Real-time Event Classification in Field Sport Videos

Rafal Kapela[a,*], Aleksandra Świetlicka[a], Andrzej Rybarczyk[a], Krzysztof Kolanowski[a], Noel E. O'Connor[b]

[a] *Department of Computer Engineering, Poznan University of Technology, Poland*
[b] *INSIGHT: Centre for Sensor Web Technologies at Dublin City University, Dublin, Ireland*

Abstract

The paper presents a novel approach to real-time event detection in sports broadcasts. We present how the same underlying audio-visual feature extraction algorithm based on new global image descriptors is robust across a range of different sports alleviating the need to tailor it to a particular sport. In addition, we propose and evaluate three different classifiers in order to detect events using these features: a feed-forward neural network, an Elman neural network and a decision tree. Each are investigated and evaluated in terms of their usefulness for real-time event classification. We also propose a ground truth dataset together with an annotation technique for performance evaluation of each classifier useful to others interested in this problem.

Keywords: real-time sports event detection, neural networks, state machines, field sports, sport broadcast

---

1. Introduction

Sport is consistently highly rated in terms of television broadcasts [1, 2] and in some countries, sports broadcast are the most watched broadcasts. This is true especially for significant sporting events like the Olympics or for the national/regional finals of the most popular sport in a given country. Across Europe soccer usually in the center of attention. Based on publicly available

*Corresponding author:
Email address: rafal.kapela@put.poznan.pl

statistics [3], one can observe that matches played in Germany's Bundesliga, the Premier League and Spain's La Liga are watched by over 10 million fans each year with a substantially larger audience watching at home on TV. However, soccer is not the only sport that enjoys significant popularity and large viewing figures. In Ireland, for example, soccer is considered to be in third position alongside rugby, after Gaelic football and hurling [4], the finals of which are guaranteed huge audiences both in the stadium but also in front of the TV [5]. Considering other countries, we can add the following to the most popular field sports around the world: basketball, rugby, cricket, field and ice hockey or many others [6]. Depending on the country, the success of the local or national team and the time of year, sport can often be considered to be users' most desirable audio-visual information.

As a result, there has been significant interest in algorithms for automatic event detection in sports broadcasts. This is motivated by potential applications such as automatic highlight generation for summarization and second screen applications, indexing for search and retrieval in archives, mobile content delivery either off-line or as an added value in-stadium user experience. However, most event detection algorithms published thus far normally focus on a particular type of the sport (e.g., tennis, soccer, cricket, etc.) and are not robust for other types of sports, thereby limiting their applicability. Like for example event detection systems presented in [7], [8], [9], [10], [11], [12] can work autonomously and some have ability to turn on themself at specific time in order to analyze broadcasted video together with web-casting text. However, this systems suffer from the lack of flexibility that would allow it to analyze more than just one type of sport. This is a very good example of the state of the art in this field – although there are plenty of examples that can be featured with high accuracy all of them work for only one type of sport. This is caused by the fact that different sports present different characteristics either in the rules for that sport of the manner in which it is captured and directed for broadcast. In addition real-time aspect is quite often neglected whereas in most application scenarios where a game is analyzed in order to provide rich content to the end users event extraction time should be one of the main parameters taken into account.

For this reason, in this paper we focus on a generic subset of all sports that can be designated as *field sports*, a term introduced in [13] to refer to any sport played on a grass pitch (soccer, rugby, field hockey, etc.) featuring two teams competing for territorial advantage. In this work, however, we extend this genre to include other sports that exhibit similar characteristics but that are not necessarily played on a grass pitch. Specifically, we extend the definition of field sports to include sports played in a playing arena that features some kind of scoring posts (e.g., goal post in soccer or basket in basketball), whereby the overall objective is territorial advancement with a view to obtaining a score.

Taking into account the diversity of the different field sports a range of event detection algorithms were presented in recent years. Even for one kind of sport the research can be conducted from different points of view. In [14] and [15] researchers pay their attention to the fact, that a low-level simple audio-visual features are often not rich enough to represent semantically complex information on the level appropriate to human perception. As a solution they propose a multi-level multimodal descriptors related to the position of the camera in relation to the players and the field. The results presented by them are impressive (recall and precision on the level of about 90%) however they do not assume that their system to analyze video content in the real-time. It has been shown in [13] that about 97% of interesting moments during a game are followed by a close-up shot presenting a player who scored or who caused some interesting action. In addition, features like end of a pitch, audio activity or crowd shot detection have been shown to be very useful in event detection [13]. The presented system is proven to work with different field sports such as soccer, rugby, field hockey, hurling and Gaelic football. In this work a Support Vector Machine (SVM) was used as a event classifier. However, mainly because of the use of the Hough transform the implementation is very time consuming and inapplicable in real-time systems. A very similar approach is presented in [16]. In order to detect an event the authors declare so called "plays" where mainly a color histogram is calculated plus some heuristics are applied about the regions of histogram detection. An event is categorized using Bayesian Network based on the sequence of camera shots. In this work events were detected in baseball, American football and Japanese sumo wrestling. Another example of work that belongs to this group is presented in [17] where, based on simple visual features like pitch orientation and close-up detection, the authors achieve good accuracy. However, again no time performance is given in the paper and there is a big drop in accuracy when the SVM is trained on the samples that do not belong to the same game. It is worth noting that the three approaches described above [13, 16, 17] are capable of extracting not only goals but also other exciting moments like penalties or close misses. In [18], very simple features like pixel/histogram change ratio between two consecutive frames, grass ratio and background mean and variation in addition to time and frequency domain audio features were used in order to detect events in soccer games. Although reporting high accuracy of the system using simple features the authors do not mention its time performance. Although the acceptance of the MPEG-7 standard in the community has been rather low, there are still approaches based on MPEG-7 descriptors. In [**?**] an event detection and tactics analysis is proposed. This kind of approach could be really useful for coaches and trainers for soccer game analysis after the game but from real-time analysis perspective it is not significantly interesting.

