Joint Depth and Alpha Matte Optimization via Fusion of Stereo and Time-of-Flight Sensor

Jiejie Zhu[†] Miao Liao[†] Ruigang Yang[†] Zhigeng Pan[‡]

[†]Center for Visualization and Virtual Environments, University of Kentucky, USA
[‡]State Key Lab of CAD&CG, Zhejiang University, China

Abstract

We present a new approach to iteratively estimate both high-quality depth map and alpha matte from a single image or a video sequence. Scene depth, which is invariant to illumination changes, color similarity and motion ambiguity, provides a natural and robust cue for foreground/background segmentation – a prerequisite for matting. The image mattes, on the other hand, encode rich information near boundaries where either passive or active sensing method performs poorly. We develop a method to combine the complementary nature of scene depth and alpha matte to mutually enhance their qualities. We formulate depth inference as a global optimization problem where information from passive stereo, active range sensor and matte is merged. The depth map is used in turn to enhance the matting. In addition, we extend this approach to video matting by incorporating temporal coherence, which reduces flickering in the composite video. We show that these techniques lead to improved accuracy and robustness for both static and dynamic scenes.

1. Introduction

Image matting refers to the problem of extracting a foreground object by recovering per-pixel opacity from its background. It has been investigated by computer vision [2, 8, 18] and computer graphics [10, 16, 7] researchers for a long time. Basically, matting is an ill-posed problem because we need to conversely estimate three unknowns from one equation:

$$I = \alpha F + (1 - \alpha)B \tag{1}$$

where the input I is a composition of a foreground image F and a background image B. Its color is assumed to be a linear combination of the corresponding foreground and background colors weighted by opacity α. Most state-of-the-art algorithms require user interactions (such as *trimap* [2, 15, 6] or *scribbles* [11, 4, 18]) to generate high-quality output. Automatic methods typically require static scenes(e.g., [16]) or fairly elaborated setups (e.g., [7]). Robust and automatic matting for dynamic scenes remains an open challenge.

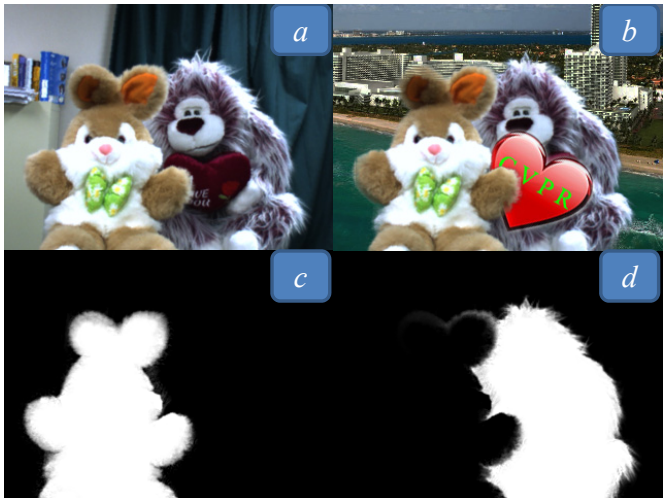


Figure 1. Piecewise Multi-layer Matting. Input a has 3 main layers: Background,Teddy and Bunny. With depth information, our algorithm can automatically calculate each layer's (Bunny (c) and Monkey (d)) matte in a recursive way. As a result, we can easily replace the background or insert images in between layers (b).

Image matting (or at least its binary version) can be considered as a crude estimation of scene depth. Therefore using scene depth is a natural way to bootstrap the process. Given the recent advances in stereo vision and active time-of-flight (TOF) sensors, a number of approaches [19, 3] have been developed to use the depth information to automatically extract alpha matte from natural images or videos.

While the depth information is typically used to generate the *trimap* for the matte and/or treated as an additional channel compliment to the RGB color channels, the resulting matte can in fact help the depth estimation process too. The matte clearly marks foreground and background boundaries where either passive or active method performs poorly.

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

Given the complementary nature of alpha matte and scene depth, we develop an iterative process to mutually enhance each other's quality. The spirit of our method is mostly related to [21, 17], which combines stereo-vision with alpha-matting. While some very impressive results have been presented, stereo matching, alpha-matte and over-segmentation rely on the color information. If there is not enough color or texture variation, neither of them can produce the correct result and fusing them does not lead to any improvement.

