# Incremental Object Part Detection with a Range Camera

**Report**

**Author(s):**
Gächter, Stefan

**Publication date:**
2007

**Permanent link:**
https://doi.org/10.3929/ethz-a-010113627

# Incremental Object Part Detection with a Range Camera

*Stefan Gächter*
*Autonomous Systems Lab (ASL)*
*Swiss Institute of Technology, Zurich (ETHZ)*
*8092 Zurich, Switzerland*
*stefan.gaechter@mavt.ethz.ch*
*http://www.asl.ethz.ch*

*2007-9-1*
*ETHZ-ASL-2006-12 — Version 2.0.2*

**Abstract**

This report presents an object part detection method using a particle filter. The method is adapted to a range camera that provides 3D information with a high data rate. However, the data is affected by considerable measurement noise and distortion. Thus, the range data is quantized to cope more efficiently with the high data volume and segmented into primitive parts with morphological operators to assure processing speed. Measurement noise, outliers and segmentation errors are handled with a particle filter used here as a soft decision tree to detect object parts over several frames.

## Contents

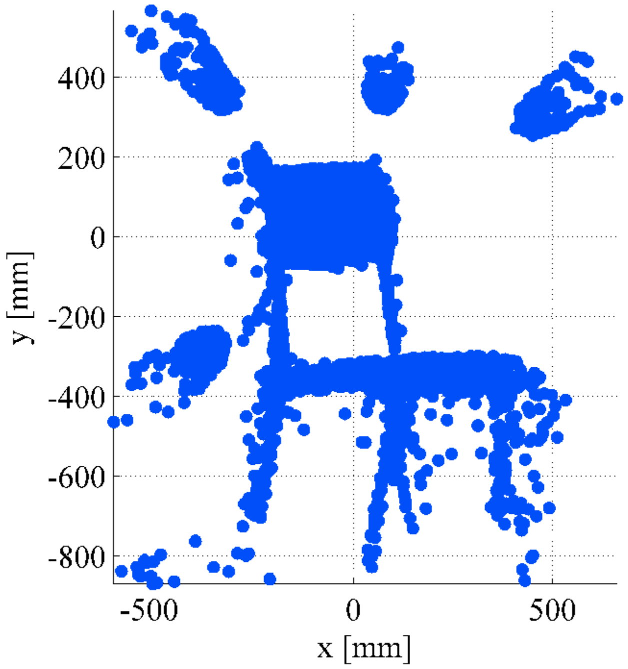
# 1 Introduction

The majority of current works in object recognition and classification is appearance based. However, geometric models can play an important role in the detection of objects in cluttered scenes. Geometric models can account for the different views of the same object, such as pose and illumination, and for variations of the object in an object class, such as structure, material, and texture. Often, the perfect reconstruction of the three-dimensional geometry of the environment is not necessary, but an approximate model is sufficient. In Sudderth et al. (2006) a method has been proposed that estimates the 3D pose of familiar objects from monocular images to enhance appearance based object detection. However, the method uses binocular images for the training process. Instead, a range camera could be used. A range camera provides depth information with a higher data rate but lower accuracy than a traditional laser scanner. As it is shown in this work, the measurement accuracy is sufficient enough to robustly detect primitive parts of an object. To do so, it is necessary to filter the registered and segmented range images to distinguish between useful and spurious parts. The approach chosen is an incremental estimation of the presence of object parts using a particle filter. Thus, the particle filter is used here as a sort of soft decision tree similar as in Schubert and Sidenbladh (2005), where a particle filter is applied to do clustering. The implementation of the particle filter is kept simple. Only a single particle set is used to track the multiple hypothesis of object part presence in a sequence of range images. The use of particle filters for object detection has been earlier proposed, see for example Czyz (2006). In the present work, a similar approach is pursued but extended to handle multiple object parts of different types in a 3D space. A chair is chosen as an example object to demonstrate the method. However, the approach presented here can be generalized and applied to a larger class of objects.

The report is organized as follows. In the next section, the range camera is introduced. The measurement principle and limits are discussed. The object part segmentation is presented in section 2. The detection method with a particle filter is presented in section 3. Section 4 is devoted to the experiment and section 5 concludes the report.

