Guided Sampling and Consensus for Motion Estimation

Ben Tordoff and David W Murray

Department of Engineering Science, University of Oxford
Parks Road, Oxford OX1 3PJ, UK
[bjt,dwm]@robots.ox.ac.uk

Abstract. We present techniques for improving the speed of robust motion estimation based on random sampling of image features. Starting from Torr and Zisserman's MLESAC algorithm, we address some of the problems posed from both practical and theoretical standpoints and in doing so allow the random search to be replaced by a guided search. Guidance of the search is based on readily-available information which is usually discarded, but can significantly reduce the search time. This guided-sampling algorithm is further specialised for tracking of multiple motions, for which results are presented.

1 Introduction

Since its introduction by Fischler and Bolles in 1981 [1] and later appearance in the statistical literature as Rousseeuw and Leroy's Least Median of Squares [4], random sampling and consensus (RANSAC) has been widely used in computer-vision — particularly in the areas of recovering epipolar geometry and 3D motion estimation [5, 7–10].

In [7], Torr and Zisserman describe a method of maximum likelihood estimation by sampling consensus (MLESAC). It follows the random sampling paradigm of its RANSAC ancestor, in that a minimal set of matches is used to estimate the scene motion and then support is sought in the remaining matches. However, whereas RANSAC just counts the number of matches which support the current hypothesis, MLESAC evaluates the likelihood of the hypothesis, representing the error distribution as a mixture model.

In this paper we seek ways to make iterative random sampling more suitable for use in applications where speed is of importance, by replacing random sampling with guided sampling based on other knowledge from the images.

We begin with a detailed review of the MLESAC algorithm, and in section 3 make observations on its performance. In section 4 we describe a scheme for resolving one issue in MLESAC's formulation and using this result guide the random sampling to reduce the search time. The possibility of incorporating multiple match-hypotheses is entertained in section 5, and section 6 shows that the search time is further reduced in the multiple motion case when information is propagated over time. Results and discussions appear in the sections to which they relate.

2 Maximum Likelihood Estimation by Sampling Consensus

As the basis for the remainder of this paper we will now describe Torr and Zisserman's MLESAC algorithm in detail. It is assumed that the feature detection and matching stages have given rise to a set of matches where each feature is only matched once, but some and possibly many, of the matches may be in error.

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2350, pp. 82–96, 2002.

As mentioned earlier, MLESAC evaluates the likelihood of the hypothesis, representing the error distribution as a mixture model. Several assumptions are made, the veracity of which are discussed in section 3:

1. The probabilities of matches being valid are independent of one-another.
2. If a match is a mismatch, the error observed is uniformly distributed.
3. If a match is valid it will be predicted by the correct motion estimate up to Gaussian noise related to the noise in feature position estimates.
4. Every match has the same *prior* probability of being a mismatch.

A minimal set of matches h is chosen to estimate a motion hypothesis M_h, for which all matches $i = 1 \ldots n$ are either valid, v_i, or invalid, $\bar{v}_i$. The probability that M_h is a correct estimate of the true motion is denoted $\mathrm{p}(M_h)$. All n features are transferred between images using this motion and the differences between the estimated and actual match positions give rise to residual errors r_i, and hence to an overall error R_h. These errors may be calculated in just the second image, or more usually in both images (see [2], sec. 3.2). The aim is to sample randomly the space of possible motions and choose the hypothesised motion M_h that has maximum posterior probability given the data available, $\mathrm{p}(M_h|R_h)$. This cannot be measured directly and Bayes' rule is used:

$$M_{\mathrm{MAP}} = \max_h \left[\mathrm{p}(M_h|R_h)\right] = \max_h \left[\mathrm{p}(R_h|M_h)\frac{\mathrm{p}(M_h)}{\mathrm{p}(R_h)}\right]$$

where $\mathrm{p}(R_h|M_h)$ is the likelihood that it is correct and $\mathrm{p}(M_h)$, $\mathrm{p}(R_h)$ the prior probabilities of the motion and residuals respectively. If these terms can be measured then the maximum a posteriori (MAP) motion can be estimated.

The prior $\mathrm{p}(R_h)$ is constant irrespective of the choice of M_h, and nothing is known about the prior probability that the motion is correct, $\mathrm{p}(M_h)$. This means that the MAP estimate cannot be found and the new aim is to maximise the likelihood and hope that the motion with maximum likelihood (ML) estimate is similar to the maximum posterior (MAP) estimate. The new aim is to find

$$M_{\mathrm{MLESAC}} = \max_h \left[\mathrm{p}(R_h|M_h)\right]$$

2.1 Evaluating the Likelihood

To convert the likelihood into a usable form it is necessary to use assumption (1), that the probability of each residual is independent

$$\mathrm{p}(R_h|M_h) = \prod_i^n \mathrm{p}(r_i|M_h) \quad .$$

Evaluation of the probability of each residual has two parts, according to whether it is a valid match or not. Assumption (2) states that if the feature is mismatched the probability of the residual is uniform, but will be related to the size of the search area w

$$\mathrm{p}(r_i|\bar{v}_i, M_h) = 1/w \quad .$$

For a valid match the residual is due only to zero-mean Gaussian noise of deviation σ (related to the feature detector localisation error), and under assumption (3) the conditional probability is

$$\mathrm{p}(r_i|v_i, M_h) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{r_i^2}{2\sigma^2}} \quad .$$

