IMPSAC: Synthesis of Importance Sampling and Random Sample Consensus

P.H.S. Torr and C. Davidson

Microsoft Research Ltd, 1 Guildhall St, Cambridge CB2 3NH, UK
philtorr@microsoft.com

Abstract. This paper proposes a new method for effecting feature correspondence between images. The method operates from coarse to fine and is superior to previous methods in that it can solve the wide baseline stereo problem, even when the image has been deformed or rotated. At the coarsest level a RANSAC-style estimator is used to estimate the two view image constraint $\mathcal{R}$ which is then used to guide matching. The two view relation is an augmented fundamental matrix, being a fundamental matrix plus a homography consistent with that fundamental matrix. This is akin to the plane plus parallax representation with the homography being used to help guide matching and to mitigate the effects of image deformation.
In order to propagate the information from coarse to fine images, the distribution of the parameters $\mathbf{\Theta}$ of $\mathcal{R}$ is encoded using a set of particles and an importance sampling function. It is not known in general how to choose the importance sampling function, but a new method "IMPSAC" is presented that automatically generates such a function. It is shown that the method is superior to previous single resolution RANSAC-style feature matchers.
Keywords: Structure from motion, Stereoscopic vision.

1 Introduction

The goal of this work is to obtain accurate matches and image relations between consecutive images, with the ultimate aim of recovering 3D structure and camera projection matrices from an uncalibrated image sequence (such as might be obtained from a hand-held camcorder) where the motion is unlikely to be smooth or known *a priori*. Once the matches and two view image relation have been recovered, they can be used for image compression, or as a basis for building 3D graphical models from an image sequence [2,22,28]. These are underpinned by the need to match tokens/features (usually interest points) successfully through image sequences with a large number of frames. It transpires that the correspondence problem is one of the most difficult parts of structure recovery, especially when these images are far apart (the *wide baseline* problem) or when they undergo large rotations (the *image deformation* problem). Small baseline image matching technology has made large advances over the past decade [1,2,3,11,17, 22,26,30], but there has been comparatively little progress in wide baseline matching technology. Furthermore, the small baseline methods do not work on every image pair. For example, feature based cross correlation methods may fail if (1) there are insufficient features in the image pair, (2) there is too much repeated structure for features to get

D. Vernon (Ed.): ECCV 2000, LNCS 1843, pp. 819–833, 2000.

a good match, or (3) there is an image deformation that causes the cross correlation to fail.

There has been some work on rectifying these problems. Pritchett and Zisserman [19] present a set of recipes for special cases, but no unified theory of how to solve the problem in general. Cham and Cipolla [5] present a multi scale method for feature matching when making mosaics. The work is valid only if there is no parallax, i.e. if the image motion is governed by a homography. Furthermore the formulation is flawed as it propagates parameters using the estimate at the coarser level as a prior for the estimate at the finer, but since the images at fine and coarse resolution are not independent, the prior and likelihood are not independent. This leads to an erroneous posterior, which is then used (in their method) as a prior for the next level, compounding the error.

The method presented here solves the image deformation and wide baseline matching problems. It also requires no camera calibration. A coarse to fine approach is adopted in which information about the epipolar geometry is passed from the coarser levels to the finer. Ideally, the information to be transmitted would be the posterior distribution of the parameters at the coarser level. Encoding this posterior distribution and its relation to the finer level is an intricate task, not least because the normalization constant of the distribution is unknown. Three powerful statistical methods are enlisted to create a solution: (1) to represent the distributions as a set of particles, (2) the use of importance sampling to generate unbiased draws from the posterior distribution, (3) RANSAC to generate the importance sampling function. In this way the posterior distribution at the coarse level is used as an importance sampling function to draw samples from the posterior distribution at the finer level. As a result, the epipolar geometry is estimated by using features at many different scales, solving the problem of having to select this scale manually.

A fundamental component of several existing algorithms is the use of epipolar geometry to simplify the search for correspondences between view pairs, particularly because epipolar geometry and matches consistent with this geometry may be computed simultaneously, using only features in each view. Two images of a rigid object are related by a fundamental matrix, or in special cases just by a homography. The types of two view relations that might arise are described in Section 2, and the likelihood of the matches given these relations in Section 2.1. Existing geometry based matching methods are reviewed in Section 3, they comprise two stages: (a) estimate best cross correlation matches, (b) estimate epipolar geometry using a robust estimator. However this approach breaks down for the image deformation and wide base line cases. In Section 4 the coarse to fine algorithm is outlined, and the wide base line problem overcome, but cross correlation still fails if there is image deformation. This is because matches are initially scored by a combination of their cross correlation score and their agreement with epipolar geometry. However in order to calculate the cross correlation the deformation of each image patch must be known. Thus an image deformation homography is estimated in addition to the epipolar geometry, leading to a plane plus parallax representation. Local patches may be warped by the image deformation homography to establish cross correlation scores. This combined set of parameters is referred to as the augmented fundamental matrix and is described in Section 5. The results are given in Section 7, where the algorithm is demonstrated on the wide baseline and the image deformation problems.

Notation The image of a 3D scene point $\mathbf{X}$ is $\mathbf{x}^1$ in the first view and $\mathbf{x}^2$ in the second, where $\mathbf{x}^1$ and $\mathbf{x}^2$ are homogeneous three vectors, $\mathbf{x} = (x, y, 1)^\top$. The correspondence $\mathbf{x}^1 \leftrightarrow \mathbf{x}^2$ will also be denoted as $\mathbf{x}^{1,2}$. Throughout, underlining a symbol $\underline{x}$ indicates the perfect or noise-free quantity, distinguishing it from $x = \underline{x} + \Delta x$, which is the measured value corrupted by noise.

