Image retrieval based on region shape similarity

Cheng Chang Liu Wenyin Hongjiang Zhang

Microsoft Research China, 49 Zhichun Road, Beijing 100080, China

{wyliu, hjzhang}@microsoft.com

ABSTRACT

This paper presents an image retrieval method based on region shape similarity. In our approach, we first segment images into primitive regions and then combine some of the primitive regions to generate meaningful composite shapes, which are used as semantic units of the images during the similarity assessment process. We employ three global shape features and a set of normalized Fourier descriptors to characterize each meaningful shape. All these features are invariant under similar transformations. Finally, we measure the similarity between two images by finding the most similar pair of shapes in the two images. Our approach has demonstrated good performance in our retrieval experiments on clipart images.

Keywords: Content-based image retrieval, shape features, shape similarity, Fourier descriptors

1. INTRODUCTION

The popularity of digital images is rapidly increasing due to significant progresses made in digital imaging technologies and high-volume secondary storage technologies. More and more digital images are becoming available every day. However, the abundance of images underscores the absence of an automatic capability of effective and efficient image retrieval, which is still an open problem puzzling lots of researchers.

Among other image retrieval methods, content-based image retrieval[4] (CBIR) is an approach that exclusively relies on the visual features, such as color histogram, texture, shape, and so forth, of the images. One of the obvious advantages of CBIR over other methods, e.g., text-based image retrieval, is that CBIR can be done in a fully automatic process since the visual features are automatically extracted. While text-based image retrieval assumes that all images are labeled with text. This process is known as image annotation. Since automatic generation of descriptive keywords or extraction of semantic information for images requires machines to understand images in general domains, which is beyond the capability of current computer vision and intelligence technologies, image annotation is usually done by humans. This is a labor-intensive process and therefore may be tedious, subjective, inaccurate, and incomplete.

However, CBIR also suffers a low retrieval precision. Among others, one main reason is that many CBIR systems handle each image as an entire semantic unit. This is usually not true since there are at least two different things—foreground and background—and usually there are several more meaningful objects coexisting in the same image. In order to retrieve those images containing the content of interest, each object should be treated as an individual semantic object during the image retrieval process. In this case, there should be some effective ways to describe these objects and region-based image retrieval has been proposed. Some region-based image retrieval systems just simply divide the entire image into several regular, and usually, overlapped regions and treat each region as a single image. Others, such as blobworld[3], just use some regular and roughly homogeneous (with respect to color or texture) regions instead of segmented regions to represent semantic units of the images. They have not solved the fundamental issue of multiple semantic objects.

Psychological experiments have shown considerable evidence that natural objects are primarily recognized by their shapes[2]. However, it is quite hard for machines to understand images as human beings do because automatic shape segmentation from general complex images is still one of the most difficult problems in machine vision. Even though images can be well segmented based on similar color or texture features, these primitive regions are usually less useful than their combinations, which represent meaningful objects. Shape-based image retrieval methods are therefore greatly dependent on how well the meaningful shapes are segmented from the images. In addition, shape similarity assessment also relies on the selection of discriminative shape features. Both problems challenge the success of shape-based image retrieval approaches.

In this paper, we present an image retrieval method based on region shape similarity and apply it to retrieval of clipart images. The key idea is to first determine some dominant and meaningful regions in an image based on region segmentation and mergence. Dominant regions are often the most important in presenting the semantic content of such clipart images. In addition, clipart images can be easily segmented into a limited number of primitive regions, each of which consists of a uniform color. Since the number of primitive regions is usually very small, it is possible to examine all combinatorial and connected regions of these primitives. Image similarity between two images is then evaluated based on the shape similarity between these combinatorial regions in the two images. We use a set of concise shape features, including eccentricity, compactness, solidity, and normalized Fourier descriptors, to measure the shape similarity. As we show in our experiments, the image similarity assessment method based on region shape similarity is effective and efficient to find similar clipart images.

The rest of the paper is organized as follows. In Section 2, we present the region segmentation and mergence approach to obtain composite and meaningful shapes. In Section 3, we present the shape features used in shape similarity assessment. We show some preliminary experimental results in Section 4 and finally, we conclude in Section 5.