The prior probabilities that match i is valid $\mathrm{p}(v_i)$ or invalid $\mathrm{p}(\bar{v}_i)$ are by definition mutually exclusive ($\mathrm{p}(v_i, \bar{v}_i)$=0) and exhaustive ($\mathrm{p}(v_i)$+$\mathrm{p}(\bar{v}_i)$=1)so that the combined probability of observing the residuals given the hypothesised motion is given by the sum rule

$$\mathrm{p}(r_i|M_h) = \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{r_i^2}{2\sigma^2}}\right) \mathrm{p}(v_i) + \left(\frac{1}{w}\right) (1 - \mathrm{p}(v_i)) \quad . \tag{1}$$

Assumption (4) states that $\mathrm{p}(v_i)$ is constant across all matches and all hypothesised motions, that is $\mathrm{p}(v_i) = \mathrm{p}(v)$, the prior estimate of the proportion of valid matches. This controls the relative importance of the valid and invalid probability distributions, examples of which are shown in figure 1. Note the Gaussian curve when the residuals are small and the non-zero tails, giving a cost function which does not over-penalise extreme outliers. This shape is characteristic of the related robust method of M-estimation [3].

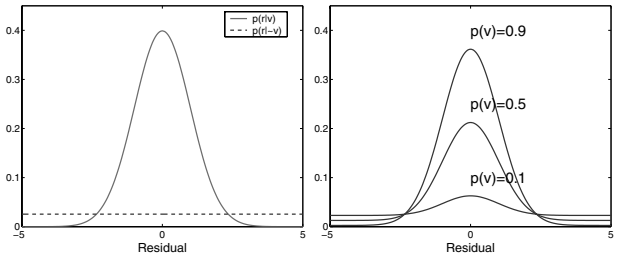


Fig. 1. (Left) the Gaussian and uniform distributions modelling valid and invalid residuals respectively. (Right) three examples of the combined distribution for different values of $\mathrm{p}(v)$. Note the long tails which reduce the penalty for large residuals.

However, even assuming $\mathrm{p}(v)$ to be the same for all features, it still must be estimated somehow. Torr and Zisserman approach this problem by using an iterative maximisation scheme to estimate $\mathrm{p}(v)$ directly from each motion hypothesis and set of residuals. The goal is to find the $\mathrm{p}(v)$ that maximises $\mathrm{p}(R_h|M_h)$ for the current hypothesis M_h. As $\mathrm{p}(R_h|M_h)$ varies smoothly with $\mathrm{p}(v)$, any suitable ascent method can be used to find the maximum. However, in this approach each hypothesis M_h will generate its own estimate of $\mathrm{p}(v)$, meaning that the comparison between likelihoods is based on different mixtures. We will return to this point in section 3.1.

We now have all the information required to calculate and compare the likelihood of each hypothesis. In practice for numerical stability the log likelihood is optimised:

$$M_{\mathrm{MLESAC}} = \max_h \left[\sum_i^n \log \left\{ \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{r_i^2}{2\sigma^2}} \right) \mathrm{p}(v) + \left(\frac{1}{w} \right) (1 - \mathrm{p}(v)) \right\} \right] \quad (2)$$

2.2 Knowing When to Stop Sampling

As in RANSAC [1] and Least Median of Squares [4], if the sampling and evaluation steps are repeated over a large number of samples it is hoped that at least one hypothesised motion will be close to the true motion. If the proportion of valid data is $\mathrm{p}(v)$, and the minimum number of features required to form a hypothesis is m, then the probability, $\mathrm{p}(M_c)$, that a correct hypothesis has been encountered after I iterations is approximately

$$\mathrm{p}(M_c) \approx 1 - [1 - \mathrm{p}(v)^m]^I \quad . \quad (3)$$

Although $\mathrm{p}(v)$ is not generally known in advance, a lower bound can be estimated from the largest $\mathrm{p}(v)$ observed in the mixture estimation step. A stopping condition is usually determined from a desired confidence level (eg. $\mathrm{p}(M_c) > 95\%$).

2.3 Example Results

Figure 2 shows some results from running MLESAC on an image pair containing a moving object, a toy locomotive. The motion is pure-translation and, as the object exhibits little thickness, a planar homography can be used to model the transformation. Thus, four point matches between the two views are used for each MLESAC sample (shown in black) and the residual error for all other matches used to score the motion hypothesis. Matches which are inlying to the hypothesis are shown in white. Note that the better motion hypotheses give residual distributions which are highly peaked at low residual error but with a significant tail. The second peak observed in the final sample is due to the second motion present in the data-set — the stationary background.

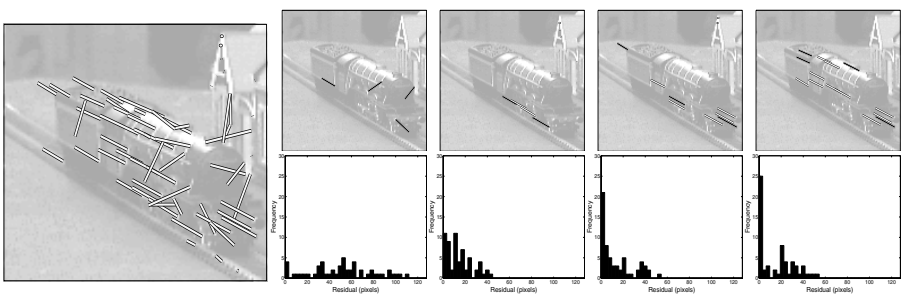


Fig. 2. The starting match-set for the train images (left) and several of the samples that MLESAC selects, along with the distribution of resulting residuals. The samples are arranged with increasing likelihood from left to right.

