Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Investigating Visual Feature Extraction Methods for Image Annotation

Rukun Hu[1], Shuai Shao[1]
[1]Image Processing and Pattern Recognition Laboratory
Beijing Normal University
Beijing 100875, China
Email: hurukun@126.com

Ping Guo[1,2]
[2]The School of Computer Science and Technology
Beijing Institute of Technology
Beijing 100081, China
Email: pguo@ieee.org

***Abstract*—In order to investigate the performance of visual feature extraction method for automatic image annotation, three visual feature extraction methods, namely discrete cosine transform, Gabor transform and discrete wavelet transform, are studied in this paper. These three methods are used to extract low-level visual feature vectors from images in a given database separately, then these feature vectors are mapped to high-level semantic words to annotate images with labels in a given semantic label set. As it is more efficient to depict the visual features of an image by the feature distribution than to resort to image segmentation technology for semantic image blocks, this paper is going to find out which of the three feature extraction methods performs better in image annotation based on the distribution of feature vectors from the image. The performance of three different kinds of feature extraction method is fully analyzed, and it is found that discrete cosine transform method is more suitable for Gaussian mixture model in automatic image annotation.**

***Index Terms*—Automatic image annotation, feature distribution, expectation maximization algorithm, Bayesian decision.**

I. Introduction

Content-Based image annotation, the problem of marking images with semantic labels according to their content, has been the subject of a significant amount of research in the last decade [1]. With the aid of the annotation, the accurate query for image will be easier, which means that the user can specify the query through a natural language description [1]. But as the increasing number of images is so huge that the manual image labeling becomes impossible, the research for automatically extracting semantic descriptors from images become necessary, and has been attracting many researchers' attention.

Basically, automatic image annotation procedure can be described as extracting low-level visual feature vectors from images and mapping them to the high-level semantic words. So it can be posed as a classification problem where each class is defined as the group of database images labeled with a common semantic label [1]. To solve a classification problem , both feature extraction method and classifier need to be concerned. According to the well-known statistical decision theory, the Gaussian mixture model (GMM) can well approximate any kinds of distribution of feature vectors and the Bayesian decision rule

$$P_{W|X}(i|x) = \frac{P_{X|W}(x|i)P_W(i)}{\sum_i P_{X|W}(x|i)P_W(i)} \tag{1}$$

can help to minimize the probability of error annotation. Then one of the key problems in image annotation turns to choose a proper visual feature extraction method based on the GMM distribution.

As to the feature extraction method, much work have been done and many methods have been proposed. Generally, visual features in an image may include color, texture, shape, color layout, etc. The color feature is relatively robust to background complication and independent of image size and orientation. A lot of work has been done with different color feature representations, like color histogram [4], color moments [8], [9], as well as color sets [10]. Meanwhile, there are also some work with texture [11] and shape [12] features. Recently, a hybrid approach, which incorporates color, shape and spatial relations among objects in a image is proposed in [13].

As we known, the color-based features are not suitable to GMM method in image annotation, because the objects with the same semantic labels may appear in totally different colors. Three most famous transform methods, namely discrete cosine transform (DCT), Gabor transform and discrete wavelet transform (DWT), have been adopted successfully to solve the image annotation problem in some other models separately in [1], [14]. Inspired to investigate the performance of the visual feature extraction method for image annotation based on the distribution of feature vectors, these three famous transform methods are investigated to extract feature vectors which can be taken as a mixture of shape and texture features of the image, and their performance is analyzed in this paper.

The paper is organized as follows: Section II defines the image annotation problems and Section III has a brief review of feature extraction methods used in this field. The improved expectation-maximization (EM) algorithm [1], [18] used to estimate the distribution of semantic class is introduced in Section IV and Section V shows the experiment and the analysis of the results.

II. Image Annotation

Following the discussion in [1], consider an image database $\mathcal{T} = \{I_1, I_2, ..., I_N\}$ and a semantic label vocabulary $\mathcal{L} = \{w_1, w_2, ..., w_N\}$. The goal of image annotation is to annotate a given image I_i with semantic labels W_i from vocabulary $\mathcal{L}$ based on the information of training set

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

$\mathcal{D} = \{(I_1, W_1), ..., (I_D, W_D)\}$ from $\mathcal{T}$ where W_i is a subset of $\mathcal{L}$ [1].

