Automatic Natural Video Matting with Depth

Oliver Wang* owang@soe.ucsc.edu
Jonathan Finger* jfinger@cs.ucsc.edu
Qingxiong Yang† qingxiong.yang@uky.edu
James Davis* davis@cs.ucsc.edu
Ruigang Yang† ryang@cs.uky.edu

*University of California, Santa Cruz †University of Kentucky

Abstract

Video matting is the process of taking a sequence of frames, isolating the foreground, and replacing the background in each frame. We look at existing single-frame matting techniques and present a method that improves upon them by adding depth information acquired by a time-of-flight range scanner. We use the depth information to automate the process so it can be practically used for video sequences. In addition, we show that we can improve the results from natural matting algorithms by adding a depth channel. The additional depth information allows us to reduce the artifacts that arise from ambiguities that occur when an object is a similar color to its background.

1 Introduction

In video production, it is common to need to remove the background from a sequence and put a new one in its place. This process requires what is called an alpha matte, which defines what percent of each pixel is occupied by the foreground. Typically the alpha matte is computed using a blue (or green) background for easier segmentation. However, this technique is not useful in all situations since it requires a calibrated studio setup with special equipment. Natural matting methods are a class of algorithms that attempt to solve the image matting problem without prior knowledge of the background.

Natural matting algorithms often require a user generated segmentation to identify background, foreground, and unknown regions. This segmentation is called a *trimap*. In general, trimaps must be drawn by hand, either for each frame, or at keyframes. This makes the algorithm difficult to extend to video because manually creating trimaps for many frames is far too costly when dealing with long sequences. Most natural matting algorithms also work only on the image domain and are therefore susceptible to errors in places where adjacent sides of a depth discontinuity have similar colors.

We present two contributions using the additional information acquired by a depth camera. First, we remove the frame-by-frame manual step from the process by automating the trimap generation. Secondly, we use the depth information to disambiguate regions that are prone to error using standard natural matting. We demonstrate our method by augmenting two commonly used natural matting techniques, Bayesian matting [5] and Poisson Matting [13], to include the depth information. Figure 1 shows the overall approach. These improvements could be applied to other natural matting algorithms as well.

2 Related Work

There has been a large body of research concerning video and photo matting in general. Smith and Blinn analyzed a commonly used technique, constant color matting, using a blue screen [12] . Several single frame natural matting algorithms [10][11][6][5][13] were developed to work with a more wide range of backgrounds. However, one problem with all of these techniques is that they are not optimal when dealing with video, as they require trimaps. In addition, while these algorithms are capable of producing high quality results, there is an inherent ambiguity with regions across depth boundaries with similar colors. Flash matting expands upon Bayesian matting by collecting two images, one with a flash on and one with a flash off [14]. This extra information helps improve the quality of the results, and reduces the likelihood of same color ambiguities.

There have been several groups that have presented solutions to video matting. Chuang et al. showed that optical flow could be used to interpolate hand drawn trimaps across time, which reduces the amount of manual input required [4]. Multiple cameras were also used to bypass the manual input trimap problem [9] [8]. These methods are able to generate trimaps automatically, but because they are dependent on specific apertures and blurring, they can be thrown off by scenes with inadequate light, motion blur, or lack of texture information. A Bayesian matting modification has been made using spatiotemporal information to create video mattes [2]. However, it assumes a moving foreground object, which is not always the case. Zitnick et al. [16] uses image segmentation-based stereo to construct a depth map, which is used to separate the foreground from background, but do not compute partial alpha values.

There have been other hardware solutions that use depth

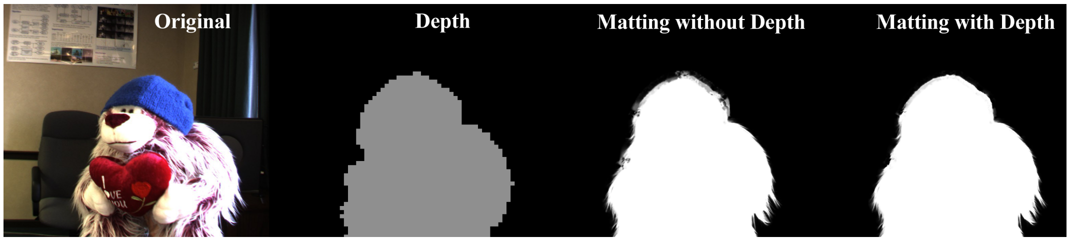


Figure 1. An overview of our algorithm showing the input color and depth images, matting without depth, and our solution incorporating depth.

