Segmentation of Multiple Motions by Edge Tracking between Two Frames

Paul Smith, Tom Drummond and Roberto Cipolla

Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK
{pas1001|twd20|cipolla}@eng.cam.ac.uk

Abstract

This paper presents a method for segmenting multiple motions using edges. Recent work in this field has been constrained to the case of two motions, and this paper demonstrates that the approach can be extended to more than two motions. The image is first segmented into regions, and then the framework determines the motions present and labels the edges in the image. Initialisation is particularly difficult, and a novel scheme is proposed which recursively splits motions to provide the Expectation-Maximisation algorithm with a reasonable guess, and a Minimum Description Length approach is used to determine the best number of models to use. The edge labels are then used to determine the the region labelling. A global optimisation is introduced to refine the motions and provide the most likely region labelling.

1 Introduction

The increasing availability of video in digital form means that there is growing demand for methods of automatically analysing video. The first stage in a semantic analysis of a video sequence is the segmentation of the frames into regions representing different real-world objects, objects with different image motion. Ideally, such a segmentation should provide a clean cut-out of the objects, and also the relative depth ordering of each object.

The approach proposed in this paper falls into the category of layered representations [6, 16], where the frame is decomposed into a series of regions, each obeying some parametric motion. The separation of small objects from the background may be achieved by robustly estimating the dominant motion and identifying non-conforming pixels [5, 8, 10]. However, the more general case of identifying multiple moving objects usually requires the simultaneous estimation of all motions and layers [11, 15, 17].

Motion estimation is poor in regions of low texture, and it is common to either use spatial coherency [10, 15, 17] or the image structure [17] as a prior when assigning pixels to layers. An alternative approach is to segment the image beforehand, into regions of similar colour and intensity (which are therefore likely to have the same motion). These regions are then merged with their neighbours if they have similar motions [1, 2, 9]. This region-merging approach is the one followed in this paper.

This paper takes the region-merging approach one stage further, and only considers the edges of these regions. If the region has been extracted because there is very little

BMVC 2000 doi:10.5244/C.14.35

(a) Initial static segmentation of frame 1

(b) Edges labelled as motion 1 or motion 2

(c) Foreground regions

Figure 1: *'Foreman' segmentation from two frames.* The foreman moves his head very slightly to the left between frames, but this is enough to accurately estimate the motions and calculate edge probabilities. The foreground motion can then be identified and the regions labelled to produce a good segmentation of the head.

structure in the interior of the region, this by definition means that the interior pixels are of little use for motion estimation. The majority of the motion information comes from the edges, and these edges are sufficient to both estimate the motions and layer the regions.

Many papers on motion segmentation do not consider the question of the depth ordering of layers (which is in front of which), or the problem of occlusion. Where it is considered, it is by identifying those outliers to the motion estimation which could be due to the occlusion of one layer by another [2, 17]. This paper shows, again, that the edges of regions are sufficient to also determine the layer ordering.

This paper extends the work presented in [13] by considering the particular problems of multiple motion estimation. The general approach is outlined in Section 2, before the details of the multiple-motion algorithm are given in Section 3. This includes details of the initialisation (Section 3.2) and optimisation (Sections 3.3.3 and 3.3.4), which comprise the main body of new work in this paper. Results are given in Section 4.

2 Motion Segmentation using Edges

Motion segmentation aims to determine the regions of a frame which belong to different real-world motions. This can alternatively be expressed as finding the edges of the moving objects in the image. If it is assumed that these edges are visible as intensity or colour edges in the frame, the task is one of deciding which edges in the image are occluding edges, and of which object.

This edge-based framework was introduced in [13] for the case of two motions, where a robust implementation using a Bayesian formulation was developed. The method can be summarised by reference to Figure 1.

(a) The frame to be segmented is first divided into regions using an edge-based static segmentation algorithm [12].

(b) The region edges are tracked into the next frame of the sequence. The motion is constrained to fit one of a fixed number of motion models, and the Expectation-

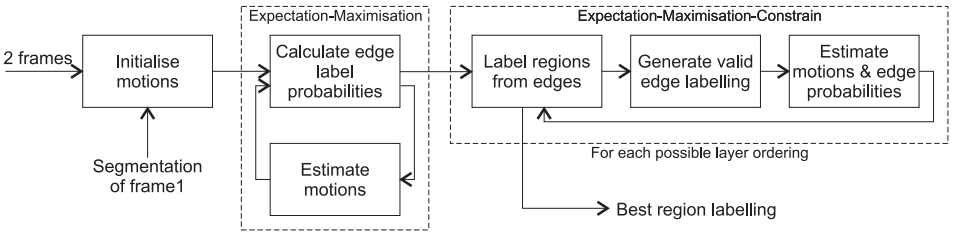

Figure 2: *Overview of the extended algorithm.* The initialisation stage determines the number of motions and a good starting point for the EM stage, which determines the motion parameters and edge labels. The region labelling is performed as part of a global optimisation step which also refines the edge labels and motions

Maximisation (EM) algorithm [3] is used to simultaneously estimate the motion model parameters and, for each edge, the probability of it fitting each model.

