Multi-sensor Fusion Using Dempster's Theory of Evidence for Video Segmentation

Björn Scheuermann, Sotirios Gkoutelitsas, and Bodo Rosenhahn

Institut für Informationsverarbeitung (TNT)
Leibniz Universität Hannover, Germany
{last name}@tnt.uni-hannover.de

Abstract. Segmentation of image sequences is a challenging task in computer vision. Time-of-Flight cameras provide additional information, namely depth, that can be integrated as an additional feature in a segmentation approach. Typically, the depth information is less sensitive to environment changes. Combined with appearance, this yields a more robust segmentation method. Motivated by the fact that a simple combination of two information sources might not be the best solution, we propose a novel scheme based on Dempster's theory of evidence. In contrast to existing methods, the use of Dempster's theory of evidence allows to model inaccuracy and uncertainty. The inaccuracy of the information is influenced by an adaptive weight, that provides a measurement of how reliable a certain information might be. We compare our method with others on a publicly available set of image sequences. We show that the use of our proposed fusion scheme improves the segmentation.

1 Introduction

Segmentation of foreground objects in video sequences is a fundamental step in many computer vision applications and has been widely studied in the last years. A popular application in movie production is the integration of virtual objects into a sequence [1]. Because of many aspects in real-world scenarios video segmentation is a very challenging task. Illumination changes or background appearance changes, caused by people walking around, are typical problems that need to be treated.

The segmentation problem can be formulated using probabilistic models like Markov or conditional random fields. It has been shown, that the maximum a posteriori solution for these models corresponds to the discrete minimization of an appropriate energy function [2–4].

Time-of-Flight (ToF) cameras are perfect candidates to simplify the problem of binary video segmentation. ToF cameras use active sensors to measure the time taken by infrared light to travel to the object and back to the camera. The travel time corresponds to a certain depth value. Thus, ToF cameras are able to determine the depth value for the pixels in an image, which can be seen as additional information for each pixel.

The proposed algorithm is related to many recent works on binary image or video segmentation [2–7]. In [2–4], the authors use a discrete energy minimizing framework for interactive image segmentation. The problem of segmentation is transferred on a graph, where the minimum cut corresponds to the minimum energy state. In [5] and

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 431–438, 2013.

Fig. 1. Example segmentation result by fusing color and depth information using Dempster's theory of evidence. The explicit modeling of uncertainty forces the algorithm to segment the person in the foreground even if the depth information of the person in the background is similar. Input data taken from [9].

in [7], stereo images where used to estimate the scene depth. They showed that the combination of estimated depth and color improves the segmentation result. However, the estimation of the scene depth is a non trivial problem that is prone to errors in real-world scenarios.

The two most related methods are [8, 9]. In [8], Scheuermann and Rosenhahn proposed to use Dempster's theory of evidence for energy minimizing segmentation. They proposed a variational energy functional, including mass functions to fuse color and texture information, and solved it using level sets. In [9], Wang et al. proposed a very similar method, the so-called ToFCut algorithm. They combine depth and color cues in a discrete energy function and weight the information adaptively.

In this paper, we propose a novel method to fuse color and depth information in a discrete energy function. Therefore we use Dempster's theory of evidence to combine the different information. Using the proposed feature fusion allows us to explicitly model inaccuracy and uncertainty. This modeling provides an elegant way to incorporate the reliability of a feature channel. The information about how reliable a feature channel might be, can be either defined manually, based on prior information, or using our proposed adaptive weighting function. The adaptive weighting uses the symmetric Kulback-Leibler divergence as a measure of reliability. Therefore we compute distances of foreground and background histograms based on the segmentation result of the previous frame.

In summary, our main contributions are:

- A novel discrete energy function including Dempster's theory of evidence for feature fusion.
- An adjustable mass function, that can either use prior information or an adaptive weighting function based on the symmetric Kullback-Leibler divergence.
- Improved color and depth models, that are more robust.

In contrast to [9], we propose to use Dempster's theory of evidence to fuse color and depth information. We show that the proposed discrete energy function is more intuitive then the ToFCut functional. Furthermore, we propose stable functions, based on the Kulback-Leibler divergence, to adaptively compute the confidence of each sensor.

