

SPORTS INDEXING THROUGH CAMERA AND CONTENT UNDERSTANDING

P. Chippendale¹, A. Pnevmatikakis²

¹Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy, chippendale@fbk.eu

² Athens Information Technology, 19,5 km Markopoulou Ave., GR-19002, Peania, Attiki, Greece, apne@ait.edu.gr

Abstract

In this paper we present an automated system for the creation of detailed indexes, tailored towards video-streams covering large scale athletic events. The level of detail that the indexing system can generate covers: camera location estimation, camera usage estimation (where it is looking and what event it is covering), who is visible in the frames and an estimation of event phase. These indexes are gleaned from the fusion of the outputs of advanced image processing modules that can estimate scene content in real-time, with the Sports Information System (SIS) metadata. The image processing modules form part of the sensing infrastructure developed inside the EU-funded My-e-Director 2012 FP7 project.

Keywords: camera motion estimation, video indexing, video processing, sports

1 Introduction

The importance of annotations in sporting events is well understood in modern sports broadcasting: whether viewers wish to receive extra information about what they are seeing or broadcasters wish to create content-rich indexes for future archive access. For major sporting events, a manual annotation system exists (namely the Sports Information System (SIS) created by ATOS Origin) that generates a temporal record of events. However, the SIS stream only superficially covers an event, hence there is a need for an all encompassing automated system, operating at the camera level, that can pick up on the wealth of interesting incidents that occur during such events. Such a system would need to understand the visual content ‘embedded’ inside a video streams and from this generate a rich stream of appropriate meta-data.

Although there are many problems associated with the extraction of useful content from video images, nevertheless video has the potential of providing us with a rich source of information, ranging from low-level features such as environmental context, e.g. track, field, crowds, etc., to more complex content features like person tracking (i.e. which athletes are visible and where they are in the frame and in the stadium), and camera deployment modelling. In this paper, a system for automatically reasoning about higher-level activities is outlined, which is capable of generating annotations like: Athlete X is preparing to make his 3rd triple jump attempt and camera y is viewing it from close to the take-off board.

2 Camera Position through Athlete Presence Estimation

Cameras covering athletic events usually place the competing athlete(s) in the central portion of the image, framing them within a range of acceptable sizes. To achieve this, he smoothly pans, tilts and zooms the camera to follow the action. This creates the effect of a quasi-static background scene being gradually moved behind the athletes (providing they are moving) who then appear quasi-spatially static in the frame. To segment a dynamic scene of the type, we apply two algorithms based on optical flow and corner detection with cross-frame correlation to identify regional motion tendencies and thus extract foreground motion blobs (generated by moving athletes) away from the static yet-translating background (exploited to estimate camera motion). This is shown in Figure 1.



Figure 1: Matching of key-points in successive frames (top-right crosses), cluttered frame difference (bottom-left) and projectively matched frame difference (bottom-right) showing potential athletes. Detected bodies are marked by blue rectangles (top-right)

Foreground regions are fused with face [1] and text detections [2] to yield athlete regions (see Figure 2).

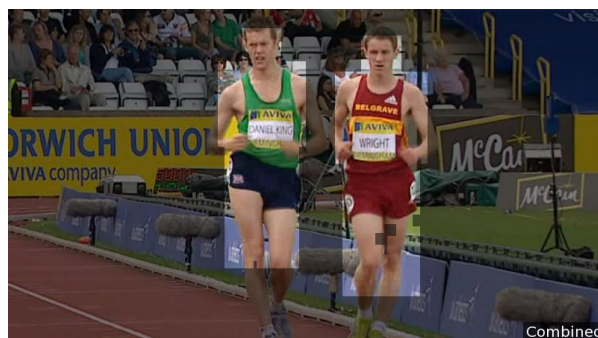


Figure 2: Athlete regions of interest produced by module fusion

The fused regions are in turn exploited to guide the athlete recognition algorithm, based on text reading (see Figure 3), face identification and body tracking through histogram and template matching.



