Random Sampling Image to Class Distance for Photo Annotation

Deyuan Zhang, Bingquan Liu, Chengjie Sun, and Xiaolong Wang

ITNLP Lab, School of Computer Science and Technology
Harbin Institute of Technology, Harbin, P.R. China
{dyzhang,liubq,cjsun,wangxl}@insun.hit.edu.cn
http://www.insun.hit.edu.cn

Abstract. Image classification or annotation is proved difficult for the computer algorithms. The Naive-Bayes Nearest Neighbor method is proposed to tackle the problem, and achieved the state of the art results on Caltech-101 and Caltech-256 image databases. Although the method is simple and fast, for the real applications, it suffer from the imbalance of the training datasets. In this paper, we extend the image to class distance which is more general, and use the random sampling technique to alleviate the situation of the imbalance of the training datasets. We perform our method on the ImageCLEF 2010 Photo Annotation task, and the results(INSUNHIT) showing that the algorithm is fast and stable. Although it does not achieving the state of the art performance, more image features can be used to improve the performance and dimension reduction techniques can be adopted to reduce the complexity of space and time.

Keywords: Image Annotation; Nearest Neighbor Classification; Random Sampling; ImageCLEF Photo Annotation Task

1 Introduction

Although human beings recognize the scenes or objects in an image easily, automatic image classification and annotation is a challenging task for computer programs. According to [1], human can recognize about 30000 categories, while discriminating even two categories is difficult for computer vision systems[2].

In recent thirty years researchers have proposed many image descriptors and learning algorithms to recognize objects and scenes. The image descriptors in the literature is roughly classified into five categories: global descriptors[3][11] that represent the image with global attributes, block based descriptors[5] which represent the image using the image blocks, the region based features[6] that is generated by image segmentation algorithms, local patch features that represent images with descriptors find by the interest point or blob operators of the image, and some other features[8] such as the text labels tagged by the internet user on photo sharing communities. Although these image features have been proposed to tackle some specific image recognition tasks, they usually failed to succeed when applying to new image concepts and datasets.

The learning algorithms have been proposed with the development of image descriptors. Various kinds of learning algorithms are proposed or adopted in the literature, including Non-parametric methods[7], kernel machines[9], generative models[4], multiple instance learning algorithms[6], distance metric learning frameworks[12], or biological inspired neural networks[13]. These models have applied to many different tasks, and achieved state of the art results.

Although these image recognition systems succeed in several tasks or benchmark databases, they remain naive for the real world applications. In the real world image databases, the visual concepts shares large intra-class and inter-class variability, and the relation of the concepts is complicated. In addition, the images used for training is imbalanced, resulting in the failure of some learning algorithms. Finally, the databases contain large scale images, and the simplicity and the running time need to be considered when design image categorization systems.

In this paper we describe our system(INSUNHIT) for ImageCLEF2010 Photo Annotation task. We use the dense SIFT[10] image descriptor based on the observation that it performs well both in object categorization and scene classification. We extend Naive-Bayes Nearest Neighbor(NBNN)[7] classification method, and propose Random Sampling Image to Class Distance(RS-ICD) for image classification. RS-ICD can deal with imbalanced dataset while preserving the simplicity of NBNN method. The ImageCLEF2010 challenge results shows that the method is very stable and fast.

2 Overview of the ImageCLEF2010 Photo Annotation Dataset

In this section we briefly discuss some difficulties of the ImageCLEF Photo Annotation dataset. For a detail overview of the dataset please refer to [15], the emphasis in this section is that why this dataset is difficult for computer vision algorithms. The focus of our system is image content, therefore we do not introduce the difficulties of text labels and EXIF information of the photos.

First, compared to the benchmark datasets used in the literature, the visual concepts(annotation keywords) is more abstract and shares more variability. There exist 93 visual concepts, including objects(such as “trees” and “flowers”), scenes(such as “desert”, “sky” and so on), abstract concepts(such as “macro”, “spring” and “summer”). Some of the concepts have visual similar properties—for example, “child” and “baby”—, and the image features can not discriminate them perfectly. Image descriptors is difficult to choose.

