Robust Motion Trajectory Estimation for Long Image Sequences

David Gibson and Michael Spann
School of Electronic and Electrical Engineering,
The University of Birmingham,
52 Pritchatts Road, Edgbaston, B15 2TT, UK.
Email: gibsond@eee.bham.ac.uk, spannm@eee.bham.ac.uk

Abstract

This paper presents a new approach for the estimation of motion trajectories [1] from image sequences. The approach is not limited to the tracking of feature points, but relies on the matching of image patches in consecutive image frames. Motion trajectories are parametrically modelled and the problem is considered from an optimisation perspective, using a Markov Random Field (MRF) framework. A three stage optimisation procedure is used to determine the trajectory parameters, and a recursive estimation approach is developed. A method for the robust detection of occlusion regions is also developed. Although robustly computed motion trajectories show great promise in many motion-related areas of computer vision [2,3,4], this paper illustrates their use in the context of video compression.

Keywords: Trajectories
Motion

1. Introduction

Traditionally the extraction of motion trajectories has been based on feature extraction and tracking. Although this approach has key benefits in terms of processing efficiency, it generally results in a sparse set of trajectories, whose distribution through the image is very dependant upon the location of feature points, which are typically on object boundaries. The approach considered here aims to track 'interesting' image patches - defined simply as those image areas which differ significantly from their neighbouring image patches. Such image patches are assumed to move based on a 3D constant acceleration model. The trajectories can be described in terms of a small set of motion parameters and thus the problem is converted from that of 'trajectory estimation' to that of 'trajectory parameter estimation'. Given a whole set of trajectories (with associated parameters), the problem can be considered from a Markov Random Field (MRF) point of view. This allows an energy function to be defined which measures the desirability of a given trajectory, given the sequence data and neighbouring trajectories. A three stage optimisation process is used to facilitate the robust handling of different image sequences. This is used in conjunction with a robust occlusion detection algorithm, which is used to temporally delimit the trajectories.

This paper is structured as follows; Section 2 provides an introduction and mathematical framework for motion trajectories. Section 3 considers the motion trajectory estimation problem in terms of a MRF model, highlighting both qualitatively and mathematically the key features (expressed as terms of an energy function). Section 4 discusses the optimisation process used to actually determine the motion trajectory set and shows how this optimisation process can be applied recursively. Section 5 details an approach developed to detect areas of occlusion and to compensate for the effects which occlusion areas have on motion trajectories. Section 6 illustrates the use of motion trajectories for video compression. The penultimate section presents a set of results comparing the accuracy of the motion trajectories estimated using the approach detailed in this paper, to that of a feature-based tracking method and assesses the success of using motion trajectories for video compression. The last section presents some conclusions.

2. Motion Trajectory Estimation

A motion trajectory can be considered as a spatio-temporal curve in a spatio-temporal space. The spatio-temporal space can be constructed from a temporal stack of image frames, extracted from the image sequence. Given this definition, a trajectory can be used to express the motion of a small image patch on the image plane as a function of time. Assuming that

the 3D object patch, corresponding to the 2D image patch, follows a motion governed by a simple constant acceleration model, then it is possible to model the position of the 3D object patch mathematically as;

$$\begin{pmatrix} X(t) \\ Y(t) \\ Z(t) \end{pmatrix} = \begin{pmatrix} a_0 + a_1 t + a_2 t^2 \\ b_0 + b_1 t + b_2 t^2 \\ c_0 + c_1 t + c_2 t^2 \end{pmatrix} \quad (1)$$

Assuming simple projective geometry, the mapping of a point on an object onto the image plane (x(t),y(t)), can be described as;

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} X(t)/Z(t) \\ Y(t)/Z(t) \end{pmatrix} \quad (2)$$

The motion of an image patch, as a function of time can then be described as;

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \left(\frac{a_0 + a_1 t + a_2 t^2}{c_0 + c_1 t + c_2 t^2}\right) \\ \left(\frac{b_0 + b_1 t + b_2 t^2}{c_0 + c_1 t + c_2 t^2}\right) \end{pmatrix} \quad (3)$$

Furthermore, as only monocular sequences are used, it is possible to set, $c_0=1$, without any loss of generality. Thus a motion trajectory p can be uniquely described by a set of 8 parameters $\phi_p=\{a_2, a_1, a_0, b_2, b_1, b_0, c_2, c_1\}$. This ignores the problem of object occlusion, which shall be considered shortly.