(c) The edge label probabilities are used to determine the most likely region labelling and layer ordering via Simulated Annealing.

3 Segmenting Multiple Motions

Although very reliable when fitting two motions (see Figure 1), applying the basic algorithm to fitting more than one foreground motion proves to be much more difficult. This paper extends the algorithm in order to provide consistent performance in these more complex situations. Figure 2 shows the extended algorithm.

The initialisation of the EM algorithm is particularly important, as local minima become more of a problem with a larger number of motions. A novel initialisation stage is introduced (Section 3.2), which recursively tries fitting more and more motions, selecting the most likely number of motions using a Minimum Description Length (MDL) method.

With the number of motions determined, and good initial guesses for these motions, the EM algorithm is used to recursively label the edges and re-estimate the motions. The motion estimation and edge labelling is the core technology used in this edge-based approach and this is reviewed first, in Section 3.1. Once an edge labelling is obtained, the regions may be labelled and the layer ordering determined (Section 3.3).

The edge labels determined by the first EM algorithm are unconstrained—they give no consideration to whether they are plausible labels for that area of the frame. Constraints cannot be reliably enforced during the first EM stage, as with an unconverged solution these constraints may be incorrectly applied. However, once an initial solution has been found these may be incorporated as part of a larger EM loop (here called ‘EMC’, Expectation-Maximisation-Constrain, Section 3.3.4) to produce a globally optimal solution.

3.1 Estimating Motions and Edge Labels

Edges are good features for tracking. Their long extent means that they can be matched between frames with good reliability, and can provide a more accurate estimate of the

motion than corner features. Along each edge in the region segmentation, tracking nodes are initialised at 10-pixel intervals. The edges are placed in the second frame according to their currently estimated motion and from each tracking node a search is then made normal to the edge to find the residual error to the nearest image edge. This residual is used both in the calculation of the edge probability and the update of the motion.

Edge labelling and multiple motion estimation is a circular problem. A motion cannot be fitted to a set of edges until those edges have been labelled, but to do a labelling a comparison must be made between different motions. This is problem resolved by using Expectation-Maximisation [3], which alternately calculates the motions from the current edge labels and estimates the edge label probabilities from the motions.

3.1.1 Edge Probabilities

The likelihood of the edge fitting the given motion is estimated from the set of tracking node residuals along that edge. It is assumed that individual node residuals along an edge are independent (which is not strictly true, but good enough for this application). The residual error under the correct motion has been modelled from experimental data, as has that under likely incorrect motions.

The likelihood is thus calculated from the product of the node likelihoods under the 'correct motion' model. The probability that an edge fits a particular motion is the product of this probability with the likelihood that, under the other motions, the nodes fit the 'incorrect motion' model. This is normalised over all motions.

3.1.2 Updating Motions from Edge Residuals

The motion of edges is modelled by a 2D affine transformation, which has been found by many to be a good fit to the small inter-frame motions [10, 16]. Traditional contour tracking techniques may be used, and the one used here is based on the group-constrained tracking method of Drummond and Cipolla [4]. This finds the best fit to the projection of the local vector field at each node.

A few modifications are made to provide robustness when fitting multiple motions. Under the EM algorithm each motion is updated by using the data from each edge in proportion to the probability that the edge belongs to that motion. A number of edges will always be incorrectly labelled and this is likely to introduce a substantial number of outliers into each motion estimation.

Gross outliers are dealt with by using an M-estimator [7] instead of simple least squares, which reduces the influence of large errors on the estimation. The other effect of outliers is, however, more subtle. If the edges of one motion are entirely confined in one small area of the frame, edges from further away can be incorporated into the motion with only a small local distortion. This problem can only be resolved by applying some form of heuristic to make it less likely that these edge will be included, and the solution adopted here is to restrict the freedom of the motion model. Image motions are most likely to be translational, but the outlier edges typically attempt to impose a rotation or shear component. A prior is therefore placed on the motion model, making it more likely to fit the translational components of the motion.