Second, the image dataset is very imbalanced. The quantity of training images of each concept ranges from 12 training images(“skateboard”), to 7484 images(“Neutral_Illumination”). This makes the classification system more difficult to train the concepts.

3 Methods

In this section we introduce our learning algorithm in detail. Our algorithm extends the NBNN method, and the RS-ICD is stable for measuring the distance of the image to a query concept.

3.1 Overview of NBNN

The NBNN method defines the distance of image to class to cope with the large intra-class variance of the class. The method is based on the Naive-Bayes and Kernel Density Estimation framework, and the image to class distance is the near optimal distance when training images is very large.

Here we only review the method of NBNN. The NBNN methods operate on the local patch based image features. For a given class $C_j, j \in (1, 2, ..., L)$, and query image Q and its corresponding image descriptors $d_1, d_2, ..., d_n$, the distance of Q and class C_j is defined as follows:

$$d(Q, C_j) = \sum_{1}^{n} NN_{C_j}(d_i) \tag{1}$$

where $NN_{C_j}(d_i)$ is the nearest distance of the descriptor d_i to class C_j:

$$NN_{C_j}(d_i) = min(distance(d_i, d_{C_j})), d_{C_j} \in C_j \tag{2}$$

The classification process is proceed:

$$C_{opt} = argmin_C(d(Q, C_j)), C_j \in C \tag{3}$$

3.2 Image to Class Distance

Although the NBNN is effective for image classification that output a single label when decision, it can not be extend to image annotation(multi label) because the image to class distance of each image is not comparable. In order to deal with this problem, the distance should be normalized:

$$dK(Q, C_j) = \frac{1}{n} \sum_{1}^{n} KNN_{C_j}(d_i) \tag{4}$$

Where the $KNN_{C_j}(d_i)$ is the average distance of the top K nearest neighbor:

$$KNN_{C_j}(d_i) = \frac{1}{K} \sum_{1}^{K} distance(d_i, d_{C_j}) \tag{5}$$

When $K = 1$, both the distance function and the decision function are the same as the NBNN method. Therefore the distance defined by NBNN is a special case of our Image to Class Distance(ICD). The most important is that ICD is comparable between images, and the distance can be used to do multi label classification.

3.3 Random Sampling Generalized NBNN

The Image to Class Distance suffers from the imbalanced training datasets. We use random sampling technique to tackle the problem. Therefore the our algorithm described as follows:

The Random Sampling Image to Class Distance Based Annoation

```
Begin
    For t = 1,2,...T
        Random sampling L images from each class Ci denoted as Ci(t);
    End
    Extract the descriptors of query image Q
    For t = 1,2,...T
        compute dK(Q, Ci(t)) of each class
    End
    the image to class distance dK(Q, Ci) = average(dK(Q, Ci))
    compute the probability p(Q, Ci)=exp(-a*dK(Q, Ci))
End
```

4 Challenge Results

4.1 Experimental Setup

Here we describe the experimental setup of our algorithm. The images are transformed into gray image, and resized to 300 pixels while keep the aspect ratio if the image's length or width are larger than 300 pixels. 128 bin Dense SIFT image features with the step of 8 pixels are extracted. The parameter T of Random Sampling process is set to 10.

We run the algorithm on a computer with Intel Core 2 Duo Q9400 CPU, 4 GB Memory, 32 bit linux operating system. The SIFT extraction matlab program is provided by Lazbnik[16], and the classification algorithm is coded by Python and Numpy toolkit. For Approximate Nearest Neighbor Search, FLANN[17] using randomized KD-Tree with Python interface is used.

To evaluate the performance of the system, Mean Average Precision results is performed as the main evaluation measure.

