An Effective Framework for Contented-Based Image Retrieval with Multi-Instance Learning Techniques

Yu Peng* Kun-Juan Wei**, and Da-Li Zhang*

Abstract — **Multi-Instance Learning(MIL) performs well to deal with inherently ambiguity of images in multimedia retrieval. In this paper, an effective framework for Contented-Based Image Retrieval(CBIR) with MIL techniques is proposed, the effective mechanism is based on the image segmentation employing improved Mean Shift algorithm, and processes the segmentation results utilizing mathematical morphology, where the goal is to detect the semantic concepts contained in the query. Every sub-image detected is represented as a multiple features vector which is regarded as an instance. Each image is produced to a bag comprised of a flexible number of instances. And we apply a few number of MIL algorithms in this framework to perform the retrieval. Extensive experimental results illustrate the excellent performance in comparison with the existing methods of CBIR with MIL.**

Index Terms — **contented-based image retrieval, multi-instance learning, image segmentation, mathematical morphology, feature extraction**

I. INTRODUCTION

With the rapid development of electronic manufacture and internet technology, multimedia information is extremely exploded, multimedia data has been widely used in today's life, and image retrieval has been applied to extensive fields. However, most of the existing CBIR systems in which only global information is used or a user must explicitly indicate what part of the image is of interest. The query may actually contain multiple, possibly heterogeneous objects, ambiguous and difficult to be perceived. If each image can be treated as a sub-images bag which represents semantic concepts of the original, and each sub-image is described with a features vector, then the target concept can be learned through Multi-Instance Learning algorithm, the problem of the ambiguity can be gracefully resolved.

Multi-Instance Learning is firstly proposed by Dietterich et al.[1] to predict the activity of drug molecules. MIL is a variation of supervised learning for problems with incomplete knowledge about labels of training examples. Comparing to the supervised learning model, MIL provides a new way of modelling the teacher's weakness. Instead of receiving a set of bags that are labelled positive or negative. Each bag contains many instances. A bag is labelled positive if at least one instance in that bag is positive, and the bag is labelled negative is all the instances in it are negative. There are no labels on the individual instances. Learning from a small collection of positive and negative examples, we can get the concept point and use it to retrieve images that contain the concept from a large database. Many MIL algorithms have been proposed, such as learning axis-parallel rectangle(APR) concept [1], diverse density[2] is to find a concept point in n-dimension feature space, the optimal concept point is defined as the one with the maximum diversity densithy, which is a measure of how many different positive bags have instances near point, it had been applied in image retrieval[3][4]. Zhang and Glodman[5] combined Expectation Maximization with DD optimization function and the computatimal time, and can avoid local maxima since it makes major changes on the hypothesis when it switches from one instance to another in a bag, and also, a kNN based Citation-kNN algorithm was proposed to solve MIL problem in [6].

The CBIR system with MIL techniques must be based on precise segmentation, accurate object detection and effective features extraction, which play a key role in developing a practical retrieval system, the nucleus of the techniques is how to transforms an image into an meaningful bag(a set of instances), we call this process bag-generator, the excellent bag generator can output a variable number of instance to represent the semantic concepts accurately and tactfully, which can perform the algorithm easily and straightforwardly, and improve the precision of the retrieval results.

In this paper, we present a superior bag generator named SuperBag in our framework, experiments show that our work achieves a satisfying results in image retrieval system in comparison with the other existing methods. The paper is organized as follows. Section 2 presents SuperBag's techniques details. In section 3 the experimental results will be demonstrated. In the end, a conclusion is given in section 4.

II. SUPERBAG TECHNIQUES

2.1 Mean Shift Analysis with Improved Fast Gauss Transform

Mean shift, proposed by Fukunaga and Hostetler[7], a non-parametric estimator of density gradient, which is associated iterative procedure of mode seeking. Given n data points $x_1, \cdots, x_n$ in the d-dimensional space R^d, the multivariate kernel density estimator with kernel function $K(x)$ and window bandwidth h, is given:

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (1)$$

Where the d-variable kernel $K(x)$ is nonnegative and integrates to one. The Gaussian kernel is a common choice. The mean shift algorithm is a steepest ascent procedure which requires estimation of the density gradient:

Manuscript received May 24, 2007 ; revised Aug. 8, 2007.