## 1.1 Range Camera

In recent years, a novel type of range camera to capture a 3D scene emerged on the market. The measurement principle is based on time-of-flight using modulated radiation of an infrared source. The emitted radiation is reflected by objects and projected onto a CMOS/CCD imager. The phase shift between emitted and received signal for each pixel results in the distance measurements. In general, the camera provides range and reflectance images. The former encodes the depth whereas the latter the received signal strength information. Most manufacturers - see for example MESA Imaging AG, Canesta Inc., PMDTechnologies GmbH, Matsushita Electric Industrial Co. Ltd, Sharp Co., 3DV Systems - use similar processes to produce the imagers or to measure the phase shift and, thus, the various cameras performance is expected to be similar. However, different efforts are made to suppress ambient light or to miniaturize the camera. One of the smallest range cameras is the SR-3000 made by MESA Imaging AG, see figure 1(b). For the work presented here, the SR-2 of the same manufacturer is used, which exhibits similar measurement performance for the application in question.



(a) Front view of a 3D point cloud acquired with a range camera.

(b) Range camera SR-3000.

Figure 1

The camera has a resolution of 124 × 160 pixels with approximately a maximum measurement distance of 8m. The intrinsic and extrinsic camera parameters have to be calibrated to increase the measurement accuracy. The intrinsic camera parameters are similarly calibrated using a checkerboard as described in Kahlmann et al. (2006). The extrinsic parameters are similarly calibrated as described in May et al. (2006). The distance measurement depends on an offset that varies with the range. This is one major parameter to calibrate. It is done by capturing a series of images of a reference plane at a distance between 0.5 . . . 1.5m. The range dependency of the offset is almost linear, but varies for each pixel as the signal strength decreases for the imager peripheral area. The range calibration can be done only approximately, because it also depends on other parameters such as the integration time. However, the integration time cannot be adjusted for each pixel individually. For the present work, a sufficient low value of the integration time is chosen to avoid signal saturation.

Despite the calibration, the range image remains affected by noise, outliers and distortions. The statistical measurement error in the center of the imager is about 1.5cm standard deviation in the range of interest. The signal-to-noise ratio is low, because the emission power is limited to comply with eye safety standards. The low emission power is especially a problem when the signal is absorbed or deviated by the object's surface. Further, the imager has a limited resolution. Thus, the range camera can hardly capture jump edges and integrates instead over the gaps. These mixed measurement points are clearly visible in the point cloud for a scene with a chair and four reference spheres, see figure 1(a).

Another source of error is the camera's settling time to reach the thermal equilibrium of more than 10 min. Therefore, the range measurement can have a drift over time. These errors make it hard to get an accurate and consistent image for different viewpoints of the same scene. However, the acquired information is still rich enough to estimate an object's structure, even though the geometry is not precise.

## 2 Object Part Segmentation

The range camera provides a stream of images. The goal is to detect in this stream primitive parts that belong to an object. In case of a simple chair, these primitive parts are leg, back, and seat. The range images are registered and transformed into a 3D point cloud. The point cloud is quantized and segmented with morphological operators. For each resulting part, spurious or not, a shape factor is computed. This factor is a measure of similarity that a certain part belongs to a certain class of primitive parts. Along with the primitive part position, this information is accumulated over time using a particle filter. An object part is detected, if enough evidence for a certain primitive part is accumulated.

