Automatic Semantic Annotation for Image Retrieval Based on Multiple Kernel Learning

Alin Hou
Collage of Computer Science and Engineering,
Changchun University of Technology,
Changchun, China
houalin@sohu.com

Liang Wu
Collage of Computer Science and Engineering,
Changchun University of Technology,
Changchun, China

Chongjin Wang
Collage of Computer Science and Engineering,
Changchun University of Technology,
Changchun, China
feiyang2010jin@126.com

Fei Li
Collage of Computer Science and Engineering,
Changchun University of Technology,
Changchun, China

Junliang Guo
Collage of Computer Science and Engineering,
Changchun University of Technology,
Changchun, China

Abstract—The image low level features have a gap with the high level semantic feature which human understand, the researches begin to focus on automatic semantic annotation image retrieval rather than on content image retrieval. The previous methods mostly base on single kernel learning, which has some limitations, which is no effective feature information processing. In this article, an automatic image annotation framework is proposed based on Radial Basic Kernel function combining Spatial Pyramid and Histogram Intersection Kernels. This framework utilizes multiple kernel learning, the k-mean clusters the training images to dictionary. The feature parameters are optimized Spatial Pyramid and Histogram Intersection Kernel. Then radical basic kernel function trains the data and predicts the labels of the images. Spatial Pyramid, reflecting features of location information, is more exact than Bag of Word. Experimental results demonstrated that the proposed framework effectively improves the performance of image annotation and outperform state-of- the-art on the multiple databases.

Keywords-semantic annotation; spatial pyramid (SP); histogram intersection kernel (HIK); Radical Basic Kernel function; multiple kernel learning (MKL).

I. INTRODUCTION

Nowadays, as the image capture devices are easy to gain and internet service for sharing images is more convenient, more and more images are available. It becomes a serious problem that how to obtain the images what we want and manage the images efficiently and effectively. There are large amount of researches on image retrieval in the past two decades. Most of the researches concentrate on content based image retrieval. However, the research method leads to a semantic gap between image low-level feature with image semantic understandable by humans. To solve this problem, some researches begin to consider semantic image retrieval. We need to annotate image firstly, how to annotate exactly image become a research focus. Some researchers used the probability of associating words with image region to annotate images automatically. Such as, Mori et al [1] used a Co-occurrence Model in which they looked at the co-occurrence of words with image regions created using a regular grid. Latterly, researchers pay attention to machine learning approaches and apply them into image annotation, Kobus Barnard [2] present a statistical model for organizing image collection which integrates semantic information with associated text and visual information of image feature and the model learns relationship between text and image feature. In recent years, Multiple Instance Learning (MIL) [3] and image annotation with relevance feedback [4] are researched. However, they still can't satisfy to us without sufficient recall ratio and precision. The method is improved by multiple kernel learning. It appears large advance, which optimize feature parameters.

The rest of the paper is organized as follows. Related work is been discussed in section 2. Then, our model and algorithms are described in detail in the section 3.This is followed by experiment setting and result evaluation in the section4. In the section 5 we show experimental results for the different models and discussion. Section 6 summaries and concludes the paper.

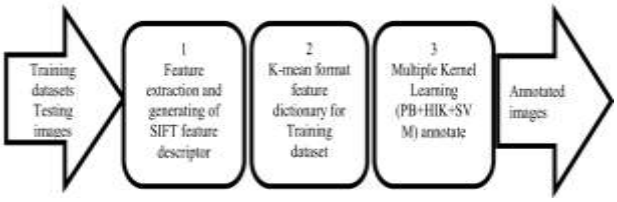
II. RELATED WORK

There has been large amount of work on the image semantic annotation for image retrieval. Some research focus on relevance models, J. Jeon [5] propose two relevance models which combine the probabilistic annotation-based cross-media relevance model (PACMRM) and direct-retrieval cross-media relevance

