3D Object Detection with Multiple Kinects

Wandi Susanto, Marcus Rohrbach, and Bernt Schiele

Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract. Categorizing and localizing multiple objects in 3D space is a challenging but essential task for many robotics and assisted living applications. While RGB cameras as well as depth information have been widely explored in computer vision there is surprisingly little recent work combining multiple cameras and depth information. Given the recent emergence of consumer depth cameras such as Kinect we explore how multiple cameras and active depth sensors can be used to tackle the challenge of 3D object detection. More specifically we generate point clouds from the depth information of multiple registered cameras and use the VFH descriptor [20] to describe them. For color images we employ the DPM [3] and combine both approaches with a simple voting approach across multiple cameras.

On the large RGB-D dataset [12] we show improved performance for object classification on multi-camera point clouds and object detection on color images, respectively. To evaluate the benefit of joining color and depth information of multiple cameras, we recorded a novel dataset with four Kinects showing significant improvements over a DPM baseline for 9 object classes aggregated in challenging scenes. In contrast to related datasets our dataset provides color and depth information recorded with multiple Kinects and requires localizing and categorizing multiple objects in 3D space. In order to foster research in this field, the dataset, including annotations, is available on our web page.

1 Introduction

Accurate 3D localization and categorization of objects is an important ingredient for many robotic, assisted living, surveillance, and industrial applications. Consumer depth cameras, such as Microsoft's Kinect, provide depth information based on an active structured light sensor combined with color images. The additional depth information simplifies estimating the position in 3D real world coordinates; however, a single (color and depth) camera is still limited especially in highly cluttered scenes, with object occlusion and objects which are difficult to distinguish from a single view. While a plate and bowl might not be distinguishable from above in a color image (see e.g. Fig. 1, 1^{st} column), additional cameras (2^{nd} column) and depth information (3^{rd} and 4^{th} column) can help to resolve these problems. Given the low price (and likely increasing accuracy in near future), multiple devices observing the same area are a viable and likely setting for smart homes or surveillance. While the benefit has been argued before

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 93–102, 2012.

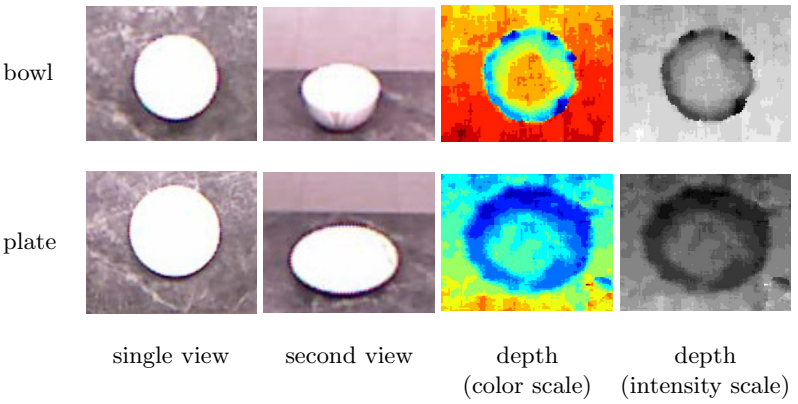


Fig. 1. Objects (here bowl and plate) are sometimes difficult to distinguishable from a certain single view (1^{st} column). Having additional views (2^{nd} column) or depth (3^{rd} and 4^{th} column) can frequently resolve the problem.

we find this scenario of combining multiple views and active depth sensors underrepresented in related work. Multiple works have shown that object recognition can be improved using stereo/depth information [17,1,12,8] or by using multiple (color) images [7,19]. However, fusing information from multiple *depth* cameras for object detection has hardly been addressed. An exception is [14], which however uses the depth information only to estimate if an object is occluded but does not combine depth from multiple views and does not use depth for representing objects. Given the emergence of low-cost depth sensors we would like to intensify research in this area and thus propose a dataset for multi-camera object recognition to explore how to use color and depth information from multiple views.

