Accepted in IEEE Int. Symposium on Circuits and Systems (ISCAS), 2015

Real-Time Visual Play-Break Detection in Sport Events using Context

Marc-André Carbonneau,
Alexandre J. Raymond and Ghyslain Gagnon
Département de génie électrique
École de technologie supérieure
Montréal, Canada
marcandre.carbonneau@gmail.com

Eric Granger
Département de génie de la production automatisée
École de technologie supérieure
Montréal, Canada
eric.granger@etsmtl.ca

***Abstract*—This paper presents a two-stage hierarchical method for play-break detection on non-edited team sports video feed. Unlike most existing methods, this algorithm uses modern action and event recognition method thus does not rely on production cues of broadcast feeds, but instead concentrates on the content of the video. Moreover, the method does not require player tracking, can be used in real-time and can be easily adapted to different sports. In the first stage, bag-of-words event detectors are trained to recognize key events such as line changes, face-offs and preliminary play-breaks. In the second stage, the output of the detectors along with a novel feature based on the number of detected spatio-temporal interest points are used to create a context descriptor. The final classification is performed on this context descriptor. Experiments demonstrate the benefits of using this context descriptor by reducing the frame classification error by 18% when compared to the baseline method. The efficiency of the proposed method is demonstrated on a real hockey game (accuracy over 88%).**

***Index Terms*—event detection; bag-of-words; play-break classification; event modelling.**

I. Introduction

Automatic video summarization is of great importance in a world producing an ever increasing quantity of visual data. For instance, Cisco forecasts that in 2018, a million minutes of video content will be transferred over the Internet every second [1]. Sport events attract large audiences and is therefore a non-negligible video category. In sport events, some sequences are less pertinent and do not catch the interest of the viewer (e.g. time-outs). When live broadcasting such events, it would make sense to detect these less interesting sequences to adjust the compression rate of the broadcast feed, replace them with advertisement or players information. Play and break detection have to be performed to achieve that goal. The detection of these events is also primordial to perform automatic editing and summarization of sporting videos.

Over the years, a lot of approaches have been proposed to tackle the play break and event detection problem in sport. However, to our knowledge none of these may be applied to unedited footage from a fixed camera because they rely on production cues. For instance [2] used production cues such as replay and close-up sequences. In [3], Wang and Zhang proposed a method to recognize shooting events in ice hockey. The brightness of the frames was used as a feature to distinguish between close-ups and global camera views. Ekin [4] also used the type of point of views in the frame as features. Qian [5] used overlaid text amongst other features. Assfalg [6] presented a method to detect important moments in a soccer game. This method uses unedited video streams from a mobile camera. The position of the ball is inferred based on the camera motion, and the part of the field covered in the frame is used as a feature. This means that the cameraman has done most of the tracking and pattern recognition job.

Recent advances in action and event recognition have made it possible to work directly on the content of the video feed instead of its editing style as the existing methods do. The method proposed in this paper makes use of modern action recognition methods to detect on-the-fly play and break segments in an unedited video feed captured by a fixed camera. Moreover, it can run in real-time, and does not need segmentation or tracking of the players. Also, many existing methods rely on rules or expert knowledge specific to the sport of interest [6]–[8], which makes them inapplicable to other sports. The proposed method could easily be adapted to other sports.

The contributions of this paper are threefold. A new method for play-break segmentation is introduced. A framework is proposed to adapt the standard bag-of-words (BOW) classification pipeline to event detection. A new context descriptor based on the output of these detectors as well as spatio-temporal interest points (STIP) detection number is introduced.

The complete system is validated on a newly introduced dataset consisting of a complete hockey game. The method is compared to a baseline method widely used in event detection [9], [10].

II. Event Detectors

The proposed method adapts the generic bag-of-words classification framework to event detection. An overview of the detector is presented in Fig. 1. The generic framework for action recognition [10], [11] is used to classify complete sequences as one of some predefined classes. The proposed adapted framework enables the detection of events on a live feed even if they are concurrent.

First, the incoming frames are grouped in video slices. STIPs are then detected and extracted. Then, principal component analysis (PCA) and whitening are applied to the STIP feature vectors. Finally, histograms are produced and detection is performed on the slices.

A. Video Slicing

To achieve temporal localization of the events, the video sequence is divided into smaller sub-sequences called slices, using an overlapping sliding window. Each of these slices is classified separately and a likelihood score is produced for each of them. The step size between each window slice determines the granularity of the detection as well as the latency of the system when used in live feed contexts.

