Contour-Based Large Scale Image Retrieval

Rong Zhou, and Liqing Zhang

MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
rongzhou@sjtu.edu.cn,
zhang-lq@cs.sjtu.edu.cn

Abstract. The paper presents a contour-based method for large scale image retrieval. With the contour saliency map of the object, it could address the shift-invariance problem, and with hierarchical and multi-scale feature extraction, it is able to deal with the scale-invariance problem to a certain extent. Different from existing algorithms, the features used in the retrieval system contain not only local information, but also global information of the object. By taking advantage of this characteristic, we could build a hierarchical index structure which helps to fast retrieval of the large scale database. Furthermore, our method allows two kinds of query image: a hand-drawn sketch or a natural image. Thus it is possible to refine the search results by choosing one image from the list of previous sketch retrieval results as the new query. It brings the better interactive user experiment and the convenience for those who aren't good at drawing. The experiment results verify the performance of our method on a database of four million images.

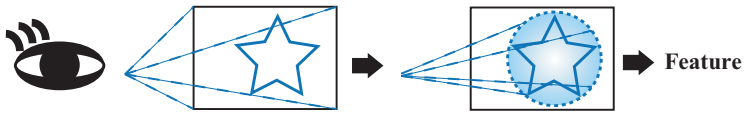
Key words: shift-invariance; contour saliency map; hierarchical structure; global-to-local feature; orientation information

1 Introduction

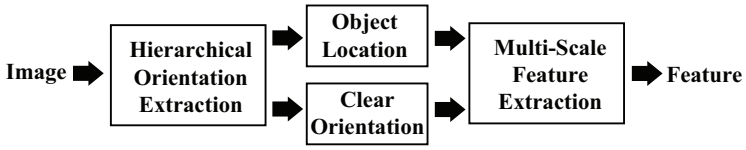
Contour is a very important channel for human being to recognize or distinguish the objects from an image or a scene. Image retrieval based on contour has been attracted great attention in the data mining society [1], but most of works mainly dealt with image retrieval in small database [2][3][4]. And in large scale database, Eitz [5] presented a method that divides an image into a fixed number of cells, and each cell corresponds to a structure tensor descriptor which stores the main direction of the gradients of the cell. Different from Eitz's method which hasn't index structure and must scan the whole database for each query, Cao [6] presented an index-able oriented chamfer matching method. But both their works have the same limitation that the shift-invariance problem still exists in their retrieval system. The objects in the query image and in the retrieved image must have the same position, this property will reduce dramatically the recall rate in image retrieval. And in most of situation, the users usually only

mind whether they could search the object they wish, and don't care where it is in the image.

To address this issue, we propose a shift-invariance method for large scale image retrieval. It comes from the fact that when human beings see an image, they usually look through the whole image for a short while and then focus their eyes on the salient place of the image, as shown in Fig. 1(a). That means in most of time, people only pay attention to a local part of an image instead of the whole image. So different from existing algorithms, we don't extract features on the whole image, instead, we first find the saliency map of the object which is usually a local part of the image, and then extract features on the part.



(a) Image processing of human visual system.



(b) Framework of feature extraction.

Fig. 1. Framework of feature extraction by simulating human visual system.

Another contribution is a contour-based image retrieval prototype system for the database including more than four million images. With hierarchical and multi-scale feature extraction, we could easily obtain not only the position of the object but also the global-to-local orientation features, which brings two advantages: shift-invariance and scale-invariance to a certain extent. These cannot be achieved by most existing retrieval systems. Moreover, the system provides users two query methods: a hand-drawn sketch or a natural image, as shown in Fig. 2. If you are a good painter, you could draw a sketch whatever you imagine, but if the sketch doesn't like what you imagine very much, you can select a natural image which is most similar to what you wish from the list of retrieval results and then make the second retrieval to achieve satisfactory images.