2.4 Summary of the Basic Algorithm

Robust computation of a single motion using MLESAC involves the following steps

1. For many trials do:
   (a) Choose a minimum number of matches and estimate a motion hypothesis M_h.
   (b) Calculate the residual error r_i for all matches when transformed by the hypothesised motion.
   (c) Estimate $\mathrm{p}(v)$ for a particular $\mathrm{p}(M_h)$ by maximising the sum-log-likelihood.
   (d) Retain the M_h which has the highest likelihood.
2. Use the estimated motion and list of valid matches as the starting point for a more accurate motion calculation, if required.

3 MLESAC Revisited

Having described the rationale underlying MLESAC and having demonstrated it in action, we now return to some specific points which merit closer examination.

3.1 Estimating the Mixture Parameter

In [7], Torr and Zisserman assume that the probability of a match being valid, $\mathrm{p}(v_i)$, is the same for all matches. Further, this constant value $\mathrm{p}(v)$ is re-estimated for each hypothesised motion. There are two deficiencies here:

1. All matches are not equal. Information which is usually freely available indicates that some matches are more likely to be valid than others (ie. we can refine the prior).
2. The prior probability that a match is valid does not depend on the hypothesised motion — it is a *prior* constant. Allowing it to vary makes the comparison of the likelihoods of the different motion hypotheses unfair.

However, as MLESAC works, and works well, these deficiencies appear to have little effect on the determination of which motion hypotheses are better. To reach an understanding of this, a large number of trials were conducted on a variety of imagery where the log-likelihood score (equation 2) was evaluated for a range of motion hypotheses. For each motion hypothesis the log-likelihood was found over the complete range of $0 \leq \mathrm{p}(v) \leq 1$.

Typical results are shown in figure 3. The interesting properties are

- Better hypotheses have higher likelihoods over all values of $\mathrm{p}(v)$ — ie. the curves do not cross.
- Better hypotheses have maxima at higher $\mathrm{p}(v)$ values.
- The maxima are always located at or below the true value of $\mathrm{p}(v)$.

Taken together, these observations suggest that there is little to be gained by re-estimating $\mathrm{p}(v)$ for each motion hypothesis — it might as well be taken to be 0.5. This saves a small amount of computation time and overcomes the first of the objections that were raised above as the estimate is now constant across all hypotheses. Even if a more accurate estimate of $\mathrm{p}(v)$ is desired, the estimate from the best hypothesis seen so far should be used, rather than an estimate from the current hypothesis.

Note from figure 3 that poor motion hypotheses give curves with maximum $\mathrm{p}(v)$ at, or close to, zero, making all matches appear as outliers.

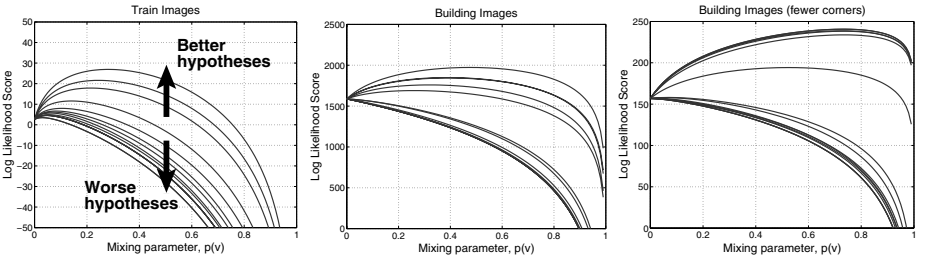


Fig. 3. The effect of varying the mixing parameter on the likelihood score of several motion hypotheses for three different sets of matching features.

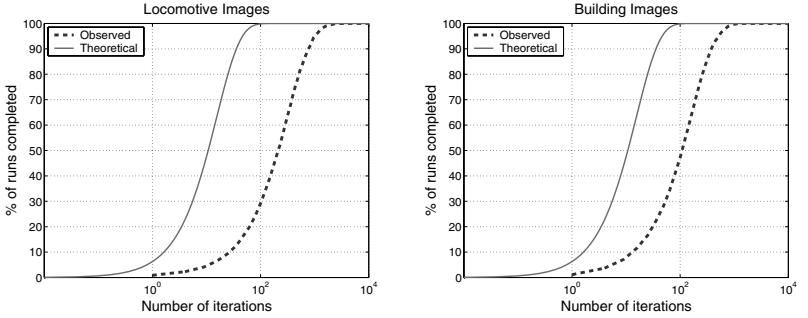


Fig. 4. MLESAC is run on the images 5000 times and the first iteration at which more than 75% of inliers is found recorded for each run. The proportion of runs that complete at or before a given iteration are shown.

3.2 Stopping Criterion

It is widely appreciated that the approximate stopping criterion (eqn. 3) specified for both RANSAC [6] and MLESAC [7] is often wildly optimistic. With about 50% valid data, equation (3) suggests that a confidence of 99% would require about 70 samples, but on the train images of the previous section the first reasonable solution was not seen until several hundred samples and the best solution at around 9500 samples.

To demonstrate, the MLESAC algorithm was repeatedly run on the locomotive images, stopping when a good solution was found, a "good" solution taken as one where at least 75% of the possible inliers are found. Figure 4 shows the number of iterations and the proportion of 5000 runs that had found a solution at or before this time for two sequences. In both cases the actual number of iterations required to observe a good motion hypothesis is significantly larger than predicted. The theoretical form of the stopping curve matches that observed from the data, but with around an order of magnitude shift.

The reason for this difference is that with noisy data it is not enough to have a sample composed only of inliers, they must be inliers that span the object so that the remaining matches are compared to an interpolated rather than an extrapolated motion. This significantly reduces the number of sample sets that will accurately hypothesise the motion. In cases where the motion is uniform over the image (eg. due to camera motion) a quick fix is to force the random sampling to pick points widely spaced, but

when the size of the object in the image is not known this approach will cause problems. A more generally applicable alternative is desirable.