* Department of Automation, Tsinghua University, Beijing, 100084, People Republic of China

** The 13th Institute, 10th Academy, China Aerospace Science and Technology Corporation (CASA), Beijing, 100854, People Republic of China
E-mail : pengy05@mails.tsinghua.edu.cn

$$\nabla \hat{f}_{h,K}(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x) g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)$$

$$= \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} x_i g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)} - x \right] \quad (2)$$

Where $g(x) = -k_N'(x) = (1/2)k_N(x)$, which can in turn be used as profile to define a Gaussian kernel $G(x)$, $C_{k,d}$ is the normalization coefficient. The first term is proportional to the density estimate at x computed with kernel G, the second term is the mean shift.

$$m(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)} - x \quad (3)$$

Which is proportional to the normalized density gradient and always points toward the steepest ascent direction of the function[8]. The mean shift algorithm iteratively performs the following two steps till it reaches the stationary point:

- Computation of the mean shift vector $m(x^k)$;
- Updating the current position $x^{k+1} = x^k + m(x^k)$.

In order to efficiently estimate sums of Gaussians, Yang and Duraiswami[8] develop an improved fast Gauss transform(IFGT) applied to the mean shift algorithm, which can dramatically reduce the computational complexity and storage cost, and speed up the kernel density estimation, as illustration in [8], which is the keystone for application in segmentation for image database retrieval procedure. We first transform the image to L*u*v color space, employ the mean shift with IFGT algorithm in the joint spatial-range domain with h=0.1 to all the points, $k = 5$ is the number of clusters, the convergence points are grouped by a simple k-means algorithm, there is none post-processing procedure as in[9]. Figure 1 shows the surprising results performed by mean shift analysis with IFGT.

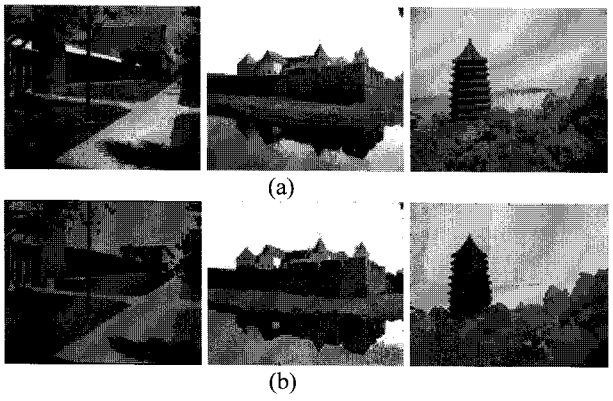


Fig. 1. Segmentation using Mean Shift Analysis with IFGT. (a) Original images. (b) Segmentation results (h=0.1 k=5).

2.2 Object Detection Using Mathematical Morphology

After segmentation, the image has been divided into several patches, every pixel $x_i (i = 1, \cdots, n)$ of the image is labelled by $L = (1, 2, \cdots, k)$, if the pixels or the patches belong to one class, they share the same class label, but they are not always connected.

- Denote the ith pixel by its coordinates (x_i, y_i), all pixels' coordinates of the same class are composed to a $2 \times N$ position matrix, where N is pixels number.
- Transfer every class to a binary image respectively according to the class's position matrix, the binary image keeping its contour and distribution.
- Employ Hit-or-miss transformation[11] of mathematical morphology to eliminate isolated pixels, and apply the opening operator: AoB = (AΘB) ⊕B, where B is a 4×4 unit square structuring element, the results reserve the large blocks and smooth the objects' contours.
- Eliminate the blocks which areas are less than 5% of the whole area, and map the large blocks to new position matrices, then crop the sub-images from the original image according to the new position matrices.

As shown in Figure 2(c), five distinct concepts have been detected, we regard the blobs detected as the ob- jects coarsely, and yet, it can satisfy the requirement of CBIR. In addition, the number of the detected objects (sub-images) can be varied adaptively and flexibly by the above methods. So we can describe the whole image with these semantic concepts.

