Scene Structure Analysis for Sprint Sports

M. Hedayati
School of Engineering
University of Waikato
Hamilton 3240
New Zealand
Email: mh267@student.waikato.ac.nz

Michael. J. Cree
School of Engineering
University of Waikato
Hamilton 3240
New Zealand
Email: M.cree@ieee.org

Jonathan Scott
School of Engineering
University of Waikato
Hamilton 3240
New Zealand
Email: scottj@waikato.ac.nz

***Abstract*—This work proposes a robust model to analyse the structure of horse races based on 2D velocity vector information. This model is capable of detecting scene breaks, classifying the view of the contenders and extracting the trajectory of the contenders throughout the race. The performance of the system is tested over six video clips from two different broadcast sources. The performance analysis shows the model achieves a high accuracy of view classification with the lowest value of 83%, all in real time.**

I. Introduction

Automated systems for the analysis of sports have drawn the attention of many commercial entities and are opening up many opportunities for researchers to study and develop a computer vision approach to sport analysis. Sports video analysis studies can be classified into two main categories: event detection and high-level analysis [1], [2]. For event detection the algorithm is designed to extract the sporting highlights. In contrast, high-level analysis deals with the extraction of complex information such as motion and the skill of players [3].

A comprehensive sports analysis system must perform three main challenging tasks: first, analyse the scene to extract the structural information from the broadcast video [4], second, apply object detection to locate the objects of interest in a frame and, lastly, track the objects and determine their motion. Sports events, unlike news or movies, have a well-defined content structure [5], however each sport has its own unique and specific structure. For this reason, most research in the area of sports video analysis is specific to a certain type of sport [6]. Systems for the automated analysis of a number of different sports have been reported [13], [14], [15], [16], [17], yet there is no significant study that has been done specifically on horse race analysis. This paper, as part of an intelligent horse race video analysis system, focuses on scene structure analysis for circular track sports such as cycling, running, skating and horse racing.

The main objectives of this paper are to classify each video frame based on view angle of the race contenders and then extract the overall trajectory of the contenders. This article is organized as follows: Section 2 is a review of prior significant work on the scene analysis model; Section 3 outlines the structure of the broadcast horse race and identifies the main information that can be extracted from the videos; Section 4 describes our proposed model for frame classification and trajectory analysis; Section 5 evaluates the performance of the system; and, finally, Section 6 concludes and provides recommendations for further work.

978-1-5090-2748-4/16/$31.00

II. Review

In sports broadcasts the cameras and the camera viewpoint change frequently. For example, in a short period a camera might show the broad view of the field, followed by a close-up of the player and then move to crowd reaction [7]. Therefore in any sports analysis system, first the video is broken down into different shots, and then the shots are analysed to extract the structural information for that particular sport. In the computer vision literature, the first task is often called shot detection and the subsequent process is known as event detection or semantic analysis.

1) Shot detection: The shot is a continuous sequence of frames taken by a single camera. The basis of any shot detection method is the fact that frames surrounding a shot boundary display a significant change in visual content [8]. The shot detection algorithms do not depend on the sports structure, and any reliable shot detection algorithm can directly apply to any sports video. The boundary detection algorithm can be categorized into four main approaches: pixel comparison, histogram difference, edge ratio, and motion vector analysis.

The pixel-based method is the simplest approach to detect shot boundaries which are identified by change in the number of pixel beyond a certain threshold [9]. In the histogram based approach the grey or colour histograms of two consecutive frames are compared and if they differ by more than a certain amount, then a change of shot is identified [10]. In the edge based difference method the edges for two consecutive frames are extracted and the similarity between the extracted edges is used to determine if the shot has changed [11]. The motion vector approaches are based on the assumption that motion exhibits discontinuities across shot boundaries, thus the frame to frame motion can be used to detect the changing shot [12], [13].

2) Event detection: While a sports event may be a few hours in duration, it may nevertheless contain only a few moments that are of interest for the viewer or the professional

sports analyst. The main task of the event detection process is to identify and extract these moments automatically.

