Visual Information Abstraction For Interactive Robot Learning

Kai Zhou, Andreas Richtsfeld, Michael Zillich, Markus Vincze, Alen Vrečko and Danijel Skočaj

***Abstract*— Semantic visual perception for knowledge acquisition plays an important role in human cognition, as well as in the learning process of any cognitive robot. In this paper, we present a visual information abstraction mechanism designed for continuously learning robotic systems. We generate spatial information in the scene by considering plane estimation and stereo line detection coherently within a unified probabilistic framework, and show how spaces of interest (SOIs) are generated and segmented using the spatial information. We also demonstrate how the existence of SOIs is validated in the long-term learning process. The proposed mechanism facilitates robust visual information abstraction which is a requirement for continuous interactive learning. Experiments demonstrate that with the refined spatial information, our approach provides accurate and plausible representation of visual objects.**

I. INTRODUCTION

Knowledge extension through interactive continuous learning is a desirable property of any cognitive robot. As the most important knowledge resource, visual perception for cognitive robots has received widespread attention in the last decades [1][2][3][4]. An interactive learning robotic system, which is capable of obtaining information from visual sensors as well as information provided by a human teacher, can facilitate and increase the robustness of the knowledge extension process. It also requires sophisticated functionality from the underlying visual system:

1) The bottom-up visual attention mechanism, required to generate focus of attention without any prior information about the objects and scene.

2) The exhaustive modelling of objects in the scene, which forms the underlying base of high-level conceptual properties, such as colour, 3D shape properties and pose.

3) A robust visual system that can handle temporary occlusions and re-identify the objects after re-appearance, since the tutor might remove or introduce new objects in the scene. Note that since there are no detailed models of the objects available, well-developed object detection/recognition/tracking techniques cannot be implemented directly.

To meet these requirements, we design a new visual attention mechanism as the underlying information abstraction system for our robot George[1], depicted in Fig. 1. Our visual information abstraction system computes 3D spatial layout and stereo line features coherently, and yields spaces of interest (SOIs) from the resulting spatial geometry. These SOIs are subsequently validated by tracking them over time, based on persistence, stability and size. As segmentation based on the stereo 3D data alone tends to be imperfect and can include background, especially for weakly textured objects, stable SOIs are augmented with a precise segmentation mask using the graph cut algorithm [5] based on combined colour and 3D information. Object properties to be learned, such as colour and shape, are then extracted based on the segmentation mask.

Fig. 1: Scenario demonstrating interactive learning between robot George and tutor.

The paper is organized as follows. In §II we introduce the background and review state-of-the-art solutions. §III gives an overview of the system competencies. In §IV we describe how to use coherent stereo line detection and plane estimation for reasoning about accurate spatial abstraction and SOIs. The detailed validation of the SOIs in continuous learning, as well as the generation of segmentation masks using SOIs, are outlined in §V. Subsequent sections present the experimental results, evaluations. Conclusions are given at the end of the paper and the future work is shortly discussed as well.

The work was supported by EU FP7 Programme [FP7/2007-2013] under grant agreement No.215181, CogX.

Kai Zhou, Andreas Richtsfeld, Michael Zillich and Markus Vincze are with Automation and Control Institute, Vienna University of Technology, Gußhausstraße 27-29, A-1040, Vienna, Austria {zhou,ari,zillich,vincze}@acin.tuwien.ac.at

Alen Vrečko and Danijel Skočaj are with the Visual Cognitive Systems Laboratory, University of Ljubljana, Slovenia. {alen.vrecko, danijel.skocaj}@fri.uni-lj.si

[1]The robot can be seen in action in the video accessible at http://cogx.eu/results/george.

II. Related Work

In this section, we will present an overview of conventional visual perception systems for robotic learning systems, then introduce recent work on holistic scene understanding from which we draw inspiration.

Due to the lack of high-level representations of visual objects, interactive robotic learning systems usually group coherent low-level features as the visual information abstraction mechanisms for segmenting irregular regions from background (e.g., coloured blobs [6], object proper motion [1], spatial reasoning [2][3] or mixture of models [7][8]). In all these approaches, planar surface estimation for spatial reasoning has attracted the most widespread attention, since the studies in multiple subjects, such as psychology [9], ecology [10], computer vision [11] and robotics [12], have provided evidence that planar surface estimation paves the way to build up the hierarchical structure of a scene which constitutes behaviour-relevant entities as well as dominates man-made real-world environments. However, the aforementioned research obtains visual information using plane estimation for spatial reasoning in *isolation*.

On the other hand, the availability of *coherent* spatial abstraction and object detection can be a crucial advantage for any visual component. This coherent processing, also known as holistic scene understanding can provide significant improvements by considering the relationships governing the structure of the scene (spatial layout, objects in the scene, etc.), thereby improving the performance of each sub-task in the integrated process [13][14][15]. Hence, we unify a generic plane estimation method and a bottom-up stereo line feature detection in a joint probabilistic model to provide refined supporting surfaces. Any parts sticking out from the supporting surface form spaces of interest (SOIs), without regard to its properties. The resulting SOIs are fed into a 2D segmentation scheme for producing accurate object masks, which are used for recognition or learning.

