete Fourier transform can be used to describe it. Before forming the edge image, the image area is normalized to a maximum size of 512×512 such that the aspect ratio is maintained. If both dimensions of the original image are larger than 512, the image is decimated with filtering. Otherwise the normalization is made by bicubic interpolation.

After edge detection, the Fourier transform is computed for the normalized image using the FFT algorithm. The magnitude image of the Fourier spectrum is first low-pass filtered and thereafter decimated by the factor of 32. The resulting number of dimensions in the feature vectors is 128. The final reduction is made after the edge detection and FFT because the resolution of the edge detection procedure would not be sufficient to extract relevant edges from a decimated image.

3.4. Polar and Log–Polar Fourier Features

The Fourier features described above are translation invariant but not rotation invariant. Our method, which is named as polar Fourier features, is rotation invariant with respect to the center of the image but not invariant to translation and scale.

At first the image is normalized and the edge image is obtained similarly as with the Fourier features. The binary edge image is then transformed to the polar coordinates by the formula

$$\mathcal{I}(\rho,\theta) = \mathcal{I}_1(\rho,\theta) \oplus \mathcal{I}_2(\rho,\theta), \tag{3}$$

where $\rho = 0, \frac{1}{512}, \ldots, \frac{511}{512}$, $\theta = 0, \frac{1}{512}, \ldots, \frac{511}{512}$, $\oplus$ means the binary OR operation, and

$$\mathcal{I}_1(\rho,\theta) = \mathcal{I}(\lfloor R\rho\cos\theta\rceil + c_x, \lfloor R\rho\sin\theta\rceil + c_y) \tag{4}$$

$$\mathcal{I}_2(\rho,\theta) = \begin{cases} 1 & \text{if } \rho = \frac{1}{512}\lfloor \frac{512}{R}\sqrt{(x-c_x)^2+(y-c_y)^2}\rceil, \\ & \theta = \frac{1}{1024\pi}\lfloor 512(\angle(x-c_x, y-c_y)+\pi)\rceil, \\ & \text{for some } (x,y) \text{ for which } \mathcal{I}_\mathrm{e}(x,y)=1, \\ & x = 0,1,\ldots,511, y=0,1,\ldots,511 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where (c_x, c_y) are the coordinates of the centroid located at the center of the image, $R = \sqrt{c_x^2 + c_y^2}$, and $\lfloor\cdot\rceil$ means rounding to the nearest integer. The above procedure prevents the formation of gaps between the edge pixels in the polar coordinate system.

For the polar image the Fourier transform and decimation are performed similarly as with the Fourier features and a 128-dimensional feature vector is obtained. The method is invariant to translation in the polar plane, and therefore rotation invariant with respect to the center of the image and translation invariant along the radius from the center.

Even more invariances can be obtained by a slight modification to the feature. Log–polar Fourier features are invariant to affine transformations i.e. to translation, rotation and scaling. These can be obtained by replacing R and ρ in (4) and (5) with $\ln R$ and e^ρ, respectively. Translation invariance is obtained by setting the centroid (c_x, c_y) to the center of mass of the binary edge image. Rotation invariance is obtained by using the magnitude spectrum of the log–polar transform as the rotations affect only the phase of the spectrum. Accordingly, the invariance for scale is obtained by taking logarithm of the radius in the polar coordinate plane.

All the Fourier-based features presented here are sensitive to occlusion: the direct use of the Fourier transform may lead to very different magnitude spectra for occluded images. In addition, if some parts of an image are missing, the calculation of the centroid will go wrong and significantly differing log–polar images will result.

4. Evaluation Methods

The presented shape features are evaluated with methods from [7]. Let N be the total number of images in the image database $\mathcal{D}$. Let $\mathcal{C} \subset \mathcal{D}$ be a *class* of similar images determined by some criteria.

In order to measure the clustering of the features the concept of *observed probability* $P_{\mathrm{o},\mathcal{C}}(n)$ is defined. It is the probability that an image $\mathcal{I} \in \mathcal{C}$ has as the n^{th} closest neighbor an image which also belongs to the class $\mathcal{C}$. For each image $\mathcal{I}_j \in \mathcal{C}$ the distance to every other image is calculated and the images are sorted in the order of ascending distance. For each image $\mathcal{I}_j$ we thus define a sequence $h_j(n)$ as

$$h_j(n) = \begin{cases} 1 & \text{if the } n^{\mathrm{th}} \text{ closest image belongs to } \mathcal{C}, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $n = 1, \ldots, N-1$.

The observed probability $P_{\mathrm{o},\mathcal{C}}(n)$ for the class $\mathcal{C}$ having $N_\mathcal{C}$ images is then defined as

$$P_{\mathrm{o},\mathcal{C}}(n) = \frac{1}{N_\mathcal{C}} \sum_{\mathcal{I}_j \in \mathcal{C}} h_j(n), \quad n = 1, \ldots, N-1. \tag{7}$$

In the optimal case $P_{\mathrm{o},\mathcal{C}}(n) = 1$ if $n \leq N_\mathcal{C} - 1$ and 0 otherwise. On the other hand the worst case performance results when $P_{\mathrm{o},\mathcal{C}}(n)$ equals the *a priori* probability of class $\mathcal{C}$ for all n.

4.1 Local Performance

Good features should have high observed probabilities for the very first indices. Therefore the *local* measure is defined as the average of the observed probability for the first 1% of indices, i.e.

$$\eta_\mathrm{local} = \frac{100 \sum_{n=1}^{0.01N} P_{\mathrm{o},\mathcal{C}}(n)}{N}. \tag{8}$$

The local performance measure has the ability to describe if the feature space is clustered such that there is a high probability that any image of the class $\mathcal{C}$ has an image of the same class near it in the feature space. Note that η_local is dependent on the *a priori* probability of the class.

