Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes

Stefan Hinterstoisser[1], Stefan Holzer[1], Cedric Cagniart[1], Slobodan Ilic[1],
Kurt Konolige[2], Nassir Navab[1], Vincent Lepetit[3]

[1]Department of Computer Science, CAMP, Technische Universität München (TUM), Germany
[2]WillowGarage, Menlo Park, CA, USA
[3]École Polytechnique Fédérale de Lausanne (EPFL), Computer Vision Laboratory, Switzerland
{hinterst,holzerst,cagniart,slobodan.ilic,navab}@in.tum.de,
konolige@willowgarage.com, vincent.lepetit@epfl.ch

Abstract

We present a method for detecting 3D objects using multi-modalities. While it is generic, we demonstrate it on the combination of an image and a dense depth map which give complementary object information. It works in real-time, under heavy clutter, does not require a time consuming training stage, and can handle untextured objects. It is based on an efficient representation of templates that capture the different modalities, and we show in many experiments on commodity hardware that our approach significantly outperforms state-of-the-art methods on single modalities.

1. Introduction

Real-time object learning and detection are important and challenging tasks in Computer Vision. Among the application fields that drive development in this area, robotics especially has a strong need for computationally efficient approaches, as autonomous systems continuously have to adapt to a changing and unknown environment, and to learn and recognize new objects.

For such time-critical applications, template matching is an attractive solution because new objects can be easily learned online, in contrast to statistical-learning techniques that require many training samples [3, 7, 2, 8, 26]. Our approach is related to recent and efficient template matching methods [12, 20] and more particularly to [11], which consider only images and their gradients to detect objects. As such, they work even when the object is not textured enough to use feature point techniques and can directly provide a coarse estimation of the object pose. However, similar to previous template matching approaches [1, 14, 9, 21], they suffer severe degradation of performance or even failure in the presence of strong background clutter such as the one displayed in Fig. 1. Adding other modalities such as depth information then becomes an attractive solution.

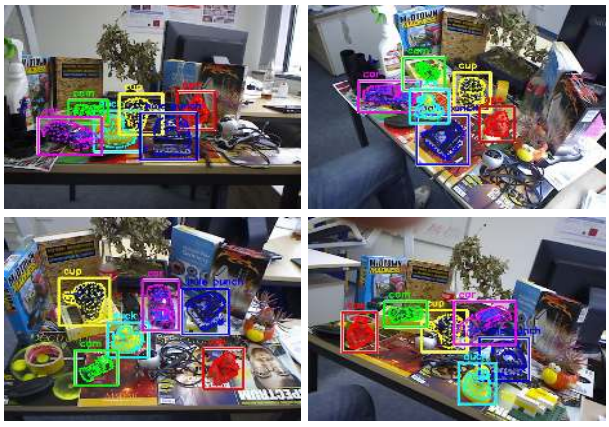

Figure 1. Our method can detect texture-less 3D objects in real-time under different poses over heavily cluttered background using an image and its depth map.

We propose an efficient method that simultaneously leverages the information of multiple acquisition modalities to define a template, and thus robustly detects known objects in difficult environments. In our approach, data from each modality is discretized into bins, and we make use of the "linearized response maps" introduced in [11] to minimize cache misses and allow for heavy parallelization. In this paper, we focus on the combination of a color image and a dense depth map. However, our approach is very generic and could easily integrate other modalities as long as they provide measurements aligned with the image that can be quantized.

For image integration, we show how to extract gradients

from the color images which are more robust to the background than gradients computed on gray value images. For depth integration, we propose a method that robustly computes 3D surface normals from dense depth maps in real-time, making sure to preserve depth discontinuities on occluding contours and to smooth out discretization noise of the sensor.

In the remainder of the paper, we first discuss related work before detailing our approach. We then present quantitative evaluations for challenging scenes that show that our multimodal templates outperform state-of-the-art approaches.

2. Related Work

The problem of multi-view 3D object detection has been widely studied in the literature. Two main categories can be distinguished: those applied to intensity images and those working on range data or depth images. Both image modalities are rarely used simultaneously. The methods operating on intensity images can be divided into two main categories: learning-based methods, and template matching methods.

Learning Based Methods These typically require a large amount of training data and a long offline training phase. As with our template-based approach, they can detect objects under different poses. For example, in [26, 13], one or several classifiers are trained to detect faces under various views. More recent approaches for 3D object detection are related to object class recognition. Stark *et al.* [23] rely on 3D CAD models and generate a training set by rendering them from different viewpoints. Liebelt and Schmid [16] combine a geometric shape and pose prior with natural images. While these approaches are able to generalize to the object class they are not real-time capable and require expensive training. By contrast our method is very fast, learns new objects in virtually no time and can recover the pose of the particular object of interest.

