A NEW GLOBAL MOTION ESTIMATION ALGORITHM AND ITS APPLICATION TO RETRIEVAL IN SPORTS EVENTS

A. Kokaram and P. Delacourt *
Electronic and Electrical Engineering Department,
Trinity College, Dublin 2, Ireland
{*anil.kokaram,perrine.delacourt*}*@tcd.ie*

Abstract - In this paper, we propose a novel global motion estimation technique based on weighted gradient and Displaced Frame Difference (DFD) associated with Wiener estimation. Then, we apply this technique to parse events of a high level of understanding in a cricket game. A user oriented analysis of the game then reveals a distinct connection between the global motion and specific events. By estimating global motion and analysing the temporal evolution of the estimated motion parameters, we present an effective process for the extraction of cricket events, leading to a succes rate of 88.9%.

INTRODUCTION

There has been considerable investigations in the role that motion plays in assisting the retrieval of information from video [1, 2, 3]. It is expected that motion should play a part in the analysis of video sequences for the purposes of information extraction principally because it is such an important feature of interesting image sequences.

Sports events have been considered in [4, 5] (basket ball, baseball). This papers deals with a highly structured sports events that has not been yet studied: cricket. As far as television coverage is concerned, the sport is typified by stylised camera motion which is matched to the behaviour of the players in the field. Without delving into the specific nature of the game, there are three "editing' events of immediate importance to the viewer: the bowler run up, the batsman's stroke and the direction of the ball after being hit (for more details, see *www.cricket.org*). Since the game play time can extend over five days for 6 hours per day, it is sensible to consider an automated mechanism for retrieving useful information. This would normally require a very high level of understanding of the event, but because the camera movement is so well matched to the game, in fact we can deduce these high level events directly from the observation of Global motion of the scene.

A new global motion estimation algorithm is described in the next section. The following section shows how the observation of motion matches the play to a very high level of abstraction. Results of an automated analysis of cricket videos based on temporal evolution of motion estimates are then presented. Finally, we make final comments and give several directions for future investigations.

*Work funded by EU Project MOUMIR (**Mo**dels for **U**nified **M**ultimedia **I**nformation **R**etrieval) *http://www.moumir.org*

GLOBAL MOTION ESTIMATION

The work here employs an image sequence model as follows

$$I_n(\mathbf{x}) = I_{n-1}(\mathbf{F}(\mathbf{x}, \mathbf{a})) + \epsilon(\mathbf{x}) \tag{1}$$

where $I_n(\mathbf{x})$ is the grey level of the pixel at the location given by position vector x in the frame n. The vector function $\mathbf{F}(\mathbf{x}, \mathbf{a})$ is an affine linear transformation of image coordinates to represent motion such as zooming, rotation and translation between the current frame n and the previous frame $n-1$. $\mathbf{a}$ is the vector formed by the motion parameters. In this paper, the transformation $\mathbf{F}(\mathbf{x}, \mathbf{a}) = \mathbf{A}\mathbf{x} + \mathbf{d}$ is used where $\mathbf{A}$ is a 2×2 matrix for affine transformation and $\mathbf{d}$ is the displacement vector. Then, the parameter vector is equal to: $\mathbf{a} = (a_1\ a_2\ a_3\ a_4\ d_1\ d_2)$, also denoted $\mathbf{a} = (\mathbf{A}, \mathbf{d})$, where a_i are the matrix coefficients of $\mathbf{A}$ ($A = [a_1 \quad a_2; a_3 \quad a_4]$ in Matlab notation) and d_i are the vector coefficients of the translational motion component $\mathbf{d}$.

An implicit function in extracting global motion is the segmentation of each frame of the sequence into areas which move according to this global motion and those which do not [6, 7]. This is achieved here by using a novel scheme for gradient based estimation which implicitly removes the effect of one of these two areas. Therefore, $\mathbf{a}$ is chosen to minimise a weighted DFD function over the whole image as follows

$$\text{WDFD}(\mathbf{x}, \mathbf{a}) = \sum_{\mathbf{x}} w^2(\mathbf{x})(I_n(\mathbf{x}) - I_{n-1}\mathbf{F}(\mathbf{x}))^2 \tag{2}$$

$w(\mathbf{x})$ can be chosen such that it is 1 at sites undergoing global motion and 0 otherwise. In effect, the problem now is to estimate the local/global segmentation of the sequence given by $w(\mathbf{x})$ implicitly *and* to estimate $\mathbf{a}$ where $w(\mathbf{x}) = 1$. This problem is solved iteratively by alternating between estimates of $w(\mathbf{x})$ and $\mathbf{a}$. Thus $w(\cdot)$ is estimated given an estimate for $\mathbf{a}$ then $\mathbf{a}$ is calculated given the previously estimated $w(\mathbf{x})$.

