# IMAGE ANNOTATION BASED ON BAG OF VISUAL WORDS AND OPTIMIZED SEMI-SUPERVISED LEARNING METHOD

**Jun Li[1], Hongmei Zhang[2] and Yuanjiang Liao[3]**
*School of Electron and Information Engineering, Ningbo University of Technology, China*
E-mail: [1]vbli@163.com, [2]zhm@nbut.cn, [3]lllin80@163.com

*Abstract*

***This paper proposes a new approach to annotate image. First, in order to precisely model training data, shape context features of each image is represented as a bag of visual words. Then, we specifically design a novel optimized graph-based semi-supervised learning for image annotation, in which we maximize the average weighed distance between the different semantic objects, and minimize the average weighed distance between the same semantic objects. Training data insufficiency and lack of generalization of learning method can be resolved through OGSSL with significantly improved image semantic annotation performance. This approach is compared with several other approaches. The experimental results show that this approach performs more effectively and accurately.***

*Keywords:*

*Image Retrieval, Image Semantic Annotation, Bag of Words (BoW), Semi-Supervised Learning*

## 1. INTRODUCTION

Due to the recent progress in the decreasing storage costs and the growing availability of broadband data connection, digital images are becoming widely used. However, the increasing availability of digital images has not been accompanied by an increase in its accessibility. This is due to the nature of image data, which is unsuitable for traditional forms of data access, indexing, search, and retrieval. If we want to study or find some interested images, at present we have to manual searching them. This is an extremely time consuming, tedious and labor-intensive process. Therefore, the demand of new technologies and tools for effective and efficient retrieval of image data has been exacerbated by recent trends.

Conventional methods to searching an image/video based on the low level features. However, the main drawback of the low level features based methods is it often fails to meet the user's information need [1], [2]. In fact, users likely use semantic-level to depict their information needs, such as “people”, “bikes”, “cars”, “rivers”, “buildings”, “stars”, “computers” and different types of events (“talking”, “running”, etc.). Therefore, the technologies and tools for effective and efficient retrieval of image data are based on semantic-level [3][4]. By image automatic semantic annotation, a image retrieval problem is turned into a text retrieval task, which can be effectively resolved by taking advantages of mature text indexing and retrieval techniques.

To image annotation, many efforts have been devoted into this area for the past years, and many different methods have been proposed. Thanks to the good function learning and generalization capability, there have been some efforts turning to the tools of graph-based semi-supervised learning recently. For example, in [6]-[8] image semantic annotation by graph-based semi-supervised learning. However, the graph-based semi-supervised image annotation has a major drawback, which is the training data insufficiency and lack of generalization always happens. In order to address this drawback, this paper proposes an algorithm for image annotation. We choose the shape context features of each image as the feature vector in image annotation system. In order to solve the problem of training data insufficiency and lack of generalization of learning method, we specifically design a novel optimized graph-based semi-supervised learning(OGSSL) for image annotation, in which we maximize the average weighed distance between the different semantic samples and the negative samples, and minimize the average weighed distance between the same semantic samples. This approach is compared with several other approaches. The testing result of the experiment shows that the method has good accuracy for image annotation.

## 2. IMAGE ANNOTATION BASED ON BAG OF VISUAL WORDS AND OPTIMIZED SEMI-SUPERVISED LEARNING METHOD

In this section, we first introduce the generation of BoW, which is used to represent the object. And then, we describe the algorithm of the proposed OGSSL technique, which can significantly improved image semantic annotation performance.

### 2.1 BOW FOR IMAGE REPRESENTATION

The basic idea of bag-of-visual-words is to depict each image as an orderless collection of image features. Each feature of the image can be as a visual word. Through mapping the feature in an image to the visual word, the image can be described as a feature vector according to the presence or count of each visual word [9]-[11].

The tradition way to choose the visual vocabulary is base on the SIFT Keypoint. While keypoint is extracted from the grey level images and do not contain global features and ignorance of spatial information (distance and orientation feature). Shape is a very important feature in computer vision. Humans can easily and accurately identify the object based on the shape feature, even in very complex environment or object deformation [12], [13]. This paper intends to choose the visual vocabulary base on shape context features to implement the image annotaion.