2 Feature Extraction

Fig. 1(b) demonstrates the basic framework of feature extraction of our method. By simulating hierarchical information processing of human visual system, it could obtain a contour saliency map of the object in an image, and at the same time, it could still extract clear orientation information of an object. With the

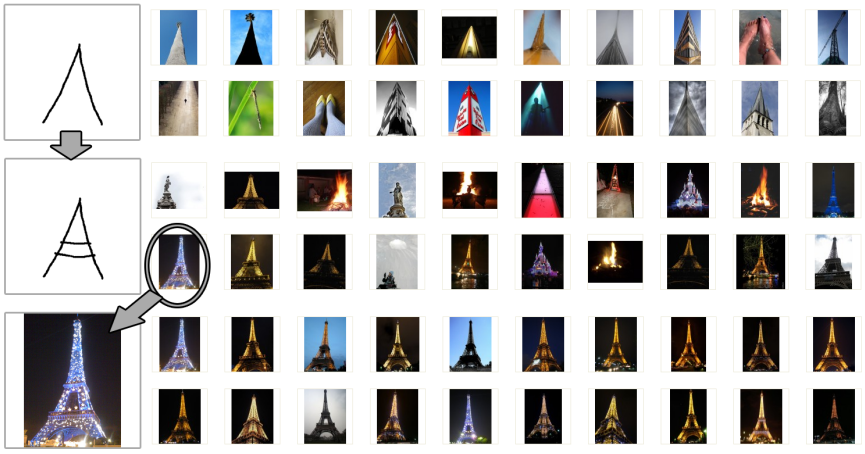


Fig. 2. Illustration of interactive retrieval process. After querying with a hand-drawn sketch, the users could choose one result image as a query and make the next retrieval.

two above mentioned, we can easily know the object's position and contour orientation information. And then, by multi-scale feature extraction, we can obtain the global-to-local feature of the object.

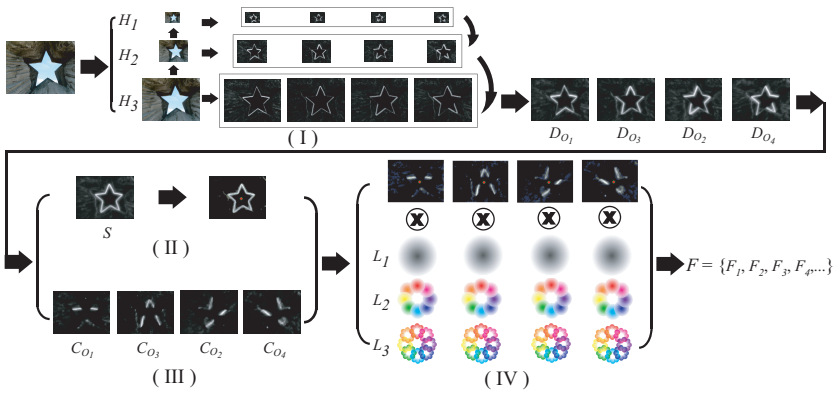


Fig. 3. Details of feature extraction. (I) Hierarchical orientation extraction. (II) Object location. (III) Clear orientation. (IV) Multi-scale feature extraction.

It is well known that human visual system processes the image with hierarchical structure. According to this, hierarchical difference image D_{O_j} and the contour saliency map S are computed from an image as:

$$S = \sum_{j=1}^{N} D_{O_j} = \sum_{j=1}^{N} \left(\sum_{i=1}^{M} [\max_{k} \{D_{H_i O_j}\}]_{m \times n} \right) \tag{1}$$

where $D_{H_i O_j}$ is the difference image of the ith level and the jth orientation, as shown in Fig. 3. And the size of $D_{H_{i-1} O_j}$ is lower than the size of $D_{H_i O_j}$ by two times. k is the red, green, blue color channel, and $\max_k\{\cdot\}$ is the maximum of difference image among the three channels. $[\cdot]_{m \times n}$ means scaling the difference image proportionably to the maximum size m×n. S is the contour saliency map of an image and is normalized to between 0 and 1. In Fig.3, $M = 3$, and $N = 4$, O_j denotes 0, $\pi/4$, $\pi/2$, and $3\pi/4$ orientation respectively. Because the minimum resolution of images for human beings is 32×32 [7], we set the maximum size of $D_{H_1 O_j} is$ 32×32, and then m×n is 128×128.

$$S_x = \arg\max_{x} \{sum(\lfloor S \rfloor_{T_s})_x \star g_x\}, \quad S_y = \arg\max_{y} \{sum(\lfloor S \rfloor_{T_s})_y \star g_y\} \tag{2}$$

where $\lfloor \cdot \rfloor_{T_s}$ denotes the value greater than T_s, in our experiment, $T_s = 0.25$. $sum(\cdot)_x$ and $sum(\cdot)_y$ are the sum along the axis x and the axis y respectively, g is the Gaussian kernel, and $\star$ denotes convolution. (S_x, S_y) are coordinates of the maximum convolution value in the saliency map, and they denote the centroid of the object in the image.