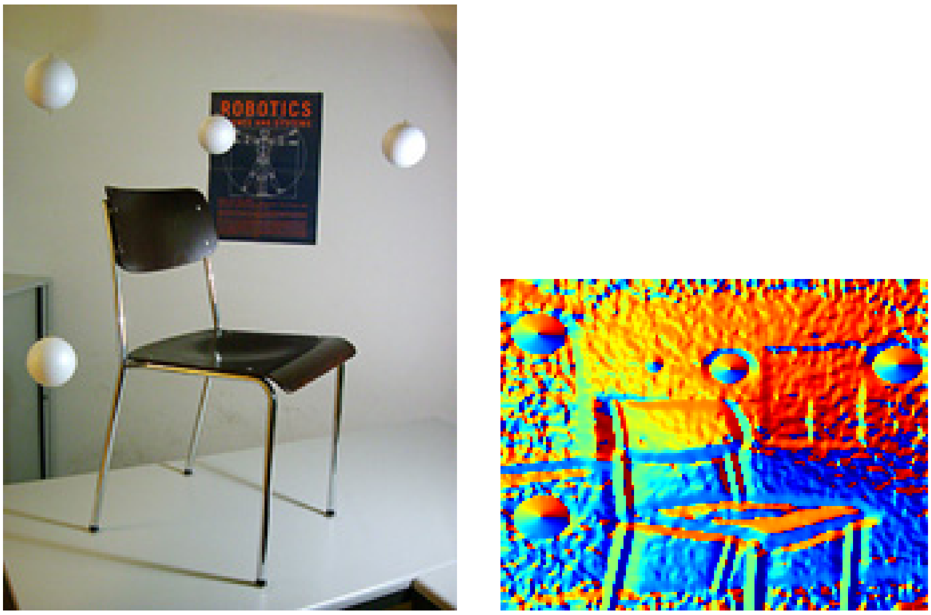
### 2.1 Registration

The object part detection has to cope with the limited field-of-view of the range camera and, more importantly, with the occlusion of object parts. This is due to occlusion from other objects and also self-occlusion. In mobile robotics, these problems can be partially solved by changing actively the camera's pose and registering the images consistently. A well known approach for the fine registration of rage images is the Iterative Closest Point (ICP) algorithm. In the present case, where the noise level and outlier ratio is high, the ICP-based approach fails or converges to a local minimum. Because image registration is not the main issue of the present work, artificial features are used to define a global reference frame, see figure 2(a). Thus, with a static camera, an image sequence of the experimental setup is acquired and the sphere's 3D position is estimated. These positions are used later on to align the range images when the camera is moving around the object.

The artificial features are spheres, because their shape is viewpoint independent. The sphere localization in the images is done by template matching in the gradient angle image. The gradient of the range and reflectance image is computed using a Gaussian kernel. The argument for each directional vector results in the gradient angle image with pixel values between 0 and $2\pi$, see figure 2(b). Because the directional vectors are all pointing toward the center of a sphere, a unique pattern results for the range and reflectance image. Then, the sphere positions can be tracked by template matching. Because only sparse and noisy measurements are captured for each sphere, fitting to estimate the center mostly failed. Therefore, only the mean value of few neighboring points on the sphere surface are used and corrected by the radius to compute the sphere's center. The $z$-direction of the reference frame is aligned with the two spheres in plumb line. The $z$-direction is later used to distinguish between the orientation of the different object parts.

### 2.2 Segmentation

Because the range camera provides a high data volume, the 3D point clouds are quantized and grouped into voxels, see figure 3(a). The voxel



(a) Experimental setup consisting of a chair and four suspended spheres used as artificial features.

(b) Gradient angle image of the experimental setup.

