Multi-View Hypotheses Transfer for Enhanced Object Recognition in Clutter

Thomas Fäulhammer Michael Zillich Markus Vincze
Vienna University of Technology, Automation and Control Institute
1040 Vienna, Gußhausstraße 27-29
{faeulhammer, zillich, vincze}@acin.tuwien.ac.at

Abstract

Despite 3D object recognition being an ongoing research field for many years, state-of-the-art methods still face problems in real-world situations with clutter, occlusion or non-textured objects. To overcome these problems, recent approaches use multi-view setups exploiting beneficial vantage points of the environment. Minimizing the assumptions on the scene and objects of interest made by these systems, we present an efficient online multi-view method, which integrates information of the captured environment merging individual single-view recognition outputs. Our method achieves state-of-the-art results for the Willow dataset at reduced computational time. Further evaluations on the more challenging TUW dataset show an increase in f-score and object pose accuracy over the number of observations.

1 Introduction

Object recognition with 3D pose estimation is an active research area, where many promising approaches have been introduced recently [3, 10–12]. Many of these methods share common problems when faced with challenging real-world scenes, where objects are partially occluded, are far away (and thus cover only a small image area) or lack distinctive features. Most of these problems can be alleviated by choosing a better view point, e.g. nearer to an object or looking behind an occluder. Particularly in robotics, the camera does not have to be static. As a mobile robot moves around in the environment, it will encounter many views of the current scene. While a particular object might be occluded or displays only a poorly textured surface from one view, chances are good that the same object will appear much more favourably in some other view.

Regarding typical every-day indoor environments, the majority of the scene is static, especially objects (it is mostly people that cause movement). Exploiting this *static* information, we propose a method working on RGB-D data that continuously transfers hypotheses constructed at various vantage points into a common framework to gather the maximum amount of information for all objects in the scene, and thus overcome the problem of single particularly poor views. Contrary to existing batch methods [1,3], the proposed approach enables the system to improve recognition online (i.e. with each new observation) using a dynamic graph representation of the observed environment. Additionally, we provide an in-depth evaluation of our proposed method on two publicly available datasets containing heavily cluttered RGB-D scenes and compare it to state-of-the-art systems. In particular, we show the improvement of the recognition performance over the number of observations.

2 Related Work

Focusing on recent methods deployed on *single-view* RGB-D data, Xie *et al* [12] proposed dense SIFT feature extraction followed by a RANSAC pose estimation stage to generate hypotheses, which are verified by means of a multimodal (color, shape and gradients) scoring scheme. Tang *et al* [11] extracts table planes from RGB-D data, clusters points above by Euclidean distance and finds candidate models of textured objects by matching SIFT features. Hinterstoisser *et al* [7] proposed a multimodal template (color gradients and surface normals) based matching approach to handle objects without texture. Although these methods show excellent performance in particular scenarios, they are either computationally expensive, have many assumptions on the scene layout, require textured objects or show limited recognition results when objects become partially occluded. For an extensive review of available approaches for object recognition and pose estimation from single images please refer to [6].

To avoid the computational cost of an accurate full *multi-view* generalized camera approach, , Collet and Srinivasa [3] obtain efficient recognition and full-pose estimation by an introspective multi-view method that uses a multi-step optimization technique.

Instead of using a static camera rig with known extrinsic parameters, general views without the prior knowledge of the relative pose to a common coordinate system allow us to use our method on more general setups covering a wider field of view or seeing the scene from a completely other perspective for instance. It shows that recognition of heavily occluded objects is still possible when a hypothesis is generated from a better viewpoint with respect to the occluded object. Furthermore, instead of using a Mean Shift clustering approach together with RANSAC for hypothesis refinement, we build a graph out of single-view hypotheses and use a 3D hypothesis verification approach [1].

Lai et al. [8] proposed a method for semantic 3D scene labelling based on single-detections, which i) reconstructs the 3D scene, ii) detects possible objects in each RGB-D frame, iii) projects the single-view scores into the reconstructed scene and iv) enforces label consistency through a voxel-based MRF. Our method enforces global consistency by a suitable 3D hypothesis verification stage and uses shared single-view recognition results among different frames to aid during the reconstruction stage. The advantage of object detection while mapping an environment has been recently shown in [5, 10] within a joint detection, tracking and mapping framework. Please note that recent RGB-D mapping methods require a continuous stream of data

which is not always available for existing recognition datasets (e.g. Challenge, TUW or Willow datasets).

