An Object Based Auto Annotation Image Retrieval System

Pei-Cheng Cheng[1], Been-Chian Chien [2], Hao-Ren Ke[3], and Wei-Pang Yang[1, 4]
[1] Department of Computer & Information Science, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.
[2]Department of Computer Science and Information Engineering,
National University of Tainan,
33, Sec. 2, Su Line St., Tainan, Taiwan 70005, R.O.C.

[3] Library and Institute of Information Management, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.
[4] Department of Information Management, National Dong Hwa University
1, Sec. 2, Da Hsueh Rd., Shou-Feng, Hualien, Taiwan 97401, R.O.C.

Abstract: In this paper, we proposed an auto annotation image retrieval system. In our system, an image was segmented into regions, each of which corresponds to an object. The regions identified by region-based segmentation are more consistent with human cognition than those identified by block-based segmentation. According to the object's visual features (color and shape), new objects will be map to the similar clusters to obtain its associated semantic concept.

The semantic concepts derived by the training images may not be the same as the real semantic concepts of the underlying images, because the former concepts depend on the low-level visual features. To ameliorate this problem, we propose a relevance-feedback model to learn the long-term and short-term interests of users.The experiments show that the proposed algorithm outperforms the traditional co-occurrence model about 19.5%; furthermore, after five times of relevance feedback, the mean average precision improves from 46% to 62.7%.

Keywords: keyword-based image retrieval, co-occurrence model, relevance feedback.

1. Introduction

The main methods of current image retrieval systems are query by example and query by keyword. When using the system of query by example, it's necessary for users to offer an image as an example for query [5] [8]. Some systems however offers simple tools of drawing that users can draw out the frame of image needed to query [1]. After this, the system will extract the image's features and contrast the similarity of its features with all other images in the image database. Then, the images with highly similarities in the database will be output to users.

Nonetheless, this method of query is not convenient for users; it is hard for users to draw an image. On the other hand, there are two main problems with traditional query by keyword. One is it takes lots of time and man force when doing image annotation for a large amount of image data. The second is that it is not easy to describe the content of an image. An image observed by different people will trigger different feelings, so when image annotation made by different people, the annotation will be different as well. It is a problem with subjectivity.

This paper proposes an auto annotate system to help explain the concept of images and provides semantic query. We apply machine learning and pattern recognition to help establish image annotation, and proposed an objected based image retrieval system. Image was segment in to several homogeneous regions. The acquired region by image segmentation is regarded as the object existed in the image and these objects will be accompanied by words added artificially to form a training data set. For images without image annotation, the system will learn the semantic concept of new images and create annotations for these new images.

By doing so, uses can query images by keyword which is familiar by users. After auto annotated mechanism, we also allow user doing relevance feedback to raise the accuracy.

In the section 2, we will review the related works of image retrieval systems. Sections 3 describe our proposed object based image retrieval system.

Section 4 shows the experiments. Section 5 has a conclusion of our work.

2. Related works

The content of images is composed of objects existed in images, so it should be assured first about what objects exist in images when doing automatic image annotation. For the time being, objects in images are segmented by image segmentation and these objects from image segmentation, one is block based [10] [13] and the other is region based [2] [4] [7]. In practice, block based is easier than region based for block based is only segment images into the same size rectangles and form objects by one or many rectangles. But objects segmented by block based are different from what people can usually recognize. For example, a picture of a tiger could be segmented into several blocks by block based, but it is different from a tiger in peoples' minds.

Region based can segment an image into an object relatively accepted by people, but region based is more difficult. Though many image segmentation techniques are proposed, it is still hard to identify which segmentation method is the best, but the objects segmented by it fit better with people's impression of a person. After acquiring objects of images, a system will infer possible semantic concepts and make image annotation through low level visual features of these objects (such as color, shape, texture, and etc.).