When the image annotation is posed as a classification problem, it becomes a mapping problem from low-level visual feature to high-level semantic label. And the performance of the image annotation can be affected by 1) the information gotten from the train set and 2) the method used to associate the visual information with the semantic labels. In most cases, the training set is weakly labeled which means 1) the absence of a semantic label from caption w_i does not necessarily mean that the associated concept is not present in I_i, and 2) it is not known which image region is associated with a specific label. Weak labeling is expected in practical annotation, since 1) each image is likely to be annotated with a small caption that only identifies the semantics deemed as most relevant to the labeler, and 2) users are rarely willing to manually annotate image regions [1]. To deal with the weak labeling problem, a supervised multiclass labeling model (SML) was proposed in [1]. Based on this model, we will investigate the effectiveness of low-level visual feature extraction methods in image annotation.

III. Feature Extraction

The feature extraction method play an important role in image annotation. According to image processing technology, Gabor transform, DCT and DWT are all well-known frequency analysis methods.

A. Discrete Cosine Transform

As a kind of separable and orthogonal transform, DCT is widely used in image analysis.

The two-dimensional DCT is defined as in (2) [5]:

$$\begin{aligned} C(u,v) &= a(u)a(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} p(x,y) \\ &\cos[\frac{(2x+1)u\pi}{2N}]\cos[\frac{(2y+1)v\pi}{2N}], \end{aligned} \tag{2}$$

$$a(i) = \begin{cases} \sqrt{1/N} & \text{if i=0,} \\ \sqrt{(2/N)} & \text{if i=1,2,...,N-1.} \end{cases} \tag{3}$$

where $p(x,y)$ is the (x, y^{th}) element of the image represented by the matrix p. N is the size of the block that the DCT is done on. The equation calculates one entry (u, v^{th}) of the transformed image from the pixel values of the original image matrix. DCT works by separating images into parts of differing frequencies. It can reduces the interruption between consecutive blocks in the image and is famous for its well-known energy compaction properties.

B. Gabor Transform

Gabor filters can serve as excellent band-pass filters for unidimensional signals [6], and here is the formula of a complex Gabor function in space domain,

$$g(x,y) = s(x,y)w_r(x,y), \tag{4}$$

where $s(x,y)$ is a complex sinusoid, known as the carrier, and $w_r(x,y)$ is a 2-D Gaussian-shaped function, known as the envelope. To specify the $s(x,y)$ and $w_r(x,y)$, we have the complex Gabor function in polar coordinates [6]

$$\begin{aligned} g(x,y) &= Kexp\{-\pi(a^2(x-x_0)_r^2 + b^2(y-y_0)_r^2)\} \\ &exp\{j(2\pi F_0(x\cos w_0 + y\sin w_0) + P)\}. \end{aligned} \tag{5}$$

where K scales the magnitude of the Gaussian envelope, (a,b) scale the two axis of the Gaussian envelope, (x_0, y_0) is the location of the peak of the Gaussian envelope, (F, w_0) is the polar coordinates form of spatial frequencies of the sinusoid carrier, and P is the phase of the sinusoid carrier.

In the experiment, we take the form of Gaussian-shape function as follows,

$$\begin{aligned} W_r(x,y) &= Kexp\{-0.5((\frac{x\cos w_0 + y\sin w_0}{a})_r^2 + \\ &(\frac{y\cos w_0 - x\sin w_0}{b})_r^2)\}. \end{aligned} \tag{6}$$

where $K = 1,\ a = 3,\ b = 4,\ w_0 = \pi/3,\ F_0 = 16\ and\ P = 0$. To apply this filter to the image by convolution, The Gabor feature vectors of the image can be achieved.

C. Discrete Wavelet Transform

Wavelet transform is a useful signal analysis method and is widely used since it was shown to be the foundation of a powerful new approach to signal processing and analysis called multiresolution theory [5]. In image processing field, the two-dimensional discrete wavelet transform play an important role in image analysis.

$$W_u(0,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=1}^{N-1} f(x,y)u_{0,m,n}(x,y), \tag{7}$$

$$W_v^{(i)}(j,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=1}^{N-1} f(x,y)v_{j,m,n}^{i}(x,y), \tag{8}$$

and

$$u_{j,m,n}(x,y) = 2^{j/2}u(2^j x - m, 2^j y - n), \tag{9}$$

$$\begin{aligned} v_{j,m,n}^{(i)}(x,y) &= 2^{j/2}v^{(i)}(2^j x - m, 2^j y - n), \\ (i) &= \{H,V,D\}, j = 1,...,J-1, \\ &m,n = 0,1,...,2^j - 1. \end{aligned} \tag{10}$$

where $u(x,y)$ is a scaling function, $v^H(x,y)$, $v^V(x,y)$, $v^D(x,y)$ are directionally sensitive wavelet functions and the size of the original image is $M \times N$. In the experiment, the Cohen-Daubechies-Feauveau 9/7 (CDF-97) wavelet transform is used to extract feature vectors.

