Stereo Matching on Objects with Fractional Boundary

XIONG, Wei

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Computer Science and Engineering

July 2007

Thesis/Assessment Committee

Professor WONG Kin Hong (Chair)
Professor JIA Jiaya (Thesis Supervisor)
Professor WONG Tien Tsin (Committee Member)
Professor TANG Chi Keung (External Examiner)

Abstract of thesis entitled:

Stereo Matching on Objects with Fractional Boundary

Submitted by XIONG Wei

for the degree of Master of Philosophy

at The Chinese University of Hong Kong in May 2007

Stereo matching and digital matting are two most basic and important topics in computer vision. Conventional stereo matching problem assumes the color constancy on the corresponding opaque pixels in the stereo images. However, when the foreground objects with the fractional boundary are blended to the scene behind using unknown alpha values, due to the difference of the spatially varied disparities for different layers, the color constancy does not hold any more. On the other side, when the color of background scene is close to the foreground object, conventional digital matting methods are always failed to achieve the alpha matte correctly without any other cues. This dissertation focuses on introducing digital matting method into stereo matching framework to improve the performance of both these two cases.

This dissertation first includes a basic survey of state-of-art narrow-band stereo matching and digital matting. To tackle the

fractional stereo matching problem, we introduce a probability framework constraining the correspondences of the pixel colors, disparities, and the alpha values in different layers, and propose an automatic optimization method to solve a Maximizing a Posterior (MAP) problem using Expectation-Maximization (EM), given the input of only a narrow-band stereo image pair. Our method naturally encodes the effect of occlusion in the formulation of layer blending without a special detection process. The depth map on the fractional area can be greatly improved using the optimized foreground and background color. Based on the relationship developed between the image pair, better alpha matte results can also be achieved. We demonstrate the efficacy of our method using difficult stereo images where comparisons with state-of-art techniques are also given.

論文題目：

在具有複雜邊界的物體上的立體匹配

論文摘要：

立體匹配和景象萃取是電腦視覺領域中最基本並且也是最重要的兩個課題。傳統的立體匹配問題通常假設在立體圖像對之間，對應的象素顏色恆定不變。然而,當我們考慮照片中具有複雜邊界的前景物體時，它的邊界顏色會以未知的混和比例和背景顏色混和。由於不同層次在不同的圖片中具有不同的空間位置，因此在複雜的邊界上象素的顏色將不再恆定不變。另一方面，當前景物體的顏色和背景場景的顏色過於接近的時候，如果沒有別的線索引入，常規的景象萃取的方法通常無法正確恢復前景物體和背景在每個象素上的混和比例值。本論文著重於將景象萃取的方法引入立體匹配的框架，從而同時提高這兩者的工作性能。

本論文首先包括一個對目前流行的立體匹配和景象萃取方法的基本調查。爲解決在複雜邊界上的立體匹配問題，我們引進了一種基於概率論的框架。其中包含了不同層次上的對應象素的顏色，空間差距，以及混和比例。同時，基於期望最大化方法(EM)，我們提出了一個自動優化方法來解決這個建立在窄差距立體匹配圖像對上的最大後驗概率(MAP)的問題。我們的方法自然的涵蓋了立體匹配中的遮擋問題。在優化後的前景和背景顏色的基礎上，在複雜邊界上求得的深度圖的結果可以得到很大的改善。同時利用建立在立體圖像對上的關係，我們可以求得更加精確的混和比例。我們用高難度的立體匹配圖像對來展示我們方法的有效性。另外，我們也展示出我們的方法和現有流行的科學方法之間的比較。

Acknowledgement

At the beginning of this thesis, I would like to thank the Department of Computer Science and Engineering of the Chinese University of Hong Kong, for giving me the permission to take the two-year study here. This period of time is very valuable and memorable indeed.

I am deeply indebted to my supervisor Professor JIA Jiaya. He is always nice, supportive and understanding. He led me walking into the world of Computer Vision and showed me the way to the brilliant future. His suggestions and encouragements helped me a lot in all the time of my course study and research. I feel really lucky for having such a good advisor in my M. Phil. study period.

I would like to thank Professor WONG Tien-Tsin. I was strongly influenced by his always energetic status and unstoppable passion towards science research. His nice direction in my first SIGGRAPH project brought me not only a much broader view and deeper thinking in the topic, but also a more clear understanding on the meaning of research. I would like to thank Professor WONG Kin Hong. He was my marker and gave me

many valuable ideas and suggestions to improve my research work from the first day of my M. Phil. study. Besides, I would also like to thank all the professors who gave instructions to me in CSE department.

I would like to thank my lab mate QU Yingge, WANG Guangyu, WAN Liang, SHI Lin, LIU Xiaopei, NI Bing, LI Wenye, MA Huiye, WU Wen, ZHANG Pingyue and PANG Wai-Man. They brought a family-like environment into our lab. Working with such excellent people is the most important factor for me having a such wonderful and fruitful life here. And the discussions with them both on research and personality are always valuable and helpful.

I would like to thank my friends FAN Bin, XU Leilei, TAO Chenjun, SHAN Qi, LIU Renting, CHUNG Hin Shun, ZHU Jianke, YANG Haixuan, WANG Hui, HU Yan, SUN Haibin, LU Shi, WEI Dan, JIANG Wenjie, MENG Wei, GAO Yan, ZHANG Fan, LI Gang, ZHANG Kun, CHEN Junzhou, DAI Hongning, XU Xuemiao, YUAN Yan, LI Shan, ZHOU Lin, CHEN Jiansheng, PENG Xiang, LI Jiexing, LE Jilin, CAI Xia, SHI Lei, MA Qiang, TENG Li, XIE Yongming, CAI Yi, CHEN Hui, DENG Hongbo, HOI Chuhong, LI Jian, LI Xiaoqi, LIN Minghong, LIU Dawei, QIN Jing, RAO Weixiong, TANG Xinmin, TIAN Ye, TU Shikui, WANG Jinfeng, WANG Yue, WU Di, XIAO Mingyu, XIAO Xiaokui, XU Zenglin, XU Yuedong and ZHOU Yangfan. I would also like to extend my gratitude to all those people who gave me help during my study and living here.

Finally, I would like give my special thanks to my parents, for their endless love and continual support in all the past years.

This work is dedicated to my dearest parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stereo matching has been an essential research topic in computer vision, and has made rapid and significant progress in recent years [20, 26, 30]. Most conventional two-frame stereo matching approaches compute disparities and detect occlusions assuming that each pixel in the input image has a unique depth value.

However, this representation has large limitations in faithfully modelling objects with fractional boundaries where pixels are blended to the scene behind with different depth values. Directly applying previous stereo matching methods on the ubiquitous hairy objects may produce problematic disparity results. One example is shown in Fig. 1.1 where the input images (a) and (b) contain a hairy fan in front of a background with similar colors. Directly applying the stereo matching method proposed in [20] generates problematic disparity result (c) along the fan's boundary without considering color blending.

Recent development on stereo matching algorithms partially

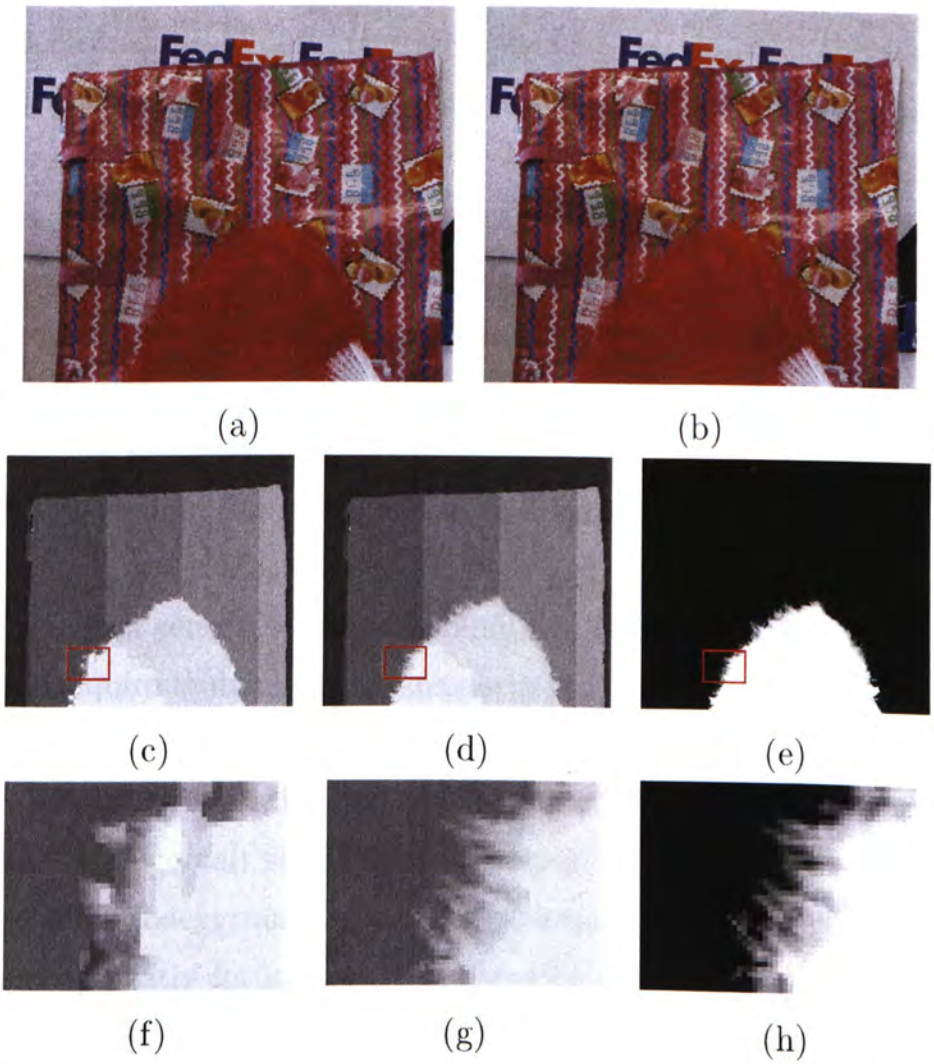

Figure 1.1: A stereo image pair containing a hairy object. (a) and (b) Input stereo images containing a hairy fan. Notice that the colors of the background scene and the hairy fan are similar. (c) Stereo matching result from Sun's method [20]. Because of the color blending, the assumption of color constancy is violated along the boundary of the fan, making the result problematic. (d) The stereo matching result obtained from our approach. The hairy structure is successfully preserved. (e) The computed alpha matte of the fan using our method. (f)-(h) Magnified regions of the results.

generalizes the above assumptions and introduces the transparency constraints. Szeliski *et al.* [23] proposed to solve the stereo matching problem with opacity using multiple input images where the color and transparency refinement are formulated as a non-linear minimization problem. However, their method has difficulties to deal with objects containing thin and long hairs or with complex alpha matte given a small number of input images. Later on, assuming a binary reflection map model, Tsin *et al.* [26] proposed to estimate the front translucent and rear background layers using graph cut. The pixel colors are further computed by iteratively reducing a difference energy in multi-frame configuration. This method is not applicable on objects with general fractional boundary. Both of the above methods require multiple input images in order to obtain satisfactory disparity maps.

In this dissertation, taking the input of only a narrow-band stereo image pair shot in a scene where the hairy objects are in front of a background scene, we solve the stereo matching problem by neatly formulate the estimation of alpha values, disparities, and pixel colors in a probability framework and robustly using Expectation-Maximization method. Unlike most previous stereo matching approaches defining a color correspondence for each pixel in image pairs, in our method, the color correspondences are established on the blended layers respectively. The two processes of transparency optimization and disparity estimation boost each other, effectively reducing the possible

errors when they are performed separately. We show the disparity maps and the alpha matte computed using our approach in Figure 1.1 (d) and (e) respectively. The comparison of the disparities are illustrated in (f) and (g). Even from the transparency point of view, our method also outperform the previous natural image matting techniques. A detailed comparison will be given later in this paper.

Our method also contributes a nice implicit formulation of pixel occlusion. In conventional stereo matching, since each pixel has at most one disparity value, the occlusion needs to be modelled separately for pixels having no correspondences [20]. In our approach, any pixel in the layer of the scene behind the hairy objects can be partially occluded, entirely occluded, or unoccluded according to the degree of transparency, which can be naturally encoded using the alpha values without special treatment.

