Assistive Sports Video Annotation: Modelling and Detecting Complex Events in Sports Video

Aled Owen[1], David Marshall[1], Kirill Sidorov[1], Yulia Hicks[1], and Rhodri Brown[2]

[1] Cardiff University, Cardiff, UK
[2] Welsh Rugby Union

Abstract

Video analysis in professional sports is a relatively new assistive tool for coaching. Currently, manual annotation and analysis of video footage is the modus operandi. This is a laborious and time consuming process, which does not afford a cost effective or scalable solution as the demand and uses of video analysis grows. This paper describes a method for automatic annotation and segmentation of video footage of rugby games (one of the sports that pioneered the use of computer vision techniques for game analysis and coaching) into specific events (*e.g.* a scrum), with the aim to reduce time and cost associated with manual annotation of multiple videos. This is achieved in a data-driven fashion, whereby the models that are used for automatic annotation are trained from video footage. Training data consists of annotated events in a game and corresponding video. We propose a supervised machine learning solution. We use human annotations from a large corpus of international matches to extract video of such events. Dense SIFT (Scale Invariant Feature Transform) features are then extracted for each frame from which a bag-of-words vocabulary is determined. A classifier is then built from labelled data and the features extracted for each corresponding video frame. We present promising results on broadcast video for a international rugby matches annotated by expert video analysts.

1 Introduction

Video analysis in sport is a growing and important field within the professional sporting environment, providing varying levels of assistance and insight depending on the sport or even the team. However, sports video analysis is currently a laborious manual process. This drastically limits the volume and quality of annotation, especially when considering the number of games that are played at a professional level. Other means of instrumentation are available at a professional level, such as relative GPS tracking. However, these are invasive (affixing physical devices to the players). These other sources often provide higher fidelity data but their invasiveness limits data capture in team based sports to only a home teams' players, thus limiting the information that you can learn from the game. This is especially true in sports such as rugby where accurate measurement of relative positions of players and constellations of players is crucial to the game's analysis.

Computer Vision in sports analysis, compared to many other fields, is in its relative infancy. Computer vision techniques have, however, been incorporated in an assistive role into broadcast packages, such as the BBC's usage of the pitch tracking to overlay pitch information [7].

Other efforts in sports analysis have focused on the tracking and identification of individual players to varying levels of success and largely fall into two categories according to the model of the camera motion. The Scale Invariant Feature Transform (SIFT) [4] has shown promise in distinguishing players form both fixed cameras [3] and from a pan-tilt-zoom camera [5]. However, in both cases these methods are applied to sports that lack the complex occlusions and structures which are present in a rugby game.

The challenge addressed in this paper is the automatic classification of video events from rugby footage. We have annotated data from past rugby matches where key events such as *scrum*, *lineout*, *ruck*, *maul*, *etc.* have been labelled. The main contribution of this paper is method to take previously unseen footage and automatically annotate each frame as one of these classes.

2 Method

In order to achieve the above classification of rugby events we have devised a pipeline that comprises the following stages:

Background Removal: The input is video is broadcast footage which often contains imagery of the crowd, advertising hoardings and other periphery. Removal of such imagery is required as features we extract and use in the later classification process are confounded by their presence. The pitch is relatively easy to robustly detect as it has a constant colour and texture. We, therefore, detect the pitch area and remove all objects not bounded by it. Following this, the pitch itself is removed leaving only the players.

In order to detect and remove the pitch we employ a technique similar to chroma-keying to remove the green component from the image. Our method here closely follows the approach of [7] for pitch tracking over multiple frames for broadcast purposes. Within each image we histogram the hue values of pixels and remove the pixels that belong to a small neighbourhood around the largest peak of the histogram. This technique is sufficient for our purposes even though it fails to remove other artefacts, such as sponsor logos on the pitch, we are not interested in and have no effect on subsequent classification. Examples of the background removal are shown in Figure 1. An exemplar histogram of the hue values in a typical image is shown in Figure 2, with a distinct peak in the green section of the colour space.