Lien et al. [14] proposed a model to classify baseball video shots into running, base, close-up and other views. They used a three stage classification: first the shot was categorized according to the proportion of soil and grass regions. Then global camera motion was used to detect the running scenes and finally skin colour detection was applied to classify whether the shot was a close-up, base, or other scene. Ye et al. [15] proposed an event detection model for soccer analysis by building feature vectors for each shot based on field view, camera motion and shot length. These vectors show the characteristic of the events and combined with a support vector machine (SVM) classifier they build a model to classify events in a soccer match.

Huang et al. [16] proposed a model that used both audio and video information to classify a tennis game into silent, close-up, wide and replay events. Silent was defined as a segment of unnoticeable audio such as short clicks. For close-up and wide shots classification, the picture outline was measured using Canny edge detection. For replay shot the logo transition was found where such a transition was often played prior to and after a replay. Montouliu et al. [17] proposed a model to perform team activity recognition and analysis in Association Football. They manually divided the football pitch into ten cells representing the most active area of playing field. Then the velocity vector for each cell was calculated over four consecutive frames resulting in feature vector representing the motion pattern in the video clip.

III. Sprint Sport Structure

In this section we focus on the analysis of horse races, and explain the kind of data that is to be extracted from the horse races, characteristic of the camera and the structure of the videos that are broadcast.

A. Key Information

After consultation with professional analysers, it was found that one of key factors in horse race is the approximate distance of jockeys from the fence, as shown in Figure 1. Another important issue is the overall flow of contenders which can be used to detect abnormal movement of the individual contenders. In order to extract this information, the system needs to detect and track the jockeys in specific frame sequences. However, as this work only focuses on the structure analysis of the race, the main aims here are to find the overall trajectory of contenders and detect the appropriate frames with respect to the contenders view angle from the camera. In general the view of the contenders with the respect to the camera angle can be classified into three categories: three quarter (3Q), side, and front views as shown in Figure 2.

B. Camera Characteristic

In sports which contenders race around a circular track, the camera view typically follows the motion of the contenders to keep them in the centre of the view, therefore the motion of the background is much more than that of the contenders in the camera view.

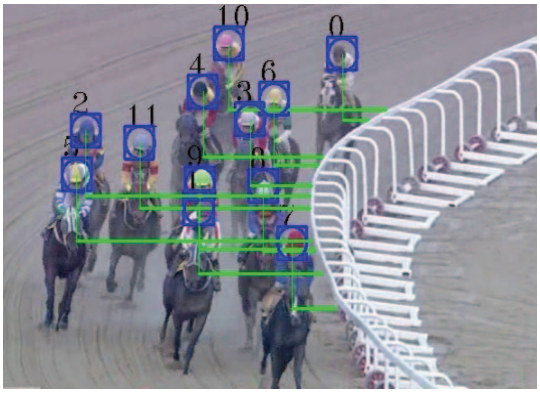

Fig. 1. The distance of jockeys respect to the fence is one of the key information for the professional analysers.

Under normal racing conditions the contenders move relatively slowly with respect to each other, and the camera is positioned sufficiently far away from the race track, so that while the camera is fixed in position and is rotated to follow the contenders along the race track, the motion of the background in the camera view can be used as a reasonable approximation of the contenders motion with respect to the real world. Therefore the contenders trajectory in the real world is the opposite of the background trajectory in the camera view. Hence accurate estimation of the background velocity is the key task in this work.

IV. Proposed Model

The diagram in Figure 3 is a flow chart of a proposed system for the shot analysis. The content of this section is divided into two parts, the first part explains the basic information on which the system is built and the second part shows how this information is applied to recognize the contenders view and the trajectory of the contenders through the video.