B. Feature Extraction

To limit the amount of data to be processed, STIPs are detected and extracted. The detection is achieved using a 3D adaptation of Harris corners [12] introduced by Laptev in [13]. Each STIP is characterized by a combination of histograms of oriented gradients (HOG) and optical flow (HOF). This descriptor has been proved to be a reliable choice for action recognition [10] because of its capacity to represent shape and motion. The STIPs are detected and extracted at different scales to compensate for perspective effects in the images captured by a far-field camera. Wang's implementation of Laptev's algorithm [10] was used in the following experiments.

Applying PCA reduction and whitening to feature vectors has been proved to boost classification performance in action recognition problems [14]. Whitening is applied after the PCA projection and dimensionality reduction.

C. Code-Word Association

In order to create a code-word dictionary, STIPs are randomly sampled from the complete training sequences. Samples taken from sequences containing the events to be recognized are also added to ensure these events appropriately are represented. If the STIPs were only sampled uniformly in the video sequences, there would be a risk of creating a dictionary lacking examples from rare classes. Once samples are collected, k-means clustering is performed in order to create N code-word prototypes. At runtime, every STIP feature vector is quantized to the nearest of these N prototypes using the Euclidean norm.

D. Detection

For each video slice, the code-words associated with the detected STIPs are pooled in a frequency histogram. This histogram represents the content of the slice. For every event detected, a likelihood score is obtained using the output of a support vector regression. The normalized χ^2 kernel is used:

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \frac{x_i y_i}{x_i + y_i}, \tag{1}$$

where x_i and y_i are the i^{th} bins of histograms $\mathbf{x}$ and $\mathbf{y}$. As mentioned earlier, N is the number of words in the dictionary.

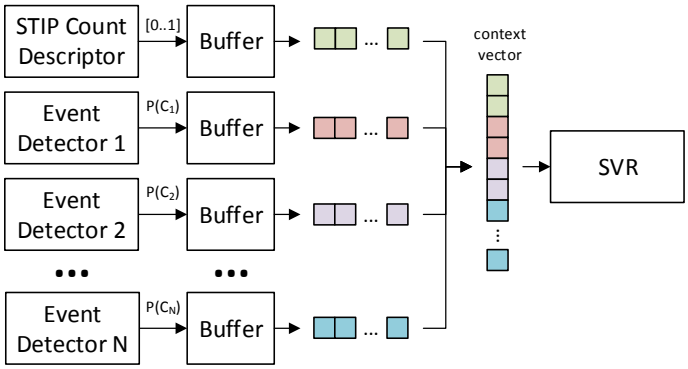

Fig. 2. Context vector construction.

III. Context Analysis

A. Context Descriptor

In order to improve the performance of the play-break recognition, a context descriptor, shown in Fig. 2, is introduced. This descriptor is constructed using the likelihood scores from the event detectors described in Section II. In the following experiments, three detectors have been trained to recognize *face-off*, *line change* and *play* sequences. The intuition behind the context descriptor is that certain events tend to precede or to indicate a play or a break sequence. For instance, knowing that a *face-off* just finished, there are greater chances that the present segment is a *play* segment. Also, if a long *line change* event is occurring, the game is probably in a *break*.

The context descriptor at time t is given by:

$$\mathbf{c}_t = \{\mathbf{f}_t, \mathbf{l}_t, \mathbf{p}_t, \mathbf{s}_t\}, \tag{2}$$

where the *face-off* descriptor $\mathbf{f}_t$ is given by:

$$\mathbf{f}_t = \{\theta_t, \theta_{t-1}, \ldots, \theta_{t-T}\}, \tag{3}$$

where T is the number of slices contained in the context window and θ_t is the event detector output at time t. The *play* and *line change* descriptors $(\mathbf{l}_t, \mathbf{p}_t)$ are constructed in a similar manner. Along with the detector outputs, a descriptor $\mathbf{s}_t$ based on the number of STIPs in a slice is also used. The elements of this vector are given by:

$$s = \begin{cases} M/\beta & \text{if } M < \beta; \\ 1 & \text{if } M \geq \beta, \end{cases} \tag{4}$$

where M is the number of STIPs detected in a slice and β is a threshold that has to be set empirically. Each time a slice is produced, a new context vector is computed. The context vector is then classified as *play* or *break* using a support vector regression (SVR) with a radial basis function (RBF) kernel.