The rest of the dissertation is organized as follows. Chapter 2 reviews previous work on stereo matching and digital image matting. A brief introduction of the Expectation-Maximization algorithm is also given. We define our model and give notations in chapter 3. The initialization step of our system is explained in chapter 4, and the detailed optimization process is described in chapter 5. In chapter 6, we show our experimental results and compare with other state-of-art methods. This dissertation will be concluded in chapter 7.

□ **End of chapter.**

Chapter 2

Background Study

Summary

This chapter gives a brief survey of the state-of-art methods in stereo matching and digital image matting. An algorithm introduction of the Expectation-Maximization method is also included.

This work is related to the research on dense stereo matching and digital image matting. The Expectation-Maximization method is used for the optimization step.

2.1 Stereo matching

There have been many methods developed to solve the conventional stereo matching problem. Middlebury College has developed an official stereo vision research page [3] to publish sample stereo image pairs and rank the results from different methods.

And a survey on the two-frame stereo matching can be found in [17].

Markov Random Field (MRF) is widely used in stereo matching to constrain the energy minimization problem with the smoothness energy function [10, 22, 8, 20]. Most of these methods solve the MRF by either Belief Propagation (BP) [5] or Graph Cut [1]. In [10], a method related to expansion move algorithm is used to find the local minimum of an energy function. Sun *et al.*[22] introduced an MAP estimation on depth and pixel occlusion situations. The system is further refined in [20], where a symmetric system combining a soft constraint on segmentation is developed. [30] proposed a Hierarchical loopy Belief Propagation algorithm to refine the results on occluded and low-texture areas. [31] developed a new MAP formulation on both depth map and MRF parameters. Rather than user defining the parameters, their system introduced an EM framework to estimate the parameters and the depth map iteratively. Computed depth result on pixels with high confidence will have stronger influence on their neighbors. Graph Cut method is applied to compute the optimal value. [4] segments the two input frames into small patches. Graph Cut is also used to find the disparity and occlusions embedded in the patches with the symmetric mapping. [24] compares the performance of *Graph Cut* and *Belief Propagation* on a set of images, and concludes that, in general, the results produced by the two algorithms are comparable.

The above methods are not proposed to solve the stereo

matching problem with color blending because of the disparity ambiguities. Szeliski and Golland [23] first proposed to solve stereo matching with boundary opacity or matting effect. The visibility is computed through re-projection, where color and transparency refinement are formulated as a non-linear minimization problem. Wexler *et al.* [28] compute alpha mattes and estimate layers from multiple images with known background information. [26] estimates depth with the consideration of layer overlapping. It uses nested plane sweep with refinement from graph cut. The attenuation factors for color blending at reflecting areas are constant. In [7] [33], to get a better result for view synthesis, boundary matting along depth discontinuities are performed after the initial disparities computed. The mistakes due to color blending will not be further corrected. Besides, [32] computes the alpha contribution on overlapping regions among segments. A more accurate optical flow estimation can then be achieved.

2.2 Digital image matting

Natural image matting is to separate the blended pixels by computing the foreground, background and the alpha matte respectively given a natural input image. Using trimaps, Bayesian Matting [2] and Poisson Matting [19] estimate the foreground and background colors by collecting samples. Wang and Cohen [27] introduced an optimization approach based on Belief

Propagation to estimate the alpha matte without trimaps. In [11], Levin *et al.* proposed a closed form solution to solve the matting problem given the user input of a few strokes. The work is further improved in [12], where spectral analysis is performed to separate the image into components. Unsupervised matting results can be got by selecting the components group with the best matting cost. To enhance the performance of video matting, Joshi *et al.* [9] used an autofocus system to first determine pixel relationships among multi-images. Based on the variation of the related pixels, they form trimap and compute alpha mattes in real-time. Sun *et al.* [21] use a pair of flash/no-flash images to extract mattes. Because the relationship between the image pair brings more constraints into the system, their trimaps can be automatically generated and the results are further improved from previous methods. However, all these methods cannot be directly applied to stereo matching without the consideration of the correspondence of colors and alpha values in input images.

2.3 Expectation Maximization

EM is an iterative optimization method to estimate some unknown parameters Θ, given some known measurement data U and taking consideration on some hidden variables J. In regular EM, the target of optimization is always an MAP problem for Θ marginalizing over J:

$$\Theta^* = \arg\max_{\Theta} \sum_{J \in J^n} P(U, J, \Theta) \tag{2.1}$$

The EM algorithm alternates between estimating the unknowns Θ and the hidden variables J. Instead of estimating the exact value of J, the EM algorithm tries to estimate the distribution of J, for a best estimation of Θ under all possible condition. One of the earliest papers in EM is by Hartley in 1958 [6]. Wu [29] discussed the convergence properties of the EM algorithm. It indicates that if the likelihood function is unimodal and a certain differentiability condition is satisfied, then any EM sequence converges to the unique maximum likelihood estimate. Another insightful explanation of EM is in terms of lower-bound maximization [13] [16]. The explanation indicates that the E-step can be interpreted as constructing a local lower-bound to the posterior distribution, whereas the M-step optimizes the bound, thereby improving the estimate for the unknowns.

EM algorithm is widely used in computer science researches. The K-Means clustering problem and the formation of mixture models, especially the Gaussian Mixture Model (GMM) [15], are some well-known examples. Besides, EM is also used in some medical imaging related works, for instance, the reconstruction of emission tomography (ET) images [14].

□ **End of chapter.**

Chapter 3

Model Definition

Summary

This chapter gives out our novel model definition, which includes foreground and background color, alpha value and foreground and background disparities.

Conventional two-frame dense stereo matching approaches estimate depth value by estimating the correspondence of pixels in the input image pair. In this dissertation, we also use two images C^r and C^m in different viewing positions, and assume that the reference image C^r and the matching image C^m are rectified [25]. Conventionally, for a pixel (x, y) in C^r and its corresponding pixel (x', y') in C^m with disparity d, we have

$$x' = x + d, y' = y. \tag{3.1}$$

The stereo matching problem is formulated as the estimation of

disparity d using the color constancy on the matched pixels in a scene with Lambertian reflectance:

$$C^r(x, y) = C^m(x + d, y). \tag{3.2}$$

In our problem definition, to model the color blending between the objects with hairy boundaries and the scene behind, we assume that each input image contains a foreground object F in front of a background scene B, both having Lambertian reflectance. The pixels in the background can be unconcluded, partially occluded, or entirely occluded by F according to the degree of transparency. Applying the equation of alpha blending, the blended color in each pixel is formulated as

$$C^k(x, y) = \alpha^k(x, y)F^k(x, y) + (1 - \alpha^k(x, y))B^k(x, y), \tag{3.3}$$

where $k \in \{r, m\}$. Accordingly, in our stereo model, instead of defining a single disparity d for each pixel in the input images, we introduce disparities d^f and d^b for latent pixels in foreground F and background B respectively. This definition largely increases the flexibility of our method to model occlusions. Hence, for each latent foreground pixel $F^r(x, y)$ (or the background pixel $B^r(x, y)$) in C^r, applying d^f (or d^b), we can obtain a matched pixel $F^m(x, y)$ (or $B^m(x, y)$) in C^m, where

$$\begin{aligned} F^r(x, y) &= F^m(x + d^f, y), \\ B^r(x, y) &= B^m(x + d^b, y). \end{aligned} \tag{3.4}$$

Moreover, since there are measurable discontinuities in depth between the foreground objects and background scene. The occlusion between them can be nicely formulated using equation 3.3 according to the corresponding alpha values without another explicit occlusion detection process:

$$\begin{cases} \alpha(x,y) = 1 & B(x,y) \text{ is entirely occluded,} \\ 0 < \alpha(x,y) < 1 & B(x,y) \text{ is partially occluded,} \\ \alpha(x,y) = 0 & B(x,y) \text{ is not occluded.} \end{cases} \tag{3.5}$$

Table 3.1: Mean Square Deviation on Alpha in two frames

	Include $\alpha \in \{0,1\}$	
	Mean Square Deviation	**Pixel Number**
Bear	0.00038764	957600
Box	0.00037437	720766
	Not include $\alpha \in \{0,1\}$	
	Mean Square Deviation	**Pixel Number**
Bear	0.013788	26922
Box	0.03524	4136

Using a narrow-band camera setup, most transparency of foreground may remain invariant fore corresponding foreground pixels. As shown in fig 3.3, only transparency of the pixels on the boundary of foreground object may have a little change via images. Fortunately, the differences of alpha between the corresponding pixel pairs of two frames are also small. Fig 3.2 shown a hairy bear toy and a box with smooth boundary. Comparing

the alpha matte got from the left and right view, the fig 3.2(c) and (f) show the difference in their alpha matte distribution, which are very similar. We manually align the two views and compute the mean square deviation (MSD) of the alpha values. As shown in table 3.1, the MSD of the alpha value is very low. Even we only consider the pixels on the boundary, i.e, discarding those totally opaque($\alpha = 1$) or transparent($\alpha = 0$) pixels, the MSD of the alpha value is still within a low level. Our model is also validated using our real data experiment. As shown in figure 6.8, the MSD error between the alpha matte extracted from our model and ground truth is only 0.00062. So, to take consider on this cue, we apply soft constraints, which will be discussed in chapter 5.2, on transparency of foreground corresponding pixel pairs. Specifically, if a foreground pixel (x, y) in C^r is matched to $(x + d^f, y)$ in C^m, we have

$$\alpha^r(x, y) \approx \alpha^m(x + d^f, y). \tag{3.6}$$

Based on this soft constraint, we can build relationships on transparency value between the image pair. Since the digital matting problem is always over-constrain, the relationship can further contributes in the reduction of ambiguities, which leads to an improvement on the alpha matte result.

In the rest of the dissertation, for simplicity, we use subscripts p, $p + d^f$, and $p + d^b$ to denote pixel in (x, y), $(x + d^f, y)$, and $(x + d^b, y)$ respectively. Substituting Equation 3.2 and 3.4 into

Equation 3.3, we obtain the following two equations for each corresponding pixel pair in the input images:

$$\begin{cases} C_p^r &= \alpha_p F_p^r + (1-\alpha_p) B_p^r, \\ C_{p+d^f}^m &= \alpha_{p+d^f}^m F_{p+d^f}^m + (1-\alpha_{p+d^f}^m) B_{p+d^f}^m. \end{cases} \tag{3.7}$$

We show one example in Fig. 3.1 where two corresponding foreground pixels are blended by different background pixels due to the disparity differences. In Equation 3.7, there are unknowns F^r, F^m, B^r, B^m, α^r and α^m to be estimated given input C^r and C^m. F^r and F^m are corresponding foreground pixels. Without loss of generality, we optimize F^r in our method. F^m , as a complement in stereo configuration, is computed by mapping the foreground pixels in C^r to C^m using the computed disparities. We estimate α^r, α^m, B^r and B^m separately in a symmetric manner. It guarantees that the unmatched background pixels due to the occlusions are appropriately handled, which in turn improves the estimation of the disparities and foreground pixels.

