Learning epipolar geometry from image sequences

Yonatan Wexler, Andrew W. Fitzgibbon and Andrew Zisserman
Visual Geometry Group,
Department of Engineering Science,
University of Oxford, United Kingdom
http://www.robots.ox.ac.uk/˜vgg

Abstract

We wish to determine the epipolar geometry of a stereo camera pair from image measurements alone. This paper describes a solution to this problem which does not require a parametric model of the camera system, and consequently applies equally well to a wide class of stereo configurations. Examples in the paper range from a standard pinhole stereo configuration to more exotic systems combining curved mirrors and wide-angle lenses.

The method described here allows epipolar curves to be learnt from multiple image pairs acquired by stereo cameras with fixed configuration. By aggregating information over the multiple image pairs, a dense map of the epipolar curves can be determined on the images. The algorithm requires a large number of images, but has the distinct benefit that the correspondence problem does not have to be explicitly solved.

We show that for standard stereo configurations the results are comparable to those obtained from a state of the art parametric model method, despite the significantly weaker constraints on the non-parametric model. The new algorithm is simple to implement, so it may easily be employed on a new and possibly complex camera system.

1. Introduction

This paper concerns the computation of epipolar geometry for stereo camera pairs. Recently many new types of camera configurations have been used which are generalizations of the conventional stereo rig with two pinhole cameras [4, 5, 6, 7, 9, 8, 10]. We are interested in such general cases and indeed in cases where a parametric model may not even be available. The approach developed here takes a fresh look at the two-camera configuration. It does not model the 3D geometric configuration of cameras, mirrors and lenses but rather learns the shape of the epipolar curves by accumulating matching evidence over multiple image pairs.

The advantages of this approach are several: it applies to any camera model, not only pinhole. This means, for example, that large radial distortion does not pose a problem and moreover, any optical configuration can be used. It is not a requirement that the epipolar curves be smooth, though we will assume that they are in the implementation here. Finally, the algorithm is extremely simple—it depends on having a reasonable model of image noise, but not on the complex bookkeeping strategies that characterize successful approaches to parametric epipolar geometry estimation. The primary practical contribution of this work is to allow automatic self-calibration of a range of stereo camera systems. Two benefits ensue: first, the method allows estimation of epipolar geometry when the parametric model is unknown, as for example with archive material. Second, even when the parametric model is known, the nonparametric algorithm may be a valuable preprocessing step.

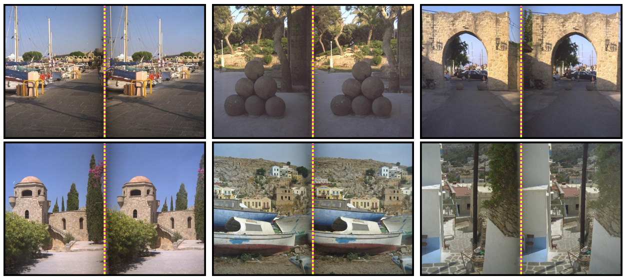


Figure 1: A collection of photos acquired from the same stereo camera. We seek a general way to learn the epipolar geometry of the camera from large sets of training images such as these, *without* an explicit parametric model of the camera system.

To expand on the latter point, it is worth briefly reviewing strategies for the estimation of epipolar geometry where parametric models are available. It is known that the epipolar geometry between two (classical) perspective pinhole cameras can be recovered using point matches [2]. The epipolar geometry is completely characterized by the fundamental matrix F, which can be computed from seven corresponding points in two images (yielding one or three

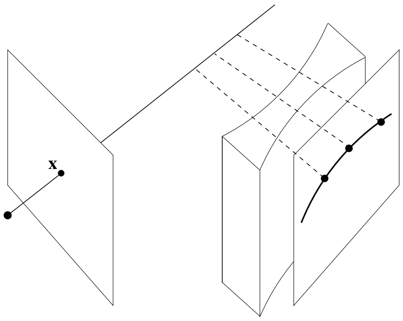


Figure 2: **A general stereo camera**. Each pixel on the left image samples light from points along a ray in space. For a given pixel, points along that ray project to some *eidolaktic set* of points in the right image. For many systems, the set of right-image points is a smooth curve, called the *epipolar curve*.

solutions). Successful techniques for estimating F from a pair of views generally have two key steps: first, pairs of left-right point correspondences are extracted from the images. Because the epipolar geometry is not yet known, these correspondences invariably include a significant number of false correspondences or *outliers*. In the second step, these outliers are identified by robust estimation of the parametric model (i.e. the parameters of the Fundamental matrix) [11, 15]. Finally, given this estimate of the epipolar geometry, a much more accurate set of correspondences can be found because the image region in which corresponding points must lie is reduced to a narrow band around the epipolar lines. Repeating the two steps refines the estimate.

The key to the parametric approach lies in ensuring that the first matching step produces enough good point correspondences that the parametric model may be accurately fitted. With non-pinhole cameras, however, this is often difficult. For example, fitting the catadioptric fundamental matrix [1] currently requires at least 15 correct correspondences, vastly increasing the combinatorics of the robust estimation algorithms. Other models [4, 6, 10] impose even more stringent requirements on the correspondences or acquisition methods, to the extent that for general cameras, no automatic system for self-calibration from image data is known to the authors.

