Scale and Rotation Invariant Color Features for Weakly-Supervised Object Learning in 3D Space

Asako Kanezaki Tatsuya Harada
Yasuo Kuniyoshi
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo Japan
{kanezaki, harada, kuniyosh}@isi.imi.i.u-tokyo.ac.jp

Abstract

We propose a joint learning method for object classification and localization using 3D color texture features and geometry-based segmentation from weakly-labeled 3D color datasets. Recently, new consumer cameras such as Microsoft's Kinect produce not only color images but also depth images. These reduce the difficulty of object detection dramatically for the following reasons: (a) reasonable candidates for object segments can be given by detecting spatial discontinuity, and (b) 3D features that are robust to view-point variance can be extracted. The proposed method lists candidate segments by evaluating difference in angle between the surface normals of 3D points, extracts global 3D features from each segment, and learns object classifiers using Multiple Instance Learning with object labels attached to 3D color scenes. Experimental results show that the rotation invariance and scale invariance of features are crucial for solving this problem.

1. Introduction

Object classification and localization are fundamental issues in various tasks, such as automatic object management, object manipulation by personal robots and so on. In these tasks, it is not only the appearance of objects in color images that is useful but also their geometric information given by depth images. The sensors in Microsoft's already globally popular Kinect capture color and depth images simultaneously, and the datasets of such RGB-D objects and scenes [15] are readily available. Therefore, in the near future it is expected that a large amount of RGB-D data will be uploaded and shared for object learning in the real world.

However, there are still significant difficulties with how to supervise objects in observed scenes. There exist datasets with ground truth of object location, such as LabelMe [24] and PASCAL VOC [16], but making such datasets has a high human cost. On the other hand, "weakly labeled" datasets in, for example, Flickr, where only the labels of objects instead of their locations in images are given, are much easier to create.

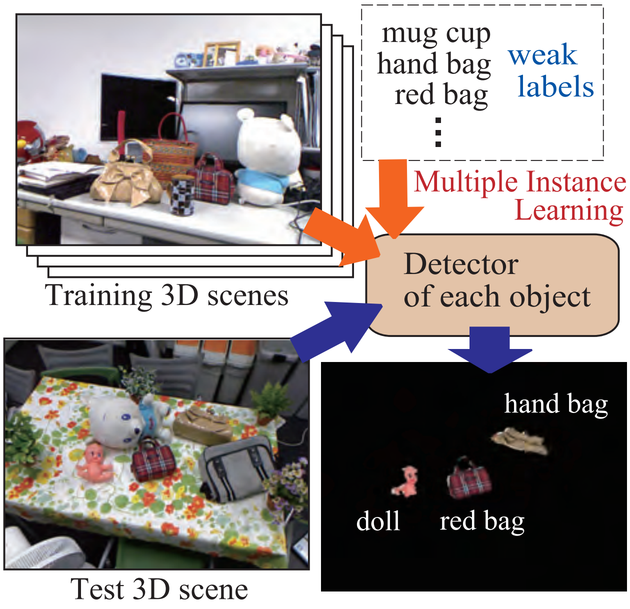


Figure 1. Overview of our system. Weak labels attached to each pair of color and depth images are used for Multiple Instance Learning, and the trained object detectors are used to localize objects in a new environment.

In this paper, we propose a joint learning method for object classification and localization that uses Multiple Instance Learning (MIL) with weakly-labeled color and depth images (See Fig. 1). Suppose there are several bags which contain multiple instances. In an MIL scenario, a positive bag label only enforces that the bag contains at least one positive instance, whereas a negatively labeled bag contains only negative instances. The proposed approach first reconstructs 3D color points from a pair of color and depth images, computes the normal of each point, and then divides the whole point cloud into multiple clusters by detecting boundaries which give large differences in normal angles. Global features are then extracted from each point cloud cluster, and they are trained as an instance of a "positive bag" if the target object's label is attached to the whole scene, or as that of a "negative bag" otherwise. In the testing process, each point cloud cluster in the whole scene is judged to be positive or negative by the learned object classifier.

The main contribution of our work is in the design of features. Since we use weakly-labeled data, no previous knowledge about each object's appearance or even its size can be used. Therefore, the features should be designed in a manner that the variation brought by a change in viewpoint is small while the differences between objects are large. We developed scale and rotation invariant features based on Circular Color Cubic Higher-order Local Auto Correlation (C^3-HLAC) Features [13], and showed better performance in experiments. Moreover, this paper is the first research on joint classification and localization learning using weakly-labeled datasets obtained by color and depth sensors.

The rest of this paper is structured as follows. Section 2 discusses related work on joint learning of object classification and localization, while the design of our 3D features is presented in Section 3. Section 4 describes the method of 3D point cluster segmentation, while Section 5 presents some experimental results. Finally, Section 6 summarizes our method and proposes ideas for future research.

