Thirteenth Eurographics Workshop on Rendering (2002)
P. Debevec and S. Gibson (Editors)

Image-based Environment Matting

Yonatan Wexler, Andrew. W. Fitzgibbon and Andrew. Zisserman

Department of Engineering Science, The University of Oxford, UK

Abstract

Environment matting is a powerful technique for modeling the complex light-transport properties of real-world optically active elements*: transparent, refractive and reflective objects. Recent research has shown how environment mattes can be computed for real objects under carefully controlled laboratory conditions. However, many objects for which environment mattes are necessary for accurate rendering cannot be placed into a calibrated lighting environment. We show in this paper that analysis of the way in which optical elements distort the appearance of their backgrounds allows the construction of environment mattes* in situ *without the need for specialized calibration.*

Specifically, given multiple images of the same element over the same background, where the element and background have relative motion, it is shown that both the background and the optical element's light-transport path can be computed.

We demonstrate the technique on two different examples. In the first case, the optical element's geometry is simple, and evaluation of the realism of the output is easy. In the second, previous techniques would be difficult to apply. We show that image-based environment matting yields a realistic solution. We discuss how the stability of the solution depends on the number of images used, and how to regularize the solution where only a small number of images are available.

Categories and Subject Descriptors (according to ACM CCS): I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—modeling and recovery of physical attributes. I.3.3 [Computer Graphics]: Picture/Image Generation—algorithms. I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—color, shading, shadowing, and texture.

1. Introduction

We wish to render images of scenes in which *optically active elements* with complex light-transport characteristics are accurately incorporated. In particular, our goal is to learn the light-transport properties of *real world* objects from images.

Of course, modeling the object's light-transport can be achieved by obtaining an accurate geometric model, and accurate refractive indices. Then, modern ray tracing techniques allow rendering even at interactive speeds[1]. However, obtaining the geometry in itself may be very difficult. In one of the examples in this paper, the transparent object is an old window, where imperfections in manufacturing has led to small deviations in shape inside the glass. No technique is known to the authors for the measurement of the internal 3D geometry of transparent objects, even if we were permitted to physically modify the window or its surround.

Recently however[2, 3], methods for obtaining environment mattes of real-world objects have been introduced. These systems illuminate the real objects with carefully calibrated backgrounds, and capture images of the appearance of the object under these backgrounds. Analysis of the images allows the objects' light-transport properties to be computed. These techniques permit the discovery of complex optical behaviour of real-world objects without explicit measurement of geometry or transmissivity parameters, and have yielded impressive composite images. However, they remain limited to situations where the object can be placed in a calibrated laboratory setting. In the example of the old window, the window would need to be removed before measurements were performed.

This paper shows that such calibration is not necessary in order to obtain realistic environment mattes. A set of images of the object *in situ* can be used to determine the op-

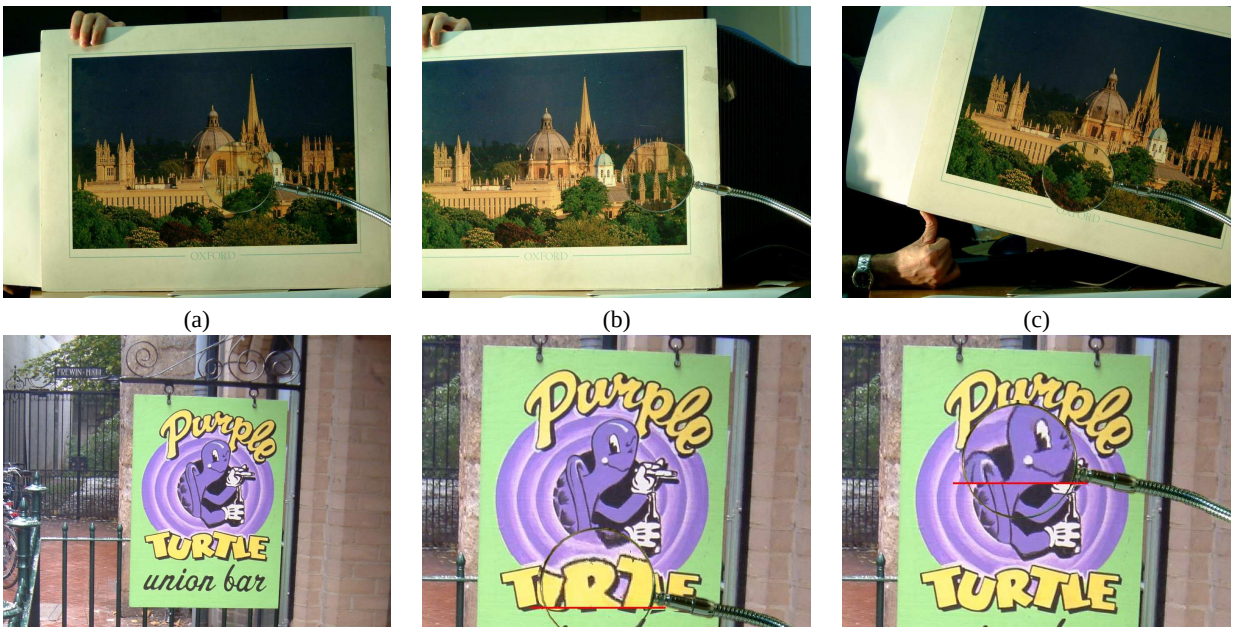


Figure 1: ***The goal of this paper****. The input is a set of images (a to c) of an optically active element—the magnifying glass—in front of a moving background. The task is to apply the light-transport properties of the imaged element to a new image (d), to generate novel composites (e) and (f) which include not just the magnifying glass, but also its light-transport properties, evidenced by the magnification of the background characters. In each image, the area above the red line shows environment matting, while the image below the line is traditional alpha matting. No* a priori *model of the lens or background was used.*

tical transport properties. We show that accurate environment mattes can be computed from natural images, without the need for specialized calibration of the acquisition. The method proceeds in two stages: first, the environment map is learnt from a set of example images containing the optical element of interest (e.g. the magnifying lens in figure 1); second, the element's environment matte is applied to a new background image.

