Bayesian Estimation of Layers from Multiple Images

Y. Wexler, A. Fitzgibbon and A. Zisserman

Robotics Research Group
Department of Engineering Science
University of Oxford
Oxford, OX1 3PJ
{wexler,awf,az}@robots.ox.ac.uk

Abstract. When estimating foreground and background layers (or equivalently an *alpha matte*), it is often the case that pixel measurements contain mixed colours which are a combination of foreground and background. Object boundaries, especially at thin sub-pixel structures like hair, pose a serious problem.
In this paper we present a multiple view algorithm for computing the alpha matte. Using a Bayesian framework, we model each pixel as a combined sample from the foreground and background and compute a MAP estimate to factor the two. The novelties in this work include the incorporation of three different types of priors for enhancing the results in problematic scenes. The priors used are inequality constraints on colour and alpha values, spatial continuity, and the probability distribution of alpha values.
The combination of these priors result in accurate and visually satisfying estimates. We demonstrate the method on real image sequences with varying degrees of geometric and photometric complexity. The output enables virtual objects to be added between the foreground and background layers, and we give examples of this augmentation to the original sequences.

1 Introduction

An important requirement for the film and broadcast industry is the insertion of virtual objects into real footage, and in turn this requires generating the correct occlusions of inserted objects by real objects in the sequence. To make an accurate composite, foreground and background layers must be estimated, and in particular, pixels which contain mixed foreground and background colour (such as at object boundaries) must be accurately modeled. This problem is difficult, as the boundary may be very complex (it is commonly required to deal with hair, for example).

This paper is concerned with the automatic extraction of occlusion masks (or "alpha mattes") which record at each pixel the proportions in which foreground and background combine to generate the image. The contribution is in the use of relative motion of the foreground and background to estimate a transparency value (or 'alpha') at each pixel. As the camera moves, a given foreground pixel will sweep over several background pixels, and so a constraint is available on the foreground's colour. The problem is poorly constrained, as foreground and background colours may be similar, so it is important to incorporate prior knowledge in order to regularize the solution. We develop a Bayesian framework (Section 3.2) which allows priors to be progressively added to

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2352, pp. 487–501, 2002.

(a)

(b)

Fig. 1. Input sequences. (a) Three frames from sequence "Railings". We wish to extract the foreground railings and an associated alpha mask for each image in the sequence in order to replace the background building. (b) Three out of ten images used in the "Monkey" sequence. (c) shows the composition of the monkey between the railing and the background building.

the information obtained from the relative motion. These priors include bounds on the image and alpha values, learned distributions of alpha values, and spatial consistency taking account of image edges (Sections 3.4 – 3.6). We demonstrate that this combination of priors facilitates the extraction of accurate alpha mattes from sequences where the foreground and background objects are approximately planar, and compare with ground truth (Section 5).

2 Background

The use of mattes to selectively obscure elements in film predates the digital computer [1]. In the 40's, complex shots combining real actors and virtual backgrounds were executed by painting the virtual components on glass. The digital equivalent is to

associate a value α with every pixel, which represents the opacity of the pixel. These values are stored in the image's *alpha channel*. Foreground and background are combined via the *compositing equation*

$$C = \alpha F + (1 - \alpha)B,$$

where F is the RGB vector at the foreground pixel, and α is the opacity associated with that foreground value, B is the background RGB and the *composite* image pixel is C.

To recover alpha mattes for real objects, two main strategies exist in the literature, one based on segmentation, the other on motion. Segmentation based techniques use a property of the background and foreground regions—for example statistics on their colour distributions—to assign each pixel a likelihood of being foreground or background. The simplest such schemes are based on shooting against a blue or green background. Then the colour distribution for the background is very well specified, and the problem is simplified. However, it is not always possible to shoot against a blue screen, so newer techniques use more sophisticated models to segment against real, *a priori* unknown, backgrounds—with impressive results [5, 10]. However, computing such segmentations on the hundreds of frames required in a typical cinema special effects shot requires significant human effort.

In a moving shot, on the other hand, more constraints are available from the relative motion of foreground and background, so it is in principle possible to automatically extract the matte. We now review previous work on motion-based methods.

In mosaicking papers which estimate the background layer by the dominant motion (e.g. see [7]) it is usual to register the background layers and superimpose the foreground (e.g. for a synopsis mosaic). Here it will be more convenient to stabilize the foreground, and have the background sweep past. The key constraint which motion adds is that foreground pixels are seen against different background pixels. Indeed we can see motion as being analogous to having several images of the same (foreground) object against different coloured backgrounds, a situation which was shown by Smith and Blinn [13] to be required for successful matte extraction.

