In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Santa Barbara, CA, pages 442-447, 1998.

A Statistical Framework for Long-Range Feature Matching in Uncalibrated Image Mosaicing

Tat-Jen Cham* and Roberto Cipolla

Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, England

Abstract

The problem considered in this paper is that of estimating the projective transformation between two images in situations where the image motion is large and feature-matching is not aided by a proximity heuristic. The overall algorithm designed is based on a multiresolution, multi-hypothesis scheme, and similarities between tracking and matching through multiple resolution levels are exploited. Two major tools are developed in this paper: (i) a Bayesian framework for incorporating similarity measures of feature correspondences in regression to specify the different levels of confidence in the correspondences; and (ii) a Bayesian version of RANSAC, which is able to utilise prior estimates and matching probabilities. The algorithm is tested on a number of real images with large image motion and promising results were obtained.

1 Introduction

Most of the present work on image mosaicing has been based on the assumption of small image motions, which significantly helps the recovery of good correspondences. For example the work carried out by Sawhney *et al.* [8] and Szeliski[9] are based on motion estimation from video sequences. The exceptions which do not require small image motion are Dani and Chaudhuri's paper[2] which is limited to separate cases of translational and rotational mosaicing, and the use of corner models for projective matching by Zoghlami *et al.* [13]. In the latter case, the matching is based on extracting a very small number of evenly-distributed corner points and computing the transformation hypotheses from pairs of correspondences. However the corners are often extracted from regions which have weak features (to maintain an even distribution) and are very susceptible to small non-linear changes in image intensities. Furthermore the transformation estimation is only based on two features without incorporating other matched features.

*Author's current affiliation is Cambridge Research Lab, Digital Equipment Corporation, Cambridge MA.

The problem which is to be addressed is that of constructing a mosaic of images which are related by the class of *projective transformations* (eg. taken from an uncalibrated rotating projective camera). Some examples of the images used are shown at the end of the paper.

The large differences in the viewing directions of the images indicate that a feature-based approach is preferred. However, most stereo or mosaicing algorithms (eg. [11, 10, 6, 9]) require the proximity criterion because of its strong advantages in matching.

- *Proximity is a very powerful criterion in eliminating mismatches when there are a large number of features*. Without this criterion, it is wrong to assume that at least half of the candidate correspondences are true. These are assumptions made in least-squares and least-median-squares[5] methods.

- *The proximity criterion significantly reduces search time*. Without this criterion, a feature selected from one image must be tested against all other features in the second image. This results in a very intensive search operation in order to recover good correspondence candidates.

However, the proximity criterion cannot be applied when image motion is large, as in the case in the type of images we want to be able to cope with. **In this paper, we adopt a number of strategies to cope with long-range feature matching**:

- **A Multiresolution, Multi-Hypothesis Approach**. A multiresolution scheme is desirable from the point of view of recovering the *globally* optimal solution[12] as well as in terms of search efficiency. While a large number of papers make use of a multiresolution scheme (eg. [11]), a rudimentary, blanket approach is usually adopted without further analysis. In this paper we consider similarities between multiresolution estimation and tracking — the former involves computing over a number of resolution levels and the latter

involves computing across a number of image frame sequences.

- **Robust Regression**. The use of similarity measures without a proximity criterion for determining matches often leads to a considerable proportion of correspondences being false matches. Hence the estimation method used must be robust to outliers. The RANSAC[3] paradigm is typically proposed, but in its original form is only suited for non-iterative estimation. A Bayesian extension to the RANSAC to cater for multiresolution, iterative schemes is described in section 2.4.

- **Incorporating Similarity Measures in Regression**. In most previous applications, similarity measures such as the cross-correlation values were simply compared against a preset threshold to decide if the candidate match can be admitted, and thereafter ignored. However since the similarity measures also indicate the amount of trust in the correspondences admitted, we show how they can be incorporated into a statistical framework for estimating the transformation in section 2.3.

2 Theory

2.1 Iterative Estimation through Multiple Resolution Levels

One of the key problems in mosaicing two images I_l and I_r obtained from a rotating uncalibrated camera is that of accurately determining the unknown projective transformation mapping I_r to I_l. The transformation, specified by some parameter vector $\boldsymbol{A}$, may be expressed in vector form via $\boldsymbol{x}_l = \boldsymbol{T}(\boldsymbol{A}, \boldsymbol{x}_r)$ where $\boldsymbol{x}_l$ and $\boldsymbol{x}_r$ are vector coordinates of corresponding image points in I_l and I_r, and $\boldsymbol{T}$ is a matrix function of $\boldsymbol{x}_r$ and $\boldsymbol{A}$ which maps $\boldsymbol{x}_r$ to $\boldsymbol{x}_l$.