Note that our visual information abstraction system is built atop the CoSy Architecture Schema (CAS) – a distributed asynchronous architecture [16], which facilitates inclusion of other components that could bring additional functionality to the system in a coherent and systematic way (such as navigation and manipulation).

III. System Competencies

The *Visual SubArchitecture (Visual SA)* of our interactive robotic system processes the scene as a whole using stereo pairs of images and provides quantitative analysis of the spaces of interest, which is followed by segmentation of potential objects and local processing. Visual features are then extracted and used for recognition and learning of objects and qualitative visual attributes. Based on the recognition results, a private belief about every object is generated in the mediative layer. The overall data flow of the entire robotic learning system is depicted in Fig. 2 and this paper will only concentrate on the quantitative layer of visual SA. (see [17] for the detailed description and evaluation of our interactive robotics learning system.)

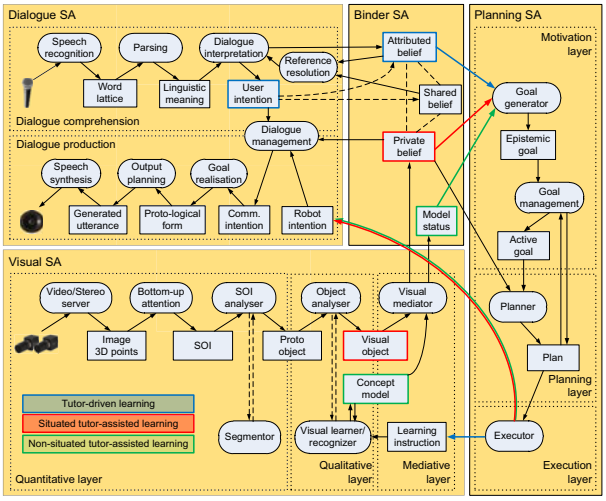


Fig. 2: Schematic system architecture. This paper focuses on the quantitative layer of Visual SubArchitecture (Visual SA).

IV. Visual Information Abstraction

We describe how to detect the stereo lines and estimate planar surfaces independently. The unification of the detected stereo lines and planes for SOI generation will be addressed in the latter part of this section.

A. Stereo Line Detection

The stereo line extraction is a strict bottom-up approach, First, edges are detected from image pairs with an adaptive canny edge detector before we fit lines into the extracted edgel chains using the method of Rosin and West [18]. To estimate 3D information, we have to match the lines of the stereo image pair. For this task, the mean-standard deviation line descriptor (MSLD) of [19] together with the constraint of epipolar lines is utilized in the calibrated stereo camera setup. We then use line-based stereo matching of specific feature points to calculate the proper geometric 3D localization of the lines.

To assess a confidence value for stereo matched lines, we take into account lines that are almost parallel to the epipolar line as well as lines pointing away from the viewpoint typically have higher errors in 3D reconstruction. The angles between the epipolar line and the matched lines in the left and right image (θ_{2Dl}, θ_{2Dr}) as well as the angle between the line and the z-coordinate in the camera coordinate frame (θ_{3Dz}), after normalization between 0 and 1 are used to generate a confidence value:

$$p(f) = \frac{\theta_{2Dl}}{\pi/2} \cdot \frac{\theta_{2Dr}}{\pi/2} \cdot \frac{\theta_{3Dz}}{\pi/2} \tag{1}$$

Note that the resulting value $p(f)$, although in the range of $[0, 1]$, is not a probability. Rather, this value denotes the quality and correctness of the reconstructed lines. Thresholding can produce a true/false judgement, which may be applied in a qualitative reasoning framework, or for learning. We use these quantities in the holistic scene understanding

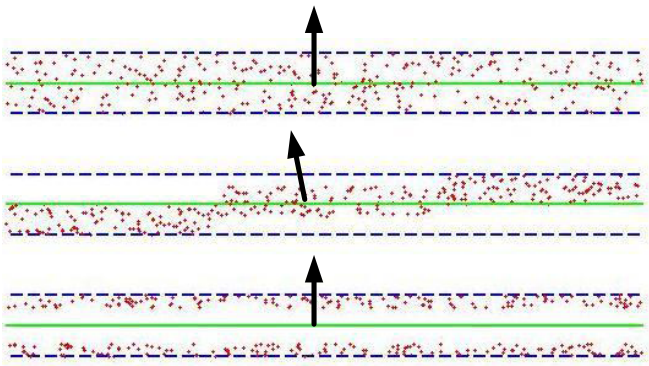


Fig. 3: Three plane estimations (each contains 300 points with Gaussian noise) are displayed. The blue dashed lines are inlier boundaries, and green lines are the side views of the estimated planes. The black arrows denote the average normal vectors $\bar{r}$ of each plane. In the top case, points are evenly distributed and the average normal vector is also approximately equivalent to the normal of the estimated plane. In the center and bottom cases, the data points are unevenly distributed but in different ways. Our assessment criterion (Eq. 2) can effectively distinguish the center uneven case and keep the bottom one as the correct estimation, while the typical evaluation criteria (e.g. the average distance of all the inliers to the estimated plane) cannot.

model as the measure of expected likelihood of the correct line detection, as discussed in §IV-C.

