Simple Motion Activity Cue for Automatic Sport Video Categorisation

Edward Jaser

Royal Scientific Society
Information Technology Centre
Amman, Jordan

William Christmas, Josef Kittler

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

Abstract

Technology has been playing a major role in facilitating the capture, storage and communication of multimedia data, resulting in a large amount of video material being archived. To ensure its usability, the problem of automatic annotation of videos has been attracting the attention of much researches. In previous work, we proposed a multistage decision making system to deal with the problem of automatic sports video classification. The system is founded on the concept of cues, i.e. pieces of visual evidence, characteristic of certain categories of sports that are extracted from key frames. In this paper we propose a simple cue generation method based on the motion activity. The cue is motivated by the fact that different sports have different level of motion activity (e.g. snooker has little activity while rugby has high motion activity). This cue, together with other cues, will be used by the decision making system to classify video shots. Experimental results on sports video materials demonstrate the benefits of the motion cue generation method.

Keywords : Sport video, dominant colour, motion, semantic cue.

1. Introduction

The generation of digital multimedia content continues to witness phenomenal growth. In the particular domain of sport, many events are taking place every day, and an overwhelming vast amount of sport video materials are being recorded and stored. Ideally, and to ensure usability, all this sports material should be annotated, and the meta-data, generated on it, should be stored in a database along with the video data. This would allow the retrieval of any important event at a later date. Such a system has many uses, such as in the production of television sport programmes and documentaries. Due to the large amount of material being generated, manual annotation is both impractical and very expensive.

In previous work [2], we propose a multistage decision making system to deal with the problem of automatic sports video classification. The system is founded on the concept of cues [7], i.e. pieces of visual evidence, characteristic of certain categories of sports that are extracted from key frames. The main decision making mechanism is a boosted decision tree which generates hypotheses concerning the semantics of the sports video content. The final stage of the decision making process is a Hidden Markov Model system which bridges the gap between the semantic content categorisation defined by the user and the actual visual content categories. The latter is often ambiguous, as the same visual content may be attributed to different sport categories, depending on the context. Different cue detection methods have been developed based on visual features such as colour and texture [5, 6, 8]. Each method can be used to form a number of different cue-detectors provided that suitable training data is available. In this paper we propose a cue generation method based on motion activity.

Motion is an important source of information used to provide valuable cues. Motion information in various forms has been deployed in many existing systems addressing sport video annotation. In [1], Gong et al. used block matching method for motion detection for the automatic parsing of football videos. Zhou et al. [9] used motion information to develop rule-based classification system addressing *Basketball* videos. They used two statistical motion descriptors: dominant motion direction and motion magnitude of the motion vectors. Kobla et al. [4] used motion information as part of a set of features for distinguishing sports video clips from other clips. Various global camera/object motion statistics were computed for this purpose. Kijak et al. [3] used motion information to calculate an activity measure that reflects camera motion for each shot. They used this measure, together with dominant colour information, to identify *global view* shots. Most of these work addressed a single discipline. Also they are computationally expensive. Different sport has different pattern of object and camera motion. For example, global view snooker sequences are usually characterised with very little object motion and almost no camera motion, while object motion in global view rugby sequences are quite high as the case with camera motion. Motivated by these observations, we describe a simple and computationally fast motion cue generation method that estimates motion activity between two frames in the same shot.

The paper is organised as follows. In Section 2 we give an algorithm for dominant colour region detection. We then describe how to generate a Motion Activity Map (MAM) from two key frames using dominant colour information in Section 3. The results of experiments designed to demonstrate the performance of the cue generation method are presented in Section 4. The paper is concluded in Section 5.

