Structure Verification toward Object Classification using a Range Camera

Stefan Gächter, Ahad Harati, and Roland Siegwart
Autonomous Systems Lab (ASL)
Swiss Institute of Technology, Zurich (ETHZ)
8092 Zurich, Switzerland
{gaechter, harati, siegwart}@mavt.ethz.ch

***Abstract*— This paper proposes an incremental object classification based on parts detected in a sequence of noisy range images. Primitive parts are jointly tracked and detected as probabilistic bounding-boxes using a particle filter which accumulates the information of the local structure over time. A voting scheme is presented as a procedure to verify structure of the object, i.e. the desired geometrical relations between the parts. This verification is necessary to disambiguate object parts from potential irrelevant parts which are structurally similar. The experimental results obtained using a mobile robot in a real indoor environment show that the presented approach is able to successfully detect chairs in the range images.**

INTRODUCTION

Object recognition and classification are long standing issues in computer vision. They date even before popularization of the mobile robotics. In the current literature, most of the approaches rely on appearance of the objects. Although appearance gives useful hints about the nature of the object, yet it alone is not enough to achieve object classification under view point and illumination changes as demanded in mobile robotic applications. More importantly is the ability of abstraction needed in object classification, where structural variations within a certain class of object should be handled. As one step beyond object recognition, in object classification, 3D perception seems necessary for dealing with objects of the real world.

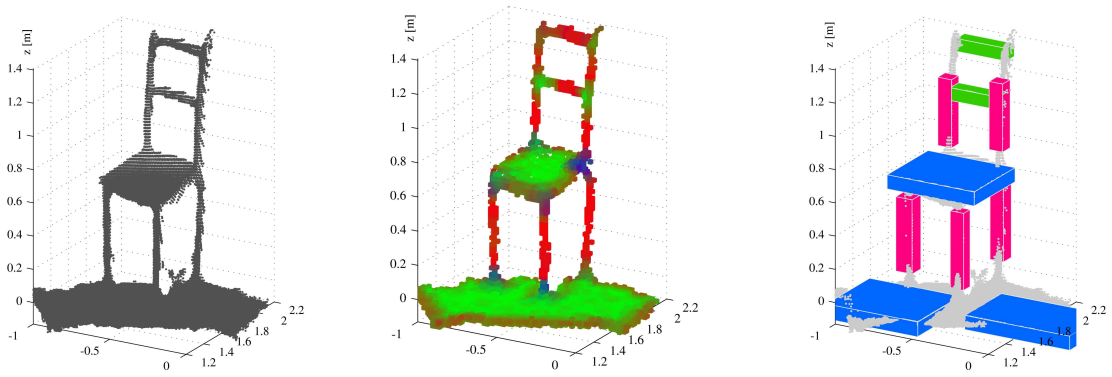
Structure variability within a class of objects may be well explained using a geometric grammar, preferably probabilistic to take care of uncertain and incomplete measurements. In such an approach, object classification is reduced to detection of some object parts and verification of the required geometric relations among them. However, robust detection of complex object parts has more or less the same nature as object classification itself. Therefore, primitive parts are used that encode the overall structure of the object parts as bounding-boxes. Such simplified geometric models are easy to detect in point cloud observations and deliver independence with regard to appearance making the object classification more practical. Considering a chair for example, stick-like shapes are primitive parts which may correspond to a chair leg or something structurally similar. Having a proper probabilistic geometric grammar as discussed in [1], it would be possible to obtain the most probable hypotheses of the object and its parts from the set of detected primitive parts.

This work was partially supported by the EC under the FP6-IST-045350 robots@home project.

In recent years, a novel type of range camera has emerged on the market which makes it possible to capture 3D scenes on mobile robotic platforms [2]. Although very compact, light and capable of measuring distances up to several meters at high frame rate, lower measurement quality in general [3] poses great challenges in using such devices in an object classification framework. Therefore, the main goal and contribution of this paper is to bring well grounded approaches from different domains together, extend and adapt them to a novel object classification framework which can work with poorly observed scenes using a mobile robot equipped with a range camera. In a broader sense, the approach presented here is motivated toward autonomous navigation and semantically annotated maps.