Related work fits into two categories: lightfields and alpha matting. Lightfield acquisition and rendering[4, 5] captures the set of light rays in a particular environment, allowing new viewpoints of the same environment to be generated which retain the light distribution within the environment. In particular the environment may contain transparent or reflective objects, new views of which may be generated providing the background does not change. However, in order to combine lightfields, or to place lightfield-captured objects in new environments, a model of light transport is needed, which existing techniques do not provide.

Recent work on extraction of α-mattes from image sequences[6, 7, 8] uses similar tools to those in this paper to compute transparency mattes for moving objects. If these papers may be considered uncalibrated extensions to the two-image matte extraction technique of Smith and Blinn[9], then the work reported in this paper is an uncalibrated extension of the original environment matte acquisition of Zongker et al[2].

Notation

Entire images, i.e. the $w \times h \times 3$ RGB array are represented as calligraphic uppercase letters $\mathcal{I}$. An individual RGB pixel from $\mathcal{I}$ is $I(x,y)$, and the (x,y) is dropped when using a single pixel as an exemplar for an image formation process which is the same at all pixel locations. A set of images—specifically now the set of input images—is denoted by subscripting, $\{\mathcal{I}_i\}_{i=1}^{n}$, as are individual pixels from a set $I_i(x,y)$.

2. The model

Our goal is to recover the action of optical elements from images. Therefore, the first desideratum is a mathematical model for that action. The model chosen is similar to earlier work[2, 3], in that the action of the optical element is modelled entirely as a 2D to 2D mapping. The observed images are considered to be the composition of (an image of) the background scene and an environment matte which encodes, for each output pixel, the set of input pixels from which it samples. Although the mapping is only from 2D to 2D, the

background "image" may live on any surface in 3D, which makes the technique quite general. In previous work[2, 3], the surfaces used were either the plane at infinity (as in environment mapping[10]) or a piecewise planar surface (e.g. cubic environment maps[11]).

To explain the model, consider the formation of a composite image $\mathcal{C}$, with RGB triple $C(x,y) \in \mathbb{R}^3$ at pixel (x,y). The composite will be the combination of the environment matte (whose form will be defined shortly) and the background image $B(u,v)$.

Each pixel in the composite collects light from a blend of pixels in B. The set of pixels which contribute to a given output pixel p is called the *footprint* of p, or p's *receptive field*. Previous researchers have defined the footprint using rectangular regions[2] or mixtures of Gaussians[3]. In this work, we must deal with complex multimodal distributions during acquisition, so we use a discrete map of source pixels, where each source pixel has an associated weight. The value of the output pixel is then computed as a weighted sum over the pixels of B. Thus if we can compute the receptive field for each pixel, we can compute the composite.

More formally, the receptive field is denoted as $r(u,v)$, and its effect is modelled as a weighted sum of background contributions

$$C = \sum_{u,v} r(u,v)B(u,v)$$

The summation is over all pixels in the background image, and there is a separate receptive field $r(u,v)$ for each foreground pixel (x,y). Figure 2 illustrates the process for a single pixel. Collecting the separate receptive fields for each (x,y) location yields the definition of the four-dimensional *environment matte*

$$w(x,y,u,v) = r(u,v) \text{ at } (x,y)$$

Recovery of $\mathcal{W}$ is the primary goal of this paper.

The development of the model to this point has ignored the contribution from reflection off the element itself (e.g. the handle of the magnifying glass), which is modelled as a foreground contribution F. This yields the complete description of the formation of the composite image $\mathcal{C}$ as follows:

$$C(x,y) = (1-\alpha(x,y))F(x,y) + \alpha(x,y)\sum_{u,v} w(x,y,u,v)B(u,v)$$

where a transparency term α is included to model partial pixel coverage. Acquiring environment mattes from images is a matter of determining $\mathcal{W}$, $\mathcal{F}$ and α given examples of $\mathcal{C}$ and $\mathcal{B}$.

In the standard formulation without environment matting[9], the background pixel $B_0 = B(x,y)$ passes straight through the optical element, and we have the standard compositing equation[6, 7, 9]

$$C = (1-\alpha)F + \alpha B_0$$

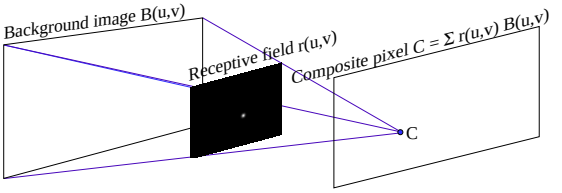


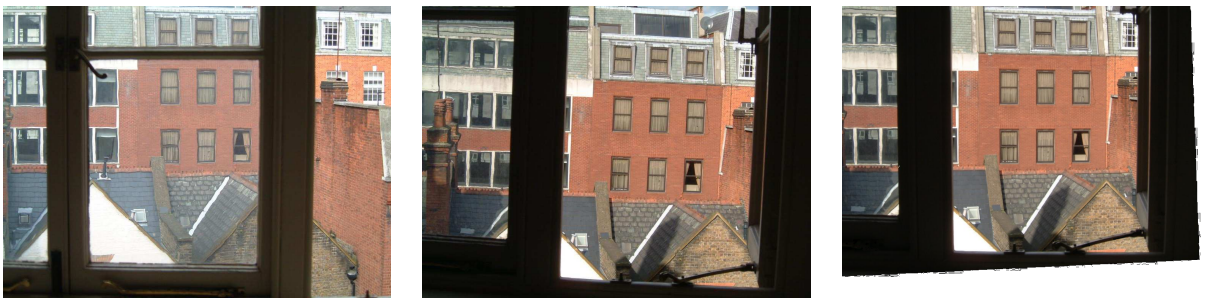
Figure 2: ***Formation of a single output pixel****. The pixel's receptive field $r(u,v)$ allows each pixel of the background to contribute to the output pixel's colour. The environment matte $w(x,y,u,v)$ is the set of all receptive fields, one per pixel (x,y) in the output image.*

The receptive field $r(u,v)$ may be thought of as the probability that a background pixel contributes to a particular composite pixel. We return to this interpretation when computing $\mathcal{W}$ in section 5.

