ONLINE PARSING OF SPORTS COACHING VIDEO THROUGH INTRINSIC MOTION ANALYSIS

Dan Ring, Anil Kokaram

Dept. of Electronic and Electrical Engineering,
Trinity College Dublin, Ireland;
dan@unworkable.org, anil.kokaram@tcd.ie

ABSTRACT

Automatic record and review of actions in sports training sessions is of great benefit to both coach and athlete. Many coaching sessions involve repetition of particular actions to hone technique, such as a swing from a tennis racket, golf club, or cricket bat. These actions can be defined by unique motion signatures. A method is proposed to parse the training video using motion into browse-able actions. The method aims to avoid the intensive explicit computation of player silhouette and motion vector fields, allowing for a real-time, online application on standard hardware.

Index Terms— Intrinsic, Motion, Estimation, Content-based, Analysis, Sports, Coaching

1. INTRODUCTION

One-on-one coaching sessions are vital to athletic training, where technique is improved and perfected through repetition. Many actions involved in the training sessions have characteristic motion patterns associated with them. Examples include golf swings, tennis serves and cricket bats. It is vital to make the best use of the coaching session. This paper presents novel methods to analyse implicit motion features, and use it to parse live coaching video, allowing both player and coach to record and review actions instantly, as shown in Figure 1.

Motion is a good feature in parsing video, and has been used to great effect in recent research. The Caviar project [1, 2] has demonstrated the detection of unwanted activity in surveillance camera footage. Local motion fields are analysed for characteristic patterns indicating questionable activity, such as fighting and running. This work allows important footage to be brought to the attention of security guards, allowing more productive use of surveillance feeds.

In [3], the 3D motion of golf swings are extracted from single camera shots. A video segmentation step first performs background / foreground segmentation to isolate the human body. 3D motion information is found by an iterative fitting process. The end result is 3D golf swing data that can be compared numerically against other players. Research in player motion analysis demonstrated in [4] also relies on an initial silhouette segmentation stage.

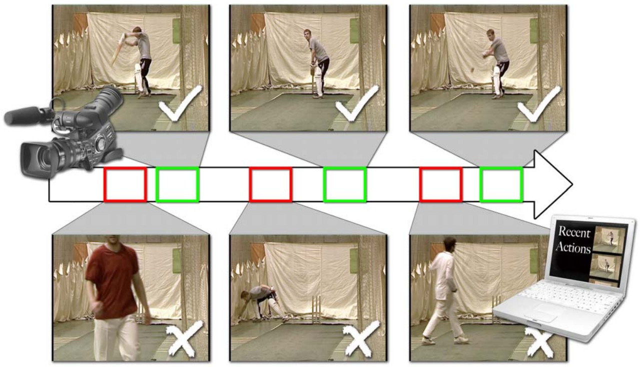

Fig. 1. The objective of the application is to create a rapid record and review system. Video is parsed in real-time during the session for interesting actions. Examples of interesting actions are shown on top, with unwanted actions below.

The application presented in this paper requires rapid record and review on "off the shelf" hardware. The Caviar project [1, 2] currently relies on the calculation of a large motion vector field. The work in [3] relies on an accurate segmentation step. These computationally expensive methods rule out use in a real-time system. The presented work avoids explicit calculation of motion fields and segmentations, allowing real-time video parsing.

Section 2 and Section 3 introduce the algorithms for intrinsic motion analysis and interesting action classification respectively. Section 4 discusses the results of the presented work, and avenues for future work are presented in Section 5.

2. ANALYSING INTRINSIC MOTION

The repetitive actions performed in a coaching session typically exhibit a characteristic motion pattern. This is often seen to be a "ready-then-action" pattern, an example for cricket is

Thanks to Irish Research Council for Science, Engineering and Technology for funding this research, Grant No. RS/2005/115.