The number of iterations required before stopping is even more important when simultaneously estimating multiple motions. In the case of a two-motion scene, a sample set of features is chosen for both foreground and background motions (M_{fh}, M_{bh} respectively) at each iteration and motion hypotheses calculated. Residuals against each motion (r_{fi}, r_{bi}) are calculated for each feature, and the evaluation of the likelihood becomes the mixture of three distributions:

$$\begin{aligned} M_{\text{MLESAC}} &= \max_h \left[\mathrm{p}(R_h|M_{fh}, M_{bh})\right] \\ &= \max_h \prod_i^n \Big\{ \mathrm{p}(r_i|M_{fh}, f_i, v_i)\mathrm{p}(f_i|v_i)\mathrm{p}(v_i) \\ &\qquad + \mathrm{p}(r_i|M_{bh}, b_i, v_i)\mathrm{p}(b_i|v_i)\mathrm{p}(v_i) + \mathrm{p}(r_i|\bar{v}_i)\mathrm{p}(\bar{v}_i) \Big\} \\ &= \max_h \left[\sum_i^n \log \left\{ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{r_{fi}^2}{2\sigma^2}} \mathrm{p}(f_i|v_i)\mathrm{p}(v_i) \right.\right. \\ &\qquad \left.\left. + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{r_{bi}^2}{2\sigma^2}} \mathrm{p}(b_i|v_i)\mathrm{p}(v_i) + \frac{1}{w}(1 - \mathrm{p}(v_i)) \right\} \right] \end{aligned} \quad (4)$$

where the extra terms $\mathrm{p}(f_i|v_i)$ and $\mathrm{p}(b_i|v_i)$ are the prior probabilities that the match belongs to the foreground or background respectively, given that it isn't a mismatch. Usually $\mathrm{p}(f_i|v_i)$ and $\mathrm{p}(b_i|v_i)$ are mutually exclusive and exhaustive so that $\mathrm{p}(b_i|v_i) = 1 - \mathrm{p}(f_i|v_i)$, and as $\mathrm{p}(v_i)$, $\mathrm{p}(\bar{v}_i)$ are also exclusive and exhaustive the parameters v_i, f_i and b_i are integrated out. As with the mixture parameter $\mathrm{p}(v)$ in the single motion case, the background/foreground priors $\mathrm{p}(b_i|v_i), \mathrm{p}(f_i|v_i)$ could be assumed uniform over all features and estimated using expectation-maximisation at each MLESAC iteration. However this has the same deficiencies as in section 3.1 and it would be preferable to find a weighting for each feature that is constant across MLESAC trials. Such a weighting is discussed in section 6.

A conservative estimate of the number of iterations I that are required to find a good pair of hypothesised motions increases dramatically over the single motion case:

$$\mathrm{p}(M_{fc}, M_{bc}) \approx 1 - \left[1 - \mathrm{p}(f|v)^{m_f}\mathrm{p}(b|v)^{m_b}\mathrm{p}(v)^{(m_f+m_b)}\right]^I \quad .$$

Consider an example with equal numbers of foreground and background features mapped by homographies between two images ($m_f = m_b = 4$), of which 25% are mismatched (ie. 37.5% are valid for each motion). The correct method of estimating both foreground and background simultaneously requires around 7660 iterations for 95% confidence and nearly 12000 for 99%.

An alternative is to estimate the motions individually. 37.5% of the data is valid for the first motion, requiring 150 iterations for 95% confidence or 230 for 99%. The second motion is estimated from the remaining matches (of which 60% is now valid) in 22 iterations for 95% confidence or 34 for 99%. These are clearly huge computational savings. However, evalutating the motions individually makes the assumption that the outliers to the first motion form a uniform distribution — visibly not the case in figure 2. In some cases it may be desirable to make this assumption and sacrifice accuracy for speed.

4 Guided Sampling

In the previous sections we have raised a couple of issues with the assumptions underlying MLESAC. We also noted that the number of iterations required for a good solution may be far higher than expected. To improve MLESAC we seek improved estimates of the following:

- $\mathrm{p}(v_i)$, the prior probability that a feature is valid.
- $\mathrm{p}(f_i|v_i)$, $\mathrm{p}(b_i|v_i)$, the prior probabilities that a valid feature belongs to the foreground or background motions.

Although the discussion that follows is explicitly concerned with point features, similar arguments can be applied to estimation based on other feature types.

4.1 Using the Match Score

An obvious source of information on the validity of a feature-match is the match score. For point features, a common measure is the zero-normalised cross-correlation between image patches around the points in the two images being compared. The score is used to select the best consistent set of matches over all possible combinations.

Figures 5(a-c) show the distribution of correlation score; for both mismatches and valid matches determined over a range of image sequences whose pixel composition is quite different. The correct match distributions are always largely the same, and, provided there is little repeated texture within the search window, so too are the mismatch distributions. Normalising these histograms gives an approximation to the probability density of observing a particular match score s_{ik} given the validity of the putative match, $\mathrm{p}(s_{ik}|v_{ik})$ and $\mathrm{p}(s_{ik}|\bar{v}_{ik})$ (each feature i has several possible matches k).

