Weakly-supervised Multi-class Object Detection Using Multi-type 3D Features

Asako Kanezaki
Grad. School of Information Science and Technology
The University of Tokyo
kanezaki@isi.imi.i.u-tokyo.ac.jp

Yasuo Kuniyoshi
Grad. School of Information Science and Technology
The University of Tokyo
kuniyosh@isi.imi.i.u-tokyo.ac.jp

Tatsuya Harada
Grad. School of Information Science and Technology
The University of Tokyo
harada@mi.t.u-tokyo.ac.jp

ABSTRACT

We propose a weakly-supervised learning method for object detection using color and depth images of a real environment attached with object labels. The proposed method applies Multiple Instance Learning to find proper instances of the objects in training images. This method is novel in the sense that it learns multiple objects simultaneously in a way to balance the scores of each training sample across all object classes. Moreover, we combine 3D features considering different properties, that is, color texture, grayscale texture, and surface curvature, to improve the performance. We show that our method surpasses a conventional method using color and depth images. Furthermore, we evaluate its performance with our new dataset consisting of color and depth images with weak labels of 100 objects and various backgrounds.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition, Range data, Sensor fusion*; I.2.6 [**Learning**]: Analogies—*Concept learning, Parameter learning*

Keywords

multi-class object detection, weakly-supervised learning, 3D features

1. INTRODUCTION

Object detection is significantly in demand for intelligent autonomous systems such as robots and smart phones. However, there remains a crucial problem on how to prepare training datasets such as bounding boxes of objects in images, which is a time-consuming job for us. To facilitate the training process, weakly-supervised learning of object detection without objects' bounding boxes is widely researched. In this approach, the object labels (i.e. names) instead of their bounding boxes are attached to each training image if there is at least one instance in the image, so that the system should estimate the corresponding instances of the objects. Multiple Instance Learning (MIL) [2] is effective to deal with this problem, which introduces a concept of "bag" containing multiple instances. A bag is positive if it contains at least one positive instance, while it is negative otherwise. In the scenario of weakly-supervised object detection, a region (or segment) in an image corresponds to an instance, whereas an image corresponds to a bag. There are a couple of state-of-the-art methods [9, 10] that approached the weakly-supervised object detection using MIL and Deformable Part Models (DPM) [3]. However, DPM is not "deformable" enough to detect objects in a messy daily living environment because it uses rotation-variant 2D features.

The most important aspects in tasks of weakly supervised learning of object detection via MIL are **how to extract instances as reasonable candidates of target objects in each image (bag)**, and **how to acquire features' robustness against viewpoint and illumination variance**. The former is difficult especially when the segmentation of image regions is difficult because of cluttered background, and the latter is difficult especially when the postures or the scales of objects are various in 2D images. To solve these problems, a method that utilized 3D information both in extracting instances and computing features from color and depth images is proposed [5]. This method achieves segments as reasonable object candidates via normal estimation of 3D points, and also realizes object detection in various postures in a cluttered environment using rotation-invariant and scale-invariant 3D features. However, it uses only color texture features, which cannot handle illumination changes. Besides, it learns each target object independently, which lacks proper balancing of the scores of each instance across all objects.

Our objective is to learn multi-class object detectors with high distinguishability from weak supervision. We propose a new MIL method that learns multi-class object detectors while balancing the scores of each instance across all object classes. Moreover, we combine several 3D features that represent color textures, grayscale textures and surface curvatures, showing that they give better results than respective single-type features. The additional contribution is that we provide a new dataset consisting of color and depth images with weak labels of 100 objects in various cluttered backgrounds.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502159.

2. METHOD

The proposed method performs over-segmentation via normal estimation, creates multiple instances by combining various numbers of the primitive segments, extracts multiple voxel features, and solves MIL to obtain multi-class object detectors. To extract instances, we used the method proposed in [5]. The details of the proposed method of learning multi-class object detectors and feature extraction are shown below.