Taking the real-time approach for a given task into consideration the amount of the work is significantly lower. However, there are works worth recommending. In [19] authors use audio-visual features (Scale Invariant Feature Transform, Spatial-Temporal Interest Points, Mel frequency cepstrum coefficients, color moments, etc.) to detect events in Internet videos. The system is capable of working in real-time under an assumption that the interval between the frames for calculation is greater than 2 seconds. The drawback of the approach is in the precision which is on the level of about 50% for all the videos. A very interesting work is presented in [20] where authors present real-time video classification based on dense Histograms of Oriented Gradients/Optical Flow. Based on the results presented there the proposed system is capable of working at speed of almost 13 fps. The results however are presented only for 320×240 resolution short (70-200 frames) videos presenting only human actions. This assumptions are quite unrealistic for wide range of different shots of the sport field, poses and numbers of the players in the shot.

Finally, there have been approaches significantly different from the "standard" low level feature-based systems. In [21] and [22] a very different approaches are taken. Both utilize the information produced by people during a game and tweeted by the popular Twitter website to detect events in different games (soccer and rugby were tested). They are, at first sight, universal approaches, however they can suffer from quite large false positive detection rates, need constant connection to the Internet and introduce some ambiguity in the form of delay between detected and real events making the detection of event boundaries more difficult. The [23] approach uses knowledge-discounted approach to detect events. By introducing a hybrid approach which integrates statistics into logical rule-based models during event detection. It seems to be applicable for not only one type of sport but time performance of the system is not given in the paper.

Our contribution in this paper is to present a novel pseudo-generic real-time system for event detection that addresses many of the limitations of the techniques outlined above. Section 2 presents the scene classification technique itself and a high level architecture of our proposed approach. In the following section we present the core event classification algorithm and three appropriate classifiers that can be used. Section 4 presents the event detection results across a large set of sports genres. This section also shows how we tested our implementation and what dataset we have chosen for this purpose. The article is concluded in section 5 where we present its main advantages and a time performance analysis.

2. The concept of the annotation system

2.1. The idea

In live broadcasts, a key challenge for a sports director is to convey to the viewer what is happening during a sporting event. This is achieved by the director switching between a variety of camera views that help describe what is happening. So for example, this could include showing a long distance shots that show a zoomed out view of the field of play, followed by a closer focus on the scoring area, followed by a close-up of the player involved, a reaction shot of the crowd or manager, etc. Whilst there is no de-facto "script" for how to present these shots, or in what order, these are the tools that a director has at his/her disposal in order to convey excitement and capture an important event. As a result, scene recognition, by which we mean recognizing what kind of camera shot is being used by a sports director at any given moment, is useful input for event detection. Although previous works [13, 16, 17, 24] are mainly based on the analysis of very simple audio- visual features like color histograms, pixel differences or audio intensities in order to detect different types of a shot in sports broadcasts we employ more powerful detection techniques. Say, for example we have to deal with videos that contain soccer and basketball games. Detecting long distance shots (i.e., shots that present the field of a game) based on the color of the pitch/court regarding the diversity of colors of the fields (e.g., muddy grass, wet grass, different colors of the court in the dead-zone region) will not be an efficient solution in practice. For this reason, in order to detect an event based on the sequence of shots we defined fourteen different scene types typically used by a director, which covered about 99% of the video footage in our database. The proposed classes are as follows:

1. close up shot head (simple background);
2. close up shot head (complex background);
3. close up shot head (mixture background);
4. close up shot waist up (simple background);
5. close up shot waist up (complex background);

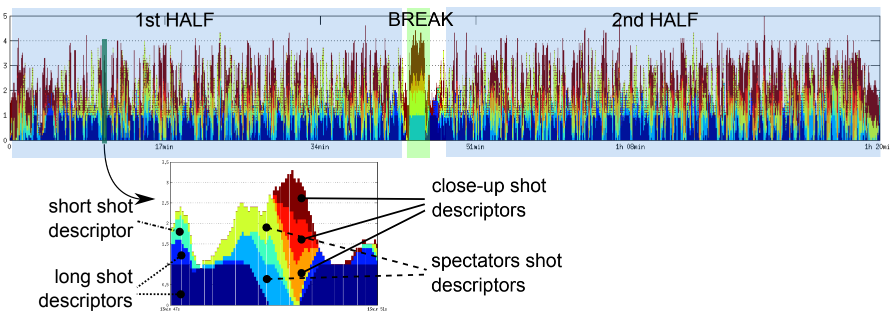

Figure 1: Visualization of an example trace of the visual features for a basketball game where the interesting moments are indicated

6. close up shot waist up (mixture background);
7. short distance shot presenting player(s) (simple background);
8. short distance shot presenting player(s) (complex background);
9. short distance shot presenting player(s) (mixture background);
10. short distance shot presenting spectators;
11. long distance shot presenting center of the field;
12. long distance shot presenting right side of the field;
13. long distance shot presenting left side of the field;
14. long distance shot presenting spectators;

In addition we have proven that these classes appear in all different genres of field sports ranging from Gaelic football to soccer.

Figure 1 presents a stacked bar graph of descriptors for an example field sport game for illustration purposes (in this particular case it is basketball). Since we described our investigation of the choice of the scene/shot detection and recognition algorithms in another paper [25] we do not repeat this here. However, we do note here that, based on our experiments the covariance of some of the descriptors is sufficiently high to omit or merge them together in order to form new ones. To this end, from the original proposed complete set of fourteen, the following 8 scene classes have been chosen along with an audio energy descriptor:

1. maximum of long distance shot presenting left/right side of the field;
2. long distance shot presenting center of the field;
3. short distance shot presenting spectators;
4. short distance shot presenting player(s) (mixture background);
5. long distance shot presenting spectators;
6. close up shot head (simple background);
7. close up shot head (complex background);
8. close up shot head (mixture background).

Each descriptor produces an output normalized to the range $[0, 1]$ that we treat as a confidence associated with that descriptor. The audio energy descriptor is simply an adaptive moving window average filter (1) over the audio intensity samples synchronized with the video stream:

$$a_k^{out} = \frac{1}{N} \sum_{i=1}^{N} a_i^{in} \tag{1}$$

where N is the width of the moving window and k is a position of the filter in the audio stream.