IV. Estimation of Semantic Class Distributions

One efficient way to the estimation of class mixture density is to adopt a hierarchical density estimation method first proposed in [15] for image annotation. This method is based on a mixture hierarchy where children densities consist of different combinations of subsets of the parents' components. In the image annotation approach, image densities are children and semantic class densities are their parents [1]. As shown in [15], it is possible to estimate the parameters of class mixtures directly from those image mixtures, using a two-stage procedure. The first stage is the estimation of image densities. Assuming that a semantic class training set has D_i images and each image mixture has K components, this leads to a class mixture of D_iK components with parameters

$$\{\pi_j^k, \mu_j^k, \Sigma_j^k\}, j = 1, ..., D_i, k = 1, ..., K.$$

The second stage is to cluster all of the image-level Gaussian components into a class-level M-component mixture, where M is the number of components desired at the class-level. Denoting the parameters of the class mixtures by $\{\pi_c^m, \mu_c^m, \Sigma_c^m\}$, ($c$=1, ..., C, m=1, ..., M, where C is the number of class.), those parameters can be estimated by the hierarchical extension of EM algorithm with the following steps [1]:

1. Initialize the parameters as means $\{\mu_c^1, ..., \mu_c^M\}$, covariances $\{\Sigma_c^1, ..., \Sigma_c^M\}$ and mixing coefficients $\{\pi_c^1, ..., \pi_c^M\}$, and evaluate the initial value of the log likelihood.

$$logP(X|\pi, \mu, \Sigma) = \sum_{c=1}^{C} log\{\sum_{m=1}^{M} \pi_c^m G(X, \mu_c^m, \Sigma_c^m)\}. \tag{11}$$

2. **E-step.** Evaluate h_{jk}^m using the current parameter values

$$h_{jk}^m = \frac{[\mathcal{G}(\pi_c^m, \mu_c^m, \Sigma_c^m)e^{-\frac{1}{2}trace\{(\Sigma_c^m)^{-1}\Sigma_j^k\}}]^{\pi_j^k N}\pi_c^m}{\sum_l[\mathcal{G}(\pi_c^l, \mu_c^l, \Sigma_c^l)e^{-\frac{1}{2}trace\{(\Sigma_c^l)^{-1}\Sigma_j^k\}}]^{\pi_j^k N}\pi_c^l}. \tag{12}$$

   where N is a user-defined number (see [15] for details) that takes the value 1 in all experiments.

3. **M-step.** Re-estimate the parameter values using h_{ij}^m

$$(\pi_c^m)^{new} = \frac{\sum_{jk} h_{jk}^m}{D_iK}, \tag{13}$$

$$(\mu_c^m)^{new} = \sum_{jk} w_{jk}^m \mu_j^k, \;\; w_{jk}^m = \frac{h_{jk}^m}{\sum_{jk} h_{jk}^m \pi_j^k}, \tag{14}$$

$$(\Sigma_c^m)^{new} = \sum_{jk} w_{jk}^m [\Sigma_j^k + (\mu_j^k - \mu_c^m)(\mu_j^k - \mu_c^m)^T]. \tag{15}$$

4. Evaluate the log likelihood and check for the convergence of the log likelihood or the parameters. If the convergence criterion is not satisfied, return to step 2.

Note that the number of parameters in each image mixture is magnitude smaller than the number of feature vectors in the image itself. Hence, the complexity of estimating the class mixture parameters will be cut down. And as the variances on the left-hand side can never be smaller than those on the right-hand side in (15), the hierarchical class density estimates are much more reliable than those obtained with direct learning [15]. To speed up parameters estimation in GMM, we adopt an integrated method [18] for EM algorithm.