Szeliski *et al* [14] considered the automatic extraction of foreground and background layers combined with a single (rather than *per-pixel*) alpha value. Their approach allows the simultaneous estimation of the source images and the image motion, for a pair of translucent planes. Szeliski and Golland [15] compute 3D reconstructions which include alpha values at each point in a 3D voxel grid, but the difficulty of computing such a dense sampling means that their results are not yet suitable for compositing. This paper regularizes the problem using data-defined priors to generate a well constrained, accurate, solution.

3 The Problem

The problem this paper solves is the following. We assume that we are given a sequence of images comprising the composite image C and the motion parameters of two objects B and F. In this work, we will assume that the motion is modeled by a plane projective transformation, so that the motion parameters are given by 3×3 homogeneous transformation matrices H^b and H^f. We may assume that the sequence has been registered

using H^f, so that the foreground object is fixed in the image, and the background is sweeping past it. Further we assume that the background has been extracted by registering the background plane so that the background image B is known.

Image formation for the composite is assumed to be the following. The measured value at each pixel is some linear combination of background and (unknown) foreground:

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

In turn, F and B are integrals over the (unknown) area of some CCD sensor. Given that the background is known, this gives an equation for the two unknowns α and F for a monochrome image. In the case of colour images the α value is assumed to apply equally to each of the RGB channels. This results in three equations in four unknowns at each pixel. As will be seen these equations are often not independent.

Clearly, in a single image, the values of α and F are conflated. A foreground pixel can be mixed with a background for two main reasons. Either the foreground object is transparent or the foreground object does not occupy a whole pixel and so the pixel samples both foreground and background colours. In this case, the alpha value is related to the proportion of the pixel's cone of view occupied by the foreground object.

It will be useful to keep in mind the qualitative range of possibilities for α and F. For example: suppose the background is known to be white and in one image we see a pixel which is a 50% combination of white and red. Then there are three general possibilities for the foreground object:

1. The object is opaque pink
2. The object is red, but 50% transparent
3. The object is opaque red, but does not cover the whole background.

In order to resolve this ambiguity we can use multiple views of the scene, so that each foreground pixel is seen against different backgrounds. figure 2 illustrates the set of constraints that three views provides for a single foreground pixel, which will now be formally defined.

3.1 Problem Definition

We are given N images $\Psi_1 \ldots \Psi_N$, and wish to compute the alpha mask for a reference view, say Ψ_1. We use this reference view to induce the geometrical parametrization that will be described in section 4. We have dense motion correspondence for the background between the views, and the images are registered so that there is no motion of the foreground object throughout the sequence. The background motion is such that for each pixel $\mathbf{x}$ in the reference view we know the corresponding pixels $\mathbf{x}_i$ in the ith image that corresponds to a 3D foreground point (noted F in figure 2). We can rewrite (1) for each pixel in each view to give the generative model:

$$\Psi_i(\mathbf{x}) = \alpha(\mathbf{x}) \cdot F(\mathbf{x}) + (1 - \alpha(\mathbf{x})) \cdot B_i(\mathbf{x}) \tag{2}$$

where $B_i(\mathbf{x}) = B(H_i^b \mathbf{x})$ is the colour measurement of the point of intersection of the ray from the ith camera centre to $F(\mathbf{x})$ (see figure 2), as measured in image Ψ_i. Thus at

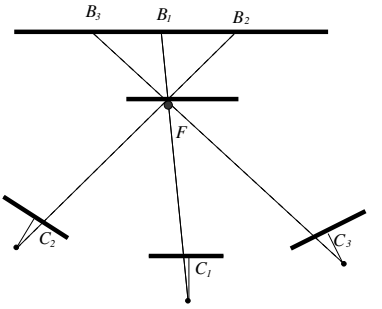


Fig. 2. Multiview camera setting. Each possible foreground point F is projected onto the ith image, combined with a different background colour B_i. When the measured B_is are different the foreground colour and transparency can be estimated. Note, this situation is equivalent to that of using different colour background screens [13].

each pixel we have N equations in 2 unknowns for monochrome, or $3N$ equations in 4 unknowns for colour pixels.

Note that this model assumes that the transparency and colours are independent of the viewing parameters, which is true for a Lambertian scene and when the occluding layer's depth is small compared to the distance from the cameras. In practice, these approximations are often good for real world sequences.

