Automatic Image Annotation and Retrieval Using the Latent Dirichlet Allocation Model

Tong Zhang¹, Zhe-Ming Lu², Kap Luk Chan³ and Zhen Li¹

¹Department of Automatic Test and Control, Harbin Institute of Technology Harbin, Heilongjiang, China
E-mail: tong.zhang12@yahoo.com
²School of Aeronautics and Astronautics, Zhejiang University Hangzhou, Zhejiang, China
E-mail: zheminglu@zju.edu.cn
³School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
E-mail: eklchan@ntu.edu.sg

Abstract: Content-based image retrieval faces a vital problem, namely “semantic gap” that exists between low level features and semantic concept. In order to solve this problem, image automatic annotations that allow users to access a large image database with textual queries are put forward. In this paper, the main study concentrates on an automatic image annotation method based on vector quantization (VQ) algorithms and Latent Dirichlet Allocation (LDA) model. VQ is used as a clustering and condensing technique. LDA model is introduced from text based information retrieval, which is widely studied nowadays and proved to be effective on discrete data processing and demission reduction. The experiment is carried out on the platform of MATLAB and applied to the Corel database of 400 images. The presented experiment results demonstrate the effectiveness of the proposed method in its application of automatic image annotation and keyword based image retrieval.

Key words: Automatic Image Annotation, LDA, VQ, Semantic Gap.

1. INTRODUCTION

The availability of huge amounts of multimedia documents necessitates a careful design and an efficient implementation of information retrieval systems that facilitate storage, retrieval and browsing of not only textual, but also image, audio and video files. To acquire the description of image content manually is difficult, since hundreds of words can hardly represent an image wholly. Traditional keyword-based multimedia retrieval systems cost a large amount of human labours to annotate images; thereby the content-based image retrieval (Feng et al. 2003) is brought forward, which tries to retrieve images directly and automatically based on their visual contents such as colour, texture and shape. However, the so-called “semantic gap”, limits the information communication between low level visual characters and high level information. Specifically, the contents of image include not only colour, texture, shape, structure and other low level visual characters, but also object, objective and other middle level features and some high level information such as scene, emotion and so on. In

the general field of retrieval, the relationship between the similarities at low level and high level is very hard to deduce (Eakins 1996). In order to solve this problem, semantic based image annotation and retrieval becomes a research focus.

In order to realize semantic based image retrieval, there are two problems that should be fixed. The first is the method that can map low level visual character to high level semantic meaning. The second is the depiction way of high level semantic meaning. Actually, the mapping process is a knowledge deduction process of high dimensional data. The application of the image segmentation can divide the image into meaningful regions. As an effective image compression algorithm, vector quantization (VQ) has received interests from many researchers (Sun and Lu 2002). In this paper, an image region characteristics clustering and classification algorithm based on VQ is adopted.

With regard to the second problem, Li and Wang (2003) used categorized images to train a dictionary of hundreds of concepts automatically. Wavelet-based features were used to describe local color and texture in the images. Chen et al. (2005) introduced cluster-based image retrieval to capture
This paper is to explore the combination of LDA model and VQ for image clustering, annotation and retrieval in an image database. Compared to text data, amazing similarities can be found in image information: both of their data are very huge and have high dimensions; moreover, similar objects can be easily found in the same image collection and the synonyms always exist in one document corpus. To fulfill this assumption, the “word” and “document” should be defined first. There are many choices for a “word”, namely a pixel, a block, a segmented image region and an image itself. If “word” is determined, “document” is confined. In this paper, normalized feature of a segmented image region is defined as “visual word”, then, an image will be a “visual document”. Every classification of images will be corresponded to “Image Corpus”.

2.2. The Application of VQ

VQ is not only widely used in the image compression domain, but also an effective technique for generative clustering and data classification (Sun and Lu 2002). In this paper, VQ is applied to cluster region features by known image group, which can be regard as an unsupervised learning process.

Assume that codebook $C = \{c_0, c_1, c_2 \ldots c_{N-1}\}$, where $N$ denotes the length of codebook. Then, the vector $x = \{x_0, x_1 \ldots x_i\}$, corresponds a codeword, which is also the mapping codeword $c_i$, should satisfy the following conditions: $d(x, c_i) = \min_j d(x, c_j)$. The distortion measure between vector $x$ and $c_i = \{c_{i,0}, c_{i,1}, c_{i,2} \ldots c_{i,k-1}\}$ is described as square error, which is $d(x, c_i) = \sum_{j=0}^{k-1} (x_j - c_{i,j})^2$.