2.1 Learning of multi-class object detectors

Our method learns linear classifiers for multi-class object detection in a MIL setup. We use Passive Aggressive (PA) [1], which is an online (thus efficient and scalable) learning algorithm to obtain multi-class classifiers. Here, we describe PA algorithm for the multi-class setting. Let N be the number of target classes and D be the feature dimension. Let us define the weight matrix as $W \in \boldsymbol{R}^{D\times N}$, the k-th column of W as $\boldsymbol{w}_k$, and the k-th element of $f(\boldsymbol{x})$ as $f_k(\boldsymbol{x})$. Now we optimize the function $f_k(\boldsymbol{x}_t) = \boldsymbol{w}_k \cdot \boldsymbol{x}_t$ that gives the score of the t-th sample $\boldsymbol{x}_t$ belonging to the k-th target object. First, we pick up the r-th class when we observe a training sample $< \boldsymbol{x}_t, y_t >$:

$$r = \underset{y\neq y_t}{\operatorname{argmax}}\, \boldsymbol{w}_y^{(t)} \cdot \boldsymbol{x}_t. \tag{1}$$

Then the hinge loss, which represents the classification error, becomes $\max(0, 1 - f_{y_t}(\boldsymbol{x}_t) + f_r(\boldsymbol{x}_t))$. Introducing a slack variable ζ, the model parameters are updated by the following equation (PA-I in [1]).

$$W^{(t+1)} = \underset{W}{\operatorname{argmin}}\, \frac{1}{2}\|W - W^{(t)}\|^2 + C\zeta, \tag{2}$$

$$\text{s.t.}\quad \max\left(0, 1 - \boldsymbol{w}_{y_t} \cdot \boldsymbol{x_t} + \boldsymbol{w}_r \cdot \boldsymbol{x_t}\right) \leq \zeta, \zeta \geq 0, \tag{3}$$

while C is called "aggressiveness parameter", which controls the aggressiveness of fitting $\boldsymbol{w}$ to each training sample. Solving the Lagrange multiplier, we obtain the following update function.

$$\boldsymbol{w}_{y_t}^{(t+1)} = \boldsymbol{w}_{y_t}^{(t)} + \tau\boldsymbol{x}_t \quad \text{and} \quad \boldsymbol{w}_r^{(t+1)} = \boldsymbol{w}_r^{(t)} - \tau\boldsymbol{x}_t, \tag{4}$$

while τ is computed as follows.

$$\tau = \min\left\{C, \frac{1 - \boldsymbol{w}_{y_t}^{(t)} \cdot \boldsymbol{x_t} + \boldsymbol{w}_r^{(t)} \cdot \boldsymbol{x_t}}{2\|\boldsymbol{x}_t\|^2}\right\}. \tag{5}$$

To deal with weakly-supervised learning, the system should estimate which instance is positive as a target object. We alternate the estimation of positive instances and the update of model parameters. Here, we describe the processing "Phase 1" where we obtain the model parameters of each class independently by selecting positive instances randomly, and the processing "Phase 2" where we refine the model parameters of all the classes simultaneously with balancing samples' scores across the object classes. Although the PA solver itself is an existing technique, the design of overall processing in a MIL setup described below is novel.

Phase 1 – independent –

We setup samples for the k-th class as follows.

- label all the instances in the images which do not have k-th object label as "negative".
- pick up a single instance randomly out of all the images which have k-th object label, duplicate it γ times, and label them as "positive".

We set γ to one-tenth of the number of negative samples. Then we optimize model parameters by applying PA with $N = 2$. We repeat this cycle M times and adopt the best model with the minimum hinge loss.

Phase 2 – simultaneous –

Letting K be the number of target classes, we setup samples as follows.

- label all the instances in the images which do not have k-th object label as $k + K$.
- pick up a single instance with the highest score $f_k(\boldsymbol{x}_n)$ out of each image which has k-th object label and label them as k. However, if this instance has the label $k'(k' \neq k)$ and $f'_k(\boldsymbol{x}_n)$ is larger than $f_k(\boldsymbol{x}_n)$, we pick up another instance with the next highest score and label it as k.