In what follows, without special annotation, we will use F to denote F^r. Thus, substituting Equation 3.4 into Equation 3.7, $C_{p+d^f}^m$ can be rewritten as

$$\begin{aligned} C_{p+d^f}^m &= \alpha_{p+d^f}^m F_{p+d^f}^m + (1-\alpha_{p+d^f}^m) B_{p+d^f}^m \\ &= \alpha_{p+d^f}^m F_p + (1-\alpha_{p+d^f}^m) B_{p+d^f}^m \end{aligned} \tag{3.8}$$

□ **End of chapter.**

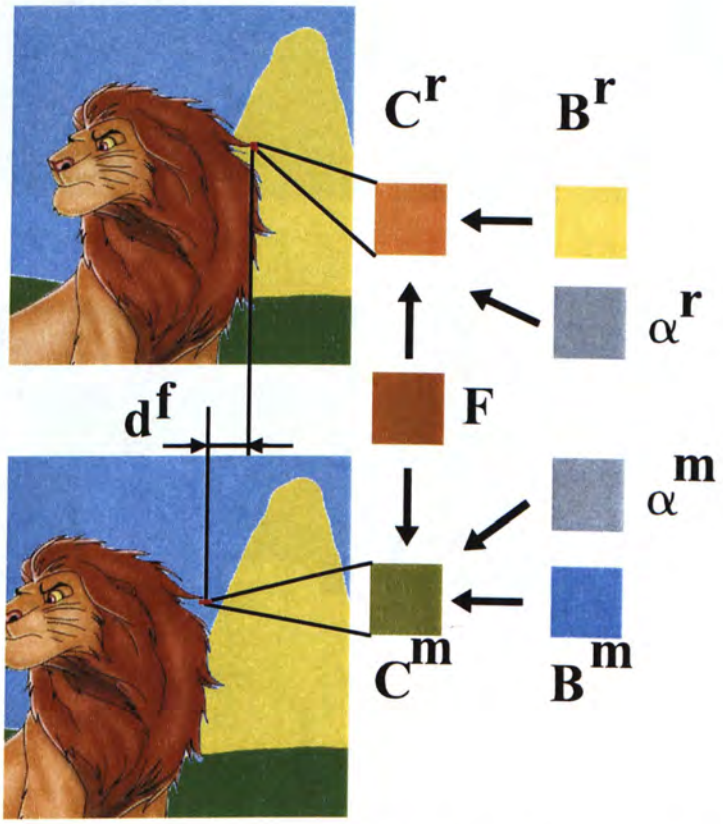

Figure 3.1: Color constancy on blended pixels. Given the input stereo image pair as shown, the semitransparent pixels C^r and C^m in the hair are blended by the foreground and the background. Since C^r and C^m are matched in foreground layer with disparity d^f, they have similar foreground color F and alpha value α^r & α^m. However, the partially occluded background pixels are different as shown in B^r and B^m.

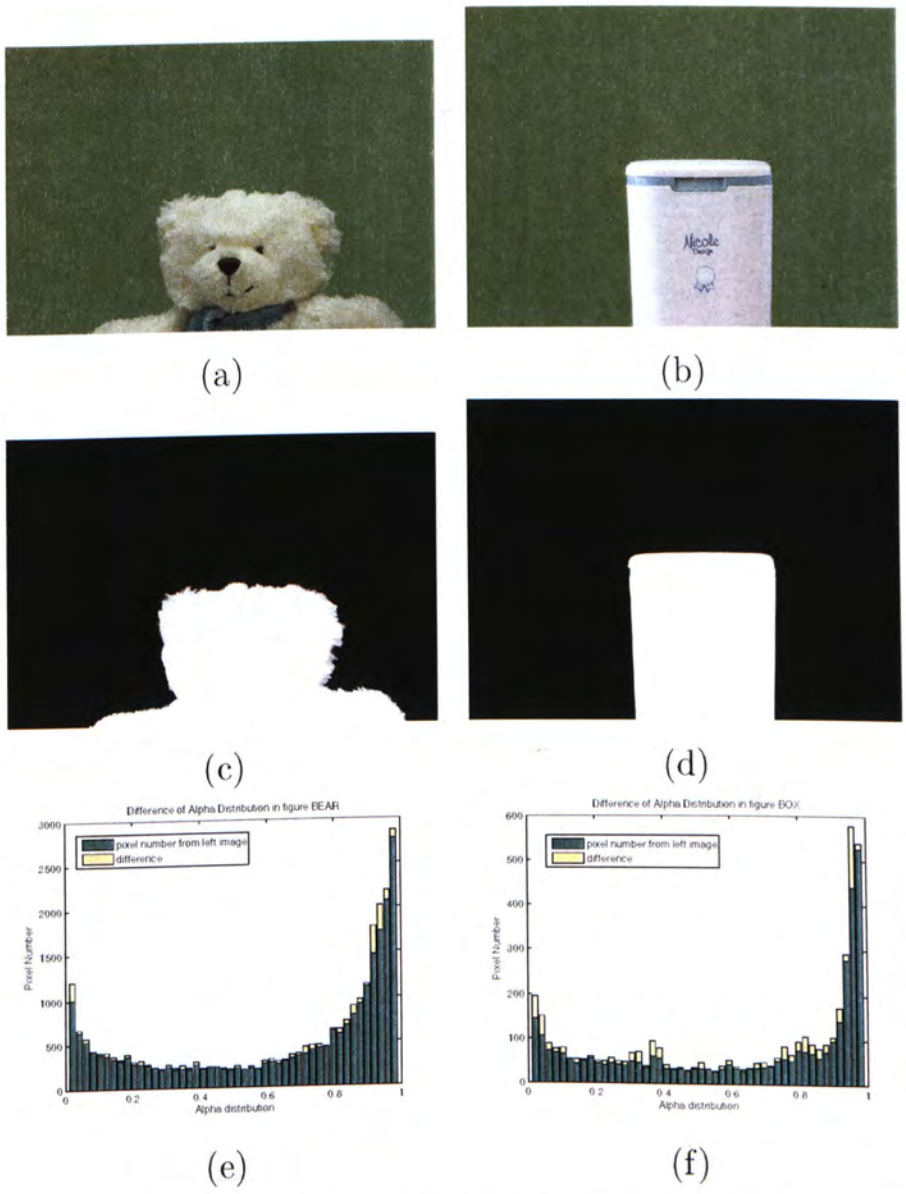

Figure 3.2: Alpha distribution. (a),(b) Left view of a typical hairy toy bear and a pink box. (c),(d) The alpha matte of (a) and (b). (e),(f) The difference of alpha distribution between the left and right image of (a) and (b).

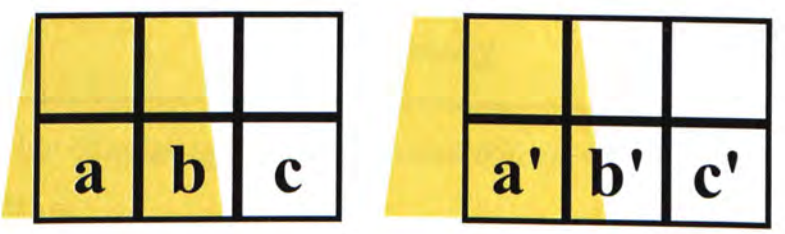

Figure 3.3: Different cases of alphas. a,b,c are pixels in image C^r and a',b',c' are their corresponding foreground pixels in image C^m. Due to the different position of the yellow foreground object under narrow-band camera assumption, the transparency of the pixels on the boundary,i.e., pixel b and b', may have a little change. However, most pixels will remain opaque (pixel a and a') or totally transparent (pixel c and c').

Chapter 4

Initialization

Summary

This chapter specifies the initialization step of our algorithm.

4.1 Initializing disparity

We initialize a single disparity d_p for each pixel p in images C^r and C^m using the previous stereo matching method [20]. However, some disparities are mistakenly computed due to the lack of consideration on color blending. Similar to [30], we form the confidence map, as shown in figure 4.2 (a), by computing the color-weighted difference for each pixel in C^r and C^m. For each pixel p, we first define a $N \times N$ window $N(p)$ centered at p. Typically, N is set to be 3 or 5. The weight of pixel $s \in N(p)$

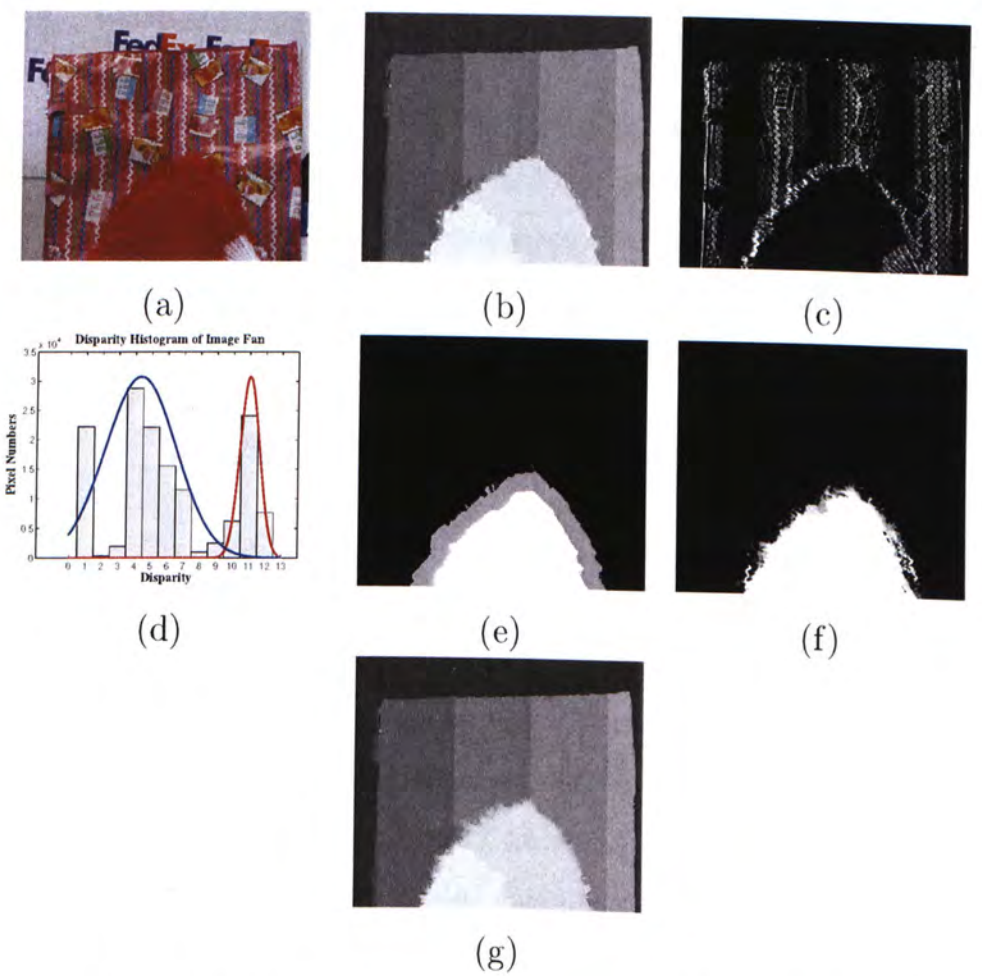

Figure 4.1: Work flow. (a) The input reference image. (b) The initial disparity map computed using [20]. (c) The computed confidence map. 'Reliable' pixels are marked darker. (d) The initially computed disparity histogram. The two fitted Gaussians are also shown. (e) Initial trimap computed. (f) Initial alpha values computed using the trimap in (e). (g) Final computed disparity map using our iterative optimization method.

is defined on both color and spacial difference:

$$w_s^p = exp(-\lambda_c \|C_s - C_p\|^2 - \lambda_s((x_s - x_p)^2 + (y_s - y_p)^2)) \tag{4.1}$$

Denote the pixels in $N(p)$ are $s_1, s_2, ..., s_{N^2}$, defined by their position. Similarly, $s'_1, s'_2, ..., s'_{N^2}$ are the corresponding pixels in the window centering at the matching pixel $p' = p + d^f$ in C^m. So the color-weighted difference for p and p' is defined as:

$$D(p, d^f) = \frac{\sum_{i=1}^{N^2} w_{s_i}^p w_{s'_i}^{p'} \|C_{s_i}^r - C_{s'_i}^m\|^2}{\sum_{i=1}^{N^2} w_{s_i}^p w_{s'_i}^{p'}} \tag{4.2}$$

Pixels with large color-weighted difference here are more likely in the color-blending region or with mistakenly computed disparities. So they are given a low confidence on the reliability of their disparities computed initially. We set a threshold T_D here to pick out those pixels and mark them as 'unreliable'.

