BUAA AUDR at ImageCLEF 2012 Photo Annotation Task

Lei Huang, Yang Liu

State Key Laboratory of Software Development Enviroment, Beihang University,
100191 Beijing, China
huanglei@nlsde.buaa.edu.cn
liuyang@nlsde.buaa.edu.cn

Abstract. This paper presents the participation of the BUAA AUDR group at ImageCLEF 2012 in the Photo Annotation and Retrieval task. We selected Flickr photos as data set to perform visual concept detection, annotation and retrieval. In this task, we had proposed multi-modal approaches that considered visual information and Flickr user tag information. We presented our visual-based and tag-based photo annotation methods, and also proposed Annotation Refining Algorithm (ARA), which attempted to make use of the relation between concepts to improve the annotation result. It was our first time to participate the Photo Annotation and Retrieval task. We submitted 3 runs totally and the purely visual submission using the global visual features and BoW features get better performance.

Keywords: ImageCLEF, Photo Annotation, Flickr

1 Introduction

This paper presents the first participation of the BUAA AUDR group at ImageCLEF photo annotation and retrieval task.

ImageCLEF 2012 includes four types of tasks: medical image retrieval, photo annotation, plant identification and robot vision. In the photo annotation task, the aim is to analyze a collection of Flickr photos in terms of their visual or textual features in order to detect the presence of one or more concepts. The detected concepts can then be used for the purpose of automatically annotating the images or for retrieving the best matching images to a given concept-oriented query. This task provides 15000 images for training and requires the annotation of 10000 images in the provided test corpus according to the 94 pre-defined categories.

We proposed multi-modal approaches that considered visual information and Flickr user tag information. We presented our visual-based and tag-based photo annotation methods, and also attempted to make use of the relation between concepts to improve the annotation result.

The remainder of this paper is organized as follows. In section 2 we describe our approaches in detail. And our submitted runs are discussed in section 3. Then we conclude in section 4.

2 Approaches

For the visual concept annotation task, we proposed multi-modal approaches that considered visual information and Flickr user tag information. We presented our visual-based and tag-based photo annotation methods in this section.