A multiresolution approach to the estimation of transformation parameter vector $\boldsymbol{A}$ involves using data obtained from a low resolution level to compute a weak hypothesis for $\boldsymbol{A}$ and refining the hypothesis by working through increasing levels of resolution. Unlike estimation carried out at a single resolution level, the data is utilised in a sequential manner, which is similar to the scenario encountered in tracking.

Suppose the resolution levels are $r = 1, \ldots, R$ with increasing resolution. Let the data obtained at resolution r be denoted by a vector $\boldsymbol{z}_r$, and the cumulative data obtained from resolution levels 1 to r be denoted by vector $\boldsymbol{Z}_r = [\ \boldsymbol{z}_r\ \boldsymbol{Z}_{r-1}\ ]^T$. When more data becomes available, we would expect the range of probable solutions to become smaller, ie. that the posterior probability density $p(\boldsymbol{A}|\boldsymbol{Z}_r)$ becomes more concentrated in a number of peaks.

Consider a peak in $p(\boldsymbol{A}|\boldsymbol{Z}_r)$ representing the most probable hypothesis. When advancing from one resolution level to the next, *three* possible types of transitions may occur at the peak in $p(\boldsymbol{A}|\boldsymbol{Z}_{r+1})$ (see figure 1), for which different estimators are suited for determining the form of $p(\boldsymbol{A}|\boldsymbol{Z}_{r+1})$:

1. *The peak becomes more pronounced*.
   In this case the initial estimation is accurate, and there is only one dominant hypothesis. The correspondences obtained at the lower resolution level is sufficient to localise the initial estimate in the vicinity of the correct solution, and a method based on iterative (ie. sequential) least-squares may be used to refine the estimate at the higher resolution level.

   Preferred estimator: Kalman filter.

2. *The peak diminishes, together with the formation of other new peaks*.
   In this case the initial estimation is erroneous, due to a lack of good correspondences at the lower resolution level. The estimation process must be re-applied to correspondences obtained at the higher resolution level. Since there may be a number of probable alternate solutions, a method capable of handling multiple hypotheses is desired.

   Preferred estimator: standard RANSAC.

3. *The peak divides into a number of sub-peaks*.
   The initial estimate is accurate as to the *region* of the correct solution, but is insufficient to distinguish between separate hypotheses within the region. In this case, a Kalman filter will fail to distinguish between the separate hypotheses, while a traditional RANSAC method operating on the data from the higher resolution level will be computationally inefficient in having to test a large number of improbable solutions as well. The answer is to incorporate Bayesian analysis into RANSAC to improve hypothesis testing (section 2.4).

   Preferred estimator: Bayesian RANSAC.

2.1.1 Multiple-Hypothesis Search Tree

The difficulty lies in deciding which of the three cases best fits the current state of an estimation process, and therefore which estimator to use. Selecting standard RANSAC is the most non-committal decision, but which is also the most computationally intensive. On the other hand, the Kalman filter is highly efficient but may result in an incorrect solution.

In our case, a *multiple-hypothesis approach* is adopted. We select the estimator according to computational efficiency (ie. Kalman filter, Bayesian RANSAC, standard RANSAC in that order), but retain a *hypothesis tree* of different states reached in the estimation process. Each leave

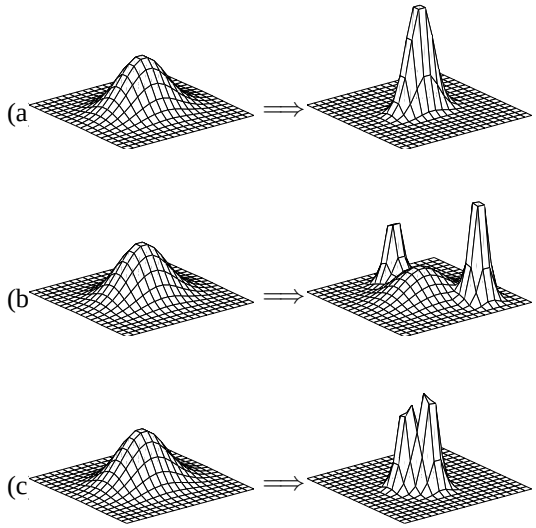


Figure 1: The figures show the 3 categories of transitions to a probability peak which can occur in the posterior probability density function $p(\boldsymbol{A}|\boldsymbol{Z}_r)$ (assumed 2D in the examples) when advancing from a lower resolution level to a higher level. In (a), a Kalman filter is preferred for estimating the optimal solution; in (b), standard RANSAC is recommended; in (c) we propose the use of Bayesian RANSAC. See the text for more details.

