Automatic Natural Video Matting with Depth

Oliver Wang
Department of Computer Science
University of California Santa Cruz
owang@soe.ucsc.edu

Jonathan Finger
Department of Computer Science
University of California Santa Cruz
jfinger@cs.ucsc.edu

Qingxiong Yang
Department of Computer Science
University of Kentucky
qingziong.yang@uky.edu

James Davis
Department of Computer Science
University of California Santa Cruz
davis@cs.ucsc.edu

Ruigang Yang
Department of Computer Science
University of Kentucky
ryang@cs.uky.edu

Abstract

Video matting is the process of taking a sequence of frames, isolating the foreground, and replacing the background with something different in each frame. This is an under-constrained problem when the background is unknown. Matting techniques exist to approximate these values using manual input cues. 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. Our goal is to make it easy for a user to do this with an arbitrary background. This process requires an alpha matte, which is an image that defines the percent of each pixel that is occupied by the foreground (this value is often called the alpha (α) value). Typically the alpha matte is computed using a film studio and a blue (or green) screen for easier segmentation. With a known background, the matting problem becomes much easier. However, this method is not useful in all situations, since it requires a calibrated studio setup with special equipment. Not all videos that we would like to remove the background from are taken in such isolated environments. Natural matting methods are a class of algorithms that attempt to solve the image matting problem without prior knowledge of the background. They are 'natural' in the sense that the image capture can take place outside of a studio with unrestricted visual components. The problem is under-constrained, so some external information is required. One of the goals of such algorithms is to produce a mask used to separate the foreground from the background with the least amount of additional information possible.

Natural matting algorithms often require a user generated segmentation to identify background, foreground, and unknown regions. This segmentation is called a *trimap*. The algorithms then use information given by the trimap to disambiguate the unknown regions. One problem with trimaps is that they must generally 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 timely a task when dealing with long sequences. In addition, most natural matting algorithms work only on the image domain and therefore are susceptible to errors in locations where two 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 automat-

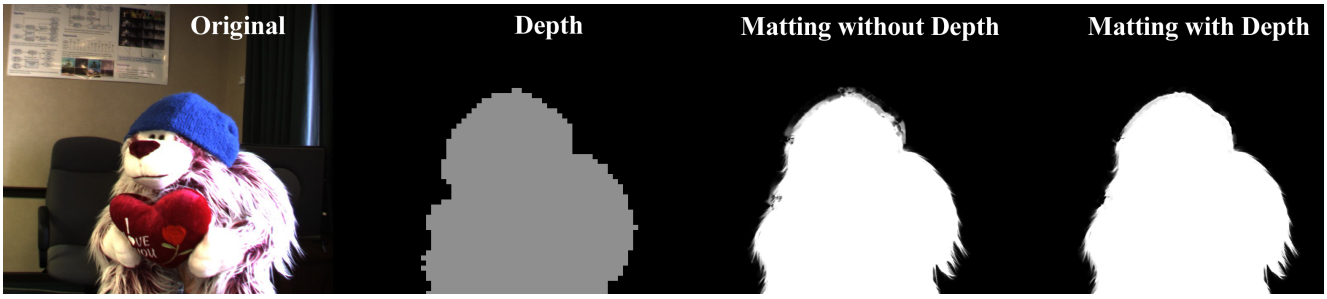


Figure 1. An overview of our algorithm showing the input images, current existing solution, and solution using our method.

ing the trimap generation. Secondly, we use the depth information to disambiguate regions that are prone to error using standard natural matting, such as areas with similar foreground/background color. The user is required to set a few parameters only once for a video sequence. These parameters must be set by the user since they represent preferences for the desired output.

Using a time-of-flight sensor that can capture full-frame scene depth at video rate, we have extra information at depth edges that can prevent bleed over artifacts visible in existing matting techniques. 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 effectiveness of our approach. These improvements could be applied to other natural matting algorithms as well.

2 Related Work

There has been extensive research done concerning video and photo matting in general. The stage was set by Smith and Blinn [12] when they analyzed a commonly used technique, constant color matting, using a blue screen. This method is still frequently used today as it is relatively simple and effective. However, this technique requires that video be filmed in front of a studio setup with a controlled background and lights.