Beginning with the model parameters obtained in "Phase 1", we optimize $2K$ model parameters $\boldsymbol{w}_k, b_k(k = 1, \ldots, 2K)$. Here, we define positive labels as $Y^+ = \{y^1, \ldots, y^K\}$ and negative labels as $Y^- = \{y'^1, \ldots, y'^K\}$. When an observed sample's label is $y_t \in Y^+$, we replace Equation (1) by the following one.

$$r = \underset{y\in\{y'_t\cup Y^+\setminus y_t\}}{\operatorname{argmax}}\, \boldsymbol{w}_y^{(t)} \cdot \boldsymbol{x}_t. \tag{6}$$

On the other hand, if an observed sample's label is $y'_t \in Y^-$, we set $r = y_t$. In this way, the decision surface is optimized among the target classes if a positive sample is observed, and otherwise it is optimized among one target class and all the rest. The former works for achieving distinguishability and the latter works for achieving detectability.

2.2 3D features

We use multiple 3D voxel features of different types; Circular Color Cubic Higher-order Local Auto Correlation (C^3-HLAC) features [6] that consider color texture, Intensity Spin Image [7] that considers gray-scale texture, and Global Radius-based Surface Descriptor (GRSD [8]) that considers surface curvature. The details of how to extract respective features are described below.

C^3-HLAC

C^3-HLAC descriptor [6] is represented by color 3D textures. Let $\boldsymbol{x}$ be the position of a 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 $\boldsymbol{f}(\boldsymbol{x}) \equiv [r_1\ r_2\ g_1\ g_2\ b_1\ b_2]^T \in \mathbb{N}^6$. Note that $\boldsymbol{f}(\boldsymbol{x})$ becomes a zero vector when the voxel has no measure point inside. Similarly to [5], we use the redefined rotation-invariant C^3-HLAC descriptors. Letting the color property of a voxel obtained when $r(\boldsymbol{x})$, $g(\boldsymbol{x})$ and $b(\boldsymbol{x})$ are binarized in a preprocessing be $\boldsymbol{f'}(\boldsymbol{x})$, a rotation-invariant version of C^3-HLAC descriptor is calculated as presented below.

$$\boldsymbol{z}_1 = \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{z}_2 = \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})^T, \boldsymbol{z}_3 = \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{a})^T, \tag{7}$$

$$\boldsymbol{z}_4 = \boldsymbol{f'}(\boldsymbol{x}), \boldsymbol{z}_5 = \boldsymbol{f'}(\boldsymbol{x})\boldsymbol{f'}(\boldsymbol{x})^T, \boldsymbol{z}_6 = \boldsymbol{f'}(\boldsymbol{x})\boldsymbol{f'}(\boldsymbol{x}+\boldsymbol{a})^T. \tag{8}$$

Therein, $\boldsymbol{a}$ is a displacement vector for a neighboring voxel, in which at least one of the x, y, and z coordinates is one. Eventually, a C^3-HLAC descriptor is obtained as a 117-dimensional vector by concatenating all elements in Equation (7) and Equation (8).

GRSD

GRSD [8] is represented by surface curvature. This descriptor is calculated by counting transitions from a Radius-based Surface Descriptor (RSD) of a voxel to that of a neighboring voxel. The RSD consists of the following six surface types defined by the highest curvature r_{min} and the lowest curvature r_{max} among voxels within a sphere of a certain radius: "planes" (large r_{min}), "cylinders" (medium r_{min}, large r_{max}), "edges" (small r_{min} and r_{max}), "rims" (small r_{min}, medium to large r_{max}), "spheres" (similar r_{min} and r_{max}), and "free space". Distinguishing all pairs of these six surface types except for the pair of "free space" and "free space", GRSD is obtained as a 20 $(= 6 \cdot 7/2 - 1)$-dimensional histogram.

C^3-HLAC and GRSD are originally defined as global features, where a feature vector is obtained by summing up local descriptors in a whole point cloud.