mode (DRCMRM). The relevance model is built that the probabilities between the images blobs with vocabulary. There are other researches in literature regarding image retrieval and feature vector identification using Bayesian classifiers[6], which propose a novel Bayesian hierarchical method for estimating mixture models of Gaussian components. Recently, machine learning method has been used extensively in automatic image annotation, such as SVM [7], which proposed a novel approach to Automatic Image Annotation which combines both Hidden Markov Model (HMM) and Support Vector Machine (SVM). Wang Chong [8] that based on probabilistic models develop two models: multi-class sLDA (supervised Latent Dirichlet Allocation) and multi-class sLDA with annotation. It also guided that classification and annotation have some relationship. A lot of image retrieval systems adopt the scale-invariant feature transform (SIFT) descriptor [9] to capture local information and adopt BoW model to conduct object matching, but it can't represent the information of location. All above researches have made a progress on image annotation, but there are still a distance for application in the real environment.

III. THE MODEL AND ALGORITHMS

SIFT is used to exact image feature as descriptor, and feature dictionary of training images is built by k-mean algorithms, then a spatial pyramid and HIK are used to SVM annotation. The processing is shown in the Fig 1.



A. Feature Descriptor

Images are composed of array of pixels, so they are represented by low level features in the image classification and annotation. The first step in the image semantic annotation is to extract efficient and effective visual feature from these pixels. Appropriate feature representation significantly improves the performance of the rest of progress. These low level descriptors are generally divided into two categories: global and local descriptor. The global features are mainly color, texture, shape, contours, etc. The local features include Discrete Fourier Transform, Harris, Scale Invariant Feature Transform (SIFT), etc. Ot ́avio A. B. Penatti [10] presents a comparative study of color and texture descriptors considering the Web as the environment of use. They are easy to be exacted, but the problem is low precision and stabilization. SIFT [11] features are invariant to rotation, illumination, scaling, translation and even affine transformation. It has been empirically proven to be one of the most robust among the local invariant feature descriptors with respect to different geometrical changes. The basic idea is to look for the extreme points in the scale space, then filter these extreme points to find the stable feature points known as key points, and finally compute local attribution of orientation gradient and describe the key points by $4 \times 4 \times 8$ vectors. In this paper, SIFT features is chosen as the local semantic descriptors.

B. Spatial Pyramid Matching and Histogram Intersection Kernel Model

After local features are exacted, the number of key points varies in the image, researchers such as Li Fei-Fei from the Stanford University were the first to apply Bag of Words model into computer image process as a sort of features [12]. But this model neglects the connections and relative positions of features, so the result may be not precision, K. Grauman proposed Spatial Pyramid Model [13]. Let x and y to be two set of vectors which represent two image features in a d-dimensional feature space. Spatial pyramid matching places a sequence of increasingly grids over the feature and taking a weighed sum of the number of matches that occur at each level of resolution. Let us construct a sequence of grids at resolutions $0,1,...,L$ such that the grid at level ℓ has 2^{ℓ} cells along each dimension, for a total of $D = 2^{d\ell}$ cells. Let H_x^{ℓ} and H_y^{ℓ} denote the histograms of X and Y at this resolution, so that $H_x^{\ell}(i)$ and $H_y^{\ell}(i)$ are the numbers of points from X and Y that fall into the i th cell of the grid. Then the number of matches at level ℓ is given by the histogram intersection function:

$$I(H_x^l, H_y^l) = \sum_{i=1}^{D} min(H_x^l(i), H_y^l(i)) \qquad (1)$$

In the following, we abbreviate $I(H_X^{\ell}, H_Y^{\ell})$ as I^{ℓ}.