Our dataset is recorded in a kitchen scenario which we found to be typical to contain large amounts of similar objects, frequently occluding each other. We have four Kinects attached to the ceiling as can be seen in Fig. 3. The idea of our approach is to combine strong visual cues from color images by using the deformable part model (DPM) [3] together with the viewpoint feature histogram (VFH) [20] extracted from depth based point clouds. We first detect the object in all views with DPM and vote for the location of objects in 3D space, using the depth information. Additionally we validate the 3D location with the VFH, extracted on the registered point clouds from all cameras.

Our contributions are as follows: First, we show the benefits of multiple depth cameras for multi-class object detection. To our knowledge we are the first to show the benefit of multiple cameras for point cloud-based object recognition. As the second main contribution we recorded a novel, challenging dataset of 9 object classes with four Kinects to evaluate our approach. As we discuss in our related work section, this setting and kind of dataset has hardly been explored previously. Additionally we show improved performance for classification on multi-camera view point clouds and improvements over related work for object detection on the Multi-View RGB-D Object dataset [12].

The remaining paper is structured as follows: We first discuss related approaches, scenarios, and datasets in Section 2. Next we introduce our approach (Section 3) and our newly recorded dataset (Section 4). For evaluation we first report the performance of our chosen feature on the RGB-D dataset (Section 5.1) and then examine the detection performance of our proposed approach on the multiple depth camera setting (Section 5.2). We conclude with a summary and ideas for future work in Section 6.

2 Related Work

We organize this section into related feature representations for depth information, multi-camera approaches, and available datasets recorded with (consumer) depth cameras.

Several ways to represent depth information and to extract feature representations have been proposed. The most common or basic way is to represent depth information as an image of depth values. On such depth images one can extract features, such as HOG [2], which have shown good performance on color images. Depending on the application HOG on depth images was found to perform worse [17] (for pedestrian recognition) or better [12] (for object recognition) than on 2D intensity images but in all cases a combination of both yielded significantly better results indicating the obvious complementary information contained in depth and color/intensity images. Combination of depth and intensity has also shown to be beneficial for template based method allowing real time object detection [8]. Another direction is to compute features on 3D point clouds. Spin image [11], VFH [20], or 3D shape context [5] are prominent examples. An alternative to represent 3D objects are 3D meshes: The 3D surf descriptor in combination with spatial pyramids in 3D achieves significant performance improvements for 3D object classification [16]. Depth has also been used to constraint the size of objects for improved detection [22]. Laser range scanners provide more accurate but less dense information and can be used to define multi-modal 3D features [6]. In our work we similarly want to benefit from the complimentary information contained in color and depth images. We choose point clouds for representing depth, as point clouds allow to easily integrate depth information from multiple views. We extract the VFH descriptor from the point clouds which has shown good performance [20].

With respect to related approaches, we first would like to note that this work focuses on a multi-camera setup, which is sometimes denoted as multi-view [4]. In most cases however, *multi-view* refers to multiple views of an object category. Although our work looks at multi-view categorization we additionally want to explore how beneficial multiple depth cameras are for object detection in 3D space. There exists several prior works which use registered color/intensity images to improve 3D object detection: E.g. [1] shows that using two cameras can improve performance. [19] detects cars, buses, and people on street scenes with six cameras, while [4] localizes objects in 3D from four X-Ray cameras. [4] uses, similar to us, a voting based approach to combine multiple views. Multiple camera views have also been used for marker-less motion capture, e.g. [13] is able

to distinguish two interacting people in a studio (green box) scenario. 3D object detection can be achieved from a single monocular camera using video and structure from motion [14,25].

Our work is most similar to [7], that detects and classifies objects in an indoor environment using multiple camera views. In a follow up work the authors use depth information to estimate which object is an occluder of another object [14]. In contrast to this we build a feature representation on the depth information and use not only depth from each view individually but use registered 3D point cloud data from multiple cameras in addition to color images. Furthermore we distinguish 9 object categories while [7,14] only distinguish 4 and 2, respectively.

Fusing point cloud videos from a single moving Kinect has previously been used to build high-resolution 3D representations [9] but without the direction of object detection but rather interaction and rendering. With a similar use case, [24] uses multiple Kinects to increase the interaction space.