Several factors contribute to making the task of learning the environment difficult, and the remainder of the paper discusses how to address these. Briefly, the difficulties include:

- We may not know in advance the background image $\mathcal{B}$. Section 3 describes how to compute a "clean plate" background image given a set of overlapping images. Section 4 shows how to precompute $\mathcal{F}$ and α.
- We will almost always have too few images to completely determine $\mathcal{W}$. Assumptions must be made about the form of $\mathcal{W}$ in order to obtain a tractable solution. Previous work[2, 3]has made these assumptions by choosing distribution models with small numbers of parameters. In section 5 we show how a non-parametric assumption about the general *behaviour* of $\mathcal{W}$ rather than a parametric model of its form suffices to give excellent estimates of the light transport properties of non-translucent objects.
- Sometimes the number of images available is extremely limited. Section 7 shows how the incorporation of an *a priori* coarse model of the distortion field can regularize the problem, and we show an example of an environment matte extracted from the minimum number possible—a single image pair.
- The environment matte $\mathcal{W}$ is a large four-dimensional linear operator. Its discrete representation would, if implemented naively, occupy $O(N^2)$ storage for N-pixel images, or about 100 gigabytes in our examples. We show how this storage cost is avoided.

Figure 1 shows an example input sequence, containing a magnifying glass for which an environment matte is to be computed. The application of that matte to a novel background is demonstrated in the figure, and the following sections describe the computational steps.



(a) (b) (c)

Figure 4: ***Two-image example****. (a) Base image. (b) A single reference view of the background, taken by moving the camera. (c) The reference view is warped so that the pixels on the red wall in the background lie approximately under their counterparts in the foreground. The environment matte will describe the remainder of the transformation.*

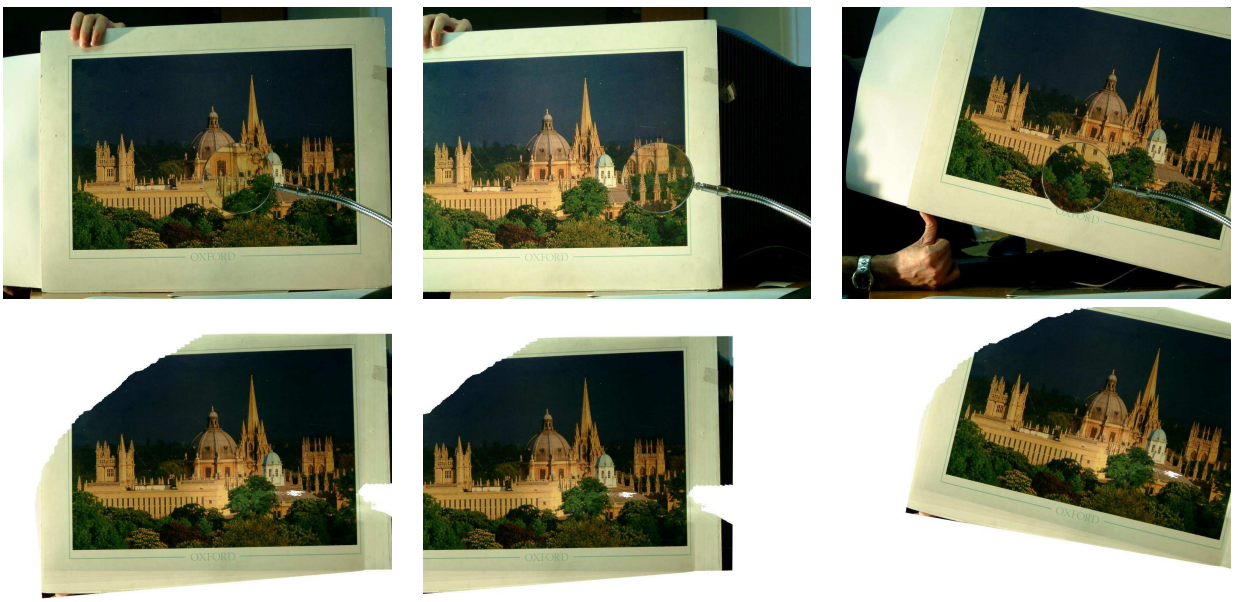


Figure 5: ***Computed*** $(\mathcal{C},\mathcal{B})$ ***pairs after clean plate extraction****. The top row shows the original images, the bottom row the image background image (the plane) replaced by the computed clean plate. The optically active element (the lens) and foreground occlusion (the handle of the lens where* $\alpha \neq 1$*) have been removed.*

3. Getting a clean plate

Computing the background image may be achieved by mosaicing the moving-background sequence[12] or moving the camera. The example in Figure 1 shows an example where the background is moving relative to the foreground, and figure 4 illustrates the effect of moving the camera, with a planar background. In both of these cases, the motion of the background is modelled by a plane projective transformation—a planar homography.

In the first instance, where the background exhibits motion relative to the optical element, no single background image is available, but we can automatically compute a "clean plate" by assembling unoccluded pixels from several images. By tracking points on the background (see section 8), a set of homographies are computed which register all images to a canonical reference image, say image 1. Call these homographies H_t, and define the function $\pi([x,y,z]^\top) = (x/z, y/z)$ and operation $\mathsf{H} * (x,y) = \pi(\mathsf{H}[x,y,1]^\top)$. Then, the registering homographies mean that for (x,y) a background pixel in both image 1 and t,

$$I_1(x,y) \approx I_t(\mathsf{H}_t * (x,y))$$

Then we warp all the images to frame 1, compute the median colour at each location and assume that will be a reasonable

