LARGE SCALE NATURAL IMAGE CLASSIFICATION BY SPARSITY EXPLORATION

Changhu Wang[1], Shuicheng Yan[2], Hong-Jiang Zhang[3]

[1]MOE-MS Key Lab of MCC, University of Science and Technology of China
[2]Department of Electrical and Computer Engineering, National University of Singapore
[3]Advanced Technology Center, Microsoft Research, Beijing, China

ABSTRACT

We consider in this paper the problem of large scale natural image classification. As the explosion and popularity of images in the Internet, there are increasing attentions to utilize millions of or even billions of these images for helping image related research. Beyond the opportunities brought by unlimited data, a great challenge is how to design more effective classification methods under these large scale scenarios. Most of existing attempts are based on k-nearest-neighbor method. However, in spite of the optimistic performance in some tasks, this strategy still suffers from that, one single fixed global parameter k is not robust for different object classes from different semantic levels. In this paper, we propose an alternative method, called ℓ^1-nearest-neighbor, based on a sparse representation computed by ℓ^1-minimization. We first treat a testing sample as a sparse linear combination of all training samples, and then consider the related samples as the nearest neighbors of the testing sample. Finally, we classify the testing sample based on the majority of these neighbors' classes. We conduct extensive experiments on a 1.6 million natural image database on different semantic levels defined based on WordNet, which demonstrate that the proposed ℓ^1-nearest-neighbor algorithm outperforms k-nearest-neighbor in two aspects: 1) the robustness of parameter selection for different semantic levels, and 2) the discriminative capability for large scale image classification task.

Index Terms— Image classification, sparsity, ℓ^1-nearest-neighbor, k-nearest-neighbor, WordNet.

1. INTRODUCTION

With the prosperity of the Web, overwhelming amounts of data are now freely available online. On one hand, it is necessary to develop effective index and search techniques for directly enhancing user experience in information search and management [1]. On the other hand, the huge deposit of multimedia data makes it possible to provide solutions to many problems that were believed to be *unsolvable* [2].

In recent years, some attempts of utilizing unlimited Web image data have been done on different image related research directions, such as image annotation [3][4], scene recognition [4], and content-based image retrieval [5]. On the one hand, the dense sampling of the visual world makes many image related problems solved without the need for sophisticated algorithms. On the other hand, unlimited images (millions of images) and classes (thousands different words) bring us more challenges on how to design effective algorithms under these large scale scenarios. In fact, most of existing attempts in large scale are based on k-nearest-neighbor method [3][4][5].

The k-nearest-neighbor method has been widely used in many computer vision problems, such as interest point matching [6], pose estimation [7], character recognition [8], and object recognition [6]. In spite of the success of k-nearest-neighbor method in both large scale image related tasks and traditional computer vision problems, its sensitivity to the value of neighbor number k makes it somewhat limited when applied to large scale image related problems. First, the visual space might be too complicated that a single fixed global parameter is not guaranteed optimal for individual datum. Second, the number of images for different classes is usually quite diverse, which makes the situation even worse. Third, an image may belong to multiple classes on different semantic levels. For example, an image of "flatfish" could belong to classes on different semantic levels such as "flatfish", "fish", and "animal". It is evident that it is not reasonable to fix k for classes on different semantic levels. Thus, it is necessary to develop a more robust alternative for these complicated large scale problems.

Recently, the concept of sparse representation attracts more and more attention in signal processing and computer vision owing to its powerful discriminative ability. Sparse representation is widely used in statistical signal processing community, whose original goal is to represent and compress signals. It is computed with respect to an overcomplete dictionary of base elements or signal atoms [9][10]. The resulting optimization problem is similar to the Lasso in statistics [11][12], which penalizes the ℓ^1-norm of the coefficients in the linear combination. Recently, Wright et al. [13] proposed a robust face recognition method based on this sparse representation, which can handle occlusion and corruption well

---

and thus achieve striking recognition performance in the face recognition task.