Estimation of the weights w is derived from a robust histogram based technique (see [7]). Assuming most of the area of the frame is occupied by the 'global object', a histogram of the |DFD| across all pixels should show a large mode near a low error corresponding to this object. This is assuming of course that the estimation of the motion is influenced more by the larger amount of data associated with this object. Therefore, the weights could be determined by deriving a threshold from this histogram that excludes the highest T%. Then, $w(x) = 0$ if $\mathbf{z}(\mathbf{x}) \geq T$, otherwise, $w(\mathbf{x}) = 1$.

An estimate for $\mathbf{a}$ is derived as the solution to the weighted least squares problem presented in 2. Using a Taylor series expansion around an initial guess for the motion parameters $\mathbf{a}_0 = [\mathbf{A}_0, \mathbf{d}_0]$ such that $\mathbf{a} = \mathbf{a}_0 + \mathbf{u}$, results in

$$\begin{aligned}\text{WDFD}(\mathbf{x},\mathbf{a}) = \sum w^2(\mathbf{x})(I_n(\mathbf{x}) - I_{n-1}(\mathbf{A}^0 x + \mathbf{d}^0) \\ - \mathbf{u}^T \nabla I_{n-1}(\mathbf{A}^0 x + \mathbf{d}^0) + e_{n-1}(\mathbf{x}))^2\end{aligned} \tag{3}$$

where $e_{n-1}(\mathbf{x}, \mathbf{a^0})$ represents the higher order terms of the expansion and $\epsilon(\cdot)$ lumped together as a Gaussian r.v. with variance σ_e^2. The ∇ operator is the usual multidimensional gradient operator. The update term $\mathbf{u}$ is intended to be small for the Taylor expansion to be valid. Gradient based motion estimation is only valid for small motion,

therefore three levels of a multiresolution pyramid are used to estimate the motion using coarse to fine refinement.

The equation can be rewritten in matrix form as $[\mathbf{z}_w - \mathbf{G}_w\mathbf{u}]^T[\mathbf{z}_w - \mathbf{G}_w\mathbf{u}]$ where $\mathbf{z}_w$, $\mathbf{G}_w$ are weighted forms of DFD and Gradient matrices respectively. The Wiener solution to this weighted least squares problem yields the following estimate for $\mathbf{u}$.

$$\hat{\mathbf{u}} = [\mathbf{G}_w^T\mathbf{G}_w + \mu_w I]^{-1}\mathbf{G}_w^T\mathbf{z}_w \quad \text{with} \quad \mu = \left(\frac{\sigma_{ee}^2}{\sigma_{uu}^2}\right)_w \tag{4}$$

where σ_{uu}^2 is the variance of the estimate for $\hat{\mathbf{u}}$ and μ is also weighted. In practice, μ is adaptively set at each iteration: $\mu = ||\mathbf{z}_w||\frac{\lambda_{MAX}}{\lambda_{MIN}}$ where $\frac{\lambda_{MAX}}{\lambda_{MIN}}$ is the condition number of $\mathbf{G}_w^T\mathbf{G}_w$. This solution is similar to that introduced in [8] with the notable exception that a weighted error criterion is used which changes the importance of each data point included in the estimation process.

PARSING CRICKET EVENTS (RESULTS)

Figure 2 (left plot) shows the global motion information extracted as a function of time across several notable events in a game of Cricket[1]. QCif image sizes were used. As far as the game itself is concerned, when the bowler is running into deliver the ball to the batsman, the camera zooms in on the 'pitch' or play area from a wide angle to a tight shot of the batsman himself. After the batsman makes a stroke, the camera then attempts to follow the ball using a combination of a zoom out to a wide angle to find the ball then a panning operation to keep the ball in the field of view. There is a delay between the batsman's stroke and the reaction of the cameraman or the editor to then track the ball. Also, the editor of the live event may in real time cut after the stroke to another camera view that is about to track the ball. That view is initially full of motion transients as the camera operator tries to find the best view. This may take just 10 frames (on the order of the reaction time of the cameraman).