Defining a complete set of motion trajectories, $\Phi=\{\phi_1, \phi_2, \phi_3, \ldots, \phi_n\}$, it is possible to describe the motion throughout the entire image, with the exception of areas of insufficient spatial texture[1]. The estimation of a complete set of trajectories is essentially that of estimating all of the parameters, for all of the trajectories in Φ.

3. MRF Approach

This parameter estimation procedure is performed by casting the problem in a MRF framework. Each trajectory is considered as an MRF site, with the collection of trajectory parameters, $\Phi=(\phi_p : p=1\ldots n)$, considered as random variables.

The maximum a posteriori (MAP) criterion has been used such that,

$$\hat{\Phi} = \max_{\{\Phi\}} p(\Phi) \quad (4)$$

Where $p(\Phi)$ is the joint probability function. Using the well known MRF-Gibbs equivalence relationship [6],

$$p(\Phi) = \frac{1}{Z} e^{-E(\Phi)} \quad (5)$$

Where $E(\Phi)$ is the energy function of the MRF.

$$E(\Phi) = \sum_{c \in C} V_c(\Phi) \quad (6)$$

The set C denotes the cliques associated with some neighbourhood system v. This neighbourhood definition is based on spatial and temporal separation of two trajectories. For trajectories to be considered neighbours, their mean spatial separation must be below a given threshold, and they must overlap temporally. Hence trajectory p' is a neighbour of trajectory p if, (See Appendix A for a list of mathematical notation)

$$\left\| X_p(0) - X_{p'}(0) \right\| < \alpha_D$$

$$\frac{\sum_{t=0}^{T-1} D_p(t).D_{p'}(t)}{\frac{1}{2}\left(\sum_{t=0}^{T-1} D_p(t) + \sum_{t=0}^{T-1} D_{p'}(t)\right)} > \alpha_N \quad (7)$$

In this case the set C is defined as,

$$C = C_1 \cup C_2 \quad (8)$$

The sets C_1 and C_2 are defined such that C_1 contains the set of all trajectories, and C_2 contains the set of all neighbouring trajectory pairs (See Fig. 1)

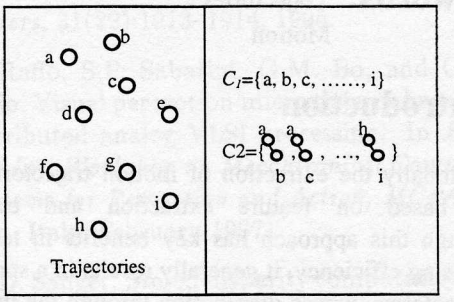


Figure 1 : Illustration of clique sets C_1 and C_2.

Hence the energy function, $E(\Phi)$, can be re-written as[2] a sum of unary (dependant on C_1) and binary (dependant on C_2) energy terms;

[1] Typically this involves areas of the image in which motion estimation would prove counter productive to the global estimation of motion.

[2] Note the implied dependency of $E_1(\phi_p)$ upon the image data.

$$E(\Phi) = \sum_{p \in C_1} E_1(\phi_p) + \sum_{\{p,p'\} \in C_2} E_2(\phi_p, \phi_{p'}) \qquad (9)$$

The key unary energy terms relating a trajectory to the image sequence data can be summarised as follows:

- Trajectory Suitability (TS) - This considers the match between image patches of consecutive frames. Referring to Fig. 2, it can be seen that the trajectory intersects each of the image frames of the sequence. If the trajectory accurately describes the motion present in the image sequence, the small image patches should be similar to each other. The approach taken in this research is to consider the match of image patches in consecutive image frames using the covariance function as a measure of image patch similarity.

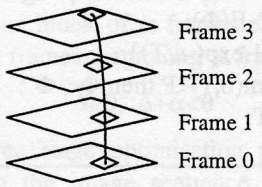


Figure 2 : Illustration of the 'Trajectory Suitability' energy term.

Bi-linear interpolation is used to allow non-pixel spaced image patch matching to occur. Furthermore, as the 3D velocity of the image patch is estimated explicitly, the covariance is performed on differently sized image patch grids, to compensate for expansion or contraction of the image patch.

- Spatial Uniqueness (SU) - To avoid trajectories passing through untextured areas, this energy term measures the spatial uniqueness of an image patch. Ideally an image patch should be significantly different from surrounding image patches as illustrated in Fig. 3.