In this paper, to avoid the sensitivity of parameter selection in k-nearest-neighbor method and improve the discriminative capability, we propose an alternative method, called ℓ^1-nearest-neighbor, based on a sparse representation computed by ℓ^1-minimization. We first treat a testing sample as a sparse linear combination of training samples, and then consider the related training samples as the nearest neighbors of the testing sample. Finally, we classify the testing sample based on the majority of these neighbors' classes. In the ℓ^1-nearest-neighbor method, the number of neighbors of a testing sample is adaptively determined by the ℓ^1-norm cost function. Although there is a parameter, it is much more robust than the case in k-nearest-neighbor. We conduct extensive experiments on a 1.6 million natural image database on different semantic levels defined based on WordNet [14]. Experimental results show the superior discriminative capability of ℓ^1-nearest-neighbor over k-nearest-neighbor in large scale natural image classification task. The main contribution of this paper is a near-neighbor-number adaptive method for large scale natural image classification problem.

2. ℓ^1-NEAREST-NEIGHBOR CLASSIFIER

Assume that there are N natural images in the training set, represented as a matrix $X = [x_1, x_2, \cdots, x_N], x_i \in \mathbb{R}^m$, where m is the feature dimension. The class label of the image x_i is assumed to be $l_i \in \{1, 2, \cdots, N_c\}$, where N_c is the total number of classes. Let us denote the testing image as x^t.

2.1. Motivations

The k-nearest-neighbor method has been widely used in recent large scale image related works, such as image annotation, scene recognition, and content-based image retrieval. With a fixed k, this method first retrieves k nearest neighbors to the query image x^t from the training set, and then sets the label of x^t as the most frequent label of these k nearest neighbors. In spite of the simplicity and effectiveness of k-nearest-neighbor in many applications, however, under our scenario, e.g. millions of images and thousands of words, we could not expect that data are evenly distributed in the data space, and thus a single fixed global parameter is not guaranteed optimal for individual datum. Moreover, it is also not reasonable to fix k for different class labels from different semantic levels.

Another intuitive way to obtain the proper number of available neighbors for individual testing image is to use the data reconstruction method. That is, for a testing image x^t, the best coefficients to reconstruct x^t with all the training data are computed by

$$\min_{a} \quad \|x^t - \sum_{i=1}^{N} a_i x_i\|_2 \tag{1}$$

where $\|\cdot\|_2$ is the ℓ^2-norm of a vector. However, this solution is too dense and may easily overfit to the noise in the data.

Sparse coding or sparse representation is a more natural choice, since it can provide sparse combination of training data for a testing datum, and also the adaptive neighbor set. It was proposed in [15] that the human vision system is near to optimality in the representation of natural scenes only if optimality is defined in terms of sparsely distributed coding rather than compact coding. Olshausen et al. [16] employed Bayesian models and imposed ℓ^1 priors to the coefficients a_i for deducing the sparse representation. Instead of using the generic dictionaries, Wright et al. [13] represented the test sample in an overcomplete dictionary whose base elements are the training samples themselves, and achieved striking performance in the face recognition problem. Beyond small scale face recognition problem explored in [13], we focus on the large scale natural image classification problem based on sparse representation.

2.2. Sparse Representation by ℓ^1 Optimization

Sparse coding is to compute the linear sparse representation with respect to an overcomplete dictionary of base elements. It is well known that the sparsest representation problem is NP-hard in general case. However, recent results [9] have shown that if the solution is sparse enough, the sparse representation can be recovered by a convex ℓ^1-minimization. Suppose we have an underdetermined system of linear equations: $x = D\alpha$, where $x \in \mathbb{R}^m$ is the vector to be approximated, $\alpha \in \mathbb{R}^K$ is the vector for unknown reconstruction coefficients, and $D \in \mathbb{R}^{m \times K}(m < K)$ is the overcomplete dictionary with K bases. If the solution for x is sparse enough, it can be recovered by the following convex optimization:

$$\min_{\alpha} \quad \|\alpha\|_1, \quad s.t. \quad x = D\alpha. \tag{2}$$

In practice, the exact $x = D\alpha$ may not be satisfied due to noise and m may be larger than K. To overcome this issue, Wright et al. [13] proposed to reformulate the reconstruction relationship as

$$x = D\alpha + \zeta, \tag{3}$$

where ζ is the noise term. Then the sparse representation problem is redefined by minimizing the ℓ^1-norm of both coefficients and reconstruction error, and turned to solve

$$\min_{\alpha'} \quad \|\alpha'\|_1, \quad s.t. \quad x = B\alpha', \tag{4}$$

where $B = [D, I] \in \mathbb{R}^{m \times (K+m)}$ and $\alpha' = [\alpha^T, \zeta^T]^T$. This problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution. In our experiments, we convert the original constrained optimization problem into an unconstrained one, with an extra regularization coefficient:

$$\min_{\alpha'} \quad \lambda\|\alpha'\|_1 + \frac{1}{2}\|x - B\alpha'\|_2^2 \tag{5}$$

2.3. ℓ^1-Nearest-Neighbor Classifier

Based on the sparse representation described as above, the ℓ^1-nearest-neighbor classifier is given as follows:

1. **Input:** The training data $X = [x_1, x_2, \cdots, x_N], x_i \in \mathbb{R}^m$. A testing image $x^t \in \mathbb{R}^m$.

2. **Sparse Representation:** The sparse representation is obtained by solving the optimization problem

$$\min_{\alpha^t} \quad \lambda\|\alpha^t\|_1 + \frac{1}{2}\|x^t - B\alpha^t\|_2^2,$$

   where $B = [X, I] \in \mathbb{R}^{m \times (N+m)}$ and $\alpha^t \in \mathbb{R}^{N+m}$.

3. **ℓ^1-Nearest-Neighbor:** The nearest neighbors of the testing image x^t is $\{x_i | \alpha_i^t > 0, i = 1, \cdots, N\}$.

4. **Output:** The label of x^t could be set as the most frequent label in the nearest neighbor set.

3. EXPERIMENTS

The proposed algorithm was evaluated for natural image classification on a 1.6 million image database[1] provided by Torralba et al. [4]. We first introduce the setup of these experiments, and then provide the detailed experimental results.

3.1. Experiment Setup

3.1.1. Data Collection

This database contains 1,608,326 color images of size 32×32 pixels, labeled by 53,650 non-abstract nouns in English, which are a subset of the lexicon of WordNet. There exist about 30 images averagely for each word (or label). All images are crawled from the Web, which are the first images returned by the online search tools with words as queries.

Each image was first transformed to a 3072-dimensional vector by concatenating the three color channels. To improve speed, images were then represented by the first 200 principal components of the features, followed by a ℓ^2 normalization for each image. 5000 images randomly selected from the database are used as testing images, and the other ones are used as training images.

3.1.2. Classes on Different Semantic Levels

All the words in the database are from the nouns lexicon of WordNet, and hence we could utilize WordNet to divide all the words into different semantic levels.

WordNet provides semantic relationships between more than 110,000 nouns. For simplicity, we reduce the initial graph-structured relationships between words to a forest-structured one by only taking the most common meaning of each word as well as considering only the "IS-A" relation.
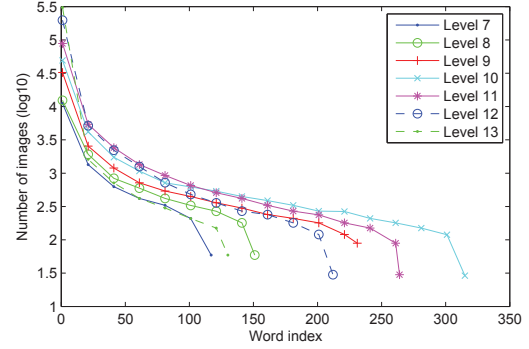

Fig. 1. The number of images labeled for each word on different semantic levels. The numbers of words on level 7 to level 13 are 130, 212, 264, 315, 231, 151, and 117, respectively. For each semantic level, words are sorted according to the number of labeled images.