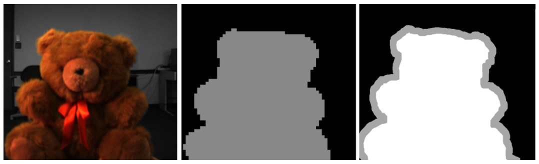


Figure 2. Creating a trimap. Left: The original image, Middle: Depth map, Right: Automatically generated trimap

information for matting purposes. 3DV Systems [1] developed a depth and image camera combination called the ZCam [7] which is able to perform foreground/background segmentation, but uses a simpler alpha value computation. Our work is similar to the ZCam in that we both use time-of-flight sensors to acquire depth data.

3 Method

In order to perform our matting, we first create a trimap from the depth image. The trimap is used as input to our modified matting methods to generate an alpha matte. The second step is to modify Bayesian and Poisson matting to use the depth information.

3.1 Automatic Trimap Generation

We use our depth information to automatically generate an accurate trimap. This is done in 3 steps: upsampling, thresholding, and dilating.

For each frame we have a high resolution picture taken with a digital camera and a low resolution depth map taken with the depth camera. The depth information that we used for our dataset was collected using the CanestaVision [3] depth camera. This camera computes a 64x64 resolution depth image. We use a super resolution method presented by Yang et al. [15] to generate high resolution depth images. This method is able to upsample the depth map up to 100 times the original resolution with little visible error.

We then compute a background-foreground segmentation using the depth information. We require that the user define a dividing plane that separates objects that lie in the foreground and objects that lie in the background. This step can not be automatic because it is a user's decision as to what is considered background for a given scene. We then compute a threshold on the distance plane over a video sequence.

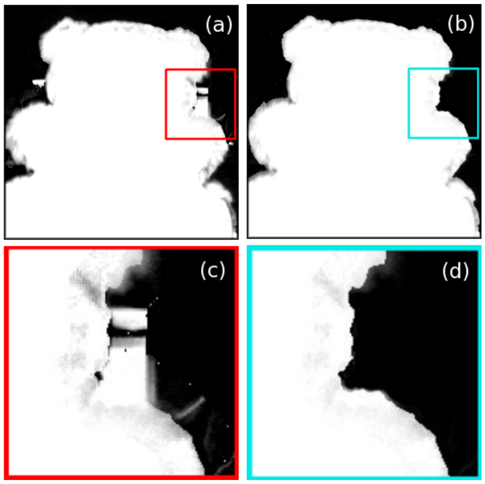


Figure 3. Left: Standard Bayesian matting. Right: Our improved Bayesian matting, showing the removal of undesirable artifacts.

The two-color image is the beginning of our trimap. We need to first determine the unknown region around an object. To do this we erode and then dilate the foreground. The exact amount of erosion and dilation is specified by the user and is dependent on the "fuzziness" of the object in the foreground. We now have a trimap that we can use to compute the alpha matte for each frame of video. This step is shown in Figure 2.

3.2 Improving Natural Matting

Natural matting algorithms generally work by estimating the unknown background, unknown foreground and unknown alpha value. Different algorithms use different methods to approximate these parameters. We perform our tests on two separate algorithms: Bayesian and Poisson matting. However, our method could be added to any natural matting method that operates on RGB images.

Natural matting techniques operate on the RGB image domain and therefore do not always produce desirable results when there are similar colors in the foreground and background, which can cause large false positive regions outside the object and false negative regions inside the object. By using the depth information in the error minimization step, we are able to prevent this bleeding.

Bayesian Matting Bayesian matting maximizes a joint probability expressed using Bayes Rule as follows:

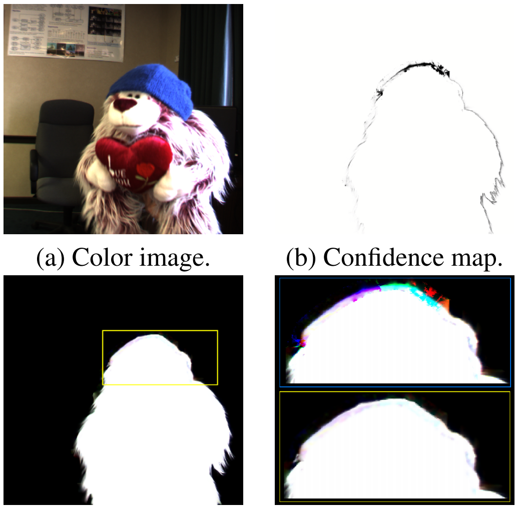
$$\begin{aligned} &\arg\max_{F,B,\alpha} P(F,B,\alpha|C) = \\ &\quad \arg\max_{F,B,\alpha} L(C|F,B,\alpha) + L(F) + L(B) + L(\alpha) \end{aligned} \tag{1}$$

