DEPTH-EMBEDDED MULTIPLE POOLING FOR IMAGE CLASSIFICATION

Zhen Zhou, Yongzhen Huang, Liang Wang, Tieniu Tan

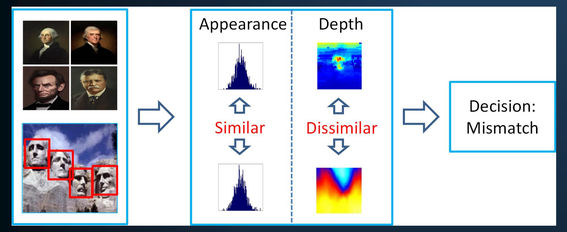
Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences

Abstract

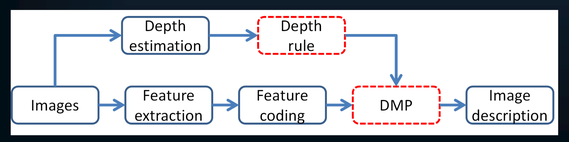
Most existing methods of image classification ignore the role of depth information hidden in 2-D images. However, the depth information is important for visual perception, especially when the appearance information does not perform well. In this paper, we propose to embed depth information within multiple 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.

Motivation

Consider the illustration in the following figure, 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.



Framework



Depth estimation

To extract depth from a single image, our way is based on the Saxena etal.'s method [1], which uses a discriminatively trained Markov Random Field that incorporates multiple scale local and global image features, and models both depths at individual points as well as the relation between depths at different points.

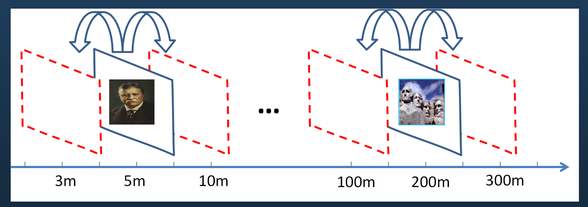
[1] A. Saxena, S.H. Chung, and A. Ng, "Learning depth from single monocular images," Advances in *Neural Information Processing Systems*, vol. 18, pp. 1161, 2006.

Depth Multiple Pooling (DMP)

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

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

where $\{\mathbb{R}^{2,i}\}_{i=1,2,\cdots,K}$ is a series of quantized depth planes. The following figure is an illustration of this depth projection.



Therefore, we choose the nearest two depth planes and assign them with soft weights.

The main idea of multiple pooling [2] 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.

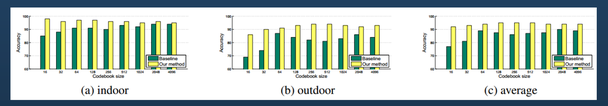
[2] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: multi-way local pooling for image recognition," in *International Conference on Computer Vision*, 2011, pp. 2651–2658.

Results

We choose two typical categories from the Sun database [3],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 as the following figure shows.

[3] 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.

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.



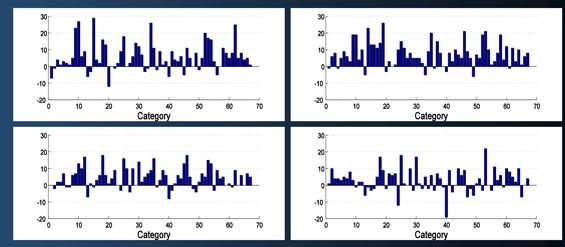
Expanding this dataset to a larger one, we choose the Indoor Scene Recognition database [4] for our experimental analysis. This database contains 67 indoor categories consisting of 15620 images, with at least 100 images per category. The reason why we choose this database is that the depth of the indoor 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 [4], 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.

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

Presented at the International Conference on Image Processing (ICIP) in the year of 2013.

Results (continued)



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.

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**

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

Conclusion

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.