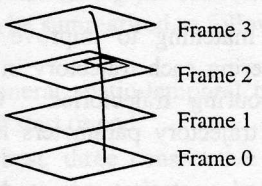


Figure 3 : Illustration of the trajectory image patch and image patches adjacent to it.

These two energy terms can be expressed mathematically as shown in Table 1.

Energy Term	Equation
Trajectory Suitability (TS)	$(E_{TS})_p = \frac{K_{TS}}{1 + \frac{\sum_{t=0}^{T-2} D_p(t).D_p(t+1).S(\underline{X}_p(t),\underline{X}_p(t+1))}{\alpha_{TS}.(a_p^e - a_p^s)}}$
Spatial Uniqueness (SU)	$(E_{SU})_p = \frac{-K_{SU}}{1 + \frac{\sum_{t=0}^{T-1} U(\underline{X}_p(t)).D_p(t)}{\alpha_{SU}.(a_p^e - a_p^s + 1)}}$ $U(\underline{X}_p(t)) = \frac{1}{u*v-1} \sum_{i=-u}^{u} \sum_{j=-v}^{v} S\left(\underline{X}_p(t), \underline{X}_p(t) + \begin{pmatrix} i \\ j \end{pmatrix}\right)$ excluding $i = j = 0$

Table 1 : Summary of unary energy terms. (See Appendix A for a list of mathematical notation)

The key binary energy terms defined for all pairs of neighbouring trajectories can be stated as follows;

- Trajectory Compatibility (TC) - This energy term considers the similarity (in terms of velocity and acceleration) between neighbouring trajectories. (See Fig. 4) This 'smoothes' out the estimated velocity field to help constrain the velocity field in areas which are not heavily textured. This is similar in approach to the smoothing effect often used when determining optical flow from two frames. However for the multiple frame case, less smoothing is required and so over smoothing is not such a problem as it often is in the two frame case.

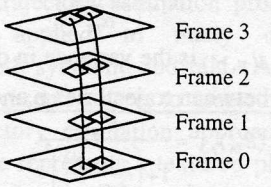


Figure 4 : Neighbouring trajectories should be similar.

- Trajectory Adjacency (TA) - If two trajectories represent exactly the same curve, one of the trajectories is clearly redundant. This is clearly not desirable and so the trajectory uniqueness energy term penalises trajectories which are considered too close to be useful. (See Fig. 5)

- Discontinuity Coherence (DC) - If an occlusion or disocclusion region is present in the sequence, a trajectory passing through this region will terminate prematurely (or begin belatedly). It is expected that

adjacent trajectories should also terminate prematurely (or begin belatedly) in approximately the same frame. If this is not the case, it is likely that the trajectory discontinuity[3] is misplaced and hence will incur a penalty in the energy function. (See Fig. 6)

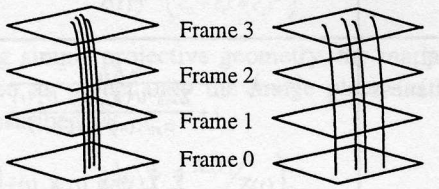


Figure 5 : Spatially separated trajectories (right) are more desirable.

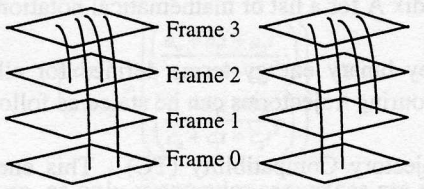


Figure 6 : Neighbouring trajectories with vastly separated (temporally) discontinuities (left) are penalised.

These three energy terms are expressed mathematically in Table 2.

Energy Term	Equation
Trajectory Compatibility (TC)	$(E_{TC})_{p,p'} = \frac{K_{TC}}{1+\left(\frac{l(p,p')}{\alpha_{TC}}\right)}$ $l(p,p')$ is the variance in distance between trajectories p and p'.
Trajectory Adjacency (TA)	$(E_{TA})_{p,p'} = \frac{K_{TA}}{1+\left(\frac{m(p,p')}{\alpha_{TA}}\right)^2}$ $m(p,p')$ is the mean distance between trajectories p and p'.
Discontinuity Coherence (DC)	$(E_{DC})_{p,p'} = K_{DC}\cdot\left[\begin{array}{l} r\left(\left\|\underline{X}_p(a_p^s)-\underline{X}_{p'}(a_{p'}^s)\right\|,(a_p^s-a_{p'}^s)\right) \\ +r\left(\left\|\underline{X}_p(a_p^e)-\underline{X}_{p'}(a_{p'}^e)\right\|,(a_p^e-a_{p'}^e)\right)\end{array}\right]$ $r(b,c)=\begin{cases}0 & \text{if } b<\alpha_{SD} \text{ and } c<\alpha_{FD}\\ 1 & \text{otherwise}\end{cases}$

Table 2 : Summary of binary energy terms. (See Appendix A for a list of mathematical notation)

[3] Defined as the temporal start or end of a trajectory.