Several single frame natural matting algorithms [10] [11] [6] have been developed to reduce the dependence on a known background. They use different methods to estimate the background, foreground and alpha values at each pixel. Chuang et al. [5] presented a Bayesian method that has played an important role in this field. However, a problem with all of these natural matting techniques is that they are not optimal when dealing with video since creating trimaps for thousands of frames is too tedious to do by hand. In addition, while these algorithms are capable of producing high quality results, they operate on the image intensity domain, so there is an inherent ambiguity with regions across depth boundaries with similar colors. Poisson matting [13] attempts to reduce these problems by introducing a set of image based processing steps that are manually adjusted to improve results, assuming a smoothness constraint and analyzing the gradient domain rather than the image domain. However, the inherent intensity ambiguity still exists. Flash matting by Sun et al. [14] expands upon Bayesian matting by collecting two images, one with a flash on and one with a flash off. This extra information helps improve the quality of the results, and reduces the likelihood of same color ambiguities. Flash matting is specific to single images, and would be hard to extend to video since every frame must have symmetric flash and non-flash images.

There have been several groups that have presented solutions to video matting. Chuang et al. [4] added to their previous work by showing that optical flow can be used to interpolate hand drawn trimaps across time. This reduces the amount of manual input, but does not get rid of it all together. McGuire et al. [9] bypasses the manual input trimap problem by using aligned cameras, where each is at a different level of focus. The blurred background can be automatically converted into a good trimap from which alpha matting can be applied. Because it is dependent on specific apertures and blurring, it can get thrown off by scenes with inadequate light or motion blur. In addition, depth from defocus requires objects to have high frequency texture information, which is not always available. Apostoloff et al. [2] presents a Bayesian matting modification with learned priors using spatiotemporal information and loose trimap generation to matte video. Their background estimation technique is based on the idea that movement between frames gives a hint of which object is in the foreground. Since the foreground object is not necessarily moving in all cases, we prefer not to make that assumption. Joshi et al. [8] performs video matting by using an array of cameras. They create a synthetic aperture image and analyze image statistics to determine what is in the foreground. This method generates

automatic trimaps for video, but requires a textured background.

There have been other hardware solutions that use depth 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. It uses depth information to split the scene into distance regions based on dividing planes set by the user. However, their method does not compute partial foreground (α) values, and therefore has noticeable artifacts for objects with fuzzy borders such as hair. Our work is similar to the ZCam in that we both use time-of-flight sensors to acquire depth data.

Our method incorporates automatic trimap generation and is able to disambiguate regions where traditional natural matting fails. We use an active depth sensor in conjunction with a camera to capture our data. The depth sensor provides depth information by sending a pulse of light and counting the time that it takes for the reflection to come back and therefore works independently of object, color or texture.

3 Method

In order to perform our matting method, we first create a trimap from the depth image. The trimap is used as input to our modified matting methods (Bayesian or Poisson based) to generate an alpha matte. The matte is used to separate the foreground from the image in order to put a new background behind it. This offline process can be applied to a full sequence of frames making it ideal for video data.

Matting can be generally described by the formula:

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

where C is the observed pixels in the composite image, F is the unknown foreground, B is the unknown background, and α is the percent of each pixel that is occupied by the foreground. The α value is visually important with fine objects, such as hair, where the color of a single pixel may be a combination of both the background and foreground.

Because there are three unknowns and one known, the problem is highly under-constrained. Fortunately, only a small amount of additional information is required to achieve convincing results. Different matting techniques solve the problem by predicting foreground, background and alpha values that minimize some constraint, usually determined by the color of neighboring pixels which are known to be foreground or background. Therefore, the user must specify where these known regions are. Our method is able to use a depth camera to perform this segmentation so the user does not have to.

3.1 Automatic Trimap Generation

A trimap is a three color image that contains information about what is foreground and what is background. In our implementation, white represents the foreground, black represents the background, and gray represents the unknown region. We show that we can use our depth information to automatically generate an accurate trimap. This is done in 3 steps: upsampling, thresholding, and dilating. Figure 2 shows the results from these steps.