2. Dominant Colour

Most sports are played on a field that have a uniform colour (e.g. tennis court). A significant portion of most sport videos represent *global views* in which a single colour (i.e. playing field) usually occupies a significant area of the image. Computing the occupancy of the dominant colour in an image serves as a good feature to train cues to detect such characteristic views. Also it is used to compute the motion activity map (as described in the next section). We first need to generate the dominant colour map (DCM) of an image. This involves two steps. In the first step, we construct a colour cube from the image. We assume that each pixel in this image is represented by ($N \times 3$) bits in the *RGB* colour space (i.e. N bits for each component). A quantisation over each colour component is performed to group all possible colours into $2^{n\times 3}$ different colours by considering only the n most significant bits of each component. The grouping is done using the following equation:

$$(r\ g\ b)^T = (\lfloor \frac{R}{2^{N-n}} \rfloor\ \lfloor \frac{G}{2^{N-n}} \rfloor\ \lfloor \frac{B}{2^{N-n}} \rfloor)^T \qquad (1)$$

A $2^n \times 2^n \times 2^n$ colour cube $\mathcal{C}$ representing the image is then created using the quantised version of each pixel. Each cell in this cube represent one colour $(r\ g\ b)^T$ and contains the frequency of this colour in the image, i.e.:

$$\mathcal{C}(r,g,b) = \sum_{y=1}^{H} \sum_{x=1}^{W} M(f(x,y),\ (r\ g\ b)^T), \quad 0 \le r,g,b < 2^n \qquad (2)$$

where W is the width of the image, H is its height, $f(x,y)$ is the quantised colour components of a pixel in the original image at x and y position (using equation 1) and $M(C(i,j), rgb)$ is defined as follows:

$$M(f(x,y),(r\ g\ b)^T) = \begin{cases} 1 & if \quad f(x,y) = (r\ g\ b)^T \\ 0 & \quad otherwise \end{cases}$$

Once the colour cube is created, we identify the dominant colour region in this cube $\mathcal{C}^d_{\hat{r},\hat{g},\hat{b}}$ that contains pixels representing the dominant colour region. This can be done as follows:

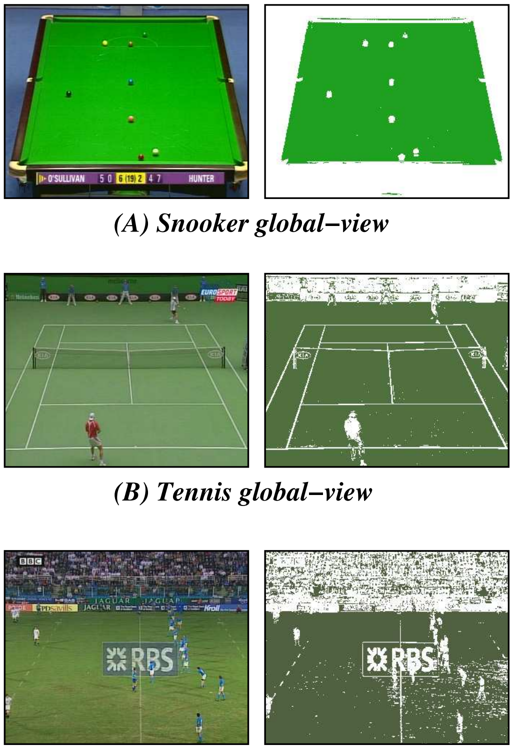
$$\mathcal{C}^d_{\hat{r},\hat{g},\hat{b}} = \arg\max_{r\ g\ b} \sum_{x=r-d}^{r+d} \sum_{y=g-d}^{g+d} \sum_{z=b-d}^{b+d} \mathcal{C}(x,y,z) \qquad (3)$$

where $d \le r,g,b < 2^n - d$, and d is a constant that controls the width and height of the dominant colour region.

Once we have identified the dominant colour region, the DCM can then be generated as follows:

$$DCM(x,y) = \begin{cases} 1 & if \quad f(x,y) \in \mathcal{C}^d_{\hat{r},\hat{g},\hat{b}} \\ 0 & \quad Otherwise \end{cases}$$

The process of generating the DCM is illustrated illustrated in Figure 1. Figure 2 shows example images representing *global view* of three different sports and the corresponding DCM.