Finally we will discuss the most relevant and related datasets: The Multi-View RGB-D Object dataset provides object classification and object detection task [12]. The dataset includes 51 classes with color image, depth image, and 3D point cloud data. While for the classification challenge there exist multiple views per object, the detection scenes are only recorded from a single camera/view. Another dataset of 75 scenes with over 50 classes was recently proposed [10]. In contrast to these datasets our dataset contains test scenes recorded by multiple depth (and RGB) cameras.

The UBC VRS dataset [7] is most similar to ours. In an indoor detection setting, recordings from multiple view-points of four object categories in 26 scenes have to be distinguished. In [14] a few scenes with Kinect recordings of mugs and bowls have been added. In contrast to our dataset the recordings are performed from a mobile robot platform exploring the room, while our cameras are installed statically. Our dataset contains more object categories (9) but a similar number of scenes (33).

3 Approaches for Recognition with Multiple Kinects

In this section we first introduce how we perform object classification on point clouds from a single and multiple cameras, and then we discuss several approaches to perform 3D object detection.

3.1 Object Classification on 3D Data

We represent depth information as point clouds, i.e. as a set of points with 3D coordinates. This allows easy aggregation if depth information is available from multiple registered cameras views.

For a given bounding box we extract the 3D point cloud descriptor viewpoint feature histogram (VFH) [20]. We do this for the single-camera and multi-camera setting. For the multi-camera setting we first register multiple object point clouds

of different cameras into a single point cloud (see Section 4.2). Only for experiments on the RGB-D dataset we combine multiple views by averaging classifiers scores due to the unavailability of registration information.

3.2 3D Object Detection

We use the deformable part model (DPM) [3] to detect objects in 2D color images. To detect objects in 3D we distinguish three approaches. The first is a baseline based on DPM, in the second we verify single camera DPM detections with the 3D VFH object classifier, while the third combines DPM detections from multiple RGB cameras and depth-based VFH verification.

sDPM (single-camera DPM): As a baseline we evaluate the detection score from a single camera and compute the 3D bounding box by using the 3D information from the depth image of the same camera.

sDPM+mVFH (single-camera DPM + multi-camera VFH): In this approach we first detect the object in the 2D color image with DPM and back-project the detected bounding box into 3D space as described above. We then verify the detection bounding box by using the 3D VFH object classifier computed on the registered point cloud from multiple cameras. For a given hypothesis h we combine the DPM scores $s_{sDPM}(h)$ and the nearest neighbor distance of the 3D VFH classifier $d_{mVFH}(h)$ with the following function, setting $\alpha = 2$ and $\beta = 100$ to provide a similar score range.

$$s_{sDPM+mVFH}(h) = s_{sDPM}(h) + \alpha \exp\left(-\frac{d_{mVFH}(h)}{\beta}\right) \tag{1}$$

mDPM+mVFH (multi-camera DPM + multi-camera VFH): Finally we want to benefit from both, multi-camera color and multi-camera depth information. We combine the sDPM+mVFH hypotheses from multiple cameras by voting for a 3D location.

The single-camera hypotheses are given in 2D bounding boxes. Fusing these 2D bounding boxes in 3D space forms polyhedrons, which are not simple to work with. Therefore, we use the center point of each object, which is defined by the center pixel of the 2D bounding box and the depth value at the corresponding pixel in the depth image. We transform all hypotheses from different cameras into a single 3D coordinate system. Each hypothesis votes for all its neighbors $\mathcal{N}$, which are closer than a threshold t with 50% of its own score. Thus the accumulated score of a hypothesis is given by:

$$s_{mDPM+mVFH}(h) = s_{sDPM+mVFH}(h) + 0.5 \sum_{i \in \{j \in \mathcal{N}(h) | \, \|h-j\| \leq t\}} s_{sDPM+mVFH}(i) \tag{2}$$

We then perform 3D non-maximum suppression with the distance threshold t in 3D space to yield the final set of hypotheses. We set $t = 10cm$ in all experiments.