Taking into account the characteristics of the interesting moment in any type of the field sport game we can distinguish three higher level *phases* of camera activity, where the director uses the various camera shots available (figure 1):

1. Center/side of the field shot;
2. Zoom-in on the player who has possession of the ball/puck (optional);
3. Close-up on the player who scored/caused interesting action.

Our proposed descriptors effectively continually monitor different aspects of these three phases of camera activity in terms of the different kinds of shots being used. The various descriptors are "triggered" by different aspects of the three phases, allowing us to build classifiers to differentiate the different phases and on this basis recognize events. In the first phase, a camera usually pictures a large part of the pitch (descriptors marked in dark blue in figure 1) or court with multiple players on it but then pans to one of the sides of the arena where

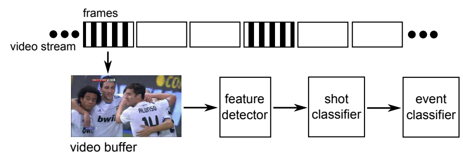

Figure 2: The block architecture of the proposed system

the event is taking place. Since in these types of shots feature spectators sitting on the sides of the pitch descriptors responsible for audience detection become dominant (colors: light blue and green in figure 1). In some sports with very high pace this could be a very quick transition (e.g., basketball) whereas in others it may take a longer (e.g., soccer). Since the data shown in the figure is from a basketball game, this transition is almost immediate (the long center shot descriptor – medium blue color – is visible only at the beginning of the magnified area). This specific action the camera is zoomed on the players at the end of the field, the short shot descriptor (cyan color) becomes more active too. The interesting moment itself ends up in the the final phase where three descriptors responsible for close-up detection (colors: orange, red and brown) are triggered since the camera focuses on the player who scored. Also, one can easily noticed that during a break usually camera focuses on players and spectators. Utilizing this structured appearance of data as the basis of an event means that we can build a classifier which is able to detect these events based on these audiovisual descriptors that differentiate these phases.

2.2. Architecture

The general architecture of our approach is presented in figure 2. It can be seen that the video decoding process is independent from the video annotation procedure thus, enabling the display of the decoded video frames on the user screen but also allowing storage of every k^{th} frame (every 5^{th} frame in our implementation) in the buffer for further annotation analysis. At the bottom of the figure one can observe an analysis pipeline responsible for feature extraction, scene recognition [25] and finally classification of events potentially interesting for the user. This modular architecture makes the system applicable for mobile devices and embedded systems (such as set-top boxes) where, for example, the decoding process usually takes place in a separate hardware acceleration unit because of CPU limitations. This way both processes can work in parallel without introducing any additional delays. Thanks to the modular approach taken in the system design process it is also possible to replace any of the existing modules with new, improved versions that for example utilize additional hardware external to the CPU (e.g., Graphics Processing Unit on dedicated extension card). For example, in our implementation we were able to replace some parts of the algorithm with their CUDA implementation improving the overall performance by 5-7%.

3. Feature extraction & Scene recognition

To solve a problem of efficient description of the video scene sequence we used a technique based on global image description with use of Fast Fourier Transform (FFT). Since in our case we do not have to deal with scene rotation or scale invariance global description based on color distribution provide sufficiently high precision. Our work in [25] shows that non-binary local feature detection algorithms like Scale Invariant Feature Transform (SIFT) [26] and Histogram of Oriented Gradients (HoG) [27] algorithms, that are characterized with the highest efficiency of scene recognition are too slow to be part of a system that has to work under real-time constraints. This work was a precedence to look for less sophisticated, but still of high efficiency algorithms for image description. We analyzed most state-of-the-art key-point extraction algorithms suitable for real time applications like Features from Accelerated Segment Test (FAST) and Features from Accelerated Segment Test – Enhanced Repeatability (FAST-ER) [28] also binary local description algorithms like: Binary Robust Independent Elementary Features (BRIEF) [29], Fast Retina Keypoint (FREAK) [30], Binary Robust Invariant Scalable Keypoints (BRISK) [31], all available in OpenCV library [32]. In our task of field sport scene recognition all of them respond with very similar effectiveness (less than 2% of difference in accuracy between the least and the most efficient). The technique proposed in [25] has one major advantage comparing to the local image descriptors which is robustness to the compression artifacts and video/image quality in general. The underlying idea of the algorithm is to treat the color in the image as it had meaningful of layout. Then particular range of colors is extracted and the very well-known Fourier transformation is used to describe this layout characteristics. Let I be the input image where colors are coded in HSV color space. We convert each pixel to its address representation where each pixel is represented as a single 10-bit value according to the following formula:

$$I^A_{x,y} = 64H_{x,y} + 16S_{x,y} + V_{x,y} + 1 \tag{2}$$

where x and y are the Cartesian coordinates of the given pixel, $H_{x,y}$, $S_{x,y}$ and $V_{x,y}$ are quantized H, S, V coefficients to 16 (4 bits), 4 (2 bits) and 16 (4 bits) levels respectively. Therefore the resulting histogram has 1024 bins (10 bits). It has been called an address representation since the calculated value points to the respective bin of the histogram (i.e., it is an address of the histogram bin). Note that histogram calculated in this way group similar colors with respect to their hue coefficient since H goes to the most significant bits of the address.

Now, let g be a radial basis function (RBF) that traverses the above histogram, so that in a single step i the processed address image (i.e., the image with pixel values converted according to the equation (2)):

$$I^g_i = \exp\left[-\frac{\left(I^A - A_i\right)^2}{\sigma^2_G}\right] \tag{3}$$

where A_i is the address at a given algorithm iteration and σ_G is chosen experimentally [25]. Note, that in a particular iteration A_i the resulting image I^g_i will have nonzero values only in the pixels which values fall into the span of the RBF (i.e., since hue component is the most significant – the pixels with similar hue value). This idea has been visualized in the figure 3. Note, that for field sports we usually deal with very convenient situation where the object and background are very contrastive and are composed of limited number of colors. Thanks to

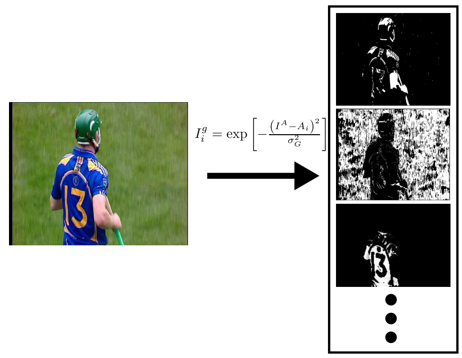

Figure 3: The idea of filtering the image with the address approach

this method we can capture layout of the colors in the image and then analyze it in further processing.

