DEPTH-EMBEDDED MULTIPLE POOLING FOR IMAGE CLASSIFICATION

Zhen Zhou, Yongzhen Huang, Liang Wang, Tieniu Tan

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, P.R.China, 100190
{zzhou, yzhuang, wangliang, tnt}@nlpr.ia.ac.cn

ABSTRACT

Most existing methods of image classification ignore the role of depth information hidden in 2-D images. However, the depth information is important for visual perception, especially when the appearance information does not perform well. In this paper, we propose to embed depth information within multiple pooling into the classic platform of image classification, namely bag-of-features. The proposed method quantifies depth diversity by projecting objects to their nearby depth planes, resulting pooling features in the 3-D space indirectly. Experimental results on the MIT Indoor Scene database demonstrate that our proposed depth-embedded multiple pooling is effective to enhance the accuracy of image classification, especially when the appearance features alone are not so discriminative.

Index Terms— Image Classification, Depth Estimation, Multiple Pooling

1. INTRODUCTION

In the field of image classification, a number of approaches have been proposed to study the role of appearance information. However, the importance of depth information is not paid enough attention. Consider the illustration in Figure 1, which shows a toy example of image matching between two groups of images, i.e., four real face photos (upper left) and an picture of the Mount Rushmore (bottom left). The appearance representations of these faces are similar. Their depth information, however, is of big difference, according to which it is easier to differentiate between these two groups of images. Therefore, appearance-based image classification methods are not so powerful to handle such problems. Motivated by this observation, we aim to build a model to embed depth information into appearance-based image classification in this paper.

Very recently, Redondo et al. [1] propose to make use of depth information for object recognition. In their method, the depth of images is directly extracted by a RGB-D sensor, which is used to recognize object categories in point clouds. Our task in this paper is different from theirs. We focus on

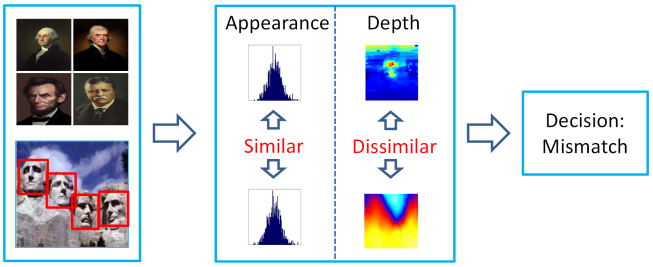


Fig. 1. An illustration of image matching with both appearance and depth information. Although the appearance information provided by BoF is similar between two kinds of images, their depth responses are much dissimilar. So the resulting decision is that they are mismatched.

2-D image classification and extract depth information from a single image. Obtaining depth information from a single image is a challenging problem in computer vision, and several related works have been proposed. Liu et al. [2] estimate the depth of each pixel from a single monocular image by using semantic segmentation and labels of scenes. Hoiem et al. [3] construct the surface layout of scenes, by which they generate a depth map with satisfying visual effect. Saxena et al. [4] apply Markov Random Field for depth estimation, and achieve good performance on both quantitative and qualitative evaluation.

In the real environment, an object's depth generally changes in a reasonable range in order to keep its appearance representation unchanged. Inspired by this consideration, we adopt the multiple pooling technique [5] to model the depth information. Here, we choose multiple pooling as the embedded method because it is effective and efficient, even with a small size of codebook. We call this process as depth-embedded multiple pooling (DMP), which can be understood as, in an intuitive perspective, grouping features into a number of levels with respect to their depth. With depth information, two features with different levels cannot be matched even if they are very similar in appearance representation. In this way, it is easy to filter objects with unreasonable depth information, e.g., a clear face picture with the depth of more than 200 meters (see the picture of the

Mount Rushmore in Figure 1). The DMP is integrated into the classic bag-of-features (BoF) model [6] for image classification. Experimental analysis demonstrates that the proposed method can largely enhance the classification accuracy.