1-4244-1437-7/07/$20.00 ©2007 IEEE

shown in Figure 2. In the case of a tennis serve, the set-up or "ready" state is the throw of the ball into the air, and the action burst is the over-arm swing that follows. The objective is to detect "ready-then-action" pairs to signal the application to begin and end recording. The computational burden of full frame motion estimation can be substantial [5]. Efficient implementations of motion estimators have long occupied the image processing community. Of particular interest to this paper is the use of integral projections [6], enabling real-time motion estimation and compensation applications [7, 8].

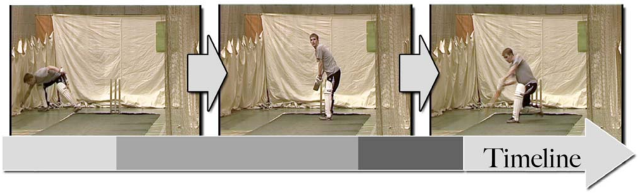

Fig. 2. A typical "Ready-then-Action" motion pattern of a cricket training session. The motion pattern begins with the athlete inactive or setting-up for the action (left image, light-grey on time-line). This is characterised by erratic motion, usually outside the region where player motion is found. Following this, the athlete will get ready, anticipating what is to follow, and is usually stationary or moving fairly little (centre image, middle-grey on time-line). When the action then occurs, there is a sudden burst of motion (right, darker-grey on time-line), and very quickly the athlete goes back to the "ready" state.

In the scenario addressed in this paper, the image context is reasonably constrained. Hence there is usually just a single foreground object against a mostly static background. What is interesting, is to consider alternatives to explicit segmentation that may yield features which are sufficiently correlated with motion to allow usable parsing. The inter-frame difference is clearly a good feature, but it is too crude by itself for reliable parsing. However, by acknowledging the constraint above, it is possible to avoid the explicit segmentation schemes in [3, 4] and [1, 2]. In this paper we introduce therefore the notion of an "intrinsic motion" feature.

The integral projections of a frame are as follows.

$$\rho_{h,n} = \sum_v I_n(h,v) \text{ and } \rho_{v,n} = \sum_h I_n(h,v) \qquad (1)$$

where $I_n(h,v)$ is the intensity of the n-th image at location (h,v). The difference between the projections $\Delta\rho_n$ is correlated with vertical and horizontal motion. This is shown in Figure 3, and $\Delta\rho_n$ is defined as follows.

$$\Delta\rho_{h,n} = |\rho_{h,n} - \rho_{h,n-1}| \text{ and } \Delta\rho_{v,n} = |\rho_{v,n} - \rho_{v,n-1}| \quad (2)$$

Observe Figure 3. What is interesting is that the "centre" of the lobes of $\Delta\rho_n$ roughly tracks the horizontal and vertical motion in the frames. Consider then that the value of $\Delta\rho_n$ at a row is proportional to the probability of motion along that row. Thus the expected value $E(\Delta\rho_n)$ or $\langle\Delta\rho_n\rangle$ is the centre mass of the lobe which is related to the location of the object.

This however is only the case when the motion of the central athlete is the dominant motion in the scene. To lower the impact of background motion on the later analysis, it is desirable to "window" or weight the area of foreground motion. Taking the temporal average of the differential projections $\overline{\Delta\rho_n}$ provides a distribution of the most likely regions of motion in the projected direction, shown as follows.

$$\overline{\Delta\rho_{h,n}} = \frac{1}{n}\sum_n \Delta\rho_{h,n} \qquad (3)$$

The expected value of $\overline{\Delta\rho_n}$ corresponds to the centre of the region exhibiting the most motion over time. In our constrained scene, this point should correspond to the centre of the region of the athlete. Weighting of the projected foreground athlete region is performed by applying 1D Gaussian distributions $g_{h,n}$ and $g_{v,n}$ generated from data of $\overline{\Delta\rho_n}$. From above, the mean μ is therefore the expected value of $\overline{\Delta\rho_n}$. μ and variance σ are calculated as shown.