2. REGION SEGMENTATION AND MERGENCE

2.1. Primitive Region Segmentation

First of all, we need to segment an image into a set of primitive regions based on pixel similarity. Generally, image segmentation is a subjective task and is difficult for machines to perform well. Fortunately, we focus on segmentation of clipart images. Since an individual clipart image usually consists of a limited number of regions, each of which contains (almost) uniform pixel values, we choose a straightforward region growing method among many color image segmentation techniques in existence and apply it to region segmentation of clipart images. In our application, a primitive region is a connected region, in which the pixel variation of each color component in the RGB color space is less than a predefined threshold.

The number of primitive regions generated using this straightforward way may be very large due to over-segmentation and many of them may be very small. Hence, we limit the number of primitive regions in a single image to a small number k and remove other smaller regions. Another reason for limiting the number of regions is to avoid the combinatorial explosion issue in the subsequent region mergence process based on the adjacency of primitive regions. Suppose we have k primitive regions, we may obtain more than 2^{k-1} merged regions in the worst case. It is not realistic to handle so many combinations if k is very big.

2.2. Region Mergence for Meaningful Shapes

After we get the segmented primitive regions, we have to merge some of them into meaningful shapes, which are semantic objects in the image. For simplicity, we require that each meaningful shape should also be connected. In order to test the connectivity of each subset of primitive regions, we first build the connectivity graph represented by its adjacency matrix for all these primitive regions and then test the connectivity of the sub-matrix containing corresponding elements.

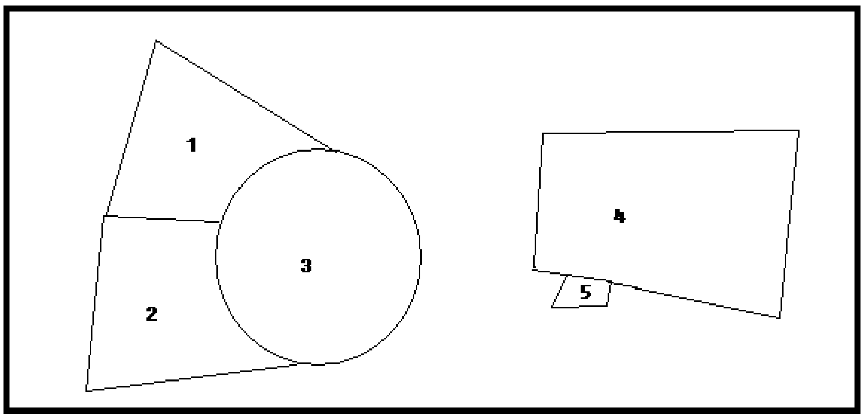
Suppose we obtain k primitive regions from the region segmentation process, we build a k-dimension adjacency matrix A, where

$A(i,j)=0$, if the *ith* region is not connected with the *jth* region, and

$A(i,j)=1$, if the *ith* region is connected with the *jth* region or $j=i$.

We use S to denote the entire set of these primitive regions. If we want to judge whether a subset of S is connected, we only need to extract the corresponding elements of A and form a new adjacency matrix B. If B is connected, we can combine the subset to obtain a merged region, which may be a meaningful shape to human vision. We test the connectivity of B by counting the number of elements in a connected component of B. We can find such a connected component using the breadth-first search strategy in a graph traversal starting from its first element. If the number of elements in the connected component resulted from the traversal is exactly the dimension number of B, we can say that it is connected. Otherwise, B is not connected.

Figure 1 is an example of region segmentation and mergence in our application. In Figure 1(a), five primitive regions, labeled 1, 2, 3, 4, and 5, respectively, are yielded from image segmentation. Region 5 is removed since it is too small to attract human vision attention. Hence, only regions 1, 2, 3, and 4 remain and form the adjacency matrix A in Figure 1(b), the sub-matrixes of which are used to test the possibility of region mergence. Among all of the 16 possible combinations of the 4 primitive regions, we finally obtain 8 meaningful shapes. They are 1, 2, 3, 1-2, 2-3, 1-3, 1-2-3, and 4. The contours of these merged regions are used in the shape similarity assessment of this image and others.



$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(a) (b)

Figure 1. Illustration of region segmentation and possible combinations of primitive regions, (a) primitive regions (b) the adjacency matrix of these primitive regions.