Upsampling

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. While high resolution images are cheap due to CCD technology, it is not as simple to capture a depth value at each pixel. Therefore, most depth cameras do not provide the same level of resolution as the color 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. For our algorithm, we desire a trimap at the same resolution as our original image, so we must first upsample the depth map to get the result in Figure 2c. Naive upsampling methods such as bi-linear interpolation or nearest neighbor will create depth edges that cross boundaries. Instead, 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.

Thresholding

We must then compute a background-foreground segmentation using the depth information, which is provided as an image where each intensity value represents a depth from the camera. To do the segmentation, we require that the user provide a dividing plane that defines which objects lie in the foreground and which objects lie in the background. Without this choice, it would be impossible to tell what is meant by foreground and background. As with the ZCam, this step is not automatic because we feel it is fundamentally a user's decision as to what is semantically defined as the background for each situation. This manual input must only be given once per capture session. We then compute a threshold on the distance plane, which gives us a background-foreground segmentation of the image. The thresholded image is shown in Figure 2d.

Dilating

The two-color image is roughly accurate around the edges. However, it is not precise, and does not take into account

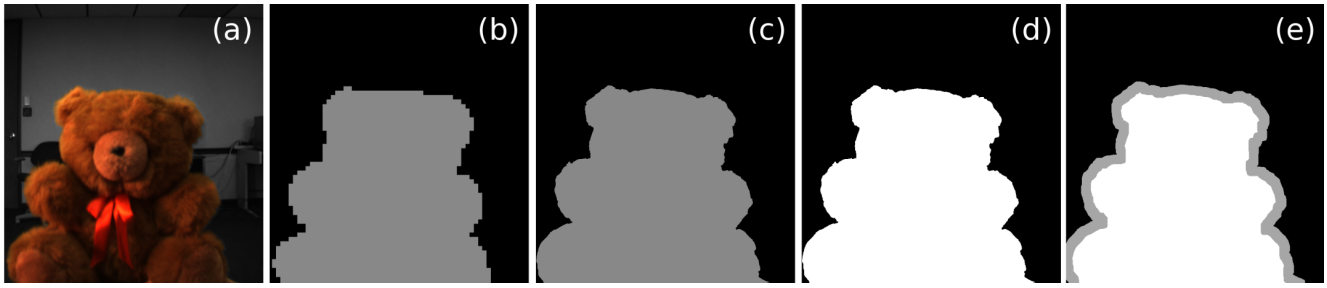


Figure 2. Creating a trimap. (a) The original image, (b) low resolution depth map, (c) supersampled depth map, (d) thresholded depth map, (e) dilated to create a trimap.

partial alpha values. We need to determine the region around the object that is unknown foreground and background. To do this we erode and then dilate the foreground. Anything that was modified in either of these steps is considered unknown. 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 (as shown in Figure 2e) that we can compute for each frame in very little time. Using these trimaps, we can generate an alpha mask at each frame in the video.

3.2 Improving Natural Matting

Natural matting algorithms generally work by computing the unknown background, unknown foreground and unknown alpha value in Equation 1. 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 three channel images.

Natural matting techniques operate on the image intensity domain and therefore do not always produce desirable results when there are similar colors in the foreground and background. In such cases many algorithms are unable to differentiate between background and foreground colors and produce a bleed of the predicted foreground color into the background. The result is the creation of large false positive regions outside the object and false negative regions inside the object. Some of these artifacts can be seen in Figure 3c. By using the depth information in the error minimization step, we are able to prevent this bleeding.