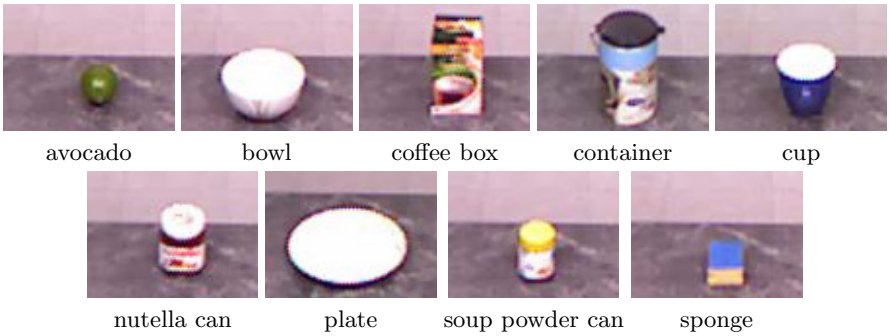


Fig. 2. Sample object images for each class in the MPII Multi-Kinect dataset

4 MPII Multi-Kinect Dataset

Our *MPII Multi-Kinect* dataset is collected with four Kinect cameras at training and test time in the same kitchen as [18]. We recorded RGB color and depth images as well as 3D point clouds per camera and registered multi-view point clouds computed from the depth images. The dataset consists of 9 classes of kitchen objects shown in Fig. 2. Class *bowl* and *cup* have two instances, all others only one. Each class in the dataset is captured in different locations and poses. The dataset consists of two parts. In the first part (classification challenge) only one object is present at a time. There are 560 shots taken from 4 cameras giving a total of 2240 pairs of color and depth images (Note we have a 10^{th} class, namely orange, is this first part of the dataset). The second part (detection challenge) is again recorded from 4 camera views. Here we recorded six to ten objects in each of a total of 33 scenes, which partially/fully occlude each other, see Fig. 3 for an example scene. We annotate each object with a bounding-polygon, providing a segmentation of the object in all RGB camera views. The dataset is publicly available on our web page including the annotations. In our detection experiments we use the camera installed on top of the scene only for ground truth as it contains no occlusions. We evaluate the multi-camera detections with a distance-based criterion with a threshold of 20 cm between the detection center and the ground truth center in 3D space (the largest extension of objects in the dataset range from 10-25 cm).

4.1 Data Preprocessing

We found the depth data provided by Kinect to be very noisy and incomplete. Noisy refers to variations between 2 to 4 different discrete depth levels. We smooth the depth data by taking the mean over 9 depth frames. For incomplete regions we median-filter with a 5x5 pixel window.

Groundtruth camera Camera A Camera B Camera C

Fig. 3. Example scene from four different views in our MPII Multi-Kinect dataset

4.2 Calibration

We calibrate the extrinsic matrix between different Kinect cameras using ICP from the PCL [21]. We choose one camera as reference to which we calibrate the other cameras. Since the cameras look at the scene from very different viewpoints, we initialize the transformation matrix between two cameras by manually rotating the point clouds before running ICP. Due to the noisy, inaccurate depth data, and position of our Kinect cameras, we found the different calibrated views can still disagree by up to about 13 cm in 3D space. The active light patterns of multiple Kinects might potentially interfere with each other. However, during our experiments we could not observe any disturbance between different Kinects, most likely due to the large angle between the different Kinects.

5 Evaluation

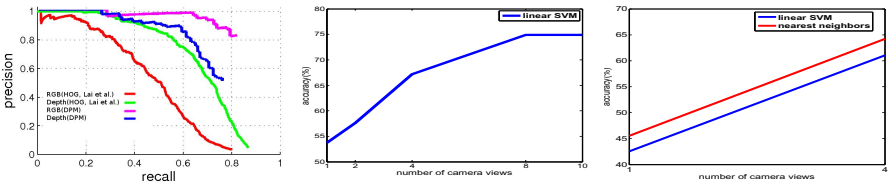
First, in Section 5.1, we present four initial experiments: (1) we extract the VFH descriptor from point clouds and compare it to related work on the RGB-D dataset [12]; (2) we repeat this for the DPM detector on color image scenes; (3) we show the benefit of multi-camera information for classification on the RGB-D dataset; and (4) on our dataset using registered point clouds.

In Section 5.2, after having shown the effectiveness of our individual components, we test our detection approaches (detailed in Section 3) using multi-view, depth, and color information on our multi-camera color and depth dataset.