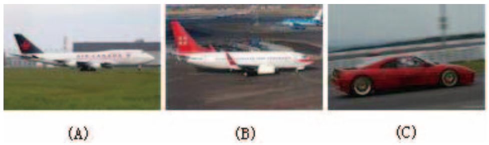

Fig. 1. Images from the training set. Image A and B are from the set of airplane and image C is from set of car.

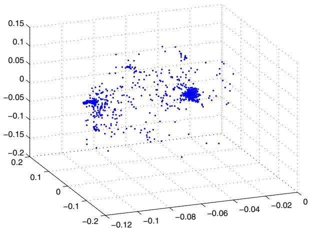

Fig. 2. Distribution of feature vectors extracted by DWT from image A above in the three-dimensional PCA space.

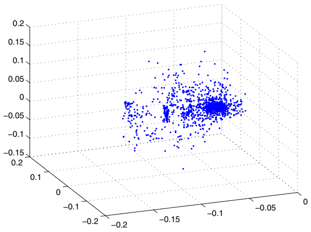

Fig. 3. Distribution of feature vectors extracted by DWT from image B above in the three-dimensional PCA space.

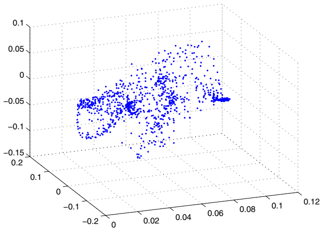

Fig. 4. Distribution of feature vectors extracted by DWT from image C above in the three-dimensional PCA space.

V. Experiment

A. Summary of Experiment

The evaluation of an image annotation system requires three components: an image database with manually produced annotations, a strategy to train and test the system, and a set of measures of annotation performance. In the experiment, the image database is consisted of about 2,000 images selected from the database used in Visual Object Classes Challenge 2008 (VOC2008), and there are 15 labels in the semantic label set, including "aeroplane", "bicycle", "car", "sky", etc. As the size of the semantic label set is small, not all of the objects in an image can be annotated. In the experiment, each image can be annotated with a caption of one to three semantic labels.

In this approach, all images are decompose into a set of overlapping 8×8 pixel regions, moves by four pixels between consecutive samples (note that all images were represented in the YC_bC_r color space). And at each region, a feature vector is achieved by the application of the three feature extraction methods mentioned above separately. Then the dimension of all feature vectors from each image is reduced to [64, 32, 16] by principle component analysis (PCA) method respectively. At last, these feature vectors in lower dimensional space are used to estimate the Gaussian mixture model parameters. For the GMM, the component number at the image-level is 8, and the number at the class-level is 16. Image annotation performance is evaluated by comparing the captions automatically generated for the test set. We define the automatic annotation as the three semantic classes of largest posterior probability, and compute the precision of every word in the test set. For a given semantic descriptor, assuming that the system annotates w_{auto}, of which w_c are correct and the precision is given by $precision = \frac{w_c}{w_{auto}}$ [1]. For each semantic classes, there are 100 images in the test set, which are all contain the corresponding objects for that semantic. All the performance of the three methods is recorded and the average precision for each method in different dimensional feature space is calculated.

B. Algorithm Description

In this section, the training and annotation algorithms used in this work is reviewed and the parameters of the training algorithm that affect the performance of the annotation tasks are identified.

In the training phrase, a training set $\mathcal{D}$=$\{(I_1, W_1)$, ..., $(I_N, W_N)\}$ of image-caption pairs is assumed, where $I_i \in \mathcal{T}$ and $W_i \subset \mathcal{L}$. In order to make our approach self-contained, we rewrite the steps of the training algorithm in the following, more details can refer to [1]:

For each semantic class $w_i \in \mathcal{L}$,

1. Build a training image set $\widetilde{\mathcal{T}} \subset \mathcal{T}$, where $W_i \subset \mathcal{L}$ for all $I_i \in \mathcal{T}$.
2. For each image $I_i \in \widetilde{\mathcal{T}}$,
   i. Decompose I_i into a set of overlapping 8×8 pixel regions, moves by four pixels between consecutive samples.
   ii. Compute a feature vector at each region of the three YC_bC_r color channels by the application of the three feature extraction methods mentioned in Section III respectively. Then the image can be represented by a set of 192×1 dimensional feature vectors as in [1]. Then all vectors are transformed into different feature spaces of lower dimension.
   iii. Assuming that the feature vectors extracted from the regions of image I are sampled independently, find the parameters of Gaussian mixture of 8 components that maximizes their likelihood using the EM algorithm (in all experiments, the Gaussian components had diagonal covariance matrices).
3. Fit a Gaussian mixture of 16 components by applying the hierarchical EM algorithm of (13)-(15) to the image-level mixtures achieved in step (2). This leads to a conditional distribution $P_{\mathrm{X}|W}(\mathrm{x}|w)$ for class w.