The shape context was introduced by Belong et al [12]. It describes an object shape by a discrete set of sample points from the object contours. These discrete sample points are not required to have special meaning, only require to be uniform and consistent distributed in the outline of the object contours. Given a sample points set $P = \{p_1, p_2, \ldots, p_n\}$ on a shape. The basic definition of shape context at point $p_i$ is a local histogram of edge points in a radius-angle-direction polar grid. Suppose there are $n_r$ (radial) by $n_\theta$ (angular) bins and the edge map $E$ is divided into $E_1, \ldots, E_{n_e}$ by $n_c$ orientations, for a point at $p_i$, its *SC* is defined as $h_i$, where,

$$h_i(k) = \#\{q \neq p_i, (q\text{-}p_i) \in bin(k)\}, k = 1,2,\ldots,n_r n_\theta n_c \quad (1)$$

Due to the data of shape context is huge and contains a lot of noise and redundancy, which is difficult to be stored and applied directly. Hence, in order to improve the efficiency of image annotaion, visual words are further generated by performing k-means cluster shape context features. Through features clustering, each visual word usually corresponds to a cluster of shape context features. Different visual word represents different position of the object. All of the visual words form a visual vocabulary.

## 2.2 THE FRAMEWORK OF OGSSL MODEL

For the OGSSL model, we first introduce the construction of the proposed OGSSL technique then show it can effectively resolve training data insufficiency and lack of generalization for the conventional graph semi-supervised learning.

### *2.2.1 OGSSL Construction:*

Assume we have a finite set of labeled training sample $T_1 = \{(x_1, y_1),\ldots, (x_l, y_l)\} \in X_1 \times Y_1$ and a finite set of unlabeled training sample $T_u = \{X_{l+1},\ldots,X_{l+u}\} \in X_u$, where $X$ is represented by a d-dimensional feature vector, i.e., $X_i \in R_d$. Labeled training sample $T_l$ is partitioned into labeled positive samples $X_i^+\left(1 \le i \le l^+\right)$ and labeled negative samples $X_i^-\left(l^+ < i \le l^+ + l^-\right)$, $y_i$ is the class label of $x_i$, e.g., $y_i \in \{-1, 0, +1\}$. We assume that class labeled 1 corresponds to the positive samples, class labeled -1 to the negative samples, and class labeled 0 to unlabeled training sample.

Shape context features are extracted from the images to represent image semantic. The shape context features that are the same semantic objects are considered as relevant to each other, and will be used as the similar pairwise in our learning task; and any two shape context features that are different semantic objects are considered as irrelevant, and will be treated as the dissimilar pairwise in our learning task. We maximize the average weighed distance between the different semantic objects, and minimize the average weighed distance between the same semantic objects. In this paper, $W_{i,j}$ denote the average weighed distance between two samples. We maximize the average weighed distance between two positive samples as Eq.(2), and minimize the average weighed distance between two negative samples as Eq.(3)

$$\min \sum_{i=1}^{l^+} \sum_{j=1}^{l^+} W_{i,j} \quad (2)$$

$$\max \sum_{i=1}^{l^+} \sum_{j=1}^{l^+} W_{i,j} \quad (3)$$

$$W = \arg\min_{W_{i,j}} \sum_{i=1}^{l} \sum_{j=1}^{l} h_{i,j} W_{i,j} \quad (4)$$

The Eq.(2) and Eq.(3) can be put together to generate the function Eq.(4) in which the variable $h_{i,j}$ indicates whether the data $x_i$ and data $x_j$ are in the same semantic class. If both $x_i$ and $x_j$ are on the same semantic class, then $y_i$ =1. If $x_1$ and $x_j$ are of different semantic class, then $y_i$ = -1. Otherwise $h_{i,j}$ = 0. Eq.(4) can be written as the following:

$$h_{ij} = \begin{cases} 1, & \textit{if } 1 \le i, j \le l^+ \\ -1, & \textit{if } 1 \le i \le l^+; l^+ < j \le l^- \\ 0, & \textit{otherwise} \end{cases} \quad (5)$$

According to [5], the graph semi-supervised learning function is defined as following:

$$f^* = \arg\min_{f(x_i)} \left\{ \sum_{i,j=1}^{n} W_{ij} \left( \frac{f(x_i)}{\sqrt{d_i}} - \frac{f(x_j)}{\sqrt{d_j}} \right)^2 + \gamma \sum_{i=1}^{1} L\left(f(x_i), y_i\right) \right\} \quad (6)$$

The first term of the above function is a smoothness regularize. The second term of the above function is the loss function.