The experimental validation on the data set used in [9] shows that the proposed method outperforms ToFCut.

2 Segmentation by Discrete Energy Minimization

The problem of binary segmenting an image or image sequence can be formalized by minimization of a discrete energy function $E : \mathcal{L}^n \rightarrow \mathbb{R}$. Usually the energy function is written as the sum of unary φ_i and pairwise $\varphi_{i,j}$ potentials.

$$E(x) = \sum_{i \in \mathcal{V}} \varphi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \varphi_{i,j}(x_i, x_j)\,, \tag{1}$$

where $x \in \mathcal{L}^n$ is a labeling, $\mathcal{V}$ corresponds to the set of all image pixels and $\mathcal{E}$ is the set of all neighboring pixels. In case of binary segmentation, the label set $\mathcal{L}$ consists of foreground (FG) and background (BG) labels. The unary potential φ_i is given as the negative log-likelihood of a probability measure, e.g. a standard Gaussian mixture model (GMM) [4]:

$$\varphi_i(x_i) = -\log p(I_i \mid x_i = L)\,, \tag{2}$$

where I_i is the feature vector of pixel i, e.g. RGB values. L is either FG or BG and p is the likelihood. The pairwise potential is usually given by a contrast sensitive Ising model, defined by

$$\varphi_{i,j}(x_i, x_j) = \gamma \cdot \mathrm{dist}(i,j)^{-1} \cdot [x_i \neq x_j] \cdot \exp(-\beta ||I_i - I_j||^2)\,. \tag{3}$$

Here γ specifies the impact of the pairwise potential, $[\cdot]$ is the indicator function and $\mathrm{dist}(\cdot)$ is the Euclidean distance between neighboring pixels. The parameter β is defined as $\beta = (2\langle ||I_i - I_j||^2 \rangle)^{-1}$, where $\langle \cdot \rangle$ indicates expectation [10].

In [9], the energy function is extended by means of additional depth information. Therefore, the unary potential takes the form:

$$\varphi_i(x_i) = -\lambda_c \cdot \log p_c(I_i \mid x_i = L) - \lambda_d \cdot \log p_d(D_i \mid x_i = L)\,, \tag{4}$$

where D_i is the depth of pixel i. The likelihood p_c is a GMM learned using 3D histograms with 8^3 bins in the RGB color space and the likelihood for depth p_d is modeled by two Gaussian distributions. The parameters λ_c and λ_d are used to adaptively weight the impact of both cues. They are based on the discriminative capabilities of the two likelihoods. The color confidence is computed using the Kulback-Leibler divergence (KL) between the grayscale histograms of frames I^{t-1} and I^t (denoted by δ^{KL}_{lum}) and the KL divegence between foreground and background color histograms of frame I^{t-1} (δ^{KL}_{rgb}) . This yields the confidence of the color term

$$\mathcal{R}_c = \exp\left(-\frac{\delta^{KL}_{lum}}{\eta_{lum}}\right) \cdot \left(1 - \exp\left(-\frac{\delta^{KL}_{rgb}}{\eta_{rgb}}\right)\right)\,, \tag{5}$$

with parameters η_{lum} and η_{rgb}. The depth confidence $\mathcal{R}_d$ is computed using the distance between the average depth values for foreground and background in frame I^{t-1}

($\Delta\chi = |(\chi^f + \chi'^f) - (\chi^b + \chi'^b)|/2$). Here, χ^f, χ'^f, χ^b and χ'^b are the mean values of the Gaussian distributions p_d. This yields

$$\mathcal{R}_d = 1 - \exp\left(-\frac{\Delta\chi}{\eta_d}\right), \tag{6}$$

with the additional parameter η_d. Finally the adaptive weights are defined as $\lambda_c = \mathcal{R}_c/(\mathcal{R}_c + \mathcal{R}_d)$ and $\lambda_d = \mathcal{R}_d/(\mathcal{R}_c + \mathcal{R}_d)$. For more details on the likelihood terms and the adaptive weighting the reader is referred to [9].