In the auto annotation phases, Mori [13] has proposed a co-occurrence model and calculates the frequency of keywords and objects in the same cluster. In [4], it provides a translation model. It regards the translation model as a dictionary and uses the machine translation to translate the objects of images into another language. In [7], it points out that an object has no equivalence with a specific word. Some words sometimes will occur only when many objects co-occur. For instance, when "landscape" is entered as a keyword, the picture of landscape will appear only when objects of mountain, rivers, falls, and fields co-occur.

For relevant studies nowadays, there are three major representations of helping images establish annotation: Fixed Annotation Model [7], Semantic Network Model [11], and Probability Annotation Vector Model [7]. Fixed Annotation Model represent image annotation is to match a keyword with an image when the keyword and a certain image have high relevance. Semantic Network Model adopts semantic network established by images and keywords to represent the semantic concept of images. Probability Annotation Vector Model, probability vector is used to represent the semantic concept of images. Every element of vectors represents a keyword indicating the possibility of the occurrence of an image.

3. Modified Co-occurrence Region Based Model (MCORM)

In this paper, an approach of automatic image annotation and query is proposed which is called as MCORM (Modified Co-occurrence Region Based Model). This model improves and adjusts the co-occurrence model by the following: (1) Acquire objects by Region Based instead of Block Based in [13]. (2) An object will have a probability to map into similar clusters to get more specified semantic concepts. (3) Emphasize more on the importance object of an image.

And in the Relevance Feedback process, it combines users' concepts and habits to raise the accuracy of query. Make use of this system can avoid the problems caused by image annotation made artificially.

There are three models in this system: Training Module, Annotation Module and Query and Feedback Module.

3.1 Training Module

This paper use color and shapes as image visual features of images. Before getting the color features of two regions, this system will find a basic rectangle which can enclose the whole region and views the image region as a sub-image. After this, the color contrast list in [9] will be adopted to quantify the color of all dots into twenty five colors. After quantifying the color, the color features acquiring method in [3] will be used to get the features in the basic rectangle and do the calculation of color similarity. In terms of the shapes, the similarity of shapes will be calculated by the method mentioned in [6].

The formula of calculating similarity of two objects in images is showed as Eq.(1): $sim\_c_{i,j}$ and $sim\_s_{i,j}$ are the similarity of color and shape of two objects respectively. Here, w_c and w_s are the weights of color and shape respectively.

$$sim_{i,j} = w_c \times sim\_c_{i,j} + w_s \times sim\_s_{i,j} \quad (1)$$

There are three main steps of the training module in this paper: 1. image segmentation; 2. K-Means Clustering [12]; 3. Infer the frequency of keywords in each cluster. The training process is as follows: first, make annotation for all images in the training data base; second, Segment all images in the training data base and get the regions of images; third, All regions segmented from image *"I"* from the training database will inherit all annotation of *"I"* as its own annotation; fourth, Take out all the features which had been segmented with the method mentioned; fifth, Cluster all regions with K-Means Clustering; sixth, Calculate the probability of every keyword of all clusters. In step 6, the formula used to calculate the probability of keyword " w_N " appeared in Cluster K is as Eq.(2):

$$P_{K,N} \approx \frac{|w_N|}{\sum_{i=1}^{M} |w_i|} \quad (2)$$

$|w_i|$ is the times of keyword " w_i " appeared in a certain cluster; M is the number of all keywords.

3.2 Annotation Module

In traditional co-occurrence module, the acquired object is corresponded to the most similar cluster in the training module. In view of that this way is quite inappropriate for its extremeness. This paper does normalizations of the latest 3 clusters of the objects. The probability " $P_{r,N}$ " of keyword w_N appeared in region "*r*" can be obtained by Eq.(3), where s_j^r is the correspondence similarity with cluster *j* in region *r*; $P_{C_j,N}^r$ is the probability of keyword w_N appeared in cluster *j* of region *r*.

$$P_{r,N} = \frac{\sum_{j=1}^{3} s_j^r \times P_{C_j,N}^r}{\sum_{i=1}^{3} s_i^r} \quad (3)$$

Besides, because the center objects in an image would be considered the theme or important in this image, if region r is in the centroid of image *I*, the weight of its semantic concept will be highlighted.