Notice we require two initial disparities for each pixel for foreground and background respectively. With a set of more reliable disparities here, we then compute the histogram of the disparities. Since we assume that there's a distance gap between the background and the foreground, it is possible to partition the histogram into two disjoint segments. In extreme cases, for instance, all the foreground colors blending with background objects, most pixels will be labelled as 'unreliable' at first and

the initial distribution of the disparities will be totally unreliable. Fortunately, these kind of cases seldom happen in our real world. In most cases, we can assume that we have enough reliable disparity samples for statistics. Thus, cues on the disparities of the foreground objects and background scenes can be approximately gained from the initial histogram. For robustness, we fit the histogram into a two-component Gaussian mixture model. The parameters of the two Gaussians for foreground and background are denoted as $\{\overline{d^f}, \sigma_{d^f}\}$ and $\{\overline{d^b}, \sigma_{d^b}\}$ respectively. They also will be used later to form likelihoods on foreground and background disparities. One example is shown in Figure 4.1 (b). Then we use the Bayes classifier to partition the histogram

$$\begin{cases} d \text{ is in foreground} & N(h(d); \overline{d^f}, \sigma_{d^f}) \geq N(h(d); \overline{d^b}, \sigma_{d^b}) \\ d \text{ is in background} & N(h(d); \overline{d^f}, \sigma_{d^f}) < N(h(d); \overline{d^b}, \sigma_{d^b}) \end{cases}$$

where $h(d)$ the value of the dth bin in the histogram. For each pixel p, if the initialized d_p is classified as the foreground disparity d_p^f, then we set d_p^b to be the background disparity Gaussian mean $\overline{d^b}$, and vice versa:

$$d_p^f = \begin{cases} d_p & d_p \text{ is in foreground} \\ \overline{d^f} & d_p \text{ is in background} \end{cases} \tag{4.3}$$

$$d_p^b = \begin{cases} \overline{d^b} & d_p \text{ is in foreground} \\ d_p & d_p \text{ is in background} \end{cases} \tag{4.4}$$

4.2 Initializing alpha matte

We use Bayesian Matting [2] method to solve the matting problem initially on both images. However, this method requires a trimap to indicate whether one pixel in the input images is definitely foreground ($\alpha = 1$), definitely background ($\alpha = 0$), or unknown. We generate trimaps for each image in following steps.

Firstly, note that Equation 4.3 and 4.4 produce a binary segmentation in input images according to whether $d_p = d_p^f$ or $d_p = d_p^b$. The disparity of the pixels around the segmentation boundaries are obviously unreliable since these pixels are more likely to be mixtures of foreground and background. We then automatically select all these boundary pixels, and dilate them by 2 to 15 pixels to form an initial 'unknown' region in the trimap. All other pixels are then marked as 'foreground' or 'background' due to their initial disparities.

Secondly, we take consider on the pixels marked as 'unreliable' at the previous step. Since the color blending situation caused between background objects or between foreground objects may also be included in the 'unreliable' pixel set, we discard those pixels which have no connection with the 'unknown' region and set the rest into 'unknown'. Two initial trimaps on C^r and C^m, as shown in Figure 4.2, are thus created.

Based the trimaps, the foreground $F^{(0)}$, background $B^{(0)}$, and alpha matte $\alpha^{(0)}$ are automatically computed using Bayesian

(a) (b) (c)

Figure 4.2: The formation of trimap. (a) The confidence map based on initial disparity computed. (b) Build trimap based on confidence map and disparity map. Pixels along the segmentation boundaries of the foreground and background are marked in orange. The 'unreliable' pixels are marked in red and blue. Note the blue pixels are those caused by the self occlusion by foreground object or background scene. They will be discarded from the final trimap. (c) The final trimap.

Matting in the two input images. Of course, since the initial matting is performed separately in two images, there are inevitable errors, as shown in Figure 4.1(d).

□ **End of chapter.**

Chapter 5

Optimization

Summary

This chapter gives out our optimization step on the MAP problem using the EM algorithm. The E step for the expectation of probabilities on foreground and background disparities and the M step for the optimal foreground and background color as well as alpha matte are iteratively performed. The final result is computed on a MRF based on the optimized parameters.

In this chapter, we describe our optimization method to solve the fractional stereo matching problem formulated in Equation 3.7 and 3.8.

Given the observation $U = \{C^r, C^m\}$, we separate the unknowns into a parameter set $\Theta = \{F, B^r, B^m, \alpha^r, \alpha^m\}$ and hidden data $J = \{d^f, d^b\}$. In this section, we aim at estimating the

parameters using Expectation-Maximization

$$\begin{aligned}\Theta^* &= \arg\max_{\Theta} \log P(U, \Theta) \\ &= \arg\max_{\Theta} \log \sum_{J \in J^n} P(U, J, \Theta), \end{aligned} \tag{5.1}$$

where J^n is the domain of J. After we have obtained the optimized parameters, we compute an optimal J combining the spatial smoothness constraint.

5.1 Expectation Step

In iteration $n+1$, given the estimated $\Theta^{(n)}$, for each pixel p, we compute in this step the expectation of $P_p(d^f = d_1, d^b = d_2|\Theta^{(n)}, U)$ where $d_1, d_2 \in \{0, 1, ..., L_d\}$. L_d is the maximum disparity. Since d^f and d^b are statistically independent, we have

$$\begin{aligned} & E(P_p(d^f = d_1, d^b = d_2|\Theta^{(n)}, U)) \\ = \; & E(P_p(d^f = d_1|\Theta^{(n)}, U) P_p(d^b = d_2|\Theta^{(n)}, U)) \\ = \; & E(P_p(d^f = d_1|\Theta^{(n)}, U)) E(P_p(d^b = d_2|\Theta^{(n)}, U)). \end{aligned} \tag{5.2}$$

In what follows, we describe the expectation computation on d^f and d^b respectively.

5.1.1 Computing $E((P_p(d^f = d_1|\Theta^{(n)}, U))$.

The conditional probability d^f is formulated using Bayes' theorem:

$$\begin{aligned} & P_p(d^f|\Theta^{(n)}, U) \\ \propto\ & P_p(d^f|U, B^{r(n)}, B^{m(n)}, \alpha^{r(n)}, \alpha^{m(n)}) \\ \propto\ & P_p(B^{r(n)}, B^{m(n)}, \alpha^{r(n)}, \alpha^{m(n)}|d^f, U) \cdot P_p(d^f|U). \end{aligned} \quad (5.3)$$

According to Equation 3.7, ideally, the corresponding foreground pixels in two images should have the same pixel color:

$$\alpha^m_{p+d^f}(C^r_p - (1-\alpha^r_p)B^r_p) = \alpha^r_p(C^m_{p+d^f} - (1-\alpha^m_{p+d^f})B^m_{p+d^f}). \quad (5.4)$$

Thus, we define the probability $P_p(B^{r(n)}, B^{m(n)}, \alpha^{(n)}|\ d^f, U)$ as the color similarity of corresponding foreground pixels in both input images

$$\begin{aligned} & P_p(B^{r(n)}, B^{m(n)}, \alpha^{(n)}|d^f, U) \\ =\ & exp(-\beta_f \|\alpha^m_{p+d^f}(C^r_p - (1-\alpha^{r(n)}_p)B^{r(n)}_p) \\ -\ & \alpha^r_p(C^m_{p+d^f} - (1-\alpha^{m(n)}_{p+d^f})B^{m(n)}_{p+d^f})\|^2), \end{aligned} \quad (5.5)$$

where β_f is a weight.

$P_p(d^f|U)$ models prior probability of d_f from initial input images. In the initialization step to be discussed in section 4, we model all disparities from the foreground and background pixels using two Gaussian distributions respectively. Thus we

formulate the probability

$$P_p(d^f|U) = N(d^f; \overline{d^f}, \sigma_{d^f}) = \frac{1}{\sqrt{2}\sigma_{d^f}} \exp(-\frac{(d^f - \overline{d^f})^2}{2\sigma_{d^f}}). \quad (5.6)$$

where $\overline{d^f}$ and σ_{d^f} are the mean and variance of the foreground disparity Gaussian introduced before. The expectation for the disparity value of each foreground pixel p can be written as

$$E(P_p(d^f = d_1|\Theta^{(n)}, U)) = \frac{P_p(d^f = d_1|\Theta^{(n)}, U)}{\sum_{d_i} P_p(d^f = d_i|\Theta^{(n)}, U)} \quad (5.7)$$

Since we have only a few levels for d_i, the computation of the above formula is easy.

5.1.2 Computing $E((P_p(d^b = d_2|\Theta^{(n)}, U))$.

For the background disparities, we can formulate the probability as

$$\begin{aligned} & P_p(d^b|\Theta^{(n)}, U) \\ \propto\ & P_p(d^b|U, B^{r(n)}, B^{m(n)}, \alpha^{r(n)}, \alpha^{m(n)}) \\ \propto\ & P_p(B^{r(n)}, B^{m(n)}, \alpha^{r(n)}, \alpha^{m(n)}|d^b, U) \cdot P_p(d^b|U) \end{aligned} \quad (5.8)$$

where the matching probability $P_p(B^{r(n)}, B^{m(n)}, \alpha^{r(n)}, \alpha^{m(n)}|d^b, U)$ is different from the foreground counterpart in Equation 5.5 due to the possibility of been occluded for any background pixels. Thus, we define the probability on background color matching

adapting to the alpha values:

$$\begin{aligned} & P_p(B^{r(n)}, B^{m(n)}, \alpha^{r(n)}, \alpha^{m(n)} | d^b, U) \\ = \; & exp(-\beta_b(1-\alpha_p^{r(n)})[(1-\alpha_{p+d^b}^{m(n)})\|B_p^{r(n)} - B_{p+d^b}^{m(n)}\|^2 + \alpha_{p+d^b}^{m(n)} P]) \end{aligned} \tag{5.9}$$

where β_b is a weight similar to that defined in Equation 5.5, $\alpha_{p+d^b}^m$ in C^m is the corresponding alpha value to α_p^r in C^r for the same background pixel, and P is set to give penalty when the value of $\alpha_{p+d^b}^{m(n)}$ is close to 1, i.e., the background pixel $p + d^b$ is largely occluded by the foreground in image C^m.

To understand the definition of Equation 5.9, let us analyze two extreme situations. On one extreme, if α_p^r and $\alpha_{p+d^b}^m$ both approach 0, it means that both the corresponding background pixels B_p^r and $B_{p+d^b}^{m(n)}$ are not occluded. Their color differences, with a large probability, measure if the two pixels are matched. On the other extreme, if either α_p^r or $\alpha_{p+d^b}^m$ approaches 1, one or both background pixels are occluded. Thus, the color difference $||B_p^r - B_{p+d^b}^m||$ is not reliable.

The definition of $P_p(d^b|U)$ is defined in a way similar to Equation 5.6 using initially estimated Gaussian distribution described in section 4:

$$P_p(d^b|U) = N(d^b; \overline{d^b}, \sigma_{d^b}) = \frac{1}{\sqrt{2}\sigma_{d^b}} \exp(-\frac{(d^b - \overline{d^b})^2}{2\sigma_{d^b}}). \tag{5.10}$$

Integrating the above two probability definition, the expectation

on d^b can be computed as

$$E(P_p(d^b = d_2|\Theta^{(n)}, U)) = \frac{P_p(d^b = d_2|\Theta^{(n)}, U)}{\sum_{d_i} P_p(d^b = d_i|\Theta^{(n)}, U)} \quad (5.11)$$

5.2 Maximization Step

After the expectation computation, we maximize the expected complete-data log-likelihood w.r.t. J given the observation U:

$$\begin{aligned}
\Theta^{(n+1)} &= \arg\max_{\Theta} \sum_{J \in J^n} P(J|\Theta^{(n)}, U) \log P(\Theta|J, U) \\
&= \arg\max_{\Theta} \sum_{J \in J^n} P(J|\Theta^{(n)}, U) \log P(J, U|\Theta) P(\Theta) \\
&= \arg\max_{\Theta} \sum_{J \in J^n} P(J|\Theta^{(n)}, U)\{L(J, U|\Theta) + L(\Theta)\} \quad (5.12)
\end{aligned}$$

where $L(\cdot)$ is the log likelihood $L(\cdot) = logP(\cdot)$. $L(\Theta)$ is further expanded to

$$L(\Theta) \propto L(\alpha^r) + L(\alpha^m) + L(F) + L(B^r) + L(B^m). \quad (5.13)$$

It is noted that $P(J|\Theta^{(n)}, U)$ is already computed in the Expectation step. Using Equations 3.4, 3.7 and 3.8, we define

$$
\begin{aligned}
L(J,U|\Theta) = \; - \; & \sum_{p\in C^r}\{(\|\alpha_p^r F_p + (1-\alpha_p^r)B_p^r - C_p^r\|^2 \\
& + \; \|\alpha_{p+d^f}^m F_p + (1-\alpha_{p+d^f}^m)B_{p+d^f}^m - C_{p+d^f}^m\|^2 \\
& + \; \gamma_B\|B_p^r - B_{p+d^b}^m\|^2)/2\sigma_C^2 \\
& + \; \gamma_\alpha|\alpha_p^r - \alpha_{p+d^f}^m|^2\exp(-\|C_p^r - C_{p+d^f}^m\|^2/2\sigma_C^2)\},
\end{aligned}
\tag{5.14}
$$

where σ_C is the standard deviation of the Gaussian probability distribution centered at C [2]. The first two term of the definition measures the error due to the color composition equations 3.7. The third term measures the color distance of two corresponding background pixels. The last term of the definition is the soft constraint we applied on transparency. We use the difference of the origin pixel values, C_p^r and $C_{p+d^f}^m$, to be the reference here. Once the difference is large, the corresponding foreground pixels are more unlikely to have the same transparency here. Here γ_B and γ_α are nonnegative weights for adjustment in the log likelihood.