Because the number of matches found at level ℓ also includes all the matches found at the finer level $\ell + 1$. Therefore, the number of new matches found at level ℓ is given by $I^{\ell} - I^{\ell+1}$ for $\ell = 0, \ldots, L-1$. The weight associated with level ℓ is set as $\frac{1}{2^{L-\ell}}$. After putting all the pieces together, a pyramid kernel can be expressed as:

$$\begin{aligned} K^L(x, y) &= I^L + \sum_{\ell=0}^{L-1} \frac{1}{2^{L-\ell}} (I^{\ell} - I^{\ell+1}) \\ &= \frac{1}{2^L} I^0 + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}} I^{\ell}. \end{aligned} \qquad (2)$$

The image space breaks down multiple scales with the method of the construction pyramid. Feature vectors are quantized all into M discrete types, each channel m gives us two–dimensional vector found in the respective images. The final kernel is then the sum of the separate channel kernels:

$$K^{\ell}(x, y) = \sum_{m=1}^{M} k^{\ell}(x_m, y_m) \qquad (3)$$

When $\ell=0$, Model would degenerate into a BOW model. Considering the fact that spatial pyramid is simply a weighted sum of histogram intersections, and $c \times min(a,b)$ for positive number. As a single histogram intersection of long vectors, K^{ℓ} is formed by concatenating the appropriately weighted histograms of all channels at all resolutions. For L levels and M channels, the resulting vector has dimensionality:

$$M \times \sum_{\ell=0}^{L} 4^{\ell} = M \times \frac{1}{3} \times \left(4^{L+1} - 1\right) \quad (4)$$

Experiments in the article have been done by using the setting of $M = 300$ and $L = 2$, according formulation (4) resulting in 6300-dimensional histogram intersections. The parameters must be efficient because the histogram vector is refined and sparse.

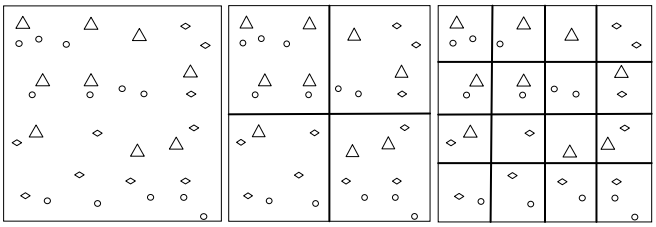


Figure 2. The constructing a three-level pyramid.

In the Fig 2, the circle, triangle, diamond represent respectively various words of an image patch after image clustered into dictionary through k-mean. Each special histogram is weighted according to eq. (2).

- The image is divided into specific size blocks, such as left to right: 1×1, 2×2, 4×4, and then count the number of different words in each box;
- Calculate histogram of each different level within the block from left to right;
- Assign weight and Connect histogram of each level in order, and the weight increases gradually from left to right. Here we use three levels spatial pyramid matching.

C. Multiple Kernel Learning

SVM theory is an effective method to solve problems in non-linear model analysis. Based on it, Radial Basic Kernel function is taken into the proposed framework. To construct multiple kernel learning model, the simplest and most common method is to consider the convex combination and the basic kernel functions as follows:

$$k(x,y) = \sum_{m=1}^{M} \beta_m k_m \left(x, y\right) \quad (5)$$

In this formulate, $k_m \left(x, y\right)$ is the basic kernel function, M is the number of total basic functions, β_m is the related weighted factor aiming at optimization. The kernel functions are combined during training stage in the model. Moreover, combinatory parameters of kernel function can be optimized for next step annotation in SVM. The features extracted from input data are transformed by mapping them to the kernel function space. Then summarizing all the features by combinatory parameters $\beta_1, \beta_2, \ldots, \beta_M$ and get the combined kernel through linear combination. At last, classification and annotation is completed by Radial Basic Kernel function, and expressed as follows:

$$k(x, y) = \exp(-\gamma \sum_{i=1}^{n} (x_i - y_i)^2) \quad (6)$$

Radial Basic Kernel function treats each dimension of feature x and y equally and often cannot represent the inner structure of feature. Multiple kernel learning can solve the problem. Suppose divide a pyramid feature into m blocks, each of the length L so that n=ML. Here each block corresponds to a block in certain layer of grid in the pyramid. Assign the initial values $d_1, d_2, ..., d_m$ to the blocks, then the following Radial Basic Kernel function is obtained:

$$k(x,y) = \sum_{i=1}^{m} d_i \, exp(-\sum_{k=(i-1)L+1}^{iL} (x_k - y_k)^2) \quad (7)$$

Both spatial pyramid and histogram intersection are applied to optimize Radial Basic Kernel function parameters and obtain the final annotation results.