The approach proposed in this work is an online multi-view object instance recognition method based on [1], which merges single-view results in a batch to generate ground-truth data of a static environment.

3 Approach

Sensing a static environment over a particular observation time by an RGB-D sensor from different vantage points, the goal of the proposed method is to recognize pre-trained models $\mathcal{M}$ and their respective 6DoF pose at each time step k using information from current and previous observations as depicted in Fig. 1.

3.1 Single-view recognition

The single-view recognizer generates for each scene point cloud $\boldsymbol{S}_k$ captured at time k a set of candidate objects (hypotheses) potentially present in the scene

$$\mathcal{H}_k = \left\{ h_k^j \right\}_{j=1}^{H_k}; \qquad h_k^j = \left(o_k^j, \boldsymbol{P}_k^j \right), \tag{1}$$

where H_k is the number of constructed hypotheses, $o_k^j \in \mathcal{M}$ the object identity and $\boldsymbol{P}_k^j$ a 4×4 homogeneous transformation matrix defining the 6DoF object pose with respect to the reference frame of $\boldsymbol{S}_k$.

To deploy the algorithm in a wide range of recognition problems, object hypotheses are obtained using the single-view recognition system proposed by Aldoma et al. [2], which uses a combination of 2D and 3D (*local* and *global*) recognition pipelines and exploits the different strengths of the individual algorithms (i.e. using texture as well as geometry information). After individual pose refinement by ICP, a final verification stage returns the subset of hypotheses $\hat{\mathcal{H}}_k$ that best represents the scene $\boldsymbol{S}_k$ with respect to a global optimality criterion [2].

3.2 Multi-view representation

To exploit the information gain from multiple views $\{\boldsymbol{S}_k\}_{k=0}^{K-1}$ of an environment, we create an undirected graph with vertices $\{\mathcal{V}_k\}_{k=0}^{K-1}$ representing single-view information and edges $\{\mathcal{E}_k\}_{k=0}^{K-1}$ connecting the views.

For each snapshot of the scene, a vertex $\mathcal{V}_K = (\boldsymbol{S}_K, \hat{\mathcal{H}}_K)$ is connected to existing vertices that share a common object hypothesis (i.e. hypotheses with the same model identity o) by edges

$$\mathcal{E}_K = \{\mathcal{E}_{K,k}\}_{k=0}^{K-1}, \quad \mathcal{E}_{K,k} = \left\{ e_{K,k}^l \right\}_{e=0}^{E_{K,k}}, \tag{2}$$

$$e_{K,k}^l = \left(o_{K,k}^l, \boldsymbol{T}_{K,k}^l, \vartheta_{K,k}^l, K, k \right), \tag{3}$$

where $\boldsymbol{T}_{K,k}^l$ is a 4×4 homogeneous transformation matrix describing the relative pose between $\boldsymbol{S}_K$ and $\boldsymbol{S}_k$ estimated by the common hypothesis $o_{K,k}^l$. The quality of the registration between $\boldsymbol{S}_K$ and $\boldsymbol{S}_k$ by $\boldsymbol{T}_{K,k}^l$ is measured by the edge weight $\vartheta_{K,k}^l$, which takes into account visibility consistency, relative overlap between views and the normal angle between corresponding point pairs[1]. $E_{K,k}$ is the total number of shared hypotheses between verified single- and multi-view hypotheses $\hat{\mathcal{H}}_K$ and $\hat{\mathcal{H}}_{k^+}$. Given $o_{K,k}^l$ is shared amongst hypotheses h_K^a and h_k^b, the transformation is estimated by

$$\boldsymbol{T}_{K,k}^l = \boldsymbol{P}_K^a \left(\boldsymbol{P}_k^b \right)^{-1}. \tag{4}$$

A fully connected graph $\mathcal{G}$ is therefore obtained if there are enough common object hypotheses for each vertex. To avoid isolated vertices in $\mathcal{G}$ (e.g. no recognized object) or, in case of weak object pose estimates, to possibly obtain a better estimate, additional edges are created by means of visual features of the scene itself. In particular, each view $\mathcal{V}_k$ is matched to $\mathcal{V}_K$ using a *first nearest neighbour* strategy with respect to their respective SIFT features yielding a correspondence set between both frames from which a rigid transformation is estimated. In our implementation, geometrically consistent correspondences [2] with a consensus set of at least 15 correspondences are used to effectively extend $\mathcal{E}_{K,k}$. While this *scene to scene* edges usually provide good transformation estimates for textured environments, transformations estimated by common object hypotheses can often improve registration performance for camera poses, which are too far away to each other to reliably match scene features.