To address this problem, new cues other than those derived from color need to be included. Encouraged by the recent success of fusion of stereo vision and TOF sensors [23], we incorporate a TOF sensor to provide *independent* measurement of depth. The main contribution of our method is to fuse information from the TOF sensor and the stereo camera to refine both the alpha matte and scene depth. Furthermore, with depth information, we can easily segment the scene into multiple layers and calculate a matte for each one of them, which allows us to not only replace the background, but also insert new image between layers, as shown in Figure 1. Finally, when dealing with video sequences, we incorporate temporal coherence in both alpha and depth estimation. All these combined lead to a more robust and fully automatic matting and depth sensing pipeline that overcomes many difficult situations such as illumination changes, moving background, color similarity, and lack of textures.

2. Related Work

There are many approaches for matting. In general, they can be categorized into two major classes: single image matting and multiple images matting. Single image based methods typically require user inputs in forms of $trimap$ or $scribbles$ to disambiguate the different regions. Bayesian [2] and gradient [15] models are probably the two most widely used methods. Basically, Bayesian methods analyze the statistical distributions of samples from foreground F and background B; gradient methods assume the gradient of mattes are co-aligned with that of colors. With global analysis, Wang [18] developed a method to add more samples in local regions by Belief Propagation and Levin [8] introduced a quadratic cost function by eliminating F and B, which leads to a sparse linear system that can be solved directly. The basic matting problem has also been extended to multiple layers [20, 13]. While some stunning results have been obtained, one of the biggest drawbacks of these single image methods is the requirement of user interaction, so their application is mainly limited in image editing.

Multiple image based approaches use multiple images to solve the basic matting equation, making it possible to create a matte automatically. The classic blue-screen technique [14] belongs to this category. More recently, Sun et al. [16] employed a joint probabilistic approach by a flash/no-flash image pair. This method assumes that the flash only causes illumination changes on the foreground object and requires a static scene. Joshi [7] introduced an array of $eight$ cameras to capture a collection of images of a scene. These images help to compute mattes by creating a synthetic aperture image that can focus on the foreground and defocus (blur) the background, leading to a better matte.

Besides illumination variations and multiple backgrounds, scene depth is another important cue to facilitate matting. McGuire [10] introduced a system using three synchronized video cameras to defocus the background. Xiong [21] employs an Expectation-Maximization (EM) framework to optimize mattes using traditional stereo geometry. It is acknowledged in the paper that given the number of unknowns, the optimization can be trapped in local minimas. Our method does not explicitly model partial transparency in the depth estimation process. We lose some capability to recover depth for some very long hairs, but in return we increase the robustness. Taguchi [17] formulates a pair-wise Markov Random Fields (MRFs) for inference depth, alpha and segments together. Similar to our approach, the matte is used in turn to optimize the scene depth and vice versa. Our method does not require over-segmentation, and by successfully fusing *independent* depth, the approach overcomes the limitation of fronto-parallel assumption in each segments.

In addition, all these depth-assisted matting methods rely on scene textures to estimate the correct depth. They will fail on textureless regions. With the availability of full-frame time-of-flight sensors, a number of techniques [19, 3] have been developed to use the independent depth measurement to automate the matting process, in particular for video. Nevertheless the depth map is used as it is or simply up-sampled.

Unlike these previous approaches, we incorporate both *passive depth* (stereo cameras) and *active depth* (TOF sensor) to jointly refine the depth map and the matte, leading to a more robust automatic matting process.

3. Algorithm Overview

Our setup is composed of stereo cameras and a SwissRanger sensor [1]. One of the cameras (the left camera in our experiments) is regarded as the reference view for which we seek to estimate the matte and the depth map. Our joint matte and depth estimation approach has two main phases (Figure 2): an initialization phase in which an initial matte is extracted from a coarse depth from the TOF sensor, and an optimization phase in which the matte and the depth are alternatively refined.

In the first phase (in section 4), we compute the depth for reference view by warping the depth from TOF sensor. We then generate a trimap by this coarse depth and initialize the

