A Modular Framework for 2D/3D and Multi-modal Segmentation with Joint Super-Resolution

Benjamin Langmann, Klaus Hartmann, and Otmar Loffeld

ZESS - Center for Sensor Systems, University of Siegen,
Paul-Bonatz-Str. 9-11, 57068 Siegen, Germany
{langmann,hartmann,loffeld}@zess.uni-siegen.de
http://www.zess.uni-siegen.de

Abstract. A versatile multi-image segmentation framework for 2D/3D or multi-modal segmentation is introduced in this paper with possible application in a wide range of machine vision problems. The framework performs a joint segmentation and super-resolution to account for images of unequal resolutions gained from different imaging sensors. This allows to combine high resolution details of one modality with the distinctiveness of another modality. A set of measures is introduced to weight measurements according to their expected reliability and it is utilized in the segmentation as well as the super-resolution. The approach is demonstrated with different experimental setups and the effect of additional modalities as well as of the parameters of the framework are shown.

Keywords: Segmentation, Image Processing, Range Imaging, Time-of-Flight (ToF), Photonic Mixer Device (PMD).

1 Introduction

Segmentation is a well known and widely studied topic in the area of image processing, computer and machine vision with many applications in associated subjects. The main sources of information are color images and several approaches have been proposed like pixel based methods, edge oriented methods, region and texture based approaches. Another research topic is how to apply these methods to other information sources, e.g. radar data, MRI scans or depth measurements.

The capability of a single modality to distinct objects is for some applications not sufficient or robust enough. The obvious approach is to utilize an additional modality acquired with another imaging sensor. But the measurements of the imaging sensors need to be registered, which is often non-trivial and sometimes needs certain assumptions. Additionally, one cannot assume that the resolutions of the imaging devices match. In the case that the lowest resolution is not sufficient for the task at hand a super-resolution method is required, since normal scaling methods are often not appropriate. In this paper we introduce a segmentation and joint super-resolution framework, which utilizes a standard

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 12–21, 2012.

segmentation method (Mean-Shift) and estimates super-resolved input images in an iterative process. The proposed method incorporates validity measures to judge the measurement quality of input data in order to account for noise and disturbances. The method is widely applicable and we demonstrate its capabilities in given test setups.

This paper is structured as follows: In section 2 the related work is discussed and in section 3 modalities used in the experiments and their acquisition are detailed. Afterwards, the segmentation and super-resolution framework is introduced in section 4. In section 5 experimental results of this framework are reviewed and this paper ends with a conclusion in section 6.

2 Related Work

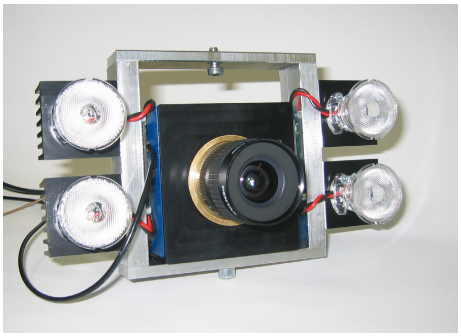
In this section research dealing with multi-image and depth image segmentation will be reviewed. In [1] the Mean-Shift algorithm is applied to color and depth images acquired with two cameras (binocular setup). The depth images are firstly resized with a bilateral filter and then segmented. Different super-resolution methods for depth imaging including several variants of bilateral filtering are compared in [2]. Color and depth are used for the task of alpha matting in [3]. Furthermore, several approaches have been studied for plain background subtraction, e.g. in [4] an approach utilizing Gaussian Mixture Models is demonstrated.

The segmentation of depth maps is performed in [5] for compression purposes and in [6] intensity and depth information of the same size is segmented with the graph-cut method and aimed at the detection of planar surfaces. In [7] a watershed based segmentation is utilized to analyze biological samples with 2D or 3D data and it combines intensity, edge and shape information. Lastly, the segmentation of ultrasound images in 2D or 3D in studied in [8].