3.3 Hypotheses projection and scene reconstruction

Using the graph representation $\mathcal{G}$, individual single-view information is merged into a global representation of the scene by the estimated camera poses. To reduce the computational complexity, we only select the edge within $\mathcal{E}_{K,k}$ with the lowest edge weight ϑ and remove all remaining ones. Since wrong pose estimates would decrease the quality of the reconstructed environment, we isolate $\mathcal{V}_K$ if the edge weight is larger than a predefined threshold effectively ignoring multi-view information at this point and building subgraphs for future observations. Choosing the subgraph containing the most recent observation $\mathcal{V}_K$, we build a 3D+RGB reconstruction of the scene as well as a multi-view model candidate set $\mathcal{H}_{K^+}$ by traversing this subgraph and compounding $\hat{\mathcal{H}}_K$ with all verified multi-view sets $\hat{\mathcal{H}}_{k^+}$ generated in previous observations, which are transformed into the reference frame of $\mathcal{V}_K$.

3.4 3D+RGB hypothesis verification

At this end, we verify our merged hypotheses $\mathcal{H}_{K^+}$ against the reconstructed environment by a 3D verification stage [1]. Since small pairwise registration errors get accumulated over multiple edges, poses of all overlapping views are optimized by a global registration [4]. To efficiently handle a large set of scene points and improve the quality of merging several clouds captured independently by noisy RGB-D sensors, finite differences are computed by appropriate nearest neighbour search in an Octree structure and the noise model of Nguyen *et al* [9] is applied. The final output is a set of verified hypotheses $\hat{\mathcal{H}}_{K^+}$, which replaces the verified single-view hypotheses stored in $\mathcal{V}_K$.

4 Results

We test our method on the two publicly available datasets TUW (static)[2] and Challenge[3] and compare

[1] For a more detailed definition of ϑ please refer to [1].

[2] https://repo.acin.tuwien.ac.at/tmp/permanent/dataset_index.php

[3] http://rll.berkeley.edu/2013_IROS_ODP/

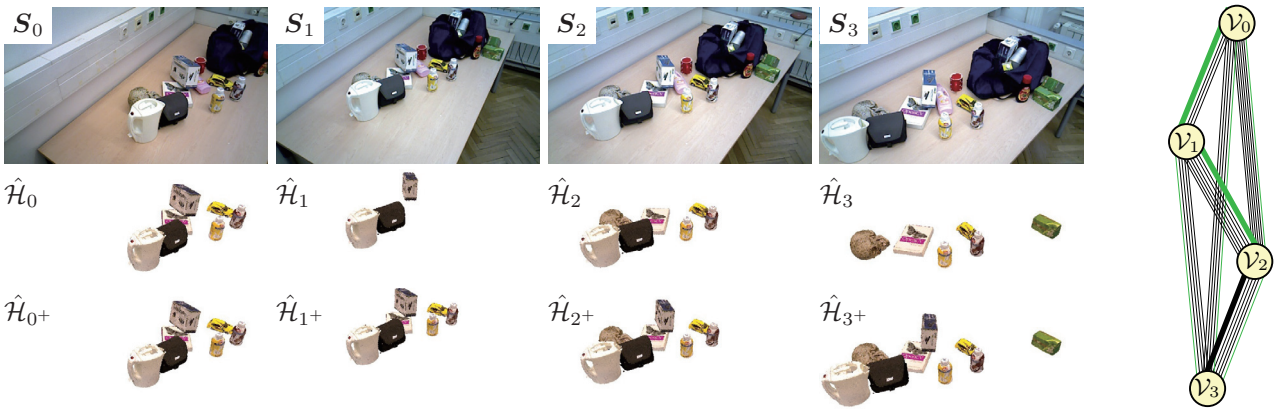


Figure 1: Workflow for a static scene captured from $K = 4$ different vantage points observed in time from left to right. *Top*: input RGB-D data; *Middle*: verified single-view hypotheses; *Bottom*: verified multi-view hypotheses considering recent observations (scenes to the left). *Right*: graph representation $\mathcal{G}$ with edges shown for common object hypotheses (black) and SIFT *scene to scene* matching (green). An acyclic graph is computed online (selected edges shown bold);

it to a single-view (SV) only recognition system [2], as well as to the results reported in [12] and [11].