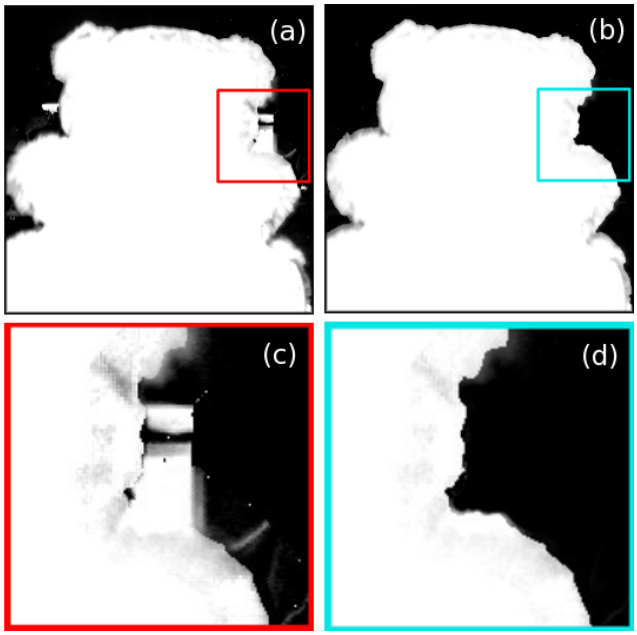


Figure 3. Improved matting. Left: Bayesian matting. Right: The improved algorithm gets rid of undesirable artifacts.

Bayesian Matting

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

$$\arg\max_{F,B,\alpha} P(F,B,\alpha|C) = \quad (2)$$
$$\arg\max_{F,B,\alpha} L(C|F,B,\alpha) + L(F) + L(B) + L(\alpha)$$

where L is the log of probability $L = log[P]$. 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 the colors F and B being the foreground and background respectively. $L(\alpha)$ is the probability of an alpha value, 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 using an inverse entropy function:

$$H'(\alpha) = 1 + \alpha \log \alpha + (1-\alpha)\log(1-\alpha)) \quad (3)$$

This function is high when the alpha value tells us that we are seeing mostly background or mostly foreground. We then include the weighted depth information as a fourth channel of information into Bayesian matting and perform the same minimization solver as presented by Chuang et al.[5].

Poisson Matting

Using depth information in the Poisson matting approach is different from the Bayesian approach. Poisson matting typically converts color images into a single-channel image. Treating the binary 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:

$$\alpha_{min} = \min(\alpha(0), \alpha(1), \alpha(2));$$
$$\alpha_{max} = \max(\alpha(0), \alpha(1), \alpha(2));$$
$$F1 = \prod_{d=0}^{2} exp(-\frac{(\alpha(d) - \alpha_{min})^2}{2\sigma^2});$$
$$F2 = \prod_{d=0}^{2} exp(-\frac{(\alpha(d) - \alpha_{max})^2}{2\sigma^2});$$
$$F = \min(F1, F2), \quad (4)$$

Figure 5. Our data capture setup, consisting of a depth camera and a color camera.

where α_{min}/α_{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:

$$\alpha' = F\alpha + (1-F)D, \quad (5)$$

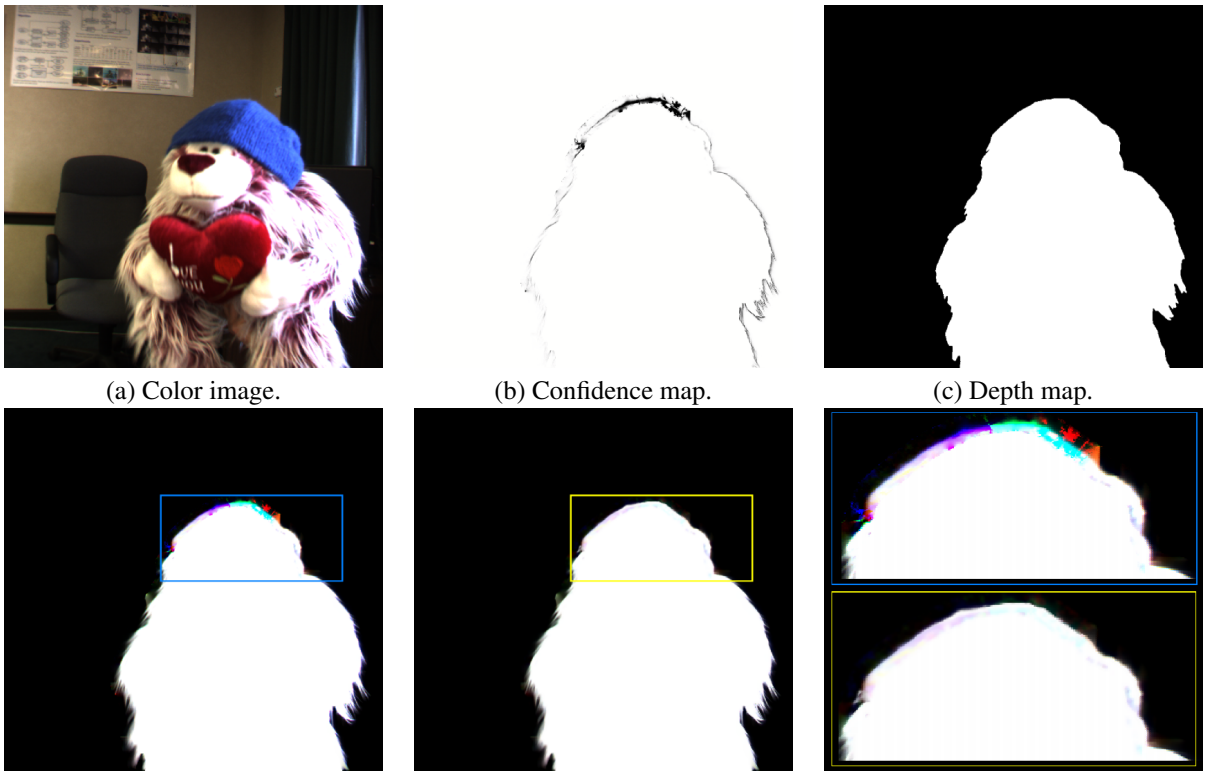
where D is the binary depth map. Figure 4 provides a visual comparison of the alpha matte with and without integrating depth information. Also provided is the corresponding color image, the confidence map, and the depth map.

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 (as the sharp black/ivory transition in the background, near the top of the blue hat in Figure 4 (a)). 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. Currently our experimental setup uses two cameras: one for depth and one for color. Figure 5 shows our data capture setup. There is a small baseline between these two sensors. 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 methods without using depth information. Figure 6 shows the output on several frames of a scene using Bayesian matting and Figure 7 shows the results from Poisson matting. The full sequences can be seen in our video. The slowdown



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

(d) Alpha matte. (e) Refined alpha matte. (f) Local alpha matte.

Figure 4. Depth-combined 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.

from incorporating the depth channel into the matting algorithm was negligible.

5 Limitations

One shortcoming from this research is that we require a dividing plane to segment the foreground and the background. Therefore, when there is an object that occurs closer than a foreground object it will be included in the foreground as well. 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 might not be desirable.

Since we are determining the background partially based upon an exact distance, the cut off is a single plane parallel to the camera. Though this is sufficient for our research, this model could be improved. The cut off area could instead be a plane at an angle or a surface with a different shape.

6 Conclusion and Future Work

Our approach represents a new way of dealing with video matting. With a depth camera we were able to speed up video processing dramatically. This makes natural matting approaches more accessible for video. We also proposed an improvement in accuracy by including the use of a depth camera. In the future we would like to see how the semantic meaning of foreground could be interpreted in a scene. It would be possible to incorporate this system into a single camera that captures both depth and color from the same viewpoint.

References

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

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

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

Figure 6. Video matting using improved Bayesian matting. Top: The original scene. Bottom: Background replaced.

Figure 7. Video matting using improved Poisson matting. Top: The original scene. Bottom: Background replaced.

[4] Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski. Video matting of complex scenes, 2002.

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

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

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

[8] N. Joshi, W. Matusik, and S. Avidan. Natural video matting using camera arrays. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, 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 2005*, 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 Computer Vision and Pattern Recognition (CVPR 2000)*, 1:18–25, 2000.

[12] A. Smith and J. Blinn. Blue screen matting. *Proceedings of the 23rd annual 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 *To appear in CVPR 2007*, 2007.