Therefore, it is typical that every bowler/batsman delivery combination is indicated by a zoom in followed by a zoom out or a cut before resuming a tracking shot, which is dominated by pan. This means that the diagonal affine transformation parameters should be lower than 1 before the stroke and a few frames afterwards. Then, they should show a noticeable discontinuity if a scene cut occurs. After the scene cut, there might be some transients in camera motion, then these parameters revert to a value of 1, indicating little or no zoom. This is indicated extremely well by the motion plots in figure 1. The observation of the behaviour of the motion parameters matches exactly with human observation.

After making the stroke, the direction of the pan can indicate whether the stroke played was on the offside (right hand side for a right hander) or onside (left hand side for a right hander). The translation parameters are therefore positive or negative depending on this motion, and the stroke direction can be diagnosed from this as well. This is extremely valuable information for coaching and viewing. Figure 1 shows this happening in two instances, verified by human observation.

[1]Bowler: MacLean, Batsman: Hussain; England Vs. West Indies 2000 in England

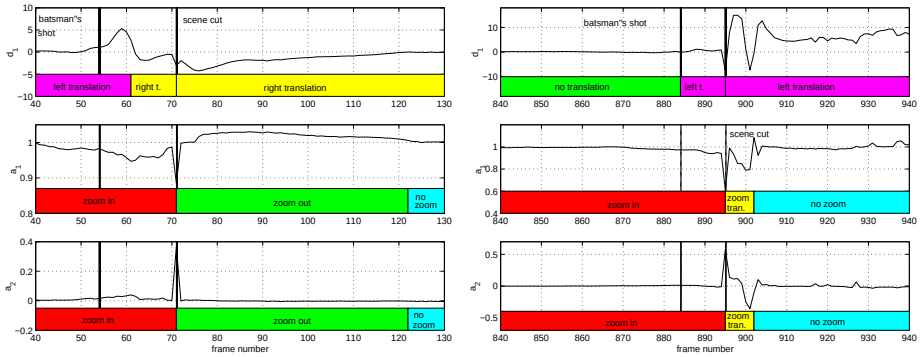

Figure 1: Left and right: two different bastsman' stroke. For each plot: translation and affine transformation parameter evolution (from top to bottom: d_1, a_1 and a_2) around a batsman's stroke followed by a scene cut and associated camera motion as seen by a human observer. The batsman's stroke is indicated by the left vertical line and the scene cut by the right vertical line.

Of course it is difficult to show these results in the paper as they relate to the temporal evolution of events. However, this is precisely the main point that must be made: i.e. it is not so important that the motion estimates be accurate. What is important is the temporal change in the motion information. It is this temporal change that indicates the onset and termination of events.

In order to fully assessed our global motion estimation algorithm, we have compared it to the one proposed in [6]. We used reference sequences (i.e. the motion parameters are well known). Both algorithms provided comparable results, close to the theoretical ones. [2]

It is interesting that scene cuts can also be indexed by the global motion parameters. Figure 2 (left plot) shows the scene cuts as detected by a human and how they correspond to impulsive events in the time domain motion tracks.

Automated analysis

As pointed out in the previous section, the mapping between the high level events and the camera motion is as follows:

1. Bowler run up and delivery: significant zoom in

2. Batsman's stroke: occurs during significant zoom in before the turning point in the zoom

3. Direction of stroke: direction of pan after bowler run up and delivery.

We approximate the zoom by the value of a_1 although it is appreciated that A allows for a more general motion model than pure zoom. It has already been established in the previous section that the events of interest in this paper corresponds to minima

[2]The authors would like to thank J.M.Odobez and P.Bouthémy, as well as IRISA, for making their *2D-motion* software available.

Actual deliveries	Nb of deliveries detected	False alarms	Missed detections
18	16	6	2

Table 1: Results of deliveries detection on 16500 frames of a cricket match.

in the track of a_1 with time. However, because the shot of the batsman as he makes his stroke is typically a tight zoom in, only significant minima are important. In this case, it is sufficient to accept minima as those corresponding to $a_1 < 0.98$.

Extrema in a_1 are detected by finding all turning points in the track $a_1(t)$. $a_1(t)$ is first processed by a median filter so that peaks corresponding to scene cuts are removed. Then, the resulting signal is low-pass filtered by a Gaussian filter and extrema are sought for in the filtered signal, by detecting zero-crossing in the corresponding gradient. Finally, only minima whose value are below a threshold (0.98) are retained. This is illustrated figure 2 (right plot). The middle graph shows $a_1(t)$ after median filter processing. The bottom graph shows the detected extrema. A value equal to 1 corresponds to a minimum and -1 indicates a maximum. The top graph displays the retained minima. These two minima correspond exactly to the two existing bowler run ups and deliveries over the first 1000 frames.