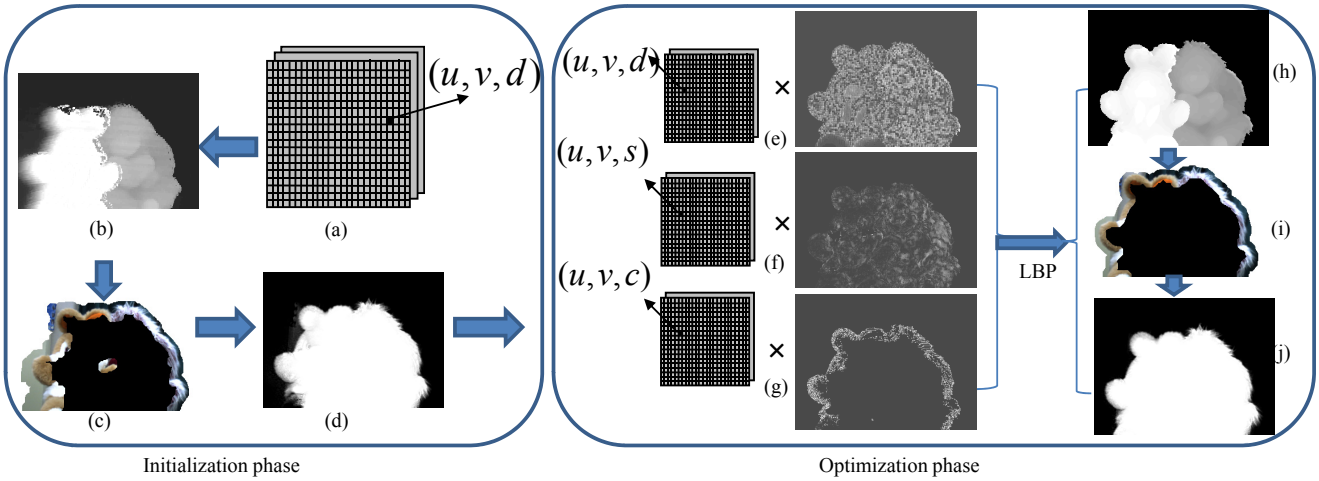


Figure 2. Overview of our algorithm for a static scene. In the initialization phase, we construct the cost volume (a) from the TOF sensor and compute its local minima (b). A trimap (c) is automatically generated by first segmenting (b) into two parts and then executing erosion and dilation operations. (c) is used to extract the initial matte (d) by the closed form solution. In the optimization phase, we construct a cost function by fusing three terms: depth cost from TOF sensor (e), pixel similarity from stereo matching (f) and confidence level from the matte (g). We resort to Loopy Belief Propagation (LBP) to infer the optimum depth (h). Then, a trimap (i) is generated from (h) and an improved matte (j) is extracted from (i) and (h). The refinement can be executed iteratively.

matte by Levin's method [8]. The initialized matte will be used in the next phase as a confidence cue.

During the optimization phase, we formulate depth inference as a MRFs and regarding it as a Maximum A Posteriori (MAP) problem. The cost function has three terms: pixel similarity from stereo matching, depth cost from the TOF sensor, and confidence level from the matte. To adaptively fuse them, we weight them by their reliability. We use Loopy Belief Propagation (LBP) [5] to do approximate inference. The refined depth will be used again to generate the trimap and consequently, high quality mattes can be extracted by an iterative process. We will explain the details in section 5.

4. Initialization

4.1. Initial Depth Acquisition

We follow up our previous work [23] and acquire initialized depth of stereo cameras by computing the local minima of a cost volume. We briefly review the method here.

We first calibrate TOF sensor with stereo cameras by regarding it as a regular camera (because it can report a gray scale image besides a full-frame depth map). Therefore, three cameras in our setup can be unified into one coordinate system. Given a range of disparity candidates, we define depth cost between passive stereo and TOF sensor. The *passive depth* is computed by stereo triangulation and the *active depth* is directly reported by the TOF sensor. Simply, the initial depth is computed as the local minima (depth with the smallest cost) from the cost volume. The volume will be used later in global optimization (see section 5.2).

4.2. Initial Matte Generation

Given the coarse scene depth, we are able to estimate the trimap automatically by foreground/background segmentation and boundary dilation/erosion.

We cluster disparities into groups using k-means.Typically we set k to two to segment foreground and background. The mean value from the two group centers is used as the binary classifier. Note that we can also set k to a value greater than two (or even use mean-shift) to segment the scene into multiple layers. We will discuss later how to take advantage of this.

To generate the trimap, we erode the foreground and background regions to remove small disconnected areas and dilate the unknown pixels inwards and outwards by 15 pixels. We will show later (in Section 5.4) that this number can be adaptively adjusted by refined depth. Given the trimap, the matte is computed using Levin's method [8]. The basic idea of this method is to derive a cost function from a linear combination instead of local smoothness on F and B. By analytically eliminating F and B, it yields a quadratic cost function only in α. The global optimum of the cost function is solved by a sparse linear system.

5. Optimization

In this section, we present our optimization method. To optimize the depth, we formulate a MAP-MRF model. To optimize the mattes, we add the depth as a weighted 4^{th} channel and adaptively narrow the unknown regions of trimap. We will explain the details below.

5.1. MAP-MRF

The energy function is composed mainly by a data term and smoothness term:

$$E = \sum_i D(d_i) + \sum_{i,j \in N} f_s(d_i, d_j) \tag{2}$$

where D is composed by three terms: f_d from stereo cameras, f_r from the TOF sensor and f_α from the mattes.

