From holistic scene understanding to semantic visual perception: A vision system for mobile robot

Kai Zhou, Karthik Mahesh Varadarajan, Andreas Richtsfeld, Michael Zillich, Markus Vincze

Abstract— **Semantic visual perception for knowledge acquisition plays an important role in human cognition, as well as in the many tasks expected to be performed by a cognitive robot. In this paper, we present a vision system designed for indoor mobile robotic systems. Inspired by recent studies on holistic scene understanding, we generate spatial information in the scene by considering plane estimation and stereo line detection coherently within a unified probabilistic framework and indicate how the resultant spatial information can be used for facilitating robust visual perception and reasoning visual elements in the scene. We also demonstrate how the proposed system facilitates and increase the robustness of two robotics applications – visual attention and continuous learning. Experiments demonstrate that our system provides plausible representation of visual objects as well as accurate spatial layout of the scene.**

I. Introduction

Holistic scene understanding has been an ultimate goal of computer vision research for more than five decades [1]. This goal starts from several sophisticated and ambitious algorithms relying on heuristics which attempted understanding 3D scene structure from a single image and in recent years much progress has been made in this field under coherent consideration of the spatial relationship between objects and scene geometry [2][3][4][5]. This combination of isolated object recognition and geometrical contextual analysis provides a robust and efficient solution to the typical chicken-and-egg problem of locating objects and reasoning about 3D scene layout simultaneously. However, this approach is implemented by many researchers for better object recognition/detection results rather than more accurate spatial reasoning, since most of the sophisticated object recognizers/detectors can provide the probabilistic representations of the recognition results and there is no general approach producing spatial abstraction with probabilistic performance measure.

On the other hand, motivated by functional interpretations of spatial language term "on", and the need for cognitively plausible and practical abstractions for mobile robots, the authors of [6] present an important functional spatial relation of mechanical support. They demonstrate that the spatial reasoning through estimation of supporting surfaces is a necessary part of linguistic interaction between robots and human beings. The accurate estimation of the supporting surfaces also paves the way to build up the hierarchical structure of a scene which helps indirect search in the context of mobile robot. Hence, we consider planar surface estimation as the general spatial abstraction and represent the plane estimations probabilistically. We then unify the estimated planes and detected features in a joint probabilistic framework to produce refined supporting planar surfaces, thereby facilitating various robotics tasks, such as robotic visual attention and interactive learning.

The paper is organized as follows. In §II we introduce the background and review state-of-the-art solutions. §III describes how to use coherent stereo line detection and plane estimation for reasoning about accurate spatial abstraction. Subsequent sections present the experimental results with synthetic scene, and real robotic applications. Conclusion is given at the end of the paper.

II. Related Work

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

Mobile robotic systems usually group coherent features as the visual information abstraction for segmenting irregular regions from background (e.g., coloured blobs [7], object proper motion [8], saliency detection [9], spatial reasoning [10] or mixture of models [11]). 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 [12], computer vision [13] and robotics [6], 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 [4][2]. 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.

Note that our visual information abstraction system is built atop the CoSy Architecture Schema (CAS) – a distributed

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

All authors 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

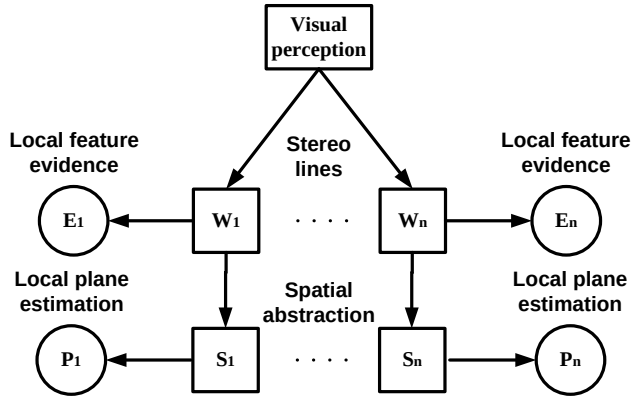


Fig. 1: Graphical model of conditional independence of competencies in our visual perception system.