B. Supporting Surface Estimation

It has been verified in [20][21] that taking into account data connectivity in evaluating hypotheses of RANSAC based approaches can significantly improve performance in plane fitting tasks. However, [20] applied CC-RANSAC to detect multiple planes in situations with only two nearby planar patches, such as steps, curbs or ramps. Unfortunately, the estimated results of CC-RANSAC might be unreliable when there are objects on the planar surfaces, especially when objects cluster together on part of the planar surface (e.g. Fig.7). We adopt CC-RANSAC [20] as the underlying plane estimator and assign confidence values to the estimated planes by calculating the average normal vector of connected points. This confidence value is used for the joint probability maximization and will be addressed in detail in §IV-C. Our plane refinement facilitates more reliable estimation than using CC-RANSAC only (experiments in §VI-A).

We start from the RANSAC hypotheses generation and evaluate each hypothesis only on a set of points $C = \{c_i, i = 1, 2, \ldots, m\}$ that belong to the same connected planar component, as in [20]. Consider three points, $X_{C_i}, X_{C_j}, X_{C_k}$, the normal vector of the plane generated by these three points is $r^t_{ijk} = V_{L_{ij}} \times V_{L_{jk}}$, where $V_{L_{ij}}$ is the vector joining X_{C_i} and X_{C_j}. The $X_{C_i}, X_{C_j}, X_{C_k}$ are removed from C and operation proceeds by considering the next three neighboring points and calculating r^{t+1}_{ijk}, which proceeds until there are less than 3 points left in C. The average normal vector $\bar{r}$ of all the points in C is computed using the collection of $\{r^1_{ijk}, \ldots, r^t_{ijk}, \ldots\}$. We define θ_{CS} as the angle between the average normal vector $\bar{r}$ and normal vector n of the estimated plane S, then we have the confidence value for the plane S,

$$Con(S) = (1 - \frac{\theta_{CS}}{\pi/2}) \cdot \frac{k}{N} \tag{2}$$

where k denotes the number of inliers belonging to the estimated plane and N is the number of points in the entire dataset. The first part of Eq. 2 measures how even the points distribute in the inlier boundary (see fig. 3 for better illustration), the second part of Eq. 2 favours planes with more inliers. Eq. 2 in essence represents the continuation and connectivity of all the inliers belonging to the estimated plane. Higher confidence values denote better quality of the estimated plane.

Again the above confidence does not explicitly represent a probability. However, we can use these confidence values to approximate a probability distribution by generating samples around the estimated plane and weighting these samples with confidences. Given the plane S returned by CC-RANSAC, and $\tilde{S}$ a generated sample near S, we formulate the probabilitiy distribution in the following way,

$$\begin{aligned} p(\tilde{S}|Con(\tilde{S})) &= \frac{p(Con(\tilde{S})|\tilde{S})p(\tilde{S})}{p(Con(\tilde{S}))} \\ &= \frac{[(Con(\tilde{S}) > t)]p(\tilde{S})}{p(Con(\tilde{S}))} \end{aligned} \tag{3}$$

Here t is a threshold and $[\,]$ denotes the Iverson bracket:

$$[X] = \begin{cases} 1, & \text{if } X \text{ is TRUE} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

With the Iverson bracket, the probability $p(\tilde{S}|Con(\tilde{S}))$ is proportional to the prior for the sampled plane $\tilde{S}$ whenever $Con(\tilde{S}) > t$, and 0 elsewhere. In other words, $p(Con(\tilde{S})|\tilde{S})$ facilitates thresholding of plane samples with low confidence. We draw samples randomly from the neighboring area of S to generate $\tilde{S}$, and $\tilde{S} \sim \mathcal{N}(\mu_n, \sigma_n)\mathcal{N}(\mu_h, \sigma_h)$, where n and h are the normal vector of plane S, and the distance of plane S to the origin. Hence, $p(\tilde{S})$ is a Gaussian distribution and assigns higher probabilities to the samples near to the estimated plane.