IV. EXPERIMENT SETTING AND RESULT EVALUATION

The experiments have been done on the dataset Caltech-256[14], corel5k and Stanford 40 actions [15]. We set SIFT patch size 16, and let grid spacing is 8. The training sets dictionary size is 300.

To evaluation the quality of an annotation image framework in a set of test images, we use many performance measures which are commonly used for image retrieval. All aforementioned are briefed as follow:

A. Precision

For a given query, its precision of the first n results of the ranking list is defined as:

$$P = \frac{N_{rel_n}}{n} \quad (8)$$

Where N_{rel_n} is the number of relevant images in first n results.

B. Mean Average Precision

Mean Average Precision (MAP) is obtained as the mean of average precisions over a set of queries. Given a query, its MAP is computed by Eq. (9), where N_{rel} is the number of relevant images. N is the number of total retrieved images, $rel(n)$ is a binary function indicating whether the n th image is relevant, and P (n) is the precision at n th image.

$$AP = \frac{1}{N_{rel}} \sum_{n=1}^{N} P(n) \times rel(n) \quad (9)$$

V. RESULTS AND DISCUSSION

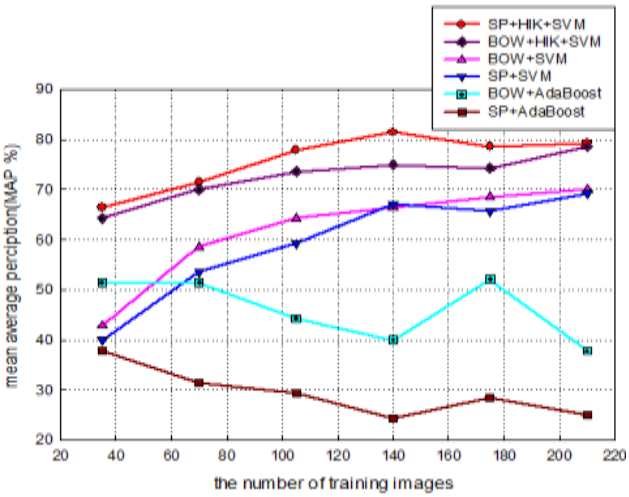


Figure 3. Comparing the performance of six image annotation methods by MAP

From the illustration, it can been seen that the Adaptive Boosting (AdaBoost) without SVM is not good for the image annotation. The precision is less 60% with the SP or BOW, and the robustness is bad. It appears unstable accuracy. Without support machine learning (SVM), the result can not satisfy us. The rest of four methods is relatively high precision and stable by using Radial Basis Kernel function. The precision improves gradually along with the increasing number of the training images. It suggests that Radial Basis Kernel function can make a distinction the image feature from the training images. There is obvious improvement in precision either combining BOW with SVM or combining SP with SVM. The spatial information affects seldom on their cooperation. When we add the histogram intersection kernel, the precision increase largely on both the above combination. It shows HIK generates better codebooks and thus improves recognition accuracy. The accuracy can increase 10% approximately in the table. Comparing these methods, we can get a conclusion that the SP+HIK+SVM is relatively better than others. It is also proved that multiple kernels model performs better than sole kernel models in the experiment.