A. Forming The Structure

As the contenders usually scatter around the centre of the frame, at the first step the incoming frames (I) are divided into four quadrants (B), $I = [B_1 + B_2 + B_3 + B_4]$. This process reduces the effect of outlier motion in the background velocity estimation. Next, the local motion vectors for each quadrant are obtained by using pyramidal implementation of Lucas-Kanade (LK) optical flow technique [18]. LK procedure starts by extracting a set of unique points at frame (n). In this paper these points are called reference points and they are extracted based on the corner detection algorithm proposed by Rosten [19]. The reference points then are matched to the corresponding location at frame $(n+1)$. Therefore the output is a set of pairs of points (Q), given by

$$Q = \{(p_i^n, p_j^{n+1}) | i \in \{1, \ldots, L_R\}, j \in \{1, \ldots, L_M\}\}, \quad (1)$$

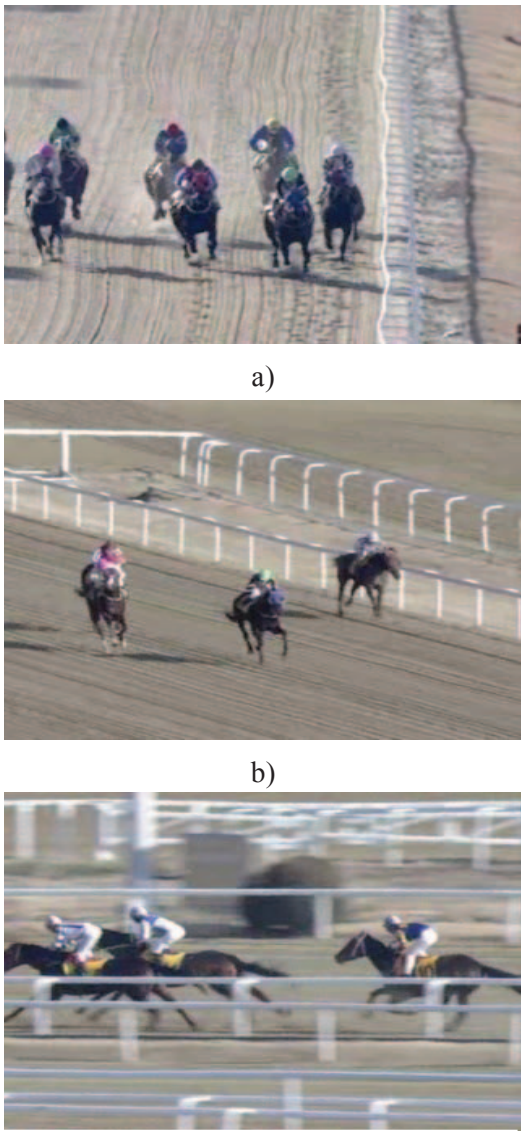

Fig. 2. Three contenders view a) front (b) 3Q c) Side.

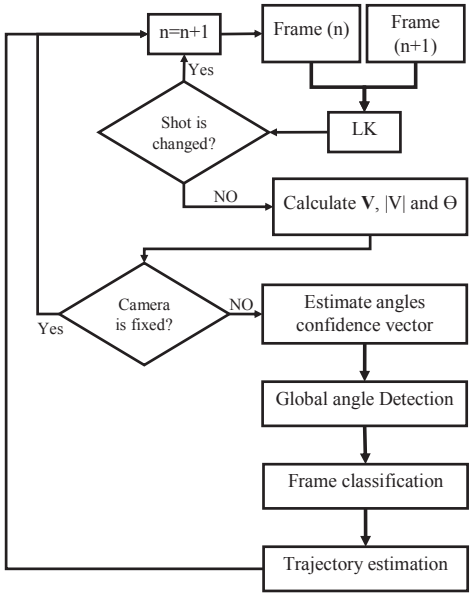

Fig. 3. The flow diagram of proposed model for shot analysis.

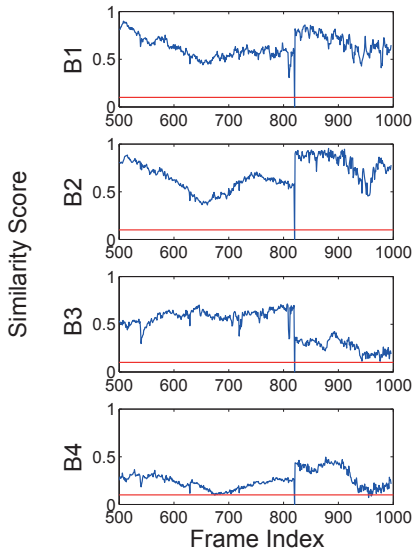

Fig. 4. The similarity score versus video frame from quadrant 1 (top left) to quadrant 4 (bottom right). The horizontal red line shows the threshold value.

where p_i^n is the reference point in the frame n and p_j^{n+1} is the corresponding match at frame $n+1$, L_R is the total number of reference points and L_M is the total number of points matched by LK to the next frame. From the set Q, the 2D motion vector, $\mathbf{V}$, velocity magnitude, $|V|$, and angle of velocity, θ, are determined for each pair, per quadrant.