The remainder of this paper is organized as follows. Section 2 describes our method in detail, including the framework of the algorithm, the process of depth-embedded multiple pooling and an in-depth analysis. Experimental results and analysis are provided in Section 3. Section 4 concludes the whole paper and discusses future work.

2. OUR METHOD

2.1. Framework

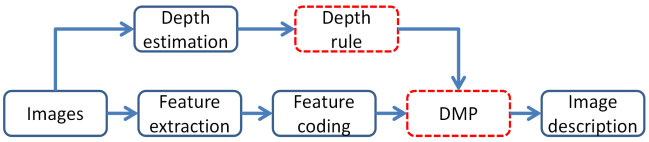


Fig. 2. The framework of our method.

Figure 2 shows the framework of our method. The BoF model is chosen as the platform of our method because it is probably one of the most successful methods for image classification. Moreover, it is easy to embed depth information into its pooling process. The BoF model includes three major components.

- **Feature Extraction.** Local features, e.g., SIFT features [7] in our method, are extracted from image patches, which are sampled from each image via the feature detector [8] or with a fixed grid [9].
- **Feature Coding.** A codebook is generated, usually obtained by clustering over local features, and then local features are encoded by this codebook. In our framework, Locality-constrained Linear Coding (LLC) [10] is chosen for feature coding. This coding method performs very effectively and efficiently in many databases for image classification [11], [12]. Other coding methods can be found in [13].
- **Feature Pooling.** This step aims to integrate all responses on each codeword into one value. Afterwards, these values are being pooled in order to produce the image representation. An in-depth analysis about feature pooling can be found in [14], [15].

To extract depth from a single image, we use Saxena et al.'s method [16][1]. As we analyze in Section 1, this method performs well in both visual effect and quantitative evaluation.

[1]The code is available from http://make3d.cs.cornell.edu.

In the step of depth-embedded multiple pooling, extracted features are divided into a number of groups according to their locations in the 3-D space. The coding responses of features are then pooled with respect to the generated groups, which will be detailed in Section 2.2.

2.2. Depth Multiple Pooling

Some notations are provided at the beginning of this section. For each input image, the feature extraction step generates a number of features $\{f_i \in \mathbb{R}^M\}_{i=1,2,\cdots,N}$, where M is the dimension of each feature and N is the number of features. Let $\{l_i \in \mathbb{R}^3\}_{i=1,2,\cdots,N}$ denotes the spatial information of features, where $l_i = (x_i, y_i, d_i)$ represents the location of the i^{th} feature in the 3-D space. d_i is the depth information, (x_i, y_i) is the image location of the specified feature.

The main idea of multiple pooling [5] is to group features by the clusters generated in the feature space. As a result, features in the same group are represented with the same bases being shared. Therefore, the rule of multiple pooling is related with the cluster in the feature space. We extend the original multiple pooling from the feature space to the 3-D space described by $l_i = (x_i, y_i, d_i)$. And accordingly, the rule is expressed as:

$$\Psi(f, l) = \begin{cases} 1, l \in C. \\ 0, otherwise. \end{cases} \tag{1}$$

where f is the feature and the corresponding location is l. C is the groups of features. Ψ is the rule which tells us a feature where it belongs. The reason considering depth in image classification can be explained. Suppose there are several kinds of features in images, represented by different colors of points. It is difficult to separate these features in the 2-dimensional space. However, with depth information it might be easy to differentiate between them by making use of their distributions in depth.