3.2 Initialisation

The EM algorithm is guaranteed to converge to a maximum, but there is no guarantee that this will be the global maximum. The most important thing in a segmentation algorithm is the initialisation [15] and, for more than 2 motions, the EM algorithm will get trapped in a local maximum unless started with a good solution. Obviously, however, if a good solution were known then the problem would already be solved!

The one case where local minima does not present a significant problem is when there are only 2 motions, and this can provide a means to approach this problem. It is found, perhaps surprisingly, that any reasonable initialisation works in the 2-motion case and, in particular, that a random initial edge labelling may be used, where each edge is either wholly assigned to one motion or the other. The EM can then start at the E-stage and produces good results in all cases tested.

To fit more than 2 motions, a sampling technique is adopted, where the EM is attempted a number of times with different initialisations. The number of motions is unknown, so these initialisations also have to include different numbers of motions. To provide a good chance of success, a hierarchical approach is used. For example, 2 motions are fitted first and then the 3 motions are initialised close to this solution.

It is worth considering what happens if 3-motion scene is labelled with only 2 motions. There are two likely outcomes:

1. The M-estimator reduces the influence of the edges belonging to one of the motions, fitting two motions correctly and leaving the edges belonging to the third motion as outliers.

2. One (or both) of the motions adjusts to fit edges from two different classes. The motion is a best fit to both sets of edges, and edges from both classes are assigned to that motion.

This provides a principled method for generating a set of 3 motion initialisations. First fit two motions, then:

1. Calculate the motion of the outlier edges and add it to the list of motions. The outlier edges are the ones for which the likelihood under the ‘incorrect motion’ statistics is greater than that under the ‘correct motion’ statistics. Or

2. Take the set of edges which best fit one motion, split these into two random groups and estimate two motions from these. Then perform EM on just these edges, to best split the two motions. Finally replace the original motion with these two. Each of the 2 initial motions can be split in this way, providing 2 further initialisations.

From each of these 3 initialisations, the EM is run to find the most likely edge labelling and motions. The likelihood of each solution is given by the product of the edge likelihoods (under their most likely motion), and best solution is the one with the highest likelihood. This solution may then be split further into more motions.

3.2.1 Determining the Best Number of Motions

As the number of motions increases, the likelihood of the solution will also increase, but this has to be balanced against the cost of a large number of motions. This is addressed by

applying the Minimum Description Length (MDL) principle, one of many information-theoretic model selection methods available [14]. This considers the cost of encoding the observations in terms of the model and any residual error. A large number of models or a large residual both give rise to a high cost.

The cost of encoding the model consists of two parts. Firstly, there are the parameters of the model; each number is assumed to be encoded to 10-bit precision, and with 6 parameters per model (2D affine), the cost is $60n_m$ (where n_m is the number of models). Secondly, each edge must be labelled as belonging to one of the models; it costs $\log_2 n_m$ to label each of the n_e edges. The edge residuals must also be encoded, and the cost for an optimal coding is equal to the total negative logarithm (to base 2) of the edge likelihoods, L_e, giving

$$C = 60n_m + n_e \log_2 n_m + \sum_e \log_2 L_e \tag{1}$$

The cost C can be evaluated after each attempted initialisation, and the smallest cost indicates the best solution and the best number of models.

3.3 Labelling Regions and Finding the Layer Order

3.3.1 The Relationship between Edges and Regions

The labelling of a region segmentation is completely defined by two parameters: $\boldsymbol{R}$, the motion to which each region belongs and $\boldsymbol{F}$, the depth ordering of the motions. Both are required to generate an edge labelling $\boldsymbol{e}$, as it is only with the depth ordering that it is known which motion is obeyed by the edge between two differently-labelled regions (it obeys the motion of the closer of the two regions, by simple occlusion reasoning). Conversely, an edge labelling encodes all that can possibly be known about the region labelling and depth ordering. Either there is only one $\boldsymbol{R}$ and $\boldsymbol{F}$ which satisfies the edge labelling, or the scene is ambiguous.