4. Optimisation

Initial experiments considering an energy function containing all of the energy terms detailed in Section 3, gave results which lacked robustness to changes in image sequence data. The approach considered here uses a three stage optimisation procedure, with each stage using an energy functions, $E(\phi)$ consisting of different energy terms, as defined in Table 3. The three optimisation stages, defined below, all rely upon the simulated annealing algorithm (See Fig. 7) to minimise their respective energy functions.

```
Initialise Φ;
Initialise T;
Repeat;
        Let Φ' be a perturbed version of Φ;
        d=E(Φ')-E(Φ);
        P=min{1,exp(-d/T)};
        if random(0,1)<P then Φ<-Φ';
        reduce T;
until(T=0)
```

Figure 7 : Illustration of the Simulated Annealing Algorithm.

Stage 1: Regularise Position

Allows pure translation of motion trajectories (maximum 2-4 pixels) to redistribute them throughout the image, whilst confining them to sufficiently textured areas.

Stage 2: Optimise Trajectory Parameters

This is the main optimisation stage, which fits the trajectories to the available image data, varying only velocity or acceleration components.

Stage 3: Refine Trajectories

Uses sub pixel matching to improve the trajectory accuracy, considering each trajectory in isolation, and ignoring neighbouring trajectories. Only a small variation in the trajectory parameters is permitted in this stage.

The three stage optimisation process can be applied iteratively to the same image data, or more usefully, can be used to provide a recursive approach to trajectory estimation.

	Unary Terms		Binary Terms		
	TS	SU	TA	TC	DC
Regularise Position		√	√		
Optimise trajectory parameters	√				√
Refine trajectories	√				

Table 3 : Illustration of the different energy terms used in the three optimisation stages.

Given an image sequence defined over θ frames, the above three stage optimisation process can be applied to a contiguous block of α frames extracted from the sequence of θ frames, beginning at frame Δ, such that,

$$\Delta \geq 0, \ \Delta+\alpha<\theta \qquad (10)$$

Once the three stage optimisation process has been performed on the image sequence subset, $\{\Delta, \Delta+1, \ldots\ldots, \Delta+\alpha-1\}$, the trajectory estimates can be temporally shifted to correspond to the image sequence set, $\{\Delta+1, \Delta+2, \ldots\ldots, \Delta+\alpha\}$. The trajectory parameters can then be re-optimised to take into account the new image data. This process can be performed recursively for the entire image sequence of θ frames.

5. Occlusion Detection and Compensation

It is unlikely that all trajectories will be appropriately defined over the entire temporal region under consideration. Trajectories intersecting occlusion regions, or exiting the boundaries of the image should be avoided, and replaced with trajectories which are temporally delimited; that is they terminate prematurely, or begin belatedly. An algorithm developed for estimating the start and end frames of a trajectory can be summarised as follows,

- Take a general spatio-temporal trajectory with no predefined start or end.
- Find the best, three consecutive frame matches - that is where the trajectory best fits the image sequence data.
- Extend the motion trajectory forwards and backwards in time until the trajectory either goes outside the image boundary, or encounters a significant drop in covariance between temporally adjacent image patches.

6. Video compression

A popular approach to the compression of video sequences is to employ a motion compensated predictor. Typically this involves the use of motion estimates to predict frame 'n+1' given any number of previous frames. Hence frame 'n+1' can be represented by a combination of motion information and a coded version of the residual[4]. The motion information can be conveniently represented by a set of trajectories. The simple approach considered here assumes that the motion of any point on the image can be described by a spatially adjacent trajectory. All neighbouring trajectories are considered as possible motion descriptor candidates, and the one which most closely describes the local image motion is used. In this way a 'pixel spaced' field of motion trajectories is created. Hence if the image frames {n, n-1, n-2, ...} of a sequence are known, it is possible to estimate the image frame {n+1} by temporally extending the trajectory set. Also, as some variation of intensity along each trajectory is inevitable, this can be modelled (in a LS sense) by a low order polynomial. This is related to an approach suggested in [7] which uses robustly computed feature correspondences (computed over two frames) to link pixels having a similar motion.