C. Unified Probabilistic Framework

Given the likelihoods for representing the correct detection of the detected stereo lines and estimated planes as shown before, $p(S)$ and $p(E|W)$ denote the prior probability of the plane estimates $S = \{s_i\}$ and probability of image evidences E produced by the stereo line candidates $W = \{w_i\}$. For each line candidate w_i, we introduce a boolean flag t_i, where $t_i = 1$ denotes positive detection of the feature. Therefore, the stereo line detection can be represented with a combination of detection result and assigned flag, i.e. $W = \{w_i\} = \{f_i, t_i\}$, where f is the collection of the feature detection results $\{f_1, \ldots, f_M\}$.

According to Bayes' theorem, $p(E|W) = p(W|E)p(E)/p(W)$, where $P(W|E)$ is the detection's confidence returned by the detector as in §IV-A. And

the $p(E)$ and $p(W)$ can be considered to be uniformly distributed, therefore $p(E|W) \propto p(W|E)$.

With the probabilistic representation of planes and stereo lines, we formulate the joint probability model of the holistic scene as follows,

$$\begin{aligned} p(S,W,E) &= p(S)\prod_{j=1}^{M} p(w_j|S)p(E|w_j) \\ &= \prod_{i=1}^{K} p(\tilde{S}_i|Con(\tilde{S}_i))\prod_{j=1}^{M} p(f_j,t_j|S)p(e_j|f_j,t_j) \end{aligned} \tag{5}$$

where K, M are the number of plane estimates and line candidates, respectively. $p(f_j, t_j|S)$ is the probability of feature detection with the underlying geometry, and denotes the relation between supporting planes and detected features. Since the boolean flag t_j is determined by both scene geometry S and feature detection results $f = \{f_1, \ldots, f_M\}$, and the feature detection process is independent with scene geometry, we have $p(f_j, t_j|S) = p(t_j|f_j, S)p(f_j|S) \propto p(t_j|f_j, S)$. Consequently Eq. 5 can be rewritten as

$$p(S,W,E) \propto \prod_{i=1}^{K} p(\tilde{S}_i|Con(\tilde{S}_i))\prod_{j=1}^{M} p(t_j|f_j,S)p(f_j,t_j|e_j) \tag{6}$$

To sum up, our joint probabilistic model consists of three parts, (1) the probability that the estimated plane is at $\tilde{S}$, (2) the likelihood of positive stereo line detection with the underlying plane estimation, (3) the confidence value of detected lines returned by the stereo line detection algorithm. The first and last probabilities are given using Eq. 3 and Eq. 1 respectively. The second probability is determined by the distance and angle between detected stereo lines and planes:

$$p(t_j = 1|f_j, S) = \begin{cases} |\cos 2\theta_j| \cdot \dfrac{\alpha\varepsilon}{d_j} & \text{if } 0 \le \theta_j < \dfrac{\pi}{4} \\ |\cos 2\theta_j| \cdot \dfrac{\varepsilon}{d_j} & \text{if } \dfrac{\pi}{4} \le \theta_j < \dfrac{\pi}{2} \end{cases} \tag{7}$$

where θ_j is the angle between line j and estimated plane, d_j denotes the distance of the mid-point of the line j to the plane. As defined in RANSAC, the inlier scale parameter ε is used to collect points, which are at a distance smaller than ε from the estimated plane. Eq. 7 in essence gives a higher confidence value to lines which are parallel or perpendicular with the estimated plane, as well as lines which are geometrically close to the plane. Since approximately parallel lines are more likely to be found on top of objects, the distances of these lines to the estimated plane are usually larger than the approximately perpendicular lines. Hence, we use a weight parameter α (empirically set to 10), which denotes that the approximately parallel lines will be taken into account when the distances of these lines to the supporting plane are less than $\alpha\varepsilon$.) to trade off these two kinds of lines.

To maximize the joint probability, we present the optimization problem as $\arg\max_{s_i,t_j}(\ln p(S,W,E))$, the logarithmic formulation can be rewritten as,

$$\begin{aligned} \ln p(S,W,E) = &\sum_{i=1}^{K} \ln p(S_i|Con(S_i)) \\ &+ \sum_{j=1}^{M} [\ln p(t_j|f_j,S) + \ln p(f_j,t_j|e_j)] \end{aligned} \tag{8}$$

where S_i, t_j are the parameters to be estimated. We select the plane which has the highest confidence value of all the plane estimation results, and only consider this plane as the scene geometry for the joint probabilistic model optimization. Then the first part of Eq. 8 is a constant and the second part can be calculated independently through M 3D matched lines comparisons of $\ln p(t_j = 0|f_j, S) + \ln p(f_j, t_j = 0|e_j)$ with $\ln p(t_j = 1|f_j, S) + \ln p(f_j, t_j = 1|e_j)$. After labeling all the stereo lines, the pose of the plane with the highest confidence is refined by searching the nearby planes $\tilde{S}$. This refined pose should satisfy the criterion of maximizing the number of stereo lines parallel or orthogonal to it.