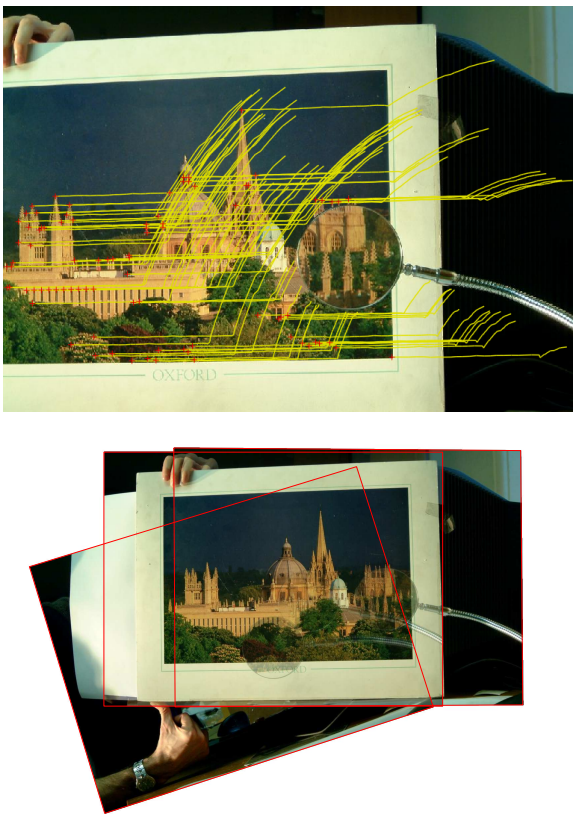


Figure 3: ***Computing a clean background for each image****. (Top) Automatically computed feature tracks on the input sequence. (Bottom) Three representative frames from the sequence assembled into a mosaic in which the background remains stationary, and the magnifying glass moves.*

estimate of the background image for image 1:

$$B_1(x,y) = \text{median}_t\, I_t(\mathsf{H}_t * (x,y))$$

Because the background is the same for each image, we inverse warp the background to generate a registered (I_t, B_t) pair at each time t:

$$B_t(x,y) = B_1(\mathsf{H}_t^{-1} * (x,y))$$

Figure 3 illustrates the process and figure 5 shows an example set of computed (composite, background) pairs.

Figure 4 shows an example where the camera is moved to obtain a clean view of the background. In the first image, which is the image into which we will place the final composite, the camera looks out through the window. The area over which the composited object (an image of a hot air balloon) will be positioned is planar, so a clean plate may be obtained by moving the camera to one side, in order to look through the open half of the window.

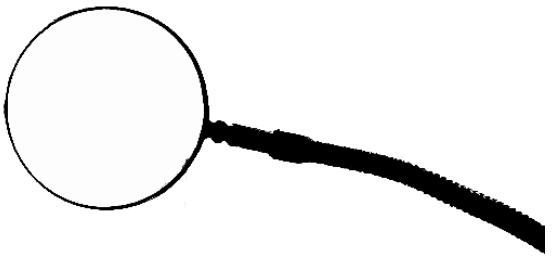


Figure 6: ***Approximate alpha matte*** *computed for foreground element.*

An approximate initial homography is obtained using four manually indicated point correspondences. Then dense point correspondences are obtained (see section 8), and a least-squares best fit homography is computed over the entire scene. Although the optical action of the window means that each correspondence includes some error due to the refraction of the light rays, the errors tend to be uniformly distributed, so the aggregate homography is sufficient for the matte computation. If the homography is wrong, the matte computation will model the error by shifting objects as they move behind the glass, but computation will not be otherwise hindered. Therefore, providing the homography is reasonably accurate, no distortion will be evident in the final composite.

4. Computing the foreground elements

Computation of $\mathcal{W}$ is significantly simplified if the parameters of the foreground elements $\mathcal{F}$ and α are computed first. This is possible if the background is reasonably heterogeneous, and moving relative to the foreground element. For the magnifying glass, we wish to recover the colour of the magnifying glass frame $\mathcal{F}$, and transmissivity values α. For example, the metal handle will have $\alpha = 0$, lens interior $\alpha = 1 - \varepsilon$ and partial pixels where the lens joins the handle will have intermediate values $0 < \alpha < 1$. For this work, an accurate alpha matte is not necessary, so a number of pragmatic schemes are applied in order to obtain the matte, as follows.

A coarse initial estimate of alpha is obtained by superimposing all images $\mathcal{I}_{i=1..n}$ and computing the per-pixel mean $m(x,y) = \frac{1}{n}\sum_i I_i(x,y)$ and variance $\sigma^2(x,y) = \frac{1}{n-1}\sum_i (I_i(x,y) - m(x,y))^2$. Because the background is changing, we expect foreground pixels to have low variance, and the background pixels to have high variance. We could then impose a hard threshold τ on the variance to separate changing (i.e. background) and constant (i.e. foreground) pixels. However, partially covered pixels will have α between 0 and 1. Roughly modelling this by passing the vari-

ance through a sigmoid

$$\alpha = \frac{1}{1+e^{-k(\sigma^2-\tau)}}$$

yields an approximate alpha matte. The tuning parameters (k,τ) were set manually in this example to give a satisfactory matte near the element boundary, where it is difficult to manually compute alpha values. However, it also includes isolated areas where the background exhibited sufficiently little change that it was marked as foreground, and these areas are manually removed using a paint package. This allows a clean matte α, as shown in figure 6, to be obtained with a few minutes of effort.

Given α, the foreground colour can be measured from the registered background images (the output of section 3) and the foreground. This estimate may be further refined after the environment matte is measured (next section).

5. Estimating the environment matte

Having reasonable estimates of α and $\mathcal{F}$, we may transform any given image $\mathcal{I}_t$ to a purely environment-matted composite $\mathcal{C}_t$ for which, given registered background image $\mathcal{B}_t$

$$C_t(x,y) = \sum_{u,v} w(x,y,u,v) B_t(u,v)$$

Thus, we ask how, given a set of $(\mathcal{C}_t,\mathcal{B}_t)$ pairs, we may obtain an estimate of $\mathcal{W}$.

In order to compute the receptive field of a given pixel p, we need at least two images: one containing the optically active element (e.g. the lens in figure 1), and one containing only the background. If the component of diffusion is small, then pixels in the background which have contributed to p's colour will have similar colour to p. In fact, for each background pixel, the similarity between its colour and the query colour is a function of the amount that background pixel contributes.