3.2 Bayesian Framework

Although the compositing equation (2) is overconstrained for three or more views, the estimation of α is still poorly conditioned. If a foreground pixel is moving over a uniform section of background, the colours B_i and the final composite values Ψ_i will all be similar, so that the N constraints reduce to just one. Therefore, the solution must be regularized, for which priors on the system parameters must be introduced. These priors are easily treated in a Bayesian framework, where the probability of the ensemble of observed images Ψ is the product of the probability of the observations given the parameters $\{\alpha, F\}$, and the prior probabilities of α and F

$$p(\Psi_1..\Psi_N) = p(\Psi_1..\Psi_N|\alpha, F)p(\alpha)p(F)$$

The likelihood (which is the first term in the product) is obtained from the error in the model (2), and modeling the noise process as a zero-mean Gaussian, gives the energy function (after taking logs)

$$L(\alpha, F) = \sum_{i=1}^{N} \int\int_{\mathbf{x}\in\mathbb{R}^2} \|C(\mathbf{x}) - \big(\alpha(\mathbf{x}) \cdot F(\mathbf{x}) + (1 - \alpha(\mathbf{x})) \cdot B^i(\mathbf{x})\big)\|^2 \mathrm{d}\mathbf{x} \quad (3)$$

To this (log) likelihood L is added the cost functions for the priors R_α and R_F yielding the combined cost

$$E(\alpha, F) = L(\alpha, F) + R_\alpha(\alpha) + R_F(F) \quad (4)$$

If there are p pixels in the image, then the cost is a function of $4p$ variables; p for the α matte, and $3p$ for the foreground colour. In the following sections we first derive the maximum likelihood estimate which minimizes (3) alone, and then derive and demonstrate three different priors that can be used to regularize the result, giving the MAP result as each is incorporated.

3.3 Maximum Likelihood

Assuming isotropic and homogeneous image noise with zero mean and standard deviation σ, the maximum likelihood solution is given by the least squares solution of (2). Because the pixels do not interact at this stage, the solution may be computed independently at each pixel, and this allows an efficient linear solution.

Equation (2) is bilinear in the two unknowns $\alpha(\mathbf{x})$ and $F(\mathbf{x})$ and so we have four times as many unknowns as pixels in the reference image. A reliable estimate for the transparency α results in a good estimate for the foreground colour $F(\mathbf{x})$ and so this equation can be solved by introducing a new variable $u(\mathbf{x}) = \alpha(\mathbf{x})\, F(\mathbf{x})$ and factoring it as a second step. The equation then becomes:

$$\Psi_i(\mathbf{x}) = u(\mathbf{x}) + v(\mathbf{x}) \cdot B_i(\mathbf{x}) \tag{5}$$

For simplicity we will write the equations for the monochrome case, for which $\mathbf{B} = [b_1 \ldots b_N]^\top$ and $\mathbf{C} = [c_1 \ldots c_N]^\top$ be the vectors of background and composite measurements—at a single pixel—and let $\mathtt{A}_{N\times 2} = [\mathbf{1}\ \mathbf{B}]$ be a matrix whose left column contains ones. Writing (5) in matrix form gives

$$\begin{bmatrix} 1 & b_1 \\ \vdots & \vdots \\ 1 & b_N \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \tag{6}$$

which we shall write as

$$\mathtt{A} \begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{C} \tag{7}$$

This overconstrained linear system can be solved in the least squares sense using the pseudo-inverse, giving $[u\ v]^\top = \mathtt{A}^+ \mathbf{C}$. We then substitute back to get the transparency $\alpha = 1 - v$ and foreground colour $F = u/\alpha$, which is valid only for visible places (i.e. $\alpha > 0$). In the case of colour images, $\mathtt{A}$ is $3N \times 4$.

This solution assumes that there is no noise in the recovery of the background $\mathbf{B}$. As the background is recovered from the images, we know that it does have the same noise characteristics (even if dampened by the mosaic) and so an alternative is to use total least squares (which is equivalent to line fitting in the foreground/composite space). The solution then minimizes the distance to both $\mathbf{C}$ and $\mathbf{B}$. To this end, we solve the following system where the mean is first subtracted from c_i and b_i

$$\begin{bmatrix} -c_1 & 1 & b_1 \\ & \vdots & \\ -c_k & 1 & b_k \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \mathbf{0} \tag{8}$$