4.1 Test setup

Each model of the two datsets is compressed into a voxelgrid with a resolution of 5 mm and views belonging to a sequence are processed online in the order of the given dataset identifiers. The single-view multi-pipeline recognizer [2] uses SIFT and SHOT features at keypoint locations computed by DoG and uniform sampling, respectively. Hypotheses for an object are created for clusters constructed by at least 3 keypoint correspondences, which are computed by first nearest neighbor matching of scene and model keypoint descriptors. Each correspondence belongs to at most 2 clusters and must be geometrically consistent (i.e. max. distance of 1.5 cm and normal dot product ≤ 0.2). All hypotheses are refined by 8 iterations of ICP. The subsequent verification stage uses the parameters from [2] and outputs the result for the single-view system used for comparison in the following.

All transferred multi-view hypotheses from previous observations are refined by another 8 iterations of ICP and the 3D verification stage uses the same parameter as verification for single-view considering all scene clouds observed so far. This produces intermediate results for every view.

The performance is measured by precision, recall and f-score, where a verified hypothesis is counted as true positive if it is within 3 cm of a ground-truth object of the same model identity with respect to its centroid. To neglect ground-truth objects outside the field of view or invisible for the current camera orientation, these values were only computed for instances with a ground-truth occlusion $\leq 95\%$ specified by [1]. Table 1 shows precision and recall averaged over all views of the given sequences. The improvement of performance over time shown in Fig. 2 is measured by averaging all intermediate results with the same number of observations taken into account by the system.

4.2 Evaluation

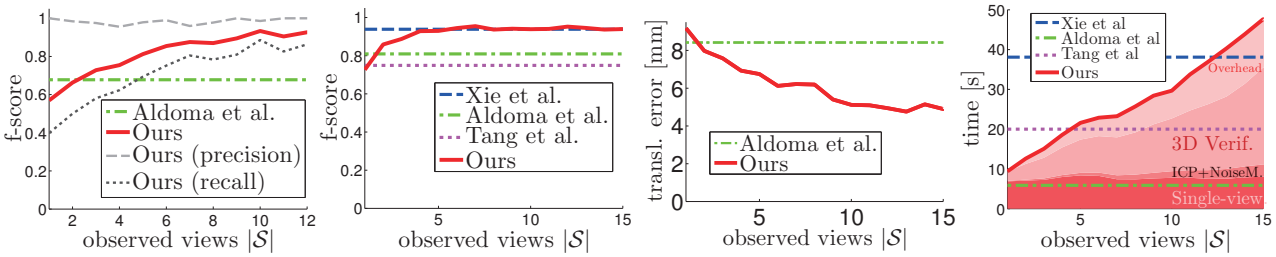
TUW dataset: As shown in Table 1, the common verification stage rejects false positives with similar precision for [2] and our method. However, the recall rate is significantly larger for the proposed multiple-view recognition. This is in particular evident for sequences with many observations (Seq. 3, 4, 6, 13, 14, 15), high clutter/occlusion (1, 3, 5, 12) or containing non-textured objects, which are usually more difficult to recognize by a single-view system (Seq. 8, 10). Fewer observations for a sequence lead to a smaller set of extended hypotheses but also mean that the first observation (multi-view equal single-view result) has a bigger influence on the overall performance averaged over all intermediate results. The poor performance in the highly cluttered Seq. 9 (five boxes stacked together, no other object present) is due to the low clutter term chosen for the 3D verification parameter set rejecting all generated hypotheses for this particular problem. Choosing individual parameters for each sequence could overcome this situation. An example result for the first four observations of sequence 14 is shown in Fig 1. Considering all ground-truth objects (occlusion $\leq 95\%$), this evaluation method achieves an overall precision and recall rate of 0.96 and 0.72, respectively. This results in an overall f-score of 0.82 compared to 0.62 for [2]. Fig. 2a shows the increase in the recall rate and, accordingly, f-score for a larger number of observations taken into account by our method.

Willow dataset: Comparing to state-of-the-art results reported in [2, 11, 12], Fig. 2b shows the average f-score of our multi-view method for intermediate results taking the same number of observations into account. On the overall Willow dataset, we achieve a precision and recall rate of 0.94 and 0.90, respectively, resulting in an overall f-score of 0.92. Another advantage of the extended set of hypotheses is the increased dimension of the solution space for the verification stage. This allows finding a better optimum if hypotheses with more accurate pose estimates get transferred into the current hypotheses set. In general, this results in a slightly lower translational and rotational error as shown in Fig. 2c.