of the hypothesis tree contains the following information: (1) the current $p(\boldsymbol{A}|\boldsymbol{Z}_r)$; (2) the last estimator attempted, if any; (3) the transformation model (either similarity, affine or projective); (4) the resolution level r; and (5) a measure ρ quantifying the similarity between the images I_l and a $\boldsymbol{A}$-warped version of I_r. A *best-first search*[7] based on ρ is then carried out until an acceptable final solution is reached.

2.2 The Standard Bayesian Framework

The standard Bayesian approach to stereo feature matching normally establishes Bayes relation between posterior and prior probabilities of the model parameters. We express this in recursive form for a multiresolution framework:

$$p(\boldsymbol{A}|\boldsymbol{Z}_r) = k\, p(\boldsymbol{z}_r|\boldsymbol{A})\, p(\boldsymbol{A}|\boldsymbol{Z}_{r-1}) \tag{1}$$

where $\boldsymbol{A}$ represent the model parameters and k is a normalisation constant. $\boldsymbol{Z}_r$ and $\boldsymbol{z}_r$ are as defined in the previous section. This is also widely used in tracking which involves sequential data.

In the case of stereo matching, this is usually implemented by (i) recovering features in the stereo images, (ii) using the prior probability $p(\boldsymbol{A}|\boldsymbol{Z}_{r-1})$ to establish search regions for matching features in the different images, followed by (iii) evaluating the posterior probability $p(\boldsymbol{A}|\boldsymbol{Z}_r)$ using the likelihoods $p(\boldsymbol{z}_r|\boldsymbol{A})$ computed from any new features found within the search regions.

The usual forms of implementation suffers from a number of problems:

1. The likelihood functions used are usually *solely* dependent on the distances between the observed positions of the features and the positions predicted from the model parameters. Measures such as neighbourhood intensity correlation values are not taken into account in the likelihood estimation. An exception to this is the similarity weighting scheme described in [4], where the similarity measure is based on estimated image noise.

2. Once features are found in the search regions (and which perhaps exceed some correlation threshold), they are usually treated as having been correctly matched.

2.3 Extending the Bayesian Framework

Suppose a number of features are extracted from two images and paired as $(f_{li}, f_{ri}), i = 1, \ldots, N$, where the pairing need not necessarily be a one-to-one mapping. The features may also be of various classes (eg. corners, lines, regions), and/or obtained from different resolution levels of the images. We further define a number of functions based on these pairings:

1. $\boldsymbol{X}_i$ represents the augmented vector containing the true (noiseless) image coordinates of the features f_{li} and f_{ri}, given by $(x_{li}, y_{li}), (x_{ri}, y_{ri})$:

$$\boldsymbol{X}_i = [\, x_{li}\ y_{li}\ x_{ri}\ y_{ri}\,]^T \tag{2}$$

   Due to noise, $\boldsymbol{X}_i$ cannot be perfectly recovered, only estimated. The corresponding estimation of $\boldsymbol{X}_i$ is denoted by $\hat{\boldsymbol{X}}_i$.

2. m_i is a boolean flag indicating the truth of the correspondence hypothesis between the features f_{li} and f_{ri}:

$$m_i = \begin{cases} 1 & \text{if } f_{li}, f_{ri} \text{ linked to the same 3D feature;} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

   m_i can only be recovered if the parameter vector $\boldsymbol{A}$ is known exactly.