Assume we are given a composite-background pair $(\mathcal{C}_t,\mathcal{B}_t)$. The composite $\mathcal{C}_t$ contains the optical element, and is therefore assumed to be the result of compositing $\mathcal{B}_t$ as above. We may obtain a (poor[†]) estimate of $\mathcal{W}$ for this pair alone—call it $\hat{\mathcal{W}}_t$—using

$$w_t(x,y,u,v) = \exp\left(-\lambda\left|C_t(x,y) - B_t(u,v)\right|^2\right) \qquad (1)$$

and then normalizing so that $\sum_{u,v}\hat{w}_t(x,y,u,v) = 1$:

$$\hat{w}_t(x,y,u,v) = \frac{w_t(x,y,u,v)}{\sum_{u,v} w_t(x,y,u,v)} \qquad (2)$$

[†] It is at this point that we are making a nonparametric assumption about the probability density $r(u,v)$, essentially saying that the probability distribution is dominated by its modes. Wide flat areas of the distribution will be suppressed by the subsequent normalization and will thus not be allowed to contribute to $\mathcal{W}$.

Here, λ is a tuning parameter, set to 10^{-2} in our experiments. Then background pixels which are similar in colour to the composite pixel $C_t(x,y)$ will be considered to be part of the receptive field of (x,y). Of course there will be many accidental similarities, so the estimated receptive field will be larger than its true extent. This is mitigated by comparing 3×3 windows rather than individual pixels, but this produces only a small improvement in signal-to-noise ratio. It is undesirable to use a larger window as this will reduce the spatial accuracy of the environment matte. Happily, however, the receptive field is constant over time, so the true receptive field must have high values for each of the N computed functions $\hat{\mathcal{W}}_t$.

Think of the estimate $\hat{\mathcal{W}}_t$ for each image as measurements from a "sensor" which returns the probability that the pixel at (u,v) contributes to (x,y). Then if the sensors are considered as independent, the estimate of $\mathcal{W}$ given all images may be computed by multiplying the per-image estimates $\hat{\mathcal{W}}_t$ and renormalizing:

$$w(x,y,u,v) = \prod_t \hat{w}_t(x,y,u,v).$$

This procedure combines the relatively poor single-pair estimates $\hat{\mathcal{W}}_t$ in a way which best uses the available information. Because the false similarities will in general not occur at the same place in the matte, but the true similarities will tend to be consistent, the procedure generates more and more accurate mattes as more images are added.

Figures 7 and 8 show the process in operation for two different choices of (x,y). The receptive fields in figure 9 show that after two images, the matte still collects from many locations in the source image, but after eight images, the receptive field for the indicated pixel has converged to a tight, accurate estimate.

At this stage, one could approximate the recovered matte using any number of schemes analogous to those used by Chuang et al[3]. Note that for the non-diffuse objects considered here, the final matte is often unimodal, however the generality of our representation is necessary in order that the intermediate stages may carry multiple hypotheses for the final mode. In fact, we do not approximate at this stage, because the matte can be directly used for composition without ever storing the full 4D array.

6. Using the environment matte

The purpose of this section is to illustrate how a general, multimodal environment matte can be computed and used to generate new composites. For concreteness, consider Figure 1. The outputs (e) and (f) are two of several hundred frames from an output movie which is to be generated by compositing the original environment matte with the novel background (d). In order to avoid storing all of $\mathcal{W}$, the output movie is generated as the environment matte is constructed.