The essential idea here is to learn epipolar curves directly from multiple image pairs acquired by a fixed configuration stereo system: consider determining the correspondence for a point in the left image from a single image pair – the true corresponding point in the right image will lie on the epipolar curve, but from any particular image pair many incorrect correspondences will be determined. However, when the correspondences from many image pairs are aggregated the incorrect correspondences will be scattered, but the true correspondences will cluster on the epipolar curve. The epipolar curve can thus be determined.

The most similar previous work is Triggs' joint feature distributions [12], which may be regarded as a parametric (in the statistical sense) analogue of our nonparametric model. The major advance of our work here is a simple robust algorithm to estimate the joint distributions of matching points in the left and right images, and demonstrations of recovered epipolar geometry on a range of central and non-central cameras.

The method is somewhat in the same spirit as [14] where many images were combined to compute an environment map which would be ambiguous if computed from a single image. It also has some similarity with the common technique for camera calibration where a dense sampling of the scene is obtained by imaging a bright light or retroreflective marker thousands of times. Beyond the obvious difference that the method described in this paper does not require a uniquely identifiable point in the scene, such techniques still require a parametric model as it is impractical to densely sample 3D space with a single point, while an image of a natural scene samples densely in two of the three dimensions.

2. The Model

We assume we have a stereo camera pair, which collects synchronized pairs of images, labelled "left" and "right".

Let $\mathbf{x}$ be some pixel coordinate in one of the images, as shown in Fig 2. The image at position $\mathbf{x}$ contains the projection of some world color onto the image plane. Assuming that the light travels in straight lines, the world point has to lie along the ray as shown. In a pinhole-camera stereo pair, the projection of the ray into the second image forms a line, called the epipolar line. When the cameras do not conform to the pinhole model the projection will not necessarily be a line. In general, we call the set of projections of the preimages of a point $\mathbf{x}$ the *eidolaktic set*[1]of $\mathbf{x}$. The task of this paper is to determine the mapping between points in each image of the stereo pair and their corresponding eidolaktic sets in the other. The eidolaktic set is neither necessarily one-dimensional nor infinite, but after describing the general model, this paper will concentrate on the case where it is 1D and smooth. In the case where the eidolaktic set is a smooth curve, we follow conventional nomenclature and call it the *epipolar curve* of $\mathbf{x}$.

Rather than define an explicit parametric representation for the curves, the eidolaktic set will be represented as an occupancy function over the pixels in the second image. Thus, each pixel $\mathbf{x}$ in one image has an associated value

[1] From eidolon ($\epsilon\iota\delta\hat{\omega}\lambda o\nu$, image reflected in mirror) and aktin ($\alpha\kappa\tau\iota\nu$, ray or beam). The difficulty with the word "epipolar" is that it implies two central projection cameras, which is a special case of the situation covered here.

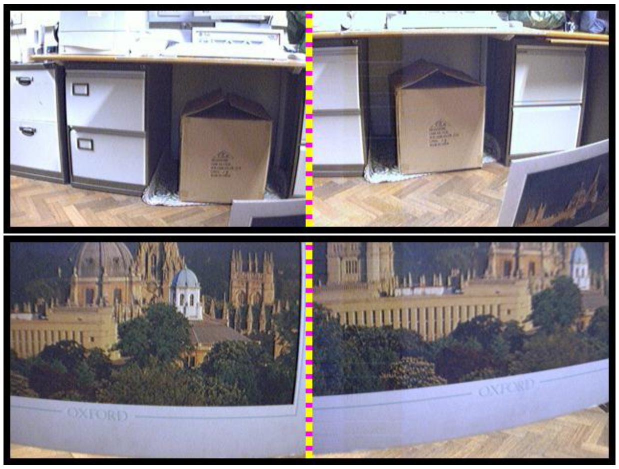


Figure 3: **Rig sequence.** Two of 220 stereo images (size 120x160) taken with two cheap PC cameras that are roughly synchronized. The cameras have substantial lens distortion and imaging noise.

$p(\mathbf{x}')$ in the other which, loosely speaking, describes the probability that pixels $\mathbf{x}$ and $\mathbf{x}'$ could be images of the same 3D point. The following sections show how a simple algorithm can be used to compute p given a set of images.

For the remainder of this discussion, we assume that each pixel $\mathbf{x}$ in the left image is treated independently, so throughout the following, $\mathbf{x}$ is fixed. A 3D point $\mathbf{X}$ represents a scene point with colour C. This point is imaged in the first camera at pixel position $\mathbf{x}$, with intensity $I(\mathbf{x})$. $I(\mathbf{x})$ is generally a 3-vector describing the pixel colour in some appropriate colourspace, for example CIE or HSV. The same point is imaged in the second camera at $\mathbf{x}'$, with intensity $I'(\mathbf{x}')$. With identical perfect cameras, and diffuse surfaces I and I' will be identical if $\mathbf{x}$ and $\mathbf{x}'$ are viewing the same point. In practice, because of distortions, noise in the imaging process, and non-diffuse surfaces, the imaged intensities will not be identical, but will be drawn from a joint probability distribution characteristic of the camera and class of scenes under investigation. Let $\nu(I, I') : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$ be the density function for this joint distribution, which we assume is available. Section 5 describes how ν is empirically computed for our experimental stereo setup (it is learnt in the manner of [3]). In the absence of empirical evidence, it may be appropriate to model the difference $\|I - I'\|$ as being Gaussian distributed with variance σ:

$$\nu(I, I') = Z_\nu \exp\left(-\frac{\|I - I'\|^2}{\sigma^2}\right)$$

where Z_ν is a normalizing constant to ensure that ν is a pdf. In addition, we assume knowledge of the prior joint density $\bar{\nu}(I, I')$ when I and I' are not the intensities of corresponding points. For quantized images, this density can often be modelled as uniform over the range of intensities, e.g.

$$\bar{\nu}(I, I') = \frac{1}{255^3}$$

Given these densities, we would like to determine the likelihood that a given pixel $\mathbf{x}'$ is the projection of a 3D point on the ray corresponding to pixel $\mathbf{x}$ in the left image. The ideal 3D ray is the set

$$R_x = \{X(\lambda), 0 < \lambda < \infty\}.$$

Parametrizing 3D space by the 2D coordinates of $\mathbf{x}'$ and the distance along the ray λ, we denote by $p(\mathbf{x}', \lambda)$ the likelihood that right-image pixel $\mathbf{x}'$ is the image of the 3D point $\mathbf{X}(\lambda)$. A given pair of images I, I' is observing a specific scene, and consequently we can compute this likelihood for the specific, but unknown, value of λ thus:

$$p_\lambda(\mathbf{x}') = \nu(I(\mathbf{x}), I'(\mathbf{x}')) + \bar{\nu}(I(\mathbf{x}), I'(\mathbf{x}')). \qquad (1)$$

Note that this is not a probability density, but merely the evaluation of a likelihood at each location in the right image. Because each pair of images gives us an estimate of $p(\mathbf{x}', \lambda)$ for a different value of λ, we can compute the sum over p_λ

$$p(\mathbf{x}') = \sum_{\lambda \in \Lambda} p_\lambda(x')$$

where Λ is simply the set of depths from which $I(\mathbf{x})$ was sampled over the training sequence. Implementation of this computation using a discrete sampling of pixels in both images leads to an algorithm to compute p, as will be described below.

3. The Algorithm

Given a set of stereo images, or a video sequence in which the world (depth map) changes smoothly, we wish to collect evidence about the eidolaktic set for each pixel. The input to the algorithm is a set of n stereo image pairs, $\{I_i, I'_i\}_{i=1}^n$. The desired output is the occupancy function $p(\mathbf{x}, \mathbf{x}')$ which encodes, for every pair of pixels in the left and right images, a measure of whether those two pixels are in corresponding eidolaktic sets. (It is clear that p is symmetric: if $\mathbf{x}'$ is in the eidolaktic set of $\mathbf{x}$, then the eidolaktic set of $\mathbf{x}'$, which is the set of pixels whose rays intersect the ray $R_{\mathbf{x}'}$, includes the pixel with ray $R_{\mathbf{x}}$, namely $\mathbf{x}$). The algorithm is described entirely in the left-right direction, computing $p(\mathbf{x}, \mathbf{x}')$ for a fixed value of $\mathbf{x}$. The other direction is directly analogous. The algorithm steps are summarized in figure 4, and we will illustrate these steps for the stereo configuration of figure 3.

First, each pair of training images gives an estimate of $p_i(\mathbf{x}, \mathbf{x}')$ for the particular scene depths in that pair. The estimate comes from application of the pixel similarity measure embodied in equation 1, and will have high values