2.3.1 Incorporating Similarity Measures

Consider the simple case where there is only a single pair of features forming a candidate correspondence. Using the above functions, (1) can be expanded in the following way:

$$\begin{aligned} p(\boldsymbol{A}|\hat{\boldsymbol{X}}_1) &= k\, p(\hat{\boldsymbol{X}}_1|\boldsymbol{A})\, p(\boldsymbol{A}) \\ &= k\, \Big\{ p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1{=}1)\, p(m_1{=}1|\boldsymbol{A}) + \\ &\qquad p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1{=}0) p(m_1{=}0|\boldsymbol{A}) \Big\}\, p(\boldsymbol{A}) \end{aligned} \tag{4}$$

By expressing the equation in this form, the likelihood function $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A})$ is separated into distinct components:

- The component $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 1)$ represents the case where the features f_{l1} and f_{r1} have been *correctly* matched (ie. they arise from the same 3D feature). This is normally assumed when Kalman filters are employed. The variance of $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 1)$ only arises from positional uncertainty.
- The component $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 0)$ represents the case where the features f_{l1} and f_{r1} have been *wrongly* matched. We would generally expect this to be a uniform function.
- The above above two components are *mixed* according to the probabilities of either case being true, $p(m_1 = 1|\boldsymbol{A})$ and $p(m_1 = 0|\boldsymbol{A})$.

The advantage in the form of (4) is that the probabilities of matching $p(m_1 = 1|\boldsymbol{A})$ can be specified. For example in many algorithms, the cross-correlation of image intensity is normally used as a similarity measure for making binary decisions on accepting or rejecting the correspondence. This implicitly assumes that the the probability $p(m_1 = 1|\boldsymbol{A})$ is a step function of the cross-correlation measure, with the step occurring at a predefined threshold. *We can instead estimate $p(m_1 = 1|\boldsymbol{A})$ as a smooth function dependent on the similarity measure*. Hence we continue to distinguish between correspondences with different probabilities throughout the estimation process.

In our case, we define the similarity measure S between image patches Ψ_l and Ψ_r in the sense of a signal-to-noise ratio

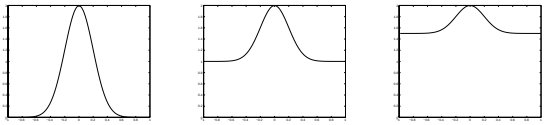
$$S = \frac{\sqrt{\text{Var}(\Psi_l)\ \text{Var}(\Psi_r)}}{\sum \left[\Psi_l(\boldsymbol{x}) - \Psi_r(\boldsymbol{T}(\boldsymbol{A}, \boldsymbol{x}))\right]^2} \tag{5}$$

and assign the probability $p(m_1 = 1|\boldsymbol{A})$ as

$$p(m_1 = 1|\boldsymbol{A}) \approx \begin{cases} 1 - \frac{S_t}{S} & \text{if } S \geq S_t \\ 0 & \text{if } S < S_t \end{cases} \tag{6}$$

where S_t is a conservative predefined cutoff threshold for the similarity measure such that the candidate match will not be included in the regression process if $S < S_t$. If $\boldsymbol{A}$ is unknown, Ψ_r is rotated such that the dominant image gradient coincides in direction with that of Ψ_l.

Figure 2 shows the variation of $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A})$ for a 1-parameter $\boldsymbol{A}$ when different $p(m_1 = 1|\boldsymbol{A})$ are used, assuming that $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 1)$ is Gaussian. Alternatively, it is also possible to approximate $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A})$ with a Gaussian $N(\boldsymbol{M_c}, \boldsymbol{S_c})$ if required. If we define the Gaussian parameters of $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 1)$ according to $N(\boldsymbol{M_d}, \boldsymbol{S_d})$, then an approximation of $\boldsymbol{S_c} = \boldsymbol{S_d}/p(m_1 = 1|\boldsymbol{A})$ may be used in the vicinity of the mean value $\boldsymbol{M_d}$.



(a) (b) (c)

Figure 2: The shape of the likelihood function $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A})$ is a mixture of a Gaussian representing $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 1)$ and a uniform function representing $p(\hat{\boldsymbol{X}}_1|\boldsymbol{A}, m_1 = 0)$. The shape varies according to the matching probability $p(m_1 = 1|\boldsymbol{A})$, which is 1 in (a), 0.5 in (b), and 0.25 in (c).

Consider the example shown in figure 3. The correspondences drawn with darker shades in figure 3(b1),(b2) are treated as more probable. If similarity measures were ignored, the two false matches in figure 3 will be allocated the same significance as the correct matches — figure 3(c) shows the poor mosaicing result of applying least-squares estimation to all the correspondences without regard to the matching probabilities. The estimation can be improved if correspondences with higher probability will be allocated greater significance, the result of which is shown in figure 3(d).