B. Building a Framework

1) Shot Detection: The literature study showed that it is essential to break the video into the shots prior to background motion estimation. This model follows the shot detection methodology proposed by Majumdar et al. [20]. They analysed the number of matched points from one frame to the next, and if that number fell below a certain threshold T, it was determined that the current shot was finished and new shot had started. For this analysis they used the similarity Score which is the ratio of total match point to the total number of reference point by,

$$Similarity\ Score = L_M/L_R. \tag{2}$$

We calculate the similarity score for the four quadrants and if the score falls below the threshold in a majority of quadrants then the shot is considered to have ended the threshold. In this experiment we empirically set the similarity score threshold to 0.1. The sample shot detection result can be seen in Figure 4.

2) Filtering: The velocity vectors $\mathbf{V}$ that have magnitude $|V|$ less than 1 pixel, referred to as static points, are removed as they are not reliable. However, it is possible that at some points in the race the camera may stop moving, and in such cases, the filtering would remove all background motion leading to an error in considering contenders motion as dominant motion.

To avoid this miscalculation and to give the system the ability to distinguish between the fixed and the moving camera, a motion score λ is calculated by taking the ratio of the total number of the static points L_F to the total number of match pairs L_M over all four quadrants, namely,

$$\lambda = \frac{1}{4}\sum_{i=1}^{i=4}\frac{L_{F,i}}{L_{M,i}}. \tag{3}$$

If λ is bigger than a predefined threshold (α) the camera is considered to have stop moving and its 2D velocity vector is set to $(0, 0)$. In this experiment we empirically choose threshold $\alpha = 0.3$.

3) Angle confidence estimation: From the remaining 2D motion vectors the frequency distribution (histogram) of velocity angles is obtained for each quadrant. The histogram bins are equally spaced at 5° interval from -180° to 180° and it is normalized locally in each quadrant so that the largest bin has value of 1.

Let $f_{b,i}$ be the normalized histogram bin and $f_{max,i}$ be the peak value of the histogram for the ith quadrant, then for each bin b the angle confidence is calculated as,

$$s_{b,i} = \begin{cases} \frac{f_{b,i}}{f_{max,i}} & \frac{f_{b,i}}{f_{max,i}} \geq .5 \\ 0 & \frac{f_{b,i}}{f_{max,i}} < .5 \end{cases}, \tag{4}$$

Therefore for each quadrant there is one vector called the angle confidence vector S where $S = \{s_b\}_{(b=1,2,\ldots,72)}$. Note that to remove insignificant values any angle that falls below 50% of the peak value is set to zero.

4) Global Angle Detection: The motion vectors might belong to the contenders or to the background and if there are too many due to contenders in some quadrant, then the confidence vector may not represent the background motion properly. It is therefore required to have some kind of mechanism to reduce the effect of these quadrants in the final angle decision. To achieve this the quadrants are weighted based on a weight factor W, given by the reciprocal of the total number of non-zero elements in the confidence vector L_S as,

$$W(i) = \frac{1}{L_{S,i}}, \tag{5}$$

L_S is the total number of non-zero elements in confidence vector S. From Equations (4) and (5) the global confidence vector (S') is formed as

$$S' = \sum_{i=1}^{4} W(i)S_i \tag{6}$$

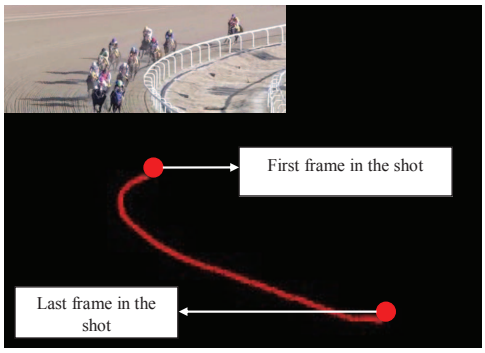

Fig. 5. Sample camera trajectory. Top left image shows the first frame of the respective shot and the red curve is the contenders' trajectory for that shot.