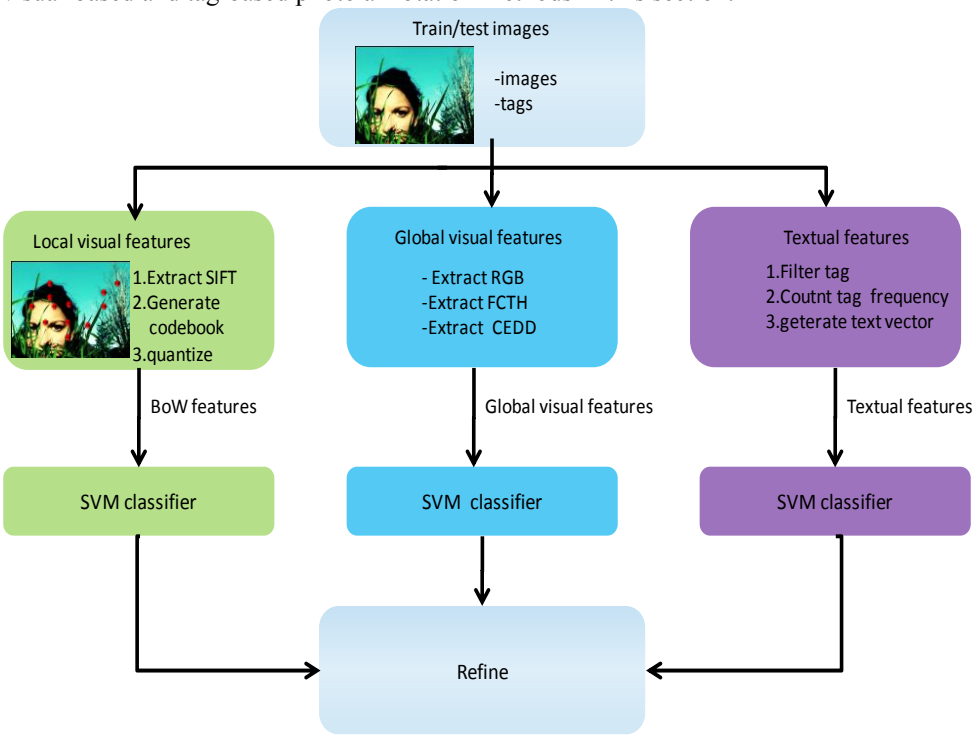

Figure 1: Phone Annotation Process

2.1 Visual-based Photo Annotation

Recently, the Bag-of-Words (BoW) [1] model has been very popular in image recognition and retrieval. In this model, the key points extracted are quantized to visual words, and an image is represented as a frequency histogram of these words. We followed BoW model. We adopted the implementation described in [2] for extracting local feature. Harris -Laplace was used to detect interest points and SIFT descriptor was extracted. These descriptors were then quantized in visual words. To form the codebook, we randomly selected approximate 1.5 million descriptors from all descriptors extracted from the training images for clustering. We used k-means

clustering method to group these descriptors into K (K=200, 1000, 2000) clusters. The codebook was formed by picking K cluster centers computed from the K clusters.

Soft assignment was used to form the feature vector. We used the mapping from [3]. Let us define V a visual vocabulary set, and v_d the visual word from its corresponding dimension d and l a local feature. Then the BoW mapping m(l) is defined as:

$$m(l)=\begin{cases}\dfrac{\exp(-\frac{1}{D_v}dist(l,v))}{\sum_{w\in V_{topN}(l)}\exp(-\frac{1}{D_w}dist(l,w))} & if\ \ Rank(dist(l,v))\le N\\ 0 & otherwise\end{cases} \tag{2.1}$$

Where dist(l,v) is the distance between the local feature l and the visual word v, Rank(dist(l,v)) is the rank of the distance between l with $v\in V$, sorted in increasing order. $V_{topN}(l) = \{v|v\in V, Rank(dist(l,v))\leqslant N\}$ .parameter D_v is estimated as the average distances to all local features from train images which had v as the nearest visual word.

For an image, the BoW mapping can be calculated as follows:

$$m(im)=\frac{1}{|F(im)|}\sum_{l\in F(im)}m(l) \tag{2.2}$$

Where $F(im)$ is the set of all local features from the image.

Besides Bow Feature, for per image we also extracted the global features such as RGB histogram, CEDD and FCTH.

For per semantic concept, we trained an SVM classifier [4] to perform a binary assignment to an image. We used the probabilistic output of the SVM.

2.2 Tag-based Photo Annotation

Since metadata of images were provided for annotation task, using text-based method in image annotation became possible. Compared with visual features of images, metadata was usually more semantic, so it could be used for identification of many abstract concepts, which was difficult with visual features.

For the experiment presented in this paper, a tag-based image annotation method, which makes use of custom tags in image metadata in determination of existence of specific concepts, was implemented and used besides visual-feature-based annotation methods. It is a rather simple algorithm, and its progress is presented below:

1. Extract unique tags from the training dataset. Since tags are attached by users, they don't have any common rules, so even tags that share the same meaning may appear different in the same dataset. Therefore, the amount of tags could be intolerably large, and it would be difficult to generate the codebook.
2. Select tags with high frequency to form a small tag set. In this experiment, totally 2400 most frequently used tags were selected from the extracted tag set.
3. Each image is given a tag-based feature vector.

4. Train SVM classifiers for each concept with tag-based feature vectors of images.

Through the above steps, SVM classifiers are obtained, and they could be used to identify the concepts in annotation task just like any other kind of classifier.

2.3 Annotation Refining Algorithm

Annotation refining algorithm (ARA) works after all previous annotation process. Its input is a complete annotation which used to be directly submitted. That is to say, Annotation refining is an extra process after common annotation process, and it is used to improve the annotation result.

It could be observed that each concept is not totally individual, e.g. "night" is usually along with "moon" and hardly with "sun". Therefore, if each concept in a picture is identified separately, the information provided by their relations could be neglected, so the annotation refining algorithm presented here attempts to make use of this usually ignored information to improve the annotation result.