5.1 Initial Experiments

We first evaluate object classification performance using VFH as a descriptor for point clouds on the RGB-D dataset [12]. We achieve a classification accuracy of 57.8% with nearest-neighbors [15] and 59.4% with χ^2 kernel SVM [23], while [12] reports 64.7% using Spin Image [11] with a Gaussian kernel SVM. We note that there is a drop in performance from Spin Image to VFH, but we found VFH faster to compute and to work very well on registered point clouds from multiple camera views as we will see below.

Second, we evaluate the DPM [3] detector on the RGB-D dataset. Not surprisingly DPM outperforms HOG (with sliding window) used in [12] on the depth (see blue vs. green in Fig. 4(a) for class *mug*) and on the color/intensity (violet vs. red) channel. In the following we use DPM as detector on the color image.



(a) DPM vs. HOG detection, class *mug*, RGB-D dataset [12].

(b) multi-camera classification on RGB-D dataset [12].

(c) multi-camera classification on our dataset

Fig. 4. Initial experiments, evaluating features and multiple cameras

In a third experiment we use multi-view object information in the classification RGB-D dataset by taking the mean of the scores from different views. There is a significant improvement from 53.7% (1 view) to 74.9% (8 views) accuracy (Fig. 4(b)) and hardly any improvement for more than 8 views.

The second object classification evaluation is performed on our MPII Multi-Kinect dataset, classification challenge. Here we register the point clouds from different views into a single point cloud. We improve the classification accuracy from 42.6% (1 view) to 61.0% (4 views) with linear SVM and from 45.5% (1 view) to 64.2 (4 views) with a nearest neighbors classifier, as can be seen in Fig. 4(c). In the following we will use the nearest neighbors classifier for VFH.

5.2 Detection with Multiple Kinects

For the 3D object detection challenge in our MPII Multi-Kinect dataset, we use three cameras for detection and one as ground truth as shown in Fig. 3.

In Table 1 we compare a DPM-baseline to our approaches introduced in Sec. 3 for all classes. We first examine the DPM-baseline present in the first three columns, comparing the mean performance in the last row. Evaluated on the color image of a single camera DPM achieves a detection mean average precision (AP) of 49.1%, 74.2%, and 61.8% for cameras A, B, and C, respectively. The benefit of camera B (see Fig. 3) is of being the one with steepest angle witnessing the least occlusion of all cameras.

Our first multi-camera approach, sDPM+mVFH, uses still single camera DPM, but verifies hypotheses with the multi-camera 3D object classifier. This significantly improves the performance of each camera to 54.9%, 85.0%, and 68.4% by 5.8%, 10.8%, and 6.6%, respectively. This consistent improvement shows the strength of adding depth information aggregated from multiple cameras.

In our second multi-camera approach we combine all cameras from the previous approach by voting for a 3D location and achieve 92.4% which is an increase by 7.4% compared to the best sDPM+mVFH (85.0% AP) and an increasy by 18.2% compared to the best DPM-baseline (74.2% AP). This shows again the benefit of multiple cameras.

In Fig. 5 we show the detection of a highly occluded *sponge* fails for the single-camera DPM (a), while we can detected it using multi-camera and depth information (b). However, the multi-view approach sometimes does not provide very accurate detections (c).

(a) Single-camera DPM fails to detect an occluded sponge.

(b) Multi-camera voting approach successfully detects the occluded sponge.

(c) Multi-camera voting locates the sponge inaccurately.

Fig. 5. Multi-camera voting approach successfully detects an occluded object

Table 1. 3D detection using multiple cameras, color, and depth (AP in %). Single-camera DPM, single-camera DPM + multi-camera VFH, multi-camera DPM+VFH.

	sDPM (baseline)			sDPM+mVFH			mDPM+mVFH
Class	cam A	cam B	cam C	cam A	cam B	cam C	multi-cam voting
avocado	50.7	85.0	56.9	54.6	100.0	81.4	100.0
bowl	3.0	75.1	48.5	2.1	76.1	51.2	87.0
coffee box	68.7	69.1	63.0	76.7	78.5	67.2	80.0
container	85.7	76.5	48.0	89.3	89.8	42.1	89.6
cup	35.7	77.3	72.6	33.2	83.6	73.5	100.0
nutella can	80.9	83.4	83.6	82.0	83.8	84.1	89.2
plate	15.4	78.1	80.2	14.6	80.2	82.5	90.2
soup powder can	70.2	65.4	62.7	69.7	75.3	70.5	98.5
sponge	31.3	58.1	40.9	71.7	97.5	63.2	97.0
mean	49.1	**74.2**	61.8	54.9	**85.0**	68.4	**92.4**

