Chapter 36
Image Annotation with Nearest Neighbor Based on Semantic Information

Wei Wu and Guanglai Gao

Abstract Most of the Nearest Neighbor (NN)-based image annotation methods do not achieve desired performances. The main reason is that much valuable information is lost when extracting visual features from image. In this paper, we propose a novel weighted NN-based method. Instead of using Euclidean distance, we learn a new distance metric with image semantic information to calculate the distance between the two images. Meanwhile, we utilize textual information of each image tagged by users to form weights of NN-based model. When introducing the semantic information, our method can minimize the semantic gap for intraclass variations and interclass similarities, and improve the annotation performance. Experiments on image annotation dataset of ImageCLEF2012 show that our method outperforms the traditional classifiers. Moreover, our method is simple, efficient, and competitive compared with state-of-the-art learning-based models.

Keywords Image annotation · Nearest neighbor · Distance metric learning · Entropy weight

36.1 Introduction

Image annotation and retrieval have drawn considerable attention in both research and practical areas. The goal of image annotation is to automatically recognize visual concepts from image semantic concepts set, and turns out to be extremely challenging due to the large intraclass variations and interclass similarities. Recently, there have been many research communities engaged in this work, such as ImageCLEF [1], TRECVID, Pascal VOC, etc., which confirm the challenges in this field.

W. Wu (✉) · G. Gao
Department of Computer Science, Inner Mongolia University,
No.235 West College Road, Hohhot, China
e-mail: cswuwei@imu.edu.cn

Z. Deng and H. Li (eds.), *Proceedings of the 2015 Chinese Intelligent Automation Conference*, Lecture Notes in Electrical Engineering 336,
DOI 10.1007/978-3-662-46469-4_36

The image annotation methods often use learning-based classifiers, and rarely use Nearest Neighbor (NN)-based classifiers, because they provide inferior performance relative to learning-based methods. But we may underestimate the effectiveness of NN-based method. Boiman et al. [2] claim that the main reason resulting in the low performance of NN-based algorithms is the information loss when extracting image visual features, particularly when constructing bag of visual words (BoVW)-based features. BoVW-based features are harmful in the case of NN-based model, which has no training phase to compensate for this loss of information. The method proposed by [2] does not use BoVW model, but directly uses local features for NN-based classifier, and achieves better performance than learning-based models.

In this paper we propose a novel NN-based method which can greatly reduce the semantic information loss, thereby improving the performance of large scale image annotation. We still use BoVW features, but introduce the image semantic information for computing distance between images. In our model, we first utilize image semantic information for distance metric learning (DML) [3, 4], and get a new distance measure. Then we generate multiple clusters for each image category using *k*-means algorithm based on this new distance. Each cluster contains a set of images and some user textual tags (We also regard them as image semantic information). We then assign a weight to each cluster according to the importance of these textual tags, and construct a semantic weighted NN-based classifier.

There are some existing works related to NN-based model. Blitzer et al. [5] learn a Mahanalobis distance metric for the traditional *k*NN model. Wang et al. [6] propose image-to-class-based NN model. Wang et al. [7] introduce the semantic relations based on WordNet for distance metric learning. Verma and Jawahar [8] present a two-step variant of the classical *k*NN algorithm. Our method is different from all the above-mentioned methods, we use a different distance metric optimization strategy, and meanwhile, we introduce the user tag information for calculating weights, and propose a novel weighted NN-based framework. Experiments on dataset of ImageCLEF2012 [1] image annotation task confirm the effectiveness of our method, the result of our model outperforms the traditional classifiers and a new baseline of NN model [9], and is competitive compared with state-of-the-art models.

The paper is organized as follows: Sect. 36.2 describes the DML using semantic information, and Sect. 36.3 introduces our NN-based model. Section 36.4 describes the experiments and results. Finally, we conclude our work and shed light on the future work in Sect. 36.5.

36.2 Distance Metric Learning

The objective of DML is to find an optimal Mahalanobis metric A from training data. In our method, we extract the pairwise constraints from training images for distance metric learning. We formalize the representation of the pairwise features constraints set as $\{(f_{i1}, f_{i2}, y_i)\}_{i=1}^{N}$, where f_{i1} and f_{i2} are two image features. And if both f_{i1} and f_{i2} belong to the same image category, then $y_i = 1$, otherwise $y_i = -1$. It is worth noting that how to select pairwise constraints can greatly affect the annotation performance. For the image semantic annotation task, there are the large intraclass variations and interclass similarities, so we comply with such selection criterion: one is that the features are of the same image category but with large variation, the other is that the features are of different image category but with large similarity.