The key points of Equation 1 exist in that how to define the groups of features, C, and how to decide whether l belongs to C. To solve the first problem, depth is quantized to a series of discrete values, which is conducted by clustering methods. We describe our scheme to address the second issue as the following form.

$$\Psi(f_i, l_i) = \begin{cases} \alpha, l_i \in C_1. \\ 1 - \alpha, l_i \in C_2. \end{cases} \tag{2}$$

$$\alpha = \frac{\tau(l_i, C_2)}{\tau(l_i, C_1) + \tau(l_i, C_2)} \tag{3}$$

$$\tau(l_i, C_1) = |d_i - {C_1}^d| \tag{4}$$

where C_1 and C_2 are the nearest two groups from l_i. $\tau(\cdot, \cdot)$ measures the distance between the two elements. ${C_1}^d$ is the depth of C_1. Therefore, we choose the nearest two groups and

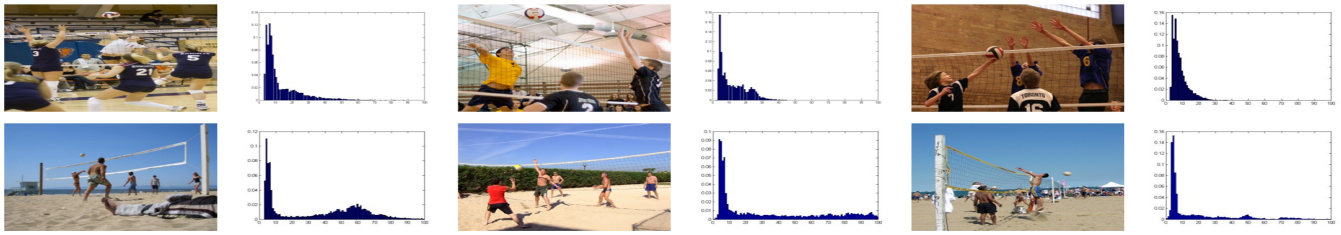


Fig. 3. Typical examples of the 2 categories from the Sun database. The first row is the samples of indoor volleyball and the second is the beach volleyball. The histogram represents the corresponding depth distribution of the image. Notice the depth shown here is limited to 100 meters.

assign them with soft weights. Considering Equation 2, it is clear to find that our multiple pooling rule is operated in the recovered 3-D space. A further analysis on depth-embedded multiple pooling will be provided in the next subsection.

2.3. In-depth Analysis

In this section, we provide in-depth analysis to the depth-embedded multiple pooling technique. Consider two different images shown in Figure 4 captured from different distances, e.g., 5 meters and 200 meters, respectively. Most existing methods of object categorization only consider appearance-based object matching, and thus it is difficult to differentiate between these two images which have similar appearance representations but belong to different categories. In our method, depth is quantized to a number of levels, and objects from the same (or nearby) levels can be matched. Our strategy actually projects original images into several depth planes, which are used to approximate the appearance representation in the real 3-D space. This process can be formulated as

$$[\mathbb{R}^2] \longrightarrow [\mathbb{R}^{2,1}, \mathbb{R}^{2,2}, \cdots, \mathbb{R}^{2,i}, \cdots, \mathbb{R}^{2,K}] \longrightarrow [\mathbb{R}^3], \quad (5)$$

where $\{\mathbb{R}^{2,i}\}_{i=1,2,\cdots,K}$ is a series of quantized depth planes.

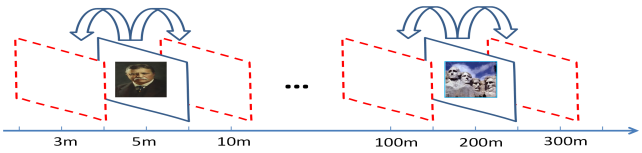


Fig. 4. An illustration of depth projection.

Due to depth quantification, this kind of 2.5-D representation is insensitive to small depth variations (usually induced by object moving in a reasonable range) and also discriminative to large depth difference (typically caused by different objects).

3. EXPERIMENTAL RESULTS

To test the performance of the proposed method, an empirical study is conducted in this section. The first part introduces the datasets and the experimental setup, followed with experimental results and analysis in the second part.