$$\mu_h = \langle\overline{\Delta\rho_{h,n}}(x)\rangle = \frac{\sum_x x\overline{\Delta\rho_{h,n}}(x)}{\sum_x \overline{\Delta\rho_{h,n}}(x)} \qquad (4)$$

$$\sigma_h = \langle\overline{\Delta\rho_{h,n}}(x^2)\rangle - \langle\overline{\Delta\rho_{h,n}}(x)\rangle \qquad (5)$$

The calculation of the mean μ is the calculation of the expectation value from a non-normalised distribution, in this case, $\overline{\Delta\rho_{h,n}}$. The Gaussian weights $g_{h,n}$ and $g_{v,n}$ are applied to the original differential projections, shown as follows, $\Delta\rho_{w,h,n} = \Delta\rho_{h,n}g_{h,n}$, and illustrated in Figure 3.

We now want to use the refined player motion feature ($\Delta\rho_{w,n}$) to identify interesting action patterns. From the examples in Figures 3 and 4, there is an increase in the amount of player motion from the "ready-to-action" stages. Observe $\Delta\rho_{w,n}$. The correlation between projection and motion has been localised to the player region. If there are no lobes in $\Delta\rho_{w,n}$, there is probably little player motion, and similarly, higher peaks indicate higher amounts of motion. The motion amount $a = (a_h, a_v)$, where a_h and a_v indicate horizontal and vertical motion amounts respectively, is calculated as follows; $a_h = \sum_n \Delta\rho_{w,h,n}(x)$. The vertical motion amount a_v is calculated similarly.

3. IDENTIFYING INTERESTING PLAYER ACTIONS

Recall from Section 2 that it is required to identify "ready-then-action" episodes in a real-world set-up. To allow the application adapt to new environments and players, statistical measurements are made upon a running history of the motion sampled for a given duration in time t. The amount of motion

