RGB-D based 3D-Object Recognition by LLC using Depth Spatial Pyramid

Toru NAKASHIKA[1,a)] Takahiro HORI[1,b)] Tetsuya TAKIGUCHI[1,c)]
Yasuo ARIKI[1,d)]

1. Introduction

Recently, high-accuracy RGB-D cameras are commertially available, which are capable of providing high quality three dimension information (color and depth). In this paper, we propose an object recognition method where the techniques of object recognition in 2D are extended to 3D.

Recent image classification systems mainly consist of the following three parts: feature extraction using scale-invariant feature transform (SIFT), coding scheme using bag-of-features (BoF) and pooling process using spatial pyramid matching (SPM) [1]. The SPM is also regarded as an extension of BoF, which partitions the image into hierarchical spatial sub-regions and computes histograms of local features from each sub-region. This spatial pyramid restricted by position has shown very promising performance on many image classification tasks. These techniques used for 2D images are applied to 3D object recognition without any changes so far. For that reason, even though the depth information captures the overall shape of an object, conventional methods use depth information only to extract the local feature.

In our proposed approach, the overall object shape is captured by the depth spatial pyramid based on depth information. In more detail, multiple features within each sub-region of the depth spatial pyramid are pooled. As a result, the feature representation including the depth topological information is constructed. We use not only SIFT, but also histograms of oriented normal vectors (HONV [2]) for the depth image, which are originally designed to capture local geometric characteristics. We also adopt locality-constrained linear coding (LLC [4]), which utilizes local constraints to project each descriptor into its local-coordinate system.

2. Methodology

2.1 Overview

Fig. 1 shows the system overview. First, the depth image and the RGB images of each channel are created from depth and color information. Feature points of each image are located by grid sampling, and features (HONV and SIFT) are extracted from each feature point. HONV is extracted from the depth image and SIFT is extracted from the RGB images. The extracted features are coded by LLC. Then, multiple codes within each sub-region of the spatial pyramid are pooled together. The pooling of the depth spatial pyramid is additionally processed for the depth image. Finally, the pooled features from all sub-regions are concatenated together for classification.

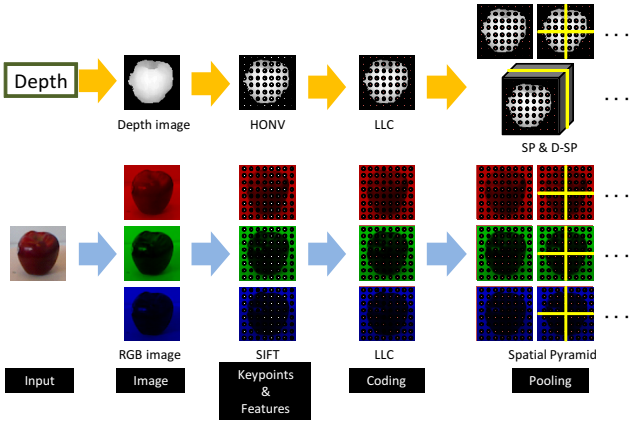

Fig. 1 System overview

2.2 Spatial pooling in 3D

Spatial pooling is the process of partitioning an image into sub-regions, and pooling multiple features within each sub-region. In this paper, we use our proposed depth spatial pyramid to divide each depth image in 3D, in addition to the conventional spatial pyramid for each image in 2D.

The depth spatial pyramid is a spatial pyramid in the depth coordinate system made from depth information. Assuming that the depth value is a coordinate, we partitions the depth image to sub-regions. However, the depth values measured disperse unlike coordinates of a usual spatial pyramid. If the space is simply divided equally like the spatial pyramid, the number of feature points within each sub-region is biased. Therefore, we divide it into sub-regions including equal number of points without dividing by coordinates. Typically, m subregions ($m = 0, 1, 2$) are used (Fig. 2). 3D space of an object is spatially constrained by

[1] Department of System Informatics, Kobe University, Rokkodai 1–1, Nada-ku, Kobe, 657–8501 Japan
[a)] nakashika@me.cs.scitec.kobe-u.ac.jp
[b)] horitaka@me.cs.scitec.kobe-u.ac.jp
[c)] takigu@kobe-u.ac.jp
[d)] ariki@kobe-u.ac.jp

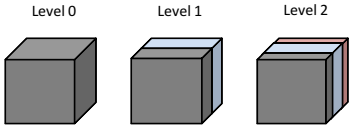

Fig. 2 Depth Spatial Pyramid

using the depth spatial pyramid and the spatial pyramid together. As a result, the overall 3D shape of the object can be expressed. In each spatial pyramid, multiple codes within each sub-region are pooled together. These pooled features from each sub-region are concatenated and normalized as the final image feature representation. We use max pooling as pooling method:

$$\boldsymbol{c}_{out1} = \max(\boldsymbol{c}_{in1}, \ldots, \boldsymbol{c}_{inH}) \quad (1)$$

where H denotes a number of feature points within the sub-region. max function in a row-wise manner returns a vector with the same size as $\boldsymbol{c}_{in1}$. These pooled features $\boldsymbol{c}_{out1}$ are concatenated as the feature vector $\boldsymbol{c}_{in}$. It is normalized by

$$\boldsymbol{c}_{out2} = \boldsymbol{c}_{in} / \|\boldsymbol{c}_{in}\|_2. \quad (2)$$

This $\boldsymbol{c}_{out2}$ is the final image feature representation. Here, for the depth image, we concatenate two feature representation made from spatial pyramid and depth spatial pyramid.