2. Related Work

Joint learning of object classification and localization by weakly-labeled images has attracted much attention recently [2, 3, 4, 6, 7, 8, 9, 10, 14, 17, 20, 22, 23, 25, 27, 28, 30]. Of the part-based approaches [7, 9, 30], where each object class is represented as a collection of salient parts, Zhang and Chen [30] achieve scale-invariant representation of objects by identifying the co-occurring high order features of two images, using the idea of the generalized Hough Transform. However, since the part-based approach depends on the stability of the interesting point detector, it is not useful for identifying textureless objects. Similarly, in the works that are more focused on how to learn resion of interest (ROI) for object categories [3, 6, 14], only salient features are taken into consideration.

Other approaches are mostly regarded as segment-based, based for example on Random Field [25], CRF [2, 8], bag of words models [4], segmentation trees [27], and others [10, 22, 23, 28]. The segment-based approach is actually compatible with depth data, since geometry-based bottom-up segmentation can list reasonable candidate objects using spatial discontinuity information.

We propose a segment-based approach using depth images, 3D color features and MIL. We use object labels attached to training images as binary bag labels for each object and learn each object's classifier, which outputs the probability of the object for each segment. To perform MIL, multiple levels of segments are listed up as instances in positive or negative bags. There are several related works that use MIL for learning objects from weakly-labeled images [5, 10, 17, 20]. Our work differs from them because Maron and Ratan [17] and Chen and Wang [5] deal with not instance-level but bag-level classification in their experiments, Nguyen *et al.* [20] uses sub windows instead of precise segmentation, and Galleguillos *et al.* [10] evaluates stability of segments and thus does not consider multiple levels of segmentation. Moreover, the novelty in our approach is to use not image features but 3D color features which are extracted from 3D color points reconstructed from color and depth images. Taking advantage of the 3D property, we design features which are robust to view point variation.

3. Features

3.1. C^3-HLAC Features

We designed our new 3D features based on the C^3-HLAC features [13] which this section describes. C^3-HLAC features are extracted from color cubic voxel data, which are obtained by quantizing 3D color points. The feature vector obtained is a histogram of local RGB correlation values between neighboring voxels, and it therefore represents the characteristics of 3D color texture. Let $\boldsymbol{x} = (x, y, z)^T$ be the position of a voxel, $p(\boldsymbol{x})$ be the flag for occupancy of the voxel and $r(\boldsymbol{x})$, $g(\boldsymbol{x})$ and $b(\boldsymbol{x})$ be its RGB values normalized between 0 and 1. By defining $r_1 \equiv sin\left(\frac{\pi}{2}r(\boldsymbol{x})\right)$, $g_1 \equiv sin\left(\frac{\pi}{2}g(\boldsymbol{x})\right)$, $b_1 \equiv sin\left(\frac{\pi}{2}b(\boldsymbol{x})\right)$, $r_2 \equiv cos\left(\frac{\pi}{2}r(\boldsymbol{x})\right)$, $g_2 \equiv cos\left(\frac{\pi}{2}g(\boldsymbol{x})\right)$, and $b_2 \equiv cos\left(\frac{\pi}{2}b(\boldsymbol{x})\right)$, a voxel status $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{N}^6$ is defined as follows:

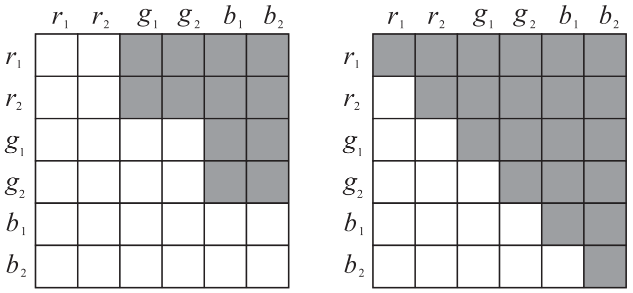
$$\boldsymbol{f}(\boldsymbol{x}) = \begin{cases} [r_1\ r_2\ g_1\ g_2\ b_1\ b_2]^T & p(\boldsymbol{x}) = 1 \\ [\ 0\,0\,0\,0\,0\,0\ ]^T & p(\boldsymbol{x}) = 0. \end{cases}$$

Although a 4-dimensional vector is enough to represent a voxel status which has RGB values and occupancy $p(\boldsymbol{x})$, a redundant 6-dimensional vector is used here. This is to make the norm of $f(\boldsymbol{x})$ constant regardless of the RGB intensity, which leads to eliminating bias in the feature space.