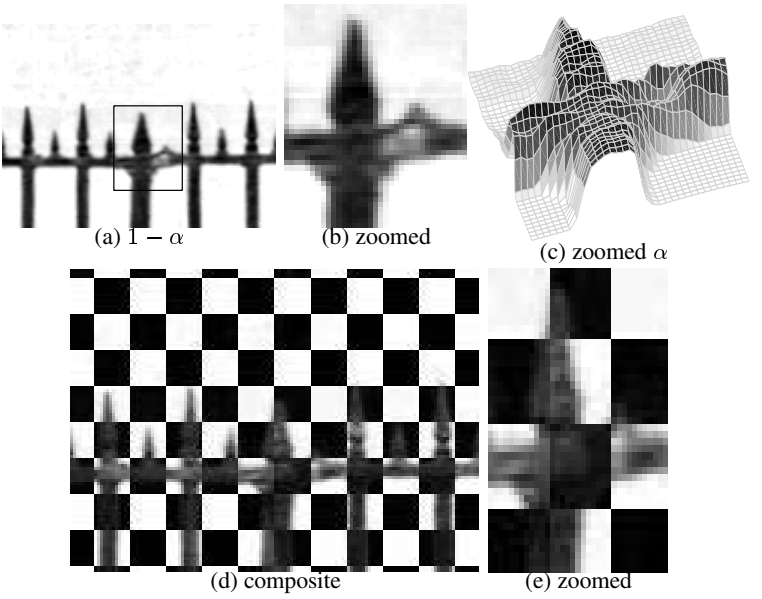
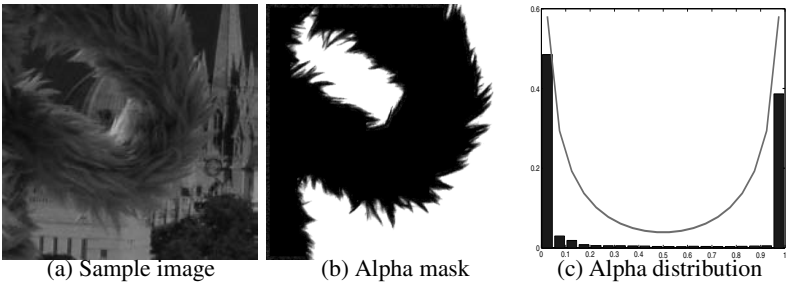


Fig. 3. Maximum Likelihood solution. The alpha mask is computed for each pixel independently by solving (8) and out of range values of α are clipped to the interval $[0, 1]$. Some residuals are still apparent where the background is uniform as seen especially in the transparent region. Colour spill from the background can be observed, i.e. light brown colour from the background is showing on the perimeter of the foreground.

The solution is then obtained using singular value decomposition by taking the singular vector belonging to the smallest singular value. In both methods weights can be incorporated into the system by premultiplying by a diagonal matrix W, and we use these to form the spatial prior explained below.

Figure 3 shows the results of this computation for the railings sequence. Section 4 describes how the images are registered and the background extracted, but here we shall just assume that these are given. The alpha and foreground values computed are clamped to the range $[0, 1]$ and then the compositing equation is used to replace the background with a checkerboard pattern. We note that while the computation is largely accurate, the alpha values are noisy, which induces a ghostly image of the original background.

The reason for this instability can be understood as follows: suppose the background colour is the same over N images (e.g. if the rays in figure 2 happen to intersect a locally homogeneous colour region of the background), then the resulting $N \times 2$ matrix $\mathtt{A}$ in (7) will have rank one instead of rank two. Consequently, there is a one parameter family of solutions for α and F for that pixel. It is exactly this situation which the priors, described in the subsequent sections, resolve.



(a) Sample image (b) Alpha mask (c) Alpha distribution

Fig. 4. Sample image with ground truth α-mask. The image is a zoomed portion of the first image of figure 1. The image and the ground truth are discussed in Section 5. The distribution of α for this figure is plotted in 4(c) with a super-imposed graph of a beta distribution.

3.4 Prior 1: Bounded Reconstruction

The first prior we apply is that the domain of α and F is bounded [5]. Our image measurements are normalized for the unit range and so is α. We thus add the following to the above constraints:

$$0 \leq \alpha \leq 1$$
$$0 \leq F \leq 1$$

Here F can be either gray-scale or a vector of image measurements (normally red, green and blue components). The optimization is no longer closed form, but as a quadratic problem subject to inequality constraints is relatively easily solved using quadratic programming [4].

3.5 Prior 2: Measured α Distribution

In a typical image the transparency is fairly constrained. Objects occupy continuous regions in the image and the background fills the rest. For the most part, only pixels on the boundary will have fractional transparency. As the count of boundary pixels is proportional to the square root of the foreground area on "average" images we expect to have far fewer of these. On images where the foreground object has more complex boundaries such as hair or fur, the boundary pixel count as a function of area will go as the fractal dimension of the curve, but will retain the tendency for most α values to be near 0 or 1. In order to incorporate this constraint, we obtained ground truth mattes for a variety of images, and computed the histogram of alpha values. Figure 4 shows a typical measured histogram.