3.3.2 Using Edge Data and Region Neighbour Prior

Having determined the edge probabilities, the most likely region labelling and motion ordering can be found. This is best done in a Bayesian framework, which allows the problem to be reversed:

$$\max_{\boldsymbol{RF}} \mathrm{P}\left(\boldsymbol{RF}|\boldsymbol{e}\right) = \max_{\boldsymbol{RF}} \frac{\mathrm{P}\left(\boldsymbol{e}|\boldsymbol{RF}\right)\mathrm{P}\left(\boldsymbol{RF}\right)}{\mathrm{P}\left(\boldsymbol{e}\right)} = \max_{\boldsymbol{RF}} \mathrm{P}\left(\boldsymbol{e}|\boldsymbol{RF}\right)\mathrm{P}\left(\boldsymbol{R}\right) \tag{2}$$

(It is assumed that the prior probability of an edge labelling is constant. It is also assumed that the priors of $\boldsymbol{R}$ and $\boldsymbol{F}$ are independent, and each depth ordering $\boldsymbol{F}$ is equally likely.)

Using this result, the best region labelling can be found by hypothesising different labellings and depth orderings and seeing (a) how well it is supported by the edge data and (b) how likely that region labelling is, *a priori*. The first term is easily calculated by taking the edge labelling inferred by the selected $\boldsymbol{R}$ and $\boldsymbol{F}$ and multiplying the edge probabilities for that labelling (assuming independence). The second term uses a Markov Random Field, which encourages neighbouring regions to be labelled the same. Rather than try every possible region labelling $\boldsymbol{R}$, Simulated Annealing is used to find the optimal solution. This labelling process is repeated for each possible layer depth ordering $\boldsymbol{F}$ (i.e. $n_m!$ times) to find the most likely solution.

3.3.3 Global Optimisation: Expectation-Maximisation-Constrain (EMC)

As described, the region labelling is determined via two independent optimisations, which use edges as an intermediate representation: first the best edge labelling is determined, and then the best region labelling given these edges. It has been previously assumed that this is a good approximation to the global optimum, but unfortunately this is not the case, particularly with multiple motions.

In the first EM stage the edges are assigned purely on the basis of how well they fit each motion, with no consideration given to how likely that edge labelling is in the context of the wider segmentation. There are always a number of edges which are mislabelled and these can have an adverse effect on both the region segmentation and the accuracy of the motion estimate. In order to resolve this, a constraint step is introduced into the EM algorithm which reassigns outlying edges to a more likely motion. This is referred to as Expectation-Maximisation-Constrain, or EMC.

Once again, initialisation is an important consideration. The constraints (i.e. a sensible segmentation) cannot be applied until near the solution, so the EMC is used as a final global optimisation stage after the basic segmentation scheme has completed.

The EMC algorithm, is perhaps more correctly named 'MEC', as the processing follows the following steps (see Figure 2):

Maximisation Calculate new motions based on the definite edge labels given by the current region labelling and depth ordering.

Expectation Estimate the probable edge labels for the motions.

Constrain Calculate the most likely region labelling, and from this the set of definite edge labels.

The process is iterated until the probability P $(\boldsymbol{RF}|e)$ is maximised.

3.3.4 Global Region Constraint

The Markov Random Field used for the region prior P $(\boldsymbol{R})$ only considers the neighbouring regions, and does not consider the wider context of the frame. This makes the Simulated Annealing tractable, but does not enforce the belief that there should, in general, be only one connected group of regions representing each foreground object. Even with EMC, it is common for a few small background regions to be mislabelled as foreground and these can again have an adverse effect on the solution.

A simple solution may be employed after the Simulated Annealing of the 'C' stage. For each foreground object who's segmentation consists of more than one connected group, region labellings are hypothesised which label all but one of these groups as belonging to a lower layer (i.e. further back). The most likely of these 'one object' region labellings is the one kept.

4 Results

Figure 1 shows the standard 'Foreman' scene, which features only two motions (although substantial depth variation in the background). The results of the model selection stage are shown in Table 1 and they agree that there are only two motions, although there is

Table 1: *MDL values.* For different numbers of motions (n_m), the total cost is that of encoding the motion parameters ('Motion'), edge labelling ('Edge') and the residual ('Residual')

	Foreman			Car & Van			Library		
n_m	1	2	3	1	2	3	1	2	3
Motion	60	120	180	60	120	180	60	120	180
Edge	0	482	764	0	322	510	0	133	211
Residual	4762	3376	3145	3829	3362	2791	2377	1617	1314
Total	4822	**3978**	4089	3889	3804	**3481**	2437	1867	**1645**