Let $\boldsymbol{a}_i$ be a displacement vector from the reference voxel to its neighboring voxel (*e.g.* $[1\ 0\ 0]^T$). The elements of a C^3-HLAC descriptor extracted from a voxel grid V are calculated by the following equations:

$$\boldsymbol{q}_1 = \sum_{\boldsymbol{x} \in V} \boldsymbol{f}(\boldsymbol{x}), \tag{1}$$

$$\boldsymbol{q}_2 = \sum_{\boldsymbol{x} \in V} \boldsymbol{f}(\boldsymbol{x})\ \boldsymbol{f}^T(\boldsymbol{x}), \tag{2}$$



(a) binalized color voxel data (b) original color voxel data

Figure 2. Elements in C^3-HLAC features calculated by (2). White grids are excluded since they are redundant.

$$\boldsymbol{q}_3(\boldsymbol{a}_i) = \sum_{\boldsymbol{x} \in V} \boldsymbol{f}(\boldsymbol{x})\ \boldsymbol{f}^T(\boldsymbol{x} + \boldsymbol{a}_i) \quad (i = 0, \dots 12). \qquad (3)$$

Note that the number of choices for $\boldsymbol{a}_i$ is 13, which is half of the 26 neighbors in a 3 × 3 × 3 grid, since each pair of symmetric $\boldsymbol{a}_i$ give the same correlation after summing over the entire grid. The matrix computed by Eq. (3) is expanded into a column vector with 36 elements. Therefore, the dimension of the vector calculated by Eq. (1) is 6, while that calculated by Eq. (3) is 468 (=36 · 13). The second part of the C^3-HLAC descriptor is computed from the binarized values of $r(\boldsymbol{x})$, $g(\boldsymbol{x})$ and $b(\boldsymbol{x})$. To determine the threshold of color binarization, we apply the histogram threshold selection method [21] to the R, G and B values respectively, using the voxel colors of all the objects in the database as sample data. The C^3-HLAC features calculated by Eq. (2) include redundant elements, such as $r(\boldsymbol{x}) \cdot g(\boldsymbol{x})$ and $g(\boldsymbol{x}) \cdot r(\boldsymbol{x})$. Excluding the redundant elements, the dimension is 12 if color values are binarized, and 21 otherwise (see Figure 2). Finally a full C^3-HLAC vector is obtained by concatenating the two vectors from the binarized color voxel data and the original color voxel data. As a result, the dimension of the C^3-HLAC feature vector becomes 981 (6+468+12 for non-binarized data plus 6+468+21 for binarized data).

3.2. Crucial Invariance

The most significant difficulty in MIL is that there is no clue to know which instance in a positive bag is positive. Therefore, features that are commonly included in positive bags but not in a negative bag should be found automatically. However, features extracted from the same object generally differ when the view point changes. To address this problem, features should be designed to have appropriate invariance against difference in view point.

First, rotation invariance is necessary to achieve robustness against different position. The C^3-HLAC features are not rotation invariant as described in Section 3.1. Therefore, we decided not to differentiate between the relative position of neighboring voxels $\boldsymbol{a}_i$, replacing Eq. (3) by the following equation:

$$\boldsymbol{q}_4 = \sum_{i=0}^{12} \boldsymbol{q}_3(\boldsymbol{a}_i) = \sum_{i=0}^{12} \sum_{\boldsymbol{x} \in V} \boldsymbol{f}(\boldsymbol{x})\ \boldsymbol{f}^T(\boldsymbol{x} + \boldsymbol{a}_i), \qquad (4)$$

which reduces the dimension of the descriptor to 117 (6+36+12 for non-binarized data plus 6+36+21 for binarized data).

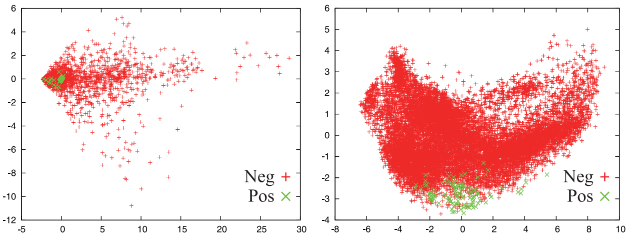


Figure 3. The two principal components of rotation-invariant C^3-HLAC features without scaling (left) and with scaling (right). Positive instances of an object are in green dots and negative instances are in red dots.

Second, scale invariance is important for robustness against an object's positional changes in depth. Unlike 2D image features, C^3-HLAC features are extracted from voxel data which are invariant to view point distance, so they already have scale invariance. However, there exists one important point. Since a C^3-HLAC feature vector is obtained by the summation of local descriptors, its norm increases as the size of a point cluster becomes larger. The two principal components of rotation-invariant C^3-HLAC features extracted from positive and negative instances of an object are shown in the left graph in Fig. 3. The features of small point clusters gather with high density, while those of large point clusters exist sparsely. This causes bias, as small segments tend to be similar to other segments while large segments tend to differ from other segments. To avoid this problem, we divide the feature vectors by $\sum_{\boldsymbol{x} \in V} p(\boldsymbol{x})$, the total number of occupied voxels in each segment. This removes the bias but still maintains scale invariance (see the right graph in Fig. 3).