Over the course of a tracking sequence, or when a camera repeatedly captures similar sequences (such as for a surveillance camera) then statistics for correct and incorrect matches can be built up online. However, where these distributions are unknown, we can approximate the distributions under the assumption of little repeated texture using simple functions. The mismatch distribution will be taken as quadratic

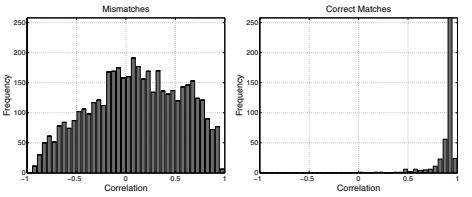
$$\mathrm{p}(s_{ik}|\bar{v}_{ik}) \approx \frac{3}{4}(1 - s_{ik})^2 \qquad -1 \leq s_{ik} \leq 1$$

and for correct matches

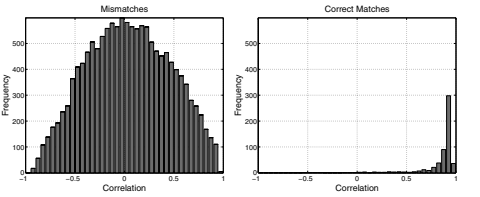
$$\mathrm{p}(s_{ik}|v_{ik}) \approx a\frac{(1 - s_{ik})}{\alpha^2} \exp - \left[\frac{1 - s_{ik}}{\alpha}\right]^2 \qquad -1 \leq s_{ik} \leq 1 \tag{5}$$

where α is a "compactness" parameter and a is a normalisation constant such that the area under the curve is unity, as in figure 5(d). These expressions are chosen for their close fit to empirical data and their simplicity. Arguing from underlying image statistics might suggest more meaningful and accurate expressions. (For the image sequences used here, $\alpha = 0.15$ and $a = 1.0$ give a reasonable fit, although the final probabilities turn out not to be particularly sensitive to small variations in α.)

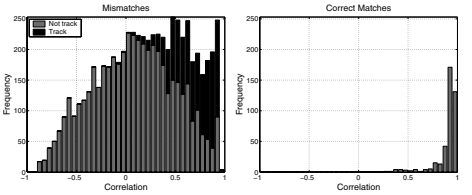
If a feature i has n_m potential matches with validity v_{ik} ($k = 1 \ldots n_m$), of which one is correct, then reasonable priors are $\mathrm{p}(v_{ik}) = 1/n_m$ and $\mathrm{p}(\bar{v}_{ik}) = (n_m - 1)/n_m$.



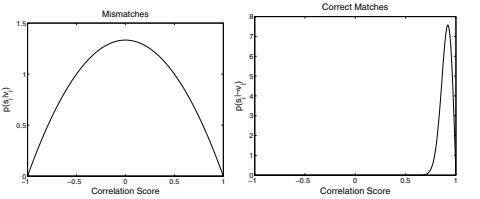
(a) Traffic images.



(b) Building images.



(c) Locomotive images (repeated structure mainly due to the track is shown in black).



(d) Approximations to the histogram shapes.

Fig. 5. Frequency of matches and mismatches against correlation score over all potential matches.

The probability of putative match k being correct when its match score is taken into account, but without considering the scores of the other matches is

$$\begin{aligned} \mathrm{p}(v_{ik}|s_{ik}) = \mathrm{p}(s_{ik}|v_{ik})\frac{\mathrm{p}(v_{ik})}{\mathrm{p}(s_{ik})} &= \mathrm{p}(s_{ik}|v_{ik})\frac{\mathrm{p}(v_{ik})}{\mathrm{p}(s_{ik}|v_{ik})\mathrm{p}(v_{ik}) + \mathrm{p}(s_{ik}|\bar{v}_{ik})\mathrm{p}(\bar{v}_{ik})} \\ &\approx \mathrm{p}(s_{ik}|v_{ik})\frac{1}{\mathrm{p}(s_{ik}|v_{ik}) + \mathrm{p}(s_{ik}|\bar{v}_{ik})(n_m - 1)} \quad . \quad (6) \end{aligned}$$

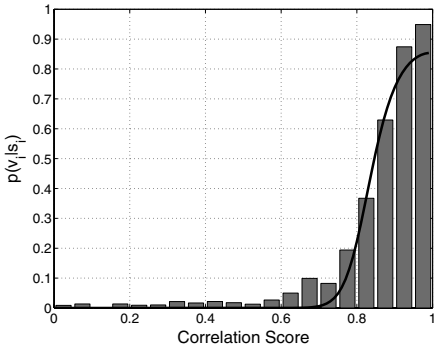
It is also desirable to consider that all the putative matches might be wrong, in which case the feature matches an extra null feature. A simple way to represent this is to increase by one the number of putative matches in equation 6.

Furthermore we can also include the additional knowledge that only one putative match per feature is correct (if any are). We calculate the probability that a putative match k is valid given the scores of all the putative matches

$$\mathrm{p}(v_{ik}|s_{i,1\ldots n_m}) = \frac{\mathrm{p}(v_{ik}|s_{ik})\prod_{j\neq k}^{n_m}\mathrm{p}(\bar{v}_{ij}|s_{ij})}{\sum_l^{n_m}\left[\mathrm{p}(v_{il}|s_{il})\prod_{j\neq l}^{n_m}\mathrm{p}(\bar{v}_{ij}|s_{ij})\right] + \prod_j^{n_m}\mathrm{p}(\bar{v}_{ij}|s_{ij})}$$

where the numerator gives the probability that this putative match is correct and all the others are wrong, and the denominator normalises by the sum of all possible match probabilities plus the possibility that none is correct. This makes the additional assumption that the probabilities $\mathrm{p}(v_{ij}|s_{ij})$ are conditionally independent. Whichever of these putative matches is selected by the matching algorithm to be the single correct match, $\mathrm{p}(v_{ik}|s_{i,1\ldots n_m})$ can be used as $\mathrm{p}(v_i)$ in equation 1 instead of using a constant $\mathrm{p}(v)$.