3 Multi-modal Sensor Data

In the area of machine vision standard color or grayscale images are the predominant source of information, but nevertheless, supplemental modalities play a growing role. In addition to a color chip we utilize in this paper a continuous wave (CW) Time-of-Flight imaging chip in which is able to provide depth and near infrared reflectivity measurements. But the lateral resolution of such imaging chips is significantly lower than those of standard color or grayscale chips and also to low to capture fine details which are searched for in many applications. On the other hand these additional modalities may be able to provide valuable clues depending on the application.

The measurements often do not provide the information we want to utilize in the segmentation directly. For color chips shadows or varying lighting is not what we want to distinguish. Usually, the same is true for depth measurements. Even planes parallel to the imaging plane do not have similar depth values. Therefore, the first step is to derive the information of the images to be used in



MultiCam characteristics	
Interface	Gigabit Ethernet
Lens adapter	C-mount
Frame rate	12 fps (up to 80 fps with reduced 2D resolution)
Color chip	Aptina MT9T031
- Resolution	2048x1536
- Chip size	6.55mm x 4.92mm
PMD chip	PMDTec 19k
- Resolution	160x120
- Chip size	7.2mm x 5.4mm

Fig. 1. The MultiCam, a 2D/3D monocular camera and its specifications

the segmentation. Color images are often transformed to L*u*v* color space to make differences perceptually uniform and to be able to eliminate the influence of lighting conditions. Concerning depth images we can derive normal vectors of the depth image and use them in the segmentation if we have mostly planar objects in the scene. In the following these transformed information will be referred to as features.

We estimate normal vectors by firstly calculating 3D points of the depth values using a range camera model. Afterwards, the surface normal $\underline{n} = (n_x, n_y, n_z)$ at each 3D point $\underline{p} = (p_x, p_y, p_z)$ can be estimated by averaging over the 8 normal vectors of triangles spanned by $\underline{p}$ and combinations of its neighbor points. This will lead to interpolation errors at the borders of objects and may need additional handling.

The measurements performed have usually varying reliability. In some cases it is useful to utilize validity measures to judge the usefulness of a measurement. The quality of lighting greatly influences color information and it can be estimated based on the Luminance. For CW ToF imaging the modulation amplitude is a measure describing the amount of active lighting at given point and serves as a valuable descriptor. The variance over time is for all measurements a reliable validity measure as long as it is available, see section 4.2 for more details.

Additionally, we need to register the different images or multi-modal information. With our monocular camera shown in figure 1 this simply consists in applying a scale factor and an offset. When using multiple cameras this is more complicated in general, but for some machine vision applications an affine transformation may suffice.

4 Multi-image Segmentation Framework

Many traditional segmentation methods perform a clustering of points in the so-called feature space consisting of coordinates and measurements. The main advantage is that these methods are typically fast. On the down side these methods require input images of the same size. In section 4.1 a well known feature space segmentation method, the Mean-Shift algorithm, is detailed.

Furthermore, a given segmentation can be utilized to generate high resolution images of the input images of lower resolution. In section 4.3 such a super-resolution method is discussed. This leads to an estimation problem, which can be formulated in the Expectation Maximization (EM) framework as follows: all labels of feature points make up the parameter set Ω and the missing measurements due to lower resolutions are unobserved latent variables Z. This is based on the assumption that the segmentation algorithm maximizes the likelihood given Z.

1. **Initialization**: Generate uniform segmentation $\Omega^{(0)}$.
2. **Estimation**: Perform a super-resolution to estimate Z.
3. **Maximization**: Perform a multi-image segmentation to retrieve $\Omega^{(t+1)}$.

The E- and M-steps are iterated until a convergence is reached or a maximum number of iterations were performed. In the following sections the segmentation and super-resolution methods are detailed.

4.1 Mean-Shift Segmentation

The Mean-Shift algorithm [9] is a feature space approach and consists of two steps. In the first one (filtering) the mean-shift vectors are calculated iteratively and the feature points are moved accordingly until a convergence is reached. These vectors are determined by calculating the weighted average of neighboring feature points. In the second step the filtered feature points are merged to form regions or segments.