4.2 Global Performance

The performance measure presented above gives a high performance value even if the images of the class $\mathcal{C}$ are clustered to many small clusters all over the feature space. This suggests using an appropriate weighting function $w(n)$ which would take global clustering into consideration. The observed probability function can be converted to a scalar by using the sum expression

$$\eta_\mathrm{global} = \sum_{n=1}^{N-1} P_{\mathrm{o},\mathcal{C}}(n)\, w(n). \tag{9}$$

It is considered that the weighting function $w(n)$ should be such that it rewards large $P_{\mathrm{o},\mathcal{C}}(n)$ in small indices and punishes the large values in large indices. For this reason the weighting function was chosen to be

$$w(n) = \frac{\cos\frac{(n-1)\pi}{N-2}}{\sum_{l=0}^{N_\mathcal{C}-2} \cos\frac{l\pi}{N-2}}, \quad n = 1, \ldots, N-1. \tag{10}$$

Then η_{global} approaches one for the optimal observed probability case and zero for the *a priori* case. In addition, η_{global} is ideally independent of the *a priori* probability.

When the weighting function is chosen this way it measures how the feature vectors of the class tend to form one global cluster in the feature space.

5. Experiments

As a test database a set of 4350 miscellaneous bitmap images was used. The images were downloaded from `ftp://ftp.sunet.se/pub/pictures`. The ground truth classes were manually picked from the database. The classes used in the experiments were *aircraft*, *buildings*, and *faces*, of which the database contains 348, 492, and 361 images, respectively. The results are summarized in Table 1 where the best results in each column are bolded.

In comparing the shape features the results showed that there is no single feature which would perform best for every image class used in the experiments. Of the local features, which are robust to occlusion and to any local disturbance in the image, both the histogram and the co-occurrence matrix of edge directions gave good results. It can also be seen that the smoothing of the edge histogram provided no advantage (EdgeHist vs. EdgeHistSm). From the global Fourier-transform-based features the magnitude spectrum of an edge image was found to work at least as well as the local shape features with the test database. In addition, the results of the experiments on the three different Fourier-based features support the assumption that with a general database the presence of all the three invariances is not beneficial.

Table 1. Performance of different features. Larger values indicate better performance.

$\eta_{\text{local}},\ 0 \leq \eta_{\text{local}} \leq 1$			
Feature	aircraft	buildings	faces
EdgeHist	0.04 ±0.02	0.07 ±0.02	0.014±0.004
EdgeHistSm	0.04 ±0.01	0.07 ±0.02	0.015±0.005
Co-occurrence	**0.06 ±0.02**	0.08 ±0.02	0.018±0.004
FFT128	0.04 ±0.01	**0.12 ±0.02**	**0.033±0.008**
PolarFFT	0.023±0.006	0.040±0.009	0.026±0.009
LogPolarFFT	0.015±0.004	0.022±0.007	0.022±0.008

$\eta_{\text{global}},\ 0 \leq \eta_{\text{global}} \leq 1$			
Feature	aircraft	buildings	faces
EdgeHist	0.43± 0.08	0.3 ±0.1	0.2 ±0.1
EdgeHistSm	0.40± 0.08	0.25±0.09	0.2 ±0.1
Co-occurrence	**0.48± 0.08**	0.16±0.08	**0.3 ±0.1**
FFT128	0.4 ± 0.1	**0.35±0.06**	**0.3 ±0.2**
PolarFFT	0.3 ± 0.1	**0.4 ±0.1**	0.2 ±0.1
LogPolarFFT	0.14± 0.05	0.23±0.09	0.19±0.09

6. Conclusions

In this work shape-describing features for general content-based image retrieval were studied. We formed various types of statistical feature vectors from the edges in non-segmented images. The best results were obtained with decimated magnitude spectrum of the edge image. Also local edge-histogram-based features, including the co-occurrence matrix of edge directions, gave good results. This indicates that both local and global information are important clues of the image shape. The experiments also suggest that with a database of miscellaneous images it is not reasonable to require the features to be invariant to affine transformations.

References

[1] M. Beigi, A. Benitez, and S.-F. Chang. MetaSEEk: A content-based meta search engine for images. In *Storage and Retrieval for Image and Video Databases*, SPIE Proceedings Series, San Jose, CA, 1998.

[2] S. Brandt. Use of shape features in content-based image retrieval. Master's thesis, Helsinki University of Technology, 1999.

[3] M. Flickner, H. Sawhney, W. Niblack, et al. Query by image and video content: The QBIC system. *IEEE Computer*, 28:23–31, September 1995.

[4] T. S. Huang, S. Mehratra, and K. Ramchandran. Multimedia analysis and retrieval system (MARS) project. In *Proceedings of the 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval.* University of Illinois at Urbana-Champaign, March 1996.

[5] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, 1996.

[6] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences.* Springer-Verlag, second extended edition, 1997.

[7] M. Koskela. Content-based image retrieval with self-organizing maps. Master's thesis, Helsinki University of Technology, 1999.

[8] J. Laaksonen, M. Koskela, and E. Oja. Content-based image retrieval using self-organizing maps. In *Proceedings of the Third International Conference on Visual Information and Information Systems, VISUAL'99*, pages 541–548, Amsterdam, The Netherlands, June 1999.

[9] W. Y. Ma and B. S. Manjunath. NETRA: A toolbox for navigating large image databases. In *Proceedings of IEEE International Conference on Image Processing*, volume I, pages 925–928, Santa Barbara, California, October 1997.

[10] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[11] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the ACM Multimedia '96*, November 1996.