2 The Two View Relations

Within this section the possible relations $\mathcal{R}$ on the motion of points between two views are summarized. Four examples of $\mathcal{R}$ are considered: (a) the Fundamental matrix [7, 12], (b) the affine fundamental matrix [18] (c) the planar projective transformation (a homography), and (d) the affinity. All these two view relations are estimable from image correspondences alone.

The epipolar constraint is represented by the Fundamental matrix [7,12]. This relation applies for general motion and structure with uncalibrated cameras; consider the movement of a set of point image projections from an object which undergoes a rotation and non-zero translation between views. After the motion, the set of homogeneous image points $\{\underline{\mathbf{x}_i}\}, i = 1, \ldots n$, as viewed in the first image is transformed to the set $\{\underline{\mathbf{x}_i}'\}$ in the second image, with the positions related by

$$\underline{\mathbf{x}}_i'^\top \mathbf{F} \underline{\mathbf{x}_i} = 0 \tag{1}$$

where $\underline{\mathbf{x}} = (\underline{x}, \underline{y}, 1)^\top$ is a homogeneous image coordinate and $\mathbf{F}$ is the Fundamental Matrix. The affine fundamental matrix $\mathbf{F}_A$ is the linear version of $\mathbf{F}$. The affine camera is applicable when the data is viewed under orthographic conditions and gives rise to a fundamental matrix with zeroes in the upper 2 by 2 submatrix[1], and it is studied in detail by Shapiro [20].

In the case where all the observed points lie on a plane, or the camera rotates about its optic axis and does not translate, then all the correspondences lie on a homography:

$$\underline{\mathbf{x}}' = \mathbf{H}\underline{\mathbf{x}} \,. \tag{2}$$

The affinity $\mathbf{H}_A$ is a special case of the homography with zeros for the first two elements of the bottom row. Again it is valid under uncalibrated orthographic conditions.

2.1 Likelihood of a Match Given a Relation

In this section, the maximum likelihood formulation is given for computing any of the multiple view relations, given a set of matches. Later this formalism will be extended to include the case when the matches themselves are unknown and must be estimated. In the following we make the assumption that the noise in the two images is Gaussian on

[1] Actually $\mathbf{F}_A$ occurs in the non-orthographic case when the optical planes of the two cameras coincide [23]. Affine reconstruction in this case gives projectively correct results.

each image coordinate with zero mean and uniform standard deviation σ. Thus, given a true correspondence, the probability density function of the noise perturbed data is

$$p(\mathbf{x}^{1,2}|\mathcal{R}) = \prod_{i=1\ldots n} \left(\frac{1}{\sqrt{2\pi}\underline{\sigma}} \right)^n e^{-\left(\sum_{j=1,2} (\underline{x}_i^j - x_i^j)^2 + (\underline{y}_i^j - y_i^j)^2\right)/(2\underline{\sigma}^2)} \quad , \tag{3}$$

where n is the number of correspondences and $\mathcal{R}$ is the appropriate 2 view relation, e.g. the fundamental matrix or projectivity.

The above derivation assumes that the errors are Gaussian. Often, however, features are mismatched and the error on the match is not Gaussian. Thus the error can be modelled as a mixture model of Gaussian and uniform distribution:-

$$p(e) = \left(\gamma \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{e^2}{2\sigma^2}) + (1-\gamma)\frac{1}{v} \right) \tag{4}$$

where γ is the mixing parameter and v is just a constant, σ is the standard deviation of the error on each coordinate. To correctly determine γ and v entails some knowledge of the outlier distribution; here it is assumed that the outlier distribution is uniform, with $-\frac{v}{2} .. + \frac{v}{2}$ being the pixel range within which outliers are expected to fall (for feature matching this is dictated by the size of the search window for matches). Therefore the error minimized is the negative log likelihood:

$$-L = -\sum_i \log \left(\gamma \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\sum_{j=1,2} \frac{\left((\underline{x}_i^j - x_i^j)^2 + (\underline{y}_i^j - y_i^j)^2 \right)}{2\sigma^2} + (1-\gamma)\frac{1}{v} \right) \right) . \tag{5}$$

Given a suitable initial estimate there are several ways to estimate the parameters of the mixture model, most prominent being the EM algorithm [6,16], but gradient descent methods could also be used. Because of the presence of outliers in the data the standard method of least squares estimation is often not suitable as an initial estimate, and it is better to use a robust estimate such as RANSAC which is described in the next section.

3 Random Sampling Guided Matching

Within this section the state of the art in feature matching is described. This computation requires initial matching of points (e.g. corners detected to sub-pixel accuracy by the Harris corner detector [10]) between two images; the aim is then to compute the relation from these image correspondences. Given a corner at position (x, y) in the first image, the search for a match considers all corners within a region centred on (x, y) in the second image with a threshold on maximum disparity. The strength of candidate matches is measured by sum of squared differences in intensity. At this stage, the threshold for match acceptance is deliberately conservative in order to minimise incorrect matches. Nevertheless, many mismatches will occur because the matching process is based only on proximity and similarity. These mismatches (called outliers) are sufficient to render standard least squares estimators useless. Consequently robust methods must be adopted, which can provide a good estimate of the solution even if some of the data are outliers.