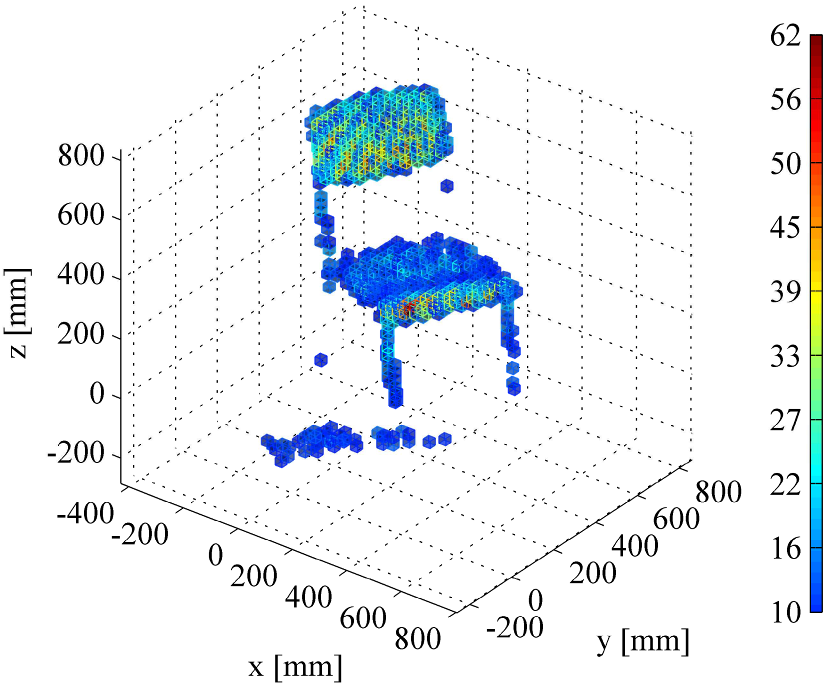
Figure 2

set is created incrementally as follows: a newly acquired point cloud is added to the voxel set, kept during ten time steps, and then removed from the voxel set. Voxels with less than ten points are discarded as noise.

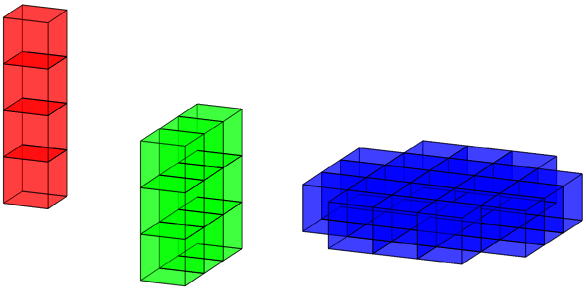
The quantized point clouds can be treated as discrete 3D gray-scale images. An efficient method to segment gray-scale images are morphological operators, see Dougherty and Lotufo (2003). The voxel set is segmented into primitive parts that constitute the objects. In the present work, the primitive parts *leg*, *back*, and *seat* are used. The object part segmentation combines the dilation of the voxel set with the structuring elements that are reflecting the shape of the primitive parts, see figure 3(b). The morphological operators are effective in the orthogonal directions. The segmentation results in an approximated structure for each part of the object. The structure is supposed to be sufficiently accurate to perform part detection. The presented segmentation method is simple and fast, but results in spurious parts. Thus, the detection algorithm has to cope with such errors.

## 2.3 Shape Factor

The segmentation results in a set of primitive parts. In the case of the chair, it is necessary to associate with each part a measure to judge, if the primitive part is a leg, back, or seat. Thus, the shape factor is computed for each part. The following shape classes are specified based on the local spatial voxel distribution of each primitive part. The local spatial distribution is captured by the decomposition of the voxel's 3D coordinates into the principal components - a set of ordered eigenvalues and -vectors. The primitive parts can have linear - stick like - or planar - plate like - shape. Therefore, the shape factor is first computed for the



(a) Voxel set of ten aligned 3D point clouds. Voxels with lower and higher point density are depicted in blue and red, respectively. The minimum number of point per voxels is ten. Voxels belonging to the background are removed automatically.



(b) The three structuring elements reflecting the shape of the primitive parts *leg*, *back*, and *seat*.

Figure 3

linear $c_l$ and planar $c_p$ case:

$$\begin{aligned} c_l &= \log_{10} \frac{s^2_{max}}{s_{med}\, s_{min}}, &(1)\\ c_p &= \log_{10} \frac{s_{med}\, s_{max}}{s^2_{min}}, &(2)\end{aligned}$$

where $s^2$ are the eigenvalues. The two shape factors tend toward zero, if there is no dominant direction, i.e. the segmented object part has a blob like shape. The two cases are further divided to distinguish between a back and seat or a leg or arm rest of a chair. If $c_l > c_p$, the weighting factor to distinguish between a vertical or horizontal stick like shape is the angle between the normalized eigenvector in the dominant direction and the unit vector in $z$-direction: $\alpha_l = 2/\pi \arccos |\mathbf{e}_{max} \cdot \mathbf{u}_z|$. Accordingly, if $c_p > c_l$, the weighting factor to distinguish between a vertical or horizontal plate like shape is: $\alpha_p = 2/\pi \arccos |\mathbf{e}_{min} \cdot \mathbf{u}_z|$.