The image address representation is then transformed with a 2-D Fourier transform which gives us a result – the frequency representation of the particular color distribution in the image (i.e., the filtered address image represents a part of an object/background texture in the image). The next step is filtering the Fourier representation with a set of Gabor filters [33]:

$$G_{K,L}(\omega, \theta) = \exp\left[\frac{-(\omega - \omega_K)^2}{2\sigma_{\omega_K}^2}\right] \exp\left[\frac{-(\theta - \theta_L)^2}{2\sigma_{\theta_L}^2}\right] \tag{4}$$

where K and L are radial and angular indexes respectively, ω_K and θ_L are the polar coordinates of the filter center. The setup of $\sigma_{\omega,\theta}$ values is the same as in in [33]. This results in 30 Gabor filters that span the Fourier space and give higher granularity for low frequencies.

The last step is a composition of the Gabor filter responses for every iteration of the I_i^g function in order to receive the overall information about the scene. This can be seen as a composition of partial informations about the layouts of the particular colors. Thanks to this composing this method works even for images with complex backgrounds. Based on the linearity of the Fourier transform we can add all the results of single step calculations for the same value of σ_G^2 into one result matrix by simply summing them and performing Gabor filtering only once at the end. So the resulting equation becomes (M is the number of steps chosen experimentally [25]):

$$F = \frac{1}{M} \sum_{i=1}^{M} \mathcal{F}\{I_i\} \tag{5}$$

Thanks to performing the filtering step only once we can achieve a very quick feature extraction method (i.e., less than 40ms). In addition, calculations for different sizes of the σ^2_{RBF} factor can be done independently, thus we can combine the results in order to train and evaluate a set of SVMs for every given class. This technique allows us to choose the best performing combination of features and SVMs for every class. Results presented in this paper show that the proposed algorithm provides the same accuracy as sophisticated and slow feature extraction algorithms like SIFT [26] or HoG [27].

4. Event classification

The event classification system comprises of a main module which is a classier and a submodule which gathers the responses of the previous one and makes the final decision about the detection of the event. The latter one is described in the section 4.2.

4.1. Event recognizer

4.1.1. Decision tree

Natural thing was to look for a classifier among deterministic methods of classification. One of the simplest seems to be the state machine, but creating it manually turns out to be impossible while there is too much data to process. That is why we focused on decision trees, which find their implementations in many computable environments (e.g. MATLAB).

Decision tree is built from the following components:

- – internal node - represents a test on an attribute,
- – nodes - where the decision is made which path will be followed,
- – leafs (branches) - which represent options of these decisions.

In our case we naturally examined two different structures, in first one we adopted 9 decision variables for 9 descriptors, while in the second - 18 decision variables, while we took into account the following moment of time. The results obtained with the first structure were not satisfying enough, hence we tried to use the bigger structure, like in the case of neural networks. Depending on the structure decision tree contains around 2 500 or 5 200 nodes.

4.1.2. Feed-forward neural network

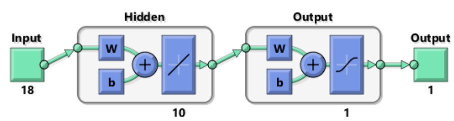

Figure 4: The architecture of the feed-forward neural network used for event detection

An intuitive choice for a neural network is the very well known feed-forward multi-layer perceptron neural network (MLP) shown in the figure 4. This was our initial choice since this kind of network facilitates a good trade-off between its generalization capabilities and complexity of the architecture [34]. The following structure for the network appeared to be the most efficient:

- eighteen inputs related with nine given descriptors and nine descriptors from the previous frame;
- ten neurons with linear activation function in the hidden layer;
- one output that refers to the attraction of the current scene of the game using the sigmoidal activation function.

In the learning process we used around 20 000 samples from six different games (chosen randomly), which gives approximately one hour and six minutes of a match. We used the Levenberg-Marquardt algorithm [35, 36] for training. The neural network reaches the minimum of the gradient after around 10 iterations, which, due to fast convergence of the training process, confirms that the network was able to learn the classification task. In order to avoid over-fitting

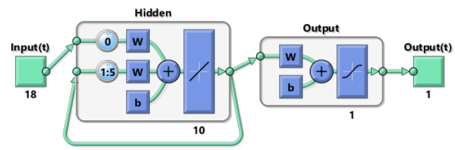

Figure 5: The architecture of the Elman neural network used for event classification

of the network to the training data we used an early stopping technique as part of the training process. Strikingly, the network seems to reach saturation for event detection since larger number of neurons in the hidden layer followed by more training samples do not increase the overall accuracy, which reached 65% in the best case. This accuracy relates to event detection in all kind of sports in our dataset. Since all the simulations of our network led to the conclusion that we had reached a saturation point in recognition capabilities, our next choice was a network which not only analyzes the current state of the game but also utilized information about previous values of descriptors. We investigated recurrent neural networks with the Elman network as a representative example of this class [37].

4.1.3. Elman neural network

The fixed back connections in the Elman network result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). Thus, the network can maintain a sort of state, allowing it to perform such tasks as sequence-prediction that are beyond the power of a standard multilayer perceptron and this is clearly desirable in our case. On the other hand, the Elman neural network has been proven to be unpredictable in terms of the generalization of any function [38] (or in other words the generalization of the Elman neural network cannot be guaranteed). However, [38] also shows that any arbitrary information, due to the existence of a hidden layer, can be encoded in the inputs since the length of the input vector is not restricted. Providing additional information about previous state of the scene description and using feedback from the hidden layer of the network any dichotomy can be stored as an input and easily used for event categorization [38]. In addition, we also know that any finite automaton can be represented by a recurrent neural network [37] making our choice a natural extension of the state machine idea. Of course the more complex architecture of the network results in higher computational complexity thus increasing its execution time [39, 40] so that the number of neurons and their activation functions must be carefully chosen in order to achieve real-time operation.