Let $P_i = \left(\underline{p}_i, \underline{f}_i\right)$ and $P_j = \left(\underline{p}_j, \underline{f}_j\right)$ be points of the feature space with lattices $\underline{p}_\cdot \in \mathbb{R}^2 = \Psi$ and features $\underline{f}_\cdot = \Phi$. For the weight $g\left(P_i, P_j\right)$ between two points we use the product of two separate Gaussian kernels but simpler and faster kernels are of course possible

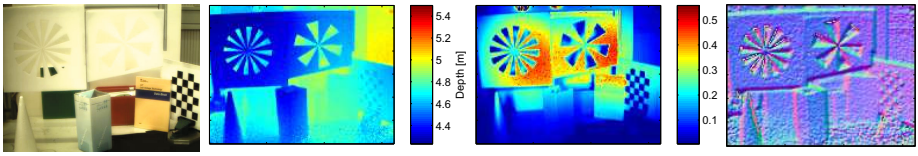
$$g\left(P_i, P_j\right) = \exp\left\{-\frac{\left\|\underline{p}_i - \underline{p}_j\right\|_2^2}{h_{space}}\right\} \cdot \exp\left\{-\frac{\left\|\underline{f}_i - \underline{f}_j\right\|_r^2}{h_{range}}\right\} . \tag{1}$$

The first term is based on the squared Euclidean distance between the two points and the second uses a (pseudo-)norm to measure differences between features. h_{space} and h_{range} are bandwidth parameters to control the influence of the kernels.

Given a point $P_i^{(t)}$ in the feature space in iteration t the next point can be computed with

$$P_i^{(t+1)} = \frac{\sum_{P \in N\left(P_i^{(t)}\right)} P \cdot g\left(P_i^{(t)}, P\right)}{\sum_{P \in N\left(P_i^{(t)}\right)} g\left(P_i^{(t)}, P\right)} , \tag{2}$$

where the set $N(\cdot)$ denotes a spatial neighborhood. The Mean-Shift vector is the difference $P_i^{(t+1)} - P_i^{(t)}$ and for each point these iterative calculations are



(a) RGB color (b) Depth $[m]$ (c) Mod. amplitude (d) Normal vect.

Fig. 2. Different modalities (colorization) acquired with the MultiCam

performed independently. Once the Mean-Shift filtering is finished, the feature points are merged using a distance threshold in the spatial domain as well as in the range domain.

4.2 Weighted Multi-modal Mean-Shift

The update formula of the Mean-Shift section can be easily extended with additional weights $\omega(P)$ leading to

$$P_i^{(t+1)} = \frac{\sum_{P \in N\left(P_i^{(t)}\right)} P \cdot \omega(P) \cdot g\left(P_i^{(t)}, P\right)}{\sum_{P \in N\left(P_i^{(t)}\right)} \omega(P) \cdot g\left(P_i^{(t)}, P\right)} \,. \tag{3}$$

The weights are specific to a feature Point P and not separate for subspaces of the feature space, since we perform a joint segmentation over all modalities. In figure 2 different modalities acquired with the MultiCam are shown. Let $P = \left(\underline{p}, \underline{f}\right)$ be a feature point and $\underline{f} = (f_L, f_u, f_v, f_d, f_{mod}, n_x, n_y, n_z, f_\sigma) \in \Phi$ with a color value (f_L, f_u, f_v) in the L*u*v* color space, a depth value f_d, a modulation amplitude f_{mod}, a normal vector $\underline{n} = (n_x, n_y, n_z)$ and a variance f_σ of the depth measurements. We can the define validity measures as follows. $\gamma_{lum}(P)$ gives penalties for low and high luminance levels with an appropriate parameter, e.g. $\alpha_{lum} = 70$, and a bandwidth $h_{lum} = 2$

$$\gamma_{lum}(P) = \exp\left\{-\frac{(f_L - \alpha_{lum})^2}{h_{lum}}\right\} \,. \tag{4}$$

Similarly, the modulation amplitude of CW ToF imaging describes reliably the noise level of the depth measurements and its influence in the validity measure $\gamma_{mod}(P) = 1 - \exp\left\{-\frac{f_{mod}^2}{h_{mod}}\right\}$ is controlled with a bandwidth parameter h_{mod}. Another indicator of the quality of the depth measurement is the variance over time, which is exploited in $\gamma_{var}(P) = \exp\left\{-\frac{f_\sigma^2}{h_\sigma}\right\}$. The complete additional weight is given by $\omega(P) = \gamma_{lum}(P) \cdot \gamma_{mod}(P) \cdot \gamma_{var}(P)$.