The basic principle of our learning task is that distances between visual feature vectors of the same semantics objects should be minimized, and meanwhile distances between visual feature vectors of different semantics objects should be maximized. To this end, we formulate our distance metric learning problem into the graph semi-supervised learning, i.e. Eq.(6), so that a better classifier can be designed for classifying test data, which can be written as the following:

$$F^* = \arg\min{}_{f(x_i) w_{i,j}} \left\{ u \sum_{i=1}^{l} \left| f(x_i) - y_i \right|^2 + \sum_{i=1}^{1} \sum_{j=1}^{1} h_{i,j} w_{i,j} + \sum_{i,j}^{n} w_{i,j} \left| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right|^2 \right\} \quad (7)$$

The Eq.(7) is the proposed OGSSL, where $D$ is diagonal weight matrix, whose entries are defined as $d_i = \sum_{j=1}^{n} w_{ij}$, $u \sum_{i=1}^{1} \left| f(x_i) - y_i \right|^2$ is the loss function, $\sum_{i=1}^{1} \sum_{j=1}^{1} h_{i,j} w_{i,j}$ is the average weighed distance between two samples and $\sum_{i,j}^{n} w_{i,j} \left| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right|^2$ is a smoothness regularize.

### *2.2.2 The OGSSL Model Analysis:*

The above optimization problem is more difficult to solve than traditional semi-supervised, for the adjacency matrix $W$ is often obtained via Eq.(8) in traditional semi-supervised, but in order to have a better classifier, we have to find an optimal $W$ in this paper.

$$W_{i,j} = \begin{cases} \exp\left( -\frac{\left\| x_i - x_j \right\|}{\sigma} \right) x & x_i \textit{ and } x_j \textit{ are neighbors} \\ 0 & \textit{others} \end{cases} \quad (8)$$

We use the learning method to optimize *W*. Let a set of option *W* training sample

$$D=\left\{\left(X_{1,2},W'_{1,2}\right),\left(X_{1,3},W'_{1,3}\right),\ldots,\left(X_{l^+,l^-},W'_{l^+,l^-}\right)\right\} \qquad (9)$$

where, $X_{i,j}\subset R^m$ is input data, which is visual features in this paper. $W'_{i,j}\in R$ is output data. The goal is to find a function $W(x)$ that can obtain targets $W_{i,j}$ for all the training data. To this end, we can start from linear decision functions. The linear decision function can be represented as,

$$W''(X_{i,j}) = \langle \omega.X_{i,j}\rangle\text{-}b \qquad (10)$$

where, $\langle \omega.\ X_{i,j}\rangle$ denotes the dot product in the space of the input patterns. Flatness in Eq.(10) means small $\omega$. For this, it is required to minimize $\|\omega\|^2$. Formally this can be written as a convex optimization problem by requiring,

$$\begin{aligned}&\min_{w,b}=\frac{1}{2}\|\omega\|^2\\&s.t.\left(\left(\omega.X_{i,j}\right)+b\right)-W'_{i,j}\le\varepsilon,i,j=1,\ldots,l\\&W'_{i,j}-\left(\left(\omega.X_{i,j}\right)+b\right)\le\varepsilon,i,j=1,\ldots,l\end{aligned} \qquad (11)$$

The standard dualization method utilizing Lagrange multipliers has been described as follows:

$$\begin{aligned}L\left(\omega,b,a^{(*)}\right)=\frac{1}{2}\|\omega\|^2-&\\ \sum_{i=1}^{1}\sum_{j=1}^{1}\alpha_{i,j}\left(\varepsilon+W'_{i,j}-\left(\omega.X_{i,j}\right)-b\right)-&\\ \sum_{i=1}^{1}\sum_{j=1}^{1}\alpha^*_{i,j}\left(\varepsilon-W'_{i,j}+\left(\omega.X_{i,j}\right)+b\right)&\end{aligned} \qquad (12)$$

In which $\alpha^{(*)}=\left(\alpha_{1,1},\alpha^*_{1,1},\alpha_{1,2},\alpha^*_{1,2}\ldots,\alpha_{1,1},\alpha^*_{1,1}\right)^T\in R^{2n}_{+}$ is trained coefficients.