The input layer of the network was constructed in two different ways. The very first and natural choice was to use nine inputs that refer to nine descriptors, but after a number of simulations we decided to improve its construction so that the information about previous scene description serve as additional inputs to the network. This way we have a network with eighteen neurons in the input layer (nine descriptors of the current state, nine descriptors of the previous state).

In the training process we used the Levenberg-Marquardt algorithm, as the most effective one (we compared the results with the simple gradient descent method which did not give satisfactory results). After only around 30 iterations the neural network reached the minimum of the gradient. This result allows us to state that despite its complexity and the unpredictability of its generalization behavior, the Elman neural network, can be trained very fast for the same task allowing even better classification results.

4.2. Final event classification

The scene recognition algorithm works in a binary fashion, so that only the 0-1 information about the scene classification is available after this phase. After this stage we applied a sliding window approach that measures the responsiveness (i.e., number of responses in the window) in order to classify the event. The approach is depicted on the figure 6. Note, that in order to capture the idea of the sequence of the shots we applied two sliding windows that are moving synchronously with constant width W and gap S between them. The event

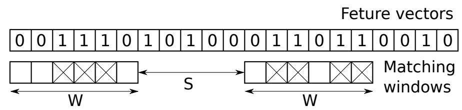

Figure 6: The visualization of the matching windows for event classification

is detected when the number of ones within the windows is equal or greater that given threshold T. The influence of the mentioned parameters for different sports and types of the events is presented in the next subsection.

5. Results

5.1. Dataset

In our experiments we used a manually annotated ground-truth dataset of sport videos. The dataset comprises of about 50 hours of sports including hurling, Gaelic football, basketball, rugby, soccer and cricket. In order to create the ground-truth our annotators analyzed the footage marking the following features:

- the time stamp of the beginning of the interesting action;
- the interesting point (if applicable) such as a goal between the beginning and end of the interesting action;
- the time stamp of the end of the interesting action;
- the information if the action included a score;
- the binary information about the level of excitement or importance of the event e.g., a goal/try vs a point/penalty in sports with different scoring mechanisms.

5.2. Classification results

Event detection and recognition is quite subjective task. This is true in particular for non-goal events where sometimes it is hard to determine if the captured moment is of high value for the viewer (especially when they are not interested in a player/team that caused the event) or sometimes simple the game does not contain many events. In this case we would have plenty non-event moments that is obviously not desirable for training the classifier since it may produce offset in the solution. This situation would affect the event standard effectiveness measuring factors like accuracy, precision and recall and make them inadequate for this task. On the other hand the factors that stand for accuracy of the system should give undoubtful information about the accuracy of the event detector. For this reason in our work we introduce different than standard accuracy measures that are focused on around the event detector performance (6)-(8):

$$\text{MA} = \frac{|\text{DE}-\text{DTE}|}{\text{NE}} \tag{6}$$

$$\text{P} = \frac{\text{DTE}}{\text{DE}} \tag{7}$$

$$\text{MP} = \frac{\text{DTE}}{\text{NE}} \tag{8}$$

where NE refers to number of events in a match, DE to number of detected events by the classifier and DTE to number of detected true events (i.e., the ground-truth size). Note, that modified accuracy (MA) tends to be close to zero for the systems that are characterized with high effectiveness and its not restricted to one for the low-performance systems. Precision (P) and modified precision (MP) were introduced in order to provide additional information about the effectiveness. They are useful in the situation where there are not many events in a analyzed game (i.e., when DTE is close to NE).

The same number of tests were performed for all three proposed classifiers. In order to verify the generalization of the proposed classifiers used data taken randomly from all the games for training. For each considered scenario we used around 20000 data samples for training classification tree and both neural networks respectively. For all experiments we used tree created and optimized with [41] algorithm available in Matlab, shown in 4.1.1 section, feed-forward and Elman neural networks with nineteen neurons in the input layer. For each game we calculated the factors presented above (6)-(8). For the event scenario we take

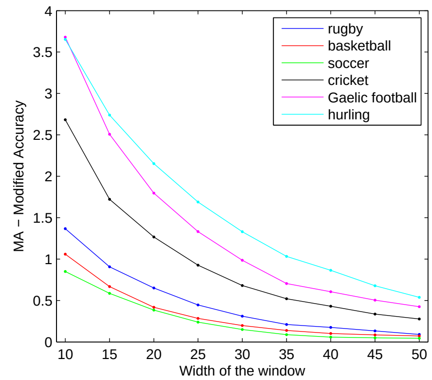

Fig. 7: Modified accuracy (MA) for different sizes of window

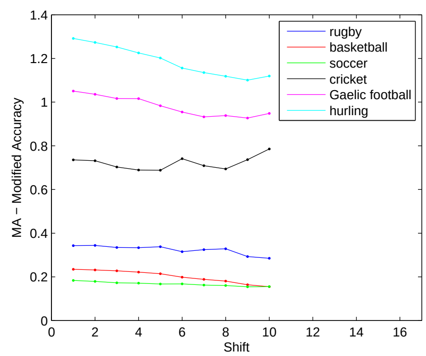

Fig. 8: Modified accuracy (MA) for different delays

all the entries in the dataset and we check the event recognition system output at every time stamp in order to calculate the measures. Tests of the system were performed on six different games: rugby, football, basketball, cricket, gaelic football and hurling.