In contrast to ToFCut, we propose to use the symmetric Kulback-Leibler divergence, since the symmetric distance does not depend on the order of the feature channels. We also use the symmetric KL divergence to measure the distance between FG and BG depth histograms in frame I^{t-1}, since the given definition using $\Delta\chi$ lacks in precision.

It has been shown that, using the defined unary and pairwise potentials, the energy (1) is submodular and can hence be represented by a graph G [10]. In this form, the global minimum of the energy function corresponds to the minimum cut of graph G that can be computed using standard maximum flow algorithms [11].

2.1 Dempster's Theory of Evidence

We continue with a brief review of Dempster's theory of evidence [12, 13], which is later used to fuse color and depth cues. Several works [8, 14, 15] applied the theory to image segmentation and showed that it can be superior to classical probability theory.

Dempster's theory of evidence is a generalization of classical probability theory, with the ability to jointly represent inaccuracy and uncertainty information. The theory is build on so-called basic probability assignments (also known as mass functions), that are defined on a hypotheses set Ω. In our case, the hypotheses set is composed by the labels FG and BG. The mass function $m(A) : \wp(\Omega) \rightarrow [0, 1]$ is defined on the power set of Ω.

The quantity $m(A)$ is interpreted as the belief strictly placed on hypothesis A. In contrast to classical probability theory, this belief is distributed on both simple and composed classes and models the impossibility to separate several hypotheses. This characterizes the principal advantage of the evidence theory.

Another particular characteristic of Dempster's theory, one which differs from classical probability theory, is: if $m(A) < 1$, then the remaining mass $1 - m(A)$ does not need necessarily refute A (i.e. support its negation). Thus we do not have the so-called additivity rule $p(A) + p(\overline{A}) = 1$.

To fuse mass functions from different feature channels we use Dempster's rule of combination, denoted by $m = m_1 \otimes m_2$. This rule combines two independent bodies of evidence, defined on the same hypotheses set Ω, into one body of evidence. Since Dempster's rule of combination has shown to be associative, we can combine information arising from more than two channels.

3 Feature Fusion Using Dempster's Theory of Evidence

In this Section we describe the details of our proposed segmentation scheme and show similarities and differences to existing approaches.

The unary potential used by ToFCut is defined as a weighted sum of negative log likelihoods, see Equation (4), and can be reformulated as:

$$\varphi_i(x_i) = -\log\left[p_c(I_i|x_i = L)^{\lambda_c} \cdot p_d(D_i|x_i = L)^{\lambda_d}\right] , \tag{7}$$

which can be interpreted as follows: if the confidence for a channel is near zero, the likelihood is near one. That means, to ignore a channel we push the corresponding likelihoods near one. This is a neither intuitive nor elegant solution. Furthermore, this non-linear solution heavily depends on a good adaptive weighting function.

In contrast to ToFCut our unary potential is defined using Dempster's basic probability assignment:

$$\varphi_i^{DS}(x_i) = -\log m(x_i = L) , \tag{8}$$

where the mass function $m = m_c \otimes m_d$ fuses the information of color and depth according to Dempster's rule of combination. Thus the complete energy function reads:

$$E(x) = \sum_{i \in \mathcal{V}} \varphi_i^{DS}(x_i) + \sum_{(i,j) \in \mathcal{E}} \varphi_{i,j}(x_i, x_j) , \tag{9}$$

Using the proposed unary potential φ_i^{DS}, we can elegantly model the uncertainty of a channel by defining the corresponding mass functions appropriately. Since we use Dempster's rule of combination, that is associative, we can also include additional information e.g. texture and motion.