(A) Snooker global–view

(B) Tennis global–view

(C) Rugby global–view

Figure 2: *Images showing a global view of three sporting events (left column) and the corresponding dominant colour map (right column). Images belonging to the global view category are frequent in most of the sporting events. The dominant colour usually occupies the centre and the bottom of an image.*

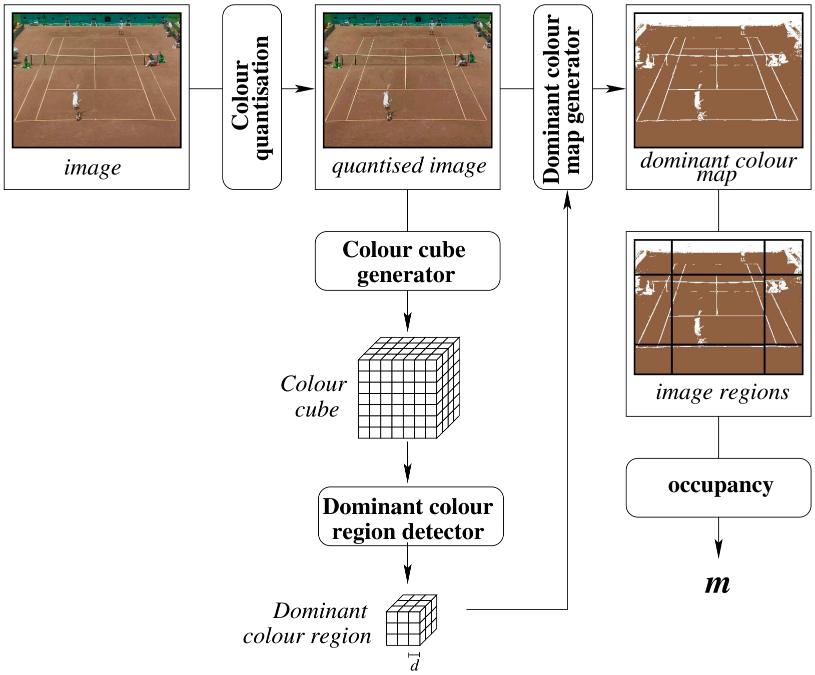


Figure 1: *The process of generating the measurement for the dominant colour cue.*

3. Motion Activity Map

From the DCM examples shown in Figure 2, we observed that objects are represented as small blobs in the dominant colour region. Comparing the position of these blobs in two adjacent frames can be a good indicator of both camera and objects motion. Some sports has little motion activity compared to others. In *snooker* for example, shots representing *global views* have, most of the time, very little motion activity. This is due to virtually no camera motion and the small size of most moving objects (*snooker balls*). In *tennis*, the camera moves in *global view* when a player close to the camera moves to one side of the court. Object motion in *tennis* is basically the movement of the players. In *rugby*, camera and object motions varies depending on what is happening in the game. For example, motion activity in *scrum* is less than that of a run.

To generate the feature vector for motion activity cue generation method, we first need to compute the motion activity map (MAM). The MAM is generated from the dominant colour map information (i.e. DCM) of two frames as follows:

$$MAM(x,y) = \begin{cases} -1 \quad if & DCM_t(x,y) = 0 \; and \\ & \quad DCM_{t+1}(x,y) = 1 \\ +1 \quad if & DCM_t(x,y) = 1 \; and \\ & \quad DCM_{t+1}(x,y) = 0 \\ 0 & otherwise \end{cases}$$

Figure 3 shows the steps needed to obtain the (MAM) using the dominant colour information of two frames. Figure 4 shows some examples of the MAM of frames selected from four different sports. As we can see from the generated MAMs, *snooker* has very little motion. Motion activity is higher when the camera is close to the subjects as we can see in Figure 4(e). Also we can see that sports differ in the number of moving objects: two in Figure 4(a), one in Figure 4(c) and many in Figure 4(g). Once the MAM is computed, we divide the MAM image into nine regions as shown in Figure 3. A 10-D feature vector is then computed containing the motion pixels ratio in the MAM and the motion pixels ratio for each of the nine regions. The motion activity based cue detector is then trained by providing a suitable training sets and a decision tree learning algorithm.

