Depth Gradient Based Segmentation of Overlapping Foreground Objects in Range Images

Andre Störmer
Bertrandt Ingenieurbüro GmbH
Lilienthalstr. 50-52
85080 Gaimersheim, Germany
andre.stoermer@de.bertrandt.com

Martin Hofmann and Gerhard Rigoll
Institute of Human Machine Communication
Technische Universität München
Arcisstr. 21, 80333 München, Germany
{martin.hofmann, rigoll}@tum.de

Abstract – *Using standard background modeling approaches, close or overlapping objects are often detected as a single blob. In this paper we propose a new and effective method to distinguish between overlapping foreground objects in data obtained from a time of flight sensor. For this we use fusion of the infrared and the range data channels. In addition a further processing step is introduced to evaluate if connected components should be further divided. This is done using non-maximum suppression on strong depth gradients.*

Keywords: background subtraction, fusion, time-of-flight camera, range image

1 Introduction

In the last years, time of flight imaging sensors delivering range images have enriched the working field of image processing. While the spatial resolution of range images is far from the resolution of typical video sensors, the information they can deliver is extremely valuable for many tasks.
In many applications a static camera observing a scene is the common case. In these scenarios, background subtraction techniques are a useful tool to either detect objects or to reduce the amount of input data for further computationally expensive processing steps. This is achieved, by removing those parts of the image data that belong to the background and thus are not of interest.
The basic principle of background subtraction is widely used. Some methods simply use an image of the empty scene to subtract from an actual image. Other methods model the background statistically or update it over time. State of the art methods try to model the probability density function for each pixels or image region separately. One successful and widely accepted approach is to use a single Gaussian, as it was used in [1]. Another approach assumes more complex distributions and uses Gaussian Mixture Models (GMM) [2]. A common method to update the GMM over time has been suggested in [3] and is further elaborated in [4]. Other references, using different methods to model the probability density function will be mentioned in Section 3.
As long as single non overlapping objects have to be detected, most methods deliver sufficient results. In difference to texture images, where the resulting foreground blobs are difficult to divide, the information in range images can be used to further examine if a resulting blob from a foreground segmentation process contains more than one object. In this paper, it is proposed to use non maximum suppressed depth gradients to find borders of objects. These borders are then used to divide foreground blobs into several parts to describe different objects.

2 Data

The data used in this paper has been recorded during the 7th Framework EU Project PROMETHEUS. In this paper, two indoor scenarios in a smart home environment are used for first experiments. In these scenarios up to 4 individuals, played by actors, portray daily behavior in a living room. The data is therefore captured in a very controlled environment. The sensor, a CMOS time of flight Camera (PMD[Vision]3k-S), delivers a range image with a spatial resolution of 64×48 pixels for a range of up to 7.5m in a depth resolution of $\approx$ 1cm as well as an 64×48-pixel NIR (Near-infrared) image. The images obtained by this sensor technology are typically very noisy. Because of this, all sensor data has been preprocessed with a 5×5-median filter, which reduces stronger noise, but some local details are lost. Examples of typical data can be seen in Fig. 1.

3 Background modeling in range and texture images

Several background modeling techniques have been suggested in the past years. Typical approaches are

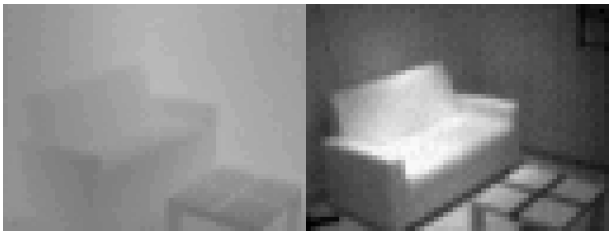

Figure 1: The sensor output for the empty scene in the "smartroom"-scenarios, the left part is the range information, the right is a NIR image.

using a pixel-wise modeling of the background, computed on a given set of training images. A pixel based background modeling leads to the decision R if a pixel $I(x,y)_t$ at time t belongs to the background (BG) or the foreground (FG)

$$R = \frac{p(BG|I(x,y)_t)}{p(FG|I(x,y)_t)} = \frac{p(I(x,y)_t|BG)p(BG)}{p(I(x,y)_t|FG)p(FG)} \quad (1)$$