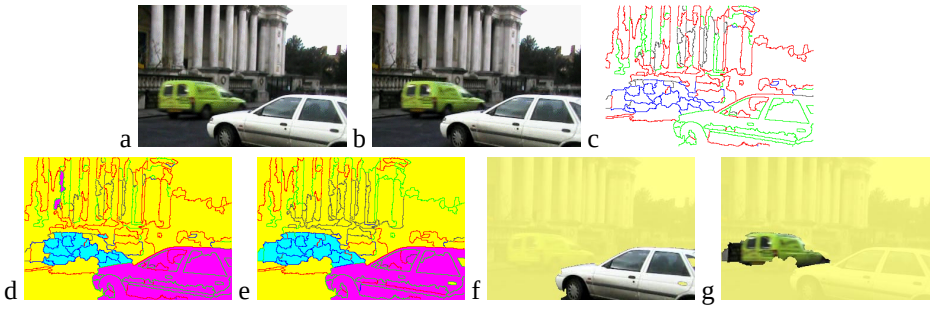

Figure 3: *'Car & Van' segmentation.* (a),(b) The two frames used for the segmentation. The car moves to the left, and the van to the right. (c) Region edges, labelled by EM. (d),(e) Edge and region labels before and after EMC. (f),(g) The two foreground layers. The car is identified as being in front of the van.

some support for fitting the girders in the bottom right corner as a third motion. The initialisation for the EM is good, and the resultant segmentation has very few errors. The refined motion generated by EMC ensures that the shoulders are labelled as foreground; under the scheme in [13], these were incorrectly labelled when using only 2 frames.

Figure 3 shows a sequence containing three motions. The camera is stationary, and the white car in the foreground begins to pull out (to the left) as the yellow van speeds by. The size of the van's motion means that under two motions, the van's edges are mainly outliers and it is here that the value of fitting a third motion to the outliers becomes apparent. The MDL is clearly in favour of fitting three motions (see Table 1).

When the edges are labelled, the car motion also fits parts of the building well, particularly due to the repeating nature of the classical architecture. This presents a few problems to the region labelling stage, as can be seen in (d) where there are a few regions on the columns which are labelled with the car. After EMC this is tidied up somewhat, but it requires the Global Region constraint to produce the clean results seen in (e),(f) and (g). The ordering of the layers is extracted correctly—the region labelling with the car in front is significantly more likely (i.e. better supported by the edge labels) than that with the order of the car and van swapped.

Another three motion sequence is shown in Figure 4. In this case the motion is due to parallax as the camera moves while viewing objects at three different depths. Once again, the correct number of motions is clearly identified by the MDL and the recursive initialisation method performs well. Labelling the motion of the horizontal lines in the scene is difficult as the camera (and hence object motions) is horizontal, and so these

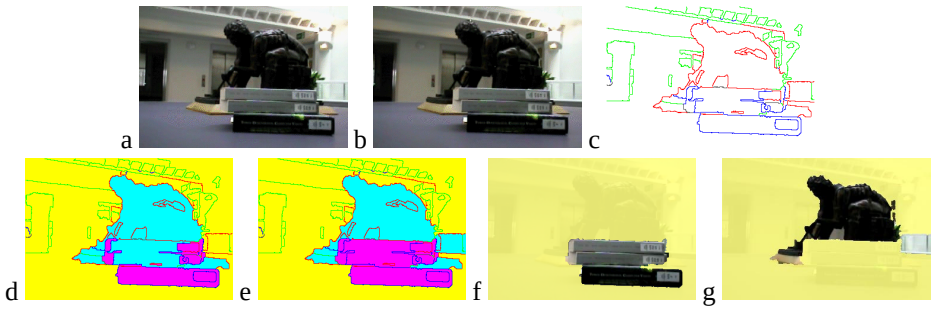

Figure 4: *'Library' segmentation.* (a),(b) The two frames used for the segmentation. The camera moves to the left, and the books, statue and background move differing amounts due to parallax (c) Region edges, labelled by EM. (d),(e) Edge and region labels before and after EMC. (f),(g) The two foreground layers. The books are identified as being in front.

edges could fit any of the motions equally well. The labelling of these edges is strongly affected by noise, and improving the edge statistics in these ambiguous cases is an area for further research. Here, the edge marking the top of the books has been incorrectly labelled and so the initial region segmentation incorrectly merges some of the books with the statue (see (c) and (d)). This is improved during the EMC stage to give the labelling seen in (e), but the edge is still very tentatively labelled. Apart from the horizontal edges, the other edges are labelled very accurately, despite the small inter-frame motion.