Template Matching This technique has played an important role in tracking-by-detection applications for many years. It is usually better adapted to low textured objects than feature point approaches. Unfortunately, this increased robustness often comes at the cost of an increased computational load that makes direct template matching inappropriate for real-time applications. This is especially true as many templates must be used to cover the range of possible viewpoints.

The methods from Gavrila and Philomin [9] and Huttenlocher *et al.* [14] are based on Chamfer matching [1] and the Haussdorf distance respectively. While somewhat faster they are very sensitive to illumination changes, noise and blur. For instance, if the image contrast is lowered, the number of extracted edge pixels progressively decreases which has the same effect as increasing the amount of occlusion. Other works do not rely on contour extraction, but directly use the image gradients. For example, Steger [24] considers the sum of dot products between the template gradients and the image gradients. This similarity measure typically provides strong peaks at the expected positions, but unfortunately also rapidly declining off-peak responses. Therefore a fine sampling must be performed to achieve good detection rates, which quickly becomes very costly.

More efficient similarity measures have been proposed recently in [12, 20]. The templates and the input images are represented using local dominant gradient orientations, which appear to yield a good trade-off between computation times and discriminative power. However, these two approaches degrade significantly when the gradient orientations are disturbed by stronger gradients coming from background clutter. In practice, this often happens in the neighborhood of the silhouette of an object, which is an important cue for texture-less objects. [11] proposed another approach based on gradient spreading that is more robust, but still not error-free. We show in this paper how considering additional modalities significantly improves the recognition performance.

Matching in Range Data This is another approach to object detection, but using depth maps instead of intensity images. An extensive review can be found in [19]. ICP [28] has been very popular for accurate registration but usually requires a good initialization. Many 3D features have also been proposed: spin-images [15], point pairs [4, 18] or point-pair histograms [22, 25]. These methods are usually expensive and assume that a full 3D CAD model of the object of interest is available. By contrast, our approach does not require a 3D model and runs in real-time.

Multimodal Detection This technique combines image and depth data, and is mainly used for pedestrian detection [5, 6, 10, 27]. While being quite effective in real applications these approaches still require exhaustive training which is inappropriate for online learning.

3. Proposed Approach

In this section, we show how we generalize the approach of [11] to easily incorporate new quantized cues. As an example we demonstrate how to integrate dense depth and color image cues.

3.1. Similarity Measure

Given a set of aligned reference images $\{\mathcal{O}_m\}_{m\in\mathcal{M}}$ of the object from a set $\mathcal{M}$ of modalities we define a template as $\mathcal{T} = (\{\mathcal{O}_m\}_{m\in\mathcal{M}}, \mathcal{P})$. $\mathcal{P}$ is a list of pairs (r, m) made of the locations r of a discriminant feature in modality m. Each template is created by extracting for each modality a small set of its most discriminant features from the corresponding reference image and by storing their locations. As

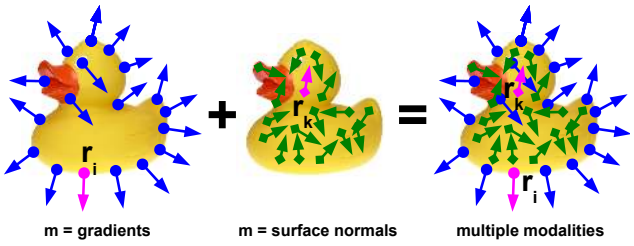

Figure 2. A toy duck with different modalities. **Left:** Image gradients are mainly found on the contour. The gradient location r_i is displayed in pink. **Middle:** Surface normals are found on the body of the duck. The normal location r_k is displayed in pink. **Right:** our approach combines multiple cues which are complementary: gradients are usually found on the object contour while surface normals are found on the object interior

shown in Fig. 2, the modalities we use in our experiments come from a standard camera and a depth sensor aligned with the camera.

Our similarity measure is a generalization of the measure defined in [11] which is robust to small translations and deformations. It can be formalized as:

$$\mathcal{E}(\{\mathcal{I}_m\}_{m\in\mathcal{M}},\mathcal{T},c)=\sum_{(r,m)\in\mathcal{P}}\left(\max_{t\in\mathcal{R}(c+r)} f_m(\mathcal{O}_m(r),\mathcal{I}_m(t))\right), \tag{1}$$

where $\mathcal{R}(c+r) = \left[c+r-\frac{T}{2},c+r+\frac{T}{2}\right] \times \left[c+r-\frac{T}{2},c+r+\frac{T}{2}\right]$ defines the neighborhood of size T centered on location $c+r$ in the input image $\mathcal{I}_m$ and the function $f_m(\mathcal{O}_m(r),\mathcal{I}_m(t))$ computes the similarity score for modality m between the reference image at location r and the input image at location t. Thus, for each feature we align the local neighborhood exactly to the associated location whereas in DOT [12], BiGG [20], HoG [3] or SIFT [17], the features are adjusted only to some regular grid. [11] showed how to compute this measure efficiently for a single modality, and we summarize below how we adapt it to multiple modalities.