Figures 7 to 9 show the influence of the sliding window solution to the accuracy of the event recognition system. As it can be seen the bigger the window size and the higher the number of positive classifier responses the better the effectiveness of the system (figure 7 and 9). This can be explained by the fact that the wider windows capture more temporal information about the sequence of the shots and event in general. The gap between the two windows (the S

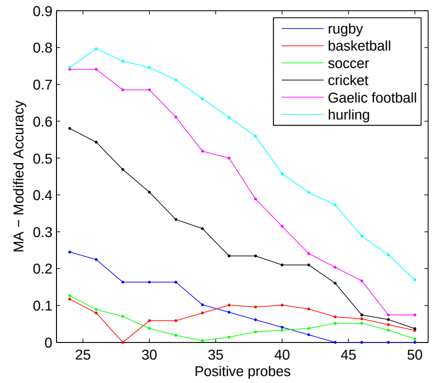

Fig. 9: Modified accuracy (MA) for specified window size(50) and different number of positive probes within

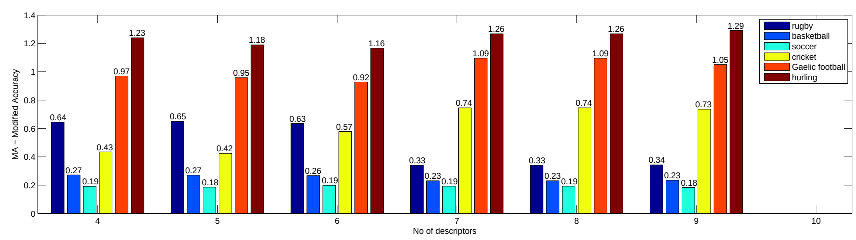

Fig. 10: Modified accuracy (MA) for different number of considered descriptors

factor) does not introduce any improvements – figure 8. Note, that we chose reasonable range of S values (10 shifts stand for 5 seconds of the video).

A natural question in the case of the classification based on any kind of features is what is the robustness of the classifier to the limited number of the features that describe the scene. Figure 10 shows the MA distribution with respect to the number of features in the vector describing the scene. The right most graph is the one for all the descriptors mentioned in 2.1 section. The each following to the left bar graph has limited number of features by one as follows:

1. audio descriptor;
2. short distance shot presenting player(s) (mixture background);
3. long distance shot presenting spectators;
4. close up shot head (complex background);

Tab. 1: True and detected events by Elman neural network set together with MA, P and MP and corresponding window size and positive probes within

	ENET							
	W	WP	NE	DE	DTE	MA	P	MP
rugby	25	15	49	47	45	0.04	0.96	0.92
basketball	10	4	188	183	182	0.01	0.99	0.97
soccer	35	31	213	199	199	0	1	0.93
cricket	35	27	81	79	78	0.01	0.99	0.96
G. football	35	23	54	52	54	0.04	1.04	1
hurling	30	18	59	59	59	0	1	1

5. close up shot head (mixture background).

For most of the sports the limitation in the number of descriptors is barely noticeable, except rugby where the event classifier seems to be correlated with the long distance shot descriptor that presents spectators. Indeed, in the footage we covered almost every event is followed by the shot that presents cheering spectators. This feature makes the rugby footage very characteristic.

Since, as we mentioned the MA factor is vulnerable to the number of the events (i.e., the same accuracy can be achieved for different number of events in the game) we would also like to present the results for the remaining proposed factors. Tables 1-3 show that for all the classifiers presented in this paper the proposed event recognition method gives very good results (i.e., all the factors that stand for the broadly defined accuracy give almost ideal results). Note, that the precision factor is sometimes bigger than one. This is due to the fact, that the two events in the video footage are very close to each other and, since the width of the sliding window covers few seconds, were classified as one event. This is correct since all this events separated by the ground truth making users consist of the genuine event and its replay (especially in soccer).

As the subject of the event recognition and classification is very popular among the academia environment we'd like also to include comparison results

Tab. 2: Example numbers of true and detected events set together with MA, P and MP and corresponding window size and positive probes within

	MLP							
	W	WP	NE	DE	DTE	MA	P	MP
rugby	25	15	49	44	45	0.02	1.02	0.92
basketball	25	23	188	177	175	0.01	0.99	0.93
soccer	30	26	213	203	202	0.01	0.99	0.95
cricket	40	32	81	77	79	0.02	1.03	0.98
G. football	25	15	54	53	54	0.0185	1.02	1
hurling	20	10	59	71	59	0.20	0.83	1

Tab. 3: Example numbers of true and detected events set together with MA, P and MP and corresponding window size and positive probes within

	TREE							
	W	WP	NE	DE	DTE	MA	P	MP
rugby	50	24	49	47	47	0	1	0.96
basketball	40	28	188	137	152	0.08	1.12	0.81
soccer	35	29	213	133	134	0.01	1.01	0.63
cricket	50	24	81	80	79	0.01	0.99	0.98
G. football	50	46	54	17	15	0.0370	0.88	0.28
hurling	50	44	59	65	57	0.14	0.88	0.97

Tab. 4: Precision comparison with other works

	other works					
	[42]	[43]	[44]	[45]	[46]	[47]
this work **0.96**	0.81	0.83	0.51	0.62	0.93	0.74

between out work and the chosen works from around the world. Table 4 presents the mentioned comparison. Note, that in this section we proposed different than standard retrieval quality factors. The only standard one is the precision which we'll compare. Also, note that the table presents the mean values of the final precision results for all the sports presented in the respective paper.

5.3. Time performance

Since we claim that our system is capable of working in a real-time environment is is crucial to investigate also the time performance of all the classifiers we proposed in this paper. In general the complexity of all the solutions is linear $O(w)$, where w is the number of parameters. For the proposed decision tree classifier this investigation is really straightforward - the tree has at maximum eighteen decision levels. This kind of operation can be done in microseconds without any sophisticated implementations. For artificial neural networks used the overall cost/time can be calculated based on the equation (9).

$$T = cA + (n - n_i)G \tag{9}$$

Where c is the number of connections, n is the total number of neurons, n_i is the number of input and bias neurons, A is the cost of multiplying the weight with the input and adding it to the sum, G is the cost of the activation function and T is the total cost. Since in both proposed networks, neurons in the hidden layers have linear activation function and we have only one neuron on the output of the network $n_i = n - 1$. This reduces the (9) equation to:

$$T = cA + G$$

leaving the total cost depending only on the number of connections and parameters of the processor (i.e., clock frequency and number of clock cycles needed for multiplication and addition operations). In the case of feed-forward ANN and Elman ANN we have $c_{FF} = 180$ and $c_{Elm} = 800$ respectively. For modern processors, multiplication and addition operations are pipelined and do not take more than a few clock cycles. In our implementation the execution time of the Elman neural network was less than 1ms leaving plenty of time for the preceding scene analysis algorithm.