2.3 Feature Extraction

For each object above, we extract a vector to represent it as an instance which is composed of color information, texture characteristics and statistical invariable moments. Calculate mean and standard deviation of R, G and B. We use Gabor wavelet transform[10] to extract the texture feature, which is based on the Gabor function, $f(x, y)$ is the impulse response of the mother wavelet :

$$f(x, y) = \left(\frac{1}{2\pi\delta_x\delta_y}\right) \exp\left[-\frac{1}{2}\left(\frac{x^2}{\delta_x^2} + \frac{y^2}{\delta_y^2}\right)\right] \cos(2\pi\mu_0 x) \quad (4)$$

Then obtain the self-similar filter dictionary by appropriate dilations and rotations of $f(x, y)$:

$$f_{m,n}'(x, y) = k^{-m} f(k^{-m}x', k^{-m}y') \quad (5)$$

Where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, $\theta = n\pi/N$, $m = 0, 1, \cdots, M-1$, $n = 0, 1, \cdots, N-1$, we use four scales M=4 and six orientations K=6, these 24 filters can be considered as the statistics of these micro-features in a given region are often used to characterize the underlying texture information [10].

But, as Figure 2(c) shown, the detected object's shape is irregular, we need to convert them into gray available subimages,

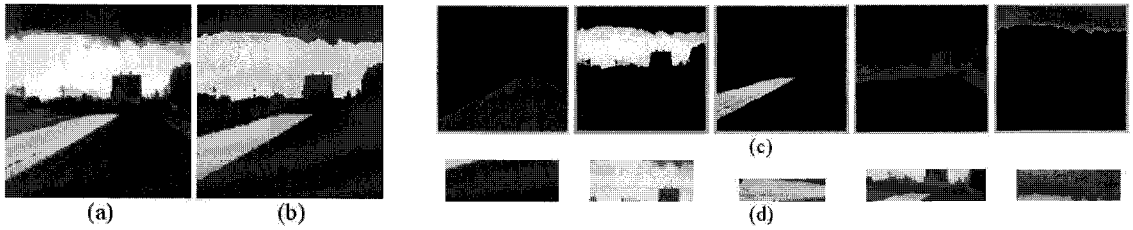


Fig. 2. Object detection and pre-treatment of feature extraction. (a) Original image. (b) Segmentation result. (c) Objects Detection by our method. (d) Gray sub-images representing the crucial parts of the first row for feature extraction.

so as to extract texture features by above method, do the following steps:

- Convert every object in Figure 2(c) into gray scale, calculate the object's mass centre and the length of major and minor axis of the ellipse which has second-order moment just the same as the object;
- For each object, obtain the angle θ between the object's major and horizontal axis, and rotate the gray image θ degrees anti-clockwise.
- Construct a rectangle frame with the mass centre, the expected sub-image's width and height are 0.8 times of the major and minor axis, crop the sub-image from the rotated image above, the final refine sub-images are shown in Figure 2(d).

In addition, the seven two-dimensional invariable moments[11] can be calculated from the gray sub-images in Figure 2(d), which are insensitive to size, movements zoom, rotation and so on.

Six color corresponding elements, twenty-four texture components, and seven invariable moments Values form the instance vector, which are normalized and represent the objects concept. Eventually, the image is converted into an image bag consisting of a variable number of 37-dimensional feature vectors(instances) collection.

III. EXPERIMENTS

Maron and Ratan[3] smoothed the image using a Gaussian filter and subsampled the image, they put forward single blob with neighbors(SBN). An SBN is defined as the combination of a single blob with its four neighboring blocks(up, down, left, right). The sub-image is described as a 15-dimensional vector, where the first three elements represent the mean R, G, B values of the central blob and the remaining twelve elements are the differences of mean color values between the central blob and other four neighboring blobs respectively. Therefore, each image bag is represented by a collection of nine 15-dimensional feature vectors.

Yang and Lozano [4] transformed color images into gray-scale images at first. Then, they divided each image into many overlapping regions. For each region, the sub-image is filtered and converted into an $h \times h$ matrix and treated as an h^2 dimensional feature vector. Each image bag generated is formed by a set of forty 64-dimensional feature vectors obtained by dividing each image into forty overlapping regions and setting h to be 8.