3. Experiments

We used the RGB-D Object Dataset for the object recognition experiments [10]. It is composed of 300 objects, 51 categories and about 42000 images containing RGB and depth information. Each object is recorded from three viewing heights (30°, 45° and 60° angles) while it rotates on a turntable. For our experiments, we used the same setup as in [5], distinguishing between category and instance recognition. Firstly, category-level classification experiments were conducted with 51 class labels. We randomly selected one object from each category for the test, and trained the classifier on the remaining objects. The accuracy averaged and the standard deviation over 10 random trials are reported for category recognition. Secondly, instance classification experiments with 300 objects were conducted. We trained the classifier on the images captured from 30° and 60° elevation angles, and tested them on the images from 45° angle. We present object recognition results on the RGB-D Object dataset with only depth features (Depth), only color features (RGB), and with both depth and color features (RGB-D). In our setup, the SIFT and the HONV features were extracted from points densely located by every 4 pixels on an image, under three scales, 8×8, 12×12 and 16×16 respectivel The codebook size was 1024, the best value in the experiment. We used as the classifier multi-class SVM (linear) to classify the vectors.

Table 1 shows the recognition results and the comparison with conventional methods. As shown in Table 1, it can be confirmed that the proposed method improved the accuracy. This result shows the effectiveness of the proposed method using HONV, LLC and depth spatial pyramid. However, only the result with depth features (Depth) for instance recognition does not outperform other methods. This is because the dataset contains many objects that the shapes are the same but the colors are different. It is generally difficult to recognize those objects only with shape information. Especially, the proposed method specialized for the shape representation was strongly influenced, and therefore its recognition accuracy was not improved. Nevertheless, the depth information contributes to the improvement of the recognition rate when mixed up with the color information (RGB-D).

Table 1 Recognition results and comparisons(%)

	Category			Instance		
	RGB	**Depth**	**RGB-D**	**RGB**	**Depth**	**RGB-D**
ICRA11[5]	74.3 ± 3.3	53.1 ± 1.7	81.9 ± 2.8	59.3	32.3	73.9
Kernel desc[6]	80.7 ± 2.1	80.3 ± 2.9	86.5 ± 2.1	90.8	**54.7**	91.2
CKM desc[7]	N/A	N/A	86.4 ± 2.3	82.9	N/A	90.4
HMP[8]	74.7 ± 2.5	70.3 ± 2.2	82.1 ± 3.3	75.8	39.8	78.9
ISER12[9]	82.4 ± 3.1	81.2 ± 2.3	87.5 ± 2.9	92.1	51.7	92.8
Proposed	**85.3 ± 1.6**	**82.9 ± 2.3**	**89.2 ± 1.6**	**93.4**	42.5	**94.2**

4. Conclusion

This paper presented a 3D object recognition method using HONV, LLC and depth spatial pyramid based on depth information. The feature representation including the topological information of shape was constructed by using depth spatial pyramid and spatial pyramid together. Our proposed method of expressing overall object shapes demonstrated the better performance compared with conventional methods in the experiments using 3D object dataset.

References

[1] 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, Vol. 2, pp. 2169–2178, 2006.

[2] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of Oriented Normal Vectors for Object Recognition with a Depth Sensor," The Asian Conference on Computer Vision, 2012.

[3] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893, 2005.

[4] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Guo, "Locality-constrained Linear Coding for Image Classification," IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367, 2010.

[5] K. Lai, L. Bo, X. Ren, and D. Fox, "A Large-Scale Hierarchical Multi-View RGB-D Object Dataset," IEEE International Conference on Robotics and Automation, pp. 1817–1824, 2011.

[6] L. Bo, X. Ren and D. Fox, "Depth Kernel Descriptors for Object Recognition," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 821–826, 2011.

[7] M. Blum, J. Springenberg, J. Wlfing, and M. Riedmiller, "A Learned Feature Descriptor for Object Recognition in RGB-D Data," IEEE International Conference on Robotics and Automation, pp. 1298–1303, 2012.

[8] L. Bo, X. Ren, and D. Fox, "Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms," Neural Information Processing Systems (NIPS), 2011.

[9] L. Bo, X. Ren, and D. Fox, "Unsupervised Feature Learning for RGB-D Based Object Recognition," In International Symposium on Experimental Robotics, 2012.

[10] RGB-D Object Dataset, http://www.cs.washington.edu/rgbd-dataset/