There are potentially a significant number of mismatches amongst the initial matches. Since correct matches will obey the epipolar geometry, the aim is to obtain a set of "inliers" consistent with the epipolar geometry using a robust technique. In this case "outliers" are putative matches which are inconsistent with the epipolar geometry. Robust estimation by random sampling (such as MLESAC, LMS or RANSAC) have proven the most successful [8,24,29,26]. These algorithms are well known and briefly summarized in Fig. 1.

Table 1. *A brief summary of all the stages of random sampling guided matching*

1. Detect corner features using the Harris corner detector [10].
2. Putative matching of corners over the two images using proximity and cross correlation to get best set of matches.
3. Repeat for a fixed number of samples or until "jump out" [25] occurs
   a) Select a random sample without replacement of the minimum number of correspondences $\{\mathbf{x}_i^{1,2}\}$ required to estimate the relation $\mathcal{R}$
   b) Estimate the unique image relation $\mathcal{R}$ consistent with this minimal set.
   c) Calculate the error $-L$ for all matches (MLESAC), or the median of residuals (LMS), or the number of inliers (RANSAC).
4. Select the best solution over all the samples i.e. that which minimizes $-L$ (MLESAC), or that which minimized the median error (LMS), or that which maximized the number of inliers (RANSAC).
5. Minimize robust cost function over all correspondences using gradient descent.

3.1 Problems with Conventional Matching

There are two types of failure mode for the class of matching algorithms in Table 1. The first is the wide baseline case, see Figure 1, which shows two images taken at the same time instant [2] where the disparity is 160 pixels. In the conventional algorithm, described above, a search window must be set for putative matches. If this search window is too large (which it must be in this case to guarantee that the correct match lies within it), then there is a combinatorial explosion of putative matches. This leads to a catastrophic failure of correlation matching as there are too many potential false matches for each corner. The second failure mode is caused when the image is rotated (see Figure 2). In this case, standard correlation matching cannot be expected to succeed, because the correlation score is not rotationally invariant. Using a rotationally invariant correlation score does not correct this problem; instead it reduces the discriminating power of the score, increasing the number of mismatches even when the second image is not rotated. The answer to both these problems, presented here, is to adopt a coarse to fine strategy. The coarse to fine strategy has been used successfully for small baseline homography matching [4], but neglected for feature matching.

[2] Kindly provided by Dayton Taylor

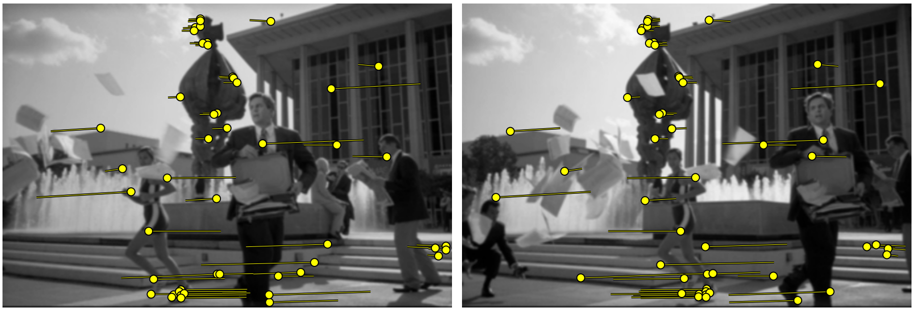

Fig. 1. Wide Baseline Failure of MLESAC/LMS/RANSAC**:** *50 matches from the first and last images of the Samsung sequence. The images are were imaged at the same time instance and are two of 50 taken from a 50 camera stereo rig. The features are shown in each image (circles) together with the line joining them to their correspondence in the other image, and are matched with an affine fundamental matrix. Although several of the features with small disparities have been correctly matched, features with large disparities are incorrectly matched. This is because, as the disparity increases, so does the number of potential mismatches.*

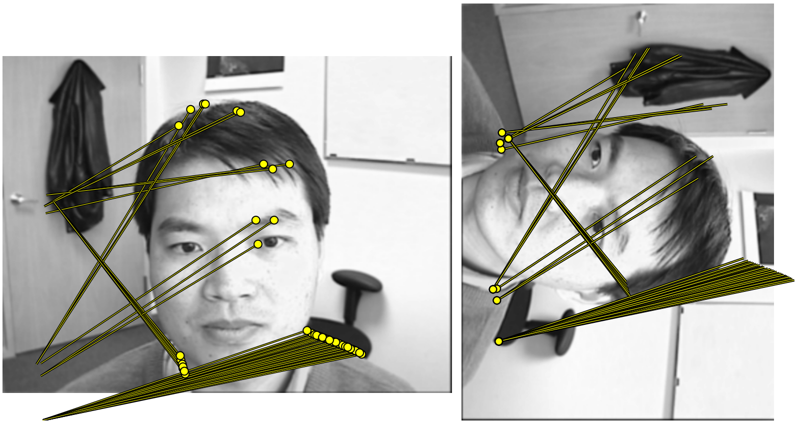

Fig. 2. Catastrophic Failure of MLESAC/LMS/RANSAC **Due To Rotation:** *the second image in the Zhang sequence has been rotated by 90 degrees, in addition there is a slight change of pose of the head. The image correlation used is not invariant to rotation, so there are too many mismatches for* MLESAC *to converge. Rotation-invariant correlation is not a solution to this problem, because it is less discriminating and thus results in too many mismatches even when the second image is not rotated.*

4 Coarse to Fine

In the coarse to fine strategy, an image pyramid is formed by subsampling the image repeatedly by a factor of 2. At the coarsest level of this pyramid (level $l = 0$), the distribution of the parameters $\boldsymbol{\theta}$ of the relation $\mathcal{R}$ given the data $\mathbf{D}_l$ is $p(\boldsymbol{\theta} | \mathbf{D}_l)$. The information contained in this posterior distribution should be propagated down to the finer levels. One way to propagate information from one level to the next is to simply propagate down the mode of this distribution. However, at the coarsest levels this distribution is not expected to have a strong peak and often propagation of the mode does not convey sufficient information. Too soon a commitment to a single hypothesis may cause the algorithm to converge to the wrong solution. Rather, it is desirable to pass as much of the distribution as possible from one level to the next.