4.2 Results

In this section we discuss the results of our system(INSUNHIT). All the results is showed on the website[14]. We submit 5 runs: the best MAP result is 23.71%, and the worst result is 22.51%. The detailed setups and the results are showed in Table1.

Our results achieved the centered of the overall results. We use different setup, and achieve similar results. This indicate that our algorithm is very stable.

Table 1. The MAP results of different setups of our algorithms

Runs	Images per random sampling process	KNN	Accuracy
1	25	1	22.86%
2	15	1	23.71%
3	20	1	23.19%
4	50	1	22.89%
5	15	3	22.51%

5 Discussion

Classifying the whole test image dataset takes about 12 hours. Although the performance of our algorithm is far behind the state of art performance, further improvements can be easily obtained. First, the 128-bin SIFT descriptor is too large, while for the image classification, we can try other image descriptors or dimension reduction techniques. Second, multiple image descriptors should be used to improve the annotation results. In recent years, most of the promising results on benchmark image datasets are performed by combining multiple image features or multiple classifiers that computed on these image features.

Acknowledgments. This investigation was supported by the project of the National Natural Science Foundation of China (grants No. 60973076), Special Fund Projects for Harbin Science and Technology Innovation Talents(grants No. 2010RFXXG003) and Microsoft Fund Projects HIT.KLOF.2009021.

References

1. N.K. Logothetis and D.L. Sheinberg: Visual object recognition, Annual Review of Neuroscience, vol. 19, 1996, pp. 577-621.
2. N. Pinto, D.D. Cox, and J.J. DiCarlo: Why is Real-World Visual Object Recognition Hard?, PLoS Computational Biology, vol. 4, Jan. 2008, p. e27.
3. M.J. Swain and D.H. Ballard: Color indexing, Int. J. Comput. Vision, vol. 7, 1991, pp. 11-32.
4. F. Li and P. Perona: A Bayesian Hierarchical Model for Learning Natural Scene Categories, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2, IEEE Computer Society, 2005, pp. 524-531
5. X. Qi and Y. Han: Incorporating multiple SVMs for automatic image annotation, Pattern Recognition, vol. 40, 2007, pp. 728-741.
6. Y. Chen and J.Z. Wang: Image Categorization by Learning and Reasoning with Regions, J. Mach. Learn. Res, vol. 5, 2004, pp. 913-939.
7. O. Boiman, E. Shechtman, and M. Irani: In defense of Nearest-Neighbor based image classification. IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.
8. Q. Yang, X. Chen, and G. Wang: Web 2.0 dictionary, Proceedings of the 2008 international conference on Content-based image and video retrieval, Niagara Falls, Canada: ACM, 2008, pp. 591-600.

9. O. Chapelle, P. Haffner, and V. Vapnik: Support vector machines for histogram-based image classification, Neural Networks, IEEE Transactions on, vol. 10, 1999, pp. 1055-1064.
10. S. Lazebnik, C. Schmid, and J. Ponce: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169-2178.
11. A. Oliva and A. Torralba: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, International Journal of Computer Vision, vol. 42, May. 2001, pp. 145-175.
12. A. Frome, Y. Singer, F. Sha, and J. Malik: Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification, IEEE 11th International Conference on Computer Vision, 2007. pp. 1-8.
13. M. Ranzato, Fu Jie Huang, Y. Boureau, and Yann LeCun: Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition, IEEE Conference on Computer Vision and Pattern Recognition, 2007. pp. 1-8.
14. IMAGECLEF Visual Concept Detection and Annotation Task 2010, `http://www.imageclef.org/2010/PhotoAnnotationMAPResults`
15. Stefanie Nowak and Mark Huiskes: New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010 In the Working Notes of CLEF 2010, Padova, Italy, 2010.
16. Spatial Pyramid Matching Code: `http://www.cs.unc.edu/~lazebnik/research/spatial_pyramid_code.zip`
17. FLANN Source Code: `http://www.cs.ubc.ca/~mariusm/index.php/FLANN/FLANN`