At last, use Eq.(4) to calculate the probability of keyword w appeared in image *I*. *N* is the number of regions in image *I*; α_j is the weight of image *I* in region *j*. $P_{rj,w}$ is the probable probability of keyword *w* in region *j*. In practice in this paper, if region *j* is the centroid region in the image, then α_j is set as 1.3 instead of 1. When using formula (Eq.4), any keyword *w* in which region that *w* has the biggest probability of occurrence will be recorded and so as its region id. This record is called *max_prw*, which is used to record the region representing keyword *w* in the whole image and can be used for user relevance feedback.

$$P(I_w) = \frac{1}{N}\sum_{j=1}^{N} \alpha_j \times P_{rj,w} \quad (4)$$

The way to represent *max_prw* is $(image, word, rid)$ where *image* is the name of this image. If the probability of keyword *w* appearing in all regions is 0, then *rid* of $(image, word, rid)$ is -1. Annotation information for an image *I* as follows:

3.3 Query and Feedback Module

After finishing automatic image annotation, there is a correlation weight of every keyword in the dictionary and every image. When query by keyword is used, users can enter many keywords to query. Due to it's hard to achieve 100% accuracy by automatic making annotation with machine learning, this paper proposed user relevance feedback to raise efficiency and effects.

On the other hand, the information adopted in the training module could be different from the image that needs annotation and limits the learning effect, so we also define and create a Cluster-Keyword Association Map-CKAM. CKAM collects the feedback information of users' query to adjust the correlation between clusters and keywords and raises the query efficiency of query system.

The correlations between clusters and words in CKAM are learned from users, so after using the system for a while, the correlation information can replace the *KPV* of each cluster in previous training module. Here, $wcs_{i,j}$ is the weight of cluster *i* and keyword *j* through feedback. At the beginning, all elements on the list are set as 1.

$$\begin{array}{c|ccccc|}
 & \mathrm{w}_1 & \mathrm{w}_2 & \cdots & \mathrm{w}_{\mathrm{n-1}} & \mathrm{w}_{\mathrm{n}} \\
c_1 & wcs_{1,1} & wcs_{1,2} & \cdots & wcs_{1,n-1} & wcs_{1,n} \\
c_2 & wcs_{2,1} & wcs_{2,2} & \cdots & wcs_{2,n-1} & wcs_{2,n} \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
c_{k-1} & wcs_{k-1,1} & wcs_{k-1,2} & \cdots & wcs_{k-1,n-1} & wcs_{k-1,n} \\
c_k & wcs_{k,1} & wcs_{k,2} & \cdots & wcs_{k,n-1} & wcs_{k,n}
\end{array}$$

Figure 1: Cluster-Keyword Association Map-CKAM

When users use query by keyword, the system will send information through CKAM to assess which clusters of regions have high correlation with the keyword and the system will provide important information of calculating the correlation of keywords and images for querying image. When users enter a keyword set $W = \{w_1', w_2', \ldots, w_m'\}$ to query, the system will calculate the correlation of any image *I* in image database and these keywords *W* by formula Eq.(5).

$$SIM(I,W)=\frac{1}{m}\times\sum_{i=1}^{m}\left\{\beta\times P\left(I_{w_i^{'}}\right)+\gamma\times\left[\frac{1}{N}\sum_{j=1}^{N}SIM\left(r_j,w_i^{'}\right)\right]\right\} \tag{5}$$