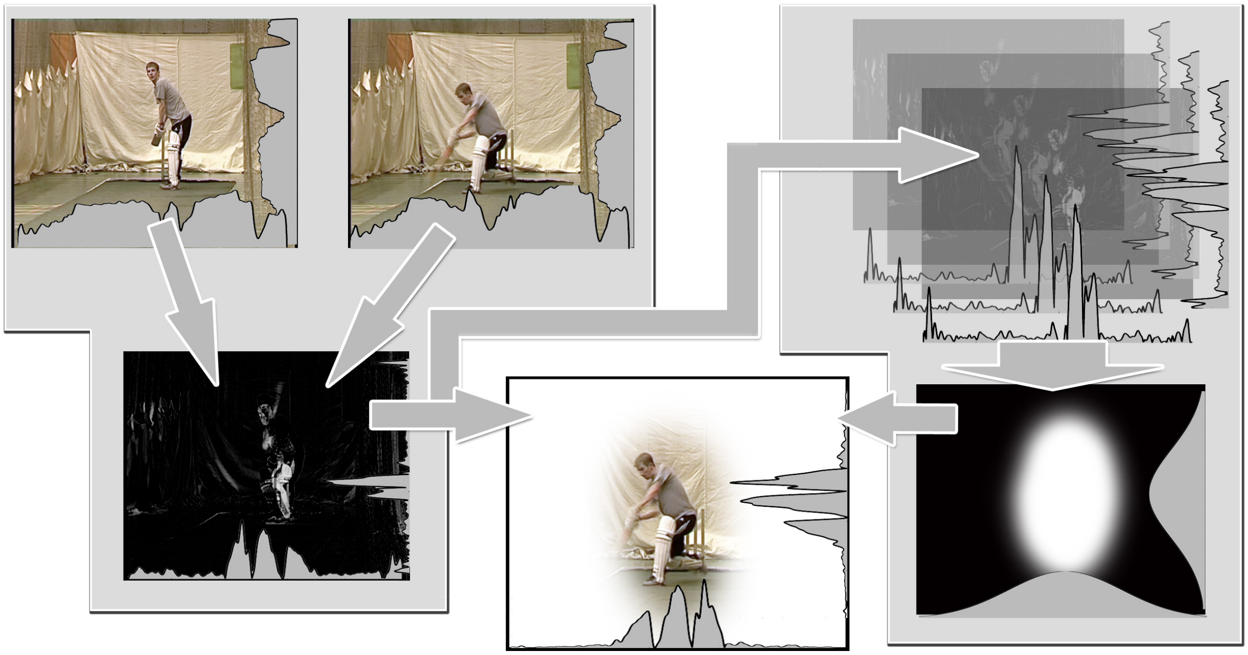

Fig. 3. A flowchart of the algorithm. Projections ρ_n of the incoming images and their differences $\Delta\rho_n$ are shown in the top- and bottom-left images. As can be seen, lobes in $\Delta\rho_n$ are correlated to motion in the image. $\Delta\rho_n$ is added to a running history (top-right). Gaussian weights g_n are derived from $\overline{\Delta\rho_n}$ are shown in bottom-right. Resulting foreground player motion isolated $\overline{\Delta\rho_{w,n}}$ in bottom-middle.

a is modelled as a running process m_a, and it is assumed the distribution of the motion is Gaussian.

$$m_a = [a_{\tau-t}, a_{(\tau-t)+1}, a_{(\tau-t)+2}, ..., a_{\tau-1}, a_\tau] \tag{6}$$

The range of samples is updated at each new frame, as each new sample a_τ arrives, the oldest $a_{\tau-t}$ is removed. To determine when to begin recording, a one-tailed Z test is performed. The most recent sample a_τ is tested for statistical significance, at a given confidence level p, against the sequence of samples m_a, as follows:

$$C(p) < \frac{a_\tau - \mu(m_a)}{\sigma(m_a)} \tag{7}$$

where $\mu(m_a)$ and $\sigma(m_a)$ are the mean and variance of the sample sequence m_a respectively, and $C(p)$ is the upper extent of the confidence interval for given probability p. In this application, a p value of .99 was used.

If the current sample a_τ tests as being significant, it is assumed that an action is taking place. The beginning of the action is found by looking backwards through the action samples m_a to find a local minimum, indicating the end of the "ready" state. The end of the action is determined when the above Z test becomes insignificant, and a minimum amount of time has passed.

4. EXPERIMENTAL RESULTS

A cricket training session was filmed in typical conditions incorporating multiple players at varying camera angles. As a result, it contained many "real-world" problems such as player occlusion, team-mates inadvertently walking across the camera shot, background motion of the net behind the foreground player, camera position and focus adjustment, and high amounts of player motion during non-interesting actions.

A 20 min portion that session was used to estimate the appropriate window size m_a. Shown in Figure 5 are the precision and recall results for this portion with varying window sizes of m_a. From Figure 5, 60 taps is a reasonable estimate of the optimum window size for m_a. With that window size, a different 30 min segment was parsed and Figure 4 shows a trace of motion amount a_h, and the corresponding parsed segments for a series of actions illustrated by the ground truth. The measured precision and recall were both .83, implying a usable system.

5. CONCLUSIONS

This paper has presented a new feature for detecting dominant motion events in sports video. The success of the analysis algorithm lies in the constrained image context outlined in Section 2. However, several areas exist where improvements can be made. Many false detections were made due to high amounts of motion during set-up or "inactive" player states.