The dual variables in Eq.(12) follows from saddle point condition that the partial derivatives of *L* with respect to the primal variables ($\omega$, *b*) have to vanish for optimality.

$$\begin{aligned}\nabla_b L\left(\omega,b,a^{(*)}\right)&=\sum_{i=1}^{1}\sum_{j=1}^{1}\left(\alpha_{i,j}-\alpha^*_{i,j}\right)=0\\&=\omega-\sum_{i=1}^{1}\sum_{j=1}^{1}\left(\alpha_{i,j}-\alpha^*_{i,j}\right)X_{i,j}\end{aligned} \qquad (13)$$

Substituting Eq.(13) into Eq.(12) yields the dual optimization problem.

$$\begin{aligned}&\min_{a,a^{(*)}}=\frac{1}{2}\sum_{i=1}^{1}\sum_{j=1}^{1}\sum_{m=1}^{1}\sum_{n=1}^{1}\left(a^*_{i,j}-a_{i,j}\right)\left(a^*_{m,n}-a_{m,n}\right)\left(X_{i,j}.X_{m,n}\right)\\&\quad-\varepsilon\sum_{i,j=1}^{1}\left(a^*_{i,j}+a_{i,j}\right)\\&\quad s.t.\sum_{i=1,j}^{1}\left(a_{i,j}-a^*_{i,j}\right)=0,\\&a^{(*)}_{i,j}\ge 0,i,\ j=1,\ldots,l\end{aligned} \qquad (14)$$

Solving above dual optimization problem, we can obtain the following regression function.

$$W''\left(X_{i,j}\right)=\sum_{i=1}^{1}\sum_{j=1}^{1}\left(\alpha^*_{i,j}-\alpha_{i,j}\right)\left(X.X_{i,j}\right)+b \qquad (15)$$

The average weighted distance between two image, $W^*_{ij}$, can be calculated via Eq.(15). Let *D* be a diagonal matrix with

$$D_{ii}=\sum\nolimits_{j=1}^{n}W^*_{ij} \qquad (16)$$

The laplacian matrix of graph can be defined,

$$L=D-W^* \qquad (17)$$

We formulate our distance metric learning problem into Eq.(7)

$$\begin{aligned}F^*=\arg\min_{f(x_i)w_{i,j}}&\left\{u\sum_{i=1}^{1}\left|f\left(x_i\right)-y_i\right|^2\right.\\&+\sum_{i=1}^{1}\sum_{j=1}^{1}h_{i,j}W^*_{i,j}+\sum_{i,j}^{n}W^*_{i,j}\left|\frac{f_i}{\sqrt{d_i}}-\frac{f_j}{\sqrt{d_j}}\right|^2\end{aligned} \qquad (18)$$

Finally, the Eq.(18) can be solved by iteration algorithm.

## 3. EXPERIMENTAL RESULTS

The performance of the proposed algorithm is evaluated on PASCAL data sets and some video images. They include ‘cat’, ‘cow’, ‘bus’, ‘dog’, ’motorbike’, ‘horse’, ‘plain’, ‘river’. The performance of image annotation algorithm is usually measured with precision. The precision are defined as following:

$$precision=\frac{N_c}{N_c+N_f}\times 100\% \qquad (19)$$

where, $N_c$ is the number of correct semantic annotation, $N_f$ is the number of false semantic annotation. A good semantic annotation should have high precision.

In order to evaluate the performance of the proposed method, we first should decide some parameter in the experiment. For the shape context feature parameter, we define $n_r$ is [0, 12, 24], $n_\theta$ is 12, $n_c$ is 4. Thus each point of the shape context feature is described by 96 local histograms. A visual vocabulary is generated by clustering the detected shape context features and treating each cluster as a unique visual word of the vocabulary. The size of visual vocabulary is determined by the number of the detected shape context clusters. In this paper, we defined the vocabularies of 1,000 visual words.

Y.G. Jiang has proposed a method of towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval [10], which is the more successful one in lots of algorithms. In the experiments, we compare our method with Y. Kawai method and the Graph-based semi-supervised learning for semantic annotation (GSSLSA). Fig.1 shows the performance of the proposed algorithm compared with Y. G. Jiang algorithm and GSSLSA. From the experimental results, we can see that the performance of our method is better than traditional algorithms.