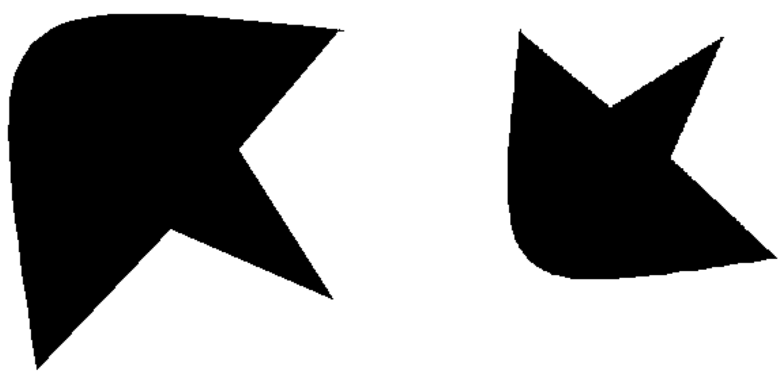


Figure 2. Illustration of two shapes that look very similar each other under similar transformations.

3. SHAPE FEATURES AND SHAPE SIMILARITY ASSESSMENT

After we obtain the meaningful shapes of the images, we measure the shape similarity between two images using a set of shape features and a shape similarity model defined in this Section. In our application, the shape features we used include eccentricity, compactness, solidity, and normalized Fourier descriptors. The first three features are global features to characterize shapes in the overall sense[7]. Fourier descriptors (FDs) are local geometric features to characterize details of shapes, which are more accurate but more noise-sensitive[7]. All of these features are invariant under similar transformations, including translation, rotation, and scaling. The reason why we use similar transformation invariants is that, in most cases, human judges two shapes as identical if one can be obtained from the other by using some similar transformation, as exemplified in Figure 2. While if the shearing coefficient of an affine transformation is big enough, those two shapes are often considered as different.

Based on these features, we define the shape similarity of two region objects using the distance model and define the shape similarity between two images as the shape similarity between a pair of the most similar meaningful regions from the two images.

3.1. Extraction of Shape Features

The shape of a region is represented using a polygon (a closed chain of points) obtained by tracing along the region's border. We further simplify the border polygon using the polygonal approximation algorithm developed by Sklansky and Gonzalez[8] to remove noises and redundant points from the polygon. The remaining points are enough to describe the contour. The number of vertexes of the simplified polygon is usually much smaller than that of the original one. The simplified polygon is used to calculate the shape features. Hence, the computation time of shape features is significantly reduced.

We represent the simplified polygon of a shape using its vertex sequence $P_0, P_1 \ldots, P_N$ { (x_0, y_0), (x_1, y_1)…(x_N, y_N) } (where $P_0 = P_N$). The shape features are calculated using the following formulas, respectively.

(1) Eccentricity is defined in Eq. (1).

$$Eccentricity = \frac{I_{\min}}{I_{\max}} = \frac{u_{20} + u_{02} - \sqrt{(u_{20} - u_{02})^2 + 4u_{11}^2}}{u_{20} + u_{02} + \sqrt{(u_{20} - u_{02})^2 + 4u_{11}^2}}, \tag{1}$$

where, $u_{p,q} = \sum_{x,} \sum_{y} (x - \bar{x})^p (y - \bar{y})^q$ is the (p, q) order central moment of the shape (($\bar{x}, \bar{y}$) is the center of the shape) and can be calculated from the polygon vertexes using the efficient method proposed by Leu[6]. As can be seen from Eq. (1), eccentricity is in fact the ratio of the short axis' length ($I_{\min}$) to the long axis' length ($I_{\max}$) of the best fitting ellipse of the shape.

(2) Compactness is defined in Eq. (2).

$$Compactness = \frac{4\pi A}{P^2}, \tag{2}$$

where, P is the perimeter of the polygon and A is the area of the polygon. Compactness expresses the extent to which a shape is a circle. A circle's compactness is 1 and a long bar's compactness is close to 0.

(3) Solidity is defined in Eq. (3).

$$Solidity = \frac{A}{H}, \tag{3}$$

where, A is the area of the polygon and H is the convex hull area of the polygon. Solidity describes the extent to which the shape is convex or concave. The solidity of a convex contour is always 1.

(4) Normalized Fourier descriptors

The above three simple features are used to characterize the region's global and overall shape. In order to discriminate two shapes in detail, we introduce a set of normalized Fourier descriptors (NFDs), which are also invariant under similar transformations.

Fourier descriptors (FDs)[5] are the coefficients of the discrete Fourier transform, which are resulted from the frequency analysis, of a shape. Although they are invariant of translation and orientation, they are not scaling-invariant. Similarly to the method of Arbter et al.[1], we normalize Fourier descriptors and make the normalized Fourier descriptors also invariant of scaling. The set of Fourier descriptors proposed by Arbter et al. are invariant under affine transformations[1] and are in complex forms. Since we only need to use some NFDs that are invariant under similar transformations, they can be defined more concisely as follows.