The coarse to fine strategy is beneficial for a number of reasons. It furnishes a solution to the wide baseline problem because the search window, and thus the number of potential false matches per corner, is reduced at the coarser levels. Furthermore, at the coarser level, it is less computationally intensive to estimate the global image deformation (e.g. cyclorotation), by testing different hypotheses for the deformation of the cross correlation between image patches.

Two problems arise with this. First, the parametric form of the distribution is not known. Second, the normalizing factor of the distribution is not known. The first problem is overcome by representing the distribution by a set of particles $\{\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_m\}$ with weights $\{w_1 \dots w_m\}$. This sort of representation has been used with a good deal of success in the tracking literature [14]. Ideally the set of particles would be drawn from the posterior distribution. One way to achieve this is via importance sampling, which is defined next.

4.1 Importance Sampling

Importance sampling [9] is a key step in drawing approximate samples from complicated high dimensional posterior distributions for which the normalization factor is unknown. Suppose it is of interest to draw samples from such a distribution $q(\boldsymbol{\theta})$, and there exists a normalized positive density function (the importance sampling function) $g(\boldsymbol{\theta})$ from which it is possible to draw samples. The algorithm proceeds as follows:

1. Generate a set of M draws $S^t = \{\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_M\}$ from $g(\boldsymbol{\theta})$.
2. Evaluate $q(\boldsymbol{\theta})$ for each element of S^t.
3. Calculate importance weights $w_i = \frac{q(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}$ for each element of S^t.
4. Sample a new draw from S^{t+1} from S^t where the probability of taking a new $\boldsymbol{\theta}_i$ is proportional to its weight w_i.

Iterating this procedure from step 2 is called *sampling importance resampling* (SIR). This process, in the limit, produces a fair sample from the distribution $q(\boldsymbol{\theta})$ [9]. The rate of convergence is determined by the suitability of the importance function $g(\boldsymbol{\theta})$. The worst possible scenario occurs when the importance ratios are small with high probability and large with low probability. There is no general purpose method for choosing a good importance sampling function, but in the next section it will be explained how RANSAC can be used to construct one.

4.2 Using RANSAC to Generate the Importance Sampling Function: IMPSAC

The success of RANSAC-style methods proves that at least some of the generated samples lie in areas of high posterior probability. It would be nice to be able to harness the RANSAC mechanism in order to generate a good importance sampling function with which to propagate information from coarse to fine levels. There are several ways in which this can be done. The method we favour is to model the importance function $g(\boldsymbol{\theta})$ as a mixture of Gaussians, each centred at a RANSAC sample, with the mixing parameters being in proportion to the posterior likelihood of each sample: $p(\boldsymbol{\theta}\,|\mathbf{D})$. This presents a new method for propagating probabilities: generate a density function $g(\boldsymbol{\theta})$ via RANSAC and use this as an importance sampling function to draw samples from the posterior. This method is dubbed “IMPSAC”.

Speed Up 1. Using all the particles to generate the mixture of Gaussians can be slow. Generally if the distribution is to be represented by L particles then a particle can be excluded from the computation if it contains less than $1/L$ of the mass of the density function.

Speed Up 2. Often the artifice of constructing the mixture of Gaussians can be computationally onerous. A simpler device can be obtained under the assumption that the initial set of particles generated by the random sampling of minimal match sets is uniform. Although this assumption is not realistic in theory, unless we are interested in calculating integrals or exact expectations under the distribution, it is safe to make in practise (when all we are interested in is finding the mode of the distribution). One case when the exact posterior would be of interest would be if one was evaluating the evidence to effect model selection (e.g. choosing whether $\mathbf{F}$ or $\mathbf{H}$ best modelled the data. This is the subject of a forthcoming paper).

5 The Augmented Fundamental Matrix

In [27] it was shown that using $\mathbf{H}$ to guide matches throughout the sequence leads to fewer matches being extracted in the part of the sequence undergoing a general motion, as might be expected since the model underfits this part. However, when a loose threshold of 3 pixels was used (as opposed to a threshold of 1.25 pixels which is the two sigma window arising from interest point measurement noise) the homography is able to carry correct matches even when the planar assumption is broken. The explanation lies in the “plane plus parallax” model of image motion [13]: the estimated homography often behaves as if induced by a ‘scene average’ plane, or indeed is induced by a dominant scene plane; the homography map removes the effects of camera rotation and change in internal parameters, and is an exact map for points on the plane. The only residual image motion (which is *parallax* relative to this homography) arises from the scene relief relative to the plane. Often this parallax is less than the loose displacement threshold, so that all correspondences may still be obtained. Thus the homography provides strong disambiguation for matching and the parallax effects do not exceed the loose threshold.