3.2. Efficient Computation

We first quantize the input data for each modality into a small number of n_o values, which allows us to "spread" the data around their locations to obtain a robust representation $\mathcal{J}_m$ for each modality. For efficiency, data spread to an image location is encoded using a binary string [11]: This can be performed very quickly by OR'ing shifted versions of $\mathcal{J}_m$, and the strings are used directly as indices of lookup tables for fast precomputation of the similarity measure. We use a lookup table $\tau_{i,m}$ for each modality and for each of the n_o quantized values, computed offline as:

$$\tau_{i,m}[\mathcal{L}_m]=\max_{l\in\mathcal{L}_m}|f_m(i,l)|\,, \tag{2}$$

where

- i is the index of the quantized value of modality m. To keep the notations simple, we also use i to represent the corresponding value;
- $\mathcal{L}_m$ is a list of values of a special modality m appearing in a local neighborhood of a value i. In practice, we use the integer value corresponding to the binary representation of $\mathcal{L}_m$ as an index to the element in the lookup table.

For each quantized value of one modality m with index i we can now compute the response at each location c of the response map $\mathcal{S}_{i,m}$ as:

$$\mathcal{S}_{i,m}(c)=\tau_{i,m}[\mathcal{J}_m(c)]\,. \tag{3}$$

Finally, the similarity measure of Eq. (1) can be evaluated as:

$$\mathcal{E}(\{\mathcal{I}_m\}_{m\in\mathcal{M}},\mathcal{T},c)=\sum_{(r,m)\in\mathcal{P}}\mathcal{S}_{\mathcal{O}_m(r),m}(c+r)\,. \tag{4}$$

Since the maps $\mathcal{S}_{i,m}$ are shared between the templates, matching several templates against the input image can be done very fast once they are computed.

For computing $\mathcal{E}$, a significant speed-up can additionally be obtained by storing data in the response maps $\mathcal{S}_{i,m}$ in the same order as they are read. Since Eq.(1) allows to consider only each Tth pixel without loosing robustness, this linear storage enables heavy parallelization and avoids cache misses that would slow down the computations. Computing the similarity measure for a given template at each sampled image location can then finally be done by adding the restructured $\mathcal{S}_{i,m}$ with an appropriate offset computed from the locations r in the templates.

3.3. Modality Extraction

We now turn to how we handle the different modalities and demonstrate this on image and depth data.

3.3.1 Image Cue

We chose to consider image gradients because they proved to be more discriminant than other forms of representations [17, 24] and are robust to illumination change and noise. Additionally, image gradients are often the only reliable image cue when it comes to texture-less objects. Considering only the normalized gradients and not their magnitudes makes the measure robust to contrast changes, and taking the absolute value of the dot product between them allows it to correctly handle object occluding boundaries: It will not be affected if the object is over a dark background, or a bright background.

To increase robustness, we compute the normalized gradients on each color channel of our input image separately

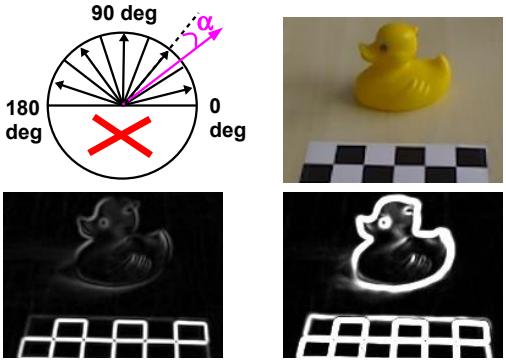

Figure 3. **Upper Left:** Quantizing the gradient orientations: the pink orientation is closest to the second bin. **Upper right:** A toy duck with a calibration pattern. **Lower Left:** The gradient image computed on a gray value image. The object contour is hardly visible. **Lower right:** Gradients computed with our approach. Details of the object contours are clearly visible.

and for each image location use the normalized gradient of the channel whose magnitude is largest. Given an RGB color image $\mathcal{I}$, we compute the normalized gradient map $\mathcal{I}_\mathcal{G}(x)$ at location x with

$$\mathcal{I}_\mathcal{G}(x) = \frac{\partial \hat{\mathcal{C}}}{\partial x} \tag{5}$$

where

$$\hat{\mathcal{C}}(x) = \underset{\mathcal{C}\in\{R,G,B\}}{\arg\max} \left\| \frac{\partial \mathcal{C}}{\partial x} \right\| \tag{6}$$

and R, G, B are the RGB channels of the corresponding color image. Our similarity measure is then:

$$f_\mathcal{G}(\mathcal{O}_\mathcal{G}(r), \mathcal{I}_\mathcal{G}(t)) = \left| \mathcal{O}_\mathcal{G}(r)^\top \mathcal{I}_\mathcal{G}(t) \right| \tag{7}$$

where $\mathcal{O}_\mathcal{G}(r)$ is the normalized gradient map of the reference image at location r and $\mathcal{I}_\mathcal{G}(t)$ the normalized gradient map of the current image at location t respectively.