Zhou and zhang [12] proposed a bag generator based on a flexible segmentation with SOM neural network, they use color and space properties of the pixels to cluster the image to 4 classes. Eliminated the isolated pixels using a gliding window, and they merged the scattered and small blocks into its similar neighbour repeatedly.

Finally, the input image is converted into a corresponding image bag consisting of 3-dimensional feature vectors(instances) which formed by the mean R, G, B of each ultimate block.

Maron and Yang utilized fixed segmentation to all images, the sub-images can't represent the accurate semantic concepts, and be sue to induce much noisy. Although Zhou[12] applied the SOM based image segmentation, the result was not satisfied to detect the distinct objects of image correctly, Figure 3(d) shows its performance, which detected four possible objects, i.e., one ear, one eye, main body, head and grass field, which obtained the defective concepts and blended the different concepts together by mistakes. Our method processes the image effectively, as shown in Figure 3(e), it is obvious that ours is more competitive to Zhou's ImaBag.

Building a medium image database consisting of 500 images derived from COREL library, which includes 4 types: waterfall, mountain, flower, tiger or lion, each type contains 125 images. We create a potential training set which consisted of 25 randomly chosen images from each of the four types mentioned above. For a given concept, we picks several positive images with target concepts and several negative images from the potential database to build a training set. Through training and learning by DD[2], sort the images by similarity degree. Specifically, the most egregious false positives and the most egregious false negatives would likely be picked and added to the training set as a feedback to the system. In Figure 4, we show a snapshot of the framework in action, 6 images contain waterfall which is the target concept and 6 negetive images(6p6n) are regarded as training set, every trial was repeated 10 times.

In the following experiments, we combine diverse density(DD), EM-DD, APR, Citation-kNN algorithms into the framework, and demonstrate their performance on the image-base, as shown in table 1, EM-DD outperforms other algorithms, because EM-DD turns a multi-instance problem into a single-instance one, and help avoid local maxima since it makes major changes on the hypothesis when it switches from one instance to another in a bag.

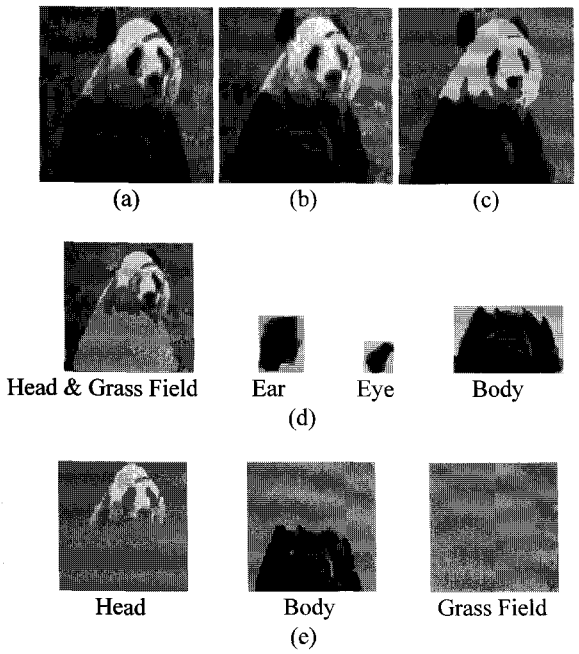


Fig. 3. Compare the processing results of SuperBag with ones by ImaBag. (a) Original. (b) Segmentation results by ImaBag. (c) Segmentation results by SuperBag. (d) Objects detected by ImaBag method(Head and Grass Field, Ear, Eye, Body). (e) Objects detected by SuperBag (Head, Body, Grass Field).

In the following experiments, we combine diverse density(DD), EM-DD, APR, Citation-kNN algorithms into the framework, and demonstrate their performance on the image base, as shown in table 1, EM-DD outperforms other algorithms, because EM-DD turns a multi-instance problem into a single-instance one, and help avoid local maxima since it makes major changes on the hypothesis when it switches from one instance to another in a bag.