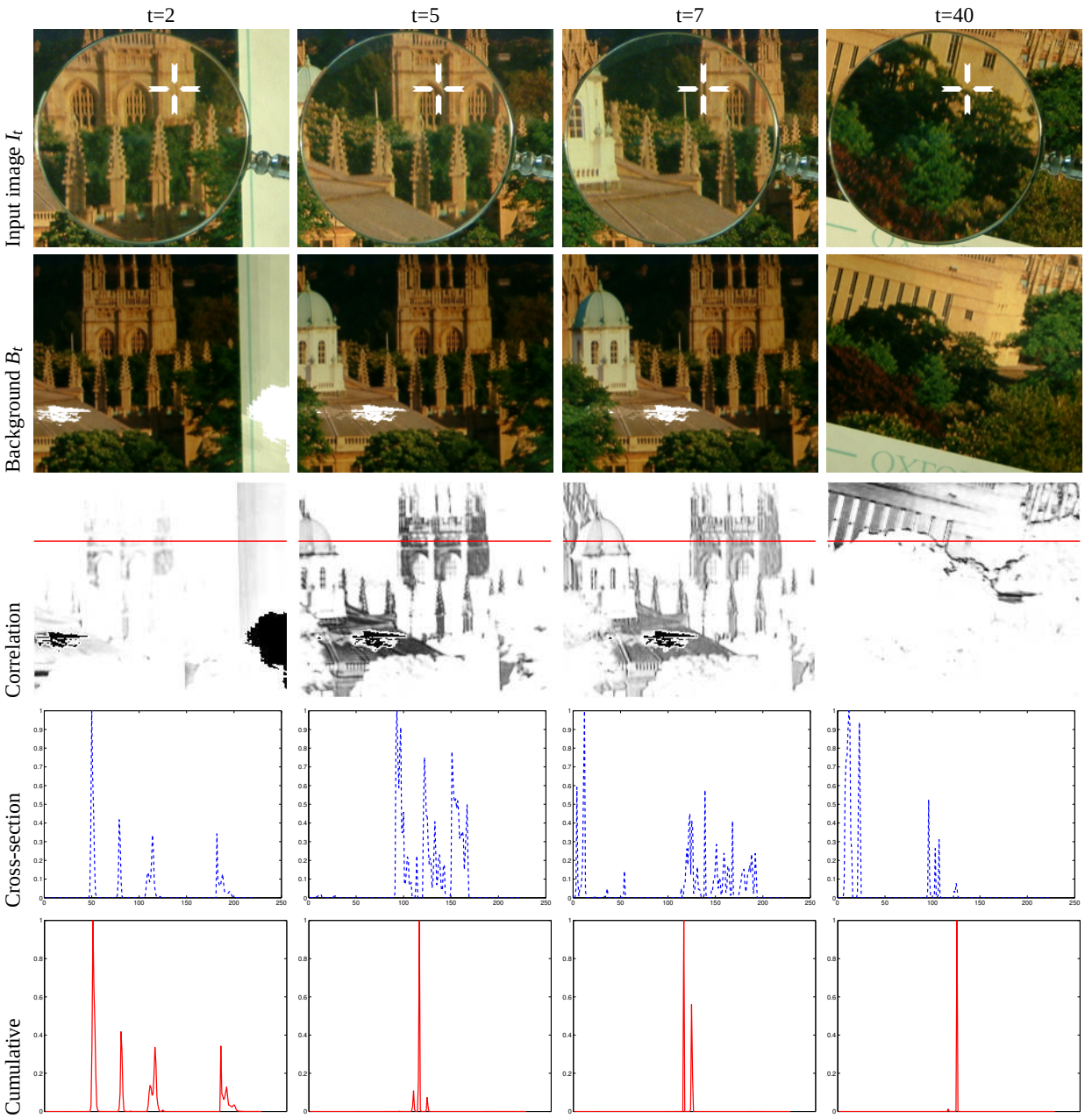


Figure 7: ***Integrating receptive fields for a single output pixel****. The query pixel is marked with a white cross. Each column corresponds to a new image pair. The first two rows show measured foreground I_t and registered background B_t. The white smudge on the background images is an area where no background colour could be computed, as it was always occluded by the magnifying glass.*
The third row shows the receptive field $r(u,v)$ of the output pixel, computed from just that view pair. The fourth row shows a cross-section through the receptive field. It can be seen that a single image does not constrain $r(u,v)$ very tightly – the curves are far from unimodal. The red curves in the fifth row, on the other hand, show the normalized cumulative products of the per-view curves. These represent the integrated receptive fields, and show that the erroneous peaks in the distribution are quickly eroded as more images are added. Furthermore, the finally accepted peak does not necessarily correspond to a maximum in any individual image.

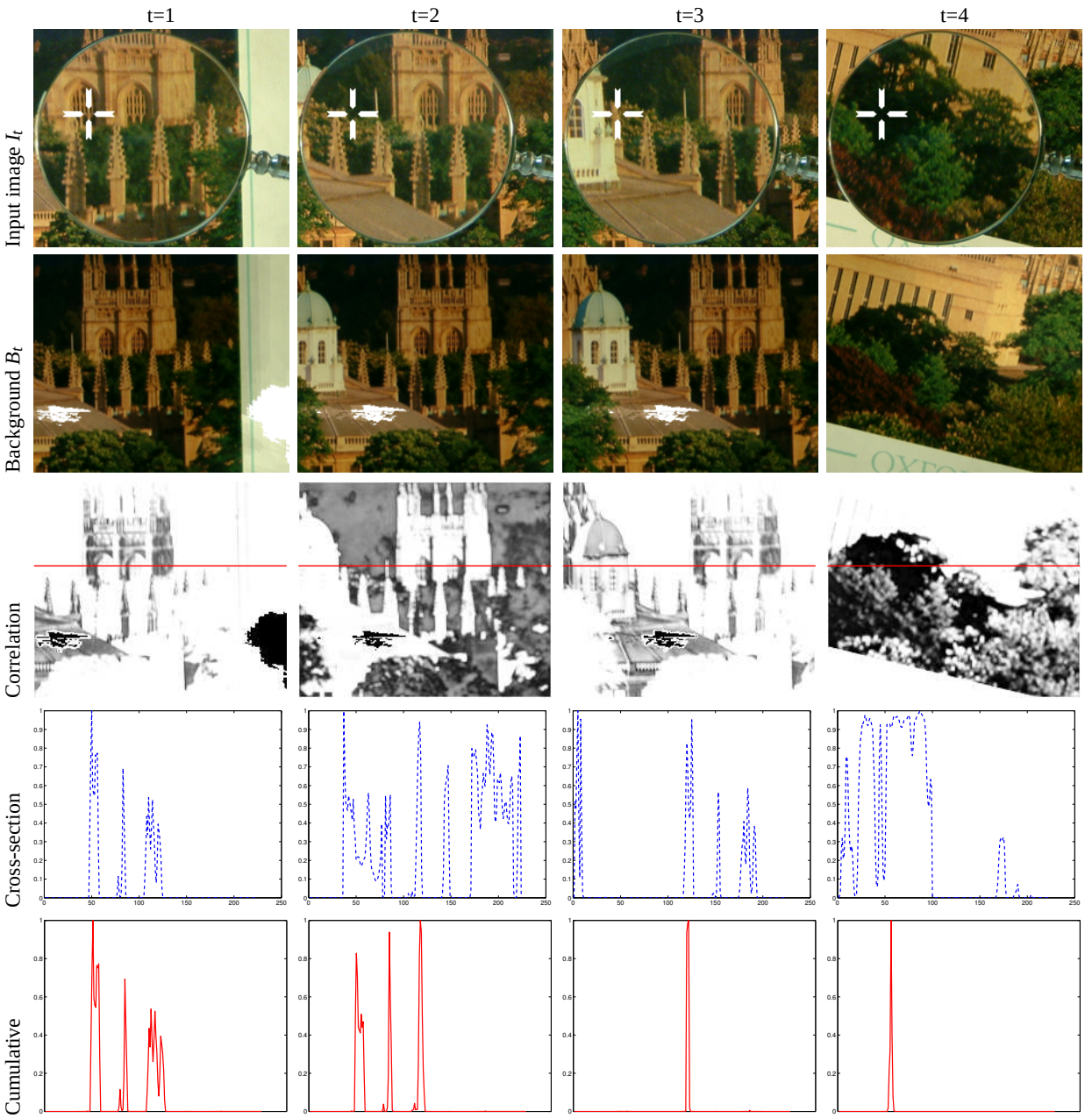


Figure 8: ***Integrating receptive fields for (another) single output pixel****. In this example, the final image moves the unimodal receptive field significantly, showing that accuracy is not simply guaranteed by unimodality.*