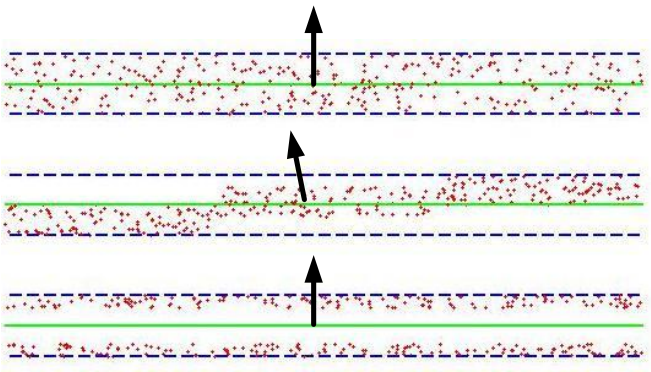


Fig. 2: Three plane estimations (each contains 300 points with Gaussian noise) are displayed. The blue dashed lines are inlier boundaries, and green lines are the side views of the estimated planes. The black arrows denote the average normal vectors $\bar{r}$ of each plane. In the top case, points are 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) can not.

asynchronous architecture [14], 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. Visual Perception System

Our visual perception system processes the scene as a whole using stereo pairs of images to detect the stereo lines and estimate 3D planar surfaces, which is followed by coherent consideration of these two elements as well as their relations. The overall processing and conditional independence of competencies are depicted in Fig. 1.

We first describe how to detect the stereo lines and estimate planar surfaces independently. The unification of the detected stereo lines and planes for producing holistic scene understanding 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 [15]. 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 [16] 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 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:

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

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

B. Supporting Surface Estimation

The RANSAC [17] algorithm is commonly applied to estimate planes with noisy point cloud data, and numerous extensions and modifications have been derived from it. and It has been verified in [18][19] that taking into account data connectivity in evaluating hypotheses of RANSAC based approaches can significantly improve performance in plane fitting tasks. However, [18] 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.4). We adopt CC-RANSAC [18] 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 §III-C. Our plane refinement facilitates more reliable estimation than using CC-RANSAC only (experiments in §IV).

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 [18]. 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. 2 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 §III-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

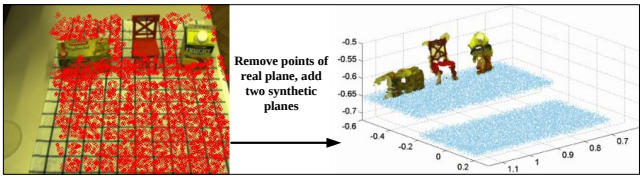


Fig. 3: Generatic synthetic data of two nearby planes.

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,

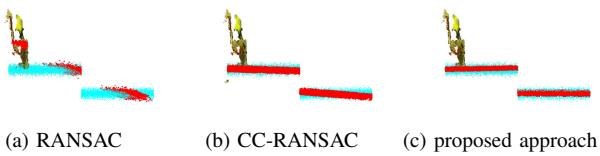
$$\begin{aligned}\ln p(S,W,E) = &\sum_{i=1}^{K} \ln p(S_i|Con(S_i)) \\ &+ \sum_{j=1}^{M} [\ln p(t_j|f_j,S) + \ln p(f_j,t_j|e_j)]\end{aligned} \tag{8}$$

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.

IV. Evaluation With Synthetic Scene

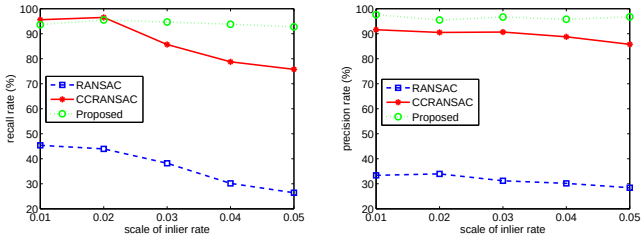
In order to compare the performance of the proposed joint probabilistic approach with CC-RANSAC [18], we generate a synthetic dataset with noisy 3D points. A simple scene consisting of one supporting plane and object clutter is used. All points belonging to the dominant plane (points shaded red in left image of Fig. 3)) 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 number of points of three objects is 8039) in right image of Fig. 3 represent the objects.

In Fig. 4 we compare the plane estimation results of RANSAC, CC-RANSAC and the proposed approach on the synthetic dataset. The red points represent the typical results of inliers belonging to the detected planes (as seen from



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

Fig. 4: 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. 5: Qualitative comparison of RANSAC, CC-RANSAC and the proposed method with various inlier noise scale.

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 scales, 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. 5 we see with increasing inlier noise scale, the proposed method produces the best plane estimation in terms of accuracy and stability.