The shape factors can be expressed as shape probabilities. Therefore, the shape factors are transformed into values with the range $[0, 1]$ using $d = \gamma(1 - \exp(-c^2/a^2))$, where $a$ defines how fast the shape probability converges to one and $\gamma$ defines a factor with the range $[0, 1]$ that can be used to account for the size and shape of the structuring elements. Thus, three shape probabilities can be defined for the three primitive parts of a chair:

$$\beta^r = \begin{cases} (1-\alpha_l)\, d_l & : \; r = leg \\ \alpha_p\, d_p & : \; r = back \\ (1-\alpha_p)\, d_p & : \; r = seat \end{cases} . \qquad (3)$$

The probability that a segmented object part is $r = noise$ is then $\beta^r = 1 - d_l$, if $c_l > c_p$, and $\beta^r = 1 - d_p$, if $c_p > c_l$. With this definition, the shape probabilities sum up to one for all possible cases.

## 3 Object Part Detection

When the method so far is applied to a registered and quantized point set, the result may be ambiguous, because only a simple segmentation and classification method is applied to noisy and distorted data. However, it is likely that the performance would be improved, if the information from several range images is used incrementally over time. Thus, the object part detection can be formulated in the framework of recursive Bayesian estimation, see Ristic et al. (2004). The type of an object part $R_k$ can be modeled as a Markov system, where, in case of the chair, the state values $r$ are associated with the three possible types of primitive parts and noise: $r = \{noise, leg, back, seat\}$. The position of the part - the centroid of the voxel coordinates - can be modeled with a random vector $\mathbf{x}_k$. The probability of object part presence at time $k$ is then the marginal of

$$P(R_k = r|\mathbf{z}_{1:k}) = \int_V p(R_k = r, \mathbf{x}_k|\mathbf{z}_{1:k}) \mathrm{d}\mathbf{x}_k, \qquad (4)$$

that is the marginal of the joint probability of part position $\mathbf{x}_k$ and type $R_k$ given a range image sequence $\mathbf{z}_{1:k}$ over a region of interest $V$. The solution to find $p(R_k = r, \mathbf{x}_k|\mathbf{z}_{1:k})$ can be done in a recursive prediction and update procedure using a particle filter with the augmented particle state $\mathbf{y}_k^{(n)} = [\mathbf{x}_k^{\mathrm{T}}, R_k]^{\mathrm{T}}$ that consists of, both, the continuous valued part position and the discrete valued part type. Thus, the particle filter approximates the posterior density $p(\mathbf{y}_k|\mathbf{z}_{1:k})$ by a weighted set of $N$ random samples or particles. The evolution of each particle state through time is defined by the transition probability function $p(\mathbf{y}_k|\mathbf{y}_{k-1})$ - the relation among particle states over time - and the observation likelihood function $p(\mathbf{z}_k|\mathbf{y}_k)$ - the relation between particle state and the measurement.

The particle filter discussed in the following is an adaption of a sampling-importance-resampling filter as presented in Isard and Blake (1998), that has been extended to multiple targets in Koller-Meier (2000), with the multiple-model approach as presented in Ristic et al. (2004). Multiple-model means in the present context that the filter deals with a particle state of continuous and discrete values. The main steps of the particle filter are *initialization*, *propagation*, *observation*, and *selection*.

### 3.1 Transition Model

In this work, the scene is static. Changes in the state are due to camera noise, drift and segmentation inaccuracies. Thus, the transition model of the part position is the linear model $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{v}_{k-1}$, where $\mathbf{v}_{k-1}$ is the process noise assumed, for simplicity, to be white, zero-mean Gaussian with covariance matrix $\mathbf{C}_u = \sigma_u^2 \mathbf{I}$. The transition probability density function is $p(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_{k-1}, \mathbf{C}_u)$.