7. Results

The motion trajectory estimation procedure has been successfully applied to many different image sequences; both synthetic and real. A single synthetic sequence is presented here, to asses the success of the motion trajectory estimation approach. (See Fig. 8) This sequence represents a textured planar object with a constant acceleration 3D motion.

To quantify the success of the motion trajectory approach, a measure of the optical flow error (in some chosen frame of the sequence) has been used,

$$\varepsilon = ||u-v|| \qquad (11)$$

Where, 'u' represents the optical flow estimated from the motion trajectory, and 'v' represents the known value of the optical flow at that point.

[4] The residual being the difference between the motion compensated prediction and the actual image data.

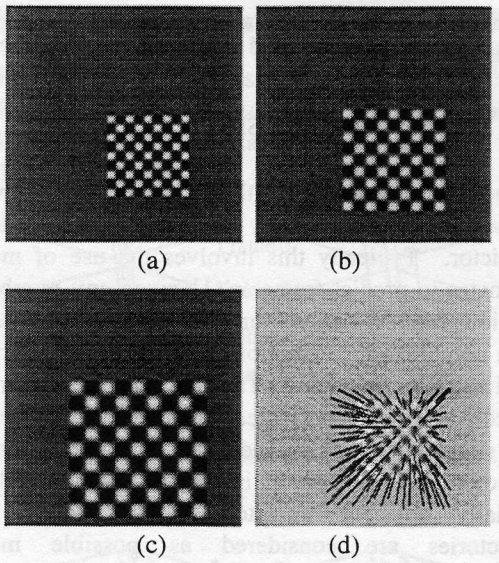


Figure 8 : (a-c) Frames 1, 4 and 7 of the synthetic sequence, (d) Estimated trajectory set. (one third of trajectories shown)

Fig. 9(a-c) shows the error (ε) probability distributions for three different optical flow estimation procedures[5]. These three procedures are,

- The motion trajectory estimation procedure considered in this paper.
- A two frame feature matching procedure developed by Zhang Z.Y. et. al. [5]
- A multi-frame version of the two frame feature matching procedure considered above. A motion trajectory is fitted, in the least squares (LS) sense, to a series of linked feature points.

As can be seen, the approach presented here produces an optical flow error of less than one-tenth of a pixel for the vast majority of trajectories for this image sequence. Note that the feature matching approach detected typically 150 features per frame, with the motion trajectory approach having a similar number of trajectories[6]. Furthermore, the periodic nature of the texture pattern in the sequence, caused some initial matching problems for the feature based approaches, which did not occur in the motion trajectory based solution.

[5] In the case of the trajectory approaches, the optical flow is derived simply from the motion trajectory parameters.

[6] Although the number of trajectories used can be increased or decreased as required.

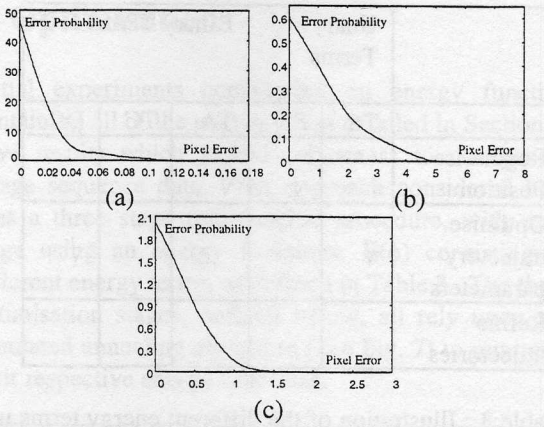


Figure 9 : Optical flow error distributions, (a) Motion trajectory approach, (b) Two frame feature matching algorithm, (c) Multi-frame feature matching and tracking algorithm.

For the purpose of video compression a pair of real, outdoor sequences have been used. Fig. 10 illustrates two frames from the first sequence, and a corresponding set of motion trajectory estimates. The scene is static, with camera motion from left to right.