In most cases nothing is known about foreground objects. It is unknown how often, when and where they will occur. Because of this the priors are set to be equal $p(FG) = p(BG)$ and a uniform distribution of foreground object appearance is assumed $p(I(x,y)_t|FG)) = c_{FG}$. If these assumptions are considered in equation 1, this leads to the decision that a pixel belongs to the background, if

$$p(I(x,y)_t|FG) > R \cdot c_{FG} = c_{thr} \quad (2)$$

where c_{thr} is a threshold. In the following, $p(I(x,y)_t|BG)$ is referred as the background model, which will be estimated by a training set X. The estimated model is then denoted as $\hat{p}(I(x,y)_t|X,BG)$ and explicitly depends on the training set. If all training samples are assumed independent, the main task is to estimate the density function for each pixel and to adapt it to possible changes. Several different density estimates have been used in the past, kernel based estimates were used in [5] and Gaussian Mixture Models have been used in a wide range of approaches, e.g. in [2] [3]. Explicit modeling of the time aspect has been considered by using Hidden Markov Models in [6] or in [7] by modeling the pixel value distribution over time as an autoregressive process.

The results of this background subtraction are usually used as input for higher level processing like tracking or recognition systems. The main focus of this paper is to take the result of the background subtraction, which delivers blobs of foreground objects, and examine for each blob if it represents a single or multiple foreground objects. This is done by searching for strong depth gradients and using them to divide blobs.

4 Gaussian Mixture Model

Since illumination in the scene could change gradually or suddenly (change in texture information) or new objects could be brought in or removed from the scene (change in texture and range information), an adaptive modeling of the background is chosen. For this the training set X is updated by adding new samples and discarding old ones. A time period T is chosen, and at time t the training set is $X_t = I(x,y)_t, .., I(x,y)_{t-T}$. Among these samples from the recent history there could be some values that belong to the foreground, thus the reestimate using this dataset is $p(I(x,y)_t|X_t, BG+FG)$. A GMM with M components is used:

$$p(I(x,y)|X_t, BG+FG) = \sum_{m=1}^{M} \hat{\pi}_m \mathcal{N}(I(x,y); \hat{\mu}_m, \hat{\sigma}_m^2) \quad (3)$$

where $\hat{\mu}_m$ are the means and $\hat{\sigma}_m^2$ are the variances of the Gaussian components. Mixing weights, denoted by $\hat{\pi}_m$ sum up to one and are non-negative. For a new sample $I(x,y)_t$ the ownership $o_{t,m}$ is computed based on the Mahalanobis distance to each Gaussian component $D_m^2(I(x,y)) = (I(x,y)_t - \hat{\mu}_m)^2/\hat{\sigma}_m^2$. If the distance is smaller than three standard deviations, the ownership $o_{t,m}$ is set to one, else it is set to zero. If there is no component which is close enough, a new Gaussian component is added with $\hat{\mu}_{M+1} = I(x,y)_t$, $\hat{\pi}_{M+1} = \alpha$ and $\hat{\sigma}_{M+1}^2 = \sigma_0^2$. The recursive update rules for existing Gaussian components, given a new data sample $I(x,y)_t$ are:

$$\hat{\pi}_m \leftarrow \hat{\pi}_m + \alpha(o_{t,m} - \hat{\pi}_m) \quad (4)$$

$$\hat{\mu}_m \leftarrow \hat{\mu}_m + o_{t,m}(\alpha/\hat{\pi}_m)(I(x,y)_t - \hat{\mu}_m) \quad (5)$$

$$\hat{\sigma}_m^2 \leftarrow \hat{\sigma}_m^2 + o_{t,m}(\alpha/\hat{\pi}_m)((I(x,y)_t - \hat{\mu}_m)^2 - \hat{\sigma}_m^2) \quad (6)$$

This method is an online clustering algorithm, usually foreground objects will be represented by some clusters with small weights $\hat{\pi}_m$. Therefore the background model can be approximated by the first B largest clusters. In this approach such a background model will be used independently for both range and NIR texture of the data. The clusters used to describe the background are the two clusters with the strongest weights. Note, that a more sophisticated method to chose the number of clusters and a detailed explanation of this background modeling technique is given in [8]. In this paper the number of Gaussians at each time step is fixed, because the focus is on the processing of the extracted foreground blobs, as it will be described in the next section.