IV. Experimental Methodology

A. Datasets

As no existing datasets met the requirements of our problem, a new one was created and has been made publicly available

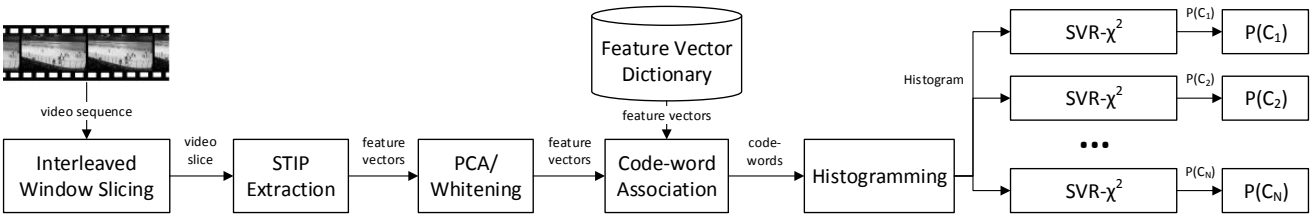

Fig. 1. Schematic block overview of the event detectors.

Fig. 3. Examples taken from the dataset.

TABLE I
PARAMETERS USED FOR EACH EXPERIMENT REPETITION.

	SVR		Context Descriptor	
	γ	C	T	β
Period 1	0.001	1000	15	100

on-line. The dataset consists of a complete university-level hockey game captured from two far-field views of the ice rink. Fig. 3 shows images taken from each camera. The video sequence from each point of view is treated as a different training instance since their appearance differs considerably. The images are in grayscale with a 480x270 pixels resolution at 30 frame per second (fps). For our application, along with play-break classification, 3 types of events are identified:

Play: A sequence is labelled as a play sequence when at least one player is visible and it is possible for a human to determine if the other players are actively playing.

Face-Off: A face-off event starts when the players are converging to their respective positions, waiting for the puck drop. It stops when the puck has been released and the players start to skate away. Some more difficult examples include images taken from afar where only two or three players are visible. The events are 2 to 12 seconds long and often contain line changes.

Line Change: A line change event usually happens during a break, but may also occur during playtime. The dataset contains both situations. The event is characterized by players coming from and going to the player bench. The event durations vary from 2 to 19 seconds.

B. Protocol

A hockey game is divided in three periods. In the dataset, the game was captured from two angles which makes 6 video sequences. The video sequences are further partitioned in six parts making 36 sub-sequences.

The experiment results are obtained using 6-fold cross-validation. For each fold, a part from each video sequence is reserved for testing. The sliding window contains 45 frames (1.5 sec) and a new one is produced each time 15 frames are produced. The dictionary contains 400 code-words. This has been arbitrarily selected based on earlier works on homogeneous datasets such as [11]. The PCA stage is set to keep 97% of the signal energy, which corresponds to 83 components out of 162.

The SVRs of the event detectors are trained on all positive examples from the training set. An equivalent number of negative examples are sampled randomly. The SVR regularization parameter C is obtained by grid searching using 8-fold cross-validation on the training set. The configuration yielding the best accuracy was retained ($C = 0.01$).

The parameters for the context descriptor and the final SVR (C and γ) classifier were determined using 8-fold cross-validation on the training set. The parameters are summarized in Table I.

The baseline method used for comparison is very similar to the one used by the authors of [9]. This method is the usual BoW pipeline [10] where the STIPs are detected with 3-D Harris corners and HOG/HOF is used as descriptor. The classification of the histograms is performed by a SVM. However, a PCA and whitening stage has been added. Also, the number of detected STIPs was not limited. These modifications have been made to make sure the comparison with our method was fair and not due to whitening or STIP count.

V. RESULTS

Fig. 4 shows the receiver operating characteristic (ROC) and precision-recall (P-R) curves obtained using the proposed and reference methods for one run of the experiment.

The result variations from one fold to another are explained by the nature of the data. For instance, some parts of the game contain less occurrences of some events than others, which might result in folds containing more training instances than another. Also, in this particular game, one team was dominating the other. Because of this, a bigger part of the action occurred in one zone thus creating imbalance in the number of event instances captured by the two camera angles.

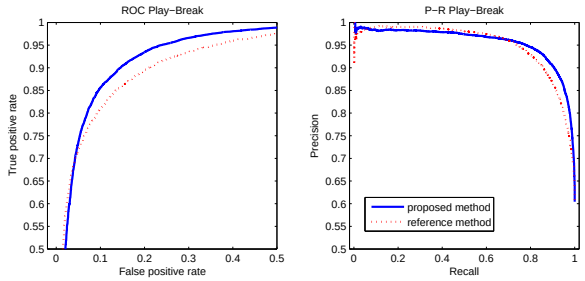

Fig. 4. Precision-recall and ROC curves for play sequence detection.