3.1 Mass Functions

The most important difference between the proposed method and ToFCut is the feature fusion using Dempster's theory of evidence instead of summing up weighted log-likelihoods. Therefore the main contribution is the definition of appropriate mass functions, that model inaccuracy and uncertainty in an elegant way. The mass functions modeling color and depth information are defined by:

$$\begin{aligned} m_c(\Omega) &= \frac{\lambda_d(1 - (p_c(I_i|x_i = \text{FG}) + p_c(I_i|x_i = \text{BG})))}{K} , \\ m_c(L) &= (1 - m_c(\Omega)) \frac{p_c(I_i|x_i = L)}{p_c(I_i|x_i = \text{FG}) + p_c(I_i|x_i = \text{BG})} \end{aligned} \tag{10}$$

for the color term and

$$\begin{aligned} m_d(\Omega) &= \frac{\lambda_c(1 - (p_d(I_i|x_i = \text{FG}) + p_d(I_i|x_i = \text{BG})))}{K} , \\ m_d(L) &= (1 - m_d(\Omega)) \frac{p_d(D_i|x_i = L)}{p_d(D_i|x_i = \text{FG}) + p_d(D_i|x_i = \text{BG})} \end{aligned} \tag{11}$$

for the depth term, where L is either FG or BG. The uncertainty $m_c(\Omega)$ and $m_d(\Omega)$ of the models is defined by summing up the likelihoods. This means that the uncertainty of a model is high, if FG and BG likelihoods are small. The normalization factor K is chosen so that $m_c(\Omega) + m_d(\Omega) = 1$, which means that the sum of modeled uncertainty is one. The parameters λ_d and λ_c are the adaptive weights coming from the histogram analysis. They can be used to further increase or decrease the importance of a feature channel.

Table 1. Comparison between the proposed method DS and ToFCut obtained on four video sequences. The mean percentage error, computed across the whole sequence, is provided. The results obtained by ToFCut are taken from [9]. The proposed method clearly outperforms ToFCut.

Seq. ID	WL		MS		MC		CW	
No. Frames	200		400		300		300	
Alg.	ToFCut	DS	ToFCut	DS	ToFCut	DS	ToFCut	DS
% Error (Equal Weight Fusion)	1.37	0.54	0.51	0.23	0.16	0.06	11.68	2.21
% Error (Adaptive Weight Fusion)	1.35	0.51	0.51	0.23	0.15	0.06	0.38	0.26

3.2 Color and Depth Likelihoods

We also use an improved color model, since the one proposed in [9] is sensitive to small bins and lacks in precision, leading to suboptimal segmentation results. Similarly to [9], we use two 3D histogram with $H = 8^3$ bins in the RGB space for FG and BG. For each bin we learn a 3D-Gaussian with mean μ_k^j, covariance matrix Σ_k^j and weight w_k^j, for $k \in 1 \dots H$ and $j \in \{\text{FG, BG}\}$. The conditional probability is now given by:

$$p(I_i \mid x_i = L) = \sum_{i \in \mathcal{N}} w_i^L G(I_i | \mu_i^L, \Sigma_i^L)\,. \tag{12}$$

In contrast to ToFCut we omit the normalization term, to make the model more robust.

To model the depth likelihoods we use the conditional probability proposed by Wang et al. [9], where two Gaussian's are used for foreground and background. Furthermore we define a threshold T on the depth map, to exclude pixels from the training of the Gaussians. This threshold forces pixels with a depth value smaller than T to be segmented as background and improves our FG and BG models. Thus, the single parameter T is intuitive and easy to adjust.

4 Experimental Results

In this Section, the evaluation of the proposed method is presented. For qualitative and quantitative analysis we use the ToFCut data set with the corresponding ground truth data [1]. In Table 1 we present the obtained results and compare them to ToFCut by means of mean percentage error of misclassified pixels [5, 9]. In the experiments we use an equal weight fusion of color and depth information by setting $\lambda_c = \lambda_d = 0.5$ and an adaptive weight fusion based on histogram analysis. The quantitative results show that for both systems, equal weight fusion and adaptive weight fusion, the proposed fusion with Dempster's theory outperforms ToFCut. Important to notice is, that we only need to adjust two intuitive parameters: γ, the weighting of neighboring discontinuities and T, the threshold of the depth map. The parameters η_{lum}, η_{rgb} and η_d, controlling the adaptive weighting, remain constant in all our experiments, while in [9] they have to be adjusted for each sequence manually. Furthermore, the results show that the proposed

[1] `http://vis.uky.edu/`

Fig. 2. Example segmentation results, on four sample frames from each of the video sequences

fusion works well on many sequences without an adaptive weighting. Qualitative results for all sequences are presented in Figure 2. They show that the small segmentation error corresponds to a high-quality segmentation.