The proposed approach can account for different views of the same object and for variations in structure, material, or texture of the objects of the same kind as far as the decomposition of the objects into its parts is known. The decomposition itself, that is the grammar, may be learned out of some training examples [1]. However, here an a-priori defined grammar is used and the focus remains on dealing with part detection and structure verification in realistic cases.

To avoid the challenging segmentation of noisy range images for each primitive part, a track-before-detect scheme is implemented using a particle filter which accumulates the information of the local structure – represented in form of shape factor, a local disparity measure – and estimates pose and extension of the potential primitive parts over time. This is a common approach in radar applications, where a target has to be jointly tracked and classified in highly noisy data [4], [5]. To realize the observation function of the particle filter, a classifier is trained using support vector machine for each part category. Therefore, different types of primitive parts are detected in parallel while a structure verification procedure based on a voting scheme periodically apply the considered grammar to come up with the pose of potential objects in the scene. Thus, although single observations are too poor to infer the presence of any object directly, the presented approach incrementally collects the



(a) Point Cloud (b) Voxel Set (c) Estimated Parts

Fig. 1. (a) Single point cloud and (b) a quantized version of a sequence of five registered range images at step $k = 25$. The colors in (b) indicate the shape factors: red for *linear* like, green for *planar* like, and blue for *spherical* like local structures. (c) Estimated primitive parts at step k = 25. The color indicates the number of hypothetical parts encoded by a particle: green for 2, blue for 3, and magenta for 5 states.

evidences from the sequence of range images and tracks the hypothetical primitive parts leading to object hypotheses.

The approach presented here is quite general in handling different object parts with simple geometry. However, through out this paper, chairs consisting of legs, a seat and a back support are chosen as example objects to demonstrate the method. In the next section, the particle filter based primitive part detection approach, originally presented in [6], is briefly explained.

I. Particle Filter based Primitive Part Detection

In this approach, part detection is formulated as tracking hypothetical bounding-boxes in a sequence of voxelized point clouds using a particle filter. The details of the algorithm can be found in [6]. Here, a brief summary is given.

A. Incremental State Estimation

When dealing with noisy observations, it may not be possible to detect primitive parts in single observations. Thus, the detection performance can be improved by tracking potential targets over time and using the accumulated information. The track-before-detect concept has already been studied in radar applications [5]. The same concept is realized here by a particle filter which is extended to handle multiple primitive parts of the same type:

$$\begin{aligned} p(\mathbf{y}_k|\mathbf{Z}_{k-1}) &= \int p(\mathbf{y}_k|\mathbf{y}_{k-1})p(\mathbf{y}_{k-1}|\mathbf{Z}_{k-1})d\mathbf{y}_{k-1} \\ p(\mathbf{y}_k|\mathbf{Z}_k) &\propto p(\mathbf{z}_k|\mathbf{y}_k)p(\mathbf{y}_k|\mathbf{Z}_{k-1}), \end{aligned} \tag{1}$$

where $\mathbf{y}_k = [R_k, \mathbf{x}_{1,k}^\mathsf{T} \dots \mathbf{x}_{r_k,k}^\mathsf{T}]^\mathsf{T}$ is the augmented state, which contains the current estimate of number of object parts present in the view R_k and their bounding-box parameters $\mathbf{x}_{i,k}$ at step k. Similar to [7], the number of primitive parts R_k is modeled by a Markov chain with a predefined transition matrix, where the state value at step k is a discrete number $r_k = \{0, \dots, M\}$ with M being the maximum number of parts expected in each view. Here, M varies between 4 and 8 depending on the part to detect.