In the weight $g\left(P_i, P_j\right)$ between two feature points each subspace is typically treated independently, i.e. the range (pseudo)-norm $\|\cdot\|_r$ consists of separate norms for each subspace. Euclidean norms are commonly used, but this is not

appropriate for some modalities, especially for comparison of two normal vectors. Here the squared sine of the enclosed angle is much more suitable. Nevertheless, for simplicity it is also possible to use the distance measure $\delta_n\left(\underline{n}_i, \underline{n}_j\right)$ between normal vectors $\underline{n}_i$ and $\underline{n}_j$ of unit length, which embeds a Euclidean distance in a Gaussian kernel

$$\delta_n(\underline{n}_i, \underline{n}_j) = \exp\left\{ -\frac{\left(\left\|\underline{n}_i - \underline{n}_j\right\|_2^2 - 2\right)^2}{h_{normal}} \right\} . \tag{5}$$

In the experiments we use the following weighting kernel to measure differences between the feature points $P_i = (\underline{p}_i, \underline{c}_i, d_i, \underline{n}_i)$ and $P_j = (\underline{p}_j, \underline{c}_j, d_j, \underline{n}_j)$ with lattices $\underline{p}_.$, color values $\underline{c}_.$, depth values $d_.$ and normal vectors $\underline{n}_.$ with associated bandwidths

$$g\left(P_i, P_j\right) = \exp\left\{ -\frac{\left\|\underline{p}_i - \underline{p}_j\right\|_2^2}{h_{space}} - \frac{\left\|\underline{c}_i - \underline{c}_j\right\|_2^2}{h_{col}} - \frac{\left(d_i - d_j\right)^2}{h_{depth}} \right\} \delta_n(\underline{n}_i, \underline{n}_j) . \tag{6}$$

4.3 Joint Super-Resolution

A given segmentation can be used to estimate high resolution images of multi-modal images of lower resolution. Super-resolution for multiple images is often performed under the assumption that data of the different images coincides, e.g. color and depth. In the area of depth imaging cross bilateral filtering working on this assumption is commonly used. The same notion is true for segments found in multi-modal data.

Let $S_1, S_2, \ldots, S_m \subset \Psi \times \underline{\Phi}$ be segments with $S_i \cap S_j = \emptyset$ for $i \neq j$ with a subspace $\underline{\Phi}$ of Φ, for which a super-resolution should be performed. Let $\underline{q}^{(t)} = (\underline{p}, \underline{f}^{(t)}) \in S_i$ be a point at iteration t of the EM algorithm with arbitrary lattice $\underline{p}$ and feature $\underline{f}^{(t)}$. Let further $g(\cdot,\cdot)$ be a (Gaussian) kernel and $N(\cdot)$ a spatial neighborhood. In the subsequent iteration $\underline{q}^{(t+1)}$ is computed with