4 Experimental Results

Experiments were conducted on a database comprising video material from three video tapes representing three different sporting events: **Tennis** (the first two sets of the 2003 Australian open men's final), **Rugby** (the second half of the 2003 Six Nations Cup match) and **Snooker** (four frames from final match of the 2003 Masters). In each experiment, two configurations are considered. In the first, 10% of the material is used for training and the remaining 90% for testing. In the second configuration, and to study the effect

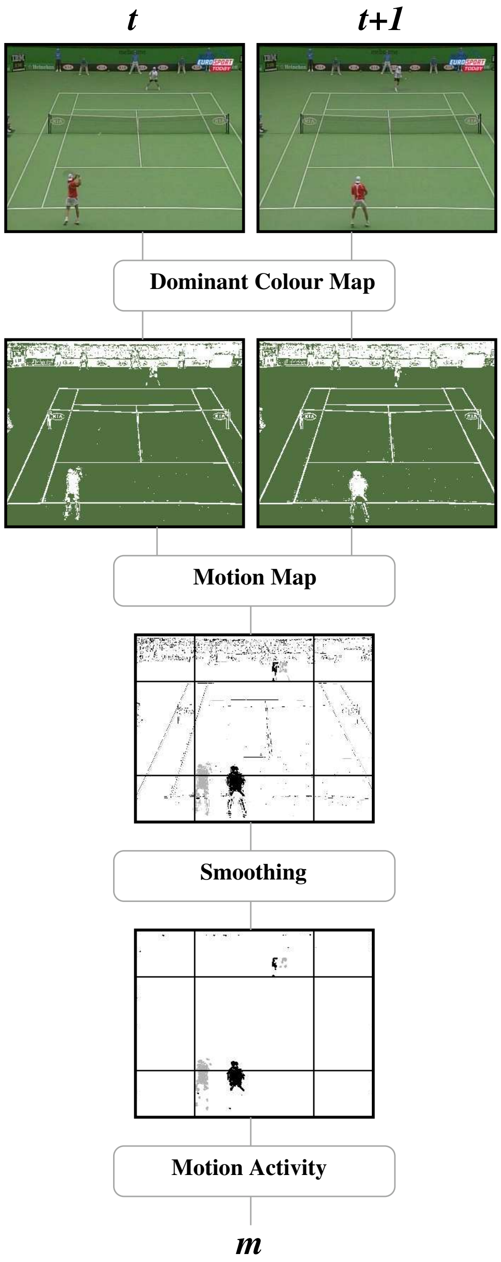
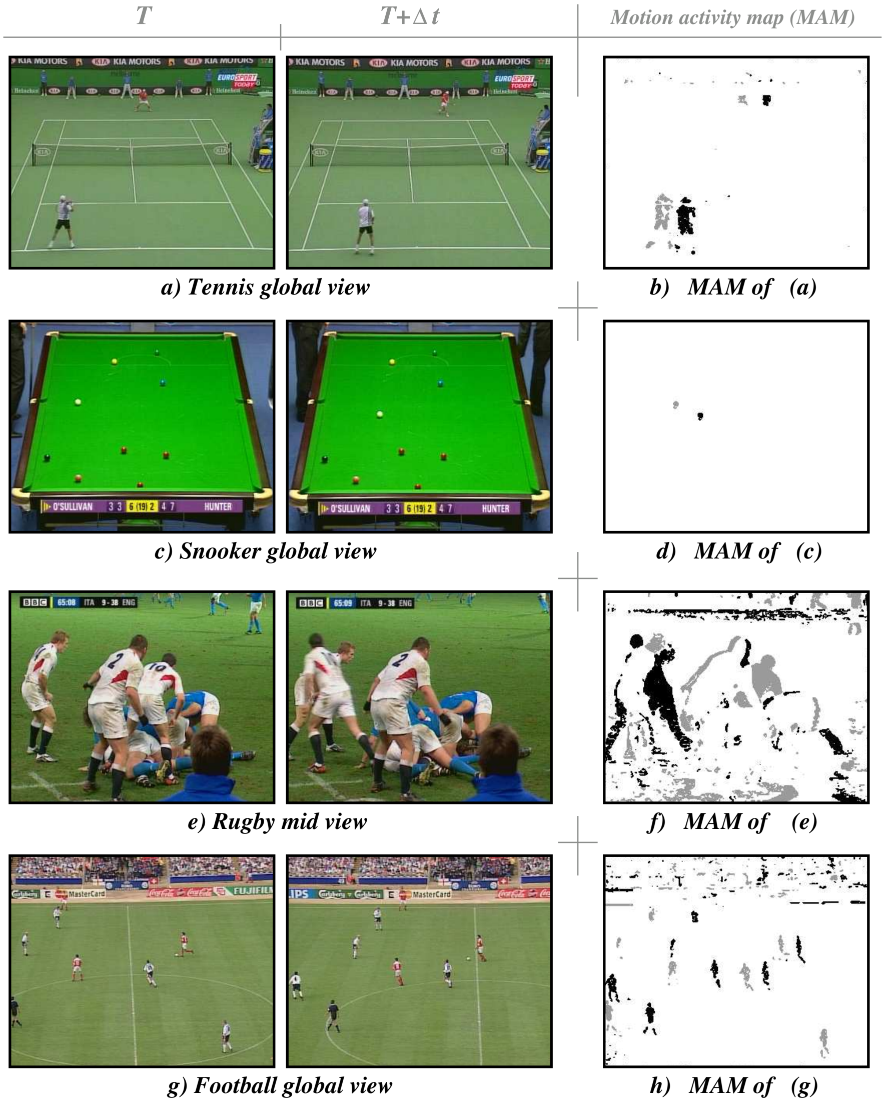