B. Feature Vector

The shape factors characterize the local part structure by its linear, planar, or spherical likeliness. They are calculated for each voxel using its surrounding spatial voxel distribution by the decomposition of the distribution into the principal components:

$$r_l = \frac{\lambda_1 - \lambda_2}{\lambda_3 + \lambda_2 + \lambda_1} \tag{2}$$

$$r_p = \frac{2(\lambda_2 - \lambda_3)}{\lambda_3 + \lambda_2 + \lambda_1} \tag{3}$$

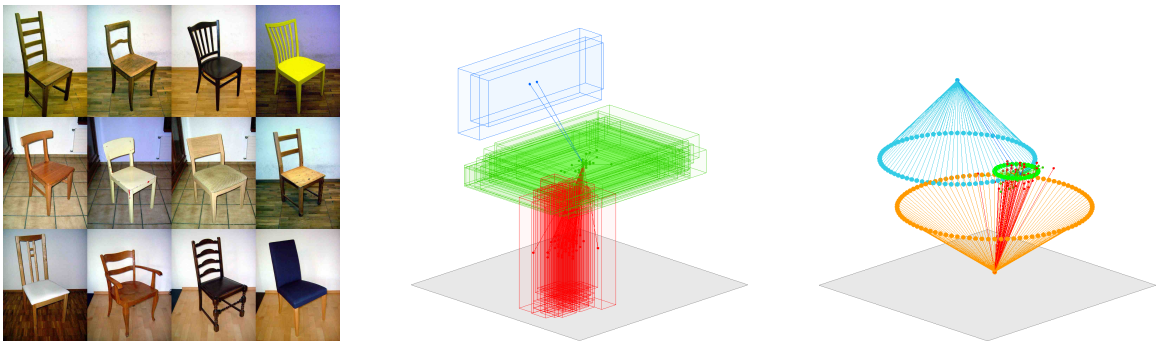
$$r_s = \frac{3\lambda_3}{\lambda_3 + \lambda_2 + \lambda_1}. \tag{4}$$

where λ_i are the ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$. r_l, r_p, and r_s express local similarity to linear, planar, and spherical shapes respectively. Figure 1(a) depicts the original point cloud acquired with a range camera and figure 1(b) depicts the corresponding shape factor colored voxel set of a chair, where for each voxel the shape factor was computed according to (4).

The shape factor distribution in the region of interest defined by the bounding-box is approximated by a histogram to obtain a unique feature vector that models an object part. This approach is inspired by the work done in [7], while color is replaced by shape factor. In addition, dimensionality reduction is applied on the histogram to obtain a compact representation. The feature vector is composed of the compact histogram and six simple geometric features to account for the occupancy and eccentricity of the voxel distribution in the bounding-box.

C. Integration of the Observations

The observation likelihood function generates the importance weights used to incorporate the measurement information $\mathbf{z}_k$ in the particle set. A support vector classifier [8] is trained to detect the primitive parts in the sparse and noisy data. In 2(a), twelve different chairs are depicted which are used to obtain voxel images from different viewpoints and to build a manually labeled training set.



(a) (b) (c)

Fig. 2. (a) Training set of twelve chairs. (b) Extracted object parts and their relative position with respect to the chair center. (c) Implicit shape model of the chair taking into account the part symmetry. Vote vectors for the leg are depicted in orange, the ones for the seat in green, and the ones for the back in blue. Darker colored vectors indicate the original votes.

In the detection framework, the observation likelihood is usually defined as a ratio of the probability that an object part is present to the probability of its absence. This is equivalent to the ratio of the classification probabilities computed with the learned classifier. Assuming that the classification can be done independently for each hypothetical part and considering the probability as a distance $a_{i,k}$ in the range of $[0, 1]$, the unnormalized importance weight $\tilde{\pi}_k$ for each particle is computed as:

$$\tilde{\pi}_k = \begin{cases} 1, & R_k = 0 \\ \exp\left(-\frac{1}{b}\sum_{i=1}^{r_k}(1-2a_{i,k}) + r \cdot c\right), & R_k > 0 \end{cases} \quad (5)$$

where b is a parameter to adjust the observation sensitivity and c accounts for the a-priori knowledge. Figure 3(a) depicts the outcome of the leg classifier for each voxel considering a fixed bounding-box; this is treated as the observation likelihood in the particle filter framework. Different primitive parts are detected with particle filters working in parallel using classifiers trained on different training sets. For the demonstrated chair example, leg, seat, and back classifiers are considered which are mainly detecting vertical stick-, horizontal plate-, and stick-like structures.