First of all, we normalize the length of the shape contour to 1 and express its polygon vertexes as $p(l) = x(l) + jy(l)$, where, $l = \int_c dt \Big/ \oint_c dl$ is the normalized parameter. We then calculate continuous integrals, as shown in Eq. (4), on all the edges of the polygon to obtain the NFDs.

$$z(k) = \oint_c p(l)e^{-j2\pi kl}\,dl = \int_0^1 p(l)e^{-j2\pi kl}\,dl = \sum_{n=0}^{N-1} \int_{l_n}^{l_{n+1}} p(l)e^{-j2\pi kl}\,dl, \tag{4}$$

where, l_0=0 and l_N=1.

Theoretically, shapes can be fully recovered from their Fourier descriptors. However, for real life shapes, the high frequency Fourier descriptors correspond most likely to noises and distort the shape. We therefore use only some low frequency NFDs of the whole set. Among all 256 (which is also the total number of points yielded from the parametric discretization of the original shape contour[7]) NFDs, we use only $z(k)$ (k=1..12) in the shape similarity assessment process in our application.

In summary, the feature vector f used in our application to characterize a shape includes 15 elements. $f(1)$ represents eccentricity, $f(2)$ represents compactness, $f(3)$ represents solidity, and $f(4)$~$f(15)$ represent the 12 normalized Fourier descriptors.

3.2. Shape Similarity Assessment

Given two regions, their shape similarity is measured as the distance between their shape feature vectors, as shown in Eq. (5).

$$d(S_1, S_2) = \sum_{i=1}^{15} w(i) \times \left\| f_1(i) - f_2(i) \right\|, \tag{5}$$

where, $f_1(i)$ and $f_2(i)$ are the *ith* components of the feature vectors of shapes S_1 and S_2, respectively. $w(i)$ is the weight of the *ith* feature component in the distance model, which can be either Euclidean distance, or city-block distance (as used in our experiments), or some other forms. The weight can be adjusted such that Eq. (5) produces the best result.

Based on the above defined shape similarity, we define the region shape similarity between two images I_1, I_2 as follows.

$$d(I_1, I_2) = \underset{i,j}{Min}\ d(S_1, S_2), \tag{6}$$

where, $S_1(i)$ is the *ith* meaningful region in image I_1, $S_2(j)$ is the *jth* meaningful region in I_2, $d(S_1, S_2)$ is defined in Eq. (5). Eq. (6) means that the region shape similarity of two images is the shape similarity of the most similar pair of regions between the two images. In other words, we consider two images as similar if and only if the two images contain similar meaningful regions. The reason why we made this assumption is that, without prior knowledge, we cannot tell which region among others should represent the image's semantics. Under this assumption, the most similar one among all the regions will be considered suitable to represent the image. The similarity function $d(S_1, S_2)$ in Eq. (6) can also be a general function of the shape similarities of all similar region pairs between the two images. A possible alternative is the average of the shape similarities of the top N most similar region pairs.

4. EXPERIMENT RESULTS

In our experiments, we apply the region shape similarity of images defined in Eq. (6) to clipart image retrieval. Our test image database contains 150 clipart images of various types selected from the Corel Gallery. Given a query image, the retrieved images are ranked by their similarities to the query.

Figure 3 shows our experiment on finding star-like clipart images. In Figure 3, the leftmost image (with label 1) is the query image containing a complex form of pentagram and others are retrieved images. The label on top-left corner of each image is the rank of the image according to its shape similarity to the query. For the image with label 2, which is the most similar image to the query, we first segment it into five primitive regions and then obtain 26 combinatorial regions in total. One of them is the pentagram containing all these five primitive regions. Therefore, it is the most similar image to the query. It is interesting that in the image with label 5, the green leaf of the carrot is very similar to the pentagram. The image is therefore in top ranks.