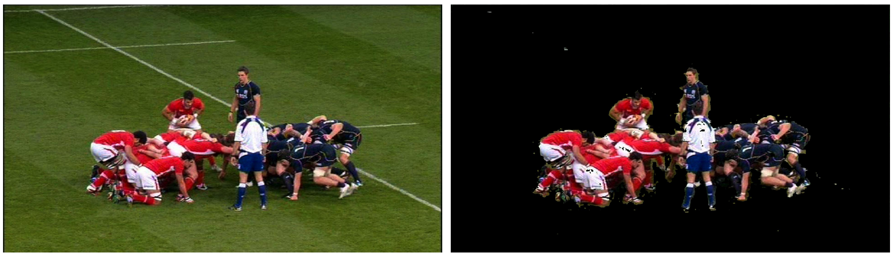

Figure 1: Images before (left) and after (right) background subtraction.

As the peak is so distinct within the colour space no advance technique is employed to extract the relevant area. Instead a simple walk from the maximum point is used, terminating when a point of inflection are detected. Once the area determined to be the pitch is decided, pixel outside of the convex hull of the pitch are also rejected. This ensures that the crowd is removed that would not be detected by the chroma-keying.

PHOW Image Description: The next stage is to compute a suitable description of the shapes of the scrums *etc.* that has sufficient discriminative power for classification. We have explored a variety of image description techniques. Initially starting with the popular SIFT feature descriptor and, in particular, the pyramid histogram of words (PHOW) [1]. It essentially

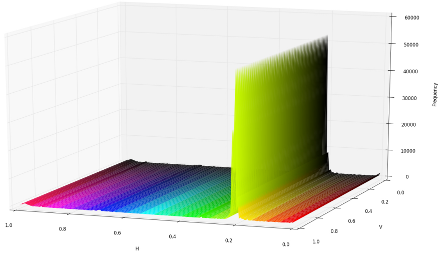

Figure 2: Histogram of hue values within the HSV colour space. A fixed saturation value of 1 is chosen for visualisation.

uses dense SIFT descriptors over a variety of scales to achieve higher scale independence. Each video frame is described in terms of these PHOW features. We then cluster these features into a fixed number of "visual words", as in [1, 6]. This quantisation yields a compact summary which is also more invariant to noise. In our experiments, we use 300 visual words to describe the scene in order to ensure we are not too coarsely quantising the large number of SIFT features that we are computing, this is similar number to those used classifying on the Caltech 101 dataset [2].

Once a suitable vocabulary is generated, the visual words are then spatially binned to produce a spatial histogram of the given image. Using a coarse 2 × 2 grid within the image, we aim to describe the key regions of the scene, while minimising the number of spatial bins used to increase the pipeline's ability to deal with minor changes in scale that cover the majority of eventualities occurring in the data. A diagram summarising the key stages of the PHOW pipeline is shown in Figure 3.

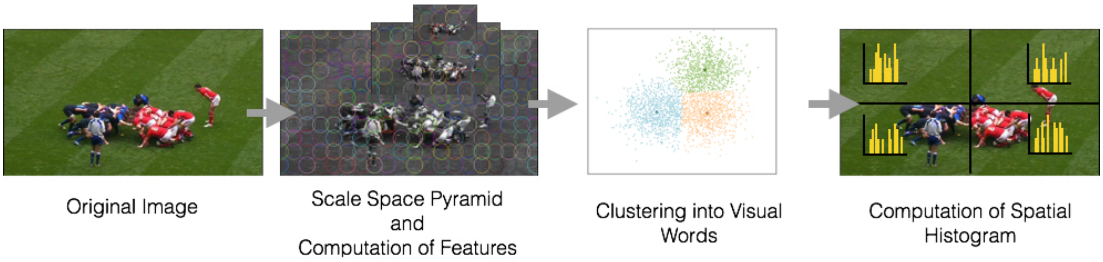

Figure 3: Diagram of the key stages with computing the PHOW descriptor.