Similar to the methods proposed to solve the natural image matting problem [2, 27], we estimate the foreground color, alpha value, and background color likelihoods for each pixel by first collecting samples from the neighboring pixels. Then we model these samples using single Gaussian or Gaussian mixtures for background and foreground respectively. In what follows,

for simplicity, we describe our method using a single Gaussian model. The formulation and optimization using Gaussian mixtures are similar.

For each pixel p, denoting the constructed Gaussian models for foreground color as $\{\overline{F_p}, \Sigma_{F_p}^{-1}\}$, we obtain

$$L(F) = \sum_p L(F_p) = \sum_p -(F_p - \overline{F_p})^T \Sigma_{F_p}^{-1}(F_p - \overline{F_p})/2 \quad (5.15)$$

Similarly, we obtain the definitions of $L(B^r)$, $L(B^m)$ and $L(\alpha)$:

$$L(B^k) = \sum_p L(B_p^k) = \sum_p -(B_p^k - \overline{B_p^k})^T \Sigma_{B_p^k}^{-1}(B_p^k - \overline{B_p^k})/2, k \in \{r, m\}, \quad (5.16)$$

$$L(\alpha^k) = \sum_p L(\alpha_p^k) = \sum_p -\frac{(\alpha_p^k - \overline{\alpha_p^k})^2}{2\sigma_{\alpha_p^k}^2}, k \in \{r, m\}. \quad (5.17)$$

For better explanation, we define

$$f(\Theta) = -\sum_{J \in J^n} P(J|\Theta^{(n)}, U)\{L(J, U|\Theta) + L(\Theta)\}. \quad (5.18)$$

Given all above definitions of probabilities, our target is to find a best unknown set Θ to minimize $f(\Theta)$. One observation here is that in f, each F, B or α value may only have cross-terms with the unknowns on the same scanline. So we can optimize them scanline by scanline. Like the traditional digital matting methods [2] [21], we optimize α and $\{F, B\}$ iteratively.

5.2.1 Optimize α, given $\{F, B\}$ fixed

When $\{F, B\}$ are fixed, given the width of image W, for each scanline y, the problem turns to be finding

$$X^* = \arg\min_X f(X), \quad X = [\alpha^r_{1,y}, \alpha^r_{2,y}, \ldots, \alpha^r_{W,y}, \alpha^m_{1,y}, \alpha^m_{2,y}, \ldots, \alpha^m_{W,y}]^T \tag{5.19}$$

Note here the unknowns α^r and α^m are sparsely nested, so it's still a multi-variable non-linear optimization problem. The Hessian Matrix of function $f(X)$ can be written as:

$$\nabla^2 f(X) = Z = \begin{bmatrix} Z_{00} & Z_{01} \\ Z_{10} & Z_{11} \end{bmatrix} \tag{5.20}$$

Here Z_{ij} are $W \times W$ matrices. Denote $p^f_{p,k} = P_p(d^f = k|\Theta^{(n)}, U)$, $p^b_{p,k} = P_p(d^b = k|\Theta^{(n)}, U)$. We define

$$p^f_{p,k} = p^b_{p,k} = 0, \quad \text{when} \quad k \notin [0, L_d] \tag{5.21}$$

And matrix Z_{ij} can be written as:

$$\begin{aligned}
Z_{00}(i,j) &= \delta_{ij}\sum_{d^f} p^f_{p,d^f}\{1/\sigma^2_{\alpha_p} + (F_p - B^r_p)^T(F_p - B^r_p)/\sigma^2_C \\
&\quad +2\gamma_\alpha \exp(-\|C^r_p - C^m_{p+d^f}\|^2/2\sigma^2_C)\}, p = \{i,y\} \qquad (5.22)\\
Z_{11}(i,j) &= \delta_{ij}\sum_{d^f} p^f_{p-d^f,d^f}\{1/\sigma^2_{\alpha^m_p} + (F_{p-d^f} - B^m_p)^T(F_{p-d^f} - B^m_p)/\sigma^2_C \\
&\quad +2\gamma_\alpha \exp(-\|C^m_p - C^r_{p-d^f}\|^2/2\sigma^2_C)\}, \quad p = \{i,y\} \qquad (5.23)\\
Z_{01}(i,j) &= -2p^f_{p,d^f}\gamma_\alpha \exp(-\|C^r_p - C^m_{p+d^f}\|^2/2\sigma^2_C)\}, \\
&\quad d^f = d^b = j - i, p = \{i,y\}, \qquad (5.24)\\
Z_{10}(i,j) &= -2p^f_{p-d^f,d^f}\gamma_\alpha \exp(-\|C^m_p - C^r_{p-d^f}\|^2/2\sigma^2_C)\}, \\
&\quad d^f = d^b = j - i, p = \{i,y\}, \qquad (5.25)
\end{aligned}$$

where δ_{ij} is the Kronecker delta. It can be seen that Z is a strictly symmetric diagonally dominant real matrix with positive diagonal entries, which is proved to be positive definite. So the function $f(X)$ is convex. We can take partial derivatives on $f(X)$ with respect to X, and set them to be zero to compute the α for global minimum. For the scanline y, the equation can be written as:

$$\begin{bmatrix} G_{00} & G_{01} \\ G_{10} & G_{11} \end{bmatrix} X = \begin{bmatrix} H_0 \\ H_1 \end{bmatrix}, \qquad (5.26)$$

Here G_{00}, G_{01}, G_{10} and G_{11} are $W \times W$ matrix:

$$\begin{aligned}
G_{00}(i,j) &= \delta_{ij} \sum_{d^f} p^f_{p,d^f} \{1/\sigma^2_{\alpha_p} + (F_p - B^r_p)^T (F_p - B^r_p)/\sigma^2_C \\
&\quad + 2\gamma_\alpha \exp(-\|C^r_p - C^m_{p+d^f}\|^2/2\sigma^2_C)\}, p = \{i, y\} \qquad (5.27) \\
G_{11}(i,j) &= \delta_{ij} \sum_{d^f} p^f_{p-d^f,d^f} \{1/\sigma^2_{\alpha^m_p} + (F_{p-d^f} - B^m_p)^T (F_{p-d^f} - B^m_p)/\sigma^2_C \\
&\quad + 2\gamma_\alpha \exp(-\|C^m_p - C^r_{p-d^f}\|^2/2\sigma^2_C)\}, \quad p = \{i, y\} \qquad (5.28) \\
G_{01}(i,j) &= -2p^f_{p,d^f} \gamma_\alpha \exp(-\|C^r_p - C^m_{p+d^f}\|^2/2\sigma^2_C)\}, \\
&\quad d^f = d^b = j - i, p = \{i, y\}, \qquad (5.29) \\
G_{10}(i,j) &= -2p^f_{p-d^f,d^f} \gamma_\alpha \exp(-\|C^m_p - C^r_{p-d^f}\|^2/2\sigma^2_C)\}, \\
&\quad d^f = d^b = j - i, p = \{i, y\}. \qquad (5.30)
\end{aligned}$$

And H_0, H_1 are $W \times 1$ vectors:

$$\begin{cases}
H_0(i) &= \overline{\alpha^r_p}/\sigma^2_{\alpha^r_p} + (F_p - B^r_p)^T (C^r_p - B^r_p)/\sigma^2_C \\
H_1(i) &= \sum_{d^f} p^f_{p-d^f,d^f} \{\overline{\alpha^m_p}/\sigma^2_{\alpha^m_p} + (F_{p-d^f} - B^m_p)^T (C^m_p - B^m_p)/\sigma^2_C\}
\end{cases}, \quad p = \{i, y\} \qquad (5.31)$$

α^r and α^m can be directly computed by solving the linear equation set 5.26.

5.2.2 Optimize $\{F, B\}$, given α fixed

Similarly, the pixel color $\{F, B^r, B^m\}$ can also be computed by scanline. We define

$$\mathbf{V} = [\mathbf{F}; \mathbf{B}_y^r; \mathbf{B}_y^m]; \tag{5.32}$$

where

$$\mathbf{F}_y = [F_{1,y}; F_{2,y}; \ldots; F_{W,y}] \tag{5.33}$$

$$\mathbf{B}_y^r = [B_{1,y}^r; B_{2,y}^r; \ldots; B_{W,y}^r] \tag{5.34}$$

$$\mathbf{B}_y^m = [B_{1,y}^m; B_{2,y}^m; \ldots; B_{W,y}^m] \tag{5.35}$$

$F_{i,j}$, $B_{i,j}^r$ and $B_{i,j}^m$ are 3×1 vectors represents the foreground and background color at pixel $\{i, j\}$. Then the Hessian Matrix of $f(\mathbf{V})$ is also can be proved to be positive definite. So we take partial derivatives on f with respect to $\mathbf{V}$, and also set them to be zero to compute $\{F, B^r, B^m\}$. For each scanline y, we have:

$$\begin{bmatrix} A_{00} & A_{01} & A_{02} \\ A_{10} & A_{11} & A_{12} \\ A_{20} & A_{21} & A_{22} \end{bmatrix} \mathbf{V} = \begin{bmatrix} M_0 \\ M_1 \\ M_2 \end{bmatrix}, \tag{5.36}$$

And A_{uv} can be written as:

$$A_{uv} = \begin{bmatrix} a_{uv}^{1,1} & a_{uv}^{1,2} & \cdots & a_{uv}^{1,W} \\ a_{uv}^{2,1} & a_{uv}^{2,2} & \cdots & a_{uv}^{2,W} \\ \vdots & \vdots & \vdots & \vdots \\ a_{uv}^{W,1} & a_{uv}^{W,2} & \cdots & a_{uv}^{W,W} \end{bmatrix}, u, v \in \{0, 1, 2\} \quad (5.37)$$

$a_{uv}^{i,j}$s are 3×3 matrices:

$$\begin{aligned}
a_{00}^{i,j} &= \delta_{ij} \sum_{d^f} p_{p,d^f}^{f} ((\{\alpha_p^r\}^2 + \{\alpha_{p+d^f}^m\}^2) I/\sigma_C^2 + \Sigma_{F_p}^{-1}), && p = \{i, y\} \\
a_{01}^{i,j} &= \delta_{ij} \alpha_p^r (1 - \alpha_p^r) I/\sigma_C^2, && p = \{i, y\} \\
a_{02}^{i,j} &= p_{p,d^f}^{f} \alpha_{p+d^f}^m (1 - \alpha_{p+d^f}^m) I/\sigma_C^2, \quad d^f = j - i, && p = \{i, y\} \\
a_{10}^{i,j} &= \delta_{ij} \alpha_p^r (1 - \alpha_p^r) I/\sigma_C^2, && p = \{i, y\} \\
a_{11}^{i,j} &= \delta_{ij} (\{(1 - \alpha_p^r)^2 + \gamma_B)\} I/\sigma_C^2 + \Sigma_{B_p^r}^{-1}), && p = \{i, y\} \\
a_{12}^{i,j} &= -p_{p,d^b}^{b} \gamma_B I/\sigma_C^2, \quad d^b = j - i, && p = \{i, y\} \\
a_{20}^{i,j} &= p_{p-d^f,d^f}^{f} \alpha_p^m (1 - \alpha_p^m) I/\sigma_C^2, \quad d^f = j - i, && p = \{i, y\} \\
a_{21}^{i,j} &= -p_{p-d^b,d^b}^{b} \gamma_B I/\sigma_C^2, \quad d^b = j - i, && p = \{i, y\} \\
a_{22}^{i,j} &= \delta_{ij} (\{(1 - \alpha_p^m)^2 + \gamma_B\} I/\sigma_C^2 + \Sigma_{B_p^m}^{-1}), && p = \{i, y\}
\end{aligned} \quad (5.38)$$