5) Object View Detection: Finally from the location of the maximum element in vector S' the optimal angle θ_{opt} is extracted and the frame is classified with respect to θ_{opt} by,

$$C = \begin{cases} \text{Front} & |\theta_{opt}| > 60 \quad \wedge \quad |\theta_{opt}| < 120 \\ \text{Side} & |\theta_{opt}| < 10 \quad \vee \quad |\theta_{opt}| > 170 \\ \text{3Q} & \text{otherwise} \end{cases}, \tag{7}$$

6) Contender's Trajectory Estimation: The optimal background velocity vector V_{opt} is estimated by taking the mean of all 2D motion vectors that falls inside b_{opt} where the b_{opt} is the bin corresponding to the optimal angle, and then the contenders trajectory can be obtained by accumulating $-V_{opt}$ over all frame in the shot (Figure 5).

V. Data Evaluation and Validation

The proposed model is implemented with OpenCV on Intel core i7- 4770 @ 3.4 GHz CPU with 16 GB RAM. This model can easily be used in real time with the processing time of 0.03 second per frame ($\approx$ 33 frames per second).

To estimate classification accuracy of the camera view six video clips are selected as ground truth data , two clips for each view. There are between 100 and 300 frames in each video clip. Figure 6 shows the sample frame for each clip and their respective classes. Using the ground truth data the classification performance is evaluated based on sensitivity,

$$Sensitivity = \frac{TP}{TP + FN}. \tag{8}$$

As can be seen in Table I the real time classification result is highly accurate with lowest sensitivity of 83%. It is worth noting that as the contenders' trajectory is estimated based on the angle of the camera view, therefore, the trajectory accuracy is correlated with the classification accuracy and the result in Table I can be considered as an approximation to the trajectory accuracy.

If the real time processing is not the issue this performance can be further improved by simple smoothing over the final output. Table II shows the accuracy after applying mode filtering with window size 31. The effect of mode filtering can be seen clearly from Figure 7 where the outputs before and after filtering can be compared.

Fig. 6. Class 0, 1, 2 which represent Front, Side, and 3Q, respectively.

TABLE I
REAL-TIME PERFORMANCE OF PROPOSED MODEL

	Class	Total frame	TP	FN	Sensitivity
RACE 1	**Front**	300	299	1	.99
	Side	104	87	17	.83
	3Q	274	256	18	.93
RACE 2	**Fornt**	230	219	11	.95
	Side	100	100	0	1
	3Q	114	114	0	1

TABLE II
THE PERFORMANCE AFTER APPLYING MODE FILTER

	Class	Total frame	TP	FN	Sensitivity
RACE 1	**Front**	300	300	0	1
	Side	104	96	8	.92
	3Q	274	261	13	.95
RACE 2	**Fornt**	230	227	3	.99
	Side	100	100	0	1
	3Q	114	114	0	1

VI. CONCLUSION

This paper proposed a statistical framework based on a 2D velocity vector. This model is able to classify video frames based on contender view and find the trajectory of contenders. To our knowledge this model is the first to approach the analysis of a horse race from broadcast videos. The results prove the reliability of the system with the lowest accuracy at 83% in real time. The validation result shows that this accuracy can be further improved in off-line processing

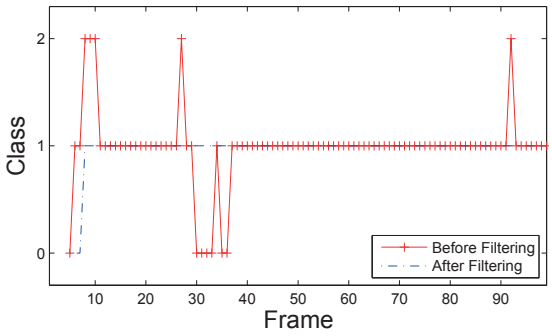

Fig. 7. Sample results for side view clip from the race one before and after mode filtering.