V. SOI Validation And Refinement

In order to autonomously learn visual object concepts, the system needs to tackle obstacles and the variation in positions of the objects due to the interaction of tutor and robot. Since initially there are no models for objects yet, the system cannot rely on model-based recognition, but requires a more general mechanism.

A. SOI Validation

The validation of the SOIs is based on their persistence, stability and size. We use three SOI features to check if the current SOI is matched with any existing one, 1) Jensen-Shannon divergence (JSD) is utilized for measuring the similarity between two normalised colour histograms of current and previous SOIs, 2) the ratio of the number of matched SIFT features in the two SOIs to the number of the SIFT features of the previous SOI; this measurement is only active when there is a reasonable number of SIFT features ($n_F > \tau$). 3) the difference between radii of the bounding spheres of two SOIs divided by the radius of the previous SOI. Given one SOI O from the current SOI list and SOI $\tilde{O}$ from the previous SOI list, then the deviation of these two SOIs $d_{O,\tilde{O}}$ can be computed as follow,

$$d_{O,\tilde{O}} = w_1 D(\mathcal{H}_O, \mathcal{H}_{\tilde{O}}) + w_2 \frac{|n_{F(\tilde{O})} - n_{F(O|}}{n_{\tilde{O}}} + w_3 \frac{|r_{\tilde{O}} - r_O|}{r_{\tilde{O}}} \tag{9}$$

where weight parameters $w_1 + w_2 + w_3 = 1$ and $w_1 = \alpha$, $w_2 = \beta\delta(n_{F(\tilde{S})} > \tau)$, $w_3 = \gamma$, α, β are set to 0.6 and 0.2. $D(\mathcal{H}_O, \mathcal{H}_{\tilde{O}})$ is the Jensen-Shannon divergence of two normalised colour histograms. JSD provides a more appropriate measure of dissimilarity between two color histograms and it is numerically more stable than other measures such as Kullback-Leibler (KL) divergence [22]. In essence, $d_{S,\tilde{O}}$ can be used to examine the similarity of tracked SOIs by considering the difference between their colours, textures and

Fig. 4: Track IDs of detected SOIs across frames.

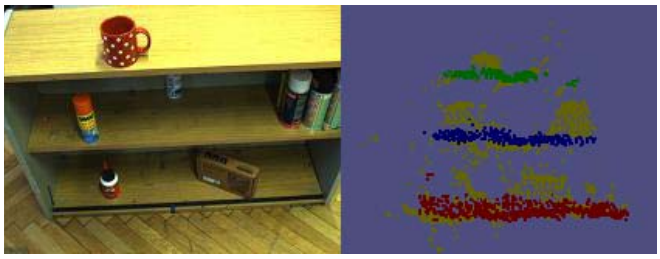


Fig. 5: 3D point cloud representation of the plane estimation results, note that the figure is best viewed in color.

sizes. Fig. 4 demonstrates correct identifications of various SOIs extracted from a video sequence of the scene.

B. Segmentation Mask

The remaining points sticking out from the estimated planes are segmented using 3D flood-filling and the resulting clusters yield SOI bounding spheres. Note that the bounding sphere is taken to be slightly larger than the actual point cluster to ensure that it also contains a part of the plane points, needed for the following segmentation step. Fig. 5 shows a multi-layer shelf scene and corresponding reconstructed point cloud. The detected planes are represented in terms of different colours and remaining sticking out points are shown in yellow. Because of the inherent limitation of stereo reconstruction at poorly textured surface parts and shadowing effects between left and right camera, the resulting SOIs require further refinement using 2D colour based segmentation.

The 2D segmentation is based on *energy minimization with graph cuts*. The back-projected 3D points within the SOI provide colour and spatial cues for the object and its background. The cost function for the object combines the *colour cost* with the *spatial cost*, while the cost function for the background consists of the *colour cost* component only. The *spatial cost* is simply the distance between the point and the object's nearest back-projected 3D point. The *colour cost*, on the other hand, is the average distance between the point's colour and the K nearest colours from the sample (K is determined based on the sample size). Besides foreground and background cost functions, there is a third cost function with a fixed cost to cover those areas where both former functions have high costs. While these areas are considered uncertain and might be resolved on higher levels of the system's cognition, they are deemed as background at this stage by the recognizer.

The distance between two colours is calculated in the HLS colour space:

$$\Delta HLS = \Delta^2 S + (1 - \Delta S)\Delta HL \quad (10)$$

$$\Delta HL = \bar{S}\Delta H + (1 - \bar{S})\Delta L, \quad (11)$$

where ΔH, ΔL and ΔS are the distances between the two HLS colour components, while $\bar{S}$ is the average saturation of the two colours. All the parameters are normalised to values between 0 and 1. The H distance has to be further renormalised and truncated because of its circular parameter space. The contribution of each colour component to the overall distance between the two colours is thus determined by the saturation difference and saturation average.

The code for the graph cut algorithm was kindly provided by Boykov, Veksler and Zabih [5].