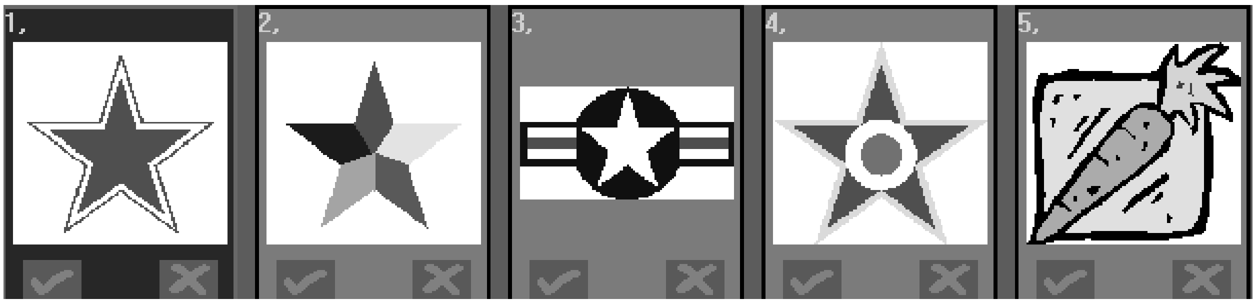


Figure 3. Clipart image retrieval result of finding star-like clipart images.

Figure 4 shows the clipart image retrieval result for a query with arrowhead shapes. In Figure 4, the leftmost image is the query image. All the images that have arrowhead shapes are found and ranked to the top of the result list. Obviously, the result is reasonable.

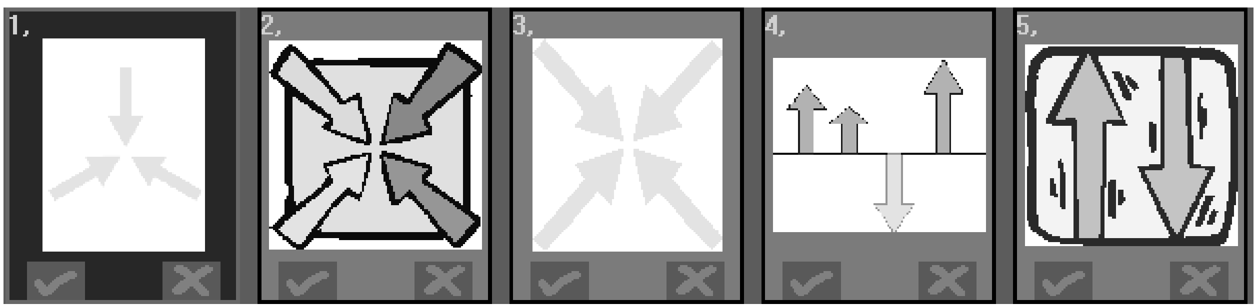


Figure 4. Clipart image retrieval result for a query with arrowhead shapes.

Figure 5 shows the clipart image retrieval result for a query with a circle and a triangle. In Figure 5, the top-left image is the query image. Images with labels 2, 7, 10, and 15 contiain only equilateral triangles while other images may contain circles or both. In this example, the ranking may not be satisfactory according to some people due to subjectivity, which also shows the difficulty of CBIR.

5. CONCLUDING REMARKS

In this paper, we presented an image retrieval approach based on region shape similarity between images and applied it to clipart image retrieval. Its performance is good for simple color images, such as those clipart images, each of which contains only a few simple regions. However, there are two potential problems with the method in handling more complex images.

The first problem is that it is quite hard for machines to determine meaningful regions. It is impossible for machines to automatically extract all of those meaningful regions identical to what a human being would do. Even different people may find different interesting shapes from the same image, as exemplified in Figure 5. Both under-segmentation and over-segmentation do harm to the determination of interesting shapes. In the case of under-segmentation, some interesting regions cannot be found and therefore cannot be evaluated in the image similarity assessment process. In case of over-segmentation, too many combinations can be generated and may mislead shape similarity assessment. Some small shapes that are not interesting to human beings may have very similar features to the query. These small shapes may be really similar to the regions of query due to scaling. Or, it may also be due to the second problem—the validity of similarity assessment models in existence. Although many features and similarity models are proposed, none of them has been proved to be identical to the human vision model, which is considered as very complicated and involves many spontaneous and subconscious processing tasks.

Hence, the success of region-shape based image retrieval systems for general images heavily depends on the success of image segmentation and feature-based similarity assessment techniques.