Frame-by-Frame Event Classification: Having produce a feature description of each frame we finally proceed to classification. We use a Support Vector Machine (SVM) in order to find the separating plane(s) between the classes. The SVM which has become a popular workhorse supervised classifier in computer vision. We have observed a relatively low performance of a classic linear SVM which suggests that our feature vectors are not linearly separable. We therefore make use of an explicit feature map using the χ^2 kernel as described in [8]. This combined approach of using explicit feature maps and linear SVMs is more efficient than using a non-linear SVM for the large corpora of available data.

3 Results

We have a performed an experimental validation of our approach using broadcast footage as described below.

Video Footage Data To train and test the classifier we took a collection of annotated games from the 2012 Six Nations rugby tournament. This data was annotated with high level labels showing the timestamps of scrums and lineouts by expert rugby video analysts.

In our preliminary experiments (the focus of this work) we focused on two games, Wales V Scotland for training and Vales V France for testing. Due to the broadcast nature of the footage, the game is captured from several cameras. However, we use only footage from single camera that covers the primary action rather than focusing on specific players or wide-angle views (see Figure 4).

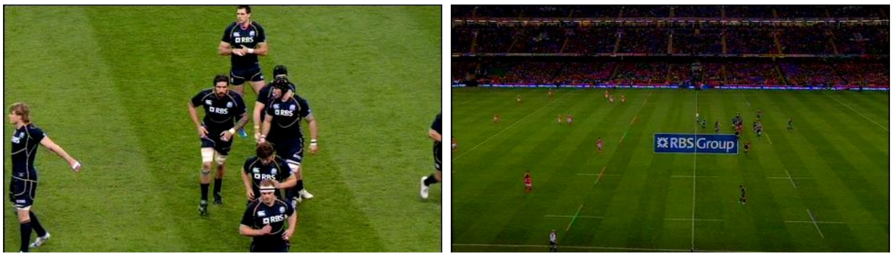

Figure 4: Example frames from a typical footage, close zoom (left) and far zoom (right).

This camera zooms tightly to the specific events on the pitch at a raised angle. Hence, it is likely that to a certain extent our classification is influenced by the actions of the cameraman and how they choose to frame the shot, although from our experience cameramen are fairly consistent in framing the action across games.

Experiment We have a carried a basic experiment to validate our approach. From the video data we extracted all examples of scrums and lineouts. The training data was exclusively extracted from the Scotland game. The test data was obtained from the France game. Features were extracted as described above and a classified built to discriminate between the two classes: scrums and lineouts. There were 750 frames per class (1500 in total) in the training set and 200 frames for each class in the test set, randomly chosen from the entire corpus. The results in Table 1 demonstrate the performance characteristics of the trained classifier.

	True	False
Positive	115 (29%)	0 (0%)
Negative	200 (50%)	45 (11%)

Table 1: Classification performance.

These are encouraging, showing an 88.75% classification accuracy overall. Close examination of failure cases revealed that some misclassifications are a result of coarse annotations where breaks or resets of actions (e.g. when a scrum is reset) are not clearly delineated in normal class boundaries. They actually look more like other plays (e.g. broken play).

4 Conclusion and Future Work

We have developed a practical solution to a novel problem of rugby event classification that has great potential to automate the current manual analysis process. Our current method does not take advantage of any temporal information. We intend to incorporate this in the future. Some initial investigation in using temporal models such Hidden Markov Models has shown promise. More refined domain specific features are also under investigation.

References

[1] Anna Bosch, Andrew Zisserman, Xavier Mu, and Xavier Munoz. Image classification using random forests and ferns. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[2] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[3] H Li and M Flierl. Sift-Based Multi-View Cooperative Tracking For Soccer Video. *Speech and Signal Processing*, pages 1001–1004, 2012.

[4] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[5] Wei-lwun Lu, Jo-anne Ting, James J Little, and Kevin P Murphy. Learning to Track and Identify Players from Broadcast Sports Videos Shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716, 2013.

[6] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470 – 1477, 2003.

[7] Graham Thomas. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing*, 2(2-3):117–132, 2007.

[8] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492, 2012.