We use a fairly standard smoothness term formed as:

$$f_s = \min[(d_i - d_j)^2, T_1], j \in N(i) \tag{3}$$

where d_i and d_j are the disparity of pixel i and its neighbors j. T_1 is the truncational value of intensity, which is set to maximal disparity value.

In this quadratic truncational model, small intensity differences cause smaller penalties and large differences cause larger penalties. This encourages a few places where nearby pixels that change their costs significantly. We explain how to calculate and fuse different data terms in following section.

5.2. Data Terms

Stereo Matching f_d encodes the color consistency. In our implementation this pixel-wise matching cost is computed by an adaptive color weight strategy [22], which makes use of both color and geometric distance to provide moderate smoothness and preserve boundary sharpness.

TOF sensor f_r encodes the depth consistency which is defined as the geometric difference between *passive depth* $X_{passive}$ and *active depth* X_{active} explained in section 4.1. In addition, we incorporate a linear truncation model to maintain large disparity variations among candidates:

$$f_r = \exp \frac{-\min[|X_i - X_{tof}|, T_2]}{\gamma_r} \tag{4}$$

where T_2 is the truncation value of depth which is set to $300mm$, and γ_r controls the shape of the weighting function.

Alpha Mattes f_α encodes the opacity consistency on the foreground object in the left and right views. Similarly, we calculate its cost using pixel-wise matching method. Since the alpha value is confined in $[0, 1]$, we therefore define f_α as following:

$$f_\alpha = \exp \frac{-|\alpha_i - \alpha_{i'}|}{\gamma_\alpha} \tag{5}$$

where i and i' are matched pixels in stereo, and γ_α controls the shape of this weighting function. Although f_α is simple, it is effective to improve depth regularization results, particularly on boundaries.

5.3. Adaptive Weight

To merge the data terms, we introduce three weighting factors w_d, w_r and w_α:

$$d = w_d \cdot f_d + w_r \cdot f_r + w_\alpha f_\alpha \tag{6}$$

Instead of manually (empirically) specifying the weights [23], we adaptively compute them as a reliability: a metric defines how much trust we should give to candidate disparities. The idea behind reliability is simple: the best depth candidate should have a low cost while others are obviously larger. Therefore, we intuitively define the matching reliability of pixel i as how distinctive its best cost c_i^{1st} and its second best cost c_i^{2nd} is:

$$R(i) = \begin{cases} 1 - \frac{c_i^{1st}}{c_i^{2nd}} & c_i^{2nd} > T_c \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

T_c is a small value to avoid c_i^{2nd} equals zeros.

With all terms defined, we do approximate inference of equation 2 using LBP [5].

5.4. Optimize the Mattes

We design an iterative procedure to refine the mattes based on the optimized depth. We first use the previous matte as a confidence map to refine the foreground boundary in the depth map, and then automatically generate the trimap introduced in section 4.2. This time, however, we can safely narrow the unknown region (by reducing the dilation/erosion band size ($2 \sim 4$ pixels)) because the optimized depth already gives a good approximation. The new matte is used to replace the previous matte (Figure 2 (d)), and the algorithm (Figure 2 (e)$\sim$(j)) will run again. Results show that our algorithm can achieve satisfactory results by only $2 \sim 3$ iterative steps.

To facilitate matting, we add depth as the 4^{th} channel with the original R, G, B channels for a color image. In detail, the off-diagonal entries $(i, j)^{th}$ of the *matting lapalacian* [8] becomes:

$$\sum_{k|(i,j)\in w_k} (\delta_{ij} - \frac{1}{|w_k|}(1 + (A_i - \mu_k)(\Sigma_k + \frac{\varepsilon}{|w_k| I_4})^{-1}(A_j - \mu_k))) \tag{8}$$

where δ_{ij} is the Kronecker delta; A_i is a 4×1 vector of R, G, B augmented with the depth for pixel i; μ_k is a mean vector of A_i in a window w_k; $|w_k|$ is the number of pixel in window k; Σ_k is a 4×4 covariance matrix; ε is added to increase numerical stability when F;B and D in w_k is constant; I_4 is a 4×4 identity matrix.

Although scene depth gives us strong evidence on F and B from the depth edges, it violates the linear assumption of color combination from F and B in opacity regions. We

therefore weight the depth channel using the previous matte by an inverse entropy function:

$$H(\alpha) = \frac{1}{1 + \alpha \log \alpha + (1-\alpha)\log(1-\alpha)} \quad (9)$$

$H(\alpha)$ is large when the alpha tells us that we are seeing mostly B or mostly F.