Those words with exactly the same meaning (in the same synset in WordNet) are considered as one word. It should be noted that after merging words with the same meaning, the number of words in the database is reduced to 36,766. We use the biggest tree in the reduced word forest in this experiment, whose root is "entity". This tree has 32,659 words, and contains most of labels in the database. There is a strict "IS-A" relation between offspring and forefather on this tree. For example, "weaverbird" is an "oscine", and an "oscine" is an "bird". All the words on this tree are all "entity", which is the root of the tree. Thus, for every image in the database, we add to it all forefather labels of its original label.

We divide the whole tree into different semantic levels, according to the distance between the corresponding node and the root. For example, we define "entity" as the highest semantic level 18, since there are 18 levels in total; the sons of "entity" are at level 17; the grandsons of "entity" are at level 16; and so on. We conduct natural image classification task on different semantic levels respectively. On the one hand, it is impossible and unnecessary to recognize all objects in the lowest semantic levels. On the other hand, it is lack of value to classify images in very high semantic levels, e.g. distinguishing "matter" with "object". Thus, we conduct experiments only in the middle semantic levels, i.e. from level 7 to level 13, in which the words with less than 5 offspring words are also removed. Fig. 1 shows the number of images for each word on different semantic levels. From Fig. 1 we can see that the numbers of images belonging to different classes are quite unbalanced, which makes the problem more difficult.

3.2. Classification Results on Different Semantic Levels

Fig. 2 shows the classification accuracies on different semantic levels, with different parameters, for different classifiers. From this figure, we can have a serial of conclusions. First, the performance of k-nearest-neighbor (kNN) algorithm is
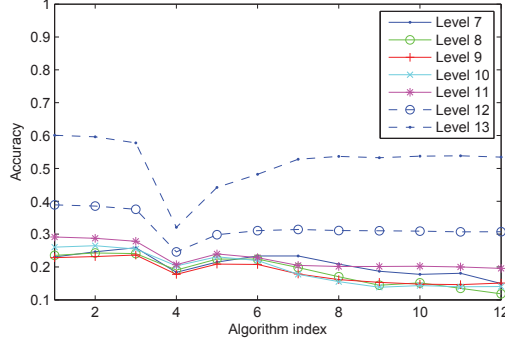
Fig. 2. Classification accuracies on different semantic levels. Algorithm 1 to 3 represent ℓ^1-nearest-neighbor algorithms with λ being 0.05, 0.1, and 0.2, respectively. Algorithm 4 to 12 represent k-nearest-neighbor algorithms with k being 1, 5, 10, 50, 100, 150, 200, 300, and 500, respectively.

quite sensitive to the parameter k within each semantic level. Second, the best ks of kNN on different semantic levels are quite different. kNN prefers larger k on higher semantic levels, while prefers relatively smaller k on lower semantic levels. Third, the ℓ^1-nearest-neighbor (ℓ^1NN) classier consistently outperforms kNN method on different semantic levels. Moreover, the performance of ℓ^1NN is much less sensitive to its parameter λ than that of kNN. The ℓ^1NN algorithm with λ being 0.2 cost about 30 seconds on average for one image using Matlab without any special optimization, on a single computer with 3.00GHz Intel Xeon CPU and 16G memory. We believe that great improvements on efficiency could be achieved if special optimization such as efficient indexing or parallel computing is adopted.