4. 3D Point Cluster Segmentation

This section describes how to produce 3D point clusters as candidate segments for objects. We primarily use the geometry-based segmentation method proposed by Mozos *et al.* [19]. This method follows a criterion based on the maximum difference in angle between the surface normals. For each point, the system calculates its normal by identifying a tangent plane at the selected point, and approximating the point's neighborhood using a height function relative to this plane in the form of a 2nd order bi-variate polynomial

Algorithm 1: 3D Point Cluster Segmentation

```
S /* input 3D scene */
$l \leftarrow l_0$ /* cluster distance threshold */
P = $\{p_1...p_m\}$ in S /* plane detection */
R = $\{r_1...r_n\}$ in S-P /* distance clustering */
foreach $r_i \in R$ do
    C = $\{c_1...c_M\}$ in $r_i$ /* normal clustering */
    if $M > M_{max}$ then
        $l \leftarrow \epsilon * l$
        R' = $\{r'_1...r'_{n'}\}$ in $r_i$ /*distance clustering */
        R ← R ∩ R'
```

defined in a local coordinate system [18]:

$$h_{(u,v)} = c_0 + c_1 u + c_2 v + c_3 uv + c_4 u^2 + c_5 v^2,$$

where u and v are coordinates in the local coordinate system lying on the tangent plane. To obtain the unknown coefficients c_i, a direct weighted least squares minimization is performed and the point is projected onto the obtained surface. For further details, please refer to [19].

This segmentation produces small segments which can be used as primitive segments. Next, the segments are connected iteratively until all the segments have been connected, and thus a number of hierarchical segments of various size is obtained. However, the total number of candidate segments tends to become too large if every combination of primitive segments is considered. Therefore, we also use plane detection and a simpler cluster detection method based on point distance to limit the number of candidate segments.

The algorithm is shown in Algorithm 1. First, we estimate and detect planes by RANSAC and do clustering for the remaining points so that the distances between nearest neighbor points among clusters becomes larger than the threshold $l = l_0$. The normal-based segmentation described before is then performed for each point cluster. If the total number of segments obtained from each point cluster M is larger than M_{max}, the point clustering is performed again for this cluster with a smaller distance threshold ϵl ($\epsilon < 1$). In this paper, we set l_0 to 20mm, M_{max} to 300, and ϵ to 0.75.

5. Experiment

We evaluated the performance of object detection learned by MIL and weakly labeled 3D color scenes, that is, pairs of color and depth images with binary signatures representing whether or not they contain each object.

5.1. MIL Methods

We selected two alternative MIL methods, the Expectation-Maximization version of Diverse Density (EM-DD) [29] and multi-instance Support Vector Machines (mi-SVM) [1]. For their implementation, we used Jun Yang's library MILL [12]. The details of these methods are described below.

EM-DD

EM-DD [29] is an EM-style extension of the Diverse Density (DD) algorithm [17]. Suppose that B_i^+ is a positive bag, and that the j-th instance in that bag is B_{ij}^+. DD estimates a concept point t in feature space that is close to at least one instance in all the positive bags and is far from all the instances in all the negative bags. By assuming that the bags are conditionally independent given the concept point t and applying Bayes rule, this is solved by maximizing the following likelihood:

$$\arg\max_t \prod_i P\left(t|B_i^+\right) \prod_i P\left(t|B_i^-\right).$$

Given that the label of a bag comes from using "logical-OR" on the labels of its instances, $P\left(t|B_i^+\right)$ is finally estimated (although not necessarily) by a Gaussian-like distribution, $\exp\left(-\|B_{ij}^+ - t\|^2\right)$.

EM-DD starts with t obtained by DD as an initial guess, and repeats E-Step, which chooses the instance most likely to be positive from each bag, and M-Step, which updates t. In this experiment, we executed ten trials with different starting points and selected the one that gave the minimum objective function, that is, the minimum likelihood of training data.

mi-SVM

Andrews *et al.* [1] proposed two types of MIL-setting SVM, one for instance-level classification (mi-SVM) and the other for bag-level classification (MI-SVM). We used mi-SVM, which maximizes the usual instance margin jointly over the unknown instance labels and a linear or kernelized discriminant function, given below:

$$\min_{\{y_i\}} \min_{\boldsymbol{w},b,\xi} \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i$$
$$s.t. \forall i : y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, y_i \in \{-1, 1\},$$

where the relationship between instance labels y_i and bag labels Y_I is expressed by the linear constraints

$$\sum_{i \in I} \frac{y_i+1}{2} \geq 1, \forall I\, s.t. Y_I{=}1, and\ y_i{=}{-1}, \forall I\, s.t. Y_I{=}{-1}.$$

We used a linear discriminant function in this experiment.