There are two parts of Eq.(5). First is the correlation of *I* and *W* calculated by automatic annotation module. Second is the correlation of regions of images and keywords entered by users through CKAM. Here, $P\left(I_{w_i^{'}}\right)$ is the probability of the *i-th* query keyword $w_i^{'}$ appears in image *I*; $SIM\left(r_j,w_i\right)$ is the correlation of keyword w_i and the number *j* region r_j in image *I*; β and γ are weights of these two parts respectively. Through CKAM, which kinds of regions are highly correlated with this keyword can be known. Here, Eq.(6) is the formula of $SIM\left(r_j,w_i\right)$ in Eq.(5), and $wcs_{i,k}$ is the weight of keyword w_i and cluster *k* in CKAM; $s_x^{r_j}$ is the similarity with the *x-th* cluster in Region r_j; $wsc_{i,rc_{jy}}$ is the weight of keyword w_n in CKAM and the *y-th* cluster with similarity in Region r_j.

$$SIM\left(r_j,w_i\right)=\frac{1}{\sum_{i=1}^{k}wcs_{i,k}}\left(\frac{\sum_{y=1}^{3}s_y^{r_j}\times wsc_{i,rc_{jy}}}{\sum_{x=1}^{3}s_x^{r_j}}\right) \tag{6}$$

Through the calculation of Eq.(5), the similarities of keywords entered by users and all images can be acquired and the image with high similarity can be sent back to users though sorting. The images through the query process could be incorrect or not the one that users want, therefore user relevance feedback mechanism is proposed to raise the efficiency of this system. There are three parts of this feedback module.

1. Modify the probability of the occurrence of keywords appeared in assigned images.
2. Modify CKAM.
3. Send back new query results to users.

In the process of modifying the probability of the occurrence of keywords appeared in assigned images is to collect all images of positive and negative examples that users send back and adjust the probability of images and keywords.

There are two parts in this process: 1. Adjust the probability of keywords of positive images. 2. Adjust the probability of keywords of negative images. Adjust the probability of keyword *w* in KPV according to the keyword *w* entered by users and the corresponding positive image *I*. The way of adjusting is to add c_1 to the original probability. All values of keywords appeared in images is between 0 to1, so after adding, they system will check if the value is over 1, if so, then the value will be 1.

Similarly, the system will adjust the probability of keyword *w* in KPV according to the keyword *w* entered by users and the corresponding negative image *I*. The way of adjusting is to subtract c_2 from the original probability. After subtracting c_2, check if the value is smaller than 0; if so, then let the value be 0 and set the rid of *max_prw* where the keyword w is corresponding to image *I* -1.

Through modifying CKAM, the type of region of keywords of query which is more accepted by users can be acquired and can be used as reference information for the next query.

In this process, for the keyword *w* in a keyword set, the system will collect related images recognized by all users and get the *max_prw* of all relevant images which are relevant to the query keywords and find out the region which can represent the image corresponding to the keyword *w*. The rid of *max_prw* in some images assigned by users is set as -1, which means images without any regions corresponding to the keyword *w* is more representable. In this process, there are two steps. First, get all positive images that their rids corresponding to the keyword w in *max_prw* are not equivalent to -1,and add its features of region to Local Dominate Region List (LDRL) and add its region id to Global Dominate Region List (GDRL) , and adjust CKAM. Second, to any positive image I_p, that its rid corresponding to the keyword *w* in *max_prw* equivalent to -1, get the region *r* which has the biggest similarity with the region feature of LDRL in I_p and set the rid corresponding to the keyword *w* of *max_prw* in I_p into the rid of region *r* and adjust CKAM. The process of adjusting CKAM through a positive image is as follow:

1. Get the region that its serial number is the rid of *max_prw* corresponding to the keyword *w* in image *I*.
2. Get the latest 3 cluster serial number (rc_1, rc_2, rc_3) and corresponding similarity (rs_1, rs_2, rs_3) of the region that its serial number is *rid.*
3. Add these three similarity values on corresponding $wcs_{rc_i,w}$ in CKAM. Here, $wcs_{rc_i,w}$ is the value corresponding to the keyword *w* in *i-th* similar cluster serial number of region *r* in CKAM.