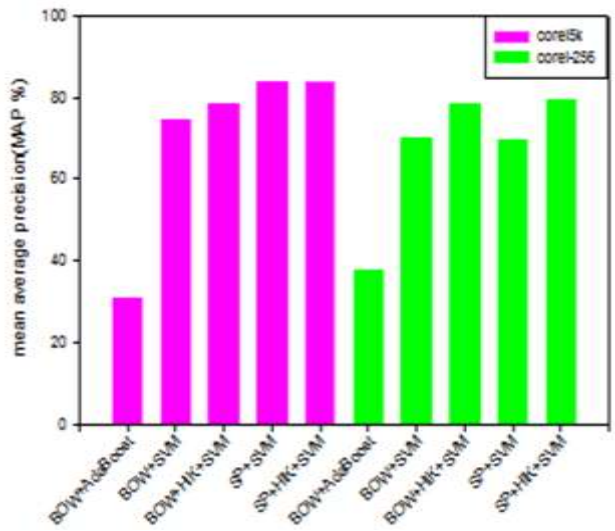


Figure 4 .Experiments on the datasets Corel5k and Corel-256

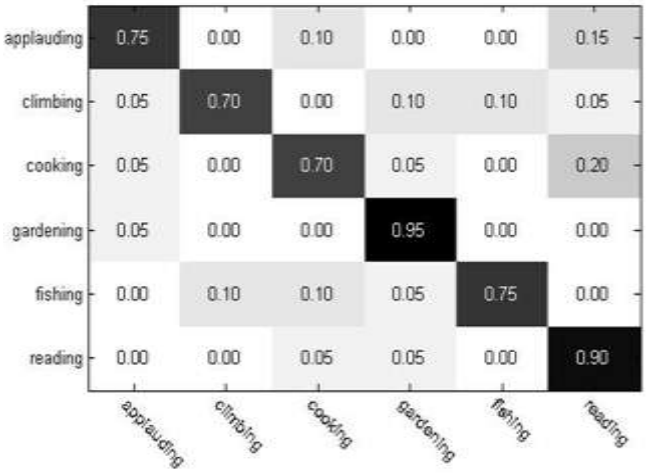


Figure 5. Confusion matrix on the Stanford40 action

It can be seen from Fig. 4 and Fig. 5, the detectors learnt was used to detect the candidate regions via MKL from all the test images, and KNN method was employed for baseline as dictionary.

As is shown in the Fig. 4, the obtained experiment results are approximately the same on different datasets by use of both corel5k and corel-256, The proposed framework is relative better than that of other methods. The precision remains at around 80%. However, the method using BOW and AdaBoost is only around 30%. The precision is more than 65% in the methods on basis of SVM. At the same time, the experiment results show that if only combining SP with SVM, the precision may be not stable, on the contrary, MKL is more robust than the previous.

Fig. 5 is a confusion matrix on the Stanford40 action. The Stanford40 action dataset contains a variety of human action such as: applauding, blowing, and cooking, etc. It is relative difficult to learn and recognize. We dealt with 300 training images and 120 testing images in the experiments. Then MKL method is applied to the datasets and still get relatively good performance. The classification precision is above 70%. The highest precision reaches up to 95%. As the results, human actions and annotate semantic labels can be recognized.

The experiments demonstrated the MKL is benefit to image classification and annotation. It not only optimizes the feature parameters, but also increases the robustness and accuracy in the classification and semantic annotation.

VI. CONCLUSIONS

In this paper, a framework for image semantic annotation by MKL is described. Firstly, the features are extracted by SIFT to generate descriptor, then these feature descriptors are clustered and a dictionary of training sets is built, and each clustering center is a visual word. Then the features are organized via Spatial Pyramid. Thereafter MKL is applied to learn an optimal combination of histogram intersection kernels. Finally, the testing image labels are predicted by Radial Basic Kernel function, which is one of the most widely used kernel functions. The kernel function has a relative better performance in the image classification and semantic annotation. During the process, the parameters are

optimized by spatial pyramid and histogram intersection kernel, aiming at machine learning (SVM). SVM is no doubt the most popular classification technique as the learning model for image annotation. Each semantic concept is treated as a class and the SVM is used to estimate the class distribution. It has been verified by experiments that our model is more accurate than the previous algorithm of the same kind.