Specifically, we firstly extract features of all the training images and use the *k*-means algorithm in Euclidean distance space to cluster the image features for each image category, with the result that *k* centers are formed for each image category. Then we regard these centers as visually different "images" in the same semantic category (namely, the images with a large intraclass variation), and for each pair of these images, we construct pairwise constraints $(f_{i1}, f_{i2}, y_i = 1)$. Last, for each center of an image category, we search for the closest image in Euclidean distance in any other image category (namely, the images with a high interclass similarity), and construct pairwise constraints $(f_{i1}, f_{i2}, y_i = -1)$.

Given the feature pairwise constraints information, the goal of our task is to learn a distance metric *A* to effectively measure distance between any two image features f_{i1} and f_{i2}, which can be represented as formula (36.1):

$$d(f_{i1}, f_{i2}) = \sqrt{(f_{i1} - f_{i2})^T A (f_{i1} - f_{i2})} \tag{36.1}$$

To find an optimal metric *A*, the distances between visual features of the same semantic category should be minimized, and meanwhile distances between features of different semantic category should be maximized. Based on this principle, we formulate this distance metric learning problem into the following optimization:

$$\begin{aligned} \min_{A,b}\ f(A,b) = \sum_i y_i(\|f_{i1} - f_{i2}\|_A^2 - b) + \frac{\lambda}{2}\operatorname{tr}(A^T A) \\ s.t. \quad y_i(\|f_{i1} - f_{i2}\|_A - b) \le 1 \\ A \ge 0, \|A\| = 1/\sqrt{\lambda} \end{aligned} \tag{36.2}$$

where $\| \bullet \|_A$ is the Mahalanobis distance between two features under metric *A*. With the first inequality constraints, minimizing this term will make the distance between two semantically identical image features closer. The second term of the objective function is the regularization term, which prevents the overfitting by minimizing this model. The second constraint is introduced to prevent the trivial solution by shrinking metric *A* into a zero matrix. Parameter λ is a constant, *b* is a threshold. We use a stochastic gradient search algorithm to solve this optimization problem [5]. The algorithm is an iterative process, and empirically, this iterative algorithm converges quickly with no more than five iterations.

36.3 Nearest Neighbor-Based Model

We first use k-means clustering method to construct clusters for each training image category. Instead of using Euclidean distance, we use our trained distance metric when running the clustering algorithm, and this is the main difference with [6]. Thus we get k cluster center features for each image category: $f_1, f_2, \ldots, f_k$. Now our work is to search out the image class C which minimizes the sum $\sum_{i=1}^{k} d(f_{\text{test}}, f_i^C)$, where the distance function $d(\cdot)$ is based on the new distance, shown by formula (36.1), f_{test} is the feature of test image, and C denotes the image category, f_i^C denotes the ith cluster feature of image class C, and k takes the same value for all the image class. We also consider that each cluster contains a set of images and some textual terms. So we can utilize this semantic information to assign a weight for each cluster. The major idea is that, the higher the frequency of a term in a cluster is, the more representative this cluster will be. In contrast, if a large number of different terms occur in a cluster, this cluster would be not well representative for related image class [10]. We can calculate entropy according to the terms in a cluster, and this entropy can be viewed as a weight for a cluster. The higher the weight of a cluster is, the greater the distance will be to this cluster. So our NN-based classifier can be changed to this form: minimizing the sum $\sum_{i=1}^{k} w_i^C d(f_{\text{test}}, f_i^C)$, where w_i^C is the entropy weight of the ith cluster for image class C. The entropy can be calculated as follows:

Considering training images for class C are divided into k clusters, $\{C_1, C_2, \ldots, C_k\}$, and there are y unique textual terms $\{t_1, t_2, \ldots, t_y\}$. Assume that a cluster contains several images and each image is assigned several textual terms (It is a truth for dataset of ImaegCLEF [1]). Hence a cluster can be viewed as a collection of terms, $C_i = \cup\, t_j$. The entropy of the ith cluster can be defined as:

$$\text{Entropy}^i = \sum_{t_j \in C_i} \left(\frac{tf_{t_j}^i}{\sum_{t_v \in C_i} tf_{t_v}^i} \log \left(\frac{\sum_{t_v \in C_i} tf_{t_v}^i}{tf_{t_j}^i} \right) \right) \tag{36.3}$$

where $tf_{t_j}^i$ denotes the frequency of term t_j in the ith cluster. Our classifier can therefore be summarized as follows:

1. Constructing k clustering centers for each image category C: $(f_1^C, f_2^C, \ldots, f_k^C)$.
2. Calculating the k entropy weights for each image category C: $(w_1^C, w_2^C, \ldots, w_k^C)$.
3. Computing the visual feature f_{test} of test image.
4. Classification result:

$$\hat{C} = \arg\ \min_C \sum_{i=1}^{k} w_i^C d(f_{\text{test}}, f_i^C) \tag{36.4}$$

When applying to the multilabel image annotation problem, we need only to compute the sum $\sum_{i=1}^{k} w_i^C d(f_{\text{test}}, f_i^C)$ for each image class, and then sort the class labels in ascending order according to these sums.