Examining the figure, we observe that the distribution of alpha values is approximated by a beta distribution and so we can incorporate this knowledge as a prior in our system by defining the following (per-pixel) error function:

$$E_{\mathrm{d}} = \frac{1}{n}\sum_i (\alpha F + (1-\alpha)B - C_i)^2 + \frac{\rho}{\beta(\eta,\tau)}\alpha^{\eta-1}(1-\alpha)^{\tau-1} \qquad (9)$$

Where the sum is over all samples that contain this information (some images may not contain this pixel) and η,τ are constants that depend on the fractal dimension of the foreground shape. The coefficient $\beta(\eta,\tau)^{-1}$ equals $\frac{\Gamma(\eta,\tau)}{\Gamma\eta\Gamma\tau}$, where Γ is the gamma function, and normalizes the integral of the prior to unity. In this paper we use $\eta = \tau = -\frac{1}{4}$ which produces the graph in figure 4. The parameter ρ is not learned from data, but is specified in order to balance the prior and error terms.

Figure 5 shows the result of optimization of the new regularized error for several values of ρ. The noisy alpha values over the background have been removed and the transitional values at the foreground-background boundary are sharpened. However the boundary remains relatively noisy, as the straight vertical edges of the railings show.

3.6 Prior 3: Spatial Consistency

The spatial dependency in image measurements (pixels) is an important prior in many computer vision algorithms. However, describing the relationship between the colour values is fraught with difficulties. Often in Markov random field models a prior encouraging spatial continuity of the image is added to the energy function [12], i.e. a prior of the form $R_F(\mathbf{x}) = \sum_{\mathbf{y} \text{ near } \mathbf{x}} \|F(\mathbf{x}) - F(\mathbf{y})\|^2$. However, this leads to smoothing in textured areas, and reduces the definition at boundaries. On the other hand, applying such a prior to the alpha channel is far more reasonable: a piecewise constant image model is sensible for alpha providing we do not smooth over boundaries. In order to implement this, we impose a directional prior on α smoothness, using the image gradient as a guide. It is most likely that the actual alpha map agrees with some of the contours in the reference image and this is the prior that we want to use.

The gradients of the intensity image Ψ are computed by convolving with horizontal and vertical difference operators. At each pixel the local gradient is the vector $\mathbf{g} = (\frac{\partial C}{\partial x}, \frac{\partial C}{\partial y})^\top$. The weight of a contribution to pixel $\mathbf{x}$ from a neighbouring pixel $\mathbf{x}+\mathbf{p}$ is

$$W_{\mathbf{p}} = \exp(-\frac{\mathbf{p}^\top \mathtt{G} \mathbf{p}}{\mathbf{p}^\top \mathbf{p}})$$

where $\mathtt{G} = \mathbf{g}\mathbf{g}^\top$. This may be incorporated in the MAP (4) as a prior on α

$$\sum_{\mathbf{x}} \sum_{\|\mathbf{p}\|<=1} (W_{\mathbf{p}}(\alpha(\mathbf{x}+\mathbf{p}) - \alpha(\mathbf{x})))^2$$

but an efficient approximation is obtained by modifying the total least squares system (8). The system is augmented to include the neighbouring pixels as additional foreground and background estimates, forming a $9N \times 3$ system (assuming a 3×3 neighbourhood), with the contributions weighted as above. Solving the augmented system yields α values which have been smoothed *along* the edges, but not across [11]. A similar mechanism is used in optic flow computation [3, 8].

Figure 6 shows the result of applying this prior to the test sequence, and illustrates the improvement in edge consistency along the straight sections, without typical smoothing artefacts such as blunting of the spikes.