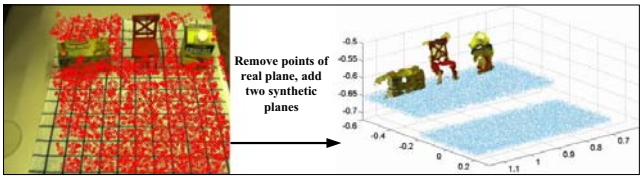


Fig. 6: Generatic synthetic data of two nearby planes.

VI. EXPERIMENTS

In order to compare the spatial abstraction results of the proposed approach with other methods, we first test this with synthetic data. Then experiments with our George robot in continuous interactive learning scenario demonstrate that our visual information abstraction mechanism provides plausible and robust visual object concepts for the continuous learning system.

A. Better Spatial Abstraction

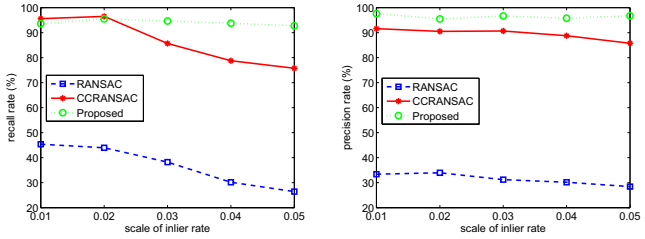
The accuracy of plane estimation is important since the plane estimates are utilized to generate SOIs which are formed by the points sticking out from the supporting planes. Incorrect estimation of supporting planes usually produces incomplete segmentation of objects, which is one of the main sources of failure for our system.

In order to compare the performance of the proposed joint probabilistic approach with CC-RANSAC, we generate a synthetic dataset with noisy 3D points. A simple scene consisting of one supporting plane and object clutter is used. All points belonging to the dominant plane (points shaded red in left image of Fig. 6)) have been manually removed and replaced with two synthetic supporting planar patches (parallel to the original plane), modeling two supporting surfaces at different heights. This synthetic scene facilitates qualitative comparison of CC-RANSAC and the proposed method with different scales of inlier noise. These planar patches have been generated with 15000 points (7500 each), corrupted by Gaussian noise of standard deviation σ. The coloured points (total amount of points of three objects is 8039) in right image of Fig. 6 represent the objects.

In Fig. 7 we compare the plane estimation results of RANSAC, CC-RANSAC and the proposed approach on the

(a) RANSAC (b) CC-RANSAC (c) proposed approach

Fig. 7: Comparison of plane estimation results of RANSAC, CC-RANSAC and the proposed method using synthetic data (side view). Points on the planes are corrupted by Gaussian noise with $\sigma = 0.01$, the height between two planes is $0.05m$. The typical estimation results of the three tested methods are illustrated with red points.



(a) Recall (b) Precision

Fig. 8: Qualitative comparison of RANSAC, CC-RANSAC and the proposed method with various inlier noise scale.

synthetic dataset. The red points represent the typical results of inliers belonging to the detected planes (as seen from the side view) and the proposed method clearly outperforms RANSAC and CC-RANSAC. The estimated plane using CC-RANSAC is tilted towards the objects because of the higher density of points in that area. The isolated plane estimation with CC-RANSAC is also worse because RANSAC based methods always converge to the largest plane near the optimum, which in this case is the diagonal plane.

We compare RANSAC, CC-RANSAC and the proposed holistic method on synthetic data with different inlier noise scale, each method is given 20 trials and the results in average are collected. The recall rate measures the proportion of estimated inliers in actual inliers of the model, and the precision rate presents the proportion of correctly estimated inliers in all the estimated inliers. From Fig. 8 we see with increasing inlier noise scale, the proposed method produces the best plane estimation in terms of accuracy and stability.

B. Robust Robotic Vision

We collected a database of 4650 image pairs of 310 objects (each object is observed from 15 views); some of them are shown in Fig. 9. The results of the subsequent segmentation step are shown in the right part of Fig. 9. On the left side, the SOIs are marked on the original image with IDs (yellow numbers) and the red circles denote the points belonging to the supporting surfaces. The following 2D graph-cut segmentation only processes in the neighbouring area of SOIs. The right side zooms on these area. The top images show the position of backprojected 3D points (light green for object, red for background, dark green for unknown) and the segmentation (grey for object, white for