This first processing lead to a sizeable number of false alarms. Therefore, we used another feature to distinguish with real ball deliveries: the zoom duration. We imposed a minimum duration of 1.6 s. We also removed detected ball deliveries too close to one another (i.e. separated by 2 s at the most). Another feature that we will investigate in our future work is the shape of the zooms associated with ball deliveries.

We have experimented our algorithm on 16500 frames (a sequence of 9 min 47 s) and results are shown in table 1. By defining the success rate as the ratio between the number of deliveries correctly detected and the actual number of deliveries, we have obtained a success rate of 88.9%.

Concerning the batsman's stroke itself, no particular event appears in the motion parameter evolution, except that it takes place at the end of a significant zoom in before a turning point. Since the zoom related to the ball delivery lasts tens of frames, it means that once significant zooms have been detected, we are able to locate batsman's strokes by showing to the user the previous tens of frames. It is expected that the use of the audio track may enable a better time localisation of the stroke itself, since the stroke is characterised by a particular sound.

Finally, we have verified that the direction of the ball after the stroke is perfectly indicated by the value of the horizontal translation parameter d_1. In our context, a value lower than 0 corresponds to a right translation of the picture (see left plot in figure 1), i.e. the ball has been hit to the left.[3]

FINAL COMMENTS

This paper has shown the importance of camera motion in the diagnosis of this particular scenario. There is a one-to-one correspondance between motion and events

[3]Indicative frames from the video sequence are shown at: *www.mee.tcd.ie/∼sigmedia*

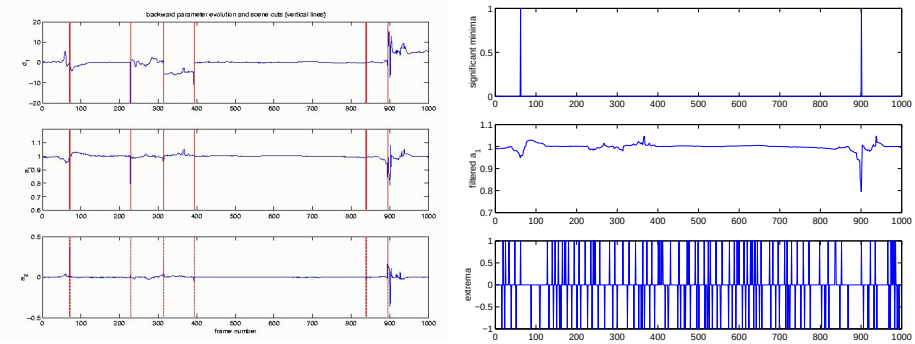

Figure 2: Left plot: temporal evolution of translation and affine motion parameters, resp. d_1, a_1 and a_2 (Vertical lines correspond to actual scene cuts). Right plot: from top to bottom: retained minima of a_1 parameter corresponding to ball deliveries, a_1 parameter filtered by a median filter, detected extrema of filtered a_1

of a high information content in this case. It is interesting that the temporal evolution of the motion parameters is more important than accurate estimation. Further work will involve the use of all the coefficients of $\mathbf{A}$, as well as the consideration of a perspective motion model. It is expected that the use of the audio track will resolve the remaining ambiguities in the video analysis.

References

[1] S. Dagtas, W. Al-Khatid, A. Ghafoor, and R.L. Kashyap. Models for motion-based video indexing and retrieval. *IEEE Transactions on Image Processing*, 9, 2000.

[2] R. Fablet, P. Bouthémy, and P. Pérez. Non parametric statistical analysis of scene activity for motion-based video indexing and retrieval. Technical report, IRISA, 2000.

[3] Y-P. Tan, D.D. Saur, S.R. Kulkarni, and P.J. Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10, 2000.

[4] D. Saur, Y.P. Tan, S.R. Kularni, and P.J. Ramadge. Automated analysis and annotation of basketbakk video. In *SPIE storage and retrieval for image and video databases*, 1997.

[5] T. Kawashima, K Tateyama, T Iijima, and Y. Aoki. Indexing of baseball telecast for content-based video retrieval. In *IEEE ICIP*, pages 871–875, 1998.

[6] J.M Odobez and P. Bouthémy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6, 1995.

[7] F. Dufaux and J. Konrad. Efficient, robust and fast global motion estimation for video coding. *IEEE Transactions on Image Processing*, 9, 2000.

[8] J. Biemond, L. Looijenga, D. E. Boekee, and R.H.J.M. Plompen. A pel–recursive Wiener based displacement estimation algorithm. *Signal Processing*, 13, 1987.