REFERENCES

[1] G. Thomas, *Sports TV applications of computer vision.* Visual Analysis of Humans. Springer London, 2011, pp. 563–579.

[2] T. DOrazio and M. Leo, *A review of vision-based systems for soccer video analysis.* Pattern recognition, 2010, vol. 43, no. 8, pp. 2911–2926.

[3] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang and H. Wang, *Automatic player detection, labeling and tracking in broadcast soccer video.* Pattern Recognition Letters, 2009, vol. 30, no. 2, pp. 103–113.

[4] Y. Hu, B. Han, G. Wang And X. Lin, *Enhanced shot change detection using motion features for soccer video analysis.* 2007 IEEE International Conference on Multimedia and Expo, IEEE, pp. 1555–1558.

[5] D. Zhong and S. F. Chang, *Real-time view recognition and event detection for sports video.* Journal of Visual Communication and Image Representation, 2004, vol. 15, no. 3, pp. 330–347.

[6] B. Li, J. H. Errico, H. Pan and I. Sezan, *Bridging the semantic gap in sports video retrieval and summarization.* Journal of Visual Communication and Image Representation, 2004, vol. 15, no. 3, pp. 393–424.

[7] R. Kapela, A. wietlicka, A. Rybarczyk and K. Kolanowsk, *Real-time event classification in field sport videos.* Signal Processing: Image Communication, 2015, vol. 35, pp. 35–45.

[8] A. Hanjalic, *Shot-boundary detection: unraveled and resolved?* IEEE transactions on circuits and systems for video technology , 2002, vol. 12, no. 2, pp.90–105.

[9] J. S. Boreczky and L. A. Rowe, *Comparison of video shot boundary detection techniques.* Journal of Electronic Imaging, 1996, vol. 5, no. 2, pp. 122–128.

[10] H. Ueda, T. Miyatake, and S. Yoshizawa,, *IMPACT: an interactive natural-motion-picture dedicated multimedia authoring system.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI. 91, ACM, 1991, no. 8, pp. 343–350.

[11] R. Zabih, J. Miller and K. Mai, *Feature-based algorithms for detecting and classifying scene breaks.* Cornell University, 1995.

[12] E. Bruno, D. Pellerin, *Video Shot Detection Based on Temporal Linear Prediction of Motion.* Proc. of IEEE International Conference on Multimedia and Exposition, Lausanne, Switzerland,2002, pp. 289–292.

[13] Hu. Yichuan, B. Han, G. Wang, and X. Lin, *Enhanced shot change detection using motion features for soccer video analysis.* In 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 1555-1558.

[14] C. C. Lien, C. L. Chiang and C. H Lee, *Scene-based event detection for baseball videos.* Journal of Visual Communication and Image Representation, 2007, vol. 18, no. 1, pp. 1–14.

[15] Q. Ye, Q. Huang, W. Gao and S. Jiang, *Exciting event detection in broadcast soccer video with mid-level description and incremental learning.* In Proceedings of the 13th annual ACM international conference on Multimedian, ACM, 2005, pp. 455–458.

[16] Y. P. Huang, C. L. Chiou, and F.E Sandnes, *An intelligent strategy for the automatic detection of highlights in tennis video recordings.* Expert Systems with Applications, 2009, vol. 36, no. 6, pp. 9907–9918.

[17] R. Montoliu, R. Martn-Flez, J. Torres-Sospedra and A. Martnez-Us. *Team activity recognition in Association Football using a Bag-of-Words-based method.* Human movement science, 2015, vol. 41, pp. 165-178.

[18] J. Y. Bouguet, *Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm.* Intel Corporation, 2001, vol. 4, no. 1, pp. 4–12.

[19] E. Rosten and T. Drummond, *Machine learning for high-speed corner detection.* European conference on computer vision, Springer, 2006, pp. 430–443.

[20] J. Majumdar, G. M. Venkatesh and A. Kamath, *Video Shot Detection using Corner Detectors and Optical Flow.* International Conference on Emerging Research in Computing, Information, Communication and Applications, 2013.