After this, the system will recalculate the correlation between all images and feedback information. The process of any image *I* through user relevance feedback and the similarity of query set *W* and *I* can be known by Eq.(7). Here, *u* and *v* are constants; m is the number of keywords in keyword *W*; $P\left(I_{w_i^{'}}\right)$ is the probability of the keyword $w_i^{'}$ in image *I*. *G* is the regions set that in GDRL while modifying CKAM; $|G|$ is the number of regions in G; S_{r_j,G_k} is the similarity between region r_j in image *I* and the *k-th* region G_k in the whole *G*.

$$SIM(I,W)=u\left(\frac{1}{m}\sum_{i=1}^{m}P\left(I_{w_i^{'}}\right)\right)+v\left(\max_{j=1\,to\,N}\left(\frac{1}{|G|}\sum_{k=1}^{|G|}S_{r_j,G_k}\right)\right)$$

(7)

Through the process of user relevance feedback to modify CKAM and after using this system for a while, the correlation between words in CKAM and each cluster can be more accepted by users. So, the information can be used to replace each KPV in the training database. The probability $P_{K,N}$ appeared in cluster *K* of any keyword *N* can be modified by Eq.(8). Here, t_i is the times of keyword w_i has been queried; $wcs_{K,N}$ is the value corresponding to the keyword w_N and the cluster C_K in CKAM.

$$P_{K,N}=\frac{\left(\dfrac{wcs_{K,N}}{t_N}\right)}{\left(\displaystyle\sum_{i=1}^{n}\frac{wcs_{k,i}}{t_i}\right)} \qquad (8)$$

4. Experiment

In this section, the practice of the system will be introduced and the efficiency of this system will also be evaluated.

To solve the problems that users might encounter while doing the image query, the multiple image query system is proposed in our system. The interface of this system is shown in Figure 2. There are three ways of query in this system: Query by example, Query by object and Query by keyword.

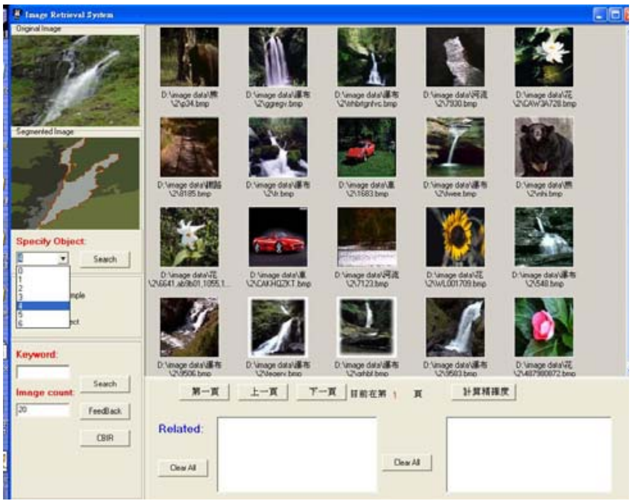

Figure 2: Search “fall” by query by object

In this experiment, there are 1200 images in total. 800 images of the 1200 are used for the data of the training data set and the rest 400 images are used for testing. In the process of training, to the images in the training data set, one to seven keywords are given to make image annotation. After making annotation, deleted those keywords not often appearing first and then there are 64 keywords can be used for making automatic image annotation.

When put MCORM into practice, K-Means is the tool of clustering in this paper and K is set as 300. Eq.(1) is used to calculate the similarity of regions. In practice, w_c of Eq.(1) is set as 0.8, and w_s is set as 0.4. In Eq.(1), the range of color similarity is between 0 and 2 and the shape similarity is between 0 and 1, so the value of similarity is between 0 and 2. In this experiment, 6 keywords are used to test.

This paper is to make image annotation by revising the traditional co-occurrence model and Region Based image segmentation. So, in order to compare the efficiency of Region Based and Block Based image segmentation while using co-occurrence model, besides MCORM, there are also two other systems put into practice to evaluate the efficiency was implemented:

(1) BCOM (Block Based Co-occurrence Model)： This model uses block based image segmentation to acquire objects in images and make annotation for images by co-occurrence model.