Considering the special characteristics of the data (range + NIR), the foreground detection is on one hand very robust for all kind of objects, with a high distance to the wall or the furniture, on the other hand it is not very robust for objects or persons which are e.g. sitting on the furniture, because the distance is then naturally

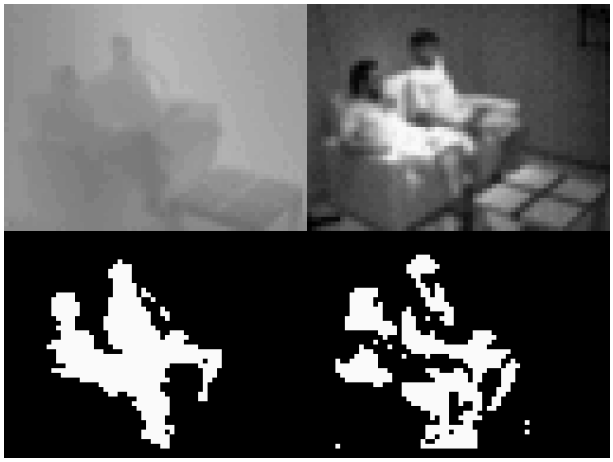

Figure 2: The results of the background subtraction on the range image (left) and on the NIR image (right)

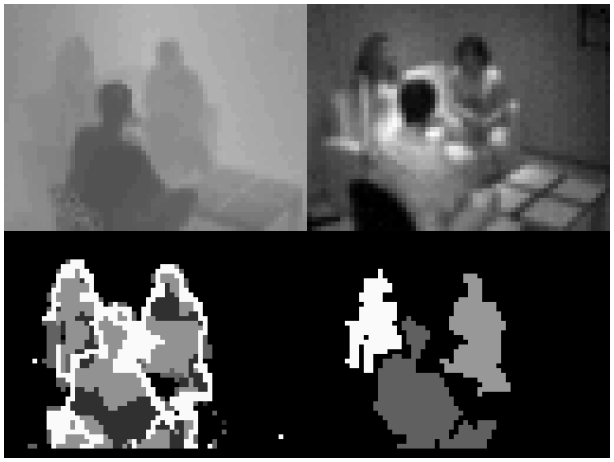

Figure 3: Depth gradient based segmentation of foreground blobs: the upper images are the range (left) and the NIR image(right), in the lower left all channels that will be fused are depicted: light gray is foreground that is detected in both range and NIR image, darker gray is foreground that is detected in the NIR image only and very dark gray means that foreground is detected in the range image only. Strong depth gradients are shown in white. The lower right image shows the resulting foreground clusters after applying the fusion rule.

small. Also the characteristic of NIR imaging leads to very similar NIR textures for all kind of textiles, this includes the clothes the people are wearing as well as the covering of the couch, which enhances the difficulty of extracting foreground in NIR images.

5 Postprocessing of extracted blobs using depth gradients

After the background subtraction has been performed, blobs of foreground objects are extracted. Note, that these foreground blobs are of similar characteristics, no matter which background subtraction method has been performed, so that the following approach can be combined with a variety of different methods.
A typical foreground blob is the result of a connected component analysis performed on the pixel wise binary foreground/background decision (see Fig. 2).

Overlapping foreground objects or those which are very close to each other result in a single blob. To divide this blobs a depth gradient based segmentation is proposed. For each pixel belonging to the foreground, the gradient is computed on the range image by applying a gradient operator or forward or backward difference. In this paper, the Sobel operator $(\mathrm{S}_x, \mathrm{S}_y)$ is chosen to estimate the partial derivatives for the x- and y- direction.

$$\hat{\nabla}(I(x,y)_t) = \begin{pmatrix} \mathrm{S}_x(I(x,y)_t) \\ \mathrm{S}_y(I(x,y)_t) \end{pmatrix} \tag{7}$$

the gradient magnitude is then computed as:

$$|\hat{\nabla}(I(x,y)_t)| = \sqrt{(\mathrm{S}_x(I(x,y)_t))^2 + (\mathrm{S}_y(I(x,y)_t))^2} \tag{8}$$

and the gradient orientation can be estimated by

$$\hat{\Theta} = \arctan \frac{\mathrm{S}_y(I(x,y)_t)}{\mathrm{S}_x(I(x,y)_t)} \tag{9}$$

After that a non-maximum-suppression is computed by checking for each pixel if it has a higher magnitude than the neighbors in the gradient direction, for this the gradient orientation is discretized to the eight orientations in a one pixel neighborhood. An absolute threshold d_{thr} for the minimal gradient magnitude is used additionally, to avoid to find depth edges that are of no interest. After depth edges of interest are found, a fusion of the foreground of the range data FG_R, the foreground of the NIR data FG_{NIR} and the depth edges $|\hat{\nabla}_{nms}(I(x,y)_t)| \geq d_{thr}$ is done. The following fusion rule leads to good results (see Fig. 5):

$$FG = (FG_R \cup FG_{NIR}) \cap (|\hat{\nabla}_{nms}| < d_{thr}) \tag{10}$$

Note, that in this equation an inversion of the depth edges is done ($< d_{thr}$).