II. Object Structure Verification

The detection algorithm provides different primitive parts as can be seen in figure 1(c). The encoded knowledge of the structure in the grammar is used to disambiguate primitive parts belonging to the object from the rest. In the case presented in figure 1(c), this means to reject planar patches detected on the ground that are structurally similar to the seat or leg like structures supporting the back. This process is called structure verification, which applies the grammar constraints. It is implemented as a voting scheme similar to the *implicit shape model* as presented in [9]. However, the algorithm has been extended to the 3D case.

A. Structure Model

The implicit shape model as used here is a probabilistic, non-parametric model which encodes 3D structure of the object in terms of relative location of every part with respect to a pre-defined reference point, here the center of the seat. The implicit shape model consists of a set of object specific parts $\mathcal{S} = \{s_j\}$ and a set of corresponding votes $\mathcal{V} = \{v_j\}$. This model is learnt by memorizing all relative locations of the parts as depicted in figure 2(b) for the chair collection.

The votes can be seen as a sample-based representation of a spatial probability distribution $p(o, \mathbf{x}_o | s_j, \mathbf{x}_j)$ for each s_j at a given position $\mathbf{x}_j$, where $\mathbf{x}_o$ denotes the reference point of the object o. In the present case, the object is a chair and the primitive parts s_j are of leg-, seat-, or back-like structure. Thus, the implicit shape model represents a set of a-priori known part types, where each part position is represented by a probability distribution. This representation can be seen as the dual of the constellation-type model [10], where the parts are defined with a fixed location but variable appearance.

Here, the spatial probability distribution for each part is assumed to be Gaussian which is represented by the collected votes during the training. Since the orientation of the object relative to the observer is arbitrary, some symmetry is added to the probabilistic model. In other words, each learnt Gaussian is rotated along the z-axis and for each part a donut shape is obtained in the voting space. The resulting shape model for the chair collection of figure 2(a) is depicted in figure 2(c).

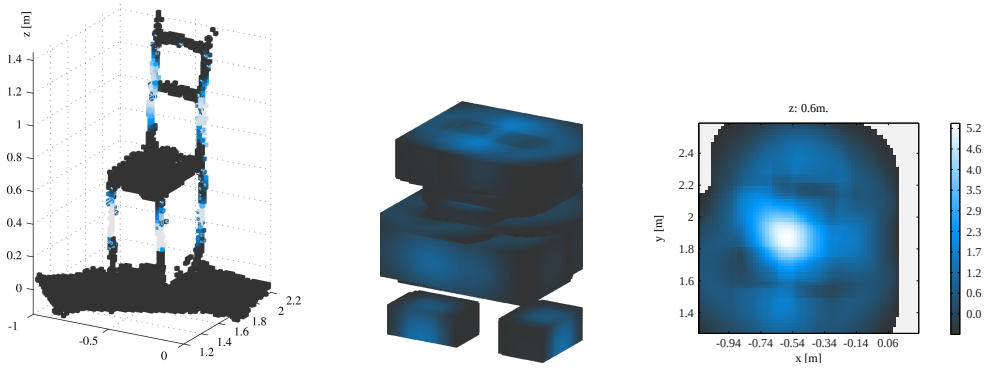
B. Structure Verification

In the original implicit shape model approach, each interest point that match with an entry in the codebook cast its votes. Similarly, according to the estimated state, any detected part primitive serves as an interest point. The estimated state $\hat{\mathbf{x}}_k^r$ for each part type is obtained as

$$\hat{r}_k = \arg\max_i \frac{\sum_{n=1}^{N} \delta(R_k^{(n)}, i)}{N}, \quad (6)$$

$$\hat{\mathbf{x}}_k^r = \frac{\sum_{n=1}^{N} \mathbf{x}_k^{(n)} \delta(R_k^{(n)}, \hat{r}_k)}{\sum_{n=1}^{N} \delta(R_k^{(n)}, \hat{r}_k)}, \quad (7)$$

where $\hat{r}_k$ is the maximum a-posteriori estimate of the number of present primitive parts at step k. For each estimated part,