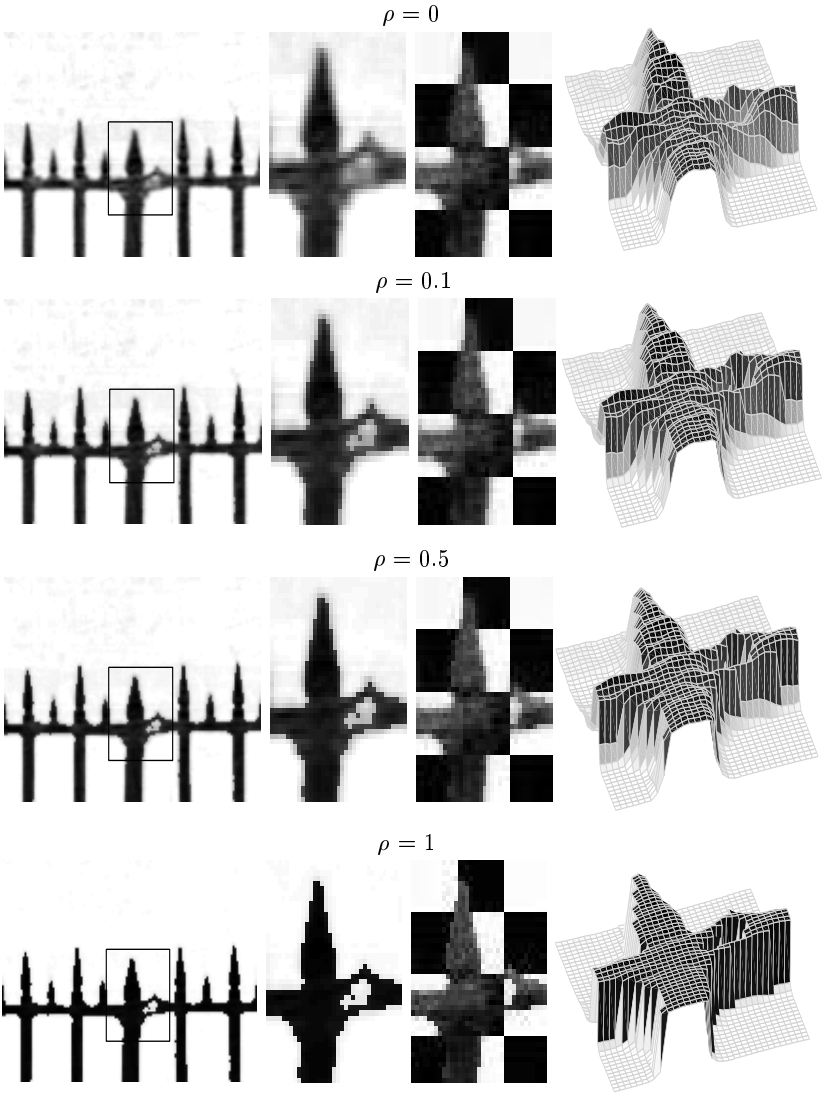


Fig. 5. Prior 2, on the distribution of α. As the prior on the distribution of α increases, the mask becomes more binary. The figure is organized in columns: Left column shows the alpha mask where white colour marks transparent pixels and black marks opaque ones. Next is a zoomed portion of the mask on the left. The third column demonstrates the use of the mask for composition on a checkerboard pattern to emphasize both 'overshooting' and 'undershooting' of the values. The fourth column shows the alpha map as a height field.
The reconstruction is shown for four values of the prior weight ρ. $\rho = 0$ is the MLE solution for which transparency is evident. As ρ increases, the matte is driven towards a binary image where pixels are labelled as purely foreground or background. Note that for $\rho = 1$, the mask is binary and a halo is visible around the spearhead. This is evident in color images.

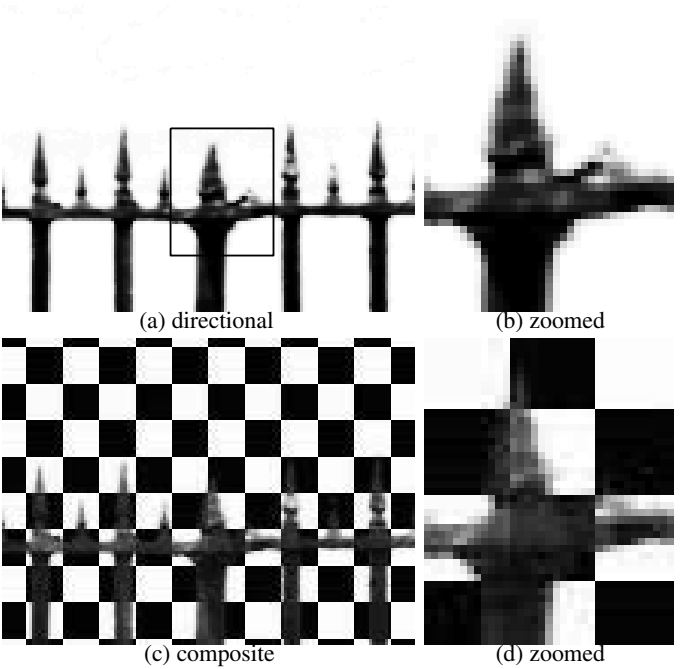


Fig. 6. Prior 3, directional: The reconstruction is regularized by adding a prior on α consistency, guided by the image gradient. The resulting mask is cleaner in the uniform areas and is not smoothed.

4 Implementation

In order to supply the data for (2) we need to relate the image measurements spatially, i.e. identify corresponding pixels across the images. If both the foreground and the background are planar the scene structure can be described by two homographies per input image, even if the objects are independently moving. When the foreground object has general shape we represent it as a depth map for the reference image as described later.