2.4 Bayesian RANSAC

The RANSAC paradigm[3] bears resemblance to the multi-hypothesis methods[1] used in tracking, in that different hypotheses are tested in order that false matches may be effectively excluded, which is not possible with Kalman filter methods. However, one of the shortcomings of the standard RANSAC paradigm is that the estimation is based solely on data present, since *the transformation hypotheses generated from the bases do not take into account apriori information on the transformation to be estimated*. Generally noise in the feature positions can have a significant effect on the hypothesized transformation if a minimal set of correspondences are used, particularly if the feature configurations are poor or degenerate. This may be so even if all the correspondences are inliers. As a result, more basis sampling may have to be carried out for successful robust regression.

On the other hand if a Bayesian approach is established, the accuracy of the generated hypotheses can be improved through the use of the apriori probability $p(\boldsymbol{A})$ by making the following modifications:

- Instead of randomly sampling the dataset with equal probability, each feature pair (f_{li}, f_{ri}) is assigned a probability $p(m_i = 1)$ of being sampled. Hence the pairs of features which are more likely to be correctly matched will be selected with greater frequency.
- The transformation parameters $\boldsymbol{A}$ for a single basis are normally computed without taking into account the apriori probability density $p(\boldsymbol{A})$. If $p(\boldsymbol{A})$ has a

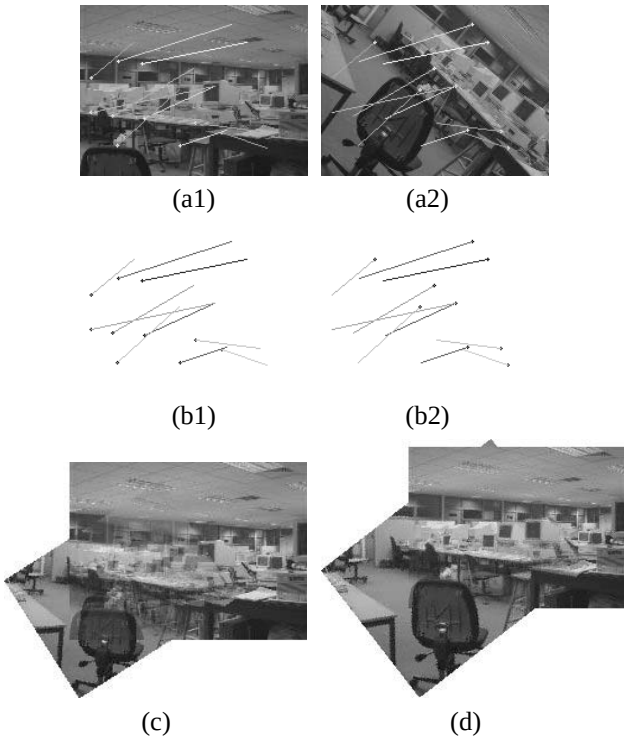


Figure 3: (a1),(a2) show the correspondences initially obtained between the two images. (b1),(b2) represent the matching probability of the correspondence according to the darkness of the shade. (c) shows the inaccurate registration computed while assuming all the correspondences are equiprobable, while (d) shows the more accurate registration computed if correspondences are assigned significances according to their matching probabilities.

strong bias, it is better to choose the MAP estimate for $\boldsymbol{A}$:

$$p(\boldsymbol{A}|\{\hat{\boldsymbol{X}}_i, i = 1, .., N\}) = k\ p(\{\hat{\boldsymbol{X}}_i, i = 1, .., N\}|\boldsymbol{A})\ p(\boldsymbol{A}) \quad (7)$$

where N is the number of features in the basis, and the likelihood $p(\{\hat{\boldsymbol{X}}_i, i = 1, .., N\}|\boldsymbol{A})$ may be estimated from the observed positions of the features.

In our experiments, we make the assumption that in the vicinity of each probable hypothesis, both the likelihood $p(\{\hat{\boldsymbol{X}}_i, i = 1, .., N\}|\boldsymbol{A})$ and probability density $p(\boldsymbol{A})$ may be approximated by Gaussians, and have means $\boldsymbol{M_l}$, $\boldsymbol{M_A}$ and covariances $\boldsymbol{S_l}$, $\boldsymbol{S_A}$ respectively. The likelihood mean $\boldsymbol{M_l}$ is the transformation parameters computed directly from the basis features, while the likelihood variance $\boldsymbol{S_l}$ may be estimated from image gradients and estimated noise variances ([4] provides detail on the computation).