6. Conclusion

The paper presents a real-time event classification solution for broadcast sports videos. There are two main novelties presented: it is the first approach (to our knowledge) that explicitly deals with the problem of sports event classification from a real-time perspective; the range of the sports that can be annotated by the system is extremely broad.

Whilst the state of the art solutions very often obtain good accuracy they do not consider time performance as an important issue despite the fact that it could be highly desirable in a range of applications (e.g., in the scenario when this kind of system works on an embedded platform like a set-top box preparing feeds for the second screen application). For this reason, the classifiers we choose have linear transfer function in all neurons from the hidden layer allowing faster execution times (i.e., the cost related to calculation of the transfer function is eliminated). We have proved that our classification is not only as good as state of the art algorithms but also takes no longer than a few milliseconds to classify whether the particular part of a game could be interesting to the user. Having well designed ground truth dataset we can distinguish not only potentially interesting content but also classify it as a goal or highly interesting/exciting event. This enables placing specific markers in a video file in future applications. As previously mentioned, apart from the other presented algorithms, the system works not only for a particular type of the sport like soccer or basketball but can be thought of as a universal platform for so called field sports.

As it can be seen all the system is designed by be capable of working in real time. All the solutions assure the optimal flow of the information with regard to the processing time. This involves the use of parallel and pipelined processing which were extensively used in the project. A good example of this is an event classification system that is based on the sliding window. In other words it is just a shift buffer (pipeline) with a simple counter on the top of it. For this reason event classification engine was not designed as a standard classification vector machine like SVM, neural network or decision tree. The use of these techniques would require much more complex and slower solutions (e.g., in the case of a tree we would have 900 levels – 50 samples times two windows times nine scene descriptors). That's simply not realistic in a system that has to work under real-time regime.

To conclude the time performance of the event classification module, all the classification methods presented herein do not affect or do not redistribute in any way the main computational burden of the processing flow. That means that in comparison to the time performance of the feature extraction and scene classification engine the time needed for event classification can be in fact disregarded.

References

[1] The most watched TV shows of all time, http://www.dailymail.co.uk/tvshowbiz/article-1071394/The-watched-TV-shows-time–old-programmes.html, accessed: 22/08/2014.

[2] Broadcasting of sports events, http://www.sportsmediawatch.com/2013/01/2012-numbers-game-the-most-watched-sporting-events-of-the-year/, accessed: 22/08/2014.

[3] Crowds on the up as League gates top 16m, again, http://www.football-league.co.uk/footballleaguenews/20120517/crowds-on-the-up-as-league-gates-top-16m-again_2293334_2775283, accessed: 22/08/2014.

[4] Sport in Ireland, http://www.irishtimes.com/sport/gaelic-games, accessed: 22/08/2014.

[5] About the GAA, http://www.gaa.ie/about-the-gaa/publications-and-resources/, accessed: 22/08/2014.

[6] World's Most Popular Sports by Country , http://mostpopularsports.net/by-country, accessed: 22/08/2014.

[7] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, Q. Huang, Using webcast text for semantic event detection in broadcast sports video, Multimedia, IEEE Transactions on 10 (2008) 1342–1355.

[8] Y. peng Guan, J.-J. Li, Y. Ye, J. Si, H. Zhang, Content based sports video sequences analysis and synthesis, International Conference on Computer Science and Service System (2011) 2170–2172.

[9] C. Perin, R. Vuillemot, J.-D. Fekete, Soccerstories: A kick-off for visual soccer analysis, IEEE Transactions on Visualization and Computer Graphics 19 (12) (2013) 2506–2515.

[10] J. quan Ouyang, R. Liu, Ontology reasoning scheme for constructing meaningful sports video summarisation, Image Processing 7 (4) (2013) 324–334.

[11] F. Sanchez, M. Alduan, F. Alvarez, J. Menendez, O. Baez, Recommender system for sport videos based on user audiovisual consumption, Multimedia, IEEE Transactions on 14 (6) (2012) 1546–1557. doi:10.1109/TMM.2012.2217121.

[12] J. Han, D. Farin, P. H. de With, A mixed-reality system for broadcasting sports video to mobile devices, MultiMedia 18 (2) (2011) 72–84.

[13] D. A. Sadlier, N. E. O'Connor, Event detection in field sports video using audio-visual features and a support vector machine, in: IEEE Trans. Circuits Systems Video Technology, Vol. 15, 2005, pp. 1225–1233.

[14] C. Shu-ching, C. Min, Z. Chengcui, S. Mei-ling, Exciting Event Detection Using Multi-level Multimodal Descriptors and Data Classification (2006). doi:10.1109/ISM.2006.7.

[15] Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. G. Hauptmann, Multimedia event detection using a classifier-specific intermediate representation, IEEE Transactions on Multimedia 15 (7) (2013) 1628–1637.

[16] M. Tavassolipour, M. Karimian, S. Kasaei, Event detection and summarization in soccer videos using bayesian network and copula, IEEE Transactions on Circuits and Systems for Video Technology 24 (2) (2014) 291–304.

[17] Q. Y, Q. Huang, W. Gao, 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 Multimedia, MULTIMEDIA '05, ACM, New York, NY, USA, 2005, pp. 455–458. doi:10.1145/1101149.1101250.
URL http://doi.acm.org/10.1145/1101149.1101250

[18] B. Han, Y. Hu, G. Wang, W. Wu, T. Yoshigahara, Enhanced sports video shot boundary detection based on middle level features and a unified model, IEEE Transactions on Consumer Electronics 53 (3) (2007) 1168–1176.