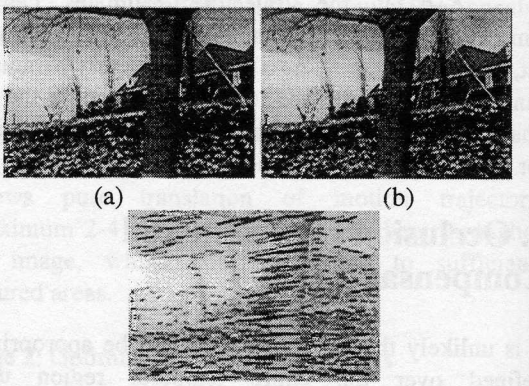
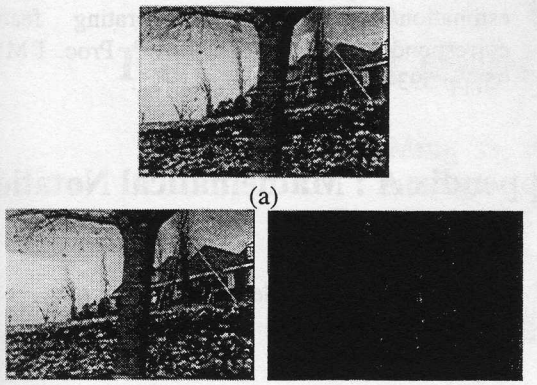


Figure 10 : (a) Frame 1 of the 'flower garden' sequence, (b) Frame 7 of the 'flower garden' sequence, (c) Motion trajectory estimate.

The set of trajectories was calculated for the first eight frames, with the motion compensated prediction based on the first seven frames of the sequence along with the set of motion trajectories. Fig. 11 illustrates the estimated eighth frame, the actual eighth frame, and the difference between the two. The motion compensated prediction effectiveness is measured using the RMS of the intensity error between the actual image data and the motion compensated prediction. The results gained from the motion trajectory approach are compared with an 8*8 block matching approach (see Table 4). This block matching approach considers translated blocks of 8*8 pixels such as to minimise the displaced frame

difference (DFD). The total number of bits required to encode the data for the block matching approach is comparable to that of the motion trajectory approach.



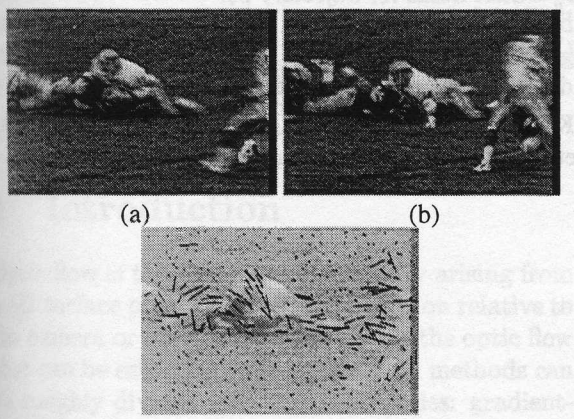
(a)

(b) (c)

Figure 11 : (a) Estimated eighth frame, (b) Actual eighth frame, (c) Difference between estimated and actual eighth frames. (exaggerated error)

	RMS error
Motion Trajectory Approach	11.33
8*8 Block Matching Approach	13.46

Table 4 : Comparison of the motion trajectory approach with a 8*8 block matching approach for the 'flower garden' sequence.



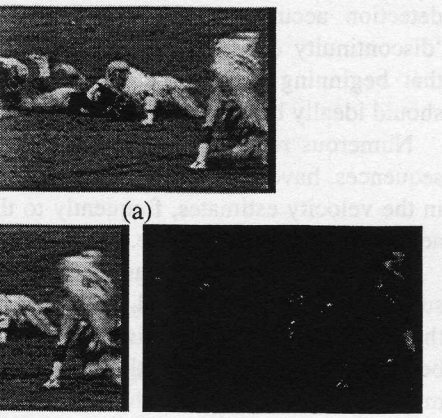
(a) (b)

(c)

Figure 12 : (a) Frame 1 of the 'football' sequence, (b) Frame 4 of the 'football' sequence, (c) Motion trajectory estimate.

Fig. 12 illustrates two frames from the second sequence, and a corresponding set of motion trajectory estimates. The sequence is that of an American football game, with large, complex non-rigid body motion and high levels of occlusion.

For this sequence the set of trajectories was computed over the first five frames, as the image motion only loosely fits the constant 3D acceleration model. The motion trajectory estimates, along with the first four frames of image data were used to generate a motion compensated prediction, as before; this time for frame five (see Fig. 13). The RMS error comparison with the BMA approach is shown in Table 5.