(a) (b) (c)

Fig. 3. (a) Observation likelihood for the leg classifier using a fixed bounding-box size. Brighter colors indicate higher classification probabilities. (b) Snapshot of the 3D voting space and (c) a slice where the maximum detection likelihood occurs. The detection likelihood is normalized such that it reflects best the contribution of each primitive part. The maximum is located where the votes converge from five parts: 3 legs, 1 seat, and 1 back part.

votes are casted according to the learnt distribution and the uncertainty obtained from the particle filter, see figure 2(c). The votes of each primitive part can be weighted according to the ratio of contributing particles, i.e. its detection certainty. However, since different classes of primitive parts are estimated independently, such weights should be normalized properly. In the experiments presented here, all votes are considered with equal weights. A sample snapshot of the voting space for the estimated parts of figure 1(c) is depicted in figure 3(b). The presented slice in figure 3(c) clearly shows the aggregation of the votes in the center of the seat as desired.

Once the voting space is populated, the local likelihood maxima indicate potential object locations. The structure verification is completed by searching for these maxima. The current implementation uses a non-maxima suppression technique. Then, the learnt object parts are projected back and compared with part estimations using a bounding-box check. If the projected and estimated parts are the same, it is assumed that the estimated part belongs to the sought object.

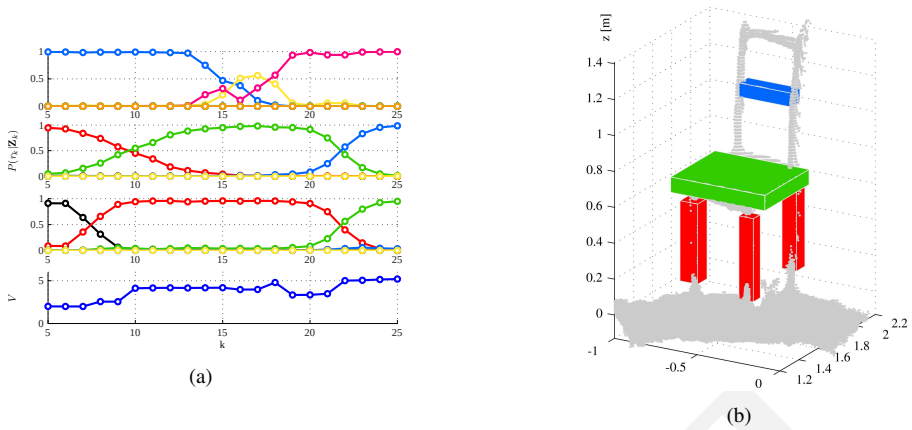
III. Experiments

The above discussed structure verification method is exemplified with a chair. Chairs in reality are designed with various shapes and structures. Here, they are modeled with three different kind of bounding-boxes to cover the stick and plate like structures of the chair legs, seats, and backs. For each object part class, an independent particle filter is used for the detection. The parameter setting follows according the recommendations in [6]. The outcome of the part detection undergoes the structure verification.

Two experiments are performed with the *range camera* mounted on a *robot* at height of about 1.1 m facing downward with a tilt angle of about 15°. In the first experiment, only one chair is in the scene while in the second and third experiment the robot is observing a round dining table, two chairs and a coffee table in the cafeteria of our lab. In all experiments, the robot slowly approaches the objects in the scene recording range images and odometry at about 2 Hz. Totally 200 and 450 range images are captured in the first and second experiment respectively. Because of occlusions and the narrow field of view of the camera, the number of hypothetical chair parts in the view varies considerably when robot moves through the scene. Hence, the algorithm should dynamically adapt to what momentarily is present in the view.