This suggests a new method for matching, in which one (or more) homographies *and* a fundamental matrix are estimated for the data. The homographies estimated at the coarser level are used to guide the search windows in order to detect matches for the features at the finer level. They can also be used to guide the cross correlation matching at the finer level in that the patches that are correlated can be corrected by transformation under the homography. This representation is referred to as the *augmented* fundamental $\mathbf{F}^+$ or affine fundamental matrix $\mathbf{F}^+_A$. For the examples presented in this paper, one homography is sufficient to guide matching. This leads to a 10 parameter estimation problem for $\mathbf{F}^+$ (8 for the homography and 2 for the epipole, alternatively: 7 for the fundamental matrix and 3 for the plane of the homography), and 7 for $\mathbf{F}^+_A$ (6 for the affinity and 1 for the epipole, alternatively 4 for the affine fundamental matrix and 3 for the plane). Future work will consider the use of several planes to augment the fundamental matrix, but for many image sequences one seems to be sufficient to get good matches.

In order to estimate the augmented relation, the likelihood for a match given this relation (Section 2.1) is decomposed into two parts: the first is the usual likelihood of the fundamental matrix (4), the second is the likelihood of the parallax in the image given the homography. This is assumed to be Gaussian with large variance. This has the effect in general that if two equally good matches happen to lie along an epipolar line the one closer to the base plane represented by the homography is favoured.

5.1 Augmented Likelihood Formulation

Previously the optimisation was done on only the "best" set of matches found under cross correlation. If the image deformation is unknown, this is no longer acceptable and the likelihoods must be extended to incorporate a term for the probability of the correlation conditioned on a given match and a given homography. Given the set of images (the data) $\mathbf{D}_l$ at level l of the image pyramid, both the parameters of the relation $\boldsymbol{\theta}$ and the set of matches δ_i, $i = 1 \ldots n$ need to be estimated. Here the ith match is encoded by δ_i, which is the disparity of the ith feature of the first image. The set of disparities of all the features is Δ. The laws of probability give:

$$p(\boldsymbol{\theta}, \Delta|\mathbf{D}_l) \propto p(\mathbf{D}_l|\boldsymbol{\theta}, \Delta)p(\boldsymbol{\theta}, \Delta) = p(\mathbf{D}_l|\boldsymbol{\theta}, \Delta)p(\Delta|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ . \tag{6}$$

Under the assumption that the errors in each match are independent, and that the the distribution of matches are independent:

$$p(\boldsymbol{\theta}, \Delta|\mathbf{D}_l) = \prod_i p(\boldsymbol{\theta}, \delta_i|\mathbf{D}_l) \propto \prod_i p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_i)p(\delta_i|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ . \tag{7}$$

This is the criterion to be optimised. However, only the augmented relation $\boldsymbol{\theta}$ is propagated from the coarser level, and the matches are encoded by the homography part of $\boldsymbol{\theta}$ and the disparity assigned to the parallax.

The probability of $\boldsymbol{\theta}$ can be calculated by integrating out the disparity parameters. Note the following identity: $\int_{-\infty}^{\infty} p(\mathbf{X}, \mathbf{Y}|\mathbf{I})d\mathbf{Y} = p(\mathbf{X}|\mathbf{I})$. Then

$$p(\boldsymbol{\theta}|\mathbf{D}_l) \propto \int p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_1)p(\delta_1|\boldsymbol{\theta})p(\boldsymbol{\theta})d\delta_1 \times \ldots \times \int p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_n)p(\delta_n|\boldsymbol{\theta})p(\boldsymbol{\theta})d\delta_n . \tag{8}$$

Since δ_i may take only a finite number of values, corresponding to the features $j = 1 \ldots m$ of the second image (see below for the case of occlusion),

$$p(\boldsymbol{\theta}|\mathbf{D}_l) \propto \prod_i \sum_j p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_i = j)p(\delta_i = j|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{9}$$

Each term in this expression is the product of three elements. First, $p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_i = j)$ is the likelihood of the image (patches) given the augmented fundamental matrix $\mathbf{F}^+$ and the match $\delta_i = j$. This is evaluated from the cross correlation score after warp under the homography part of $\mathbf{F}^+$ under the assumption that the image intensities have Gaussian error mean zero and standard deviation σ_D. The second term $p(\delta_i = j|\boldsymbol{\theta})$ is the likelihood of the match given the relation, given by equation (3) (account for occlusion is made below). The third term $p(\boldsymbol{\theta})$ is the prior on the relation, assumed uniform here, but this can be altered to include any appropriate prior knowledge.

Thus the decomposition above is useful in two ways: (1) it yields $p(\boldsymbol{\theta}|\mathbf{D}_l)$ without having to commit to a set of matches and (2) the likelihood $p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_i)$ takes account of the different hypothesised image deformations.

Occlusion To take account of occlusion, the disparity δ_i for a given match can take a null value, representing the fact no match can be found with a finite probability, that is $p(\delta_i = \emptyset) = \rho_1$. For this value of δ_i, the conditional probability of the image patch correlation $p(\mathbf{D}_l|\boldsymbol{\theta}, \delta_i)$ is also set to a constant value ρ_2. The resulting estimate of $\boldsymbol{\theta}$ remains constant over a large range of $\rho_{1,2}$. Smaller values of these constants tend to peak the distribution, while larger values flatten it.

6 Feature Matching Algorithm Using IMPSAC

The algorithm is summarized in Fig. 2. The first stage is to generate the features at all levels. Then, at the coarsest scale, cross correlation scores are generated between all features, with each patch undergoing 16 evenly space rotations (this is only necessary if image deformation is expected). Random sampling of minimal match sets is used to generate an initial set of putative solutions, each match being picked in proportion to its correlation likelihood.