Besides video segmentation, interactive image segmentation is a challenging task. Since there exists no benchmark including depth images, we use the same data set. Qualitative results are presented in Figure 3. Since color and depth models are learned from rough user strokes, the models are likely to be incomplete. By using the proposed fusion based on Dempster's theory of evidence, this is elegantly modeled in our mass functions and the segmentation result outperforms ToFCut.

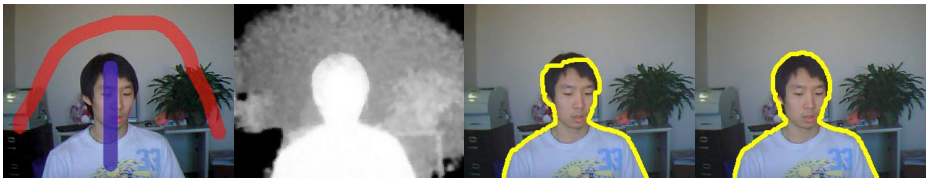

Fig. 3. Example interactive segmentation result. From left to right: Color image with initialization (FG in blue/BG in red), corresponding depth image, segmentation result using ToFCut with equal weights, proposed DS fusion with equal weights.

5 Conclusion

The paper presents a novel video segmentation scheme. It uses Dempster's theory of evidence to fuse color and depth information. With Dempster's theory of evidence we are able to define the uncertainty of a feature in an elegant way using prior information or an adaptive weight based on the symmetric Kullback-Leibler divergence. Furthermore, we propose adjusted color and depth models to improve the segmentation results. The quantitative evaluation shows that the proposed method outperforms ToFCut. In contrast to ToFCut, the proposed method has less parameters that are more intuitive and easy to adjust. An additional property of the proposed fusion scheme is the naturally given possibility to include further information like motion or user priors.

References

1. Cordes, K., Scheuermann, B., Rosenhahn, B., Ostermann, J.: Learning object appearance from occlusions using structure and motion recovery. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 611–623. Springer, Heidelberg (2013)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. TPAMI 23(11), 1222–1239 (2001)
3. Blake, A., Rother, C., Brown, M., Perez, P., Torr, P.: Interactive image segmentation using an adaptive GMMRF model. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 428–441. Springer, Heidelberg (2004)
4. Rother, C., Kolmogorov, V., Blake, A.: Grab cut: Interactive foreground extraction using iterated graph cuts. In: SIGGRAPH, vol. 23, pp. 309–314 (2004)
5. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Bi-layer segmentation of binocular stereo video. In: CVPR, vol. 2, pp. 407–414. IEEE (2005)
6. Criminisi, A., Cross, G., Blake, A., Kolmogorov, V.: Bilayer segmentation of live video. In: CVPR, vol. 1, pp. 53–60. IEEE (2006)
7. Harville, M., Gordon, G., Woodfill, J.: Foreground segmentation using adaptive mixture models in color and depth. In: EVENT, pp. 3–11 (2001)
8. Scheuermann, B., Rosenhahn, B.: Feature quarrels: The dempster-shafer evidence theory for image segmentation using a variational framework. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 426–439. Springer, Heidelberg (2011)
9. Wang, L., Zhang, C., Yang, R., Zhang, C.: Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera. In: 3DPVT (2010)
10. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient nd image segmentation. IJCV 70(2), 109–131 (2006)
11. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? TPAMI 26(2), 147–159 (2004)
12. Dempster, A.P.: A generalization of Bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological) 30(2), 205–247 (1968)
13. Shafer, G.: A mathematical theory of evidence. Princeton university press (1976)
14. Adamek, T., O'Connor, N.E.: Using Dempster-Shafer theory to fuse multiple information sources in region-based segmentation. In: ICIP, pp. 269–272 (2007)
15. Chaabane, S.B., Sayadi, M., Fnaiech, F., Brassart, E.: Dempster-Shafer evidence theory for image segmentation: application in cells images. IJSP (2009)