5.2. Database

For our training samples, we made twelve situations where five out of twelve possible objects (see Fig. 4) were

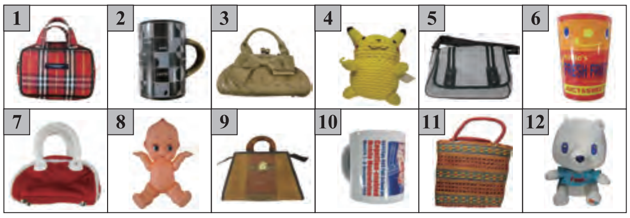


Figure 4. Images of the twelve target objects.

put in different positions, and then color and depth images were captured by a Kinect sensor. Ten pairs of color and depth images per situation were captured from different viewpoints, resulting 120 pairs of images. Note that one pair of color and depth images corresponds to one bag in MIL. There are 50 positive bags and 70 negative bags for each object.

Similarly, for testing samples, we captured ten views of each of twelve situations with five objects, but with a different environment than where training samples were collected. Example images of training samples and testing samples are shown in Fig. 5. The ground truth of object location in the testing samples was given manually. If the number of points in the output is more than 2% of the ground truth and more than a half of the number of points in the output are included in the ground truth, then the output is regarded as true positive. Note that this judgment is more optimistic than that usually used in object detection (*e.g.* PASCAL VOC [16]). This is because, in MIL, an output segment may become only a part of a whole object (*e.g.* a handle of a cup), which still works for many kinds of application. The database is available from our web site.[1]

5.3. Results

We compared several extensions of C^3-HLAC [13] features: (a) rotation-variant without scaling (original), (b) rotation-invariant without scaling, (c) rotation-variant with scaling, (d) rotation-invariant with scaling (proposed). For (a) and (c), we compressed the feature vectors into 100 dimensional vectors by doing PCA and whitening, similarly to [13]. We set the size of a voxel to 10mm × 10mm × 10mm.

Average ROC curves when using EM-DD and mi-SVM are shown in Fig. 8. For EM-DD, the decreasing order of performance was (d), (c), (b) and (a), while for mi-SVM, it became (c), (d), (b), and (a). Overall, the results with EM-DD were higher than with mi-SVM, so we conclude that (d) was the best choice among the alternative features.

A comparison of the average rate that the correct object was ranked in the top q of the target scene is shown in Fig. 6. We refer to this rate as the q-rank rate. The q-rank rate represents the probability that when the system outputs the list of q candidate segments, it includes the correct object. When we used the proposed features with EM-DD, the average rate that the system output a list of four candidates, including the correct one, was greater than 50%.

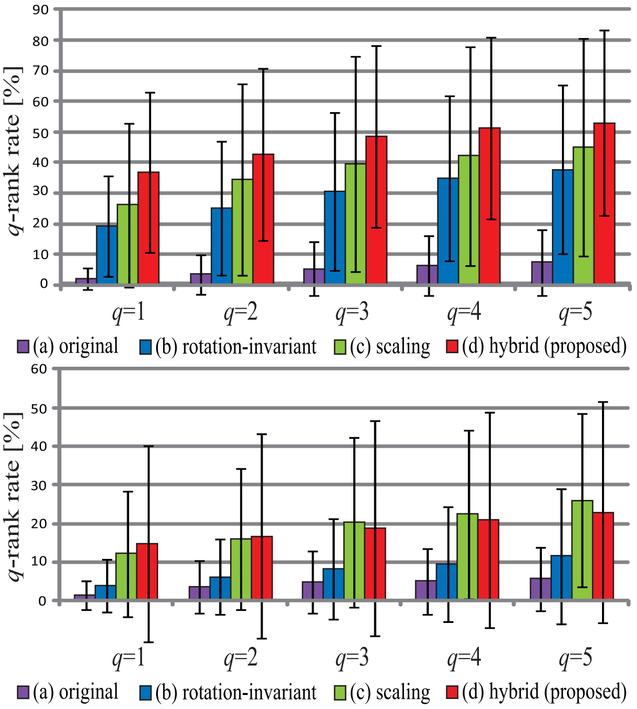


Figure 6. Comparison of average q-rank rate [%] with EM-DD (top) and mi-SVM (bottom).

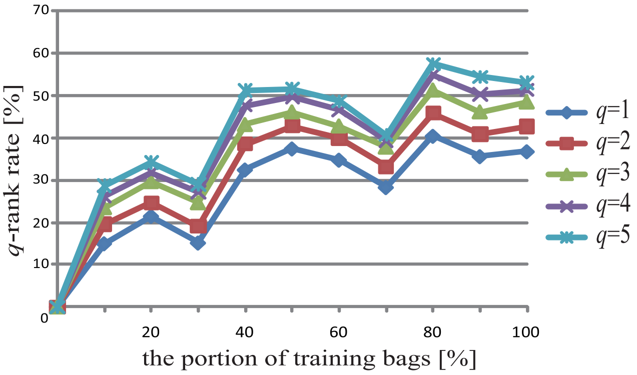


Figure 7. Average q-rank rate [%] with EM-DD vs the portion of training bags.

The average q-rank rate [%] with EM-DD versus the portion of the training bags is shown in Fig. 7. We changed the number of the training bags in our database described in Section 5.2 from 12 (10%) to 120 (100%). This indicates that larger training data will increase the accuracy.