After the coarsest level $l = 0$, two options are considered for generation of the subsequent importance sampling functions, both valid. The first method (importance sampling) is to use the mixture of Gaussian methods described above. This has the advantage that new particles are generated across the whole parameter space, the disadvantage that it is slow to compute. The second method (importance resampling) represents $g_l(\boldsymbol{\theta})$, $l > 0$ using the set of particles S^l each assigned probability $p(\boldsymbol{\theta}_i) = \pi_i$ where $\pi_i = \frac{w(\boldsymbol{\theta}_i)}{\sum_j w(\boldsymbol{\theta}_j)}$ and $w(\boldsymbol{\theta}_i) = \frac{p(\boldsymbol{\theta}_i|\mathbf{D}_l)}{g_{l-1}(\boldsymbol{\theta}_i)}$. A problem with the resampling approach is that one particle $\boldsymbol{\theta}_{max}$ may come to represent all the probability mass at a given level and hence all the particles at the finer level will be replicas of it. One solution to this problem in a different setting is justified by Sullivan and Blake [21] in which a small amount of noise (compensated for by subtracting it from the prior $p(\boldsymbol{\theta})$) is added to each particle as it is transmitted to the next level. This can be intuitively explained in this case by the fact that the resolution of the match-coordinates changes as the image is subsampled (here by a factor of 2). For instance, if the features are not represented to sub-pixel accuracy, then change of scale introduces some uncertainty into where the features should lie at the next scale of the order 0-1 pixel. Each particle was estimated from a minimal set of feature matches. Thus, to add uncertainty to $\boldsymbol{\theta}$, noise from 0-1 pixel is added to the minimal set used to estimate it. In this case, each particle represents a distribution over $\boldsymbol{\theta}$ -space, determined by the level of uncertainty in the coordinates.

Table 2. *Feature Matching Algorithm using* IMPSAC.

1. At each scale: Detect features.
2. Putative matching of corners over the coarsest two images using proximity and cross correlation under a variety of rotations.
3. At the **coarsest level**. Generate a set of particles $S^0 = \{\boldsymbol{\theta}\,^0_m\}$ and weights $\{w^0_m\}$, $m = 1 \ldots M$ as follows:
   a) Select a random sample without replacement of the minimum number of correspondences required to estimate the relation $\mathcal{R}$
   b) Calculate $\boldsymbol{\theta}\,^0_i$ from this minimal set.
   c) Calculate $w^0_i = p(\boldsymbol{\theta}\,|\mathbf{D}_0)$ for each sample.
4. For $l = 1$ to $l =$ finest level
   a) Generate an importance sampling function $g_l(\boldsymbol{\theta})$ from S^{l-1}.
   b) Generate M draws from g_l, to generate S^l.
   c) For each $\boldsymbol{\theta}\,^l_i$, calculate $w^l_i = p(\boldsymbol{\theta}\,^l_i|\mathbf{D}_l)/g_l(\boldsymbol{\theta}\,^l_i)$.
5. The particle with the maximum posterior probability is taken as the MAP estimate. This can then be used as a starting point for a gradient descent algorithm.

7 Results

The final stage of the algorithm in Table 2 is to select the most likely particle at the finest level as the most likely hypothesis. This is the particle $\boldsymbol{\theta}\,_{imax}$ which maximises $p(\boldsymbol{\theta}\,_i|\mathbf{D})$. The i^{th} feature in the first image is matched to the feature j in the second image which maximises $p(\delta_i = j|\boldsymbol{\theta}\,_{imax})$. Figure 3 shows the successful matching of two images with up to 160 pixels disparity, demonstrating the capacity of IMPSAC for wide baseline matching. Figure 4 shows how IMPSAC is robust to large rotations of the image. In figure 5, mismatches of MLESAC are corrected by rematching with the augmented likelihood, doubling the number of matched features.

8 Future Work

Due to space constraints, model selection is not the topic of this paper. However it will be briefly illustrated how importance sampling can be used to evaluate the marginal likelihoods required for model comparison. Given a set of k models $\mathbf{M}_1 \ldots \mathbf{M}_k$ that can explain the data $\mathbf{D}$ (here the models are fundamental matrix, homography, augmented fundamental matrix etc.) then Bayes rule leads us to

$$p(\mathbf{M}_i|\mathbf{DI}) = \frac{p(\mathbf{D}|\mathbf{M}_i\mathbf{I})p(\mathbf{M}_i|\mathbf{I})}{p(\mathbf{D}|\mathbf{I})}\ , \qquad (10)$$

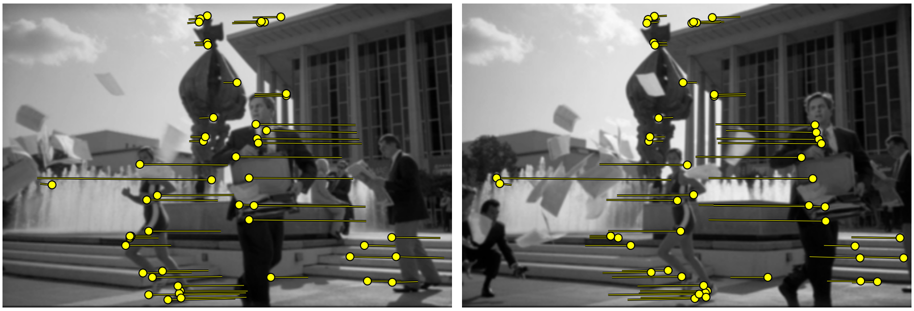

Fig. 3. Wide Baseline Success of IMPSAC**:** *the first and last images from the Samsung sequence, captured at the same time but from different positions. The disparity between the images is up to 160 pixels, yet only 3 or 4 of the 50 example matches shown are mismatched.*