The term $L(C|F,B,\alpha)$ is the log probability of the observed pixel value C given a predicted foreground F, background B, and α. $L(F)$ and $L(B)$ are the log probabilities of colors F and B being the foreground and background respectively. $L(\alpha)$ is the log probability of α, which for our implementation is assumed to be constant. The algorithm works its way from the outside in until the whole unknown area is filled.

Our depth information gives us strong evidence for whether the object is foreground or background in regions with strong depth edges. However, this information is inaccurate when the object is semitransparent (has an alpha value that is not 0 or 1). We therefore weigh our confidence in the depth channel based on the estimated alpha value, such that the weight is high when the alpha value tells us that we are seeing mostly background or mostly foreground. We include the weighted depth information as a fourth color channel into the Bayesian matting and perform the same minimization as presented by Chuang et al.[5]. This step is shown in Figure 3.

Poisson Matting Using depth information in the Poisson matting approach is different from the Bayesian approach. Poisson matting converts color images into a single-channel image. Simply treating the depth map as an additional channel leads to a poor alpha matte with an appearance similar to the depth map, since the binary depth map heavily influences the gradient field. To integrate the depth map into the Poisson matting approach, a confidence map is produced that is based upon the consistency of the three channels of the matte generated by the global Poisson matting approach:

$$\begin{aligned} \alpha_{min} &= \min(\alpha(0), \alpha(1), \alpha(2)); \\ \alpha_{max} &= \max(\alpha(0), \alpha(1), \alpha(2)); \\ F1 &= \prod_{d=0}^{2} exp(-\frac{(\alpha(d) - \alpha_{min})^2}{2\sigma^2}); \\ F2 &= \prod_{d=0}^{2} exp(-\frac{(\alpha(d) - \alpha_{max})^2}{2\sigma^2}); \\ F &= \min(F1, F2), \end{aligned} \tag{2}$$



(a) Color image. (b) Confidence map.

(c) Alpha matte. (d) Local alpha matte.

Figure 4. Improved Poisson matting. A confidence map is produced by measuring the consistency of the RGB channels of the alpha matte generated from the global Poisson matting approach. The confidence map is then used as guidance for the combination of the binary depth map and the alpha matte. Note that we show the independent matte for each color channel for illustration purpose.

where $\alpha_{min}/\alpha_{max}$ is the minimum/maximum of the matte, and F is the confidence map. The final alpha matte is the linear combination of the matte generated from the Poisson matting approach and the depth map based on the confidence map F:

$$\alpha' = F\alpha + (1 - F)D, \tag{3}$$

where D is the binary depth map. Figure 4 provides a visual comparison of the alpha matte with and without integrating depth information.

Poisson matting assumes that the gradient change in the unknown regions within the trimap is caused by foreground/background transitions only. This assumption is violated when there are textures in the foreground or background. The depth map is independent of textures and therefore provides a better estimation of the boundary in this case.

4 Experimental Results

We tested our approaches using several real sequences we captured. Our experimental setup uses two cameras: one for depth and one for color. We register these two images via an affine transformation. Given the low resolution from the depth sensor, we found that this simple method provides a good enough registration between the two cameras.

Our automatically generated trimaps worked well with both of our modified natural matting algorithms, greatly reducing the amount of noise when compared to the original

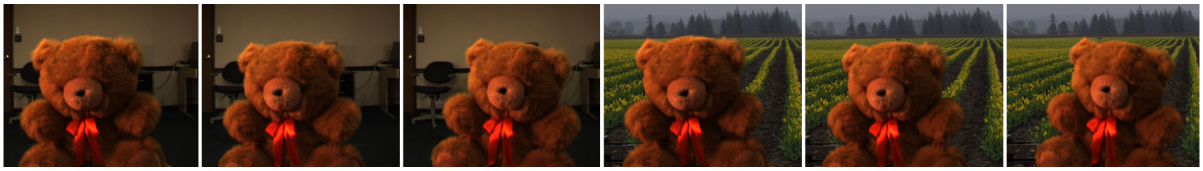


Figure 5. Video matting using improved Bayesian matting. Left: The original scene. Right: Background replaced.