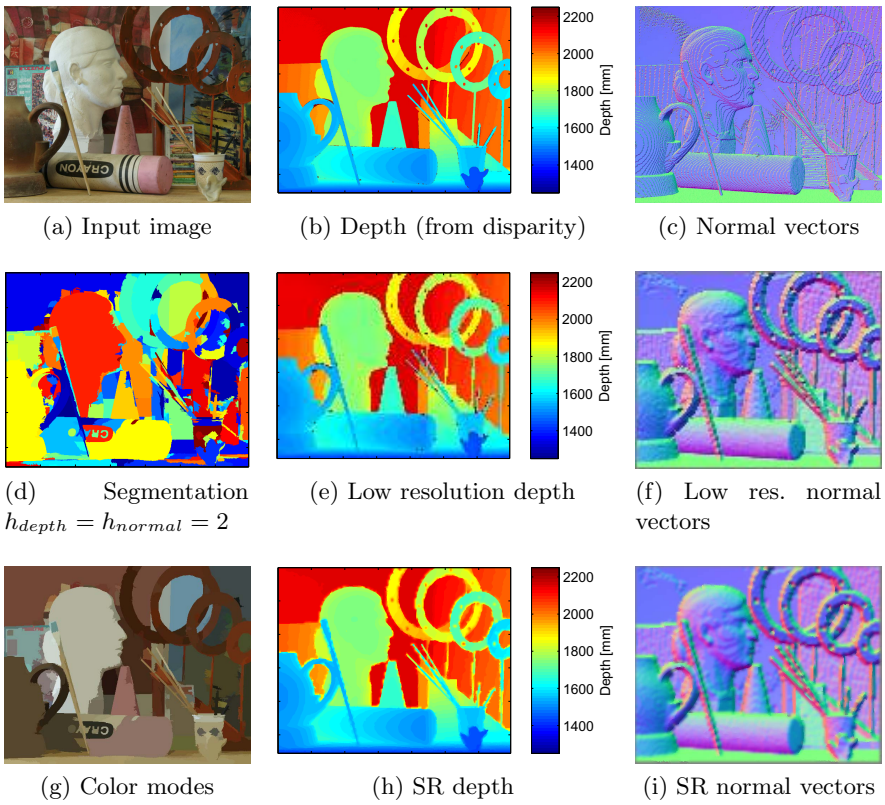
$$\underline{q}^{(t+1)} = \frac{\sum_{\underline{u} \in S_i \cap N\left(\underline{q}^{(t)}\right)} \underline{u} \cdot \omega\left(\underline{u}\right) \cdot g\left(\underline{q}^{(t)}, \underline{u}\right)}{\sum_{\underline{u} \in S_i \cap N\left(\underline{q}^{(t)}\right)} \omega\left(\underline{u}\right) \cdot g\left(\underline{q}^{(t)}, \underline{u}\right)} . \tag{7}$$

There are different possibilities to calculate this formula, e.g. the spatial neighborhood can include only the direct measurements or every feature point on a finer lattice and there are many ways to choose the kernel. In the experiments with the MultiCam we firstly create feature points on the lattice of the color image, since it has the highest resolution. To this end the other images are transformed, which includes here only a nearest neighbor scaling and a translation. Then the formula is applied with a spatial kernel for each feature point yielding a new set of points.

5 Experimental Evaluation

In order to evaluate the proposed segmentation and super-resolution framework we start with the well known high quality Middlebury benchmark dataset [10], which consists of sets of color images taken from different views and associated disparity maps. A view of one setup was chosen and is shown in figure 3, where the disparity map was converted to a depth map and the normal vectors were computed. The depth map was downsampled with a factor of 5 to simulate measurements of different lateral resolutions. One exemplary segmentation is shown and the super-resolution results are displayed also, which show weaknesses for thin objects but work in general as expected. Furthermore, the iterative estimates of the high resolution depth maps are shown in figure 4. One can observe a very smooth initial estimate and a sharpening in following iterations.

In figure 5 a similar setup, in which a variety of mostly white objects are arranged, is shown. The setup was acquired with the MultiCam and depth as well as normal maps were calculated. It shall be noted that it is usually possible to

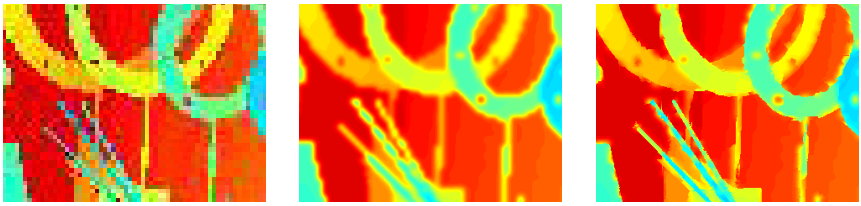


(a) Input image (b) Depth (from disparity) (c) Normal vectors

(d) Segmentation $h_{depth} = h_{normal} = 2$ (e) Low resolution depth (f) Low res. normal vectors