6. Extensions

Piecewise Multi-Layer Matting With the depth information, we are no longer limited to a single matte for the foreground. As we discussed in section 4.2, it is relatively straightforward to segment the scene into multiple depth layers. We can use a simple procedure to estimate the matte for each layer. Using a synthetic scene with three layers as an example (Figure 3), we start from the furthest layer B. We calculate a matte for R and G. Using the newly acquired RG region as input, we can calculate a matte for R, then G's matte can be calculated as $\alpha_G = (1 - \alpha_R) \cdot \alpha_{RG}$. As the last layer, R's true alpha value is $\alpha_R = (\alpha_R) \cdot \alpha_{RG}$. This procedure can be easily extended to more than three layers, though it is not likely to make any noticeable visual difference in the final composite.

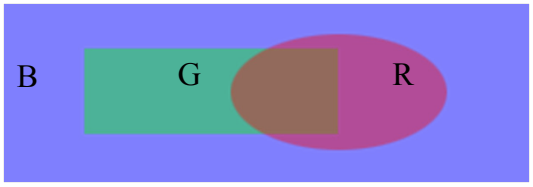


Figure 3. Synthetic scene with three layers.

Note that our method is different from recent multi-layer matting approaches [13] in which the matting equations is extended into the weighted sum of more than two layers and the alpha values for all layers are estimated simultaneously. We always solve the matte in a piecewise way, so any existing bi-layer matting method can be used. This is possible since we know the z-order. It is surprising that this has not been explored in previous depth-assisted matting methods.

Video Matting We extend current approach to video matting by introducing a new term f_t to maintain temporal coherence. f_t encodes temporal consistency by using the optical flow. To find the temporal correspondence for pixel i at time t, we locate its correspondence in previous frame $t-1$ by optical flow [9]. Considering the noise induced in flow estimation, we define a local window and use its weighted sum as the temporal evidence for pixel i:

$$f_{i_t} = \frac{\sum_{j \in N(i_{t-1})} w(j_{t-1}) \cdot C(j_{t-1})}{\sum_{j \in N(i_{t-1})} w(j_{t-1})} \quad (10)$$

$w(j_{t-1})$ is the geometrical distance between the corresponding pixel i_{t-1} and its neighbors j, $C(j_{t-1})$ is the cost from frame t.

7. Experiment Results

The evaluation of our approach is performed on static and dynamic scenes. For static scenes, we choose several challenging cases and both quantitative and qualitative comparison are presented. For dynamic scenes, we demonstrate our results on several video sequences, in which large motion, illumination changes and background movement are presented.

The two video cameras we are using are both DragonFly2 IEEE-1394 CCD cameras. The TOF sensor we have is a SwissRanger SR3000 [1], which can continuously produce a depth map of 176×144 resolution with an operational range up to 7.5 meters. In our current setup, two cameras have a baseline about $100mm$ and they are verged towards each other around 8 degrees from the parallel setup.

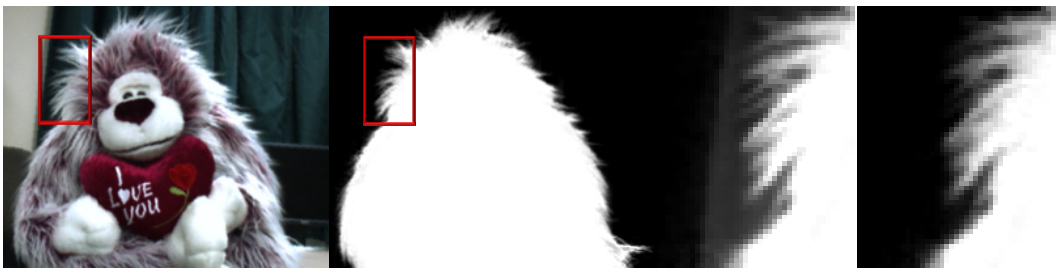
7.1. Results from Static Scenes

We test our algorithms on a number of static scenes. By applying methods in the initialization and optimization phase, we compare the trimaps and the mattes before and after the optimization phase. To evaluate the quality of depth map, we first obtain ground truth depth using structured light techniques [12], and compare it against these produced by three variations of our method: with the depth from the TOF sensor (d_t), from stereo (d_s), and from fusion (d_f).