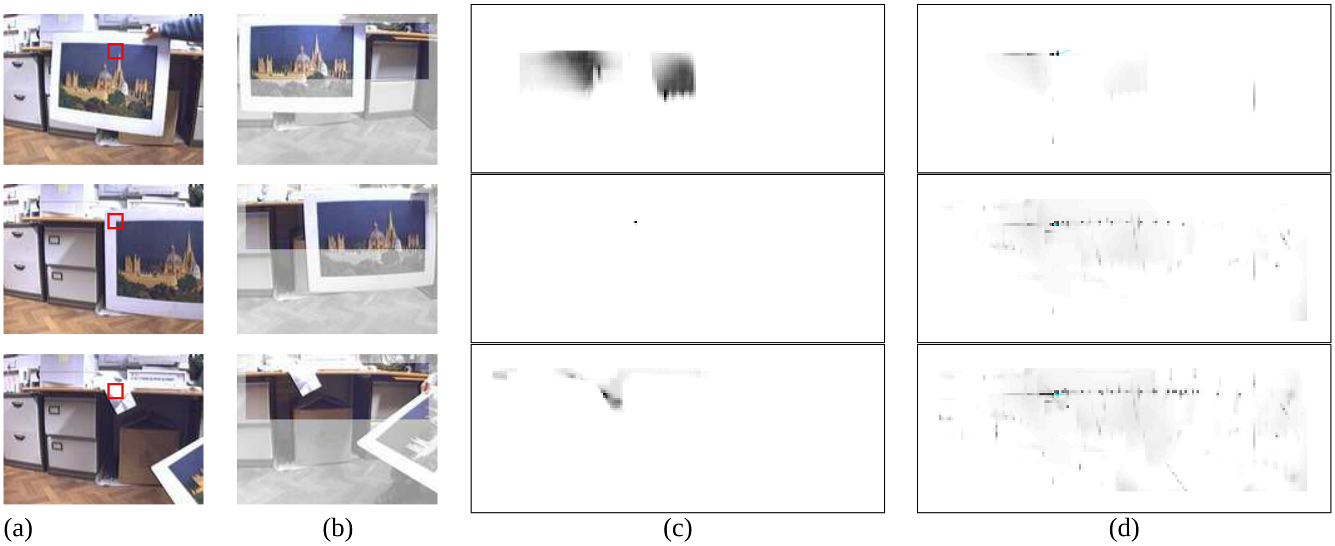


Figure 5: **Steps in the algorithm for one left-pixel** $\mathbf{x}$. For each row: (a) Pixel $\mathbf{x}$ marked in the left image. (b) Search region $\mathcal{S}$ highlighted in the right image. (c) Similarity function $p_i(\mathbf{x}, \mathbf{x}')$ plotted for $\mathbf{x}'$ in $\mathcal{S}$. (d) Accumulated eidolaktic occupancy function $p(\mathbf{x}, \mathbf{x}')$. The epipolar curve of $\mathbf{x}$ goes through the dark dots in the accumulator.

```
1. For each left-pixel $\mathbf{x}$
   (a) Initialize $P(\mathbf{x}') = 0 \quad \forall$ right-pixels $\mathbf{x}'$
   (b) For each input image pair $I_i, I_i'$
       i. For each right-pixel $\mathbf{x}'$
            Compute $p_i(\mathbf{x}')$ from eq.1(see §5)
       ii. Normalize so that $\sum_{\mathbf{x}'} p_i(\mathbf{x}') = 1$
       iii. Accumulate $P(\mathbf{x}') = P(\mathbf{x}') + p_i(x') \quad \forall x'$
   (c) Record $p(\mathbf{x}, \mathbf{x}') = P(\mathbf{x}') \quad \forall \mathbf{x}'$.
2. Impose smoothness priors $p(\mathbf{x}, \mathbf{x}') \approx p(\mathbf{y}, \mathbf{y}')$ for
   $\mathbf{y}, \mathbf{y}'$ near $\mathbf{x}, \mathbf{x}'$.
```

Figure 4: Algorithm to compute the occupancy function $p(\mathbf{x}, \mathbf{x}')$—a soft discretization of the indicator function which is 1 when $\mathbf{x}$ and $\mathbf{x}'$ are on corresponding epipolar curves, and 0 otherwise.

where image pixels have similar colours. See section 5 for precise details of its implementation. The third column of figure 5 shows the discretization of $p_i(\mathbf{x}, \mathbf{x}')$ computed over a region of interest in the second image. The sequence of images contains examples of frames where the matches for pixel $\mathbf{x}$ are highly ambiguous (large dark areas in the array), and where the match is unambiguous (a single black peak in the array).

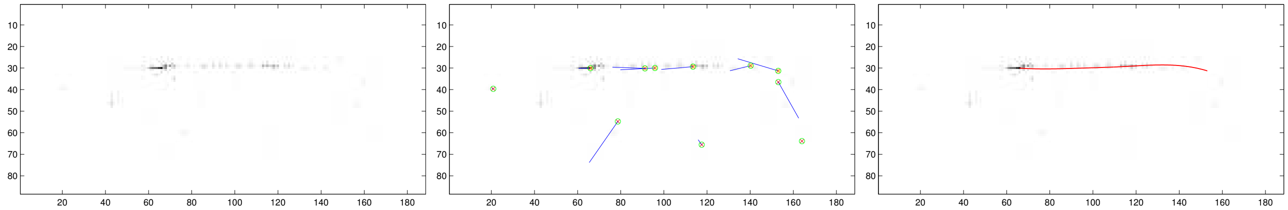
Second, at each step, the estimate of p_i is normalized to sum to one, and then added to the accumulator array $p(\mathbf{x}, \mathbf{x}')$. The key to the success of the algorithm lies in this aggregation step. False matches tend to be suppressed at this stage, due to the following observation. As the eidolaktic set is a one dimensional region in a two dimensional space (the range of p for fixed $\mathbf{x}$), correct matches will populate more densely than false matches. For example, if the length L of the epipolar curve is proportional to the image width, then each correct match has probability $\frac{1}{L}$ of accumulating whereas an erroneous match has probability $\frac{1}{L^2}$. This means that for sufficiently many training images, the eidolaktic set will be prominent in the image of $p(\mathbf{x}, \cdot)$. Figure 5, column (d) shows the evolving accumulator, with the epipolar curve (for this is a smooth camera) becoming clearly defined as more images are added.