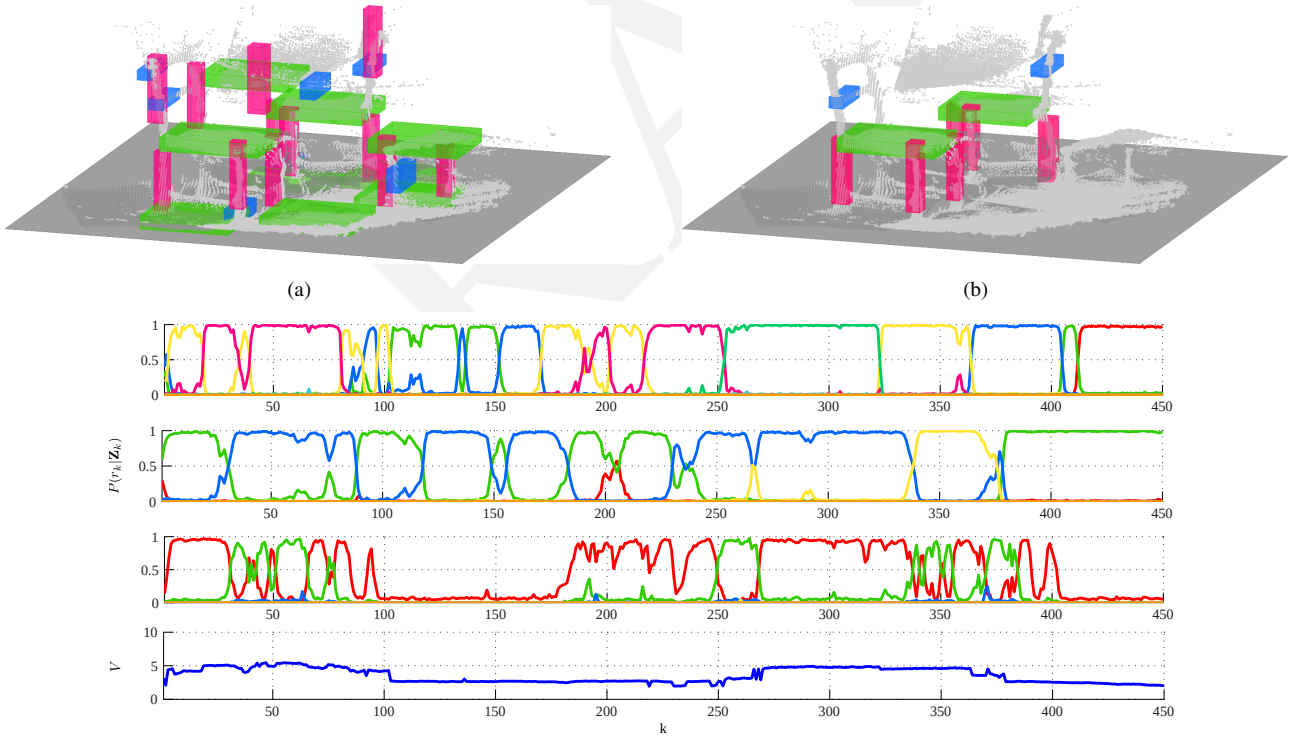
The result of the first experiment is depicted in figure 4. Figure 4(a) depicts the evolution of the part presence probability for the three primitive part types over time. The number of all potential parts – leg, seat, and back – increase over time. Accordingly, the maximum vote strength depicted in the last row increases too. Thus, the evidence of having a chair present increases when having more parts present that vote for the same object center. All detected part like structures at step 25 are depicted in figure 1(c). The corresponding votes casted for each part are depicted in figure 3(b). The chair structure is verified with respect to the position of the maximum detection likelihood depicted in figure 3(c). The resulting chair parts are depicted in figure 4(b). Chair legs, seat, and back are extracted correctly. The back consist of one stick like part as memorized by the implicit shape model depicted in 2(b).