Figure 3: *The process of generating the motion map from two key frames. The motion map captures both motion of the objects in the scene as well as the camera motion. In this example, we notice that only the tennis players are moving. Camera motion is almost non-existent.*

of increasing the size of the training set, 20% is used for training and 80% for testing. Each experiment is repeated 10 times and the average rate and the standard deviation of each measure are reported.

To evaluate the performance of the motion activity cue method, three cue detectors were created and trained: one for detecting *snooker* global view segments, one for detecting *tennis* global view segments and one for detecting *rugby* global view segments. The three cue detectors were used to detect the corresponding cues in the database. Table 1 shows the performance of the *snooker* cue. We noticed that by increasing the training set size, the recall rate did not change significantly. However, precision did improved significantly. This is because the camera motion is very low as the case with moving objects. Table 2 and Table 3 shows the result of *tennis* and *rugby* cues respectively. Unlike



a) Tennis global view ***b) MAM of (a)***

c) Snooker global view ***d) MAM of (c)***

e) Rugby mid view ***f) MAM of (e)***

g) Football global view ***h) MAM of (g)***

Figure 4: *Four examples of motion activity maps (MAM) computed from frames selected from four different sports. As we can see from the generated MAMs, Snooker has very little motion. Motion activity is higher when the camera is close to the subjects as we can see in (e). Also we can see that sports differ in the number of moving objects two in (a), one in (c) and many in (g) and their sizes.*

Experimental Setups	Overall Classification rate	*snooker*	
		Recall	Precision
Setup 1: 10% training, 90% testing	93.8%(± 2.3)	95.7%(± 3.3)	91.9%(± 4.6)
Setup 2: 20% training, 80% testing	94.9%(± 1.3)	95.0%(± 1.9)	94.6%(± 3.5)