In order to quantize the gradient map we omit the gradient direction, consider only the gradient orientation and divide the orientation space into n_0 equal spacings as shown in Fig. 3. To make the quantization robust to noise, we assign to each location the gradient whose qantized orientation occurs most often in a 3×3 neighborhood. We also keep only the gradients whose norms are larger than a small threshold. The whole unoptimized process takes about 31ms on the CPU for a VGA image.

3.3.2 Depth Cue

Similar to the image cue, we decided to use quantized surface normals computed on a dense depth field for our template representation as shown in Fig. 4. They allow us to represent both close and far objects while fine structures are preserved.

In the following, we propose a method for the fast and robust estimation of surface normals in a dense range image. Around each pixel location x, we consider the first order Taylor expansion of the depth function $\mathcal{D}(x)$:

$$\mathcal{D}(x + dx) - \mathcal{D}(x) = dx^\top \nabla \mathcal{D} + h.o.t. \tag{8}$$

Within a patch defined around x, each pixel offset dx yields an equation that constrains the value of $\nabla\mathcal{D}$, allowing to estimate an optimal gradient $\hat{\nabla\mathcal{D}}$ in a least-square sense. This depth gradient corresponds to a 3D plane going through three points X, X_1 and X_2:

$$X = \vec{v}(x)\mathcal{D}(x), \tag{9}$$

$$X_1 = \vec{v}(x + [1,0]^\top)(\mathcal{D}(x) + [1,0]\hat{\nabla\mathcal{D}}), \tag{10}$$

$$X_2 = \vec{v}(x + [0,1]^\top)(\mathcal{D}(x) + [0,1]\hat{\nabla\mathcal{D}}). \tag{11}$$

where $\vec{v}(x)$ is the vector along the line of sight that goes through pixel x and is computed from the internal parameters of the depth sensor. The normal to the surface at the 3D point that projects on x can be estimated as the normalized cross-product of $X_1 - X$ and $X_2 - X$.

However this would not be robust around occluding contours, where the first order approximation of Eq. (8) no longer holds. Inspired by bilateral filtering, we ignore the contributions of pixels whose depth difference with the central pixel is above a threshold. In practice, this approach effectively smooths out quantization noise on the surface, while still providing meaningful surface normal estimates around strong depth discontinuities. Our similarity measure is then defined as the dot product of the normalized surface normals:

$$f_\mathcal{D}(\mathcal{O}_\mathcal{D}(r), \mathcal{I}_\mathcal{D}(t)) = \mathcal{O}_\mathcal{D}(r)^\top \mathcal{I}_\mathcal{D}(t) \tag{12}$$

where $\mathcal{O}_\mathcal{D}(r)$ is the normalized surface normal map of the reference image at location r and $\mathcal{I}_\mathcal{D}(t)$ the normalized surface normal map of the current image at location t.

Finally, as shown in Fig. 4, we measure the angles between the computed normal and a set of precomputed vectors to quantize the normal directions into n_0 bins. These vectors are arranged in a right circular cone shape originating from the peak of the cone pointing towards the camera. To make the quantization robust to noise, we assign to each location the quantized value that occurs most often in a 5×5 neighborhood. The whole process is very efficient and needs only 14ms on the CPU and less than 1ms on the GPU.

4. Experiments

We compared our approach, which we call LINE-MOD (for "multimodal-LINE"), to several methods: LINE as introduced in [11], which uses only the image intensities (referred to as LINE-2D here); a variant that we call LINE-3D

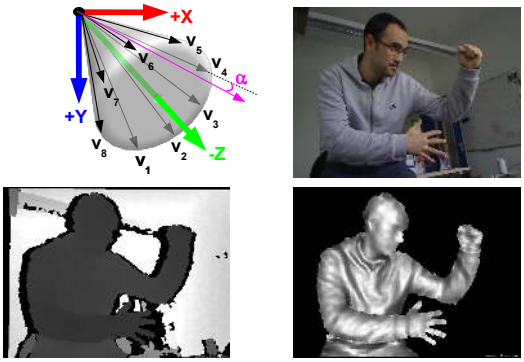

Figure 4. **Upper Left:** Quantizing the surface normals: the pink surface normal is closest to the precomputed surface normal v_4. It is therefore put into the same bin as v_4. **Upper right:** A person standing in an office room. **Lower Left:** The corresponding depth image. **Lower right:** Surface normals computed with our approach. Details are clearly visible and depth discontinuities are well handled. We removed the background for visibility reasons.

and that uses only the depth map; DOT [12]; and HOG [3]. For HOG, we used our own optimized implementation and replaced the Support Vector Machine mentioned in the original work of HOG by a nearest neighbor search. In this way, we can use it as a robust representation and quickly learn new templates as with the other methods. The experiments were performed on one processor of a standard notebook with an Intel Centrino Processor Core2Duo with 2.4 GHz and 3 GB of RAM. For obtaining the image and the depth data we used the Primesense$^{(tm)}$ PSDK 5.0 device.