While the correct layer is confidently labelled as background, the ordering of the two foreground layers is more ambiguous in this case. The poor labelling of the main horizontal edge dividing the two objects has already been mentioned, and there are very few other edges which contribute to the decision. The probability of the books being in front is calculated at 53%, so the ordering is correct, but by the smallest of margins.

The processing of the two frames can be performed fairly quickly. On a 300MHz PII, it takes about a minute to segment a 2-motion sequence, and a few minutes to segment a 3-motion sequence. The majority of time is spent in the EMC loop, which has to be repeated 4 extra times in the 3-motion case to consider all possible layer orderings.

5 Conclusions and Future Work

A system for segmenting multiple motions is presented, which uses only the edge motions between two neighbouring frames. This captures most of the motion information in the scene while providing a reduction in the amount of information to be processed. The problem of initialising multiple motions is addressed by recursively splitting solutions based on a smaller number of motions, and the Maximum Description Length is used as an effective means of identifying the correct number of models.

The edges are adequately labelled, and the motions estimated, by the Expectation-Maximisation algorithm, but then constraints are applied to refine these. The region labelling restricts the set of possible edge labels and this is used in a larger 'Expectation-Maximisation-Constrain' loop to refine the motions and generate the globally optimal region segmentation. The layer ordering is also correctly extracted and accurate segmentations of the foreground objects produced.

Future work will refine the statistics for edge labellings, addressing the sensitivity of ambiguous edges, which is currently partly due to the assumption that edge tracking nodes are independent. However, ambiguous edges can only be correctly resolved by considering information from a larger number of frames, where the motion will hopefully be such that their correct labelling becomes clear. The correct integration of information from more than two frames will enable a truly robust system to be developed, and should lead to efficient methods of performing the segmentation of a complete sequence.

Acknowledgements

This research was funded by a United Kingdom EPSRC studentship, with a CASE award from the AT&T Laboratories, Cambridge, UK. Thanks go to David Sinclair for the use of his image segmentation code and to Ken Wood for many useful discussions.

References

[1] S. Ayer, P. Schroeter, and J. Bigün. Segmentation of moving objects by robust motion parameter estimation over multiple frames. In *Proc. 3rd ECCV*, volume II, pages 317–327, Stockholm, Sweden, May 1994.

[2] L. Bergen and F. Meyer. Motion segmentation and depth ordering based on morphological segmentation. In *Proc. 5th ECCV*, volume II, pages 531–547, Freiburg, Germany, June 1998.

[3] A. P. Dempster, H. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B*, 39:1–38, 1977.

[4] T. Drummond and R. Cipolla. Visual tracking and control using lie algebras. In *Proc. CVPR '99*, volume 2, pages 652–657, Fort Collins, CO, June 1999.

[5] P.R. Giaccone and G.A. Jones. Segmentation of global motion using temporal probabilistic classification. In *Proc. 9th BMVC*, volume 2, pages 619–628, Southampton, September 1998.

[6] S. Hsu, P. Anandan, and S. Peleg. Accurate computation of optical flow by using layered motion representations. In *Proc. 12th ICPR*, pages 743–746, Jerusalem, Israel, October 1994.

[7] P. J. Huber. *Robust Statistics*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, 1981.

[8] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 12(1):5–16, January 1994.

[9] F. Moscheni and F. Dufaux. Region merging based on robust statistical testing. In *Proc. SPIE VCIP '96*, Orlando, Florida, USA, March 1996.

[10] J. M. Odobez and P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In *Video Data Compression for Multimedia Computing*, pages 283–311. Kluwer Academic Publisher, 1997.

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

[12] D. Sinclair. Voronoi seeded colour image segmentation. Technical Report 1999.3, AT&T Laboratories Cambridge, 1999.

[13] P. Smith, T. Drummond, and R. Cipolla. Motion segmentation by tracking edge information over multiple frames. In *Proc. 6th ECCV*, volume II, pages 396–410, Dublin, Ireland, June 2000.

[14] P. H. S. Torr. An assessment of information criteria for motion model selection. In *Proc. CVPR '97*, pages 47–53, San Juan, PR, June 1997.

[15] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated Bayesian approach to layer extraction from image sequences. In *Proc. 7th ICCV*, volume II, pages 983–990, Kerkyra, Greece, September 1999.

[16] J. Y. A. Wang. and E. H. Adelson. Layered representation for motion analysis. In *Proc. CVPR '93*, pages 361–366, New York, NY, June 1993.

[17] Y. Weiss and E. H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proc. CVPR '96*, pages 321–326, San Francisco, CA, June 1996.