The parameters that may affect annotation performance are: 1) number of feature vector dimensions, and 2) number of mixture components for each class in step (3).

As to the annotation phase, the algorithm is exactly the same with that in [1].

C. Distribution of Feature Vectors

To illustrate the rationality of using Gaussian mixture model in image annotation, the dimension of feature vectors extracted by DWT method from some images in the database is reduced to 3 by the application of PCA method, as shown in Fig. 2, Fig. 3 and Fig. 4, and the original images are presented in Fig. 1.

As shown in Fig. 2, Fig. 3 and Fig. 4, the feature vectors of each image roughly concentrate to several centers in the three-dimensional feature space, which means the distribution of feature vectors in each image can be fitted by GMM. And by comparing the distance between centers in Fig. 2 and centers in Fig. 3 and the distance between centers in Fig. 2 and centers in Fig. 4, we can see that the distributions in different images from the same class have more similarity than those from different classes, which means it is reasonable to describe the distribution of class feature vectors with GMM. But as the distribution of each image is different from one another, the best number of component used to describe each image is quite different. Similarly to that used in [1], we use 8 components at the image-level.

D. Feature Space of Different Dimensions

The dimension of feature vectors always plays an important role in classification problem. High dimensional feature vector means more information of the original data, but a long classification time. Fortunately, more information of the original data in the feature vector does not necessarily means a obvious increase in precision. And in practice, both precision and time cost need to be considered.

In the experiment, the original feature vectors are transformed to three lower-dimensional vectors by PCA method separately. The annotation precision based on these lower-dimensional feature vectors is shown in table I, table II, and table III where the precision recorded in percentage. As the test image is annotated with the three classes w_i of largest posterior probability in the experiment, the labels “one”, “two” and “three” mean that the class label is right annotated to the test image if only the corresponding number of labels are annotated to the test image.

Analyzing the precision recorded in each table, we can see that the average precision when using DWT and Gabor features in the 32-dimensional and 16-dimensional space is a little higher than that in 64-dimensional space. And on all 15 classes, the performance of feature vectors in 32-dimensional space is more stable than that in 64-dimensional space. The

TABLE I
COMPARISON ANNOTATION PERFORMANCE USING FEATURE VECTORS EXTRACTED BY DWT WITH DIFFERENT DIMENSIONS.

Dimen		plane	bicycle	bird	bus	car	cat	chair	dog	horse	motorbike	person	sky	sofa	train	TV	average
$d = 16$	one	51	1	27	5	7	31	26	9	3	0	50	30	7	0	8	17.3
	two	71	4	33	14	11	63	44	19	9	1	68	55	15	0	24	28.7
	three	76	5	43	21	25	78	75	28	20	3	79	59	25	4	55	39.7
$d = 32$	one	31	6	10	7	18	18	4	27	34	11	12	23	17	10	43	18.1
	two	46	8	19	15	34	30	15	41	57	22	34	32	25	31	71	32
	three	49	12	30	21	45	35	18	53	71	32	62	36	28	48	91	42.1
$d = 64$	one	0	1	57	11	14	24	0	8	0	42	2	6	14	6	15	13.3
	two	4	2	79	24	25	55	1	19	1	70	0	39	35	29	34	27.8
	three	19	3	85	32	37	74	3	31	4	78	4	67	49	47	54	38.9

TABLE II
COMPARISON ANNOTATION PERFORMANCE USING FEATURE VECTORS EXTRACTED BY DCT WITH DIFFERENT DIMENSIONS.