The model of the object part type $R_k$ is the Markov system defined by the transition matrix $\mathbf{T}_u$ assumed to be symmetric: $t_u^{i,j} = 1 - m$ for $i = j$ and $t_u^{i,j} = m/(K-1)$ for $i \neq j$, where $m$ is the probability of miss-classification of a primitive part $\{i, j\} \in \{noise, leg, back, seat\}$. For example, a $i = leg$ is classified as a $j = back$ with probability $t_u^{i,j}$. $K$ is the number of discrete states and is here $K = 4$. The Markov system can be extended to any number of primitive parts and is not restricted to the four mentioned here.

### 3.2 Observation Model

The observation likelihood function relates the current measurement with current state of object part position and type. In the present case, the observation likelihood function generates the importance factors used to incorporate the measurement $\mathbf{z}_k$ in the particle set $\{\mathbf{y}_k^{(1)}, \ldots, \mathbf{y}_k^{(N)}\}$. The particle filter has to cope with multiple segmented parts simultaneously. One possibility is to associate with each observation a particle filter. However, it will be likely that one object part may appear as hypothesis of another one and the particle filters will represent the same distribution. Then, only the computational cost will increase and a management system is necessary to deal with the individual particle filters. Another possibility is to augment the dimensionality of the particle state with each new observation. However, the size of the particle set $N$ has to be taken much larger and can become prohibitive, see Schubert and Sidenbladh (2005). Thus, a single particle filter is used for multiple parts. The probability density function $p(\mathbf{y}_k|\mathbf{z}_{1:k})$ represents multiple part states simultaneously and, therefore, is multi-modal.

The observation model of the position of an object part is the linear model $\mathbf{z}_k = \mathbf{x}_k + \mathbf{w}_k$, where $\mathbf{w}_k$ is the measurement noise assumed, for simplicity, to be white, zero-mean Gaussian with covariance matrix $\mathbf{C}_z = \sigma_z^2\mathbf{I}$. Thus, the observation probability density function is a mixture of Gaussians $p(\mathbf{z}_k|\mathbf{x}_k) = \sum_P \mathcal{N}(\mathbf{x}_k - \mathbf{z}_k, \mathbf{C}_z)/P$ over all observed object parts $P$.

The observation likelihood function must consider the object part type. Therefore, the mixture of Gaussians is weighted with the shape probabilities $\beta^r$ presented in section 2.3. The importance factor for each particle in the set is then

$$w_k^{(n)} = \sum_P \beta_k^r \exp\left(-\frac{1}{2}(\mathbf{x}_k^{(n)} - \mathbf{z}_k)^{\mathrm{T}}\mathbf{C}_z^{-1}(\mathbf{x}_k^{(n)} - \mathbf{z}_k)\right). \tag{5}$$

With this definition, the importance factor takes large values in the 3D space where a primitive part is present, otherwise the noise is favored.