6 Conclusion

In this work we explored the benefit of using multiple Kinects providing color and depth information, compared to just single color cameras. In comparison to the DPM baseline we found adding point clouds from multiple depth cameras we achieve a performance increase from 74.2% AP to 85.0% for the best camera view. Adding additionally all RGB cameras further increases the performance to 92.4% improving by 18.2% AP over the best single RGB camera view. This shows that multi-view and depth information can be very beneficial for 3D object detection.

As part of future work we plan to add the recently proposed and improved point cloud descriptors [16]. We are also interested in the important direction of human-object interactions which will further increase occlusion and thus require multiple cameras.

References

1. Coates, A., Ng, A.Y.: Multi-camera object detection for robotics. In: ICRA (2010)
2. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: CVPR (2005)
3. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2010)
4. Franzel, T., Schmidt, U., Roth, S.: Object Detection in Multi-view X-Ray Images. In: Pinz, A., Pock, T., Bischof, H., Leberl, F. (eds.) DAGM/OAGM 2012. LNCS, vol. 7476, pp. 144–154. Springer, Heidelberg (2012)
5. Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J.: Recognizing Objects in Range Data Using Regional Point Descriptors. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3023, pp. 224–237. Springer, Heidelberg (2004)

6. Gould, S., Baumstarck, P., Quigley, M., Ng, A.Y., Koller, D.: Integrating visual and range data for robotic object detection. In: M2SFA2 (2008)
7. Helmer, S., Meger, D., Muja, M., Little, J.J., Lowe, D.G.: Multiple Viewpoint Recognition and Localization. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 464–477. Springer, Heidelberg (2011)
8. Hinterstoisser, S.H.S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: ICCV (2011)
9. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In: UIST (2011)
10. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3-D object dataset: Putting the kinect to work. In: ICCV (2011)
11. Johnson, A., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. PAMI (1999)
12. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: ICRA (2011)
13. Liu, J., Shah, M., Kuipers, B., Savarese, S.: Cross-view action recognition via view knowledge transfer. In: CVPR (2011)
14. Meger, D., Wojek, C., Little, J.J., Schiele, B.: Explicit occlusion reasoning for 3D object detection. In: BMVC (2011)
15. Muja, M., Lowe, D.: Fast approximate nearest-neighbors with automatic algorithm configuration. In: VISAPP (2009)
16. Redondo-Cabrera, C., López-Sastre, R.J., Acevedo-Rodríguez, J., Maldonado-Bascón, S.: Surfing the point clouds: Selective 3D spatial pyramids for category-level object recognition. In: CVPR (2012)
17. Rohrbach, M., Enzweiler, M., Gavrila, D.M.: High-Level Fusion of Depth and Intensity for Pedestrian Classification. In: Denzler, J., Notni, G., Süße, H. (eds.) DAGM 2009. LNCS, vol. 5748, pp. 101–110. Springer, Heidelberg (2009)
18. Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., Schiele, B.: Script Data for Attribute-Based Recognition of Composite Activities. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 144–157. Springer, Heidelberg (2012)
19. Roig, G., Boix, X., Shitrit, H.B., Fua, P.: Conditional random fields for multi-camera object detection. In: ICCV (2011)
20. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D Recognition and Pose Using the Viewpoint Feature Histogram. In: IROS (2010)
21. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: ICRA (2011)
22. Saenko, K., Karayev, S., Yia, Y., Shyr, A., Janoch, A., Long, J., Fritz, M., Darrell, T.: Practical 3-D object detection using category and instance-level appearance models. In: IROS (2011)
23. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. In: CVPR (2010)
24. Wilson, A.D., Benko, H.: Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In: UIST (2010)
25. Wojek, C., Roth, S., Schindler, K., Schiele, B.: Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 467–481. Springer, Heidelberg (2010)