Similarly the posterior probability density $p(\boldsymbol{A}|\{\hat{\boldsymbol{X}}_i, i = 1, .., N\})$ is also Gaussian, with a mean of $\boldsymbol{M_p}$ and covariance $\boldsymbol{S_p}$ which can be calculated from $\boldsymbol{S_p}^{-1} = \boldsymbol{S_l}^{-1} + \boldsymbol{S_A}^{-1}$ and $\boldsymbol{M_p} = \boldsymbol{S_p}\ (\boldsymbol{S_l}^{-1}\boldsymbol{M_l} + \boldsymbol{S_A}^{-1}\boldsymbol{M_A})$ In Bayesian RANSAC, $\boldsymbol{M_p}$ is used as the transformation hypothesis, as opposed to the use of $\boldsymbol{M_l}$ as in standard RANSAC.

3 Experimental Results

We tested the algorithm on a number of images which are taken both with a digital camera (indoor scenes) and a disposable camera. In the latter case as we do not model the radial distortion, the image registration is not totally precise, although gross registration is achieved. Computation was carried on a multi-user Sun UltraSparc 1 workstation using the matrix manipulation package Octave.

Figures 4,5 and 6 show mosaics constructed from the original images (top). The mosaics shown in figures 4 and 5 took approximately 12 and 10 minutes each, because the RANSAC algorithms were almost always activated because the lack of radial distortion adjustment (we were using a disposable camera after all!) and uneven film development (resulting in nonlinear changes in image intensity) meant that the average similarity measures computed were small. In comparison, the mosaic of 9 images taken from a digital camera in figure 6 took approximately 6 minutes running time. This is faster as the Kalman filter mode was used most of the time because the projective approximation is better, and the image intensities of corresponding pixels in the images did not differ much.

Figure 4: Result 1: A mosaic of 6 images taken with a disposable camera. The original images are shown at the top row.

4 Conclusions

In this paper, we designed an algorithm based on a multiresolution, multiple-hypothesis scheme to cope with

Figure 5: Result 2: The top row shows the original images which are taken with a disposable camera. The bottom row shows the mosaic. The mosaic is difficult because of the considerable independent motion taking place.

Figure 6: Result 3: A mosaic of 8 images taken with a digital camera. The original images are shown in the top 2 rows.

long-range feature matching. Central to the algorithm is a Bayesian framework for incorporating similarity measures of feature correspondences in regression. This is required because correspondences which are admitted into the estimation process have different levels of probability of being correct. If this is ignored, all admitted correspondences will be considered equiprobable, and the erroneous matches may skew the computed estimate from the true solution. Additionally, a Bayesian approach to RANSAC was developed to utilise prior and matching probabilities, as would become available as the estimation iterates through multiple resolution levels. Application of the algorithm to the real test images show that these methods perform satisfactorily.

References

[1] I.J. Cox. A review of statistical data association techniques for motion correspondence. *Int. Journal of Computer Vision*, 10(1):53–66, 1993.

[2] P. Dani and S. Chaudhuri. Automated assembling of images: Image montage preparation. *Pattern Recognition*, 28(3):431–445, 1995.

[3] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, June 1981.

[4] R.M. Haralick and L.G. Shapiro. *Computer and Robot Vision*. Addison-Wesley, 1993. Vols I and II.

[5] P. Meer, D. Mintz, and A. Rosenfeld. Robust regression methods for computer vision: A review. *Int. Journal of Computer Vision*, 6(1):59–70, 1991.

[6] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *Proc. Conf. Computer Vision and Pattern Recognition, Puerto Rico*, pages 338–343, 1997.

[7] E. Rich and K. Knight. *Artificial Intelligence*. McGraw-Hill, 2nd edition, 1991.

[8] H.S. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):814–830, 1996.

[9] R. Szeliski. Image mosaicing for tele-reality applications. Technical Report CRL 94/2, Cambridge Research Lab, Digital Equipment Corporation, May 1994.

[10] P. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. To appear in ICCV98, 1997.

[11] J. Weng, N. Ahuja, and T.S. Huang. Matching two perspective views. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(8):806–825, 1992.

[12] A. Witkin, D. Terzopoulos, and M. Kass. Signal matching through scale space. *Int. Journal of Computer Vision*, 1(2):133–144, 1987.

[13] I. Zoghlami, O. Faugeras, and R. Deriche. Using geometric corners to build a 2D mosaic from a set of images. In *Proc. Conf. Computer Vision and Pattern Recognition, Puerto Rico*, pages 420–425, 1997.