V. Robotic Visual Attention Application

Typical robotic visual attention mechanisms generate saliency maps from 2D image features, often over several scales [20][9]. We combine our spatial abstraction from the estimated planes with the 2D saliency-based method, implemented by suppression of saliency belonging to the plane area (most likely to be the texture on the supporting surfaces) and encouraging the saliency near the objects features. First, the filtered features are separated into two categories – potential object features and plane texture features, according to the distances of features' center points to its supporting plane. Then all the potential object features are smoothed with a 2D Gaussian filter to generate object likelihood as in [21]. Fig. 6 illustrates the generation of object likelihood for a multi-layer shelf scenario. Our CogX

robot[1] was supposed to search for various objects located on arbitrary supporting surfaces using the proposed approach. The search results, processing data flow and comparison with 2D-feature saliency [20] are shown in Fig. 7.

VI. Interactive Robotic Learning Application

An interactive learning robotic system 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) *Instantaneous* knowledge acquisition of objects at the first available learning opportunity.

The proposed visual perception mechanism meets all the aforementioned requirements – it produces the accurate supporting planar surfaces by considering the *bottom-up* stereo line detection and plane estimation coherently, it facilitates *automatic* trigger of interactive robotic learning by grouping the points "sticking out" from the estimated planes. These remaining points are segmented using 3D flood-filling and the resulting clusters yield the space of interest (SOI) bounding spheres, which contain *exhaustive* information of the potential objects.

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. 8 shows a multi-layer shelf scene and corresponding reconstructed point cloud. The detected planes are represented in terms of different colours and remaining points "sticking out" 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.

Fig. 9 illustrates the test in the multi-layer shelf scene, it demonstrates sample object segmentations. In each pane, the top images show the position of reprojected 3D points (light green for object, red for background, dark green for unknown) and the segmentation (grey for object, white for 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 reprojected 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. 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 reprojected 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 detail how it performs in two real robotic tasks: robotic visual attention and continuously learning. We generate spatial information in the scene by considering plane estimation and stereo line detection coherently within a unified probabilistic framework, and show how the resultant spatial information can be used indirectly for facilitating more accurate visual perception or be used directly for reasoning visual elements in the scene. Experiments demonstrate that our system can produce more accurate spatial information, thereby providing robust and plausible representation of visual objects.

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

Fig. 8: 3D point cloud representation of the plane estimation results, note that the figure is best viewed in color.

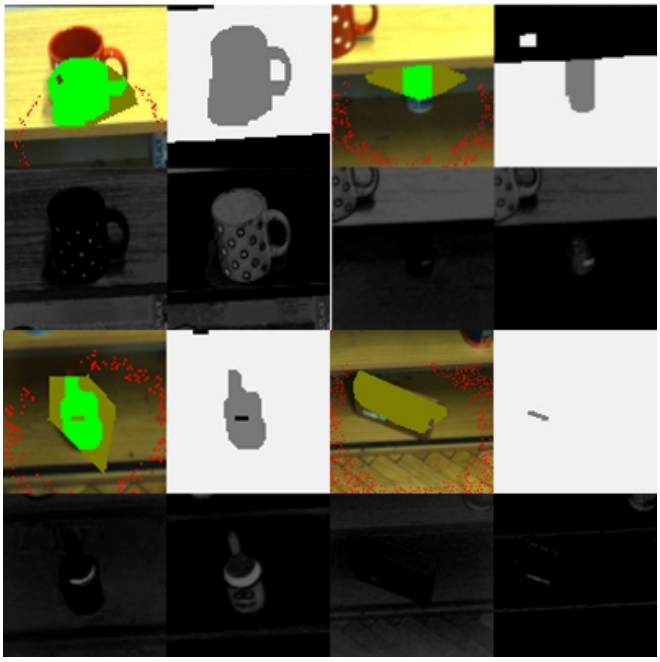


Fig. 9: Sample objects with segmentation results in multi-layer shelf scene.

References

[1] L. G. Roberts, "Machine perception of 3d solids," Ph.D. dissertation, Dept. of Electrical Engineering, Massachusetts Institute of Technology., 1963.

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

[3] M. Sun, S. Y.-Z. Bao, and S. Savarese, "Object detection with geometrical context feedback loop," in *BMVC*, 2010, pp. 1–11.

[4] D. Hoiem, A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, pp. 151–172, 2007.

[5] S. Helmer and D. Lowe, "Using stereo for object recognition," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, May 2010, pp. 3121–3127.