In this paper, $I_i$ ($i = 1, 2, \ldots, M$) denotes an image, and $I_{ij}$ ($i = 1, 2, \ldots, M, j = 1, 2, \ldots, N$) denotes a segmented image region. Using K-means algorithm, centroid point of each group is calculated, and then the codebook is generated.

The extracted region features $x = \{x_0, x_1 \ldots x_i\}$ can be regarded as points in the high dimensional space. Region features after clustering corresponds to codewords $c_i$ in the codebook $C = \{c_0, c_1, c_2 \ldots c_{N-1}\}$. The condensed codebook can be regarded as a semantic lexicon in the image database.
3. NUMERICAL EXPERIMENTS

400 images in Corel Image database are selected for simulation, which includes 20 categories. The image categories are respectively: train, cloud, beauty, mushroom, sky, face, ship, horse, grass, dolphin, waterfall, stalactite, cat, car, slide, mountain, texture, flower, dog, sun.

3.1. Image Segmentation

In order to describe the meaningful region of images, the segmentation process is necessary. After comparison of the computation speed and segmentation results by different traditional image segmentation algorithm, the watershed algorithm is chosen. Using Watershed algorithm, the 400 images are segmented into 3683 regions, each of which includes 18 dimensions of features: three dimensions of colour features, 12 dimensions of texture features and three dimensions of shape features.

3.2. Region Features Clustering by Known Information

Based on the segmented results, 3683 region features are extracted; all the features are normalized by Guass normalization. Then VQ is applied to cluster these features by the known image category information.

By using K-means algorithm, the region features are clustered into 200 categories, each of which represents some image objectives with the same characteristics. After that, each region category is corresponded with a codeword in the VQ codebook. Every region in every image can be found a corresponding codeword in the codebook. In the simulation, a dictionary with 200 “visual words” is generated.

After classification, the region code for each image can be statistically obtained. Taking the mth image \( I_m \) for example, the region codeword \( C_{m,n} \) of this image is the image region features after VQ classification. Code words in the code book are the proposed “visual words”.

3.3. LDA Model Training and Automatic Image Annotation

After that, the occurrence frequency of every “visual word” in every image is calculated, which can be further recorded into the training samples of the LDA model. The structure of training sample set can be depicted as follows. Each row represents an image. So there are 400 rows of input data for 400 images. Each image make up of several regions, which can be represented by a unique codeword in codebook. Therefore, the data form is \( \alpha : \beta \), where \( \alpha \) denotes image region code, while \( \beta \) represents the occurrence time of region \( \alpha \) in this image.

After computing based on LDA model, a similarity matrix \( \beta \) between each image and “visual words” can be obtained. The similarity between latent topic and “visual words” and the similarity between some images with known categories and “visual words” are compared by using histogram algorithm. And thus, the similarity between each topic and image can be annotated by a number of keywords. Given a testing image, based on the similarity between image and “visual words”, the automatic annotation results of this image can be obtained. Fig. 1 shows four randomly selected images and their annotation results by image automatic annotation system.

![Figure 1: Experiment Results of Automatic Annotation](image)

In the training module, a probability model is produced by using LDA model. In this model, the similarities between images of the same type are relatively big, while those between images of different types are smaller. Once the occurrence relationship between an image and its region code is recorded, latent parameters are computed by using LDA model. After 400 training samples are recorded, LDA model can obtain parameters \( \alpha \) and \( \beta \) (Blei et al. 2003). The first four similar texts between a “visual word” and a latent topic are chosen to annotate the images in the database. After the annotation, every image will be corresponded to four keywords. The automatic annotated four keywords are recorded in order by their similarity (Shown in Fig. 1). It is easily observed that not all the annotation results are correct.
3.4. Keyword based Image Retrieval

After the training part of LDA model, every image has been corresponded to four keywords. For example, when users want to search keyword “sailboat”, the images annotated as “sailboat” are retrieved, ordered by their first keywords and then the second ones. Fig. 2 shows the first 16 images with “sailboat”. Fig. 3 shows the search results when the keyword is “beauty”. VQ also applied in this part, the generated codebook makes contribution to the retrieval accuracy and speed.

4. CONCLUSIONS

This paper proposed an image annotation and retrieval method with combination of VQ and LDA model. VQ is used for the region features clustering and generated more effective data information for the later application of LDA model. After LDA model is generated, VQ is applied for the classification of the unknown images. The experiment results show that it is an effective method for automatic image annotation and keyword based image retrieval. VQ also accelerates the automatic unknown image annotation and the keyword retrieval processes.

REFERENCES