After all images have been processed, for all left-pixels $\mathbf{x}$, the array $p(\mathbf{x}, \mathbf{x}')$ encodes the set of pixels where correspondence has been observed. For a dense sampling of space, this is the entire epipolar geometry of the system.

Two further observations about this process concern ambiguous pixels, and unmatched pixels:

- Because $\sum_{\mathbf{x}'} p(\mathbf{x}, \mathbf{x}')$ is normalized to unity at each step, pixels $\mathbf{x}$ which have many matching pixels in the right image will have a reduced contribution to the final estimate. Therefore ambiguous pixels, such as those in areas of low texture, are automatically downweighted without any artificial thresholding.

- A left pixel $\mathbf{x}$ may have no correct match in the right image due to occlusion or unmodelled specular lighting effects. If there is no strong incorrect match, $p(\mathbf{x}, \mathbf{x}')$ is dominated by pixels for which



(a) PDF collected from 200 images (b) Cluster centres and line directions. (c) Fitted eidolaktic curve.

Figure 6: **Imposing smoothness constraints**. Epipolar curve for one pixel of the "rig" sequence. (a) Shows the eidolaktic occupancy function for one pixel. (b) Shows the weighted-covariance cluster centers along with their principal directions. (c) A spline-fitted estimate of the epipolar curve.

the pixel similarity measure ν has very small values, and normalization might yield a pdf with randomly placed large peaks. However, the addition of the small "no-match" term $\bar{\nu}$ to every estimate means that everywhere-small estimates are normalized to an uninformative uniform distribution.

In the absence of further information, the above algorithm can determine calibration for arbitrary camera geometries extremely simply. However, in order to obtain a dense sampling of the eidolaktic sets, a large number of images are required. Each pixel must see a reasonable number of unambiguous colours at each depth λ in the stereo system's workspace. Although this is relatively easily achieved by waving a colourful poster in front of the stereo head for a minute, we should like to calibrate from fewer images of less constrained scenes, which is the subject of the next section.

4. Smoothness constraints

For many systems the eidolaktic sets are smooth epipolar curves, and the shapes of the epipolar curves vary slowly across the image. Integrating this constraint into the estimation allows the use of fewer images, with less informative texture, for the calibration.

The smoothness constraints fall into two classes: smoothness of the epipolar curve itself, and slow variation of the shape of the curve as the pixel (of whose ray it is the image) moves in the image plane.

Smoothness of the epipolar curve. For camera systems such as a smooth lens observing a smooth mirror, the image of a 3D ray will form a smooth curve in the image. Even for cameras which observe multiple mirrors, inducing cusps in the epipolar curves, the positions of the cusps are easily indicated, and are constant over the training set, so the curves are piecewise smooth. Therefore these curves can be extracted from the sparse eidolaktic occupancy function $p(\mathbf{x}, \mathbf{x}')$ by interpolating the occupancy function. Consider the sampled array of values of $p(\mathbf{x}, \mathbf{x}')$ for a given value of $\mathbf{x}$, and denote it $v(i, j)$. The strategy we have found effective is to fit lines in small windows of v, and use the line directions to build a directional "flow" field over the right image. Points on the lines are given by the weighted centroids $\mathbf{c}(i, j) = \sum_{(i',j')\in N(i,j)} v(i, j)(i, j)$, and the line directions $\mathbf{d}(i, j)$ are the short eigenvectors of the 2×2 weighted covariance matrices $\mathbf{\Lambda}(i, j) = \sum_{(i',j')\in N(i,j)} v(i, j)(i, j)^\top(i, j)$. Here the notation $N(i, j)$ denotes the set of pixels within a neighbourhood of (i, j). The size of the neighbourhood is chosen such that the eidolaktic set is expected to be sampled at least twice in the neighbourhood. Note that this does not assume video input, as the order of the frames doesn't matter here. The neighbourhood should be large enough to include many positive matches. Larger neighbourhoods mean more smoothness in the recovered curve but increased chance of including outliers. In the limit, N is the size of the image and the direction field is defined by one main axis (i.e. it is a straight line).

Given the direction field, there are at least two strategies to fit smooth curves. Splines may be fitted, as illustrated in figure 6, where the sparsely sampled eidolaktic occupancy function is converted into a smooth epipolar curve. On the other hand, if a complex parametric model of the system is available, which could not be fitted by RANSAC-like strategies because of a surfeit of degrees of freedom, it may well be sufficiently constrained by the flow field to be usefully estimated.

Smoothness in the image plane. A second class of smoothness constraint arises when it is known that adjacent image pixels sample from rays which are similar in space. In this case, the constraint we wish to maintain is that occupancy functions for adjacent left-pixels are similar. Specifically, if $\mathbf{x}$ and $\mathbf{y}$ are neighbours in the left image, then $p(\mathbf{x}, \cdot)$ and $p(\mathbf{y}, \cdot)$ should be similar 2D functions over the right image. A rigorous imposition of this constraint is somewhat complex, but a good approximation can be computed by taking the occupancy functions for all left-pixels, morphologically dilating, and using these as a prior distribution for neighbouring pixels in a second pass of the algorithm.