3.1. Datasets and Setup

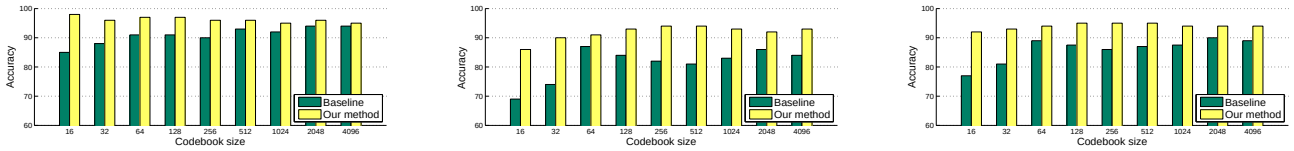
We choose two typical categories from the Sun database [17], the indoor volleyball and beach volleyball, which have similar appearance features, e.g., features of both the volleyball and the net. It is difficult to distinguish one from another exactly only by the appearance representation. This experiment can be used to verify the effectiveness of using depth information in image classification, which gives both satisfying visual effect and quantitatively accurate results. Representative samples are shown in Figure 3. Expanding this dataset to a larger one, we choose the Indoor Scene Recognition database [18] for our experimental analysis. This database contains 67 indoor categories consisting of 15620 images, with at least 100 images per category. The reason why we choose this database is that the depth of the indoor scene can be estimated with a relatively high accuracy, because of which we can concentrate more on the effect of depth information.

Following previous work [18], for each category, we use 80 images for training and 20 images for testing. To quantify depth, $K-means$ clustering is used over features' depth values, and the parameter K is determined by cross validation. The baseline is the BoF algorithm with multiple pooling. In all cases the performance is denoted by the average accuracy.

3.2. Results

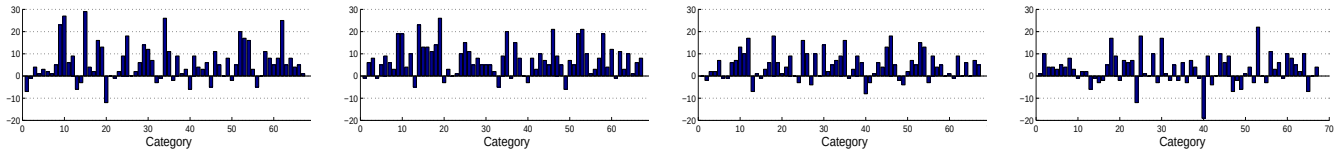
The experimental result on the 2 categories of the Sun database is summarized in Figure 5. The accuracy of our method is higher than that of the baseline both on average and for each category. As shown in Figure 3, the depth distribution of these categories is different, i.e., the indoor mainly centers on a limited depth while the outdoor has a long tail. With the depth information, it helps to distinguish between these two categories which have similar appearance representations.

We notice that there are some previously published methods that have given the accuracy the Indoor Scene Recognition database as listed in Table 1. Generally speaking, our proposed depth-embedded multiple pooling is comparable to



(a) indoor (b) outdoor (c) average

Fig. 5. Classification accuracy of the 2 categories dataset from the Sun database (%).



(a) codebook 16 (b) codebook 64 (c) codebook 256 (e) codebook 1024

Fig. 6. The differences of performance on the Indoor Scene Recognition database of our method and the baseline for each category. The horizontal axis indicates the 67 categories, and the vertical one represents the values of the accuracy by our method minus the corresponding one of the baseline. Positive values are the major part for each subgraph which implies that our method outperforms the baseline on most of categories for each codebook size.

these methods. Further more, Figure 6 and Table 1 summarize quantitative results of the Indoor Scene Recognition database, from which the following conclusions can be obtained.

1. Our method is overall effective to enhance the classification accuracy. The average accuracy of the proposed method outperforms the baseline algorithm by at most 7.4% (when codesize = 64). This result demonstrates that depth information is indeed beneficial for image classification.

2. The enhancement is more obvious when a small number of codewords are used. This is probably caused by two reasons. Firstly, the baseline algorithm performs poor with a small number of codewords. In this case, it is easier to be improved. Secondly, as the codebook size increases, the dimensionality of the final representation expands accordingly, and the over-fitting risk becomes larger.