36.4 Experiments and Results

Experimental images are from the image annotation dataset of ImageCLEF2012 [1]. There are a total of 94 concept categories for annotation. The range of these concepts is fairly wide, including natural elements, environments, people, impression, transportation, etc. There are 15,000 images for training and 10,000 images for testing, within the range of 94 concepts, and each of these images has several user tags provided by organization. We need to allocate each test image with multiple concept labels, and then sort these labels according to the similarities between the image and labels. The evaluation measurement is the MiAP (Mean interpolated Average Precision) which is widely used in the field of image annotation and retrieval.

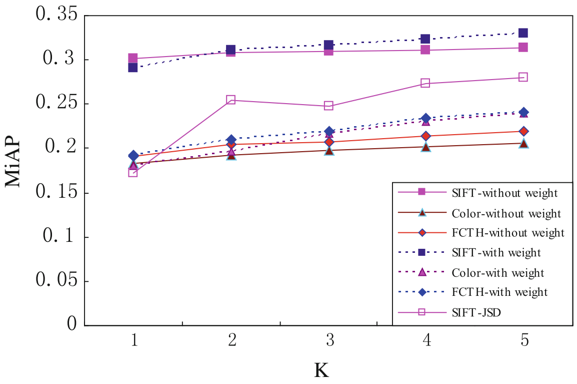
36.4.1 Experiments with Different Features

We select three features for experiments. They are Color Histograms, Fuzzy Color and Texture Histogram (FCTH), and BoVW based on SIFT local features. The size of BoVW is fixed at 500 considering the balance between the annotation performance and computational cost [11].

First, we test the above three features using the traditional *k*-NN classifier. The best result we obtained is using SIFT features, and the value of MiAP is 0.2702 when parameter *k* takes 50. Then we test our method using the same features. The experimental results of ours are plotted in Fig. 36.1. In Fig. 36.1, we use formula (36.1) to calculate distances for all the features except SIFT-JSD. SIFT-JSD denotes JSD distance for SIFT local features. The parameter *k* in Fig. 36.1 is the number of clusters for each image category.

We can see that the SIFT local features get the best result in our method, MiAP reaches 0.3143 without entropy weight, which is higher than traditional *k*-NN

Fig. 36.1 Results of our method (*k* is the parameter of *k*-means algorithm)



method, the MiAP is 0.2702. And MiAP achieves 0.3297 with entropy weight, 1.5 % higher than without entropy weight, which confirms that our weighted strategy is effective.

And we also learn that the curve is relatively flat in Fig. 36.1, which means that the parameter *k* has not much effect on performance. From the point of view of computational cost, the value of *k* of our method is far less than the traditional method. Actually, when the value of *k* is 1, the performance is much better than the traditional *k*-NN. In our experiments, we test only the value of parameter *k* to 5, and it is shared by all the image classes.

Finally, we also learn that the use of semantic distance indeed increases the performance. We can see that the result using SIFT local features with semantic distance is better than JSD distance. This shows that the introduction of semantic distance is effective.

36.4.2 Experiments with Other Methods

We also compare our method with other classifiers, as shown in Fig. 36.2. The methods for comparison are traditional *k*-NN, distance weighted *k*-NN (dw-*k*NN), Naive Bayesian (NB), NBNN method proposed by Boiman et al. [2], Baseline [9], and SVM model. The results obtained by these models use the same SIFT local features as our method, and the kernel function of SVM we used is Histogram Intersection Kernel (HIK). The model of ImageCLEF means the method which achieved the best result published by ImageCLEF2012 [1] using multiple visual features.

In Fig. 36.2, we learn that the performance of *k*-NN is close to NB, and the performance of NBNN is very close to SVM. We can see that the result of our method outperforms all the other methods except the best result by Image CLEF2012. This best result using multiple visual features by ImageCLEF2012 achieved 0.3481, slightly better than ours, that is, our method is competitive.