[6] K. Sjöö, A. Aydemir, T. Mörwald, K. Zhou, and P. Jensfelt, "Mechanical support as a spatial abstraction for mobile robots," in *2010*

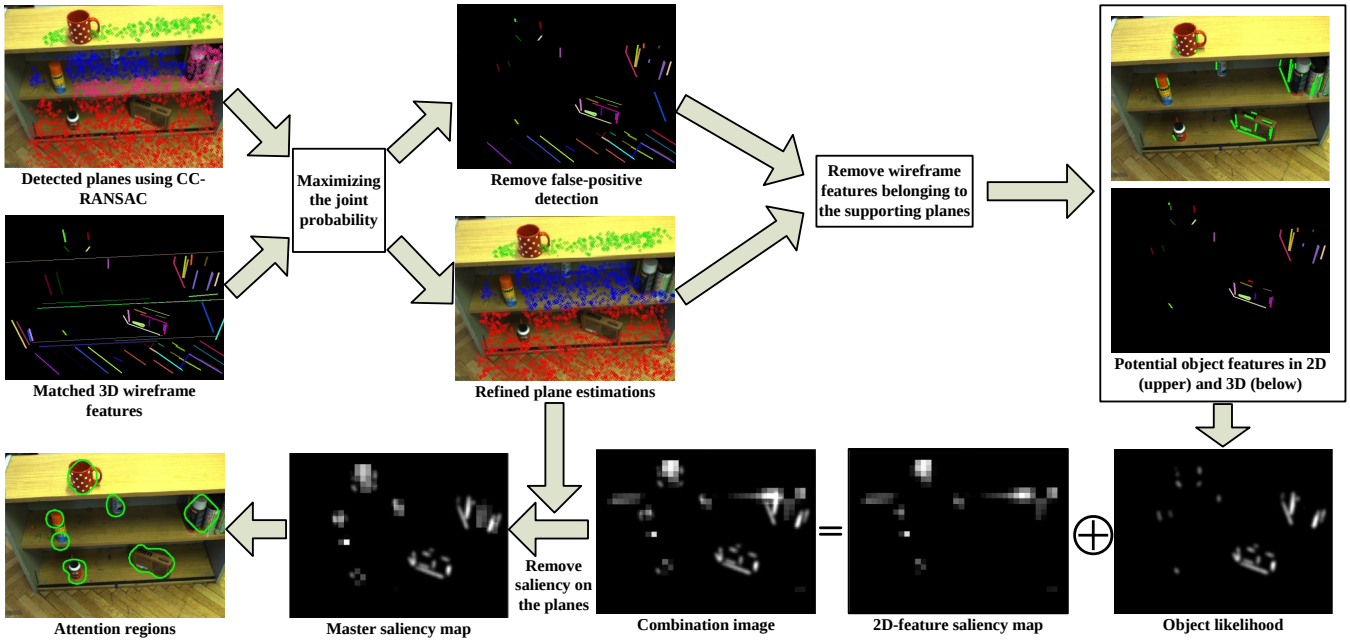
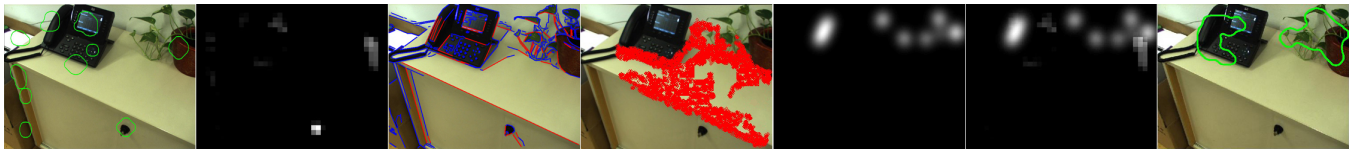


Fig. 6: By combining stereo line detection and plane estimation within a unified probabilistic framework, we generate the potential object features and refined plane estimations as spatial abstraction. The 2D saliency attention mechanism is then improved with the abstracted spatial information.



(a) Scene 1: telephone & plant

(b) Scene 2: beverages

Fig. 7: From left to right: attention using 2D saliency; 2D-feature saliency map; stereo lines (matched features marked in red); estimated plane; object likelihood; master saliency map; attention region returned by our approach

IEEE/RSJ International Conference on Intelligent Robots and Systems, October 2010.

[7] 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.

[8] 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.

[9] C.-K. Chang, C. Siagian, and L. Itti, "Mobile robot vision navigation & localization using gist and saliency," in *Intelligent Robots and Systems, 2010. IROS 2010. IEEE/RSJ International Conference on*, Oct 2010.

[10] 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.

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

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

[13] 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.

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

[15] 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.

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

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] 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.

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

[20] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[21] C. Choi and H. Christensen, "Cognitive vision for efficient scene processing and object categorization in highly cluttered environments," in *IROS*, 2009, pp. 4267–4274.