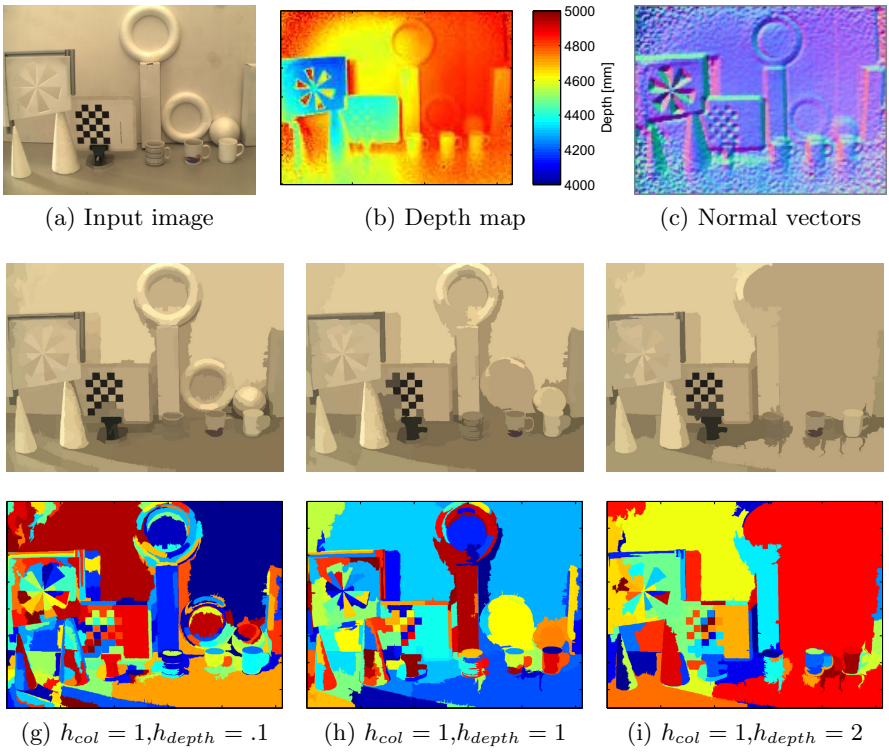
(g) Color modes (h) SR depth (i) SR normal vectors

Fig. 3. Multi-modal segmentation results for a benchmark image of the Middlebury dataset and super-resolution images



(a) Low res. depth map (b) Initial estimate (c) Third iteration

Fig. 4. Incremental estimation of the super-resolved depth map

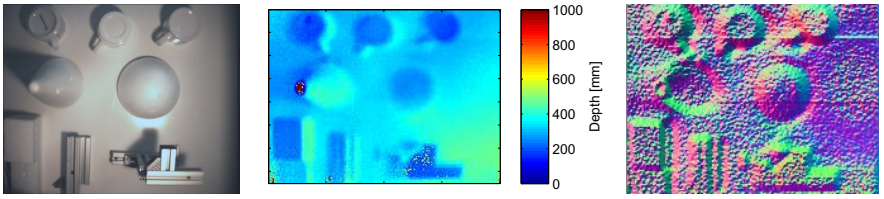


(a) Input image (b) Depth map (c) Normal vectors

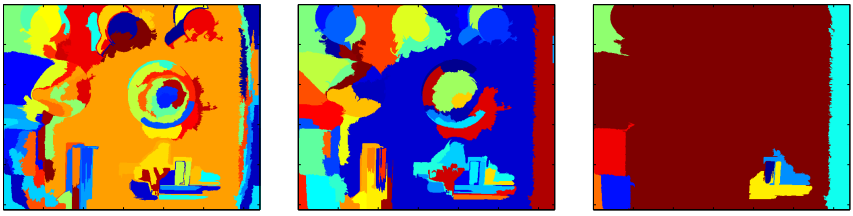
(g) $h_{col} = 1, h_{depth} = .1$ (h) $h_{col} = 1, h_{depth} = 1$ (i) $h_{col} = 1, h_{depth} = 2$

Fig. 5. Experimental setup with mostly white objects and segmentation results using depth map and estimated normal vectors