3. We note that in some classes, our method performs very well. For example, in the class of poolinside, our algorithm achieves 22% enhancement even when the codebook size is 1,024. However, for some cases, our method does not achieve satisfying performance. We suppose that, in most of these unsatisfying cases, the appearance is enough discriminative.

4. CONCLUSION AND FUTURE WORK

We have proposed depth-embedded multiple pooling to embed depth information into the BoF platform. Meanwhile, we have explained the underlying mechanism of the proposed method from depth projection and the approximation to 3-D space. The experimental results support that adding depth information can enhance the classification accuracy, especially when the appearance information performs relatively poor. Future work will mainly focus on investigating more reasonable rules for grouping features through depth-embedded multiple pooling.

Table 1. Accuracy on the Indoor Scene Recognition database from some previous works and our experiment(%).

Method		Accuracy
Quattoni et al., CVPR 2009. [18]		26.5
Zhu et al., NIPS 2010. [19]		28.0
Li et al., NIPS 2010. [20]		37.6
Wu et al., PAMI 2011. [21]		36.9
Pandey et al., ICCV 2011. [22]		30.4
Pandey et al., ICCV 2011. [22]		**43.1**
Parizi et al., CVPR 2012. [23]		37.9
#Codes = 16	Baseline	11.9
	Our Method	**17.8**
#Codes = 64	Baseline	19.0
	Our Method	**26.4**
#Codes = 256	Baseline	29.7
	Our Method	**34.4**
#Codes = 1024	Baseline	38.1
	Our Method	**41.0**

5. ACKNOWLEDGMENT

This work is jointly supported by National Natural Science Foundation of China (61175003, 61135002, 61203252), Hundred Talents Program of CAS, National Basic Research Program of China (2012CB316300), National Key Technology R & D Program (2011BAH11B01), and Tsinghua National Laboratory for Information Science and Technology Cross-discipline Foundation.

6. REFERENCES

[1] C. Redondo-Cabrera, R.J. López-Sastre, J. Acevedo-Rodriguez, and S. Maldonado-Bascón, "Surfing the point clouds: Selective 3D spatial pyramids for category-level object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3458–3465.

[2] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Computer Vision and Pattern Recognition*, 2010, pp. 1253–1260.

[3] D. Hoiem, A.A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.

[4] A. Saxena, S.H. Chung, and A.Y. Ng, "3D depth reconstruction from a single still image," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 53–69, 2008.

[5] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: multi-way local pooling for image recognition," in *International Conference on Computer Vision*, 2011, pp. 2651–2658.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, vol. 1, p. 22.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[9] M. Marszałek, C. Schmid, H. Harzallah, J. Van De Weijer, et al., "Learning object representations for visual object class recognition," in *Worshop on the PASCAL VOC Challenge, ICCV*, 2007.

[10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.

[11] Yongzhen Huang, Kaiqi Huang, Yinan Yu, and Tieniu Tan, "Salient coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1753–1760.

[12] Yongzhen Huang, Kaiqi Huang, Chong Wang, and Tieniu Tan, "Exploring relations of visual codes for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1649–1656.

[13] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference*, 2011.

[14] Y.L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

[15] Y.L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *International Conference on Machine Learning*, 2010, pp. 111–118.

[16] A. Saxena, S.H. Chung, and A. Ng, "Learning depth from single monocular images," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1161, 2006.

[17] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[18] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[19] J. Zhu, L.J. Li, L. Fei-Fei, and E.P. Xing, "Large margin learning of upstream scene understanding models," *Advances in Neural Information Processing Systems*, vol. 24, 2010.

[20] L.J. Li, H. Su, E.P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," *Advances in Neural Information Processing Systems*, vol. 24, 2010.

[21] J. Wu and J.M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[22] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *International Conference on Computer Vision*, 2011, pp. 1307–1314.

[23] S.N. Parizi, J.G. Oberlin, and P.F. Felzenszwalb, "Reconfigurable models for scene recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2775–2782.