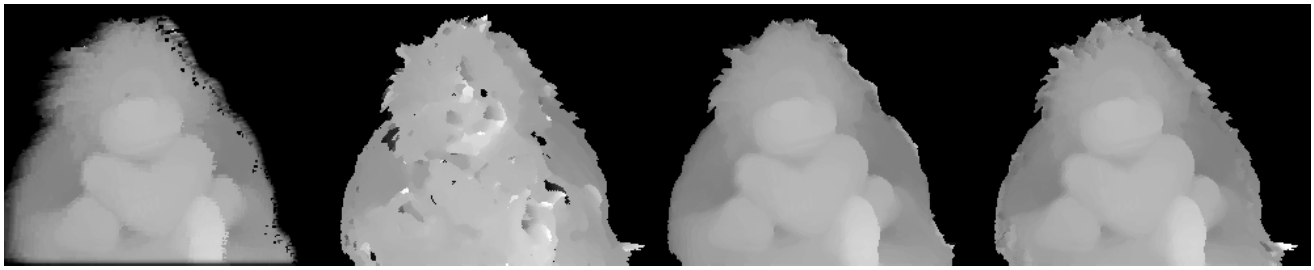
We show a full comparison of *scene Monkey* in Figure 4. There are many outliers (black holes) and many depth errors on boundaries in the initial depth (e). These are caused by false local minima from the cost volume. The matte from the trimap generated from this coarse depth has artifacts in both the noisy background and errors near the foreground boundary (see the enlarged images (c)). f, g, h are depth maps after optimization. Compared with the ground truth, we found the best is h. The final matte (b) is computed based on h by 2 iterations. All the results are generated automatically without any user interaction.

We also test our algorithms on three other cases (*scene Planar, Bear and Flower*. We show results of *Bear* and *Flower* in Figure 5. The numerical comparison of depth accuracy is presented in Table 1.

We can see that the result from fusion d_f is always the best, reducing the error from $5 \sim 7$ to $1 \sim 2$ disparity pixels on average as compared to initial depth d_i. The errors in d_i are mainly located near boundaries (see Figure 6). This again shows that incorporating a matte can efficiently reduce the depth error.