The results shown in Fig. 8 were not as impressive as

[1] `http://www.isi.imi.i.u-tokyo.ac.jp/software/color_depth_dataset_with_labels.zip`

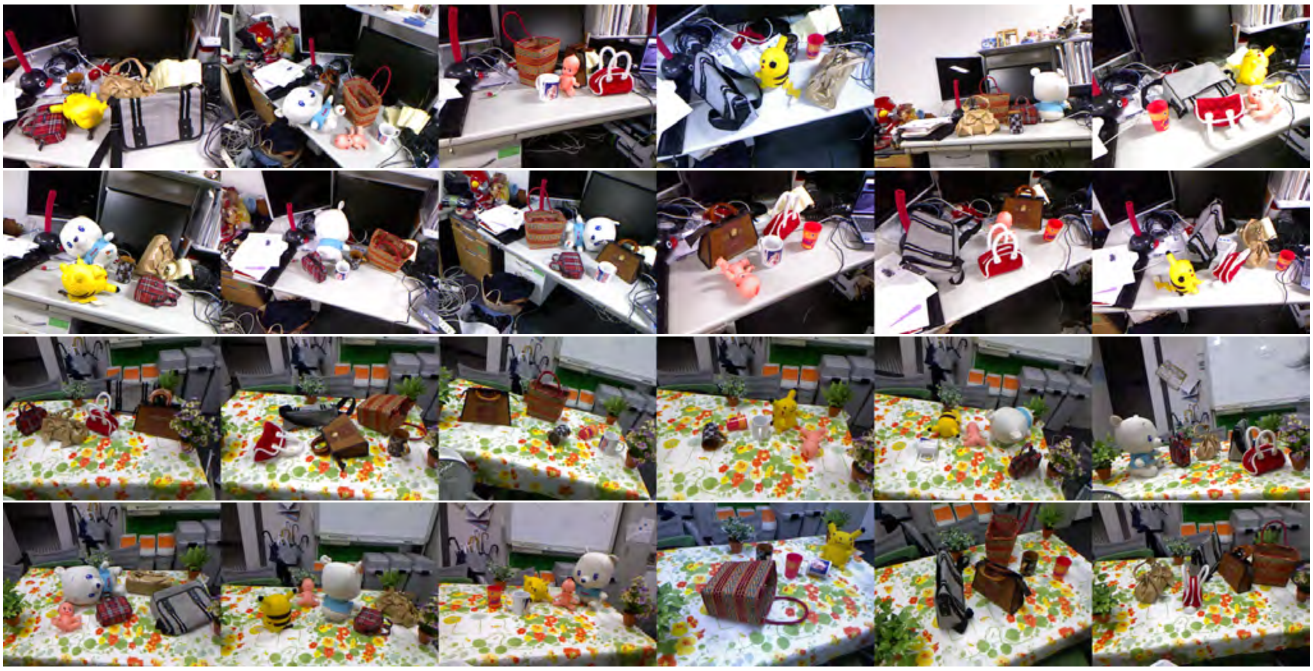


Figure 5. Example Images of the training scene (top two rows) and test scene (bottom two rows).

hoped, unfortunately. This was because training of some objects completely failed by finding the wrong local minimum, which greatly affected the total performance. Such failures are expected to be avoided by adding more training samples in various environments. The ROC curves for twelve objects respectively are shown in Fig. 9. The most successful results were given when object #4 (Fig. 4) was targeting, which recorded almost perfectly ideal ROC curves with both EM-DD and mi-SVM.

Some examples of object detection with the proposed features and EM-DD are shown in Fig. 10. The one most likely segment of each target object is shown. Successful results are shown in the middle row while failures are shown in the bottom. Various objects in varied positions seem to be correctly detected, while some failures are seen among similar segments (*e.g*. a white cup and a part of white fan).

6. Conclusion

We have proposed a method for joint learning of object classification and localization from weakly-labeled pairs of color and depth images, using geometry-based segmentation, 3D color features and Multiple Instance Learning. In this learning task, view point invariance of features and removal of bias in feature space caused by variance in object size are crucial. We showed in our experiment that rotation and scale invariant features recorded the highest performance.

Although a number of interesting results where the correct objects were detected in various positions in a new environment were achieved, we also had a few cases where the training of an object completely failed, so our method has room for improvement. We collected training samples in one specific environment, while more samples in different environments are expected to increase the probability that the correct features of objects are learned. Furthermore, more experiments to comparison to the proposed method with the well-known Spin Image [11] and its latest features (*e.g*. [26]) are required.

Acknowledgment

The authors thank Z. C. Marton from the Technische Universität München for his support on developing the software of object segmentation.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Proc. Neural Information Processing Systems 15 (NIPS)*, 2004.

[2] H. Arora, N. Loeff, D. Forsyth, and N. Ahuja. Unsupervised segmentation of objects using efficient learning. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007.