From Fig. 3 we can see, D_{O_j} cannot represent contour orientation information of the object clearly. Considering 0 and $\pi/2$, $\pi/4$ and $3\pi/4$ are orthogonal respectively, we make the following operation:

$$\begin{aligned} C_{O_1} &= \lfloor D_{O_1} - D_{O_3} \rfloor_0, C_{O_3} = \lfloor D_{O_3} - D_{O_1} \rfloor_0 \\ C_{O_2} &= \lfloor D_{O_2} - D_{O_4} \rfloor_0, C_{O_4} = \lfloor D_{O_4} - D_{O_2} \rfloor_0 \end{aligned} \tag{3}$$

where C_{O_j} denotes clear orientation map.

The final feature of an image is:

$$F_{L_p O_j t} = \sum C_{O_j}(x_{L_p t}, y_{L_p t}, r_{L_p}) \cdot G(r_{L_p}) \tag{4}$$

where $F_{L_p O_j t}$ denotes the tth feature of the L_pth level and the O_jth orientation, and $G(r_{L_p})$ is the Gaussian kernel which radius is r_{L_p}, and $C_{O_j}(x, y, r)$ is the region of the clear orientation map which centroid is (x, y) and the radius is r, and $r_{L_p} = 2r_{L_{p+1}}$, $r_{L_3} = 32$. When $p = 1$, $t \in \{1\}$, and when $p = 2$, $t \in \{1, 2, ..., 8\}$, and when $p = 3$, $t \in \{1, 2, ..., 64\}$. So the feature $F = \{F_{L_p O_j t}\}$ has $1 \times 4 + 8 \times 4 + 8 \times 8 \times 4 = 4 + 32 + 256 = 292$ dimensions. And finally, values of the feature are normalized to between 0 and 1.

So the similarity measure of two images is given by:

$$Dist(F, F^{'}) = sim\left(\{F_{L_p O_j t}\}, \{F^{'}_{L_p O_j t}\}\right) \tag{5}$$

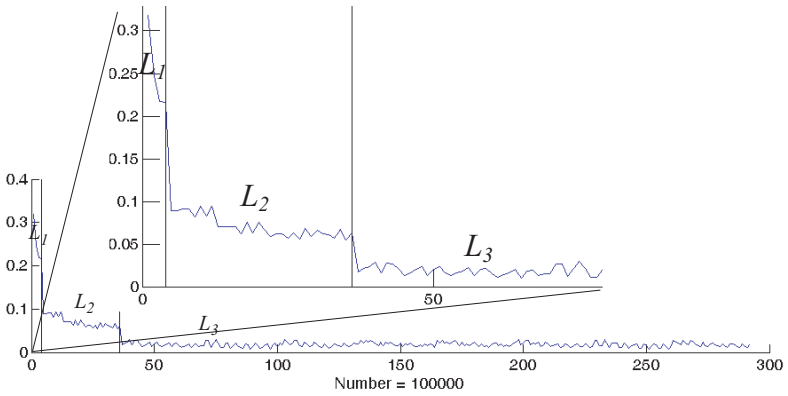


Fig. 4. Histogram of average feature values from 100,000 images. The first 4 values of F belong to region L_1, the first 5 to 36 values of F belong to region L_2, and the rest 256 values of F belong to region L_3.

where $sim(\cdot)$ could be any similarity measure, for example, Euclidean distance or cosine similarity.

Fig. 4 is the histogram of average value of feature F from 100 thousand images. Region L_1 denotes the first 4 values of F, and region L_2 denotes the following 32 values of F, and region L_3 denotes the rest 256 values of F. L_1, L_2, L_3 are corresponding to L_1, L_2, L_3 in Fig. 3. From Fig. 4, we can see the step-by-step descending trend of F. That is why we call F the global-to-local feature. Values from L_1 to L_3 denote the information which is from global to local respectively, so values in L_1 will occupy a large proportion in distance computing of equation(5). If objects in two images are very different in contour, the difference of values in L_1, L_2, L_3 must be all large, and as a result the similarity score in equation(5) is very low. But if two objects are only a little different, in other word, they should have almost the same global information and are just different in local parts, then only some values' difference in L_3 (maybe still in L_2) is large, but in L_1 must be small, and finally the similarity score in equation(5) is high.