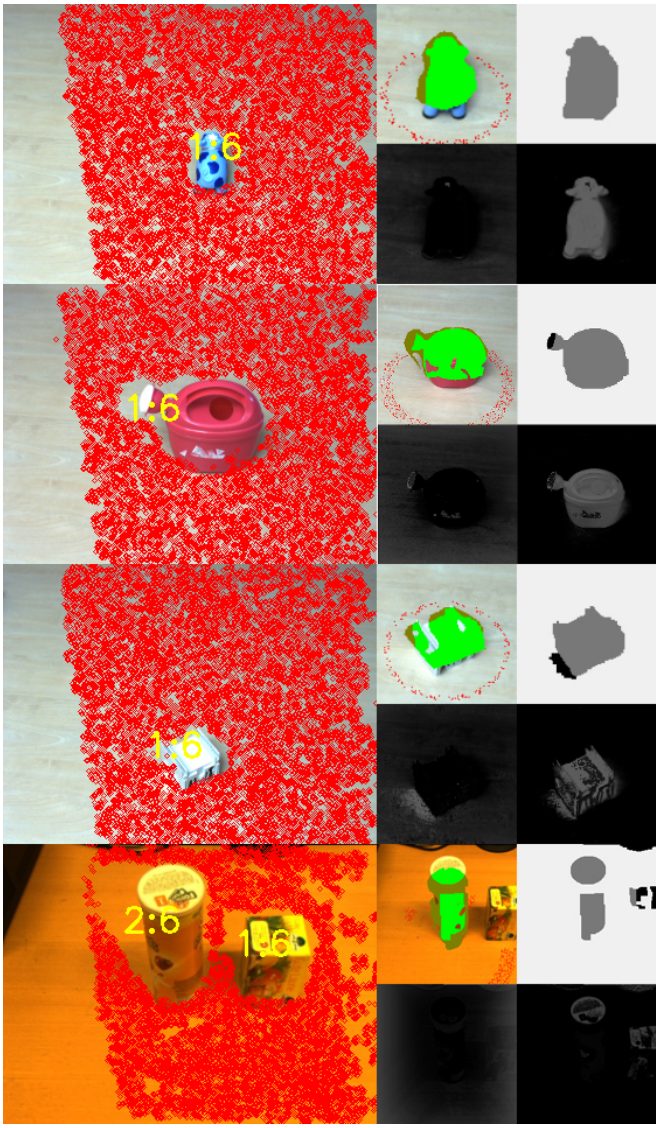


Fig. 9: Observed scene and sample objects with segmentation results.

background), the bottom images represent the graph cut cost functions for object and background where the brighter colour denotes greater cost. We can see that despite the fact that the backprojected 3D points are not very precise due to rather large noise, the graph-cut segmentation can be successfully initialised and provides a precise object contour.

Fig. 10 illustrates the test in the multi-layer shelf scene. The top image shows the backprojected 3D points belonging to the estimated planes, the bottom part of the figure demonstrates sample object segmentations. We observe that the yellow carton box is neglected due to the inherent limitation of the color-based 2D graph-cut segmentation. So we can use the backprojected SOI directly as the object mask in case the graph-cut segmentation returns trivial mask.

VII. CONCLUSION

In this paper, we present a visual information abstraction mechanism and how it performs in a continuously learning

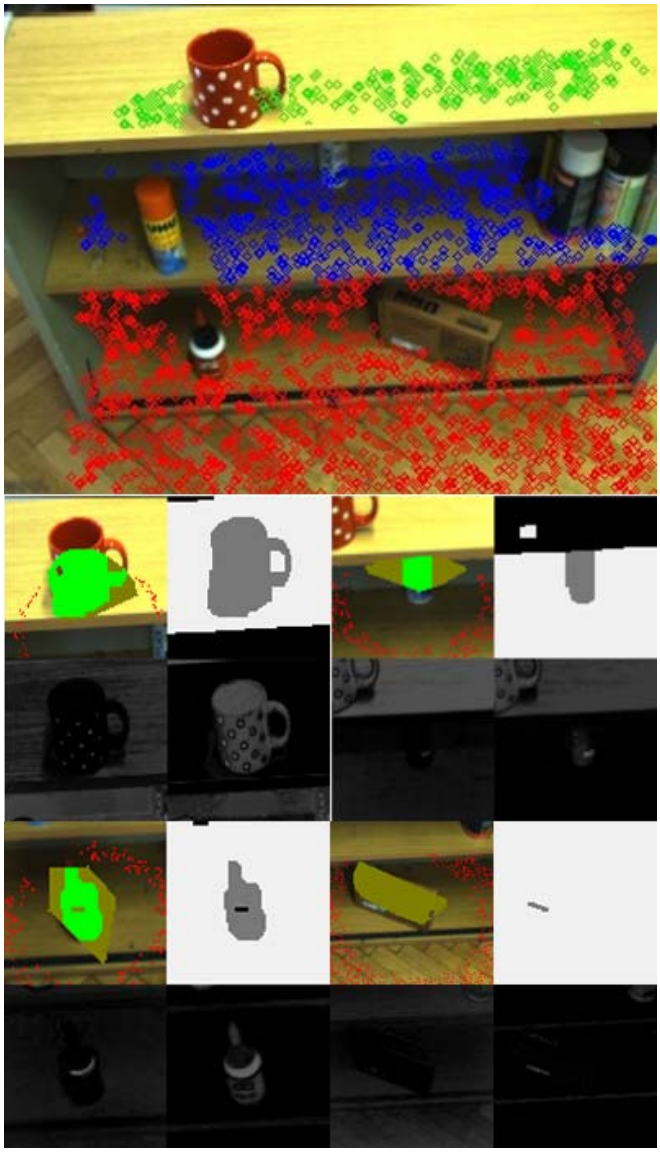


Fig. 10: More complex scene (a multi-layer shelf with sparse objects inside) and sample objects with segmentation results.