Intensity Spin Image

The Intensity Spin Image descriptor [7] is represented by a grayscale texture. This descriptor is based on the idea of Spin Image [4]. Spin Image is a 2D histogram of the projection angles α and β of surrounding points to the tangent plane of a point. Intensity Spin Image is a 2D histogram of the distances between a point and its surrounding points within a certain distance and their intensity level. Letting the quantization level of the distance between a point and its surrounding point be L_1 and that of its intensity be L_2, the dimension of this descriptor becomes $L_1 \cdot L_2$. In this work, we set L_1 to four and L_2 to five. Therefore the dimensions become twenty. We set the maximum distance to 50 mm. The Intensity Spin Image is originally a 2D image descriptor and is extracted from affine normalized patches. However, we skip this process because we use 3D point clouds, which are fundamentally affine invariant. Similarly to C^3-HLAC and GRSD, we sum up the Intensity Spin Image descriptors in a whole point cloud of an instance to obtain a global feature vector. Although C^3-HLAC, Intensity Spin Image, and GRSD are 3D features of different types represented respectively by color textures, grayscale textures and surface curvatures, they are computed efficiently in a uniform manner that extracts local descriptors of 3D points and sums them up to obtain a global feature vector.

3. RESULTS

3.1 Comparison to a conventional method

We used a color and depth image dataset[1] [5] to evaluate our method. Mean Average Precision (MAP) values of all the twelve target objects with EM-DD [11], which is a conventional MIL method, and with the proposed method are shown in Table 1. We set the number of the initial instances M to 100 both for EM-DD and our method. The putative detection was considered correct if the intersection of its bounding box with the ground-truth bounding box was larger than 50% of their union, as in Pascal VOC[2]. The results with C^3-HLAC, GRSD, Intensity Spin Image, and their combination are shown. When we permit a single instance to be attached with multiple labels so as to evaluate the performance of each class independently ("inclusive"), our method was superior to EM-DD with C^3-HLAC but was inferior with other features. However, when we attach to each instance a single label with the top score among all the target objects ("exclusive"), the performance with the combination features was also superior with our method to EM-DD. Because our method balances the scores of each instance across all the target objects while EM-DD learns each object detector independently, our method is more proper in the "exclusive" case where the distinguishability is important. Note that C^3-HLAC and EM-DD were used in a conventional work [5], and that the MAP in the "exclusive" case with our method (with "Combined" features and "Ours" learning method) was nearly twice larger than that with C^3-HLAC and EM-DD.

Table 1: MAP values in the 1st experiment.

Features	inclusive		exclusive	
	EM-DD	Ours	EM-DD	Ours
C^3-HLAC	0.0979	**0.1146**	0.0679	**0.0886**
Intensity-SI	**0.0530**	0.0411	**0.0415**	0.0349
GRSD	**0.0615**	0.0140	**0.0519**	0.0090
Combined	**0.1395**	0.1373	0.0932	**0.1227**

Table 2: MAP values in the 2nd experiment.

Features	inclusive		exclusive	
	Phase 1	Phase 2	Phase 1	Phase 2
Combined	0.218	0.217	0.222	0.250

3.2 Evaluation on a new dataset

We provide a new dataset[3] consisting of color and depth images inside a laboratory room captured by a Kinect sensor. There are 1,398 pairs of color and depth images for training and 315 pairs of them for testing (Fig. 1). The conventional dataset [5] used in Section 3.1 has a fixed background respectively for training and testing, whereas our dataset includes various backgrounds all around the laboratory room. Moreover, there are 100 object labels in our dataset while there are only 12 in the conventional one [5]. We used the combination features of C^3-HLAC, GRSD and Intensity Spin Image. We set M to 300 in this experiment.

Three instances with the highest scores of each target object are shown in Fig. 2. The ground truth objects are shown in the top row and the first, second and third ranked instances are shown below. The segments of the objects with 14 lowest hinge loss are shown from the first column to the 14th, and those with 6 highest hinge loss are shown from the 15th column to the end. Whereas there were 60 out of 100 objects whose regions were correctly ranked as top when we used the object detectors learned in "Phase 1", there were 70 when we used the object detectors learned in "Phase 2". It implies that balancing the scores of each instance across all the target objects is important to find correct objects in a weakly-supervised learning. MAP values of all the 100

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

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[3] http://www.mi.t.u-tokyo.ac.jp/kanezaki/color_depth_dataset_100

Figure 1: Example Images of the new dataset used in the 2nd experiment.