Consider an example feature from the locomotive images. There are ten putative matches, and using $\alpha = 0.15$ and $a = 1$ in equation 5 the correlation scores map to probabilities as shown in figure 6.



s_{ik}	$\mathrm{p}(v_{ik}\|s_{ik})$	$\mathrm{p}(v_{ik}\|s_{i,1\ldots n_m})$
-0.34	0.000	0.000
0.21	0.000	0.000
0.43	0.000	0.000
0.57	0.002	0.000
0.60	0.006	0.000
0.62	0.012	0.001
0.68	0.069	0.007
0.72	0.174	0.020
0.91	0.812	0.408
0.93	0.832	0.468

Fig. 6. Converting match scores into probabilities. Left: The observed histogram and approximation to it (from fig. 5) are transformed by equation 6 with $n_m = 10$. Right: The ten putative matches, their individual and combined probabilities of being valid.

4.2 Using $\mathrm{p}(v_i)$ for Guided Sampling

We now have an estimate of the probability that a match is valid based on its match score and the scores of the other putative matches with which it competed. Whilst useful as a replacement for gradient descent estimation of an overall mixing parameter, we can also use it to guide the search — if we have evidence that one match is more likely to be valid than others, then it should be selected more often.

Each feature i has a single match selected using an optimal matcher which also delivers $\mathrm{p}(v_i)$ for the selected match. The matches are sampled for use in the minimal motion computation set using a Monte-Carlo method according to $\mathrm{p}(v_i)$. The increased cost of guiding the selection is marginal when compared to the other computations performed at each MLESAC iteration.

Figure 7 shows the "time to solution" test of section 3.2 repeated with guided sampling and the individual mixing parameters. This indicates that the computational cost of converting the correlation scores into validity likelihoods and performing weighted selection is easily offset by a dramatic reduction in the number of iterations required for a given confidence level. In real-time systems where the number of iterations may be fixed by time constraints, the confidence level is significantly increased — for instance in the building images 100 iterations gives 47% confidence for random sampling, but 99% when the sampling is guided by the match score.

5 Multiple Match-Hypotheses versus Rematching

Another possible change to the algorithm is that instead of choosing only the best putative match for use in the MLESAC algorithm, we include multiple matches for each feature, weighted in importance by their probabilities. This requires two small increases

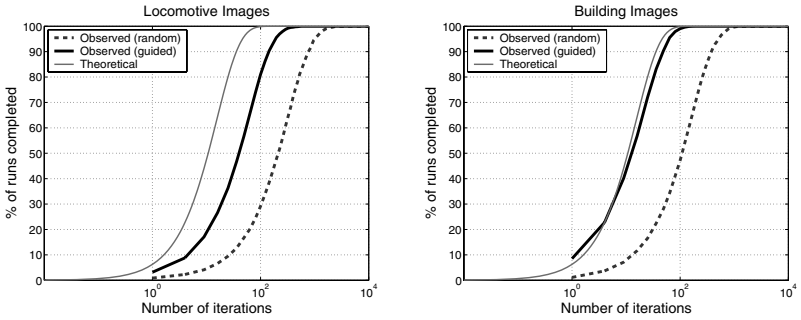


Fig. 7. The test of figure 4 is repeated with guided sampling using the match probability as in section 4.2. The dotted lines show the time-to-solution for normal random sampling and the thick solid line for guided.

in the amount of computation performed at each iteration of the algorithm, but makes a computational saving in that a global optimisation of the initial match set is not necessary.

At each MLESAC iteration when a match is selected to be part of the basis set all other putative matches for the matched features must then be removed from the list. The motions are calculated from minimal sets as before, but when evaluating the support for the motion hypothesis each putative match for each feature is tested, and the most likely one selected. In this way, even if the true match does not have the highest prior, it will still eventually be found in the MLESAC algorithm.

Using this technique, the total number of inliers that can be found by the algorithm increases by 25% for the locomotive images, and 7% for the building images. The increase depends on how well the prior $\mathrm{p}(v_i)$ reflects the correctness of the matches — if $\mathrm{p}(v_i)$ were a perfect indicator there would be no gain.

Whilst this method can significantly increase the number of matches that MLESAC finds, it also increases the sampling space from which seed matches are drawn — the average number of potential matches per feature was 1.4 for the locomotive images and 1.9 for the building, nearly doubling the sample space in the latter case. Furthermore, the extra matches that are included contain a large proportion of erroneous matches — we gain a little extra "signal" at the expense of a lot of extra "noise". Figure 8 shows the effect on the time-to-solution test for the locomotive and building images when all match hypotheses for ambiguous matches are included (ie. we have reduced the number of samples by only including multiple hypotheses when two or more matches score similarly).

We contrast this with the approach described in much of the sampling consensus literature, that of rematching. As before only one match-hypothesis is found per feature, but once MLESAC has been used and estimates of the motion(s) have been found, the likelihoods of all putative matches for all features are evaluated using these motions. For each feature, the putative match with highest likelihoods is kept in a manner similar to the initial match optimisation based on correlation score. Using this technique the time to solution remains as for the single hypothesis case, but again the total number of matches found increases. The totals observed are typically within one or two matches of the totals achieved by the multiple-hypothesis MLESAC, and there is little to choose between them on that basis.