3 Index Structure

Our feature contains an object's global-to-local information, so we select only the first 36 values of F which belong to region $L1$ and $L2$ as shown in Fig. 4 and include most of important information of the object. And for each value, we separate it into some parts, and for each part, there is a corresponding inverted list of images, as shown in Fig. 5. With the index structure, we could select top N1 ($\leqslant T$) candidate results from the database quickly, and then, we select top N2 results from N1 candidate results with similarity measure of first 36 values of F. Finally, we rank the N2 results with similarity measure of all values of F and take them as the finally retrieval results. Thus we could build a hierarchical top-down retrieval structure. In our experiment, $T = 50000$, $N2 = 2000$.

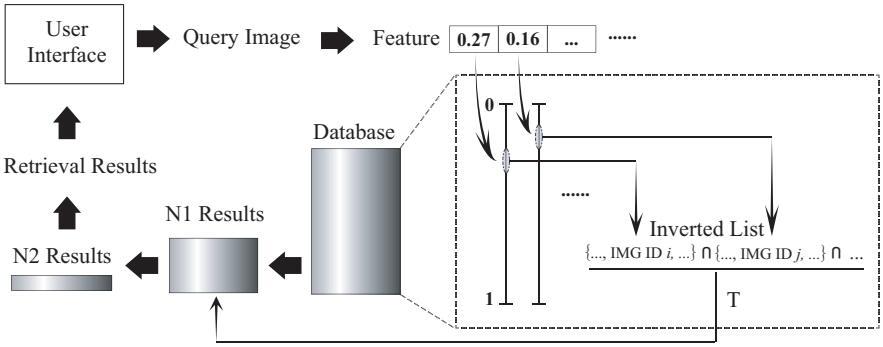


Fig. 5. Index process of database. For each query, N1 results are first selected from 4 million images by inverted files, and then top N2 results are selected from N1 results with similarity measure of first 36 values of F, and final results are from N2 results with similarity measure of all 292 values of F.

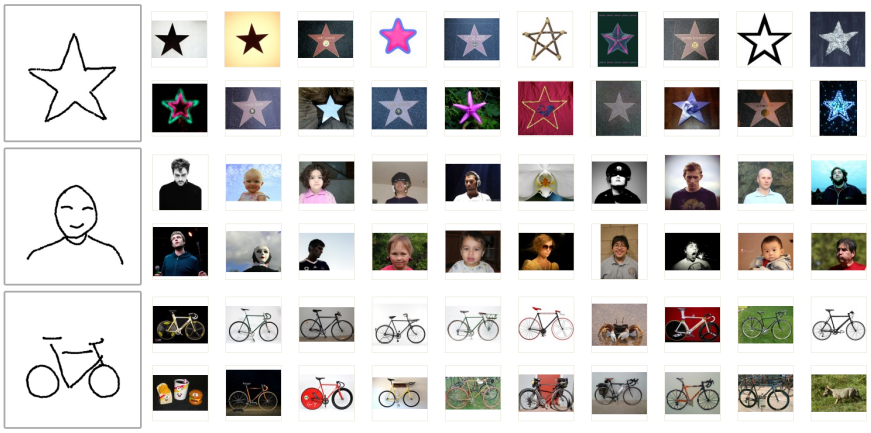
4 Experiment

To evaluate our retrieval method, we built a prototype system which database has more than 4 million Flickr images and run it on the server with 2 Intel Xeon 2.4GHz Quad Core processors and 8GB memory. Because the feature of an image has only 292 dimensions, and it takes less than 2KB memory per image, and the memory cost of our system including features of the database and the inverted file is not more than 7GB in total. So a normal server is powerful enough for our system. The average retrieval time is about between 2 and 3 seconds.