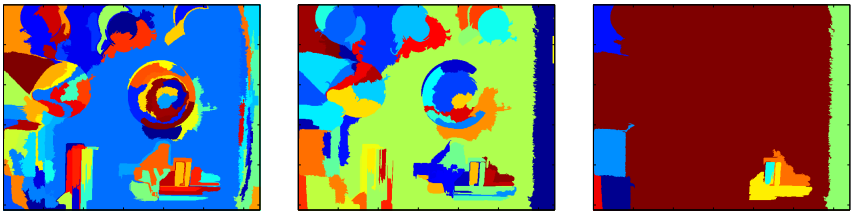
find parameters for the Mean-Shift algorithm to perform a valuable segmentation of such uniform scenes. Nevertheless, this segmentation is not robust and thus not reliable. The normal vectors do not provide valuable segmentation hints for this scene due to the parallel planes and curvatures. Segmentation results based on color and depth for three different parameter settings are shown with colored labels and average color of the segments. The depth bandwidth h_{depth} was changed to demonstrate the influence of the depth measurements and its



(a) Input image (b) Depth map (c) Normal vectors



(d) $h_{col} = .5, h_{depth} = .1$ (e) $h_{col} = .5, h_{depth} = 1$ (f) $h_{col} = .5, h_{depth} = 5$



(g) $h_{col} = 1, h_{depth} = .1$ (h) $h_{col} = 1, h_{depth} = 1$ (i) $h_{col} = 1, h_{depth} = 5$

Fig. 6. Test scene acquired under difficult lighting conditions and a set of different segmentations

significant value can be observed. In figure 6 a similar setup was acquired and it is demonstrated that color information provides in this case only small hints for the segmentation. The depth measurements can be utilized in conjunction with normal vectors to perform the segmentation and the color information is exploited mainly in the super-resolution task.

6 Conclusion

In this paper a modular multi-image segmentation framework for multi-modal data is introduced. Since the acquisition of multi-modal data is usually performed with different imaging chips, the segmentation approach needs to account for different resolutions and necessary transformations to align the different modalities. The proposed framework uses an estimation approach to jointly perform a segmentation and super-resolution, in which results of the segmentation influence the super-resolution and vice versa. The generated high resolution images are not only needed to accomplish the segmentation at borders of objects but can also

be utilized in subsequent processing steps. A set of validity measures is defined to give measurements of expected lower quality lower weights in the segmentation and super-resolution. The proposed multi-modal segmentation framework is demonstrated by applying it to 2D/3D segmentation and different experimental setups are utilized to evaluate the validity measures as well as the influence of the parameters of the approach. This should give valuable hints in which area of application the method can be applied successfully.

Acknowledgments. This work was funded by the German Research Foundation (DFG) as part of the research training group GRK 1564 'Imaging New Modalities'.

References

1. Bleiweiss, A., Werman, M.: Fusing Time-of-Flight Depth and Color for Real-Time Segmentation and Tracking. In: Kolb, A., Koch, R. (eds.) Dyn3D 2009. LNCS, vol. 5742, pp. 58–69. Springer, Heidelberg (2009)
2. Langmann, B., Hartmann, K., Loffeld, O.: Comparison of depth super-resolution methods for 2d/3d images. International Journal of Computer Information Systems and Industrial Management Applications 3, 635–645 (2011)
3. Wang, O., Finger, J., Yang, Q., Davis, J., Yang, R.: Automatic natural video matting with depth. In: Proceedings of the 15th Pacific Conference on Computer Graphics and Applications, Maui, Hawaii, pp. 469–472 (2007)
4. Langmann, B., Ghobadi, S.E., Hartmann, K., Loffeld, O.: Multi-modal background subtraction using gaussian mixture models. In: ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis (PCV 2010), pp. 61–66 (2010)
5. Jager, F.: Contour-based segmentation and coding for depth map compression. In: IEEE Visual Communications and Image Processing (VCIP), pp. 1–4 (November 2011)
6. Kahler, O., Rodner, E., Denzler, J.: On fusion of range and intensity information using graph-cut for planar patch segmentation. International Journal of Intelligent Systems Technologies and Applications 5(3), 365–373 (2008)
7. Wählby, C., Sintorn, I.M., Erlandsson, F., Borgefors, G., Bengtsson, E.: Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. Journal of Microscopy 215(Pt 1), 67–76 (2004)
8. Boukerroui, D.: Segmentation of ultrasound images: multiresolution 2d and 3d algorithm based on global and local statistics. Pattern Recognition Letters 24(4-5), 779–790 (2003)
9. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 603–619 (2002)
10. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (June 2007)