The computational cost of rematching is one evaluation of the likelihood for every putative match. If there are an average of two putative matches per feature, then this is roughly equivalent to two iterations of MLESAC. The cost of the initial match optimisation is similarly between one and two iterations of MLESAC. When these costs are compared to the increased time-to-solution of the multiple-hypothesis approach, rematching is clearly faster in almost all cases. As each iteration of MLESAC is also slower in the multiple-hypothesis case, when speed is the goal rematching is preferable.

There are two cases where using the multiple-hypothesis approach proves advantageous. The first is in offline processing when computation time is irrelevant. The second is when the number of true matches that get mismatched is exceedingly high and if alternative matches are not included the correct motion may not be found. Although correlation matching is inadequate in many respects, enough correct matches were produced in all the sequences tested for multiple-hypothesis MLESAC to perform no better than rematching.

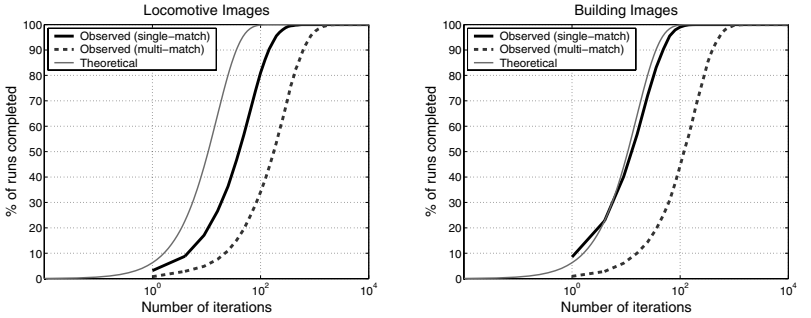


Fig. 8. Time-to-solution for guided sampling with and without multiple match-hypotheses. Including multiple match-hypotheses increases the number of inliers found, but at great computational cost.

6 Temporal Propagation of Information

Temporal consistency is a strong cue in the assignment of features to motion hypotheses — if a feature obeys the foreground motion in one frame it is likely to do so in the next. The output of MLESAC can be tailored to this purpose.

Following a successful run of guided MLESAC, we have maximum-likelihood estimates of the motions M_f,M_b, based on the overall residuals R and match scores S. As a by-product we have also evaluated the posterior probability that each match belongs to either motion, or is invalid. ie. although MLESAC is a maximum likelihood estimator of the motions, it is a maximum a posteriori estimator of the segmentation.

Without extra calculation we already have maximum likelihood estimates of the motion and the maximum a posteriori probability of each feature residual,

$$\max_h \{\mathrm{p}(R_h|M_{hf}, S),\ \mathrm{p}(R_h|M_{hb}, S),\ \mathrm{p}(r_i|M_{hf}, M_{hb}, s_i)\}\ .$$

An extra rearrangement provides a posteriori estimates of the foreground-background-mismatch segmentation (see equation 4):

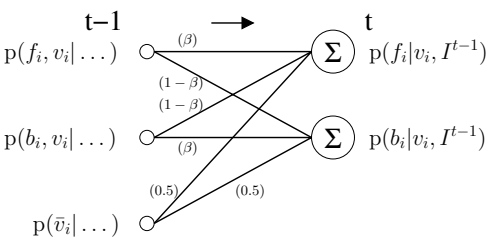
$$\mathrm{p}(f_i, v_i|M_f, r_i, s_i) = \mathrm{p}(r_i|M_f, f_i, v_i, s_i)\frac{\mathrm{p}(f_i|v_i, s_i)\mathrm{p}(v_i|s_i)}{\mathrm{p}(r_i|M_f, M_b, s_i)}$$
$$\mathrm{p}(b_i, v_i|M_b, r_i, s_i) = \mathrm{p}(r_i|M_b, b_i, v_i, s_i)\frac{\mathrm{p}(b_i|v_i, s_i)\mathrm{p}(v_i|s_i)}{\mathrm{p}(r_i|M_f, M_b, s_i)}$$
$$\mathrm{p}(\bar{v}_i|M_f, M_b, r_i, s_i) = \mathrm{p}(r_i|\bar{v}_i, s_i)\frac{\mathrm{p}(\bar{v}_i|s_i)}{\mathrm{p}(r_i|M_f, M_b, s_i)}$$

where all the probabilities shown are natural products of guided MLESAC and the score $s_{i,1\ldots n_m}$ has been abbreviated s_i. To make use of this information in subsequent frames it is necessary to allow the following possibilities:

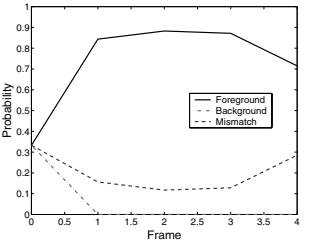
- **Propagation**. A foreground feature stays foreground, or background stays background, with probability β.
- **Cross-over**. A foreground feature becomes background, or background becomes foreground, with probability $1 - \beta$.
- **Absence**. If a matched feature did not exist in the previous frame, or was previously designated a mismatch it provides no prior information.

These relationships are summarised in the information graph of figure 9(a), where β measures the temporal efficacy of the information being propagated. Setting $\beta = 0.5$ indicates that previous assignments provide no information about current assignments, and $\beta = 1.0$ that the information is perfect. $\beta < 0.5$ indicates that previous information is in contradiction. Whilst β should be learned from observation of the extended sequence, here we fix it at 0.9 for experimentation.

At input to the first frame of a sequence we assume that features are equally likely to be background or foreground, from which point guided MLESAC increases or decreases these probabilities at each frame. (An alternative is that some heuristic initialisation might be used, such as weighting the foreground probability higher for features close to the centre of the view.) The evolution of the probabilities for a feature from the locomotive sequence is shown in figure 9(b).