Figure 3: Text extraction results (highlighted in red)

Based on foreground motion and identity estimates the system understands who is in the scene and how large they are (in terms of image size). An ‘athlete presence’ list is generated by the ID modules, which is then cross-referenced against the SIS stream to estimate the rough location of the camera that is viewing the particular scene. For example, if one or more 5000m athletes are detected in a frame and the SIS tell us that these people crossed the start/finish line around 25 seconds ago and the average 3rd lap time for this event is about 60 seconds, then the system estimates that the camera is watching a portion of track almost half-way round from the start line. We can infer the general location of the camera by estimating how frontal the athletes’ aspects are (through face and text identification scores) and by analysing recent camera motion (e.g. panning left, zooming out, etc.).

3 Phase of Event Estimation

Once the system knows who is present in an image and some other general estimations about camera activity and location, the output of the video analysis reasoning block can be used to enrich the SIS stream through the inference of events phases. For example, a long jump attempt has four phases: 1) Preparation phase (cues: an athlete is clearly visible and is usually full screen, plus there is virtually no camera motion); 2) Running phase (athlete identification cues are sparse and motion blur is detected, plus camera motion corresponds to the estimated camera position, e.g. side cameras will be panning and face-on cameras will be zooming-out); 3) Landing phase (camera stops moving and the athlete becomes the primary object in the image once more); 4) Idle phase, between attempts (camera pans back to the initial state waiting for the next athlete, hence no athletes are visible). The reasoning for these phases is demonstrated in Figure 4.

4 Conclusions

We have outlined a system that can accept standard, manual sporting annotations and enrich them, in terms of the location and identity of athletes, phases of the events and incidents of interest, by processing the feeds of the broadcasters’ cameras using video processing tools and an a-prior knowledge of natural event progression.

Such automatic annotations can be exploited to create new means for consuming broadcast sports, offering individuals or groups of users the tools to generate new ‘automatically directed’ video stream based on personal preferences.

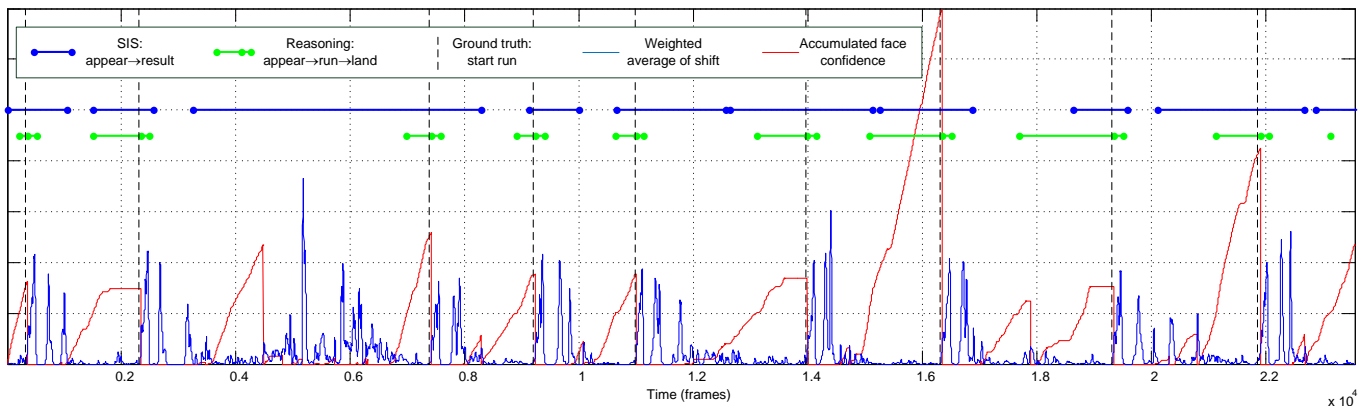


Figure 4: Automatic decomposition of long jump event: Vision processing metrics (red waveform - accumulated face confidence ; blue waveform weighted average of camera shift). Incident reasoning: (green bars - person presence). SIS incidents (blue bars – appearance of athlete to distance of jump measured). Ground truth: (actual moment of beginning of a run marked by the vertical dashed black line)

¹ A. Pnevmatikakis and L. Polymenakos, ‘Subclass Linear Discriminant Analysis for Video-Based Face Recognition’, *Journal of Visual Communication and Image Representation*, Volume 20 , Issue 8, pp. 543-551, 2009
² S. Messelodi and C. M. Modena, ‘Automatic Identification and Skew Estimation of Text Lines in Real Scene Images.’ *Pattern Recognition*, 32:791–810, 1992