Dimen		plane	bicycle	bird	bus	car	cat	chair	dog	horse	motorbike	person	sky	sofa	train	TV	average
$d = 16$	one	37	1	13	14	19	45	1	2	13	0	5	38	24	30	44	19
	two	66	4	22	26	29	63	13	8	29	1	11	51	45	33	62	30.8
	three	71	5	31	42	41	79	26	10	33	4	25	56	67	37	72	40
$d = 32$	one	34	7	24	18	14	24	3	12	18	7	42	37	31	8	14	19.5
	two	49	13	40	22	21	36	5	20	31	7	18	56	43	11	25	30.4
	three	59	17	48	28	31	56	8	29	43	9	16	58	61	19	50	41.1
$d = 64$	one	38	4	6	2	6	71	2	6	1	0	14	68	4	3	15	17.6
	two	88	11	12	9	13	83	6	14	6	3	79	86	10	17	31	32.4
	three	92	18	44	18	29	90	9	39	8	7	90	90	16	27	51	42.4

TABLE III
COMPARISON ANNOTATION PERFORMANCE USING FEATURE VECTORS EXTRACTED BY GABOR TRANSFORM WITH DIFFERENT DIMENSIONS.

Dimen		plane	bicycle	bird	bus	car	cat	chair	dog	horse	motorbike	person	sky	sofa	train	TV	average
$d = 16$	one	32	11	24	1	12	58	7	0	1	11	33	27	3	1	25	16.4
	two	46	24	31	8	31	81	9	3	0	26	47	44	24	7	63	29.6
	three	49	31	37	19	35	93	13	8	5	46	55	49	45	19	81	39
$d = 32$	one	33	24	17	22	4	3	36	0	19	15	58	24	1	44	2	20.1
	two	43	33	29	36	11	6	83	3	40	33	66	34	5	58	5	32.3
	three	46	39	31	58	30	8	93	8	52	45	71	38	14	76	13	41.4
$d = 64$	one	0	0	0	0	0	0	0	100	100	0	0	0	0	0	0	13.3
	two	66	0	0	0	0	55	0	100	100	0	0	26	6	0	0	23.5
	three	92	0	0	0	0	73	0	100	100	0	0	91	35	0	0	32.7

reduction of feature space dimension does not cause the decline in annotation precision. Even in some cases, the feature vectors with lower dimension can produce a higher precision. That is because the Gaussian mixture with 8 components at image-level and 16 components at class-level can describe the distribution of the feature vectors in lower dimensional space better in these cases. This means once the number of components for the model can be correctly chosen, the image annotation problem can be handled in a low-dimensional feature space, which will obviously cut down the time cost.

E. Different Feature Extraction Methods

To put the data in the three tables together, we can see that in different classes, the performance of the three extraction methods is quite different. In some classes such as “aeroplane”, “cat”, and “sky”, all the three feature extraction methods can perform well, but in some other classes like “bicycle”, the performance is bad. And it seems that the performance of feature extraction methods have something to do with the complexity of the shape and texture of the target object.

When annotating some classes like “motorbike”, the DWT feature vectors can get better precision than DCT method in the 32-dimensional and 16-dimensional feature space, so does Gabor method. But the performance of DWT method and Gabor method is not as stable as that of DCT method. There are four classes on which the annotation precision of DCT method in 64-dimensional feature space is below 10 percent and nine for Gabor method, but only three for DCT method. When come to the 16-dimensional feature space, the number for DWT method and Gabor method is three, but only two for DCT method. On average, the performance of DCT method in different classes is more stable than the other two method.

When come to the average precision, we can see that the performance of DCT method is a little better than that of the other two method almost in all the three different dimensions of feature space. Considering both the precision and stability, we can conclude that among the three methods, the DCT method is more suitable for GMM in image annotation.

VI. CONCLUSION

In this work, the performance of three different visual feature extraction methods applied to image annotation based on the distribution of feature vectors is investigated. According to the experiment results, we can get conclusions as follows:

- For some classes, such as “car” “sofa” and “TV”, the reduction of feature space dimension does not cause the decline in annotation precision. This good quality will bring computational efficiency to the annotation process.

- The performance of the three feature extraction method is unstable over all classes. When choosing feature extraction method for image annotation, we have to consider the character of the main targets we focus on.
- Among the three feature extraction methods, there is no one that performs better than the other two methods in all classes. But on average, the image annotation precision aided DCT method is a little higher than the other two methods in all the three kinds of dimension of feature space we investigated.

The component number K and M in this experiment are pre-settled. And they seem not to be suitable for all classes according to the experiment results. Only when the number of components can right match the distribution character of the class, the image annotation can get a well performance. In the future, as an extension of this work, self-adoptive component chosen process should be considered.

Acknowledgment

The research work described in this paper was fully supported by the grants from the National Natural Science Foundation of China (Project No. 60675011, 90820010). Prof. Ping Guo is the author to whom the correspondence should be addressed, his e-mail address is pguo@ieee.org.

