NUDT's Participation in the Robot Vision Challenge of ImageCLEF2014

Yi Zhang, Jian Qin, Fanglin Chen, and Dewen Hu

Department of Automatic Control, College of Mechatronics and Automation
National University of Defense Technology, Changsha, Hunan, China
{gfkdzhangyi,qinjian714}@126.com
{fanglinchen,dwhu}@nudt.edu.cn

Abstract. This working note describes the method of the NUDT team for scene classification and object recognition in the ImageCLEF 2014 Robot Vision Challenge. The method is composed of two steps: 1. spatial pyramid match (SPM) and a Pyramid of HOG (Histograms of Oriented Gradient) are incorporated to represent an indoor place image. 2. a multi-class SVM (Support Vector Machine) is utilized to classify an image by one-versus-all binary SVMs. Based on the method, our system wins the championship this year.

Keywords: indoor robot localization, indoor object recognition, SPM, PHOG, multi-class SVM

1 Introduction

In the 6th Robot Vision Challenge of ImageCLEF 2014, the task is to address the problem of robot localization in parallel to object recognition by use of visual and depth images. Both problems are inherently related: the objects present in a scene can help to determine the room category and vice versa. In this new edition of the task, strong variations are introduced between training and test scenarios, for instance, a non-previously seen building in the test sequence, variant lighting conditions and different indoor environment distribution.

This paper describes NUDT's participation in the Robot Vision Challenge. For the image representation, only visual images are utilized and the state-of-the-art SPM and PHOG are both applied. In addition, nonlinear SVM is used for classificaiton.

The rest part of this paper is organized as follows. In Section 2, we briefly give a brief overview of the image representation in the task. In Section 3, we describe in detail the classifier which we use in both two problems. In Section 4, experimental setup and results are introduced. In Section 5, we conclude our work.

2 Image Representation

The proposed image representation incorporates appearance and shape. For appearance, we follow the approach of SPM [1]. SIFT descriptors [2] of 16×16

pixel patches are computed over a dense regular grid with spacing of 8 pixels. We perform k-means clustering of a random subset of patches from the training set to form a visual codebook of V codes, and the corresponding descriptors are assigned to their closest vocabulary codes using the Euclidean distance. In the pyramid layer, multiple codes from inside each sub-region are pooled together into a histogram. Finally, the histograms from all sub-regions are concatenated together to generate SPM. Local place shape is described by HOG [3] within an image block quantized into B bins. Orientations of edges within a certain angular range are counted, forming each bin in the histogram. Incorporating spatial pooling like SPM, PHOG is formed [4]. The final image representation is formed by concatenating SPM and PHOG. The total cells at level l in the each pyramid is 4^l. The entire image representation is a vector with dimensionality $V\sum_{l=0}^{L} 4^l + B\sum_{l=0}^{L} 4^l$. For example, levels up to $L = 1$, $V = 10$ visual words and $B = 10$ bins it will be a 100-vector.

3 Multiclass classification

We believe that adopting an appropriate multi-class classifier meets the requirement of promoting the recognition performance. The multi-class SVM using one-versus-all role proves its discrimination in many practical applications and is adopted in our system. For each class s, after the responding SVM classifier has been obtained through training, the final decision function $f_s(x)$ for a test sample x has the following form:

$$f_s(x) = max(\sum_{i=1}^{m} a_i^s y_i \ker(x_i, x) + b^s) \tag{1}$$

Where $\ker(x_i, x)$ is the kernel function, a_i^s, b^s are the learned model parameters of each one-versus-all binary SVM, y_i is the label of the training sample x_i, m is the number of the training samples. Then the ultimate label belongs to the class with the maximal value. In practice, we find that the one-versus-all SVM with nonlinear kernel is suitable for classifying the scenes and objects.

4 Experiments

In the Robot Vision Challenge this year, only one training sequence is provided with 5000 labeled images. An additional validation sequence with 1500 labelled images is provided, and includes 500 images of a non-previously seen building that presents similar room categories and objects as for the training sequences and 1000 images of the old building in the training sequence. The test sequence contains 3000 unlabeled images. In these three sequence, only visual information is considered irrespective of depth information.