[3] M. B. Blaschko, A. Vedaldi, and A. Zisserman. Simultaneous object detection and ranking with weak supervision. In *Proc. Neural Information Processing Systems 23 (NIPS)*, 2010.

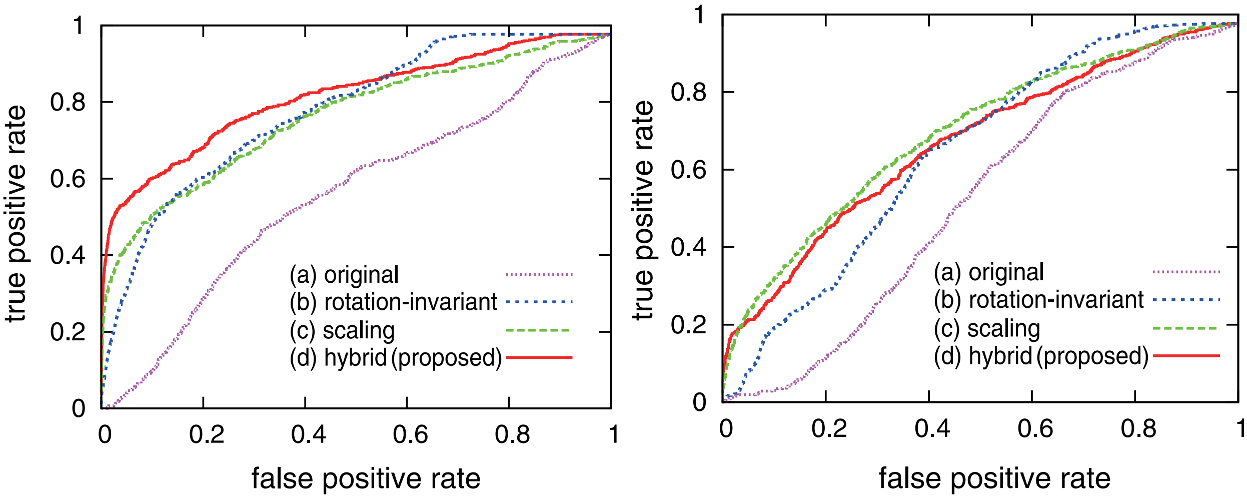


Figure 8. Comparison of average ROC curves with EM-DD (left) and mi-SVM (right).

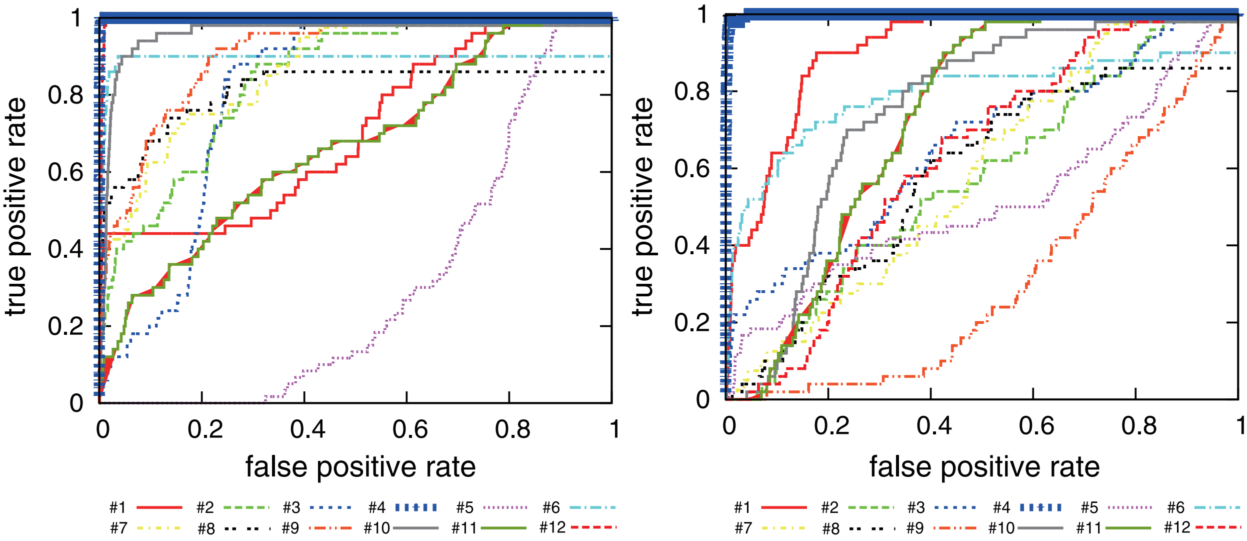


Figure 9. ROC curves for twelve target objects by the proposed features with EM-DD (left) and mi-SVM (right).

[4] L. Cao and L. Fei-Fei. Spatial coherent latent topic model for concurrent object segmentation and classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2007.

[5] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[6] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007.

[7] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proc. European Conference on Computer Vision (ECCV)*, 2006.