And M_0, M_1 and M_2 can be written as:

$$M_u = [M_u^1; M_u^2; \ldots; M_u^W], u \in \{0, 1, 2\} \quad (5.39)$$

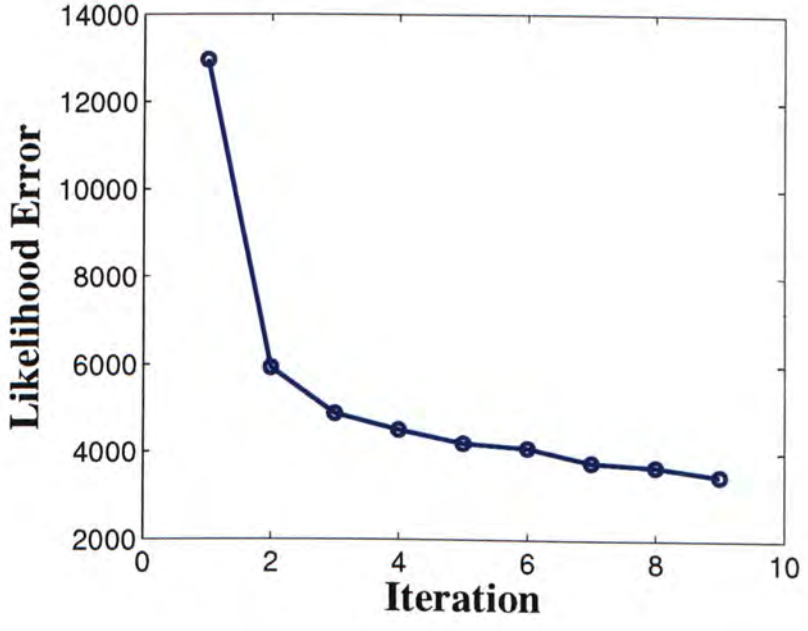

Figure 5.1: The likelihood error is decreasing in iterations for the example shown in Figure 1.1.

where M_u^i are 3×1 vectors:

$$\begin{aligned} M_0^i &= \sum_{df} p_{p,df}^f (\{\alpha_p^r C_p^r + \alpha_{p+df}^m C_{p+df}^m\}/\sigma_C^2 + \Sigma_{F_p}^{-1}\overline{F_p}), \quad p = \{i, y\} \\ M_1^i &= (1-\alpha_p^r) C_p^r/\sigma_C^2 + \Sigma_{B_p^r}^{-1}\overline{B_p^r}, \qquad p = \{i, y\} \\ M_2^i &= (1-\alpha_p^m) C_p^m/\sigma_C^2 + \Sigma_{B_p^m}^{-1}\overline{B_p^m}, \qquad p = \{i, y\} \end{aligned} \tag{5.40}$$

Here I is a 3×3 identity matrix. Then given the linear equation set 5.36, the F, B^r and B^m can be solved directly.

Using the estimated $\alpha^{(n)}$,$F^{(n)}$ and $B^{(n)}$ to be good initializations, the above optimization processes on $\{\alpha^r, \alpha^m\}$ and $\{F, B^r, B^m\}$ are iteratively performed until convergence.

We plot in Figure 5.1 the likelihood error,

$$e = -\sum_{J \in J^n} P(J|\Theta^{(n)}, U) \log \frac{P(\Theta^{(n+1)}|J, U)}{P(J|\Theta^{(n)}, U)} \tag{5.41}$$

for the example shown in Figure 1.1. It is decreasing in iterations.

5.3 Computing Final Disparities

After the optimization using the EM described, we obtain the estimated parameters Θ^* and a probability distribution of the hidden data d^f and d^b. Similar to the classical stereo matching algorithm, we form a MRF on images and define an energy, which integrates a data term and a smoothness term [20], to minimize:

$$E(d^k|U, \Theta^*) = E_d(d^k|U, \Theta^*) + E_s(d^k), \tag{5.42}$$

where $k \in \{f, b\}$. $E_s(d^k)$ is the smoothness term defined similar to that in [20],

$$E_s(d^k) = \sum_{s,t,t \in N(s)} \gamma_d |d_s^k - d_t^k|^2. \tag{5.43}$$

And $E_d(d^k|U, \Theta^*)$ is the data term

$$E_d(d^k|U, \Theta^*) = \sum_p -\log P(d_p^k = d^k|\Theta^*, U). \tag{5.44}$$

We use Belief Propagation to minimize the energy and compute the final optimal disparities. A summary of our whole algorithm process is described in algorithm 1.

Algorithm 1 Whole work flow

1. Compute initial disparity map using Sun *et al.* method [20].
2. Compute initial trimap.
   (a) Compute Confidence map and marked out 'unreliable' pixels.
   (b) Form two-component Gaussian mixture for foreground and background disparities based on reliable pixels. Initialize the foreground and background disparities.
   (c) Select pixels on the boundary of foreground and background pixels. Take consideration on 'unreliable' pixels and then form the trimap.
3. Perform Bayesian Matting [2]. Get $F^{(0)}$, $B^{(0)}$,$\alpha^{(0)}$.
4. Perform EM algorithm iteratively till converge for Θ and J.
   (a) E-step. Compute expectation of d^f and d^b given fixed $F^{(n)}$,$B^{(n)}$ and $\alpha^{(n)}$.
   (b) M-step. Compute α, F and B iteratively to maximize the log-likelihood given computed expectation J.
5. Form a MRF and integrate the smoothness term. Use BP to minimize the energy and compute the final disparities.

□ **End of chapter.**

Chapter 6

Experiment Results

Summary

This chapter gives shows our results on challenging examples. Some quantitative comparisons between our method and other state-of-art methods are also shown.

We have shown one difficult example in Figure 1.1. Since in our approach, each pixel has at most two disparities, only for visualizing the hairy object boundary, we construct the *blended disparity map* similar to the color blending

$$d_p^{show} = \alpha_p d_p^f + (1 - \alpha_p) d_p^b, \tag{6.1}$$

which has already been used in Figure 1.1 (d) and 4.1 (e). Due to the complexity of the EM algorithm, it take a Pentium(R) 3.20GHZ CPU, 1GB RAM computer more than 1 hour to compute the result.

Figure 6.1 shows another difficult example where two stereo

images contain a toy bear with long hair. (a) and (b) are two input images. (c) is the disparity result using the method in [20], which obviously causes errors around the object boundary. (d) and (e) are our *blended disparity map* and alpha matte through optimization. The complex alpha structure is preserved.

Our approach can also be applied to the traditional stereo image pairs to improve the object details. We show the "Tsukuba" example in figure 6.6. In our experiments, the lamp is automatically segmented as the foreground objects since it has largest disparities. We show our optimized alpha matte and extracted foreground in (b) and (c) respectively. Note that the boundary of the extracted lamp is smooth and natural. Using the optimized alpha matte, we compute the disparities and compare them with those generated in [4] and [20] in 6.6 (d-f) using the following formula to reduce the two disparities for each pixel to one:

$$d_p^{refine} = \begin{cases} d_p^f & \alpha_p \geq 0.5 \\ d_p^b & \alpha_p < 0.5 \end{cases} \tag{6.2}$$

Obviously, our result has clearer boundary of the lamp. In fig 6.6(h), we show our improvement along the boundary of the lamp comparing with Sun *et al.*'s method [20]. The evaluation result from the Middlebury Stereo Vision Page, figure 6.5, has also shown the advantage of our method, especially at pixels near depth discontinuities. Even though the evaluation discards a set

of occluded pixels when compare to the ground truth, which are precisely recovered via our method, our result still outperforms among all recent methods.

Table 6.1: Alpha matte difference comparison of synthetic examples

	Bayesian [2]	**Wang & Cohen [27]**
Bear	0.02424	0.02801
Girl	0.01555	0.01552
	Levin *et al.* [11]	**Our Method**
Bear	0.00516	0.00349
Girl	0.00345	0.00258

Besides, our method can also produce better matting results comparing to previous single natural image matting methods. In figure 6.2 and 6.3, we first compare our method with two state-of-art natural image matting methods [2, 27, 11] in the two synthetic 'bear' and 'girl' examples. The MSD errors between the alpha matte got from different methods and the ground truth is listed in Table 6.1, and plotted in figure 6.4.

In figure 6.7, we compare our result with the matting methods [27, 11] on the difficult "fan" example. The background has complex patterns and similar colors as the foreground, which make the foreground and background color estimation unstable. In (b) and (c), it is observable that the background patterns are mistakenly estimated as the foreground. Our result in (d) has less errors in the alpha matte thanks to the stereo configuration and the joint optimization.

Another example is shown in figure 6.8. In (e), the result got using Sun *et al.* [20] contains visible errors on the boundary region. Our result (f) successfully fixed those parts due to the consideration of alpha matte. The ground truth of the alpha matte is got using the blue screen method [18]. The MSD error between our result and the ground truth is only 0.00062218.

□ **End of chapter.**

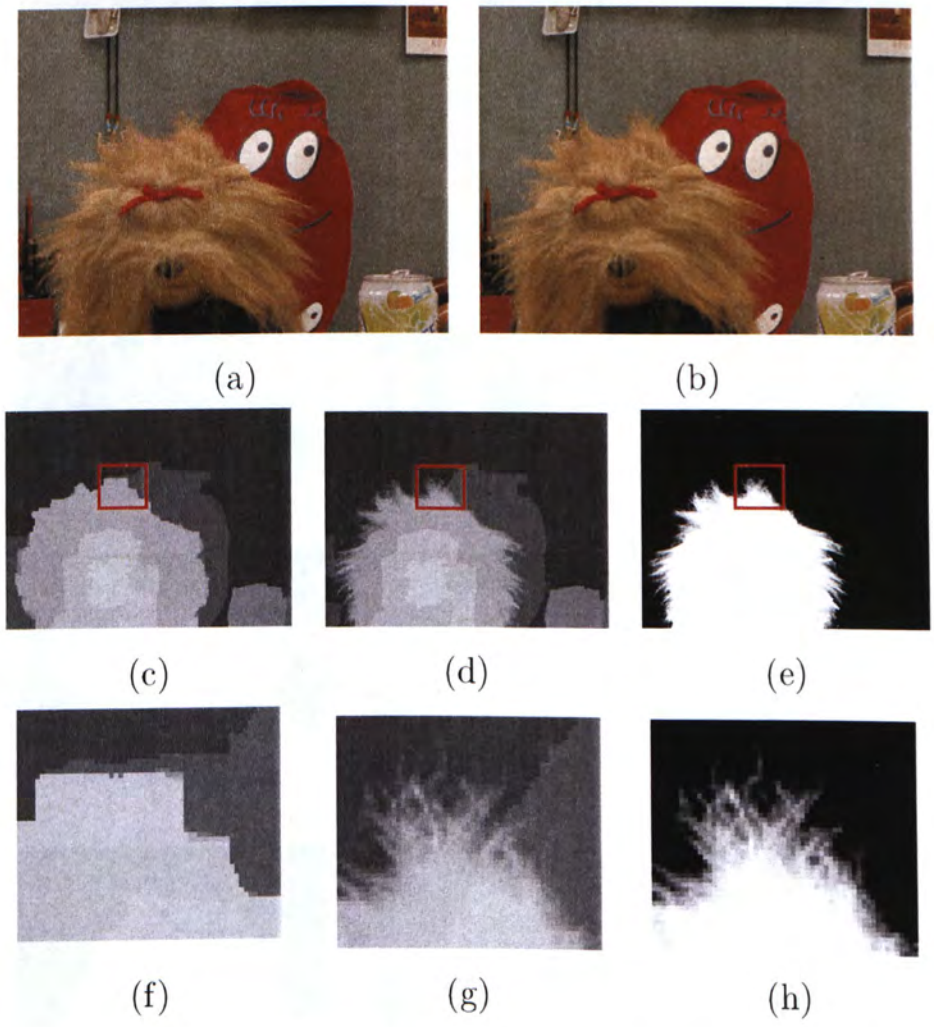

Figure 6.1: Bear example. (a) and (b) The input stereo images. (c) Stereo matching result using Sun's method [20]. (d) The *blended disparity map* computed from our method. The structures are well preserved. (e) The alpha matte computed from our approach. (f)-(h) Magnified regions in (c), (d), and (e).