The input to the algorithm is a set of N views. We choose one as a reference view, performing all computations in a world coordinate system which is attached to this view. In particular, the α map is defined with respect to this view. This confers a number of practical advantages: the resolution of the 3D representation is automatically matched to the problem, and degradation due to resampling is minimized.

We compute the projective transformations that align the foreground and background objects from each view onto the reference view. This has the advantage that accurate algorithms are available which can perform such registration (see below). Let $H_i^f : \Psi_r \mapsto \Psi_i$ be the 3×3 homographies mapping pixels in the reference image onto pixels in the ith image through the foreground object. Similarly, let $H_i^b : \Psi_r \mapsto \Psi_i$ be the homography mapping pixels from the reference view onto the ith view through the

background plane. Assuming that the background image Ψ^B is available, and is aligned with the reference image, we can now supply (2) with data measurements as follows:

$$\Psi_i(H_i^f \mathbf{x}) = \alpha \cdot F + (1-\alpha) \cdot \Psi^B(H_i^b (H_i^f)^{-1} \mathbf{x}) \tag{10}$$

This is solved using the different priors as described above.

4.1 Recovering the Background

Sometimes the background can be obtained using a photograph where the foreground object has been physically removed, and a "clean plate" has been captured. In such a case the background plate must be registered to the reference frame using, say, point correspondences between the reference and the background plate.

When a clean plate is not available, as in our test sequences, it must be recovered from the data. Given the background homographies H_i^b and a rough segmentation of the scene we can compute the background image B via a mosaic [7, 9]. Setting mosaic pixels by taking the median colour value gives the results shown in figure 7(b). We compute and store the background in the reference coordinate frame (i.e. $\Psi_r - \Psi^B = 0$ on background regions).

(a) Reference image (b) Background image

Fig. 7. The "Railings" sequence. 7(a) is the reference frame and the background image 7(b) is a mosaic of the background regions in all seven images.

4.2 Aligning the Images

For the "Railings" images in figure 7, we used interest-point matching to get an initial 3D reconstruction [2]. This gives initial camera matrices and 3D points. From this reconstruction we extracted two dominant planes—one including the railings and the other the background building. These plane induce two homographies which are used for a rough background/foreground segmentation. The background plane homography

is further refined using a direct minimization over image intensities [7, 9] with a robust kernel [6] on the intensity comparison. In this case the direct method produces a very good alignment of the background portion of the various images. We use these homographies to compute an image mosaic containing only background pixels.

5 Comparison with Ground Truth

In this section we compare the proposed methods to a ground truth image. In order to obtain a sequence for which ground truth could be extracted, we locked off the camera and foreground object (a furry animal), and moved a textured background plane behind. This sequence emulates the more complex case of the previous example, but allows us to compute ground truth. The set of test images is shown in figure 1. In addition, two further images were obtained of the object against two different backgrounds of near-constant colour. Combining these images using a technique based on [13] allowed a ground truth mask to be obtained with relative ease. Figure 8(b) illustrates the recovered ground-truth mask. Comparing figure 8(d) shows that simply requiring a bounded reconstruction as in §3.4 yields an alpha matte which is visibly far from the ground truth. The subsequent figures illustrate that the beta-distribution prior allows for some more veridical reconstructions. On the other hand, the directional prior has been confused by the wealth of interior edges in the animal's fur, and simply weakens the constraints on the reconstruction, yielding a visibly poorer mask. The next section discusses these points.

6 Conclusions

We have demonstrated the automatic extraction of alpha mattes from multiple images of moving planes, or a static 3D scene. By expressing the problem as a maximum a posteriori (MAP) estimation, the poorly constrained problem has been successfully regularized, allowing accurate estimation of colour and transparency.

Although the paper shows that adding appropriate priors can allow a clean matte to be successfully pulled from a video sequence, offering a significant reduction of effort over single-frame methods, there remains the issue of how to set the priors. As shown in the examples, different combinations of weighting between smoothness and directional priors are required for success on different sequences. Some rules of thumb can be discerned. For example, prior 2 on the distribution of α controls the sharpness of the resulting mask, and according to our experiments, it performs well when ρ is in the range $0.2 \ldots 0.5$. Finally, however, the quality of the pulled matte can only be evaluated by the operator. This situation should not be a cause for dismay, however. The goal of our algorithm is to reduce effort on the part of the matte puller, and in that it succeeds well.