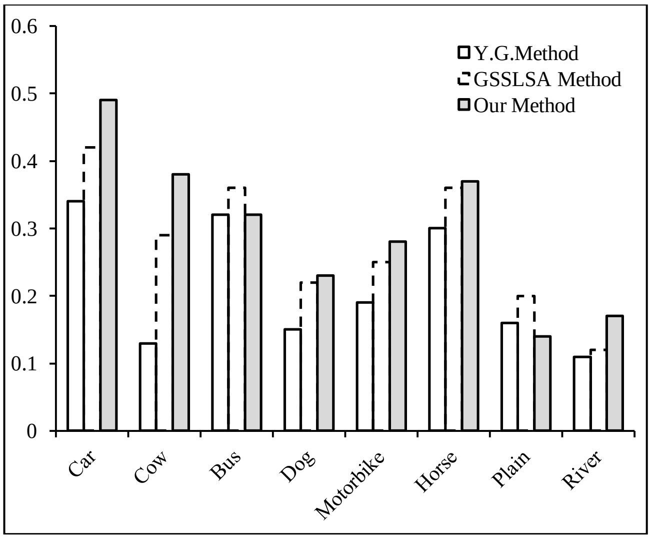

Fig.1. Experiment Comparison with Traditional Algorithms

## 4. CONCLUSION

This paper proposes a new approach to annotate image. First, shape context feature of each image is represented as a bag of visual words. Then, we specifically design a novel optimize graph-based semi-supervised learning for image annotation, in which we maximize the average weighed distance between the different semantic objects, and minimize the average weighed distance between the same semantic objects. Compared with traditional graph-based semi-supervised for image annotation, this approach can effectively resolve training data insufficiency and lack of generalization.

## ACKNOWLEDGEMENT

This work is supported by scientific research projects of Zhejiang Education Department (No. Y201224456).

## REFERENCES

[1] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber and M. Worring, "Adding Semantics to Detectors for Video Retrieval", *IEEE Transactions on Multimedia*, Vol. 9, No. 5, pp. 975-986, 2007.

[2] M. R. Naphade, S. Basu, J. R. Smith, Ching-Yung Lin and B. Tseng, "Statistical modeling approach to content-based video retrieval", *IEEE Proceedings of 16th International Conference on Pattern Recognition*, Vol. 2, pp. 953-956, 2002.

[3] J. Tang, S. Yan, C. Zhao, T.S. Chua and Ramesh Jain, "Label-specific training set construction from web resource for image annotation", *Signal Processing*, Vol. 93, No. 8, pp. 2199-2204, 2013.

[4] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram and Di Zhong, "A fully automated content-based video search engine supporting spatio-temporal queries", *IEEE Transactions on Circuit and Systems for Video Technology*, Vol. 8, No. 5, pp. 602-615, 1998.

[5] C. H. Hu, "Graph-based Semi-supervised Machine Learning", Zhejiang University, 2008.

[6] M. Guillaumin, J. Verbeek and C. Schmid, "Multimodal semi-supervised learning for image classification", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 902-909, 2010.

[7] A. Subramanya and P. P. Talukdar, "Graph-based semi-supervised learning algorithms for NLP", *Tutorial Abstracts of ACL 2012, Association for Computational Linguistics*, pp. 6-6, 2012.

[8] L. Wu, S. C. H. Hoi and N. Yu, "Semantics-preserving bag-of-words models and applications", *IEEE Transactions on Image Processing*, Vol. 19, No .7, pp. 1908-1920, 2010.

[9] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 524-531, 2005.

[10] Y. G. Jiang, C. W. Ngo and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval", *Proceedings of the 6th ACM international conference on Image and Video Retrieval*, pp. 494-501, 2007.

[11] Y. G. Jiang, C. W. Ngo and J. Yang, "Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval", *Proceedings of the 16th ACM International Conference on Image and Video Retrieval*, pp. 495-501. 2007.

[12] S. Belongie, J. Malik and Jan Puzicha, "Shape Matching and Object Recognition Using Shape Contexts", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 24, pp. 509-522, 2002.

[13] G. Mori, S. Belongi and J. Malik, "Efficient Shape Matching Using Shape Contexts", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 11, pp. 1832-1837, 2005.