To show the importance of adaptively choosing the number of neighbors, we compared three types of experimental results in Fig. 3: 1) accuracy from the kNN with k being 1, 2) accuracy from ℓ^1NN, and 3) accuracy from ℓ^1NN and for the testing images with exactly one neighbor in ℓ^1NN (denoted as "ℓ^1NN(k=1)"). The average number of neighbors is about 30 for ℓ^1NN with λ being 0.1, and about 1/6 testing images have only one neighbor. The superior performance of ℓ^1NN(k=1) over kNN(k=1) shows that it is very useful to adaptively determine the number of neighbors in ℓ^1NN. We can see that, ℓ^1NN will totally believe the nearest neighbor of the testing image if it finds that this neighbor is a very reliable one; otherwise, it will automatically turn to more neighbors. This strategy makes ℓ^1NN more robust in such a complex scenario where a fixed neighbor number is not enough to handle all classes of natural images on different semantic levels.

4. CONCLUSIONS

In this work, we have presented an alternative of k-nearest-neighbor classifier, called ℓ^1-nearest-neighbor, to solve the large scale natural image classification problem. The chal-
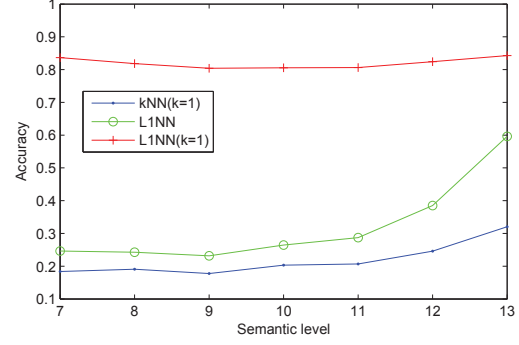

Fig. 3. Classification accuracies with only one neighbor on different semantic levels. "kNN(k=1)" represents the k-nearest-neighbor method with k being 1. "L1NN" is the ℓ^1-nearest-neighbor algorithm with λ being 0.1. "L1NN(k=1)" means that we only calculate the average accuracy of testing images which have exact one neighbor in L1NN.

lenges of this problem, e.g. the unbalance of samples in different classes and in different semantic levels, make a fixed number of neighbors not robust enough in k-nearest-neighbor based classification. The proposed ℓ^1-nearest-neighbor classifier, however, could adaptively determine the number of neighbors for a testing image, and thus be more robust and have more discriminative capability for natural image classification on different semantic levels. Extensive experiment results on a 1.6 million image database validated the algorithmic effectiveness under the large scale scenario.

5. REFERENCES

[1] F. Jing, et al. IGroup: Web Image Search Results Clustering. *ACM MM*, 2006.

[2] T. Yeh, K. Tollmar, and T. Darrell. Searching the Web with Mobile Images for Location Recognition. *CVPR*, 2004.

[3] C. Wang, F. Jing, L. Zhang, and H. -J. Zhang. Scalable Search-Based Image Annotation of Personal Images. *ACM MIR*, 2006.

[4] A. Torralba, R. Fergus and W. T. Freeman. Tiny Images. *Technical Report, Computer Science and Artificial Intelligence Lab, MIT*, 2007.

[5] C. Wang, L. Zhang, and H. -J. Zhang. Learning to Reduce the Semantic Gap in Web Image Retrieval and Annotation. *ACM SIGIR*, 2008.

[6] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. *CVPR*, volume 1, pages 26-30, June 2005.

[7] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. *ICCV*, 2003.

[8] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. *NIPS*, pages 831-837, 2000.

[9] D. Donoho. For most large underdetermined systems of linear equation the minimal l_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Math*, vol. 59, no. 6, pp. 797-829, 2006.

[10] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies?. *IEEE Trans. Information Theory*, 2006.

[11] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, no. 7, pp. 2541-2567, 2006.

[12] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267-288 1996.

[13] J. Wright, et al. Robust Face Recognition via Sparse Representation. *IEEE Trans. PAMI*, 2008.

[14] C. Fellbaum. Wordnet: An Electronic Lexical Database. *Bradford Books*, 1998.

[15] D. Field. What is the Goal of Sensory Coding? *Neural Computation*, 1994.

[16] B. Olshausen and D. Field. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research*, 1997.