[19] Y.-G. Jiang, SUPER: Towards Real-time Event Recognition in Internet Videos, in: Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval, ICMR '12, ACM, New York, NY, USA, 2012, pp. 7:1–7:8. doi:10.1145/2324796.2324805.
URL http://doi.acm.org/10.1145/2324796.2324805

[20] J. Uijlings, I. Duta, E. Sangineto, N. Sebe, Video classification with Densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off, International Journal of Multimedia Information Retrieval (2014) 1–12doi:10.1007/s13735-014-0069-5.
URL http://dx.doi.org/10.1007/s13735-014-0069-5

[21] J. Lanagan, F. A. Smeaton, Using Twitter to detect and tag important events in live sports, in: Proceedings of the fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 542–545.

[22] J. Nichols, J. Mahmud, C. Drews, Summarizing sporting events using twitter, in: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, IUI '12, ACM, New York, NY, USA, 2012, pp. 189–198. doi:10.1145/2166966.2166999.
URL http://doi.acm.org/10.1145/2166966.2166999

[23] D. Tjondronegoro, Y. Chen, Knowledge-discounted event detection in sports video, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 40 (5) (2010) 1009–1024. doi:10.1109/TSMCA.2010.2046729.

[24] H. Kim, S. Roeber, A. Samour, T. Sikora, Detection of goal events in soccer videos, in: Proceedings of SPIE, Vol. 5682, 2005, p. 317.

[25] R. Kapela, K. McGuinness, N. E. O'Connor, Real-time field sports scene classification using colour and frequency space decompositions, Journal of Real-Time Image Processing.

[26] D. Lowe, Object recognition from local scale-invariant features, IEEE International Conference on Computer Vision (1999) 1150–1157.

[27] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), Vol. 1, 2005, pp. 886–893.

[28] E. Rosten, T. Drummond, Machine Learning for High-Speed Corner Detection, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision ? ECCV 2006, Vol. 3951 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, pp. 430–443. doi:10.1007/11744023_34.
URL http://dx.doi.org/10.1007/11744023_34

[29] M. Calonder, V. Lepetit, C. Strecha, P. Fua, BRIEF: Binary Robust Independent Elementary Features, in: Lecture Notes in Computer Science, 2010, pp. 778–792.

[30] A. Alahi, R. Ortiz, P. Vandergheynst, FREAK: Fast Retina Keypoint, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[31] S. Leutenegger, M. Chli, R. Siegwart, BRISK: Binary Robust Invariant Scalable Keypoints, in: Computer Vision (ICCV), 2011 IEEE International Conference on, 2011, pp. 2548–2555. `doi:10.1109/ICCV.2011.6126542`.

[32] Open Computer Vision Library, www.opencv.org, accessed: 22/08/2014.

[33] Y. Ro, M. Kim, H. Kang, B. Manjunath, J. Kim, Mpeg-7 homogeneous texture descriptor, ETRI journal 23 (2) (2001) 41–51.

[34] L. Franco, Generalization ability of boolean functions implemented in feedforward neural networks, in: Neurocomputing, Vol. 70, 2006, pp. 351–361. `doi:10.1016/j.neucom.2006.01.025`.

[35] K. Levenberg, A method for the solution of certain problems in least squares, Quarterly of Applied Mathematics 5 (1944) 164–168.

[36] D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, SIAM Journal on Applied Mathematics 11 (2) (1963) 431441.

[37] M. Roisenberg, J. Barreto, F. D. Azevedo, Neural network complexity classification based on the problem, in: Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, Vol. 3, 1998, pp. 2413–2418 vol.3. `doi:10.1109/IJCNN.1998.687240`.

[38] B. Hammer, Generalization of elman networks, in: Artificial Neural Networks – ICANN'97, Springer, 1997, pp. 409–414.

[39] S. Mertens, A. Engel, Vapnik-Chervonenkis dimension of neural networks with binary weights, Phys. Rev. E 55 (1997) 4478–4488. `doi:10.1103/`

PhysRevE.55.4478.
URL http://link.aps.org/doi/10.1103/PhysRevE.55.4478

[40] P. Orponen, Computational complexity of neural networks: a survey, Nordic J. of Computing 1 (1) (1994) 94–110.
URL http://dl.acm.org/citation.cfm?id=640186.640192

[41] D. Coppersmith, S. J. Hong, J. R. M. Hosking, Partitioning nominal attributes in decision trees, Data Mining and Knowledge Discovery 3 (1999) 197–217.

[42] J. Wang, C. Xu, E. Chng, X. Yu, Q. Tian, Event detection based on non-broadcast sports video, in: Image Processing, 2004. ICIP '04. 2004 International Conference on, Vol. 3, 2004, pp. 1637–1640 Vol. 3. doi:10.1109/ICIP.2004.1421383.

[43] S. Miyauchi, A. Hirano, N. Babaguchi, T. Kitahashi, Collaborative multimedia analysis for detecting semantical events from broadcasted sports video, in: Pattern Recognition, 2002. Proceedings. 16th International Conference on, Vol. 2, 2002, pp. 1009–1012 vol.2. doi:10.1109/ICPR.2002.1048476.

[44] H.-S. Chen, W.-J. Tsai, A framework for video event classification by modeling temporal context of multimodal features using {HMM}, Journal of Visual Communication and Image Representation 25 (2) (2014) 285 – 295. doi:http://dx.doi.org/10.1016/j.jvcir.2013.12.001.
URL http://www.sciencedirect.com/science/article/pii/S1047320313002150

[45] D. Tjondronegoro, Y.-P. Chen, Using decision-tree to automatically construct learned-heuristics for events classification in sports video, in: Multimedia and Expo, 2006 IEEE International Conference on, 2006, pp. 1465–1468. doi:10.1109/ICME.2006.262818.

[46] L. Hua-Yong, H. Tingting, Z. Hui, Event detection in sports video based on multiple feature fusion, in: Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on, Vol. 2, 2007, pp. 446–450. doi:10.1109/FSKD.2007.278.

[47] N. Babaguchi, Y. Kawai, T. Kitahashi, Event based indexing of broadcasted sports video by intermodal collaboration, Multimedia, IEEE Transactions on 4 (1) (2002) 68–75. doi:10.1109/6046.985555.