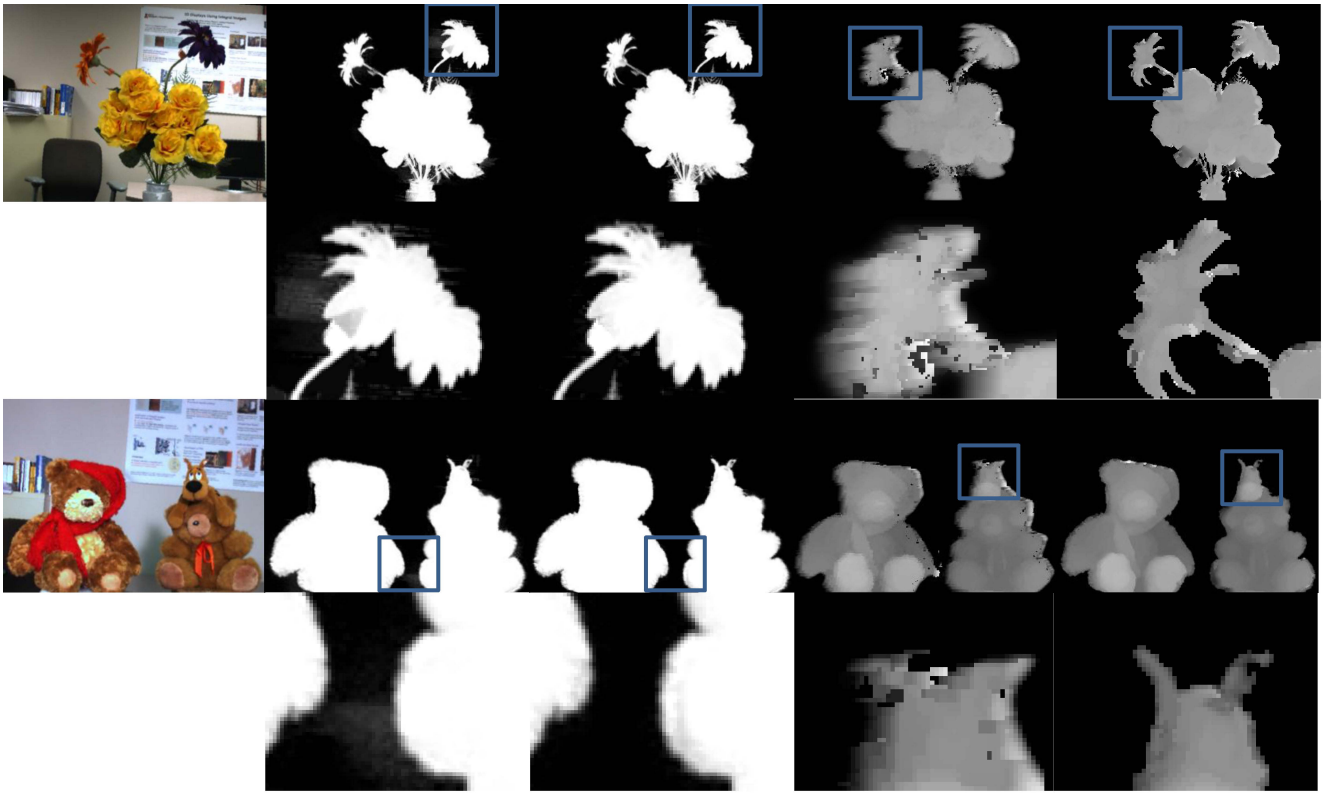


(a) Reference View (b) Optimized Mattes (c) Initial Mattes (d) Optimized Mattes



(e) Initial Depth (f) Optimized Depth from stereo (g) Optimized Depth from TOF (h) Optimized Depth from Fusion

Figure 4. Results from scene Monkey.



(a) Reference View (b) Initial Mattes (c) Optimized Mattes (d) Initial Depth (e) Optimized Depth

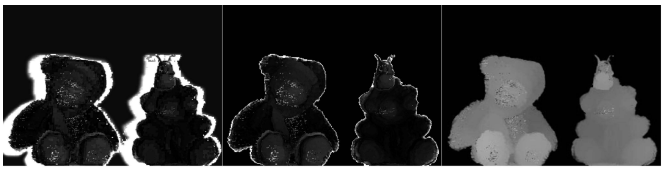
Figure 5. More results from static scene. Optimization phase reduces the opacity noise and provides better results on depth discontinuities.

7.2. Results from Dynamic Scenes

To verify the effectiveness of enforcing temporal smoothness f_t, we generate a set of 5-frame ground truth data of a moving scene. The object is manually rotated and acquired in stop motion. The ground truth matte is obtained by the blue screen approach [14]. Figure 7 shows the ground truth of the third frame. With temporal coher-

Table 1. Numerical comparison of depth against ground truth. Mean disparity error is presented.

	Monkey	Planar	Flower	Bear
d_i	4.5	3.54	10.4	6.15
d_s	0.95	1.88	2.9	1.77
d_t	1.9	1.50	4.5	1.55
d_f	0.75	1.32	2.35	1.1



(a) Initial Depth (b) Optimized Depth (c) Ground Truth

Figure 6. Visualized depth error against ground truth. High intensity means high error.

Table 2. Numerical comparison of mattes (mean α-error normalized in $[0, 255]$) and depth (mean disparity error) for a 5-frame sequence against ground truth.

	Mattes		Depth	
	without f_t	with f_t	without f_t	with f_t
frame 2	1.25	1.16	1.17	1.12
frame 3	1.09	1.07	1.04	1.04
frame 4	0.77	0.67	1.24	1.15
frame 5	1.10	1.02	1.26	1.20

ence, our algorithm is actually estimating depth from multiple shots instead of one single shot. It therefore increases the Signal to Noise Ratio (SNR) efficiently. As shown in Table 2, both mattes and depth are improved. One frame of qualitative comparison of matting with/without f_t from video sequence can be found in Figure 8.

(a) Reference View (b) Ground Truth Mattes (c) Ground Truth Depth

Figure 7. Example of ground truth data.

We further tested our algorithm on several video sequences and show part of results in the paper. The entire sequences can be found in the supplementary material.

Figure 9 shows three cases of dynamic scenes: large motion, background movement and illumination changes. The first row shows two frames of results from a hand moving sequence. The composite image shows a replaced background. We can see both the hair and the moving hand is corrected matted, and the depth discontinuity of the foreground person is preserved well. The second row shows results from a person moving behind another. We can see even if the foreground and background colors are similar (black hair and black jacket), our algorithm can still generate acceptable results without any explicit background model. The last row shows acceptable results with illumination changes in which we keep moving several red flashlights.

8. Discussion and Conclusion

Currently, our experiments are limited in indoor environments. We are unable to move our setup out because of poor depth reported from the TOF sensors. This is due to the fact that TOF sensors are too sensitive to strong background illumination. Nevertheless, moving TOF sensors out is an interesting topic and we envision to apply our methods with more robust active sensors from outdoor environments.

Another interesting extension in our approach is to include optical flow in the MRF model. However, inferring both optical flow and depth requires expensive computational resource because of the huge labeling space (number of optical flow candidates times number of disparity candidates). We think more efficient linear algebra methods are needed to resolve this problem.

We have proposed a new approach to jointly optimize depth map and alpha matte iteratively. We discussed initializing and optimizing phases, and we also extended our approach to piecewise multi-layer and video matting. Experimental evaluation shows that our approach can (1) reduce the depth error by nearly 70% compared to that directly reported from a TOF sensor; (2) provide visually pleasing matting results both from static and dynamic scenes; (3) is robust to many difficult situations.