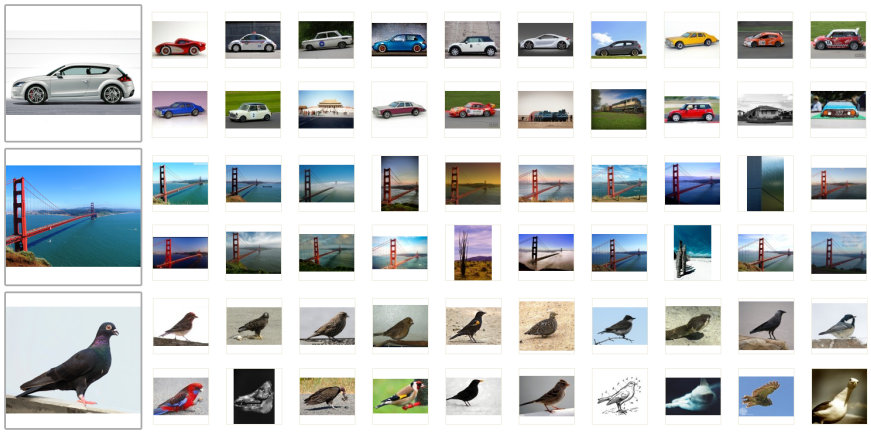
Fig. 6. Some examples of shift-invariance. Top row: the original image. Middle row: corresponding contour saliency map S and the centroid of the object (S_x, S_y) (red point). Bottom row: Canny edge detection.

To better explain why our method could deal with the shift-invariance problem, we display some examples and the corresponding contour saliency maps in Fig. 6. Our method extracts contour information of the object clearly, and further, obtains the centroid of the object. It is hard to be achieved by existing edge detection methods, e.g. Canny edge detector [8], as shown in the bottom

row of Fig. 6. And for most of saliency detection methods [9][10], it is still hard to be achieved. Because existing methods are almost based on information maximization [11], and if there are lots of contours having the same orientation in the image, these contours would be not salient in the saliency map.



(a) A hand-drawn sketch as a query image.



(b) A natural image as a query image.

Fig. 7. Some example queries and their top 20 retrieval results from the database of 4 million images.

Because no existing algorithm or image database is available for us to compare the performance, we just display the retrieval results from hand-drawn sketch and natural image as the query respectively, as shown in Fig. 7. From the results, we can see our method is shift-invariance. And for similar objects with different scales, their proportion of global feature at four orientations would

be almost same, thus their similarity score in equation(5) will be high. So our method deals with the scale-invariance problem to a certain extent.

5 Conclusion

We propose a simple and efficient contour-based method for large scale image retrieval. With hierarchical top-down index structure, our method can search the results from 4 million images quickly. Furthermore, it can use not only a hand-drawn sketch but also a natural image as the query image, which brings better interactive query method and the convenience for the users who don't do well in drawing. And our retrieval method is shift-invariance and scale-invariance to a certain extent, which could not be performed by any existing system having been published.

Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 90920014) and the NSFC-JSPS International Cooperation Program (Grant No. 61111140019).

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys 40(2), 1–60, (2008)
2. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Context. IEEE Trans. PAMI 24(4), 509-522, (2002)
3. Tieu, K., Viola, P.: Boosting Image Retrieval. IJCV 56(1/2), 17–36, (2004)
4. Shechtman, E., Irani, M.: Matching Local Self-Similarities across Images and Videos. CVPR 1–8, (2007)
5. Eitz, M., Hildebrand , K., Boubekeur , T., Alexa , M.: An Evaluation of Descriptors for Large-Scale Image Retrieval from Sketched Feature Lines. Computers & Graphics 34, 482–498, (2010)
6. Cao, Y., Wang, C.H., Zhang, L.Q., Zhang, L.: Edgel Index for Large-Scale Sketch-based Image Search, CVPR accepted, (2011)
7. Torralba, A., Fergus, R., Freeman, W.T.: 80 Million Tiny Images: A Large Dataset for Non-Parametric Object and Scene Recognition. IEEE Trans. PAMI 30(11), 1958–1970, (2008)
8. Canny, J.: A computational Approach to Edge Detection. IEEE Trans. PAMI 8(6), 679–698, (1986)
9. Gao, D., Mahadevan, V., Vasconcelos, N.: On the Plausibility of the Discriminant Center-Surround Hypothesis for Visual Saliency. Journal of Vision 8(7):13, 1–18, (2008)
10. Seo, H.J., Milanfar, P.: Static and Space-Time Visual Saliency Detection by Self-Resemblance. Journal of Vision 9(12):15, 1–27, (2009)
11. Bruce, N.D.B., Tsotsos, J.K.: Saliency Based on Information Maximization. NIPS 18, 155–162, (2006)