We compare our framework to the existing methods based on MIL have reported. Each concept of the image-base is performed 10 times using different methods, i.e. SuperBag, ImaBag $(n = 4)$, Maron and Ratan's SBN, Yang and Lozano's method $(h = 8)$. Precision and recall can evaluate the performance of the image retrieval.

Precision is the ration of the number of correctly retrieved images to the number of all images retrieved so far.

Recall is the ratio of the number of correctly retrieved images to the total number of correct images in the test set[12]. According to the table 2, we can conclude that the retrieval results by SuperBag method achieves better performance than other methods.

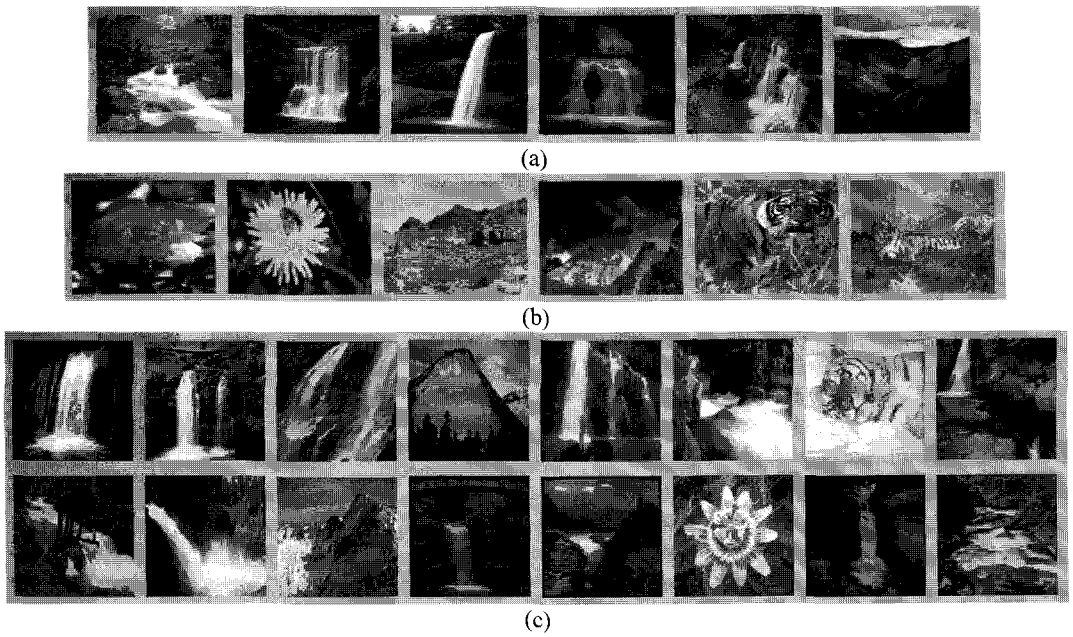


Fig. 4. Results of the waterfall concept using the method in the paper. (a) User-Selected positive examples(6p). (b) User-Selected negative examples(6n). (c) Final retrieval from test set (top 16 images).

Table 1. Compare the performance of different MIL algorithms in our framework using same sets (6p6n)

MIL Algorithm	DD	EM-DD (it=12, TH=10^{-5})	Citation-kNN (k=5;C=7)	Iterated-discrim APR
Flowers	0.540	0.630	0.610	0.520
Mountains	0.517	0.614	0.580	0.542
Waterfalls	0.558	0.622	0.571	0.540
Lions	0.486	0.578	0.562	0.533
Average	0.526	0.608	0.580	0.549

Table 2. Results of training scheme 10p10n, comparison of four methods using same sets

Image Type	SuperBag		ImaBag		Maron & Ratan (SBN)		Yang & Lozano	
	precision	recall	precision	recall	precision	recall	precision	recall
Flowers	0.712	0.793	0.689	0.746	0.707	0.781	0.655	0.676
Mountains	0.655	0.711	0.601	0.701	0.620	0.713	0.589	0.601
Waterfalls	0.758	0.806	0.731	0.761	0.729	0.813	0.670	0.737
Tigers&Lions	0.691	0.716	0.656	0.670	0.678	0.719	0.558	0.628
Building	0.726	0.736	0.659	0.710	0.690	0.699	0.587	0.650