(2) RCOM (Region Based Co-occurrence Model)： This model uses region based image segmentation to acquire objects in images and make annotation for images by co-occurrence model.

2 to 5 keywords in each category are taken for testing. Table 2 is the whole MAP of above 3 approaches.

	BCOM	RCOM	MCORM
MAP	25.55%	36.84%	46.00%

Table 1: The whole MAP of above 3 approaches

From above experiment results, it is obvious that in overall efficiency, MCORM is the best, RCOM is the second and BCOM is the worst. The objects acquired by Region Based image segmentation are better than through Block Based in terms of the visual acceptance to people. Therefore, the effects made by RCOM are better than BCOM. MCORM gets the objects of images by Region Based and through above modification proposed in this paper, it indeed raises the efficiency of the system as we can see from the experiment results.

At last, query is made through the keywords in Table 2 and evaluation is also conducted to see whether let system learn by interacting with users can raise the efficiency of the system or not. In the process of every query, users can decide whether the result from the system is relevant with the objective or not and users will send this information back to the system.

User Relevance Feedback can raise efficiency of the whole system and after the system learns from the relevant feedback from users for 4 to 5 times, the advancement of the system is getting smooth and reaches a good efficiency level. After the fifth feedback, the overall map of the system has raised from 46.00% to 62.70%.

5. Conclusions and Future Work

This paper proposed a Modified Region Based Co-occurrence Model (MCORM) which revised from [13]. In the experiment in Chapter 4, it shows query in the six main categories by RCOM is better than BCOM by 11.29% in the MAP. MCORM has better efficiency than BCOM and RCOM. It progresses by 20.45% and 9.16% respectively in the MAP. User Relevance Feedback proposed in this paper has also made achievements in raising the efficiency of the system. Over all, after using the feedback system, the value has raised from 46% to 62.7% in the MAP.

Reference

[1] A. D. Bimbo and P. Pala, “Visual Image Retrieval by elastic matching of user sketches,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, Issue 2, 1997, pp. 121-132.

[2] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, “Blobworld: A system for region-based image indexing and retrieval,” *In Third International Conference on Visual Information Systems, Lecture Notes in Computer Science*, 1999, pp. 509-516.

[3] L. Cinque “Color-based image retrieval using spatial-chromatic histograms,” *Image and Vision Computing*, Vol. 19, Issue 13, 2001, pp. 979-986.

[4] P. Duygulu, K. Barnard, N. D. Freitas, and D. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” *In Seventh European Conference on Computer Vision*, 2002, pp. 97-112.

[5] M. Flickner, H. Sawhney, W. Niblack, and J. Ashley, “Query by Image and Video content: The QBIC System,” *IEEE Computer*, Vol.28, Issue 9, 1995, pp. 23-32.

[6] D. Zhang and G. Lu, “Improving retrieval performance of Zernike moment descriptor on affined shapes,” *IEEE International Conference on Multimedia and Expo*, vol.1, 2002, pp. 205-208.

[7] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic Image Annotation and Retrieval using Cross-Media Relevance Models,” *ACM Conference on Research and Development in Information Retrieval*, 2003, pp. 119-126.

[8] J. R. Smith and S. F. Chang, “Visualseek: a fully automated content-based image query system,” *In Proceedings of ACM Multimedia*, 1996, pp. 87-98.

[9] K. C. Ravishankar, B. G. Prasad, S. K. Gupta, and K. K. Biswas, “Dominant color region based indexing for cbir,” *International Conference on Image Analysis and Processing*, 1998, pp. 887-892.

[10] J. H. Lim, “Learnable visual keywords for image classification,” *Proceedings of the fourth ACM conference on Digital libraries*, 1999, pp. 139-145.

[11] Y. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang, ”A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems,” *ACM Multimedia*, 2000, pp. 31-37.

[12] J. McQueen, “Some methods for classification and analysis of multivariate observations,” *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[13]Y. Mori, H. Takahashi, and R. Oka, “Image-to-word transformation based on dividing and vector quantizing images with words,” *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.