(a) β indicates the temporal efficacy of the data, and I^{t-1} the "previous frame's information".



(b) The evolution of the probabilities for one feature from the locomotive sequence.

Fig. 9. Propagating the assignment of a feature to a motion.

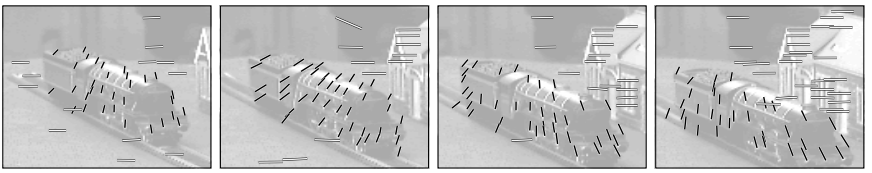


Fig. 10. Part of the locomotive sequence where both the locomotive and camera motions are approximated by planar homographies. Propagating the assignment of features to motions increases speed and ensures that the two motions are not exchanged.

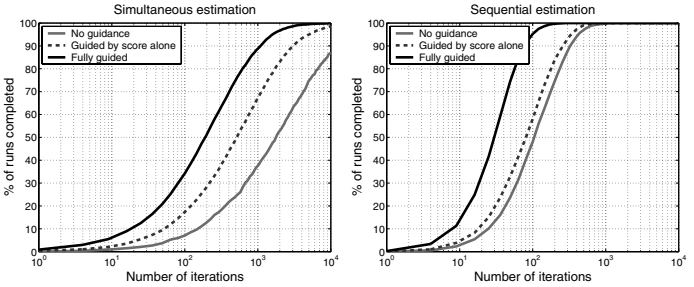


Fig. 11. An example of the improvement in speed when propagation is included. (a) simultaneous estimation of the multiple motions and (b) separate estimation of foreground then background. These tests were performed on the second frame from figure 10.

As with the improved prior on $\mathrm{p}(v_i)$ (section 4.1), these improved estimates of $\mathrm{p}(f_i|v_i)$ and $\mathrm{p}(b_i|v_i)$ can be used to weight the selection of seed matches for the foreground and background motions. However, the speed advantage this brings is harder to quantify, depending entirely on the success of the segmentation in the previous frame.

6.1 Final Results: Multiple Motion Guided MLESAC

Figure 10 shows the first few frames from a sequence with motions tracked using guided MLESAC with assignment propagation. Both foreground and background motions are approximated by planar homographies, and feature matches are initialised with equal probability of belonging to either motion. Although simultaneous estimation of the two motions is the correct segmentation method, the results shown are for separately finding foreground then background motions using just 75 and 50 iterations respectively. Propagating the feature assignments helps to prevent the foreground and background motions interchanging during the sequence.

To get an idea of how much difference propagating the assignments makes to the speed, figure 11 shows the time-to-solution test performed on the second frame of the sequence. The test is repeated with no guidance, guided by match score, and guided by both the match score and previous assignment (as before, completion requires each of the two motions to find at least 75% of the maximum possible number of inliers). For simultaneous estimation of the motions (figure 11a), around 10000 iterations are needed for 90% confidence without guidance, compared to 1000 iterations when guided by score and previous data. When estimating the two motions sequentially (figure 11b), 90% confidence requires only 75 iterations when guided (this is the total for the two estimations, with around two-thirds of the iterations needed by the first estimation).

7 Conclusions

We have introduced two simple extensions to Torr and Zisserman's MLESAC algorithm which guide the selection of features, reducing the number of iterations required for a given confidence in the solution. This has been achieved without additional image or feature measurements, and with marginal increase in computational complexity. Including multiple match-hypotheses in the sampling is straightforward, but rarely yields benefits over a final re-evaluation of matches.

Through extensive experimentation it is clear that in sampling and consensus methods, the number of iterations required to find a "good" motion estimate is far higher than the number of iterations to find a sample which consists of only valid data. This *must* be taken into account when calculating stopping criteria, and is a particular problem for simultaneous estimation of multiple motions.

We have shown that solving for a global mixing parameter is unnecessary, and that individual priors $\mathrm{p}(v_i)$ for each feature can be estimated from the number of possible matches and the match scores. Using $\mathrm{p}(v_i)$ to weight the sampling gives around an order of magnitude increase in speed, and in the multiple motion case further gains are made by propagating assignment information.

Many other cues are available in tracking sequences, not least of which is the spatial grouping of features belonging to a foreground object. Incorporating these into a framework such has been described is simple and where the information is strong would further speed the search.

Acknowledgements

This work is supported by Grant GR/L58668 from the UK's Engineering and Physical Science Research Council, and BJT is supported by an EPSRC Research Studentship.

References

1. M.A. Fischler and R.C. Bolles. Random sample concensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.
2. R. Hartley and A. Zisserman. *Multiple View Geometry, 1st edition*. Cambridge University Press, 2000.
3. P.J. Huber. *Robust statistics*. John Wiley and Sons, 1985.
4. P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection*. Wiley, New York, 1987.
5. L. Shapiro. *Affine analysis of image sequences*. Cambridge University Press, Cambridge, UK, 1995.
6. P. H. S. Torr and A. Zisserman. Robust detection of degenerate configurations while estimating the fundamental matrix. *Computer Vision and Image Understanding*, 71(3):312–333, 1998.
7. P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78:138–156, 2000.
8. P.H.S. Torr and D.W. Murray. Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, 11(4):180–187, 1993.
9. 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, September 1997.
10. G. Xu and Z. Zhang. *Epipolar geometry in stereo, motion and object recognition*. Kluwer Academic Publishers, 1996.