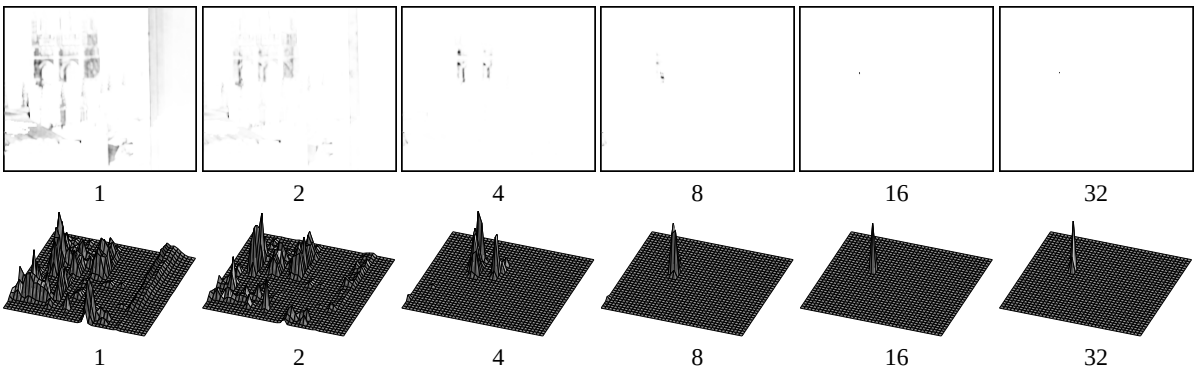


Figure 9: ***Refining receptive fields for a single output pixel****. The receptive field $r(u,v)$ for a single (x,y) pixel as more views are added. After the first pair, the receptive field is far from accurate, with many false maxima (dark regions). As more views are integrated, the estimate is progressively refined. (Top row): intensity map—dark pixels have higher weights. (Bottom row): surface plot.*

Figure 10: ***Two image composite*** *using the learnt environment map from figure 4. Left: added background layer in the shape of a hot air balloon. Note the deformation of the checkerboard. Right: the texture mapped balloon.*

For each pixel within the lens, the receptive field is computed, and the novel composite is computed at that pixel for all frames of the output movie. The receptive field for this location may then be discarded, and the next pixel is processed. This procedure is fast as long as the output movie and input images occupy less memory than the entire environment matte. In these examples, each image is of the order of a megabyte in size, so the total storage requirement is of the order of 400MB rather than the 100GB required to store the environment matte. Total time to render the output movies is of the order of hours, but may be further sped up as discussed in the next section.

7. Adding priors

As we use fewer and fewer views, more prior constraints must be added to ensure a matte of sufficient quality. In figure 4 we have only one foreground/background pair, so the form of w is tightly constrained to ensure a reasonable matte. Fortunately in this case some obvious constraints present themselves. As the distortions produced by the glass are small, we may assume that each output pixel obtains a contribution only from nearby pixels. Formally, this is $w(x,y,u,v) = 0$ for $(u-x)^2 + (v-y)^2 < \tau^2$ where τ is a distance threshold in pixels, set to 5 in example 2. Second we compute w only at locations where $I_t(x,y)$ has peaks in the local autocorrelation function (identified by a Harris corner detector), thus yielding sharp estimates of the receptive field, even with only one image. Assuming spatial coherence in the less textured areas then allows the estimates from reliable regions to be propagated into the smooth areas. These are the same sort of assumptions which are used to regularize optical flow algorithms[13], and hence are suitable only when

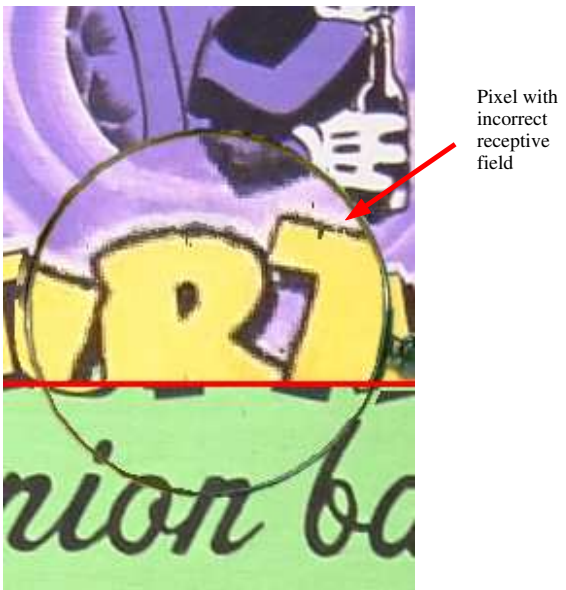


Figure 11: ***Zoomed view of lens composite*** *from figure 1. (Above red line) Environment matting. (Below red line) Alpha matting. A pixel for which the receptive field has been poorly estimated is indicated. No attempt has been made at manual cleanup.*

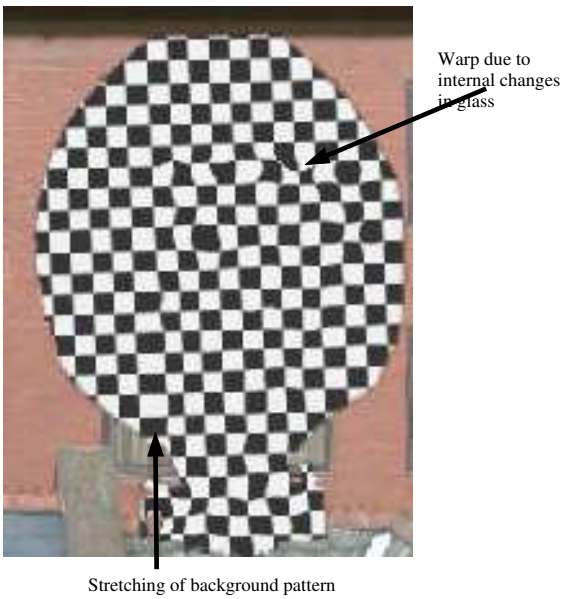


Figure 12: ***Zoomed view of checkerboard composite*** *from figure 10. Key features of the extracted matte are highlighted.*

the environment matte behaves as a small-displacement map over the image.