[8] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *Proc. European Conference on Computer Vision (ECCV)*, 2010.

[9] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2003.

[10] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object localization with stable segmentations. In *Proc. European Conference on Computer Vision (ECCV)*, 2008.

[11] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21:433–449, 1999.

[12] Jun Yang. MILL: A Multiple Instance Learning Library. `http://www.cs.cmu.edu/~juny/MILL`.

[13] A. Kanezaki, T. Suzuki, T. Harada, and Y. Kuniyoshi. Fast

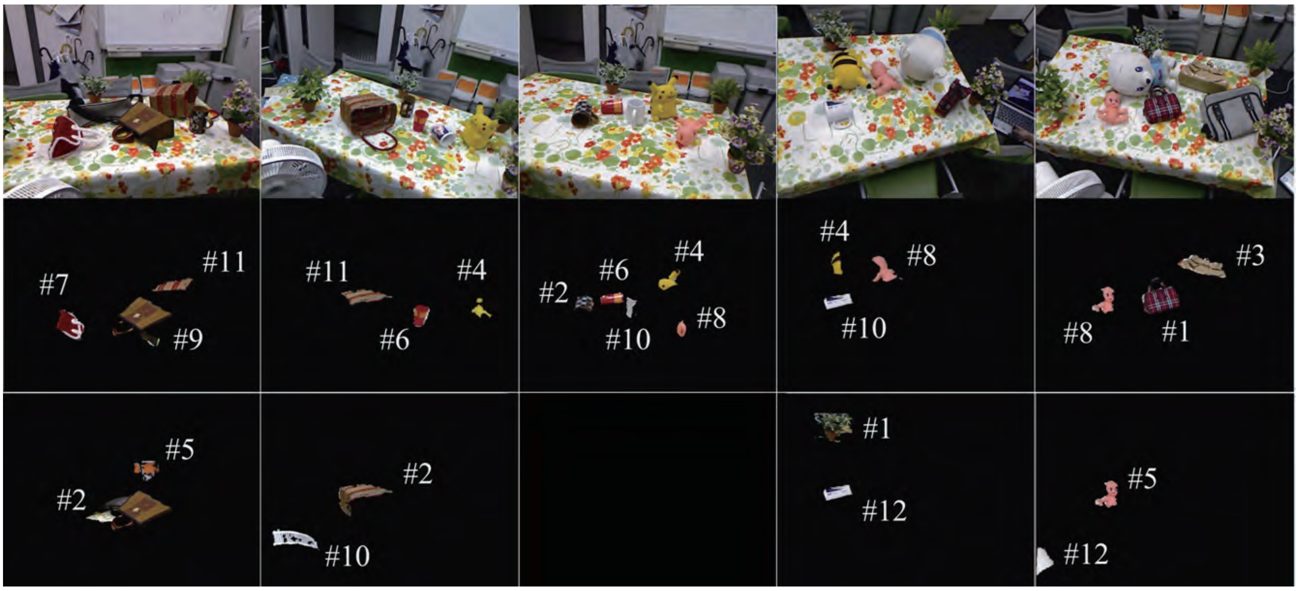


Figure 10. Example results of object detection. One segment with the largest probability of each target object (see Fig.4) in each scene is shown. The images of the test scenes are shown in the top, successful results are in the middle, and failures are in the bottom.

object detection for robots in a cluttered indoor environment using integral 3D feature table. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[14] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Proc. Neural Information Processing Systems 22 (NIPS)*, pages 961 – 969, 2009.

[15] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, May 2011.

[16] M. Everingham and L. van Gool and C. K. Williams and J. Winn and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/index.html.

[17] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. In *Proc. Neural Information Processing Systems 10 (NIPS)*, pages 570 – 576, 1998.

[18] Z. C. Marton, R. B. Rusu, and M. Beetz. On fast surface reconstruction methods for large and noisy datasets. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[19] O. M. Mozos, Z.-C. Marton, and M. Beetz. Furniture models learned from the www. *IEEE Robotics and Automation Magazine*, 18:22–32, 2011.

[20] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2009.

[21] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[22] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *Proc. European Conference on Computer Vision (ECCV)*, 2008.

[23] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2006.

[24] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision (special issue on vision and learning)*, 77:157–173, 2008.

[25] D. Singaraju and ReneVidal. Using global bag of features models in random fields for joint categorization and segmentation of objects. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[26] B. Steder, R. R. Bogdan, K. Konolige, and W. Burgard. Point feature extraction on 3d range scans taking into account object boundaries. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[27] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[28] J. Winn and N. Jojic. Locus: learning object classes with unsupervised segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2005.

[29] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *Proc. Neural Information Processing Systems 14 (NIPS)*, 2001.

[30] Y. Zhang and T. Chen. Weakly supervised object recognition and localization with invariant high order features. In *Proc. British Machine Vision Conference (BMVC)*, pages 47.1–11, 2010. doi:10.5244/C.24.47.