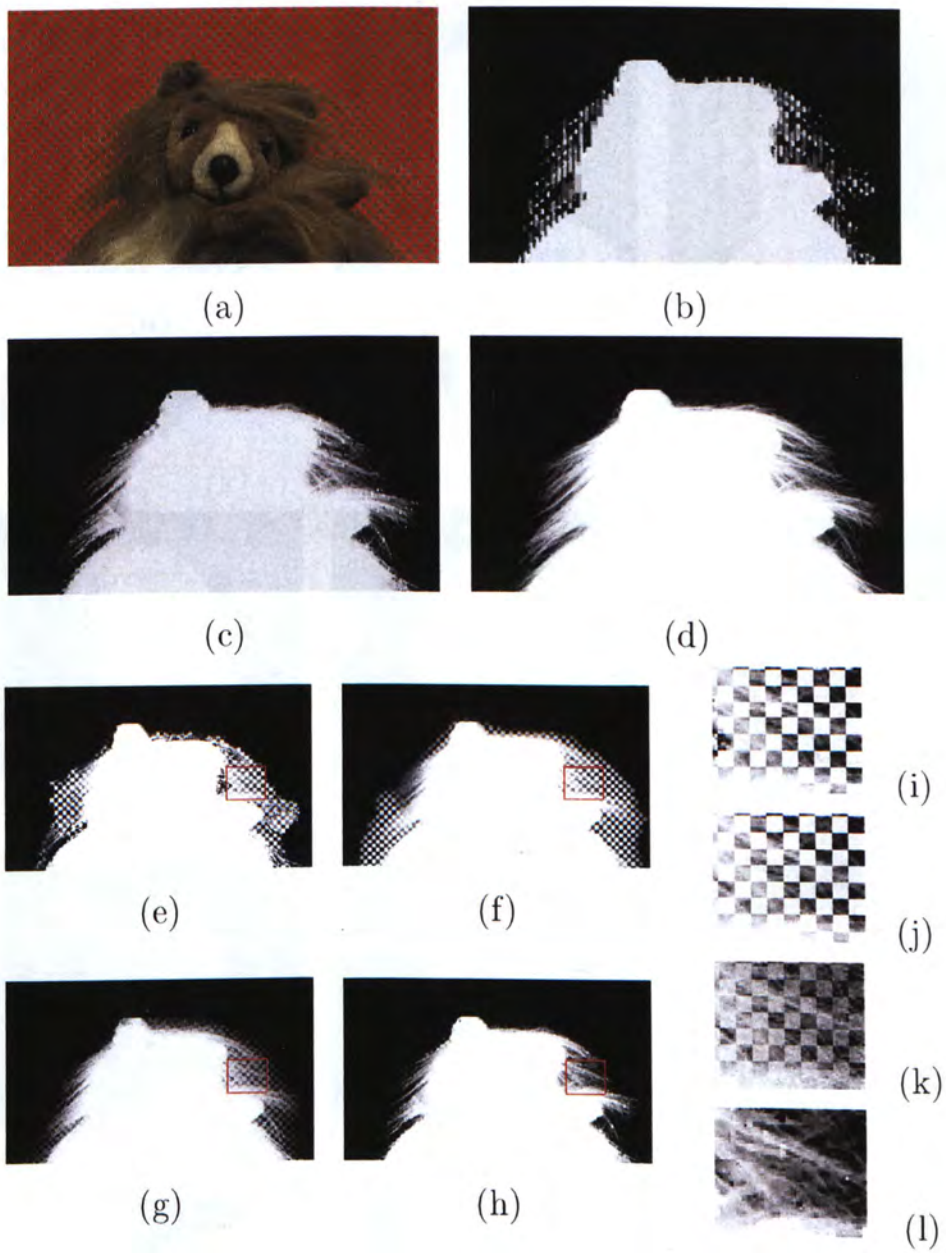

Figure 6.2: The synthetic example 'bear'. (a) Input left view. (b) Initial disparity got from [20]. (c) Our disparity map combining with alpha matte. (d) Ground truth of alpha matte. (e) Alpha matte got from Bayesian matting [2]. (f) Alpha matte got from Wang and Cohen method [27]. (g) Alpha matte got from Levin *et al.* method [11]. (h) Our alpha matte result. (i)-(l) Side-by-side comparison on the magnified region.

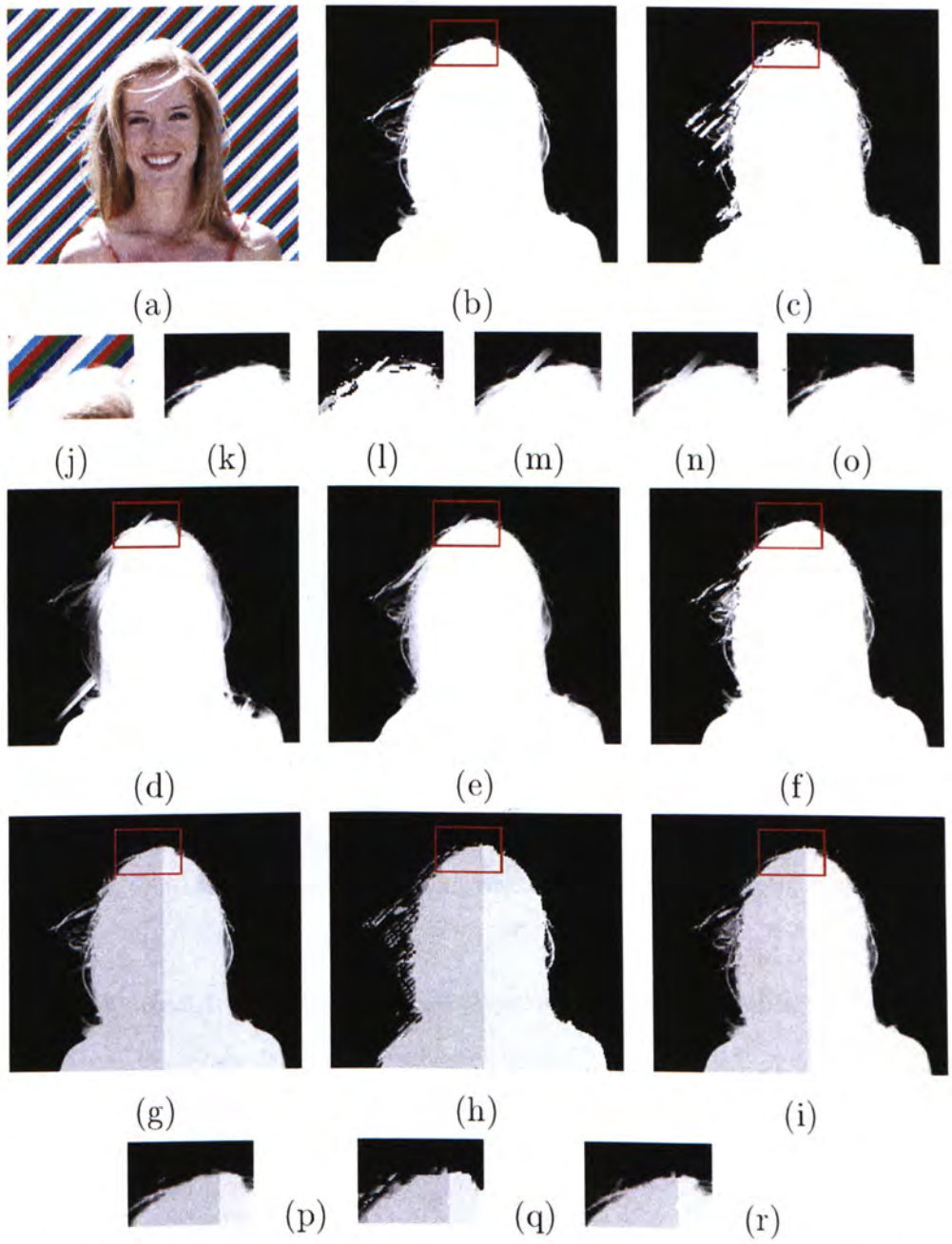

Figure 6.3: The synthetic example 'girl'. (a) Input left view. (b) Ground truth of alpha matte. (c) Alpha matte got from Bayesian matting [2]. (d) Alpha matte got from Wang and Cohen method [27]. (e) Alpha matte got from Levin *et al.* method [11]. (f) Our alpha matte result. (g) Ground truth of disparity. (h) Initial disparity got from [20]. (i) Our disparity map combining with alpha matte. (j)-(r) Magnified regions in (a)-(i), respectively.

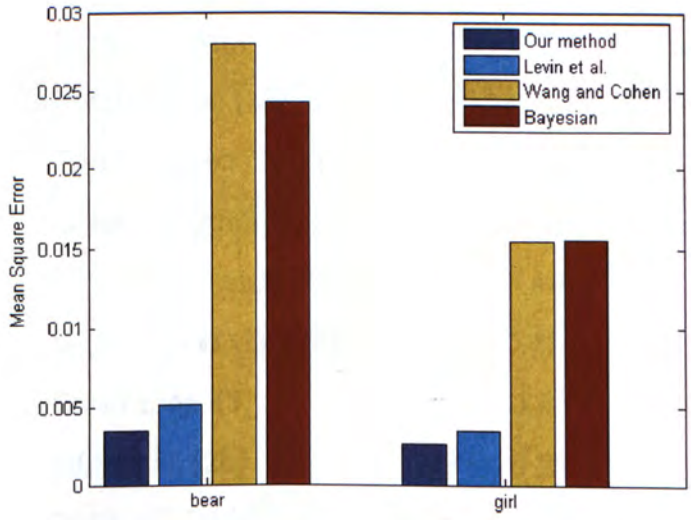

Figure 6.4: Quantitative evaluation of alpha matte results.

Algorithm	Tsukuba		
	all	untex.	disc.
Sym.BP+occl. [27]	0.97 3	0.28 4	5.45 3
Patch-based [36]	0.88 2	**0.19** 1	4.95 2
Segm.-based GC [23]	1.23 7	0.29 6	6.94 8
Graph+segm. [29]	1.39 11	0.28 4	7.17 10
GC + mean shift [34]	1.13 4	0.48 10	6.38 5
Segm.+glob.vis. [25]	1.30 9	0.48 11	7.50 13
Belief prop. [3]	1.15 5	0.42 8	6.31 4
GC+occl. [2b]	1.19 6	0.23 2	6.71 6
OUR METHOD	0.88 1	0.25 3	**4.92** 1

Figure 6.5: The quantitive comparison result for the stereo image pair "Tsukuba". Data got from the Middlebury Stereo Vision Page: "http://cat.middlebury.edu/stereo/".

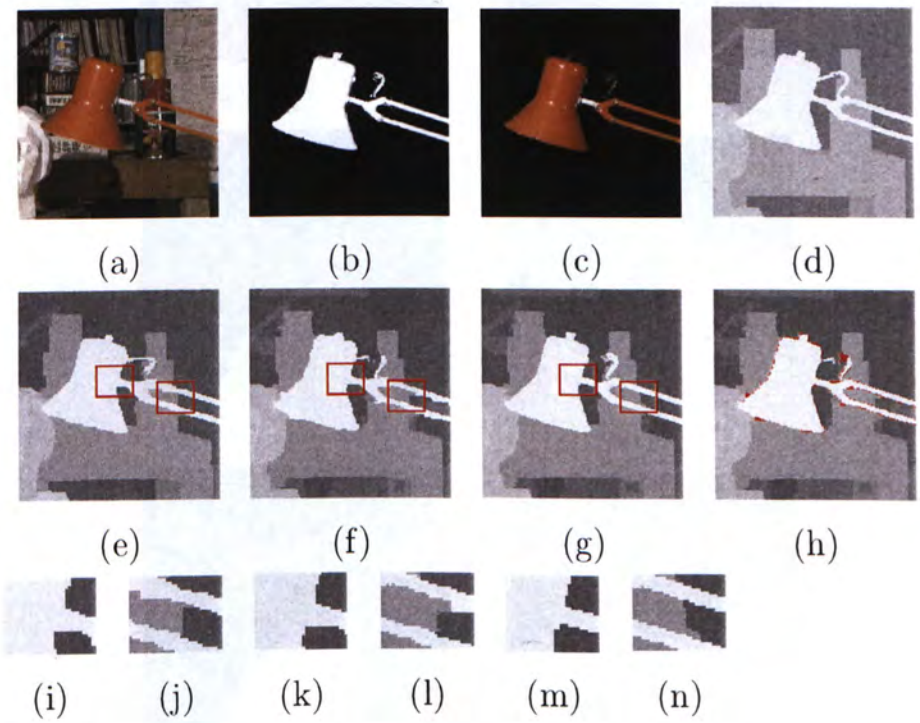

Figure 6.6: The lamp from the stereo image pair "Tsukuba". (a) Input reference image. (b) The alpha matte of the foreground lamp computed from our method. The boundary is natural. (c) The extracted foreground. (d) Ground truth. (e) Result from the patch-based method [4]. (f) Result of symmetric stereo matching [20]. (g) Our optimized disparity map. The lamp boundary has large improvement comparing to (d) and (e). (h) Our improvement along the lamp boundary from (f). Pixels marked in red are the errors fixed by our method. (i)-(n) Side-by-side comparison on the magnified regions.