4.1. Robustness

We used six sequences made of 2000 real images each. Each sequence presents illumination and large viewpoint changes over heavy cluttered background. Ground truth is obtained with a calibration pattern attached to each scene that enables us to know the actual location of the object. The templates were learned over homogeneous background.

We consider the object to be correctly detected if the location given back is within a fixed radius of the ground truth position. As depicted in the left column of Fig. 7, our new approach always outperforms all the other ones and shows only few false positives. We believe that this is due to the complementarity of the object features that compensate for the weaknesses of each other. The superiority of our new approach becomes even more obvious in Table 1: If we set the threshold for each approach to allow for 97% true positive rate and only evaluate the hypothesis with the largest response, we obtain for our new approach a high detection rate with a very small false positive rate. This is in contrast to LINE-2D, where the true positive rate is often over 90%, but the false positive rate is not negligible, which makes expensive post-processing necessary. In our method, using only the response with the largest value might be sufficient

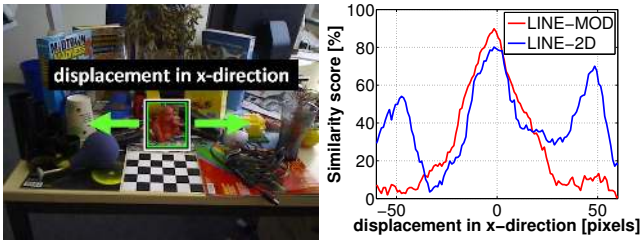

Figure 5. Combining multi-modalities results in a more discriminative response function. Here we compare LINE-MOD against LINE-2D on the shown image. We plot the response function of both methods with respect to the true location of the monkey. One can see that the response of our new method exhibits a single and discriminative peak whereas LINE-2D has several peaks which are of comparable height. This is one explanation why our new approach works better and produces fewer false positives.

in most cases.

One reason for this high robustness is the good separability of the multimodal approach as shown in the middle of Fig. 7: one can see that a specific threshold—about 80 in our implementation—separates almost all true positives well from almost all false positives. This has several advantages. First, we will detect almost all instances of the object by setting the threshold to this specific value. Second, we also know that almost every returned template with a similarity score above this specific value is a true positive. And third, the threshold is always around the same value which supports the conclusion that it might also work well for other objects. One hint why our multimodal approach has such a good separability property is given in Fig. 5. One can see that the response function has only one clear peak around the true location of the object while LINE-2D shows other peaks with almost the same height.

4.2. Speed

Learning new templates only requires extracting and storing multimodal features, which is almost instantaneous. Therefore, we concentrate on runtime performance. The runtimes given in Fig. 6 show that the general LINE approach (with 126 features) is real-time and can parse a VGA image with over 3000 templates at about 10 fps on the CPU. DOT is initially faster than our approach but becomes slower as the number of templates increases. This is because the runtime of LINE-MOD is only dependant on the number of features and independent of the object/template size whereas the runtime of DOT is not. Therefore, to handle larger objects DOT has to use larger templates which makes the approach slower once the number of templates increases.

To detect an object under a full coverage of viewpoints (360 degree tilt rotation, 90 degree inclination rotation and in-plane rotations of $\pm$ 80 degrees, scale changes from 1.0 to 2.0), we usually need less than two thousands templates.

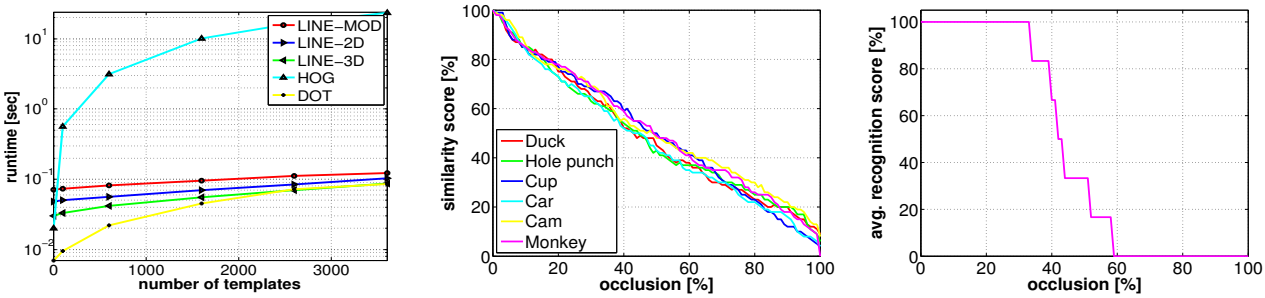

Figure 6. **Left:** Our new approach runs in real-time and can parse a 640×480 image with over 3000 templates at about 10 fps. **Middle:** Our new approach is linear with respect to occlusion. **Right:** Average recognition score for the six objects of Sec.4.1 with respect to occlusion.