(a)

(b) (c)

Figure 13 : (a) Estimated fifth frame, (b) Actual fifth frame, (c) Difference between estimated and actual fifth frames. (exaggerated error)

	RMS error
Motion Trajectory Approach	9.04
8*8 Block Matching Approach	10.11

Table 5 : Comparison of the motion trajectory approach with a 8*8 block matching approach for the 'football' sequence.

8. Conclusions

In this paper a novel approach to the extraction of motion trajectories has been proposed. Primarily this approach is concerned with the matching of small image patches between consecutive frames of an image sequence. Given this, the trajectory estimation problem has been considered from an optimisation point of view within a MRF framework. So far this approach has shown considerable robustness to different image sequence. The estimation process is also relatively insensitive to the weighting parameters $\{K_{TS}, K_{SU}, K_{TC}, K_{TA}, K_{DC}, \alpha_{TS}, \alpha_{TU}, \alpha_{TC}, \alpha_{TA}, \alpha_{SD}, \alpha_{FD}\}$ used in the MRF energy function calculation. Furthermore, the approach maintains the ability of most feature tracking algorithms to cope with large motion values, whilst providing a more dense velocity field, as would be

expected with traditional optical flow estimation procedures.

Areas of occlusion, often pose some difficulty for two and multiple frame velocity estimation algorithms. This research provides both a simple and intuitive approach to this problem, which is derived as a natural extension to the motion trajectory philosophy; that is, motion trajectories should define image patch movement over a temporal interval. Occlusion detection accuracy is improved by the use of the 'discontinuity coherence' energy term, which dictates that beginning and end of neighbouring trajectories should ideally be temporally similar.

Numerous results on both synthetic and real image sequences, have consistently shown sub-pixel accuracy in the velocity estimates, frequently to the extent of an accuracy of 0.1 pixels/frame.

Although this research has been used in the fields of surface reconstruction, and video compression; only the latter is considered in this paper. Some results have been shown to illustrate the effectiveness of using motion trajectories as the basis for a motion compensated predictor as part of a video compression algorithm.

9. References

[1] Shah M, Rangarajan K, Tsai PS, "Motion trajectories", IEEE Systems, Man and Cyber. , Vol 23, No 4, July/August 1993.

[2] Bradshaw KJ, Reid ID, Murray DW, "The active recovery of 3D motion trajectories and their use in prediction", IEEE PAMI, Vol 19, No 3, March 1997.

[3] Leduc JP, Odobez JM, Labit C, "Adaptive motion-compensation wavelet filtering for image sequence coding", IEEE trans. Image Processing, Vol 6, No. 6, June 1997.

[4] Allman MC, "Image sequence description using spatiotemporal flow curves: Towards motion-based recognition", Ph.D. Thesis, University of Wisconsin, 1991.

[5] Zhang ZY, Deriche R, Faugeras O, Long QT, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry", Technical report No. 2273, INRA, France, 1994.

[6] Li SZ, "Markov random field modelling in computer vision", Springer-Verlag, 1995.

[7] James PD, Spann M, "Multiresolution motion estimation/segmentation incorporating feature correspondence and optical flow", Proc. BMVC 95, pp 593-602, Birmingham 1995.

Appendix A : Mathematical Notation

$D_p(t) = \begin{cases} 1 & \text{if trajectory p is defined in frame t,} \\ 0 & \text{otherwise} \end{cases}$

$\underline{X}_p(t) = \begin{pmatrix} x \\ y \end{pmatrix}$ Position of trajectory p, at time t.

$S(X_p(t), X_p(t+1))$ = Similarity measure between two image patches at co-ordinates given by trajectory p, in frames (t) and (t+1).

$S\left(X_p(t), X_p(t)+\begin{pmatrix} i \\ j \end{pmatrix}\right)$ = Similarity measure between two image patches, both in frame (t), the first with co-ordinates given by trajectory p, the second with co-ordinates the same as the first except for a shift (i,j).

a_p^s = Start frame for trajectory p.

a_p^e = End frame for trajectory p.

$\{K_{TS}, K_{SU}, K_{TC}, K_{TA}, K_{DC}, \alpha_{TS}, \alpha_{TU}, \alpha_{TC}, \alpha_{TA}, \alpha_{SD}, \alpha_{FD}\}$ - A set of MRF parameters; determined empirically.