For the task, we utilize the whole images of training sequence for training SVM, and the result is similar to the result of utilizing the same number of images for per class to train SVM. In the application, only grey level cues are used. And a visual codebook of 200 codes is used for appearance, and HOG with range $[0, 360]$ using all orientation is discretized into $B = 8$ bins for shape. For multi-class classification, we utilize LIBSVM [5] with one-versus-all classifying rule. Specially, pyramid match kernel [1] is adopted. Moreover, we limit the number of levels to $L = 2$ to prohibit over fitting. All experiments are repeated 10 times with all training images. Some example images are shown in Fig. 1, Fig. 2, Fig. 3.

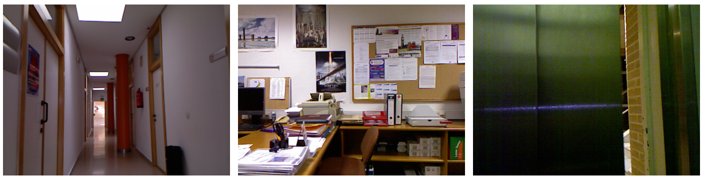

Fig. 1. Example images of scene categories in the training sequence

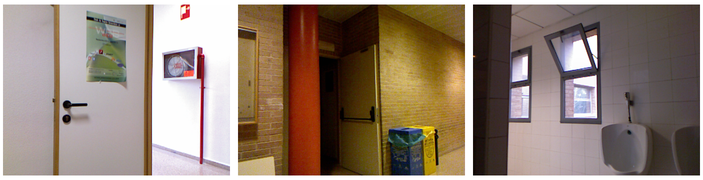

Fig. 2. Example images of object classes in the training sequence

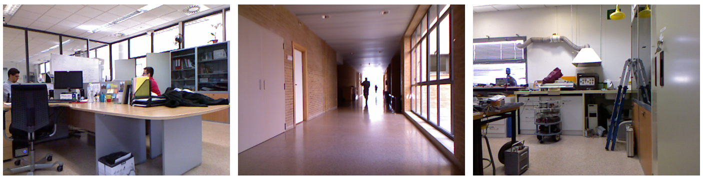

Fig. 3. Example images in the new building

4.1 Results on the validation sequence

We utilize our method on the validation sequence, and compute the respective classification accuracy for each scene category and object class. Detailed results are shown in Table 1 and Table 2.

Table 1. The accuracy on per scene category of validation data

Scene	Cor	Elev	Hall	Pro	Sec	Stu	Tec	Toilet	Visio	Ware	Total
accuracy	0.861	0.743	0.814	0.851	0.798	0.825	0.830	0.875	0.712	0.683	0.802

Table 2. The accuracy on per object class of validation data

Object	Ext	Chair	Printer	Bookshelf	Urinal	Trash	Phone	Fridge	Total
accuracy	0.842	0.815	0.713	0.741	0.831	0.804	0.672	0.710	0.798

4.2 Results on the test sequence

In the ultimate submission for classifying test sequence, we mix the training sequence and the validation sequence into a dataset for training and classify the test sequence. Scene classification and object recognition are realized separately. Finally, Our team ranks the first out of four teams, results are listed in Table 3.

Table 3. The final results

#	Group	Score Rooms	Score Objects	SCORE TOTAL
1	NUDT	1075.50	3357.75	4430.25
2	UFMS	219.00	1519.75	1738.75
3	Baseline Results	67.5	186.25	253.75
4	AEGEAN	-405	-995	-1400

5 Conclusions

In the paper, we introduce our method for the Robot Vision Challenge in ImageCLEF2014. Our method applies the SPM and PHOG for image representation and multi-class SVM for classification, which achieves the best performance among all the participants. However, From the results of the challenge, there are still some scene categories and object classes can not be processed correctly. In addition, obviously, the non-previously seen building influences the performance. In the future, we will focus on the further improvement on the classification accuracy and the generalization of our method.

References

1. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern*

Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2169–2178, 2006.

2. DavidG. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
3. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.
4. A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.
5. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
6. Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, Ismael Garcia-Varea, and Vicente Morell. ImageCLEF 2014: Overview and analysis of the results. In *CLEF proceedings*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014.
7. Jesus Martinez-Gomez, Miguel Cazorla, Ismael Garcia-Varea, and Vicente Morell. Overview of the ImageCLEF 2014 Robot Vision Task. In *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes*, 2014.