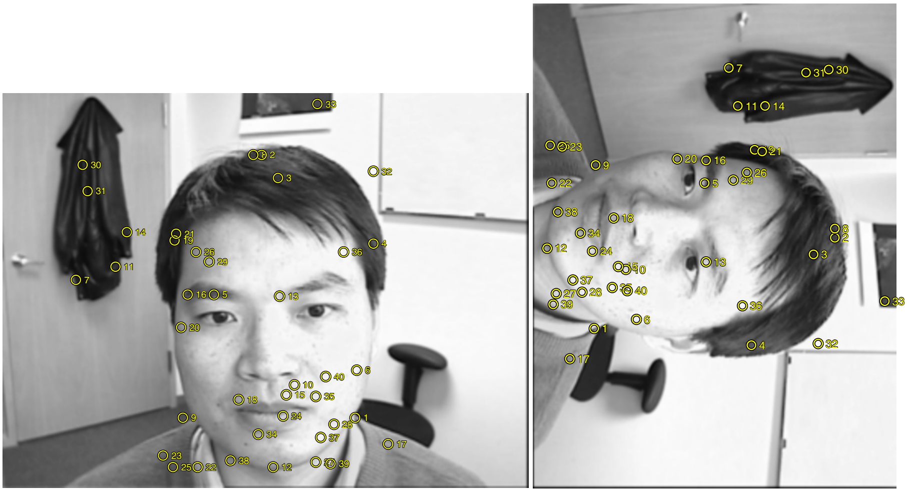

Fig. 4. Rotation Success of IMPSAC**:** *Despite the combination of a rotation of 90 degrees and the change in pose of the face, the features are correctly matched. Although just* 40 *features are shown for clarity, over* 1000 *were matched.*

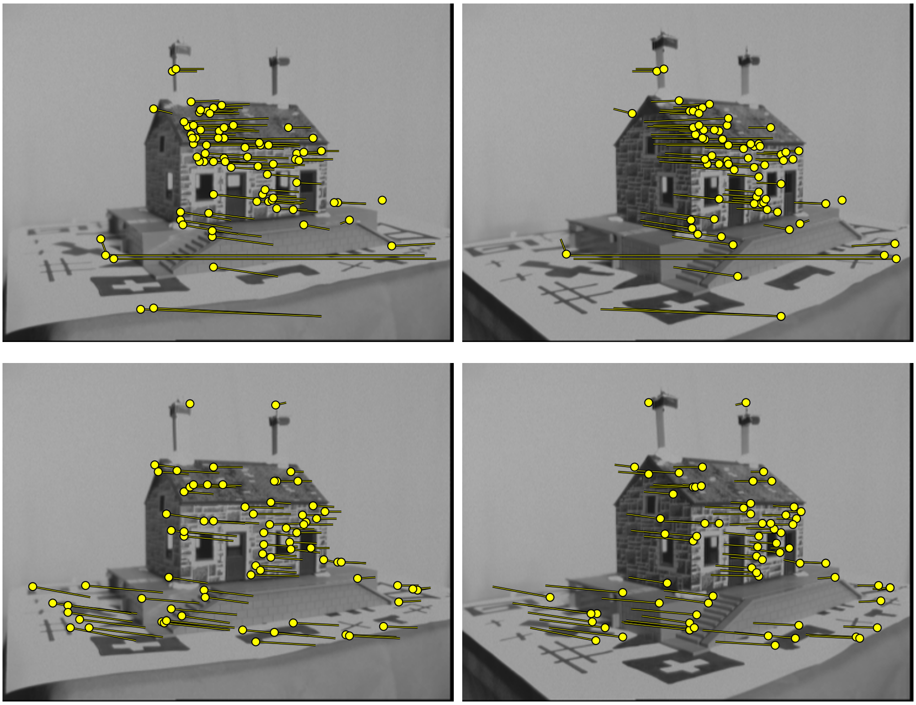

Fig. 5. MLESAC **Mismatches Corrected by Augmented Likelihood:** *(Above)* MLESAC *matches with affine fundamental matrix include numerous mismatches. (Below) From the same* MLESAC *hypothesis, rematching with augmented likelihood increases number of matches from 509 to 1274, also reducing mismatches.*

where $\mathbf{I}$ is the prior information assumed about the world. Note $p(\mathbf{D}|\mathbf{I})$ is the same for all models. Assuming that all the models are equally likely *a priori* i.e. $\mathbf{M}_i = \frac{1}{k}$, the key posterior likelihood of each model is the evaluation of $p(\mathbf{D}|\mathbf{M}_j\mathbf{I})$, which is called the evidence. This is the integral of the likelihood over all possible values of the model's parameters:

$$p(\mathbf{D}|\mathbf{M}_j\mathbf{I}) = \int p(\mathbf{D}|\mathbf{M}_j\boldsymbol{\theta}\,\mathbf{I})p(\boldsymbol{\theta}\,|\mathbf{M}_j\mathbf{I})\partial\boldsymbol{\theta} \tag{11}$$

where $\boldsymbol{\theta}$ are the jth model's parameters, and $p(\boldsymbol{\theta}\,|\mathbf{M}_j\mathbf{I})$ is the prior distribution of parameters of the model. One method for numerically evaluating this integral would be to uniformly sample the parameter space and sum the posteriors of the samples. Unfortunately the high dimensionality of the parameter space precludes this. One could draw samples from the prior and sum the posterior of these samples, but typically the prior is too diffuse to yield samples around the peak of the distribution. Importance sampling furnishes a Monte Carlo method for performing this integration [9], the advantage of which is that samples can be taken more densely around the expected peak of the posterior and less densely in areas of little interest. If the importance sampling function is $g(\boldsymbol{\theta})$ ($g(\boldsymbol{\theta})$ is a normalized density), then given a set of M particles drawn from $g(\boldsymbol{\theta})$

$$p(\mathbf{D}|\mathbf{M}_j\mathbf{I}) \rightarrow \sum_{i=1}^{i=M} \frac{p(\mathbf{D}|\mathbf{M}_j\boldsymbol{\theta}\,\mathbf{I})p(\boldsymbol{\theta}\,|\mathbf{M}_j\mathbf{I})}{g(\boldsymbol{\theta})} \text{ as } M \rightarrow \infty \qquad (12)$$