robotic system. We generate spatial information in the scene by considering plane estimation and stereo line detection coherently within a unified probabilistic framework, and show how spaces of interest (SOIs) are generated and segmented using the spatial information. We also demonstrate how the existence of SOIs is validated in the long-term learning process. Experiments demonstrate that our system can produce more accurate spatial information, thereby providing robust and plausible representation of visual objects.

Currently, we are investigating the utilization of the proposed visual information abstraction system with a mobile robotic platform which enables our robot to acquire novel information in a more active and autonomous way.

REFERENCES

[1] J. Schmudderich, V. Willert, J. Eggert, S. Rebhan, C. Goerick, G. Sagerer, and E. Korner, "Estimating object proper motion using optical flow, kinematics, and depth information," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 4, pp. 1139 –1151, aug. 2008.

[2] A. Vrečko, D. Skočaj, N. Hawes, and A. Leonardis, "A computer vision integration model for a multi-modal cognitive system," in *The 2009 IEEE/RSJ International Conference on Intelligent RObots and Systems*, October 2009, pp. 3140–3147.

[3] M. Heracles, B. Bolder, and C. Goerick, "Fast detection of arbitrary planar surfaces from unreliable 3d data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009.

[4] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmudderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *Humanoids 2008. 8th IEEE-RAS International Conference on*, Daejeon, South Korea, Dec. 2008, pp. 553–560.

[5] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[6] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, june 2005, p. 89.

[7] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmuedderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2008.

[8] S. Kirstein, A. Denecke, S. Hasler, H. Wersing, H.-M. Gross, and E. Körner, "A vision architecture for unconstrained and incremental learning of multiple categories," *Memetic Computing*, vol. 1, pp. 291–304, 2009.

[9] S. Bertel, *Spatial Structures and Visual Attention in Diagrammatic Reasoning*. Pabst Science Publishers; Lengerich, 2010.

[10] J. J. Gibson, *The Ecological Approach To Visual Perception*. Psychology Press, Sept. 1986.

[11] K. Zhou, M. Zillich, M. Vincze, A. Vrečko, and D. Skočaj, "Multi-model fitting using particle swarm optimization for 3d perception in robot vision," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2010.

[12] K. Sjöö, A. Aydemir, T. Mörwald, K. Zhou, and P. Jensfelt, "Mechanical support as a spatial abstraction for mobile robots," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2010.

[13] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *CVPR 2006*, vol. 2, 2006, pp. 2137 – 2144.

[14] S. Y.-Z. Bao, M. Sun, and S. Savarese, "Toward coherent object detection and scene layout understanding," in *CVPR*, 2010, pp. 65–72.

[15] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool, "3d urban scene modeling integrating recognition and reconstruction," *International Journal of Computer Vision*, vol. 78, pp. 121–141, 2008.

[16] N. Hawes and J. Wyatt, "Engineering intelligent information-processing systems with CAST," *Adv. Eng. Inform.*, vol. 24, no. 1, pp. 27–39, 2010.

[17] D. Skočaj, M. Kristan, A. Vrečko, M. Mahnic, M. Janicek, G.-J. M. Kruijff, M. Hanheide, N. Hawes, T. Keller, M. Zillich, and K. Zhou, "A system for interactive learning in dialogue with a tutor," in *Submitted to The 2011 IEEE/RSJ International Conference on Intelligent RObots and Systems*, 2011.

[18] P. Rosin and G. West, "Nonparametric segmentation of curves into various representations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 12, pp. 1140 –1153, Dec. 1995.

[19] Z. Wang, F. Wu, and Z. Hu, "Msld: A robust descriptor for line matching." *Pattern Recognition*, vol. Vol. 42, pp. 941–953, 2009.

[20] O. Gallo, R. Manduchi, and A. Rafii, "CC-RANSAC: Fitting planes in the presence of multiple surfaces in range data," *Pattern Recogn. Lett.*, vol. 32, pp. 403–410, February 2011.

[21] C. V. Stewart, "Bias in robust estimation caused by discontinuities and multiple structures," *IEEE Transactions on PAMI*, vol. 19, pp. 818–833, 1997.

[22] X. Huang, S. Z. Li, and Y. Wang, "Jensen-shannon boosting learning for object recognition," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, ser. CVPR '05, 2005, pp. 144–149.