Computation time: Fig. 2d shows the approximately linear increase in computational cost by the number of observations. Given a static scene and perfect single-view recognition, this is also proportional to the cardinality of the multi-view hypotheses set. Besides the approximately constant computational time for single-

Scene ID		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SV [2]	prec.	1.00	0.97	**0.98**	**0.98**	1.00	1.00	**0.98**	1.00	0.83	**1.00**	1.00	1.00	0.97	1.00	0.98
Ours	prec.	1.00	**1.00**	0.97	0.95	1.00	1.00	0.89	1.00	**1.00**	0.91	1.00	1.00	**0.98**	1.00	**1.00**
SV [2]	recall	0.51	0.61	0.53	0.53	0.47	0.64	0.85	0.41	**0.11**	0.38	0.74	0.66	0.43	0.48	0.34
Ours	recall	**0.58**	**0.63**	**0.85**	**0.74**	**0.70**	**0.84**	**0.88**	**0.80**	0.00	**0.63**	**0.81**	**0.89**	**0.57**	**0.71**	**0.52**
#views		7	7	14	16	10	13	8	8	9	7	9	8	18	16	13
#objects		53	62	78	258	186	180	48	40	45	16	53	66	82	198	122

Table 1: Precision (top) and recall (middle) for the 15 *static* environments of the TUW dataset for the single-view system [2] and the proposed multi-view approach (*Ours*). Bottom rows show the total number of views and ground-truth objects visible (occlusion $\leq 95\%$) in each sequence.



(a) TUW (b) Willow (c) Avg. trans. error on Willow (d) Test time

Figure 2: Performance over time.

view hypotheses generation ($\approx$ 8 s), the multi-view verification takes up most time in our implementation. Please note that the computation time was measured on different machines, but with similar components (multi-core i7 processor and $\geq$ 24GB RAM). Computation times for [11,12] have been taken from the respective papers. Even though this performance measure is only partially able to compare the real complexity of the methods, Fig. 2d still gives rise to significant speed ups achieved by our method compared to [12], which reports similar recognition performance.

5 Conclusion

We presented a multi-view recognition method that incorporates individual results of single-view observations into a common coordinate system and gives superior results for each new input frame compared to a single-view only system. Our method showed a significant boost in recall for the TUW dataset, consisting of heavily cluttered RGB-D frames with textured and non-textured objects, which are highly occluded in some frames. Using additional information from other views, state-of-the-art performance was shown for the Willow dataset at a reduced computational complexity. Furthermore, our method gained a more accurate pose estimate by merging hypotheses from multiple views.

As our method is independent of the single-view recognition method being used for hypotheses generation and many parameters influence the system performance, we are confident that even better recognition results are feasible with our approach. Computation time can also be reduced by further parallelization of the algorithm and implementation on the GPU.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement No. 600623, STRANDS and No. 610532, SQUIRREL.

References

[1] A. Aldoma, T. Fäulhammer, and M. Vincze. Automation of ground truth annotation for multi-view rgb-d object instance recognition datasets. In *IROS*. IEEE, 2014.

[2] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. Di Stefano, and M. Vincze. Multimodal cue integration through Hypotheses Verification for RGB-D object recognition and 6DoF pose estimation. In *ICRA*, 2013.

[3] A. Collet, M. Martinez, and S. S. Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. *IJRR*, 30(10), 2011.

[4] S. Fantoni, U. Castellani, and A. Fusiello. Accurate and automatic alignment of range surfaces. In *3DIMPVT*. Citeseer, 2012.

[5] N. Fioraio and L. D. Stefano. Joint detection, tracking and mapping by semantic bundle adjustment. In *CVPR*. IEEE, 2013.

[6] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan. 3d object recognition in cluttered scenes with local surface features: A survey. *PAMI*, 36(11), 2014.

[7] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*. Springer, 2012.

[8] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labeling in 3d scenes. In *ICRA*. IEEE, 2012.

[9] C. V. Nguyen, S. Izadi, and D. Lovell. Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *3DIMPVT*. IEEE, 2012.

[10] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *CVPR*. IEEE, 2013.

[11] J. Tang, S. Miller, A. Singh, and P. Abbeel. A textured object recognition pipeline for color and depth image data. In *ICRA*. IEEE, 2012.

[12] Z. Xie, A. Singh, J. Uang, K. S. Narayan, and P. Abbeel. Multimodal blending for high-accuracy instance recognition. In *IROS*. IEEE, 2013.