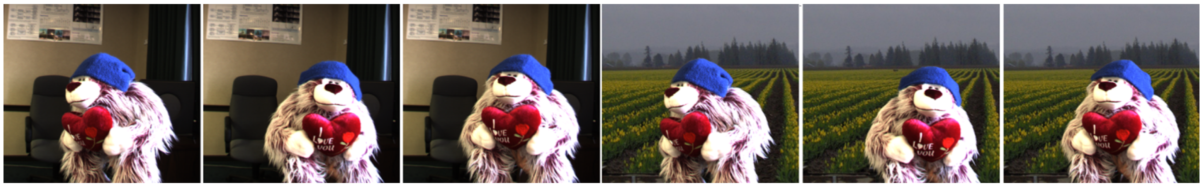


Figure 6. Video matting using improved Poisson matting. Left: The original scene. Right: Background replaced.

methods without using depth information. Figure 5 shows the output on several frames of a scene using Bayesian matting and Figure 6 shows the results from Poisson matting. The full sequences can be seen in our video. The slowdown from incorporating the depth channel into the matting algorithm was negligible.

5 Limitations and Future Work

One shortcoming of this research is that we require a dividing plane to segment the foreground and the background, this assumption is not always true. One common example of this would be feet and a floor that is visible to the camera. Since the floor's depth spans from in front of the foot to behind it, a single depth cut will divide the floor into two segments which may not be desirable. However, our approach represents a new way of dealing with video matting. With a depth camera we were able to speed up video processing dramatically by removing the manual step, making natural matting approaches more accessible for video. We also proposed an improvement in accuracy by including the use of a depth camera. It would be simple to incorporate this system into a single camera that captures both depth and color from the same viewpoint.

Acknowledgements This work is supported in part by University of Kentucky Research Foundation, U.S. Department of Homeland Security, and U.S. National Science Foundation grant IIS-0448185.

References

[1] 3DV Systems. http://www.3dvsystems.com/.

[2] N. Apostoloff and A. Fitzgibbon. Bayesian video matting using learnt image priors. *Proceedings, CVPR*, pages 407–414, 2004.

[3] Canesta Inc. http://www.canesta.com/.

[4] Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski. Video matting of complex scenes. In *Proceedings of ACM SIGGRAPH 2002*, pages 243–248, New York, NY, USA, 2002. ACM Press.

[5] Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A Bayesian approach to digital matting. *Proceedings, CVPR*, 2:264–271, 2001.

[6] P. Hillman, J. Hannah, and D. Renshaw. Alpha channel estimation in high resolution images and image sequences. *Proceedings of CVPR*, 1:1063–1068, 2001.

[7] G. Iddan and G. Yahav. 3D Imaging in the studio (and elsewhere). *Proceedings SPIE 2001*, 4298:48, 2001.

[8] N. Joshi, W. Matusik, and S. Avidan. Natural video matting using camera arrays. In *Proceedings, SIGGRAPH*, pages 779–786, New York, NY, USA, 2006. ACM Press.

[9] M. McGuire, W. Matusik, H. Pfister, J. Hughes, and F. Durand. Defocus video matting. *Proceedings of ACM SIGGRAPH*, 24(3):567–576, 2005.

[10] Y. Mishima. Soft edge chroma-key generation based upon hexoctahedral color space, Oct. 11 1994. US Patent 5,355,174.

[11] M. Ruzon and C. Tomasi. Alpha estimation in natural images. *Proceedings of CVPR*, 1:18–25, 2000.

[12] A. Smith and J. Blinn. Blue screen matting. *Proceedings of conference on Computer graphics and interactive techniques*, pages 259–268, 1996.

[13] J. Sun, J. Jia, C. Tang, and H. Shum. Poisson matting. *ACM Transactions on Graphics*, 23(3), 2004.

[14] J. Sun, Y. Li, S. Kang, and H. Shum. Flash matting. *International Conference on Computer Graphics and Interactive Techniques*, 2006.

[15] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *Proceedings, CVPR*, Minneapolis, MN, USA, 2007. IEEE Computer Society.

[16] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *Proceedings, SIGGRAPH*, pages 600–608, New York, NY, USA, 2004. ACM Press.