6 Experiments

To verify the idea to split foreground blobs based on depth gradients, experiments have been performed. For a first test, a set of N extracted foreground images (consisting of one or several foreground blobs) have been postprocessed with the given approach. After that an error measure, the average distance of the number of resulting foreground objects to the ground truth, is computed.

$$e_{num} = \frac{|N_{FG} - N_{GT}|}{n_{fr}} \tag{11}$$

Dataset	$e_{num,BG}$	$e_{num,PP}$	$\overline{N_{GT}}$	n_{fr}
smartroom1	1.087	0.346	2.067	104
smartroom2	0.4	0.37	1.55	90

Table 1: Results: $e_{num,BG}$ is the error of the foreground blobs extracted by the background model, $e_{num,PP}$ is the error of the depth gradient based postprocessed foreground blobs, $\overline{N_{GT}}$ is the average number of objects in a frame (ground truth) and n_{fr} is the number of frames used for evaluation.

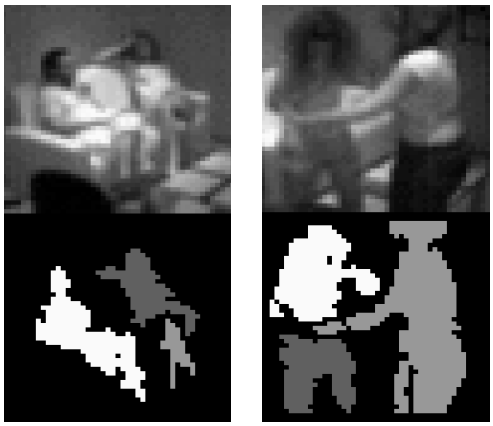

Figure 4: Typical errors that occurred are over segmentation of limbs (left image pair) and separation of occluded objects (right image pair).

where N_{FG} is the number of correctly extracted foreground objects and N_{GT} is the number of foreground objects in the ground truth. These first experiments lead to very interesting results. The first is, that obviously the depth gradient based postprocessing of extracted foreground blobs improves the performance especially on the set "smartroom1", in which up to four people are moving in a very narrow environment. Only a slight increase in performance is seen in the dataset "smartroom2" which mainly shows interaction between two persons, which are only overlapping in few cases. The number of errors that occur almost compensate the performance gain in that scenario. Typical errors are:

- Segmentation of body limbs as separate objects
- Occluded objects are separated into several parts

Both errors are expectable and can be compensated in future works by applying tracking algorithms that model the splitting and merging of blobs over time.

7 Conclusion

An effective method was shown on how to use standard background modeling techniques in conjunction with Time of Flight cameras, that deliver range images. An efficient algorithm that postprocesses foreground blobs by segmenting them based on depth gradient shows good results. Further investigations will show, if more sophisticated methods of fusing the results of the range and NIR background model with the depth gradient information will lead to better results. A combination with a tracking algorithm should lead to more robustness to avoid splitting of subparts and should enable to label the ID of split blobs over time.

8 Acknowledgments

The work described in this paper was conducted within the EU-Collaborative Project PROMETHEUS, "Prediction and interpretation of human behavior based on probabilistic structures and heterogeneous sensors", and was partially funded by the European Union Division FP7-ICT.

References

[1] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, 1997.

[2] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proc. of the Conference on Uncertainty in Artificial Intelligence*, 1997.

[3] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. of the Conference on Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[4] E. Hayman and J. Eklundh, "Statistical background subtraction for a mobile observer," in *Proc. of the International Conference on Computer Vision*, 2003, pp. 67–74.

[5] A. Elgammal, D. Harwood, and L.S. Davis, "Non-parametric background model for background substraction," in *Proc. of the European Conference on Computer Vision*, 2002.

[6] J. Kato, S. Joga, J. Rittscher, and A. Blake, "An HMM-based segmentation method for traffic monitoring movies," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, vol. 24, pp. 1291–1296.

[7] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *Proc. of the International Conference on Computer Vision*, 2003, pp. 1305–1312.

[8] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. of the International Conference on Pattern Recognition*, 2004, vol. 2, pp. 28–31.