4.3. Occlusion

We also tested the robustness of our approach with respect to occlusion. We added synthetic noise and illumination changes to the images, incrementally occluded the six different objects of Section 4.1 and measured the corresponding response values. As expected, the similarity measure used by our method behaves linearly in the percentage of occlusion as reported in the middle of Fig. 6. This is a desirable property since it allows detection of partly occluded templates by setting the detection threshold with respect to the tolerated percentage of occlusion. We also experimented with real scenes where we first learned our six objects in front of a homogeneous background and then added heavy 2D and 3D background clutter. For recognition we incrementally occluded the objects. We define our object as correctly recognized if the template with the highest response is found within a fixed radius of the ground truth object location. The average recognition result is displayed on the left of Fig. 6: Even with over 30% occlusion our method is still able to recognize objects.

5. Conclusion

We have presented a method to exploit different modalities for real-time object detection. Our novel approach is able to correctly detect 3D texture-less objects in real-time under heavy background clutter, illumination changes and noise with almost no false positives. We showed how to efficiently preprocess image and depth data to robustly integrate both cues into our approach. Additionally, we have shown that our approach outperforms state-of-the-art methods with respect to the combination of recognition rate and speed, especially in heavily cluttered environments.

Acknowledgment: This project was funded by the BMBF project AVILUSplus (01IM08002).

References

[1] G. Borgefors. Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm. *TPAMI*, 1988.

[2] A. Bosch, A. Zisserman, and X. Munoz. Image Classification Using Random Forests. In *ICCV*, 2007.

[3] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[4] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *CVPR*, 2010.

[5] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, 2010.

[6] A. Ess, B. Leibe, and L. J. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.

[7] R. Fergus, P. Perona, and A. Zisserman. Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition. *IJCV*, 2006.

[8] V. Ferrari, F. Jurie, and C. Schmid. From Images to Shape Models for Object Detection. *IJCV*, 2009.

[9] D. Gavrila and V. Philomin. Real-Time Object Detection for "Smart" Vehicles. In *ICCV*, 1999.

[10] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 2007.

[11] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, P. Fua, N. Navab, and V. Lepetit. Gradient response maps for real-time detection of texture-less objects. *under revision PAMI.*

[12] S. Hinterstoisser, V. Lepetit, S. Ilic, P. Fua, and N. Navab. Dominant Orientation Templates for Real-Time Detection of Texture-Less Objects. In *CVPR*, 2010.

[13] C. Huang, H. Ai, Y. Li, and S. Lao. Vector boosting for rotation invariant multi-view face detection. In *CVPR*, 2005.

[14] D. Huttenlocher, G. Klanderman, and W. Rucklidge. Comparing Images Using the Hausdorff Distance. *TPAMI*, 1993.

[15] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3 d scenes. *TPAMI*, 1999.

[16] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, 2010.

[17] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004.

[18] A. S. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *TPAMI*, 2006.

[19] A. S. Mian, M. Bennamoun, and R. A. Owens. Automatic correspondence for 3D modeling: An extensive review. *International Journal of Shape Modeling*, 2005.

[20] M. Muja, R. Rusu, G. Bradski, and D. Lowe. REIN - a Fast, Robust, Scalable REcognition INfrastructure. In *ICRA*, 2011.