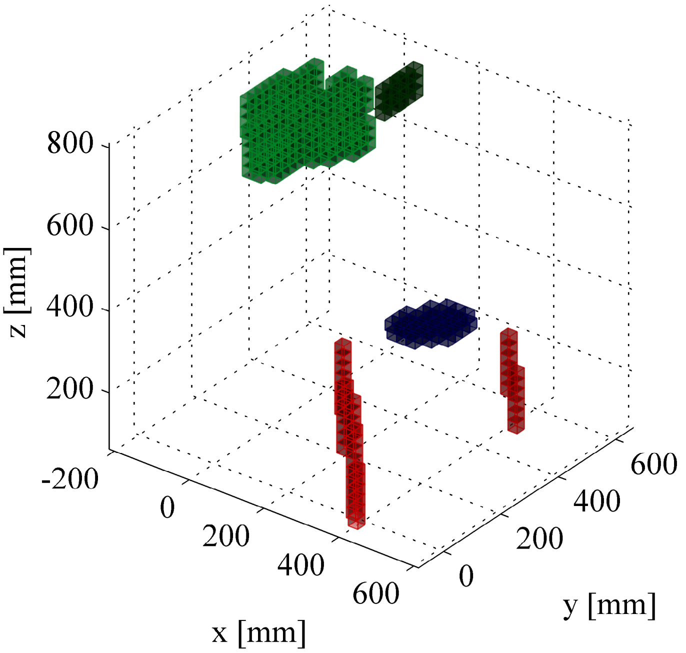
### 3.3 Initialization

Considering the output of segmentation through the time, parts will appear and disappear. Because a single set of particles is used, the filter has to be initialized appropriately to include a new observation. Therefore, an initialization density $\tilde{p}(\mathbf{y}_{k-1}|\mathbf{z}_{k-1})$, which describes the probability of having a part with state $\mathbf{y}_{k-1}$, when only the observation $\mathbf{z}_{k-1}$ is available, is computed at every time step $k$. This function is combined with the posterior density $p(\mathbf{y}_{k-1}|\mathbf{z}_{k-1})$. The particle set is augmented by $M$ particles drawn from the initialization density, then the combination is done during the factor sampling of $N$ particles from the augmented particle set.

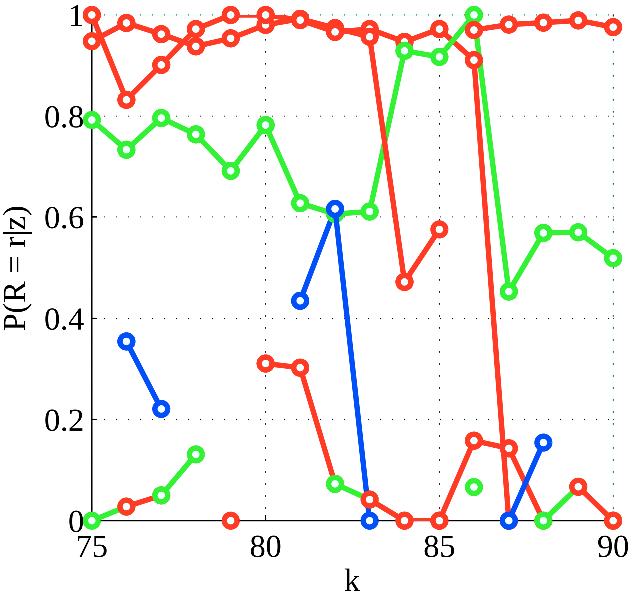
A new observation exists if in the neighborhood of $\mathbf{z}_{k-1}$ only few particles are present. In the current work, if the number of particles is below a certain threshold, the particle set is augmented with a set of samples drawn from a Gaussian distribution $\mathcal{N}(\mathbf{z}_{k-1}, \mathbf{C}_x)$, where $\mathbf{C}_x = \sigma_x^2\mathbf{I}$, for the part position and from a uniform distribution $\mathcal{U}(\{noise, leg, back, seat\})$ for the part type.

## 4 Experiment

The above described incremental object part detection method is applied to a series of about 300 range images taken of a chair by moving the camera by hand from the bottom to the top, see figure 2(a). At each time step $k$, the range image is transformed into a 3D point cloud and, with the estimated positions of the four spheres, aligned with the reference frame. The aligned point cloud is quantized and added to a voxel set that is accumulated over the last ten images. The voxel set is segmented and for each part the centroid and shape factor are computed. The part states are updated with this new observation using the particle filter. The particle filter uses $N = 500$ samples, where $M = 500$ for the initialization. The standard deviation for the transition model $\sigma_u$ and for the initialization $\sigma_x$ is chosen as 30mm, the one for the observation model as $\sigma_z = 20$mm. $\sigma_z$ is chosen slightly larger than the statistical measurement error of the range camera that has been only evaluated for the center of the imager. However, the measurement error increases for the imager peripheral area. $\sigma_u$ and $\sigma_x$ are chosen slightly larger than $\sigma_z$ to account for motion of the



(a) Segmented parts for time step $k = 76$. The color indicates the shape factor: red for *leg*, green for *back*, and blue for *seat*. The brightness of the color indicates the probability to be noise.



(b) The evolution of part detection probabilities from time step $k = 75$ until 90. The color indicates the part detected: red for *leg*, green for *back*, and blue for *seat*.

Figure 4

segmented parts, which is not considered in the transition model. The probability of miss-classification is chosen as $m = 0.3$. Thus, a segmented part can change it's type with probability of 0.1.