Current and future work is concerned with two issues. Integration of the multiple view constraint with the successful modern single view methods [5, 10] is expected to provide valuable complementary information, as single-view methods work best on background images with well constrained image statistics—for example constant backgrounds, which are the source of ill conditioning in our method. Secondly, although

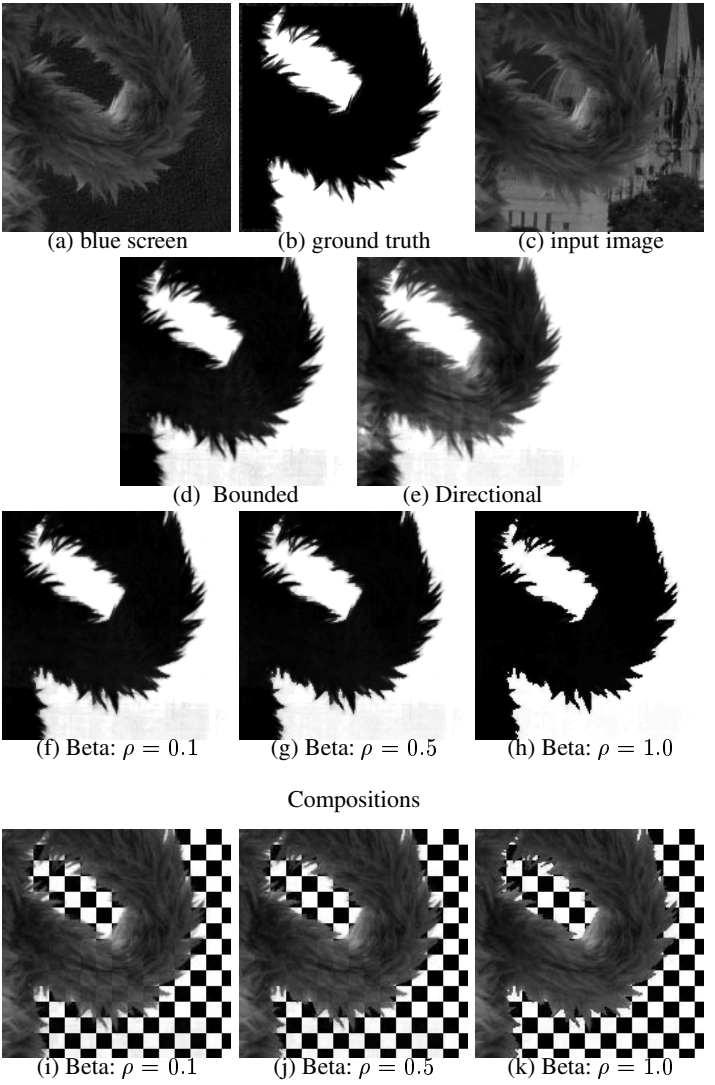


Fig. 8. Ground truth for the Monkey sequence. See section 5 for details

the theory described here applies to general scenes (i.e. nonplanar), and techniques are available to compute the pixel correspondences in these cases, much work remains before the general problem can be said to be considered solved.

Acknowledgements

Funding for this work was provided by a DTI/EPSRC Link grant. AF would also like to thank the Royal Society for its generous support.

References

1. C. Barron. Matte painting in the digital age. In *Proc. SIGGRAPH (sketches)*, 1998. `http://www.matteworld.com/projects/siggraph01.html`.
2. P. Beardsley, P. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *Proc. ECCV*, LNCS 1064/1065, pages 683–695. Springer-Verlag, 1996.
3. M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. ICCV*, pages 231–236, 1993.
4. P. T. Boggs and J. W. Tolle. Sequential quadratic programming. In *Acta Numerica*, pages 1–51. 1995.
5. Yung-Yu Chuang, Brian Curless, David Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *CVPR 2001*, 2000.
6. P. J. Huber. *Robust Statistics*. John Willey and Sons, 1981.
7. M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In G. Sandini, editor, *Proc. ECCV*, pages 282–287. Springer-Verlag, 1992.
8. H. H. Nagel. Extending the 'oriented smoothness constraint' into the temporal domain and the estimation of derivatives of optical flow. In O. D. Faugeras, editor, *Proc. ECCV*, pages 139–148. Springer–Verlag, 1990.
9. S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *Proc. CVPR*, 1997.
10. P. Perez, A. Blake, and M. Gangnet. Jetstream: Probabilistic contour extraction with particles. In *International Conference on Computer Vision*, pages 424–531, 2001.
11. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE PAMI*, 12(7):629–639, Jul 1990.
12. M. Shizawa and K. Mase. A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis. In *CVPR91*, pages 289–295, 1991.
13. A. R. Smith and J. F. Blinn. Blue screen matting. In *Proc. SIGGRAPH*, pages 259–268, 1996.
14. R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2000)*, volume 1, pages 246–253, 2000.
15. R. Szeliski and P. Golland. Stereo matching with transparency and matting. In *Sixth International Conference on Computer Vision (ICCV'98)*, pages 517–524, 1998.