9. Acknowledgment

This work is supported in part by the University of Kentucky Research Foundation, the US Department of Homeland Security, the US National Science Foundation Grant HCC-0448185 and CPA-0811647, the NSF of China (No.60533080), 863 project of China (2006AA01Z335) and Open Project of State Key Lab of CAD&CG, Zhejiang University (No.A0812).

References

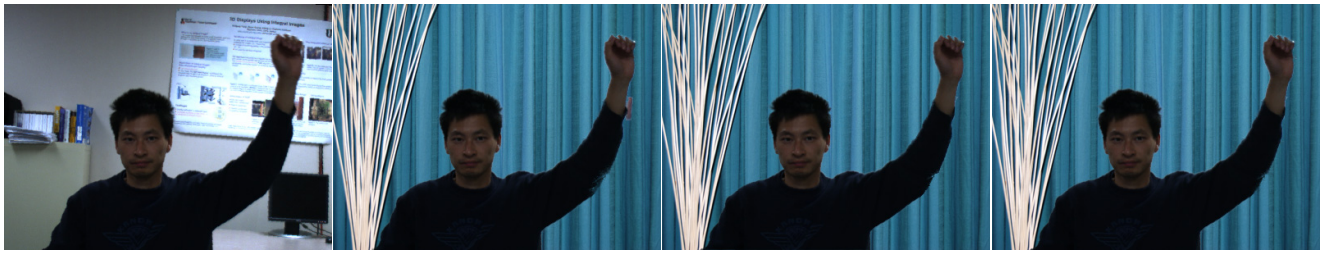
[1] Swissranger inc, sr-3. http://www.csem.ch/fs/imaging.htm, 2006.

[2] Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *CVPR*, 2001.

[3] R. Crabb, C. Tracey, A. Puranik, and J. Davis. Real-time foreground segmentation via range and color imaging. In *Workshop on Time of Flight based Computer Vision (TOF-CV)*, 2008.

[4] C.Rother, V.Kolmogorov, and A.Blake. Grabcut - interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[5] W. Freeman, E. Pasztor, and O. Carmichael. Learning low level vision. *International Journal of Computer Vision*, 40:25–47, 2001.



(a) Frame 21 (b) 3-channel Matting (c) 4-channel Matting (d) 4-channel Matting with f_t

Figure 8. Comparison of compositions using different methods.

(a) Frames (b) Mattes (c) Frames (d) Depth (e) Composition

Figure 9. Results from dynamic scenes of challenging cases.

[6] L. Grady, T. Schiwietz, S. Aharon, and R. Westermann. Random walks for interactive alpha-matting. In *ICVIIP*, pages 423–429, 2005.

[7] N. Joshi, W. Matusik, and S. Avidan. Natural video matting using camera arrays. In *SIGGRAPH*, 2006.

[8] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR*, 2006.

[9] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging understanding workshop*, pages 121–130, 1981.

[10] M. McGuire, W. Matusik, H. Pfister, J. Hughes, and F. Durand. Defocus video matting. In *SIGGRAPH*, 2005.

[11] E. Mortensen and W. Barrett. Intelligent scissors for image composition. In *SIGGRAPH*, 1995.

[12] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003.

[13] D. Singaraju and R. Vidal. Interactive image matting for multiple layers. In *CVPR*, 2008.

[14] A. Smith and J. Blinn. Blue screen matting. In *SIGGRAPH*, 1996.

[15] J. Sun, J. Jia, C. Tang, and H. Shum. Poisson matting. In *SIGGRAPH*, 2004.

[16] J. Sun, Y. Li, S. Kang, and H. Shum. Flash matting. In *SIGGRAPH*, 2006.

[17] Y. Taguchi, B. Wilburn, and C. Zitnick. Stereo reconstruction with mixed pixels using adaptive over-segmentation. In *CVPR*, 2008.

[18] J. Wang and M. Cohen. Simultaneous matting and compositing. In *CVPR*, 2007.

[19] O. Wang, J. Finger, Q. Yang, J. Davis, and R. Yang. Automatic natural video matting with depth. In *Pacific Graphics*, 2007.

[20] J. Xiao and M. Shah. Accurate motion layer segmentation and matting. In *CVPR*, 2005.

[21] W. Xiong and J. Jia. Stereo matching on objects with fractional boundary. In *CVPR*, 2007.

[22] K. Yoon and I. Kweon. Locally adaptive support-weight approach for visual correspondence search. In *CVPR*, pages 924–931, 2005.

[23] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *CVPR*, 2008.