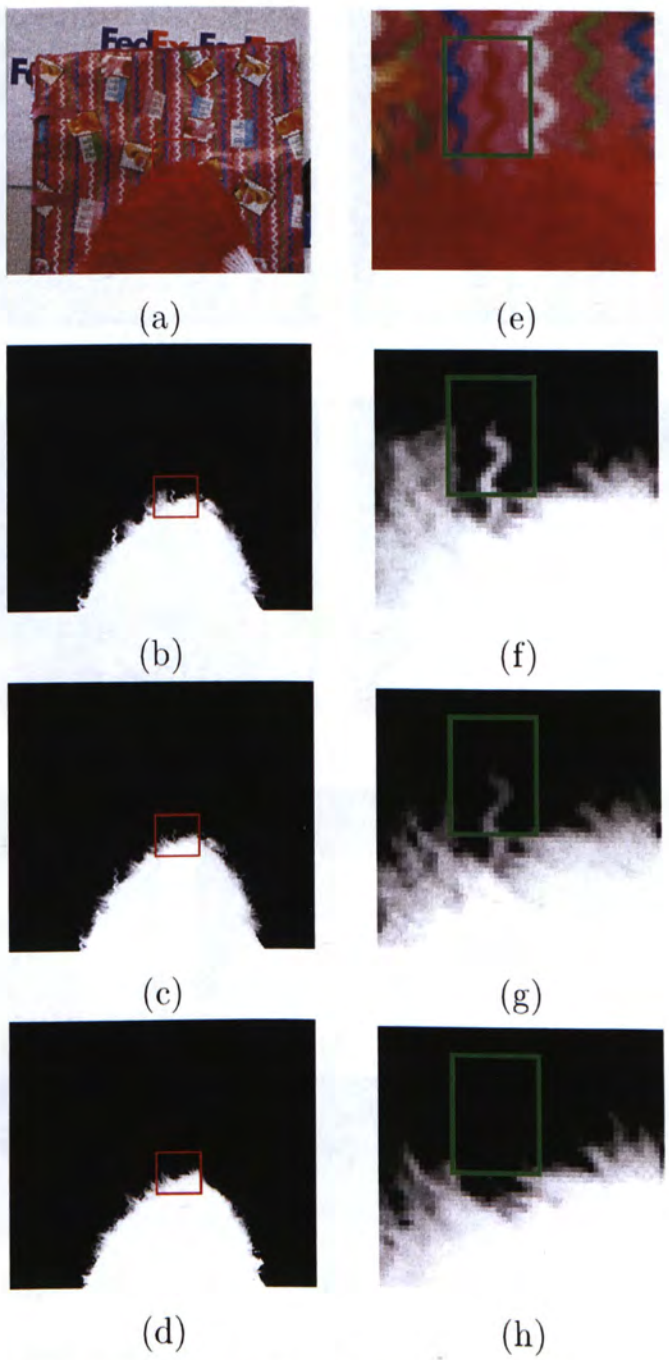

Figure 6.7: Comparison of the alpha matte. (a) Input reference image. The background and foreground have similar colors. The patterns of the background are also complex. (b) Result from the method in [27]. (c) Results from the method in [11]. (d) Our method is automatic, and does not require any user input. (e)-(h) The magnified regions for comparison. Notice that, within the green rectangle, result (f) and (g) mistakenly take the background pattern into foreground while our method produces a satisfactory alpha matte.

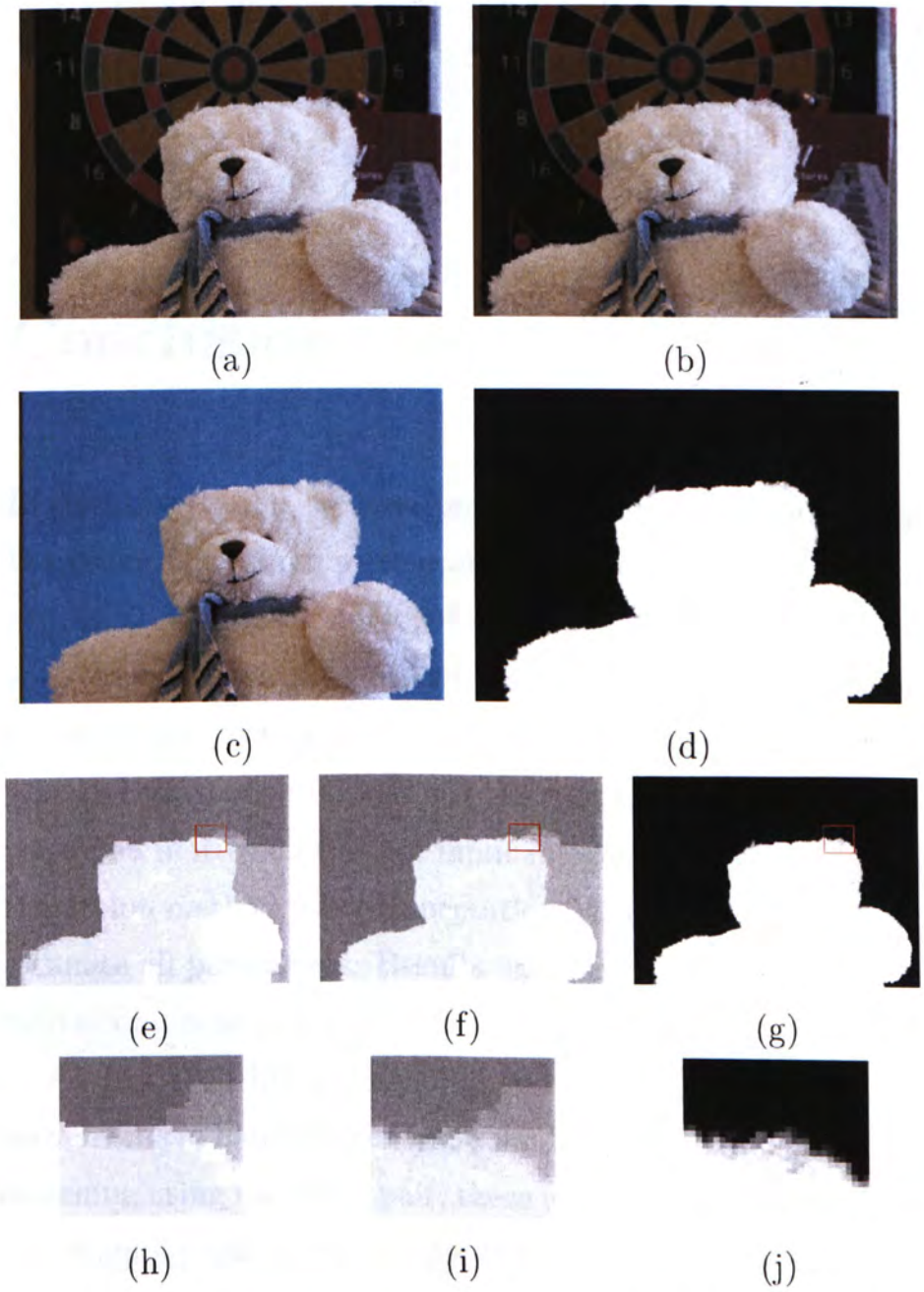

Figure 6.8: Bear example. (a) and (b) The input stereo images. (c) Foreground object with a blue screen background. (d) Ground truth of alpha matte got from blue screen method [18]. (e) Stereo matching result using Sun's method [20]. (f) The *blended disparity map* computed from our method. The structures are well preserved. (g) The alpha matte computed from our approach. (h)-(j) Magnified regions in (e), (f), and (g).

Chapter 7

Conclusion

In this dissertation, we have proposed a novel approach to solve the stereo matching problem on objects with fractional boundary using two-frame narrow-band stereo images. Each pixel, with the definition of the layer blending, is assumed to be blended by two latent pixels with different disparities. We have defined a probabilistic model constraining the colors, disparities, as well as the alpha mattes on the two input images, and designed an optimization method using Expectation-Maximization to robustly estimate all parameters. Results on a set of images and quantitative comparisons are given out.

As described before, although our method has achieved improvements in handling general boundary transparencies in stereo matching using an image pair, there are still problems unsolved. For example, our method only considers two layers, i.e., background and foreground. Then our method can only focus on refining the boundary of the foreground object, but not all the

objects in the image. We expect that if more stereo images are given, our model can be extended to handle more layers. Besides, our recent method need a textureful background scene to achieve a good separation of foreground and background. Using the method shown in [12] to separate the image into matting components may also be a good choice. These will be considered in future.

□ **End of chapter.**

Bibliography

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[2] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *Proceedings of IEEE CVPR 2001*, volume 2, pages 264–271. IEEE Computer Society, December 2001.

[3] M. College. Middlebury college stereo vision research page. *http://cat.middlebury.edu/stereo/*.

[4] Y. Deng, Q. Yang, X. Lin, and X. Tang. A symmetric patch-based correspondence model for occlusion handling. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1316–1322, Washington, DC, USA, 2005. IEEE Computer Society.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *CVPR*, 01:261–268, 2004.

[6] H. Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194, 1958.

[7] S. W. Hasinoff, S. B. Kang, and R. Szeliski. Boundary matting for view synthesis. In *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 11*, page 170, Washington, DC, USA, 2004. IEEE Computer Society.

[8] L. Hong and G. Chen. Segment-based stereo matching using graph cuts. *CVPR*, 1:74–81, 2004.

[9] 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.

[10] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. *ICCV*, 02:508, 2001.

[11] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 61–68, Washington, DC, USA, 2006. IEEE Computer Society.

[12] A. Levin, A. Rav-Acha, and D. Lischinski. Spectral matting. *CVPR*, 2007.

[13] T. Minka. Expectation-maximization as lower bound maximization. *Tutorial published on the web at http://www-white.media.mit.edu/ tpminka/ papers/em.html.*

[14] T. Minka. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, Nov 1996.

[15] T. M. Mitchell. *Machine Learning*, 1997.

[16] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, pages 355–368, 1999.

[17] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[18] A. R. Smith and J. F. Blinn. Blue screen matting. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 259–268, New York, NY, USA, 1996. ACM Press.

[19] J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum. Poisson matting. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*, pages 315–321, New York, NY, USA, 2004. ACM Press.

[20] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*

- *Volume 2*, pages 399–406, Washington, DC, USA, 2005. IEEE Computer Society.

[21] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. Flash matting. *ACM Transactions on Graphics*, 25(3):772–778, 2006.

[22] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 25(7):787–800, 2003.

[23] R. Szeliski and P. Golland. Stereo matching with transparency and matting. *International Journal of Computer Vision*, 32(1):45–61, 1999.

[24] M. F. Tappen and W. T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 900, Washington, DC, USA, 2003. IEEE Computer Society.

[25] E. Trucco, A. Fusiello, and A. Verri. Rectification with unconstrained stereo geometry. *BMVC*, 1997.

[26] Y. Tsin, S. B. Kang, and R. Szeliski. Stereo matching with reflections and translucency. *CVPR*, 01:702, 2003.

[27] J. Wang and M. F. Cohen. An iterative optimization approach for unified image segmentation and matting. In *ICCV '05: Proceedings of the Tenth IEEE International*

Conference on Computer Vision, pages 936–943, Washington, DC, USA, 2005. IEEE Computer Society.

[28] Y. Wexler, A. W. Fitzgibbon, and A. Zisserman. Bayesian estimation of layers from multiple images. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part III*, pages 487–501, London, UK, 2002. Springer-Verlag.

[29] C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[30] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2347–2354, Washington, DC, USA, 2006. IEEE Computer Society.

[31] L. Zhang and S. M. Seitz. Parameter estimation for mrf stereo. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 288–295, Washington, DC, USA, 2005. IEEE Computer Society.

[32] C. L. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on*

Computer Vision, pages 1308–1315, Washington, DC, USA, 2005. IEEE Computer Society.

[33] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 23(3):600–608, 2004.