Approximate models of the environment matte may also be used to reduce the computational effort when building the matte. In building the environment matte for the magnifying glass example, the receptive fields were computed over 200×200 regions for each of the 26000 pixels within the lens. For each of these pixels, the cost of building the receptive field was approximately 200 milliseconds (in MATLAB on a 1GHz Pentium III). Thus, the time to compute the entire environment matte is of the order of hours. This time can be reduced if an approximate bounding box of the receptive field is available for each pixel. For example, manually indicating corresponding pixels in the composite and background images allows a coarse flow field to be constructed which is then interpolated using Gaussian radial basis functions to give an approximation to the mode of the receptive field for each pixel. Applying a generous bounding box to this region allows the receptive field to be computed in 40×40 regions—yielding a 25-fold speed improvement—with indistinguishable results. In this case, human effort is traded for machine time.

8. Further implementation details

There are two approaches commonly used to automatically compute homographies in the computer vision literature. The two methods are discussed in the articles by Irani and Anandan[15] and Torr and Zisserman[16].

In the first *direct* method a cost function is defined on the raw intensities, or on the intensities after filtering, for example a gradient filter. The cost function measures the correlation under the homography between the intensities in one image and the other. This cost is optimized over the 8 parameters of the transformation, and the optimization is implemented efficiently using a coarse to fine scale pyramid search.

In the second *feature based* method, interest points are computed in each image independently and the cost function is based on the distance between the points mapped under the homography. The correspondence of four or more points defines the transformation. The correspondences are determined and the cost function optimized using a robust statistical estimator based on the RANSAC principle[14].

In order to obtain the background homographies, we used interest-point matching[17] to get an initial dense set of feature tracks. Some examples are shown in figure 3. This gives initial homographies which are used to approximately align the images. The background plane homography is further refined using a direct minimization over image intensities[8, 18] with a robust kernel[19] on the intensity comparison. In this case the direct method produces a very good alignment of the background portion of the various images.

9. Examples

Examples of the performance of image-based environment matting are shown in high resolution in figures 11 and 12. In the first example the qualitative evaluation criterion is the spatial coherence of the final composite. In this case the result is good for the majority of pixels, but at a small number of pixels (one cluster is indicated with an arrow) the receptive fields are poorly estimated. A stage where such errors were manually indicated would allow the environment matte at such pixels to be interpolated, thereby improving the composite at the cost of a small amount of operator interaction.

The second example illustrates the composite of a checkerboard pattern between the ancient window and the original background. It demonstrates that even in difficult cases, image-based environment matting allows the convincing replication of physically complex light-transport systems, and that these systems can be measured directly from natural images even when calibration is unavailable or impossible.

10. Conclusions

The examples show that, although its performance is scene-dependent, the technique can work well given sufficiently rich backgrounds, or sufficiently many images. They demonstrate that environment mattes can be captured under less stringent assumptions than have previously been described.

The issue that has not been addressed here, as with the two-image calibrated techniques of Chuang et al[3], is diffuse scattering. In the proposed technique, this weakens the approximation used to estimate the per-image probability densities, and could lead to many more erroneous receptive fields. However, a number of strategies including multi-scale analysis and improved prior models may offer a solution to the problem.

The situation where the camera moves but the background is non-planar is also difficult with current technology. In the case of the planar background, the homography provides a strong constraint on the background motion and allows a clean plate to be extracted with relative ease. For a more general background, the technique must be robust to errors in the necessary dense stereo matching.

An interesting future application is the recovery of environment mattes from archive footage, for example scenes with fairground mirrors and moving cameras, or film of destroyed glass artefacts. This would allow reflections in curved mirrors—for example sunglasses—to be replaced in footage where calibration is no longer possible.

References

1. I. Wald and P. Slusallek. *State of the Art in Interactive Ray Tracing*. The Eurographics Association, 2001. 1
2. D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin. Environment matting and compositing. In *Proceedings, Siggraph*, pages 205–214, 1999. 1, 2, 3
3. Y.-Y. Chuang, D. E. Zongker, J. Hindorff, B. Curless, D. H. Salesin, and R. Szeliski. Environment matting extensions: Towards higher accuracy and real-time capture. In *Proceedings, Siggraph*, pages 12–130, 2000. 1, 2, 3, 6, 11
4. S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *SIGGRAPH96*, 1996. 2
5. M. Levoy and P. Hanrahan. Light field rendering. In *SIGGRAPH96*, 1996. 2
6. Y.-Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *CVPR 2001*, 2000. 2, 3
7. Y. Wexler, A. W. Fitzgibbon, and A. Zisserman. Bayesian estimation of layers from multiple images. In *Proc. European Conference on Computer Vision*. Springer-Verlag, 2002. To appear. 2, 3
8. M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision, Santa Margharita Ligure, Italy*, pages 282–287. Springer-Verlag, 1992. 2, 10
9. A. R. Smith and J. F. Blinn. Blue screen matting. In *Proc. SIGGRAPH*, pages 259–268, 1996. 2, 3
10. J.F. Blinn and M.E. Newell. Texture and reflection in computer generated images. *Communications of the ACM*, 19(10):542–547, 1976. 3
11. S. E. Chen. QuickTime VR — an image-based approach to virtual environment navigation. *Computer Graphics*, 29(Annual Conference Series):29–38, 1995. 3
12. M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, 1994. 4
13. J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994. 9
14. M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. Assoc. Comp. Mach.*, 24(6):381–395, 1981. 10
15. M. Irani and P. Anandan. About direct methods. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883 of *LNCS*, pages 267–277. Springer, 2000. 10
16. P. H. S. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In W. Triggs,

A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883 of *LNCS*, pages 278–294. Springer, 2000. 10

17. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521623049, 2000. 10

18. S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997. 10

19. P. J. Huber. *Robust Statistics*. John Willey and Sons, 1981. 10