Image retrieval emphasizes real-time characteristic, and feature exacting is the most time-consuming in total process, so we try our best to improve the efficiency of feature processing and maintain high precision in image semantic annotation.

ACKNOWLEDGMENT

This work is supported by the Sate Scholarship fund (201308220163); Jilin Province Nature Science Foundation of China (20101523); International Cooperative Research Project of the Ministry of Education (Z2011138); Jilin Province Science and Technology supporting Plan of China (20100368); "Twelfth Five Year plan "science and technology research project of the Education Department of Jilin Province.

REFERENCES

[1] Y. Mori, H. Takahashi, and R. Oka. “Image-to-word transformation based on dividing and vector quantizing images with words”. In MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management,[J], 1999.

[2] K. Barnard and D. Forsyth. “Learning the semantics of words and pictures”. In International Conference on Computer Vision, [J],Vol.2, pages 408-415, 2001.

[3] Nikhil Rasiwasia, Student Member, IEEE, Pedro J. Moreno, Member, IEEE, and Nuno Vasconcelos, Member, IEEE. “Bridging the Gap: Query by Semantic Example”. IEEE TRANSACTIONS ON MULTIMEDIA, [J], VOL. 9, NO. 5, AUGUST 2007.

[4] Donn Morrison, Stéphane Marchand-Maillet, Eric Bruno, and Eric Bruno. “Automatic image annotation with relevance feedback and latent semantic analysis”. Adaptive Multimedia Retrieval: Retrieval, User, and Semantics Lecture Notes in Computer Science Volume 4918,[C] ,2008, pp 71-84

[5] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. “Automatic Image Annotation and Retrieval using Cross-Media Relevance Models”. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, [J], page 119--126. (2003)

[6] Vassilios Stathopoulos and Joemon M. Jose. “Bayesian Mixture Hierarchies for Automatic Image Annotation”. 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings.[C], pp 138-149.2009.

[7] Yinjie Lei, Wilson Wong, Wei Liu, Mohammed Bennamoun. “An HMM-SVM-based automatic image annotation approach”. ACCV'10 Proceedings of the 10th Asian conference on Computer vision - Volume Part IV, [J], Pages 115-126.(2010)

[8] Chong Wang, David Blei, Li Fei-Fei. “Simultaneous Image Classification and Annotation. Computer Vision and Pattern Recognition”. CVPR 2009. IEEE Conference on 20-25 June 2009, [J],pp 1903 - 1910 ,2009

[9] David G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. International Journal of Computer Vision Volume 60, Issue 2 , [J], pp 91-110.2004

[10] Otávio A.B. Penattia, Eduardo Vallea, b. Ricardo da S. Torresa, “Comparative study of global color and texture descriptors for web image retrieval”, Journal of Visual Communication and Image Representation [J].2012

[11] Anil Balaji Gonde, R.P. Maheshwari, R. Balasubramanian. “SIFT Feature with Relevance Feedback for Image Retrieval”. International Journal of Computing Science and Communication Technologies, [J], VOL. 3, NO. 2, Jan. 2011.

[12] LI. Fei-fei, FERGUS R, PERONA P. “Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories”, In IEEE Conference on Computer Vision and Pattern Recognition, [J], 2004.

[13] K. Grauman and T. Darrell. “Pyramid match kernels: Discriminative classification with sets of image features”. In Proc. ICCV, [J], 2005

[14] Griffin, G. Holub, AD. Perona, P. The Caltech 256. Caltech Technical Report.

[15] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. “Human Action Recognition by Learning Bases of Action Attributes and Parts”. Internation Conference on Computer Vision (ICCV), [J],Barcelona, Spain. November 6-13, 2011.