Temporal smoothness: Video sequences. The exposition to this point has taken no account of temporal dependence between the training images. This can help and hinder. It can help, because we may assume that good matches are temporally correlated, so that modelling that correlation will reinforce them, and that bad matches are sometimes uncorrelated, so they will fade. We achieve this by penalizing matches that have no temporal coherence. Video can hinder because frames which are too close may not provide a representative sampling over scene depths. In particular, the algorithm as presented here is sensitive to stationary scenes. If one frame appears 1000 times it will vastly dominate the output. Although this case is easily detected and corrected for, more subtle deviations from representative sampling may not be so easily accounted for. Note also that there is no independence assumption between frames, as the temporal process is integrating rather than multiplying the "probability densities" represented by the occupancy functions p_i.

5. Implementation

The main implementation issue is in the computation of $p_i(x, x')$. The important aspects are computation of $\nu(I, I')$ and windowing the response.

Evaluating pixel similarity. An important component of the algorithm as described is the evaluation of the pixel similarity. The joint density ν was determined empirically by manually verifying 281 point correspondences between the images of a captured stereo pair. The point correspondences yield corresponding RGB samples $\{C_i = (r_i, g_i, b_i), C'_i = (r'_i, g'_i, b'_i)\}_{i=1}^{281}$. From these correspondences, a Gaussian approximation to ν was computed. If using a Gaussian approximation, a good estimate for σ significantly reduces the amount of images needed for the algorithm. A value too big results in a blurry occupancy function whereas too small a value will reject all matches.

Neighbourhood selection. The second aspect of a successful implementation is in the use of a local window around each pixel $\mathbf{x}$ and $\mathbf{x}'$ in order to compute the match similarity. Two options in particular present themselves. The pixel intensity I can be computed as an average over a local neighbourhood, so that a low-pass filtered version of the image is used. Otherwise, similarity can be computed between windows centred on $\mathbf{x}$ and $\mathbf{x}'$, with appropriate invariances built in, as in wide-baseline stereo matching. In either case, careful modelling of the statistics of the similarity measure will improve performance.

6. Experiments

Comparison with parametric techniques. The first experiment is to compare epipolar geometry computed using the nonparametric approach to a conventional parametric method—robust fitting of a parametric model method using only one image pair. A camera was placed on a mirror and a sequence captured while a textured surface was moved in front of the system. The fundamental matrix is computed from about 250 point correspondences manually indicated. The points were chosen to lie on at least two planes in the scene. The resulting epipolar lines agreed with the fitted points to an RMS of 0.6 pixels, and an example of a point and its corresponding epipolar curve is shown in figure 7. The learnt curve from the nonparametric algorithm is shown in the lower half of figure 7, and demonstrates agreement with the parametric model to within about 0.5 pixels. This is an impressive result as it shows that the nonparametric model with 20 images can compete with the parametric model. Of course, the scene contains ideal random texture, so is a good case for the nonparametric algorithm, but the implications for more general situations remain positive.

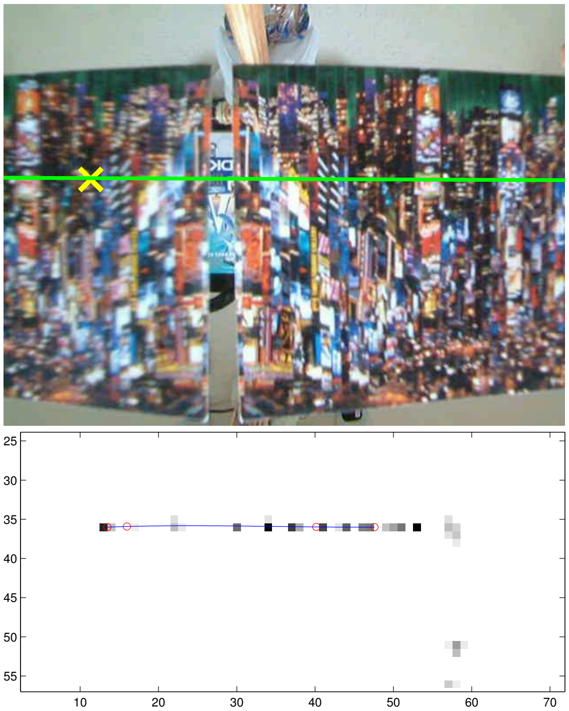


Figure 7: **Comparison with conventional Fundamental matrix model**. (a) A left image point and its corresponding epipolar line in the right image obtained by fitting the Fundamental matrix to point correspondences. (b) Zoomed view of the PDF with the fitted eidolaktic curve.

A second example in which a conventional stereo rig was used is illustrated in figure 9. In this case, 215 stereo pairs were captured by a hand-held stereo camera in an apartment building. Two epipolar curves are shown for two points, showing again that the technique accurately models the oc-

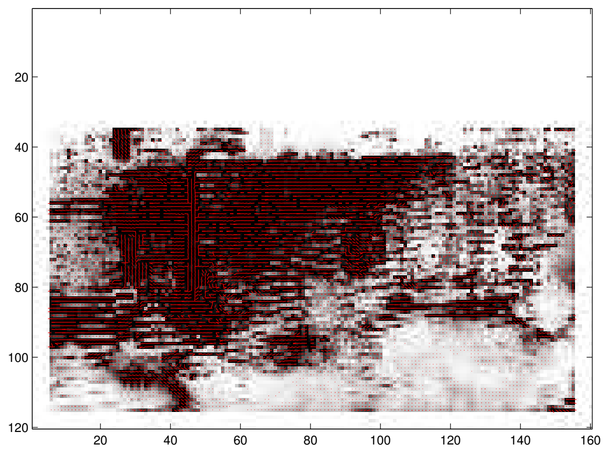


Figure 8: Dense epipolar direction field computed for “rig” sequence.

cupancy function of the joint image space.