In the second experiment with a more realistic scenario, the robot is faced with the challenge of object part detection and structure verification in the cafeteria. In figure 5(a), the detected primitive parts before the structure verification are depicted overlaid with two original point clouds. Depicted are the primitive parts with the probability larger than 0.5. In figure 5(b), the object parts after the structure verification are depicted. The verification was done for hypothetical objects with detection likelihood maxima higher than 2.9. The two chairs are successfully detected despite missing parts and considerable number of detected additional primitive parts. The algorithm has to deal with many appearing and disappearing parts as can be seen in figure 5(c). The upper three graphs depict the probabilities of the number of primitive parts present in the view for leg-, seat-, and back-like parts. The probabilities oscillate where the scene changes



(a) (b)

Fig. 4. Results of the first experiment. (a) Evolution of the part presence probabilities over time for the leg, seat, and back like parts in the first three rows. The last row indicates the maximum detection likelihood. The color indicates the number of hypothetical parts encoded by a particle: red for 1, green for 2, blue for 3, yellow for 4, magenta for 5, and cyan for 6 primitive parts. (b) Estimated object parts after the structure verification. Legs are depicted in red, seat in green, and back in blue.



(a) (b)

(c)

Fig. 5. Results of the second experiment. (a) Detected primitive parts before the structure verification step for scene with a round dining table, two chairs and a coffee table. (b) Estimated object parts after the structure verification step. Legs are depicted in red, seat in green, and back in blue. (c) Evolution of the part presence probabilities over time for the leg, seat, and back like parts in the first three rows. The last row indicates the maximum detection likelihood. The color indicates the number of hypothetical parts encoded by a particle: red for 1, green for 2, blue for 3, yellow for 4, magenta for 5, and cyan for 6 primitive parts.

considerably. As in the first case, the evidence of having a chair present is correlated with the number of primitive parts present in the view.

IV. Conclusion

This paper presented an incremental object classification based on parts. Primitive parts are detected using an extended particle filter with a support vector classifier based observation. A voting scheme is applied to verify the structure of the object. The provided experimental results show that the approach detects successfully the chairs even though the hypothetical parts vary considerably in the view.

However, the method needs further testing and improvements for its robust application in robotics. The structure verification step has to be refined to cope with higher intra class variability or with multiple objects. Therefore, the probabilistic grammar has to be extended and a sophisticated parser designed. The implicit shape model could be then used as a proposal distribution for such a parser.

References

[1] M. A. Aycinena, "Probabilistic geometric grammars for object recognition," Master's thesis, Massachusetts Institute of Technology - Department of Electrical Engineering and Computer Science, 2005.

[2] MESA Imaging AG, Switzerland, http://www.swissranger.ch/ (13.9.2007).

[3] T. Kahlmann, "Range imaging metrology: Investigation, calibration and development," Ph.D. dissertation, Eidgenössische Technische Hochschule Zürich, ETHZ, Diss ETH No 17392, 2007.

[4] Y. Bar-Shalom and X. R. Li, *Estimation and Tracking: Principles, Techniques, and Software*. Artech House, 1993.

[5] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter - Particle Filters for Tracking Applications*. Artech House, 2004.

[6] S. Gachter, A. Harati, and R. Siegwart, "Incremental object part detection toward object classification in a sequence of noisy range images," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2008.

[7] J. Czyz, B. Ristic, and B. Macq, "A color-based particle filter for joint detection and tracking of multiple objects," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 2, 2005, pp. 217–220.

[8] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science - National Taiwan University, Tech. Rep. July 18, 2007, 2007.

[9] B. Leibe, A. Leonardis, and B. Schiele, "An implicit shape model for combined object categorization and segmentation," in *Towards Category-Level Object Recognition*. Springer, 2006, pp. 496–510.

[10] R. Fergus, P. Perona, and A. Zisserman, "Semi-supervised learning and recognition of object classes," in *Cognitive Vision Systems*, H. Nagel, Ed. Kluwer Academic Publishers, in press – 2006.