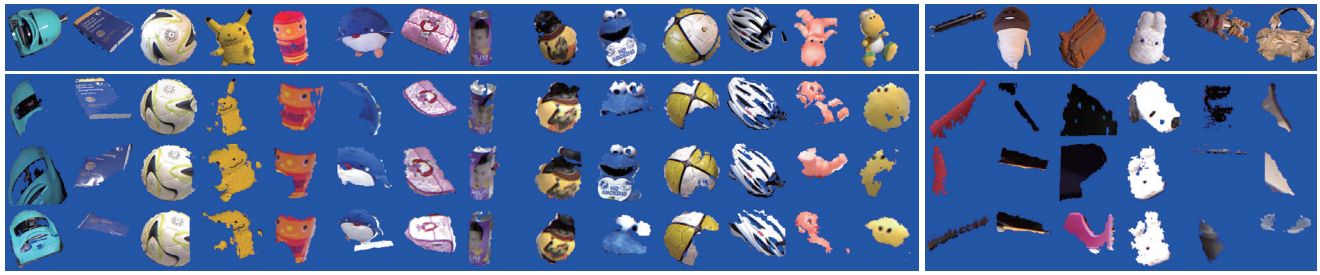

Figure 2: Obtained segments of the training dataset. Ground-truth segments are shown in the first row and the segments obtained with 1st, 2nd and 3rd ranks are shown in the rest rows. The segments of the objects with 14 lowest hinge loss are shown from the first column to the 14th, and those with 6 highest hinge loss are shown from the 15th column to the end.

target objects (with test data) are shown in Table 2. There is almost no difference between "Phase 1" and "Phase 2" in the "inclusive" case, however, the performance with "Phase 2" surpasses that of "Phase 1" in the "exclusive" case, where we attach to each instance a single label with the top score among all the target objects.

4. CONCLUSION

We proposed a weakly-supervised learning method that uses color and depth images attached with object labels to learn multi-class object detectors. Our method combines three different types of 3D features efficiently to achieve better performance. Moreover, this method is novel in that it learns multiple objects simultaneously in a way to balance the scores of each training sample across all the object classes. We showed that our method outperforms a conventional method using color and depth images with the dataset provided in its work. Moreover, we provided a new dataset with more object labels and various backgrounds, which we used to evaluate our method. We showed that our method is effective especially when we attach to each instance a single label with the top score among all the target objects, which is the case where the distinguishability is important.

5. REFERENCES

[1] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Machine Learning Research*, 7:551–585, 2006.

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89, 1997.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *Trans. on PAMI*, 32(9), 2010.

[4] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *Trans. on PAMI*, 21:433–449, 1999.

[5] A. Kanezaki, T. Harada, and Y. Kuniyoshi. Scale and rotation invariant color features for weakly-supervised object learning in 3D space. In *Proc. ICCV Workshop on 3D Representation and Recognition (3dRR-11)*, 2011.

[6] A. Kanezaki, T. Suzuki, T. Harada, and Y. Kuniyoshi. Fast object detection for robots in a cluttered indoor environment using integral 3D feature table. In *Proc. ICRA*, 2011.

[7] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *Trans. on PAMI*, 27:1265–1278, 2005.

[8] Z.-C. Marton, D. Pangercic, R. B. Rusu, A. Holzbach, and M. Beetz. Hierarchical object geometric categorization and appearance classification for mobile manipulation. In *Proc. Int. Conf. on Humanoid Robots*, 2010.

[9] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Proc. ICCV*, 2011.

[10] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *Proc. ICCV*, 2011.

[11] Q. Zhang and S. A. Goldman. EM-DD: An improved multiple-instance learning technique. In *Proc. NIPS*, 2001.