Video Matting from Depth Maps

Anonymous CVPR submission

Paper ID ****

Abstract

Image matting is the process of computing an alpha value for each pixel that corresponds to the amount of the pixel that is foreground and the amount that is background. This is often used to isolate the foreground and replace the background with a different image. Typically this is done using a special studio and a blue (or green) screen for easier segmentation. However, this method is not as robust as it requires a calibrated studio setup with special equipment. Not all videos that we would like to process are taken in such isolated environments. Our project attempts to isolate the background in videos that can be taken in the field, and do not need a special screen behind the person in the foreground. We use a bayesian image matting approach to perform the matting. We capture data using a low resolution range camera and high resolution color video camera. We use the depth information to improve the quality of the matting over a standard Bayesian matting technique.

1. Introduction

The idea behind image matting is the separation of the foreground and background images. The hardest problem with matting is being able to generate an alpha channel that will be able to isolate these layers. Another way to envision it is to separate the layers into separate images as shown in Figure 1. If we define C to represent the pixels in the composite image, B to represent pixels in the background image, and F to represent the pixels in the foreground image, we can use the compositing equation:

\[ C = \alpha F + (1 - \alpha)B \]  

Once this is solved we have an opacity value at that pixel. In the case of classical blue screen studios, C and B are known. In our case, we only know C, so we must devise a method for solving for the three unknowns.

We want to be able to specify a threshold distance and create an alpha map that bifurcates the depth map at that distance. Then the alpha map can be used to combine the foreground of the original image and the new background image.

Our project looks looks at several ways this can be done with video data. For each frame we have a high resolution picture (taken from a digital camera) and a low resolution depth field map (taken with a Canesta depth camera). To generate a proper alpha mask for the image, we begin by upsampling the depth map. We then create an three color image called a trimap where the colors represent the definite foreground, the definite foreground, and the questionable area. We then correct the edges in the map using a Bayesian approach.

Our contribution to the study of video matting is a new way to generate automatic trimaps. Instead of relying on user input to generate trimaps, we devised a new method to generate the trimaps on the fly. There is no manual input needed, and since our algorithm guesses a better trimap then a human would, it makes a very nice video matte. We also use the depth maps to speed up the processing of the animation step. We speed up multiframe processing by using a combination of optical flow and Bayesian matting instead of just relying on matting every frame individually. In addition, we use the depth information in the Bayesian matting framework to increase the knowledge of the energy minimization and improve our results.

2. Related Work

There has been a lot of work done on video matting, and photo matting in general. The stage is set by Smith [5] by analyzing the most commonly used technique, constant color matting, using a blue screen. Chaung’s [2] Bayesian algorithm has played a most important roll in this field. It requires a manual definition of a trimap, which is not optimal when dealing with video since creating trimaps for thousands of frames can take days to do by hand. Though it does not address animated images, it created an excellent alpha matte, and served as the basis of our single frame matting technique.

McGuire [4] bypasses the manual input trimap problem by using aligned cameras, where each is a different level of focus. The blurred background can be automatically con-
thresholding the depth map into a two color image. Then the next step is to generate an accurate trimap. We begin by optical image. sample the depth map so it is the same resolution as the Kolmogorov [3] cases, we prefer not to make that assumption. Since the foreground object is not necessarily moving in all frames gives a hint of which object is in the foreground. motion technique is based on the idea that movement between successive 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.

Another approach to video matting was performed by Apostoloff’s [1] research on Bayesian matting with learned priors uses spatiotemporal information and a loose trimap generation to mat 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.

Another approach to video matting was performed by Apostoloff’s [1] research on Bayesian matting with learned priors uses spatiotemporal information and a loose trimap generation to mat 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.

3. Depth Image

We can use the depth map to automatically generate a trimap for our image. The Canesta camera takes a 64x64 resolution depth image in the form of a color map. To get more accurate results we use Yang’s superresolution method [6]. This effectively uses the optical image to upsample the depth map so it is the same resolution as the optical image.

4. Bayesian Matting

Once a high resolution depth map has been generated the next step is to generate an accurate trimap. We begin by thresholding the depth map into a two color image. Then the edges are slightly blurred with a gaussian effect. The blurred edges are isolated and turned into a third color. This color represents the edge of the foreground object. Now we have a good trimap. To generate an alpha map, we solve Equation 1 using Bayes rule:

$$\arg \max L(L, B, \alpha | C) = \arg \max L(C | F, B, \alpha) + L(F) + L(B) + L(\alpha)$$

where $L$ is the log of probability $L = \log[P]$. Essentially, the algorithm looks at the questionable region and works its way from the inside out until the outer edges are sharpened. The middle most pixels are grown to encompass as much as possible until the edge is reached. Each region is a mixture of foreground, background, and previously computed values. Covariance matrices are then used to specify $P$ for the foreground and background distributions. We then have a sharp edge around the foreground area.

5. Video

Optical flow is a method to determine movement between successive frames. Each moving pixel is traced by a velocity to indicate where it is traveling. After processing a frame, we get back a field of flow map. We should then be able to track and isolate the moving foreground object.

However, researchers have had varying results with this procedure. One reason, is a robust flow method must take into account the fact that a foreground object is not necessarily moving. Another main difficulty, is dealing with outliers. Outliers are foreground regions that were not in the previous frame but are in the current one. There are no obtainable velocity vectors for these since they would have to originate from outside the picture. We can use our depth images to overcome these issues.

We create a sort of trimap that will be turned into an alpha matte by using optical flow instead of Bayesian matting.
6. Results

Our automatically generated trimap, is good and makes and works well with the Bayesian matting algorithm. Figure 2 shows our results on a single frame. Figure 3 shows the output on several frames of movement for a scene involving complex foreground and background movement.

7. Conclusion

Our algorithm represents a new way of dealing with video matting. With a depth map video we were able to do a lot more with frame matting, and optical flow. The frame to frame processing is quick. With a little refactoring and more processing power, this has the potential to work in real time.

Since the upsampling algorithm is able to produce such good images from low resolution, we could potentially take even lower resolution depth shots. This has yet to be tested, and we intend to investigate it further in future research. It would be easy enough to package a dual camera (half optical, half IR depth) as a web cam.

References