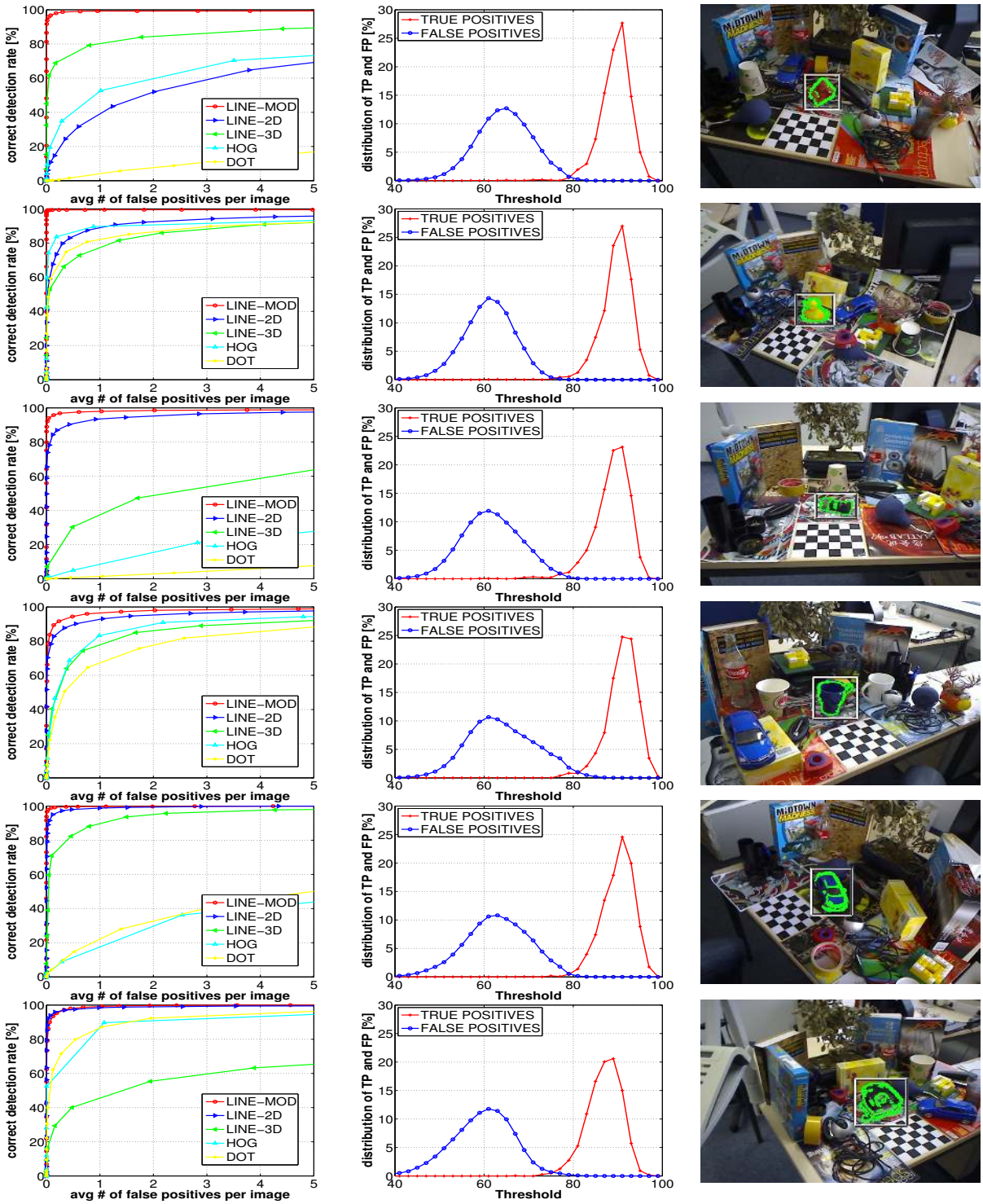

Figure 7. Comparison of our approach (LINE-MOD) with LINE based on gradients [11] (LINE-2D), LINE based on normals (LINE-3D), DOT [12] and HOG [3] on real 3D objects. Each row corresponds to a different sequence (made of over 2000 images each) on heavy cluttered background: A toy monkey, a toy duck, a camera, a cup, a toy car, and a hole punch. The approaches were learned on a homogeneous background. **Left:** Percentage of true positives plotted against the average percentage of false positives. Our multimodal templates provide about the same recognition rates for all objects while the other approaches have a much larger variance depending on the object type. Our approach outperforms the other approaches in most cases. **Middle:** The distribution of true and false positives plotted against the threshold. They are well separable from each other. **Right:** One sample image of the corresponding sequence shown with the object detected by our new approach.

Sequence	MOD-LINE	LINE-2D	LINE-3D	HOG	DOT
Toy-Monkey (2164 images)	**97.9%—0.3%**	50.8%—49.1%	86.1%—13.8%	51.8%—48.2%	8.6%—91.4%
Camera (2173 images)	**97.5%—0.3%**	92.8%—6.7%	61.9%—38.1%	18.2%—81.8%	1.9%—98.0%
Toy-Car (2162 images)	**97.7%—0.0%**	96.9%—0.4%	95.6%—2.5%	44.1%—55.9%	34.0%—66.0%
Cup (2193 images)	**96.8%—0.5%**	92.8%—6.0%	88.3%—10.6%	81.1%—18.8%	64.1%—35.8%
Toy-Duck (2223 images)	**97.9%—0.0%**	91.7%—8.0%	89.0%—10.0%	87.6%—12.4%	78.2%—21.8%
Hole punch (2184 images)	**97.0%—0.2%**	96.4%—0.9%	70.0%—30.0%	92.6%—7.4%	87.7%—12.0%

Table 1. True and false positive rates for different thresholds on the similarity measure of different methods. In some cases no hypotheses were given back so the sum of true and false positives can be lower than 100%. Our MOD-LINE approach obtains very high recognition rates at the cost of almost no false positives, and outperforms all the other approaches. The corresponding best values are shown in bold print.