IV. CONCLUSION

Much work has been developed in CBIR based on multi-instance learning, which dealed with the image ambiguous and improved the retrieval performance. The CBIR system combine the techniques, such as image segmentation, object detection, feature extraction, and machine learning algorithm. In our framework we employ improved Mean Shift algorithm to segment the image, detect the possible objects using mathematical morphology, and then extract crucial features from the image. Which converted the image into several integrated and effective instances. Every instance was a feature vector extracted from the possible object in the image. So the training and learning works will be easier and straightforward. The experiments have demonstrated its strength and excellence. Among the MIL algorithms we have employed in above experiments, EM-DD algorithm have the best performance in practical CBIR system.

There are some deficiencies in our work, the number of the clustering and the bandwidth of Mean Shift algorithm are predefined, another problem is how to accurately detect the highly complex objects in the image, and describe the objects effectively by feature extraction in the future.

REFERENCES

[1] Dietterich, R.H. Lathrop, and T.Lozano-Pérez, 1997. "Solving the multiple-instance problem with axis-parallel rectangles". *Artificial Intelligence, vol.89.*

[2] O.Maron, 1998. "Learning from Ambiguity". *Doctoral Thesis, Dept. of Electrical Engeineering and Computer Science, M.I.T.*

[3] O.Maron and A.L. Ratan, 1998. "Multiple-instance learning for natural scene classification". *The 15th International Conference on Machine Learning, Madison.*

[4] C.Yang and T. Lozano-Pérez, 2000. "Image database retrieval with multiple-instance learning techniques". *Proceedings of the 16th International Conference on Data Engineering, San Diego, CA.*

[5] Q. Zhang and S. A. Goldman, 2001. "EM-DD: An improved multiple-instance learning technique". *Neural Information Processing Systems.*

[6] Wang J, Zucker J-D. "Solving the multiple-instance problem: a lazy learning approach". In: *Proceedings of the17th International Conference on Machine Learning*, San Francisco, CA, 2000, 1119-1125.

[7] Fukunaga K. Hostetler L D, 1975. "The Estimation of the Gradient of a Density Function". *With Applications in Pattern Recognition. IEEE Trans on Information Theory.*

[8] C. Yang, R. Duraiswami, N. Gumerov, and L. Davis, 2003. "Improved fast Gauss transform and efficient kernel density estimation". *In Proc. ICCV 2003*

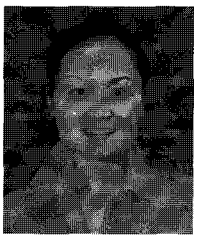
[9] D. Comaniciu and P. Meer, 2002. "Mean shift: A robust approachtoward feature space analysis". *IEEE Trans. Pattern Anal.Mach. Intell.*

[10] Manjunath B. S, Ma W. Y, 1996. "Texture Features for Browsing and Retrieval of Image Data". *IEEE Trans on Pattern Analysis and Machine Intelligence.*

[11] Rafael C. Gonzalez Rechard E.Woods, 2004. "Digital Image Processing". *Pearson Education, Inc., Prentice Hall.*

[12] Zhi-Hua Zhou, Min-Ling Zhang, Ke-Jia Chen, 2003. "A Novel Bag Generator for Image Database Retrieval With Multi-Instance Learning Techniques". *In 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03).*

Yu Peng received the B. S. degree in electrical engineering from China JiLiang University of HangZhou in 1999. As a visiting scholar at Northwestern Polytechnical Univer-sity of China from 2001 to 2002. And as a candidate for M. S. degree at Department of Automation of Tsinghua University from 2005.



Kun-Juan Wei received the B. S. degree in automation from North China Institute of Aerospace Engine-ering in 1999. As a engineer in the 13th Institute, 10th Academy, China Aero-space Science and Technology Corporation (CASA), Beijing from 2000 till now.

Da-Li Zhang received the B.S degree in department of electrical engineering in 1970. And received Ph.D. degree in Stuttgart University, Germany, in 1992.He is currently a Professor in department of automa-tion of Tsinghua University, and as a standing director of China Society of image and Graphics(CSIG).