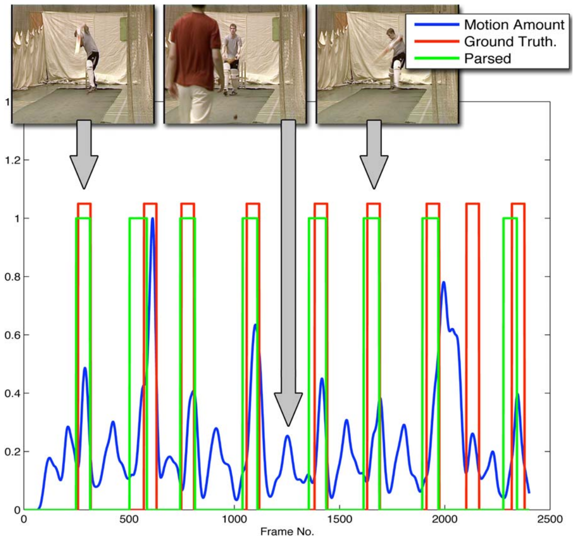

Fig. 4. Results of the parsing of several actions. The "ready-then-action" pattern can be observed in the motion amount, with motion peaks corresponding to "inactive" player states (centre image) and interesting actions (left and right images). Note the distinct minima of the "ready" state preceding interesting actions.

This could be corrected through analysis of the projected area of the background (i.e. the non-foreground player) for cross-over between foreground and background regions.

Unfortunately, the constrained image context is not always enforceable in the real world. To improve the accuracy of action detection, the projections can be used to estimate camera motion [7] and compensate accordingly. In the final application, the shape of the foreground weights can be used to indicate to the user if the current camera set-up will give good results, i.e. elliptical shapes indicate a good foreground segmentation. These additions will be addressed in future work.

6. REFERENCES

[1] José Santos-Victor Pedro Canotilho Ribeiro, "Human activity recognition from video: modeling, feature human activity recognition from video: modeling, feature selection and classification architecture," in *http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm*. BMVP, HAREM, October 2005, pp. 61–70.

[2] Filiberto Pla, Pedro Ribeiro, José Santos-Victor, and Alexandre Bernardino, "Extracting motion features for visual human activity representation.," in *IbPRIA (1)*, 2005, pp. 537–544.

[3] Jenq-Neng Hwang Ibrahim Karliga, "Analyzing human body 3-d motion of golf swing from single-camera video sequences," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006*, May 2006, vol. 5, pp. 493 – 496.

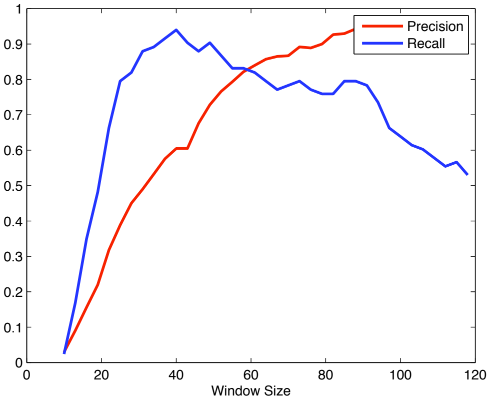

Fig. 5. Results of Precision and Recall against size of window of recorded motion data m_a. Longer window sizes provide more support to the significance test, giving a boost in precision, as seen in window sizes from around 10 to 50. However at longer values, m_a will appear noisy, causing more peaks to be identified as actions, thereby reducing precision.

[4] Shihong Xia, Xianjie Qiu, and Zhaoqi Wang, "A novel framework for athlete training based on interactive motion editing and silhouette analysis," in *VRST '05: Proceedings of the ACM symposium on Virtual reality software and technology*, New York, NY, USA, 2005, pp. 56–58, ACM Press.

[5] A. C. Kokaram, *Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video*, Springer Verlag, ISBN 3-540-76040-7, 1998.

[6] Peyman Milanfar, "A model of the effect of image motion in the radon transform domain," *IEEE Transactions on Image Processing*, vol. 8, pp. 1276–1281, September 1999.

[7] F. Pitie A.J. Crawford H. Denman F. Kelly and A.C. Kokaram, "Gradient based dominant motion estimation with general projections for real-time video stabilisation," in *IEEE International Conference on Image Processing*, October 2004, vol. 5, pp. 3371– 3374.

[8] D. Robinson and P. Milanfar, "Fast local and global projection- based methods for affine motion estimation," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 35–54, January 2003.