Table 1: *The performance of the "snooker" cue based on the motion activity.*

Experimental Setups	Overall Classification rate	*tennis*	
		Recall	Precision
Setup 1: 10% training, 90% testing	91.0%(± 2.9)	85.5%(± 9.1)	82.3%(± 7.3)
Setup 2: 20% training, 80% testing	94.7%(± 1.9)	92.2%(± 3.7)	88.4%(± 4.1)

Table 2: *The performance of the "tennis" cue based on the motion activity.*

Experimental Setups	Overall Classification rate	*rugby*	
		Recall	Precision
Setup 1: 10% training, 90% testing	89.3%(± 2.2)	77.6%(± 8.6)	81.6%(± 7.8)
Setup 2: 20% training, 80% testing	92.5%(± 1.8)	82.5%(± 4.6)	88.4%(± 6.6)

Table 3: *The performance of the "rugby" cue based on the motion activity.*

snooker, *rugby* and *tennis* recall and precision rates are significantly improved by providing more training samples, as the motion activities for both cameras and objects in these two sport are more sophisticated.

5 Conclusion and Future Work

In this paper we have described a cue generation method based on motion activity information. Dominant colour information is used to compute the motion activity map from two adjacent key frames. The map is then used to generate the feature vector that are used to train the cue. The cue is simple and computationally fast. This cue, together with other cues, is used by a hierarchical decision making system to classify video shots and generate semantic annotation of sport videos.

References

[1] Y. Gong, L.T. Sin, and C.H. Chuan. Automatic Parsing of TV Soccer Programs. In *In IEEE International Conference on Multimedia Computing and Systems*, page 167 174, 1995.

[2] E. Jaser, W. Christmas, and J. Kittler. Temporal Post-Processing of Decision Tree Outputs for Sports Video Categorisation. In A. Fred, T. Caelli, R. P.W. Duin, A. Campilho, and D. Ridder, editors, *Proceedings of the IAPR Inter. Workshops on Structural, Syntactic, and Statistical Pattern Recognition (S+SSPR 2004)*, volume 3138 of *Lecture Notes in Computer Science*, pages 495 – 503. Springer-Verlag, August 2004.

[3] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Audiovisual Integration for Tennis Broadcast Structuring. In *International Workshop on Content-Based Multimedia Indexing (CBMI'03)*, September 2003.

[4] V. Kobla, D. DeMenthon, and D. Doermann. Identification of sports videos using replay, text, and camera motion features. In *Proceedings of the SPIE Conference on Storage and Retrieval for Media Databases* , volume 3972, pages 332–343, 2000.

[5] B. Levienaise-Obadia, J. Kittler, and W. Christmas. Defining Quantisation Strategies and a Perceptual Similarity Measure for Texture-Based Aannotation and Retrieval. In *In IEEE, editor, ICPR'2000*, volume III, 2000.

[6] J. Matas, D. Koubaroulis, and J. Kittler. Colour Image Retrieval and Object Recognition Using the Multimodal Neighbourhood Signature. In *D Vernon, editor, Proceedings of the European Conference on Computer Vision LNCS*, volume 1842, pages 48–64, 2000.

[7] K. Messer, W.J. Christmas, E. Jaser, J.Kittler, B.Levienaise-Obadia, and D.Koubaroulis. A Unified

Approach to the Generation of Semantic Cues for Sports Video Annotation . *Signal Processing*, 83:357–383, February 2005. Special issue on Content Based Image and Video Retrieval.

[8] K. Messer and J. Kittler. A Region-Based Image Database System Using Colour and Texture. In *Pattern Recognition Letters*, page 1323 1330, 1999.

[9] W. Zhou, A. Vellaikal, and C. C. Jay Kuo. Rule-based video classification system for basketball video indexing. In *ACM Multimedia Workshops*, pages 213–216, 2000.