Evaluation of this leads to the selection of an augmented fundamental matrix model for the Samsung sequence shown in Figure 3, a homography model for the Zhang sequence shown in Figure 4, and an augmented affine fundamental matrix for Figure 5.

9 Conclusion

Within this paper coarse to fine estimation of structure and motion has been demonstrated. This has been achieved through the synthesis of powerful statistical techniques. The concept of using a random sampling estimator to generate the importance sampling function, IMPSAC, is a general mechanism that can be used in a wide variety of statistical problems beyond this. It provides a solution to the general problem of how to create importance sampling functions for outlier corrupted data. The coarse to fine strategy helps overcome the wide baseline problem, and this combined with the plane plus parallax representation (the augmented fundamental matrix) overcomes the image deformation problem. The resultant is a general purpose and powerful image matching algorithm that can be used for 3D reconstruction or compression. Finally how the importance sampling can also be used for automatic model selection is explained.

References

1. Ayache N. *Artificial vision for mobile robots.* MIT Press, Cambridge, 1991.
2. Beardsley P., Torr P., and Zisserman A. 3D model acquisition from extended image sequences. In *Proc. European Conference on Computer Vision*, LNCS 1064/1065, pages 683–695. Springer-Verlag, 1996.
3. Beardsley P., Zisserman A., and Murray D. Navigation using affine structure and motion. In *Proc. European Conference on Computer Vision*, LNCS 800/801, pages 85–96. Springer-Verlag, 1994.
4. J. R. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proc. 2nd European Conference on Computer Vision, LNCS 588, Santa Margherita Ligure*, pages 237–252, 1992.
5. T. Cham and R. Cipolla. A statistical framework for long range matching in uncalibrated image mosaicing. In *Conference on Computer Vision and Pattern Recognition*, pages 442–447, 1998.
6. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Statist. Soc.*, 39 B:1–38, 1977.
7. O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision, LNCS 588, Santa Margherita Ligure*, pages 563–578. Springer–Verlag, 1992.
8. M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, vol. 24:381–95, 1981.

9. A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1995.
10. C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Conf.*, pages 189–192, 1987.
11. Harris C. Determination of ego-motion from matched points. In *Third Alvey Vision Conference*, pages 189–192, 1987.
12. R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Proc. 2nd European Conference on Computer Vision, LNCS 588, Santa Margherita Ligure*, pages 579–587. Springer-Verlag, 1992.
13. Irani M. and Anandan P. Parallax geometry of pairs of points for 3d scene analysis. In Buxton B. and Cipolla R., editors, *Proc. 4th European Conference on Computer Vision, LNCS 1064, Cambridge*, pages 17–30. Springer, 1996.
14. M. Isard and A. Blake. Condensation — conditional density propagation for visual tracking. *International Journal of Computer Vision*, 28(1):5–28, 1998.
15. E. T. Jaynes. Probability theory as extended logic. Not yet published a postscript version of this excellent book is available at ftp://bayes.wustl.edu/pub/Jaynes/, 1999.
16. G.I. McLachlan and K. Basford. *Mixture models: inference and applications to clustering*. Marcel Dekker. New York, 1988.
17. McLauchlan P. and Murray D. A unifying framework for structure from motion recovery from image sequences. In *Proc. International Conference on Computer Vision*, pages 314–320, 1995.
18. J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT press, 1992.
19. P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. 6th International Conference on Computer Vision, Bombay*, pages 754–760, January 1998.
20. L. S. Shapiro. *Affine Analysis of Image Sequences*. PhD thesis, Oxford University, 1993.
21. J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Bayesian correlation. In *Seventh International Conference on Computer Vision*, volume 2, pages 1068–1075, 1999.
22. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorisation approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.
23. P. H. S. Torr. *Outlier Detection and Motion Segmentation*. PhD thesis, Dept. of Engineering Science, University of Oxford, 1995.
24. P. H. S. Torr and D. W. Murray. Outlier detection and motion segmentation. In P. S. Schenker, editor, *Sensor Fusion VI*, pages 432–443. SPIE volume 2059, 1993. Boston.
25. P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *Int Journal of Computer Vision*, 24(3):271–300, 1997.
26. P. H. S. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In U Desai, editor, *ICCV6*, pages 727–732. Narosa Publishing House, 1998.
27. P.H.S. Torr, A. Fitzgibbon, and A. Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *IJCV*, 32(1):27–45, 1999. Marr Prize Paper ICCV 1999.
28. Zeller, C. *Projective, Affine and Euclidean Calibration in Compute Vision and the Application of Three Dimensional Perception*. PhD thesis, RobotVis Group, INRIA Sophia-Antipolis, 1996.
29. Z. Zhang, R. Deriche, O. Faugeras, and Q. T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *AI Journal*, vol.78:87–119, 1994.
30. Z. Zhang and O. Faugeras. *3D Dynamic Scene Analysis*. Springer-Verlag, 1992.