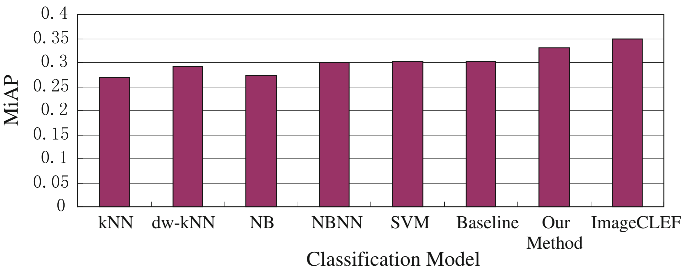


Fig. 36.2 The results comparison with different models

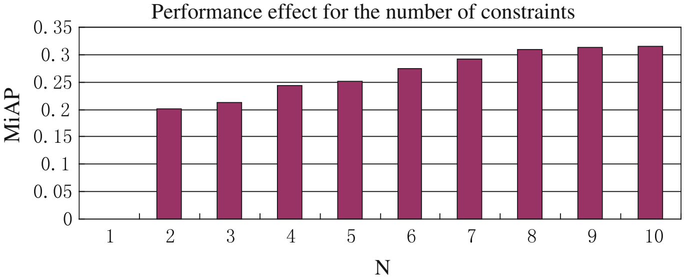


Fig. 36.3 Performance effect of the number of pairwise constraints

36.4.3 Experiments for Distance Metric Learning

Finally, we test the impact on annotation accuracy of the number of pairwise constraints on distance metric learning. We first build *N* visually different "images" using a clustering method (actually, *N* denotes the number of clustering centers) for each category. And we extract a pair of features from each of *N* × (*N*-1)/2 pairwise "images" for each category. Then for each of *N* "images" of each category, we select five different categories of images to construct five pairwise features. Thus for all 94 image categories, we totally have $94 \times (N \times (N - 1)/2 + 5 \times N)$ pairwise feature constraints. We let *N* take values from 2 to 10, with the result that the number of pairwise constraints varies from 1,034 to 8,930. We use SIFT local features to carry out experiment (see Fig. 36.3).

We can see from Fig. 36.3 that the performance gets better with the increase of the number of pairwise constraints. And in our experiment, we find that the trend of improvement increases slowly when *N* value exceeds 8. So when considering the trade-off between the computational cost and performance, we take $N = 10$ for the above experiments. And we use the same number of constraints as other visual features. When efficiency is not cared, we think that if we take greater *N*, we can achieve better performance.

36.5 Conclusion

In this paper we described a novel semantic weighted NN-based classifier based on semantic distance. Our experiments on the ImageCLEF2012 image dataset achieved good results. This confirmed that our method is suitable to large-scale image classification task with high intraclass variations and interclass similarities. In the future, if we can explore more efficient and automatic selecting method of pairwise constraints for DML, it would be more effective for performance.

Acknowledgments This work is supported by Inner Mongolia NSF 2014MS0606.

References

1. Thomee B, Popescu A (2012) Overview of the ImageCLEF 2012 flickr photo annotation and retrieval task. In: CLEF 2012 working notes, Rome, Italy
2. Boiman O, Shechtman E, Irani M (2008) In defense of nearest-neighbor based image classification. In: Proceedings of CVPR, pp 1–8
3. Wang S, Jiang S, Huang Q, Tian Q (2012) Multi-feature metric learning with knowledge transfer among semantics and social tagging. In: Proceedings of CVPR, pp 2240–2247
4. Grauman K, Sha F, Hwang SJ (2011) Learning a tree of metrics with disjoint visual features. In: Advances in neural information processing systems, pp 621–629
5. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
6. Wang Z, Hu Y, Chia LT (2010) Image-to-class distance metric learning for image classification, computer vision–ECCV. Springer, Heidelberg, pp 706–719
7. Wang F, Jiang S, Herranz L et al (2012) Improving image distance metric learning by embedding semantic relations, advances in multimedia information processing–PCM. Springer, Heidelberg, pp 424–434
8. Verma Y, Jawahar CV (2012) Image annotation using metric learning in semantic neighbourhoods, computer vision–ECCV. Springer, Heidelberg, pp 836–849
9. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation, computer vision–ECCV 2008. Springer, Heidelberg, pp 316–329
10. Su JH, Chou CL, Lin CY, Tseng VS (2011) Effective semantic annotation by image-to-concept distribution model. Multimedia IEEE Trans 13(3):530–538
11. Jia Y, Huang C, Darrell T (2012) Beyond spatial pyramids: receptive field learning for pooled image features. In: Proceedings of CVPR, pp 3370–3377