In short, the process of annotation refining is simply minimization of an evaluation function. Moreover, in order to learn parameters in this evaluation function in specific dataset, ARA needed a training process. The training process and evaluation function form the main part of ARA.

The ARA training process is a series of solution of a Linear Programming (LP) problem, which is

$$x_j = \mathbf{b}_j^T \mathbf{x} + a_j \,. \tag{2.3}$$

In equation (2.3), $\mathbf{x}$ is a vector with the jth element being x_j, which means the possibility of the existence of the jth concept in the given picture. In this experiment, x_j is the score of the jth concept in *annotations_raw* dataset. For training, $\mathbf{x}$ is given by the training dataset. $\mathbf{b}_j$ and a_j are parameters that need to be learned. They are simply the parameters in general LP problems. By solving this problem in the whole training dataset, a pair of parameters could be obtained for each concept.

The minimization problem needed to be solved in ARA can be presented as below:

$$\begin{aligned} \min \quad & f(\mathbf{x}) = \mathbf{a}^T(\mathbf{B}^T - \mathbf{I})\mathbf{x} + \frac{1}{2}\mathbf{x}^T(\mathbf{B} - \mathbf{I})(\mathbf{B}^T - \mathbf{I})\mathbf{x} + \frac{1}{2}\alpha\mathbf{x}^T\mathbf{x} - \alpha\mathbf{x}'^T\mathbf{x} \\ \text{s.t.} \quad & x_i \in [-1,1], i = 1,2,\cdots,n \end{aligned} \tag{2.4}$$

In the above problem, $\mathbf{a}$ and $\mathbf{B}$ are learned parameters in training stage.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \mathbf{B} = \left(\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n\right) \tag{2.5}$$

x' is the input annotation result. α is tuning parameter used to balance the importance of previous annotation process and refining, i.e. the refining result would be closer to **x**' with a bigger e α , and vice versa.

This is a convex optimization problem and has only one solution. Refined annotation result could be obtained as the solution **x***.

3 Experiments and Results

We submitted three different runs. The purely visual submission (BUAA_AUDR_1 in Table 1) was adopted using the global visual features and BoW features. For each concept, we selected separately the best classifier from a set of classifiers by MAP values obtained on 5-fold cross-validation on the training data. BUAA_AUDR_1 obtained the rank 52 on 80 submissions, with a MAP value of 0.142. This value was 0.08 lower than the median value for these runs, 0.228.

The purely text submission (BUAA_AUDR_2) obtained the rank 76. The result was poor, but it got a better result than the visual method for the valid set which we adopted to validate these methods. Maybe the tag-based photo annotation method was not robust.

BUAA_AUDR_3 was adopted as a multi-modal approaches with ARA. However, it didn't obtain better results than BUAA_AUDR_1. For two reasons:

1. We used a linear integration which might lead to a worse result than purely visual submission for that our purely text submission had a poor result and was not robust.
2. ARA had a shortcoming that poorly recognized concepts could spoil classification rates of better performing classes.

Table 1. Results by MAP for the photo annotation

Submission	Result File	Modality	MAP
BUAA_AUDR_1	result1_visualOnly	V	0.1423
BUAA_AUDR_2	result2_textOnly	T	0.0723
BUAA_AUDR_3	result3_textAndVisual	V&T	0.1307

4 Conclusion

This article describes the approaches and results of BUAA AUDR group at ImageCLEF 2012 photo annotation task. We submitted 3 runs totally and their results were not competitive among all the submitted runs. As this is our first time to participate in this task, we will investigate our methods to find its weak and improve its performance.

References

1. Leung T, Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV 43 (2001) 29–44.

2. Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, Evaluating Color Descriptors for Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence volume 32 (9), pages 1582-1596, 2010.

3. Alexander Binder, Wojciech Samek, Marius Kloft. The joint submission of the TU Berlin and Fraunhofer FIRST(TUBFI) to the ImageCLEF2011 Photo Annotation Task.

4. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.