References

[1] G. Carneiro, A. B. Chan, P. J. Moreno and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, 2007.

[2] R. Picard, "Digital Libraries: Meeting Place for High-Level and Low-Level Vision", Proc. Asian Conf. Computer Vision, pp. 1-5, 1995.

[3] J. Fan, Y. Gao, H. Luo and G. Xu, "Statistical modeling and conceptualization of natural images", Pattern Recognition, vol. 38, no. 6, pp. 865-885, 2005.

[4] J. Park, Y. An, G. Kang, W. Rasheed, S. Park and G. Kwon, "Defining a new feature set for content-based image analysis using histogram refinement", Int. J. Imaging Systems Technology, vol. 18, no. 2-3, pp. 86-93, 2007.

[5] Yujin Zhang , "Image Engineering (vol. 1)", Beijing, Tsinghua University Press, 1999. (In chinese).

[6] J. R. Movellan, "Tutorial on Gabor Filters", Technical Report, 2002.

[7] J. Luo, M. Boutell, and C. Brown, "An overview of exploiting context for semantic scene content understanding", IEEE Signal Process, vol. 23, no. 2, pp. 101-114, 2006.

[8] M. Striker and M. Orengo, "Similarity of color image", Proc. SPIE, Storage and Retrieval for Image and Video Databases, vol. 2420, pp. 381-392, 1995.

[9] J. L. Shih and L. H. Chen, "Color image retrieval based on primitives of color moments", Proc. IEEE the Vision, Image, and Signal Processing, vol. 149, no. 6, pp. 370-376, 2002.

[10] J. R. Smith and S. F. Chang, "Single Color Extraction and Image Query", Proc. IEEE Int. Conf. Image Processing, vol 3, pp. 528-531, 1995.

[11] J. R. Smith and S. F. Chang, "Automated binary texture feature sets for image retrieval", Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 4, pp. 2239-2242, 1996.

[12] B. M. Mehtre, M. S. Kankanhalli and W. F. Lee, "Shape Measures for Content Based Image Retrieval: A Comparison", Information Processing and Management, vol. 33, no. 3, pp. 319-337, 1997.

[13] T. K. Shih, J. Y. Huang, C. S. Wang, J. C. Hung and C. H. Kao, "An Intelligent Content-based Image Retrieval System Based on Color, Shape, and Spatial Relations", Proc. National Science Council, R. O. C., Part A: Physical Science and Engineering, vol. 25, no. 4, pp. 232-243, 2001.

[14] A. Makadia, V. Pavlovic, S. Kumar, "A New Baseline for Image Annotation", European Conf. Computer Vision, pp. 316-329, 2008.

[15] N. Vasconcelos, "Image Indexing with Mixture Hierarchies", Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 3, pp. 3-10, 2001.

[16] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation", Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 1002-1009, 2004.

[17] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures", Proc. the Seventeenth Annual Conf. on Neural Information Processing Systems, vol. 16, pp. 553C560, 2003.

[18] P. Guo, "Studies on Effective Computation of Mixture Model Parameters for Bayesian probabilistic Image Automatic Segment", Computer Sceince, vol. 29, no. 8, pp 101-103, 2002. (In Chinese).

[19] D. S. Zhang and G. Lu, "Shape based image retrieval using generic Fourier descriptor", Signal Process: Image Communication, vol. 17, no. 10, pp. 825-842, 2002.

[20] G. Carneiro and N. Vasconcelos, "Formulating Semantic Image Annotation as a Supervised Learning Problem", Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 163-168, 2005.

[21] P. Carbonetto, N. de Freitas, and K. Barnard, "A Statistical Model for General Contextual Object Recognition", Proc. European Conf. Computer Vision, vol. 1, pp. 350-362, 2004.

[22] M. Cavazza, R. J. Green and I. J. Palmer, "Multimedia semantic features and image content description", Proc. 1998 Multimedia Modeling Conf. Lausanne, IEEE CS Press, pp. 39-46, 1998.

[23] D. S. Zhang and G. Lu, "Content-based shape retrieval using different shape descriptors: A comparative study", Proc. IEEE Conf. Multimedia and Expo, pp. 317-320, 2001.

[24] Z. Qiankun, M. Prasenjit and C. L. Giles, "Image annotation by hierarchical mapping of features", Proc. Int. Conf. World Wide Web 2007, pp. 1237-1238, 2007.