This establishes that the method produces epipolar curves of satisfactory quality for conventional stereo rigs with little radial distortion. Its power is exemplified in the following two cases where more perverse camera configurations appear.

Stereo rig with radial distortion. This is the case dealt with in figure 6 for the stereo rig sequence of figure 3. The camera pair has significant radial lens distortion and the recovered epipolar curve models this reasonably accurately. In this case the spline model is probably too general, as the kink in the curve in the figure indicates. However the recovered geometry is more than adequate to guide a point-based matching algorithm, greatly improving its reliability. Figure 8 shows the dense direction field for a section of the right image. The motion was computed over the area marked in figure 5(b) and the resulting tangent fields (directions of the epipolar curves) are shown overlaid on the confidence collected from the algorithm where dark regions denote more confident areas. Most computed tangents are correct though there are regions that are wrong due to the flat areas of the photograph and the constant background. Future work is to apply a restoration algorithm to this field to obtain a smooth set of epipolar curves.

Stereo rig and spherical mirror. This is an example of a camera system where the cameras do not have well defined centres of projection. Figure 10 shows the various stages of the algorithm in operation on this sequence. As reflection off the sphere shrinks the image, there is little parallax which makes it hard to recover the correct epipolar direction. Figure 10 shows the computation of one pixel from only 30 images. At the 10th frame, the match measure is ambiguous and is thus smeared. At the 20th frame it is still ambiguous but now the ambiguity is over different regions. After adding 30 frames we get a sharp peak with the correct position. The small parallax is enough to get an estimate of the direction, even for this short sequence.

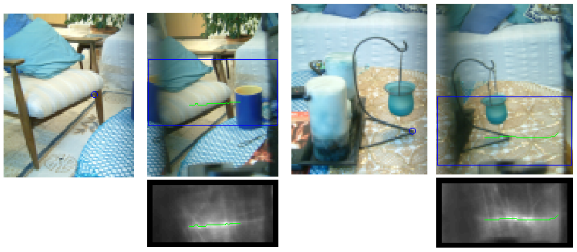


Figure 9: **Room Sequence**. Two frames from 215 captured in a domestic setting. In each case a point in the left image is shown in blue and a curve of local maxima of the eidolaktic occupancy function is highlighted in green in the right. The blue rectangle is the region in which the occupancy function was computed.

7. Discussion

This paper has described how the epipolar geometry for a stereo rig can be learnt from a set of captured image pairs. The novelty of the approach is that no parametrized model of the system geometry is used—the epipolar curves (or their generalization, the *eidolaktic sets*) are represented as a probabilistic occupancy grid over the *joint image* [13]. In one sense, we have discovered the embedding of 3D space in $\mathbb{R}^4$. The primary implication of the work, apart from satisfaction of scientific curiosity, is the simplification of point matching for non-pinhole camera configurations—because the search space for point correspondences is limited to the 1D eidolaktic set, the correspondence problem is drastically simplified. Because the model need not be known, the process can apply to archive footage of stereo sequences, for example, sports events.

The technique depends on having a large number (tens to hundreds) of images in order to obtain a representative calibration, which may not be possible in some environments. The paper has discussed some techniques for imposing various types of smoothness on the recovered geometry, in order to reduce the number of images required, and more work on these might be a profitable area of endeavour.

One feature of the procedure is that the recovered epipolar curve may be finite in length in the image. This occurs for two reasons. First, the training sequence may not contain a representative sample of the depth range. Second, even with a representative sample, the epipolar curve may be finite in length. Consider railway tracks going into the distance. The infinite line from “here” to the horizon projects to a finite line segment in the image. For many camera configurations, this is the case, and our algorithm