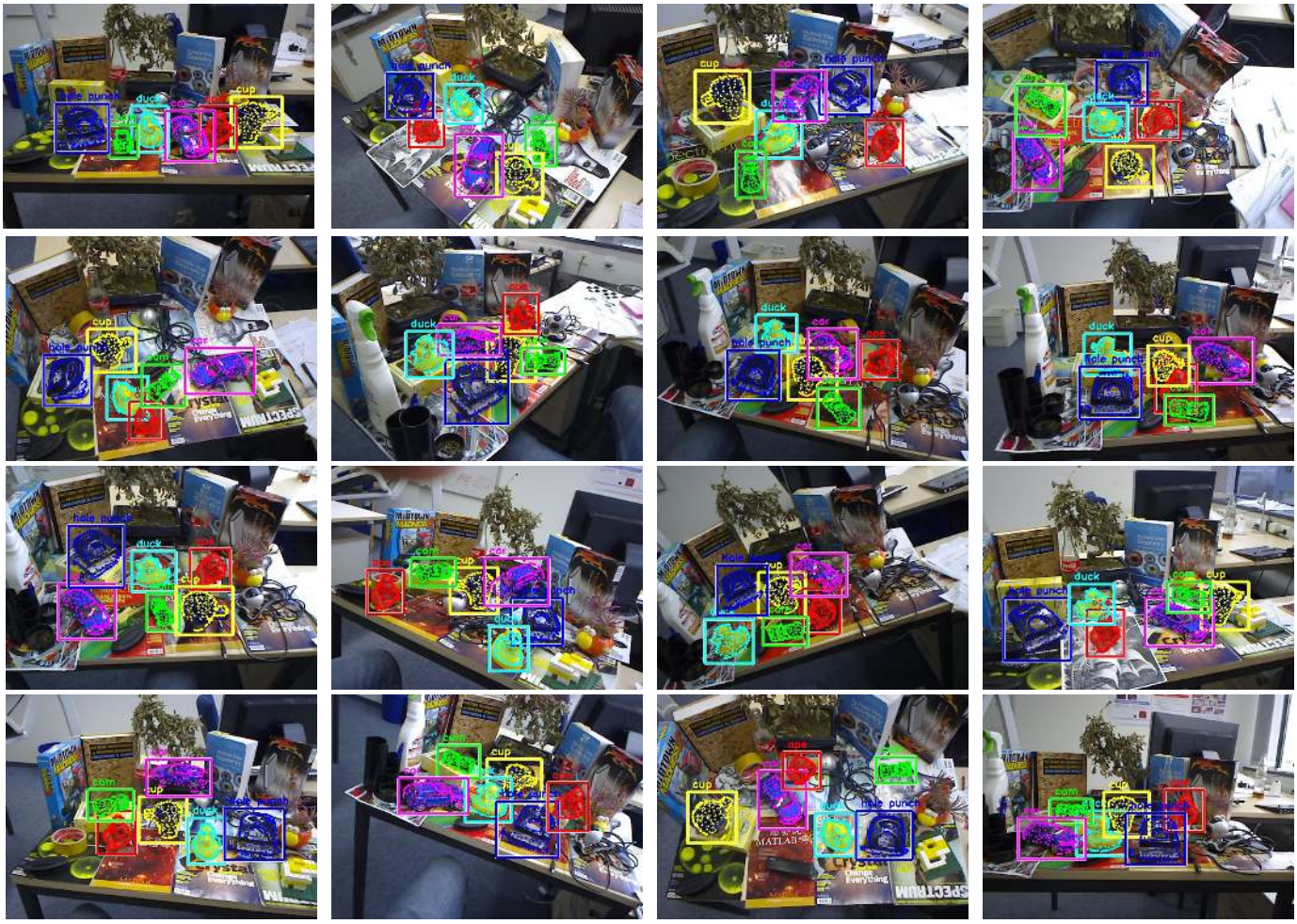

Figure 8. Different texture-less 3D objects detected simultaneously in real-time by our MOD-LINE method under different poses on heavily cluttered background with partial occlusion. **See also the supplemental video on http://campar.in.tum.de/Main/StefanHinterstoisser**

[21] C. F. Olson and D. P. Huttenlocher. Automatic Target Recognition by Matching Oriented Edge Pixels. *TIP*, 1997.

[22] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, 2009.

[23] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In *BMVC*, 2010.

[24] C. Steger. Occlusion Clutter, and Illumination Invariant Object Recognition. In *ISPRS*, 2002.

[25] F. Tombari, S. Salti, and L. D. Stefano. Unique signatures of histograms for local surface description. In *ECCV*, 2010.

[26] P. Viola and M. Jones. Fast Multi-view Face Detection. In *CVPR*, 2003.

[27] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009.

[28] Z. Zhang. Iterative point matching for registration of free-form curves. *IJCV*, 1994.