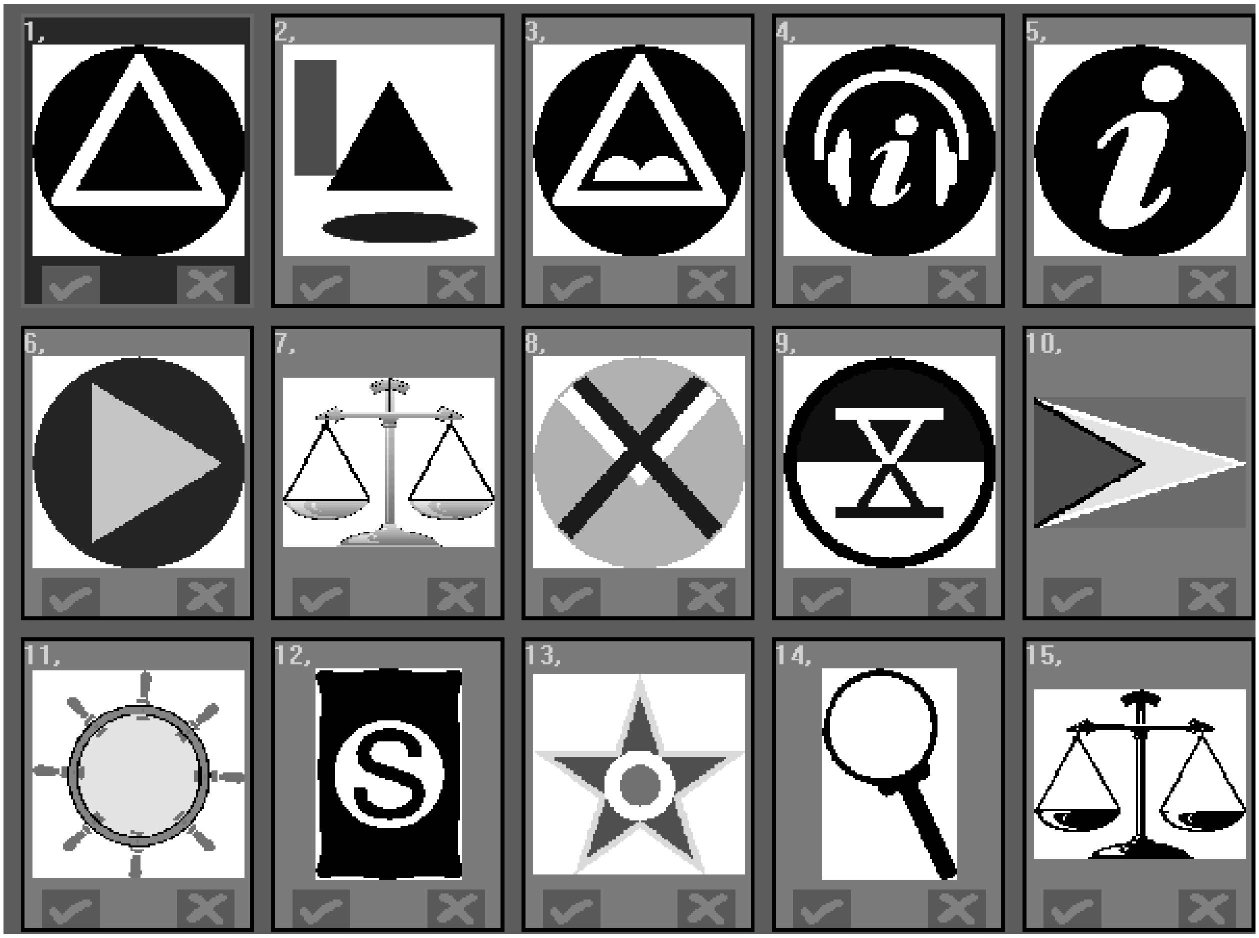


Figure 5. Clipart image retrieval result for a query with circle and triangle.

ACKNOWLEDGEMENT

The authors thank Mr. Tao Wang for his help on development of the shape-based image retrieval framework.

REFERENCE

1. Arbter K et al. (1990) Application of Affine-Invariant Fourier Descriptors to Recognition of 3-D Objects. *IEEE Trans On PAMI* 12(7):640-647
2. Biederman I (1987) Recognition-by-Components: a Theory of Human Image Understanding. *Psychological Review* 94(2):115-147
3. Carson C et al. (1997) Region-based Image Querying. In: *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 42-49.

4. Flickner M et al. (1995) Query by Image and Video Content. *IEEE Computer* 28(9):23-32.
5. Jain AK (1989) *Fundamentals of Digital Image Processing*, Prentice-Hall.
6. Leu J-G(1991) Computing a Shape Moments from Its Boundary. *Pattern Recognition* 24(10):949-957
7. Liu W et al. (2000) A Hierarchical Characterization Scheme for Image Retrieval. In: *Proc. of ICIP*.
8. Sklansky J and Gonzalez V (1980) Fast Polygonal Approximation of Digitized Curves. *Pattern Recognition* 12:327-331