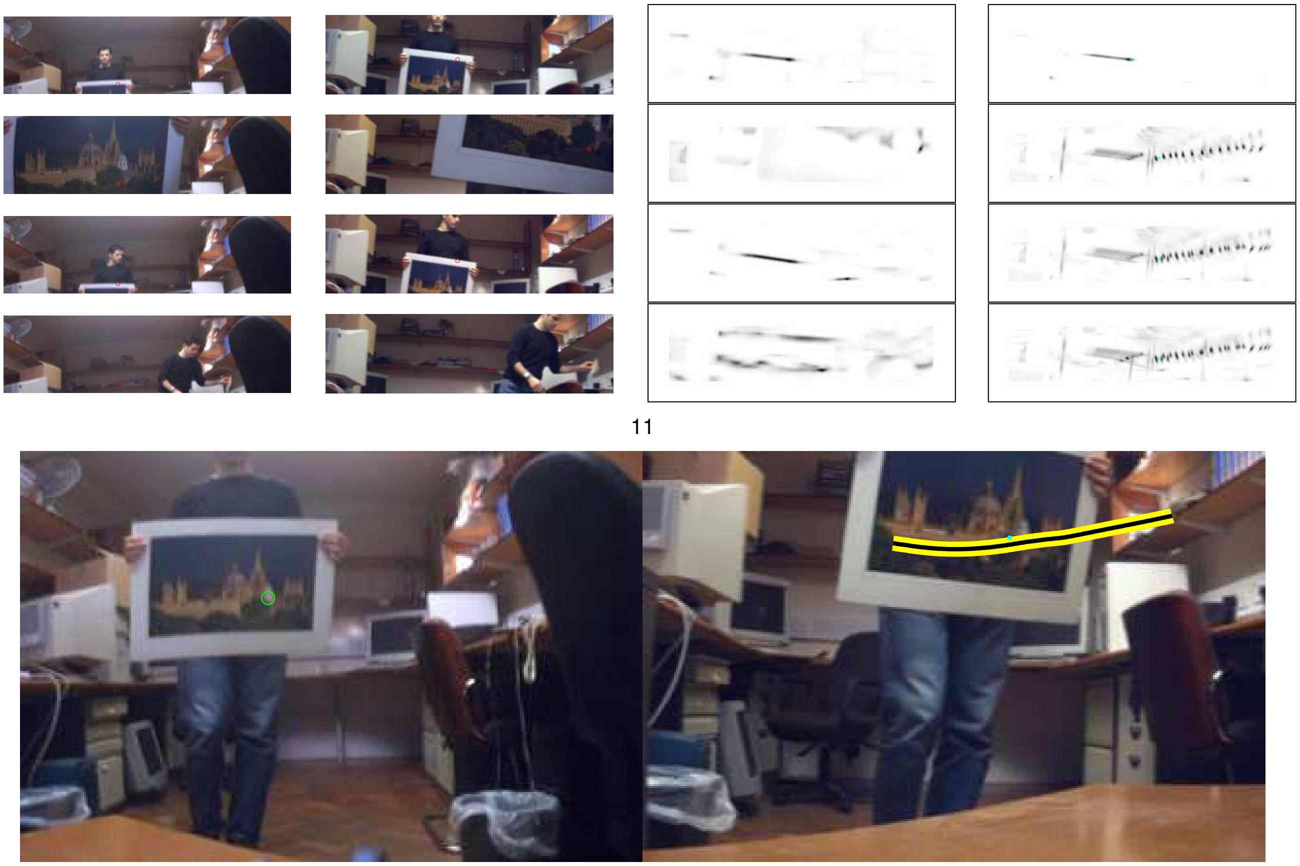


Figure 10: (Top) Algorithm applied to Sphere sequence. (Bottom) Recovered epipolar curve for one point.

will identify only the finite extent of the line. On the other hand, if the output is to be used as a matching constraint it is sensible to restrict search to the areas where matching points can be found.

Acknowledgements

Funding for this work was provided by the DTI/EPSRC Link grant V2I. AF would also like to thank the Royal Society for its generous support.

References

[1] C. Geyer and K. Daniilidis. Structure and motion from uncalibrated catadioptric views. In *Proc. CVPR*, pages 279–286, 2001.

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

[3] S. Konishi, A. Yuille, J. Coughlan, and S. Zhu. Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues. In *Proc. CVPR*, pages 573–579, 1999.

[4] S. A. Nene and S. K. Nayar. Stereo with mirrors. In *Proc. ICCV*, pages 1087–1094, 1998.

[5] S. Peleg and M. Ben-Ezra. Stereo panorama with a single camera. In *Proc. CVPR*, 1999.

[6] R. Pless. Discrete and differential motion constraints for generalized cameras. In *Workshop on Omnidirectional Vision*, 2002.

[7] S. Seitz. The space of all stereo images. In *Proc. ICCV*, pages 26–33, 2001.

[8] H. Y. Shum, A. Kalai, and S. Seitz. Omnivergent stereo. In *Proc. ICCV*, 1999.

[9] H. Y. Shum and R. Szeliski. Stereo reconstruction from multiperspective panoramas. In *Proc. ICCV*, 1999.

[10] Tomás Svoboda, Tomás Pajdla, and Václav Hlavác. Epipolar geometry for panoramic cameras. In *Proc. ECCV*, pages 218–232, 1998.

[11] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.

[12] B. Triggs. Joint feature distributions for image correspondence. In *Proc. ICCV*, pages 201–208, 2001.

[13] W. Triggs. The geometry of projective reconstruction I: Matching constraints and the joint image. In *Proc. ICCV*, pages 338–343, 1995.

[14] Y. Wexler, A. Fitzgibbon, and A. Zisserman. Image-based environment matting. In S. Debevec, P. Gibson, editor, *Proc. EuroGraphics Workshop on Rendering*, pages 289–299, Pisa, Italy, June 26–28 2002. Eurographics / ACM Siggraph.

[15] Z. Zhang, R. Deriche, O. D. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, 1995.