The probability of object part presence is computed according to (4). In case of the particle filter, the probability is approximated by $P(R_k = r|\mathbf{z}_{1:k}) \approx \sum_{N_V} \delta(R_k^{(n)}, r)/N_V$, where $N_V$ is the number of particles in the region of interest. The object part presence probabilities for a sequence of range images are depicted in figure 4(b). At time step $k = 75$, two legs and a back are present. It can be observed that the legs are present until one leg disappears at time $k = 84$, but appears again at $k = 86$, and the other disappears at $k = 87$. The back remains present with varying probability over the whole sequence. Further, a part of the seat is detected, but with low probability, see figure 4(a). Over the whole sequence, spurious parts are detected such as the cuboid near the back, see figure 4(a). Their probabilities remain low and, therefore, these parts can be classified as noise. Thus, it is possible to distinguish between noisy and primitive parts of the chair. Similar results are obtained for different range image sequences of the same and other chairs.

## 5 Conclusion

The report presented an algorithm for object part detection with a particle filter. The algorithm can handle multiple parts of different types. The experiment showed that the approach can estimate the probability of part presence in the current range image given the measurement history. Thus, segmentation errors and measurement noise and outliers are successfully pruned. The particle filter can be further improved to cope better with the temporary loss of an object part in the image stream. The next step would be to extend the approach along the lines of Sudderth et al. (2006) to exploit the object structure to do object classification.

## Acknowledgement

Thanks to Luciano Spinello for suggesting using particle filters for part detection and to Ahad Harati and Viet Nguyen for the detailed discussions.

## References

3DV Systems. Israel, `http://www.3dvsystems.com/` (13.12.2006).

Canesta Inc. USA, `http://www.canesta.com/` (13.12.2006).

J. Czyz. Object detection in video via particle filters. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 1, pages 820–823, 20-24 Aug. 2006. doi: 10.1109/ICPR.2006.877.

Eward R. Dougherty and Roberto A. Lotufo. *Hands-on Morphological Image Processing.* SPIE Press, 2003.

Michael Isard and Andrew Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

T. Kahlmann, F. Remondino, and H. Ingensand. Calibration for increased accuracy of the range imaging camera SwissRanger. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information*

*Sciences, ISPRS Commission V Symposium*, volume XXXVI, pages 136–141, 25-28 July 2006.

Esther B. Koller-Meier. *Extending the Condensation Algorithm for Tracking Multiple Objects in Range Image Sequences.* PhD thesis, Eidgenössische Technische Hochschule Zürich, ETHZ, Diss ETH No.13548, 2000.

Matsushita Electric Industrial Co. Ltd. Japan, `http://biz.national.jp/Ebox/kyorigazou/index.html/` (13.12.2006).

Stefan May, Björn Werner, Hartmut Surmann, and Kai Pervölz. 3D time-of-flight cameras for mobile robotics. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 9-15 October 2006.

MESA Imaging AG. Switzerland, `http://www.swissranger.ch/` (13.12.2006).

PMDTechnologies GmbH. Germany, `http://www.pmdtec.com/` (14.12.2006).

Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter - Particle Filters for Tracking Applications.* Artech House, 2004.

J. Schubert and H. Sidenbladh. Sequential clustering with particle filters-estimating the number of clusters from data. In *Proceedings of the 8th International Conference on Information Fusion*, volume 1, pages 122–129, 25-28 July 2005.

Sharp Co. Japan, `http://www.sharp.co.jp/corporate/news/060323-a.html/` (13.12.2006).

E.B. Sudderth, A. Torralba, W.T. Freeman, and A.S. Willsky. Depth from familiar objects: A hierarchical model for 3D scenes. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2410–2417, 2006.

# 6 Version History

- 2006-12-10 Version 1.0 by Stefan Gächter Draft version of this document.
- 2006-12-14 Version 2.0 by Stefan Gächter First review of this document.
- 2006-12-19 Version 2.0.1 by Stefan Gächter Corrected measurement equation in section 3.2.
- 2007-9-1 Version 2.0.2 by Stefan Gächter Corrected some spelling mistakes.