TABLE II
RESULTS OBTAINED ON THE DATASET.

	AUC ROC		Accuracy	
Algorithm	mean	st. dev.	mean	st. dev
Baseline method [9]	XX.XX	± X.XX		
Proposed method	**XX.XX**	± **X.XX**		

To further assess the benefits of using the context descriptor stage, the accuracy was obtained from 10 runs of the algorithm on the entire dataset. Several replications of the experiment are needed since the recognition results depend on the quality of the dictionary. To create the dictionary, STIPs are randomly sampled and the seeds of the *k-means* are also randomly selected. The area under the ROC curve (AUC) of each run is averaged and presented in Table II. An average performance boost of 0.014±0.009 can be observed, which confirms the benefits of considering the temporal context in play-break classification. When using the optimal threshold, the accuracy rises by 2.7±1.32% which translate in a 18.07±9% error reduction.

Most of the misclassified video slices are situated at the start and end of a play event. This means that the proposed algorithm often disagree for 15 frames (500 ms) with the manually obtained labels. If the slice right before or after a play event is considered as a reasonable margin of error, the AUC rises by X.XX ± X.XX%. Let it be said that even for a human annotator, it is difficult to determine the exact duration of a play sequence, especially in the frequent situation where only one player is visible.

It is possible to assess the usability of the method in real-time settings by measuring the processing time. Using a single processor, the algorithm detects STIPs, extracts the HOG/HOF features and saves them to a file at an average rate of 5 fps. Since image processing is highly parallelizable, one could expect to attain a frame rate greater than 30 fps using 8 cores. Once the STIPs are extracted, the analysis of a complete 20 minute period captured from 2 angles takes under 150 seconds using a MATLAB implementation. In lights of these results, meeting real-time requirements should not be a problem.

VI. CONCLUSION

In this paper, we presented an efficient method for play-break detection. Unlike previous efforts in the field, our method does not require an edited video sequence or camera tracking of the action. Moreover, the method can be implemented in real-time, enabling its integration in automated capture systems. Experiments demonstrated the applicability of the algorithm to a real-life setting. The use of the temporal context information proved to be beneficial to play-break segment recognition.

More experiments are needed in order to assess the suitability of this method to other sports and venues. Also the detection of other types of event should be explored to further increase the performance of the method.

ACKNOWLEDGMENT

The work is supported by Quattriumm Inc. and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index : Forecast and Methodology , 2013 – 2018," Tech. Rep., 2014.

[2] D. W. Tjondronegoro and Y.-P. Chen, "Knowledge-Discounted Event Detection in Sports Video," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Trans. on*, vol. 40, no. 5, pp. 1009–1024, Sep. 2010.

[3] X. Wang and X.-P. Zhang, "Ice Hockey Shooting Event Modeling with Mixture Hidden Markov Model," *Multimedia Tools and Applications*, vol. 57, no. 1, pp. 131–144, 2012.

[4] A. Ekin and M. Tekalp, "Generic Play-break Event Detection for Summarization and Hierarchical Sports Video Analysis," in *Multimedia and Expo, 2003. ICME '03. Proc.. 2003 Int. Conf. on*, vol. 1, Jul. 2003, pp. 169–72.

[5] X. Qian, G. Liu, H. Wang, Z. Li, and Z. Wang, "Soccer Video Event Detection by Fusing Middle Level Visual Semantics of an Event Clip," in *Advances in Multimedia Information Processing*. Springer Berlin Heidelberg, 2011, vol. 6298, pp. 439–451.

[6] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati, "Semantic Annotation of Soccer Videos: Automatic Highlights Identification," *Comput. Vis. Image Underst.*, vol. 92, no. 2-3, pp. 285–305, Nov. 2003.

[7] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A Decision Tree-Based Multimodal Data Mining Framework for Soccer Goal Detection," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE Int. Conf. on*, vol. 1, 2004, pp. 265–268 Vol.1.

[8] Y. Ariki, S. Kubota, and M. Kumano, "Automatic Production System of Soccer Sports Video by Digital Camera Work Based on Situation Recognition," in *Multimedia, 2006. ISM'06. Eighth IEEE Int. Symp. on*, 2006, pp. 851–860.

[9] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A Large-Scale Benchmark Dataset for Event Recognition in Surveillance Video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conf. on*, Jun. 2011, pp. 3153–3160.

[10] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," in *BMVC 2009 - British Machine Vision Conf.*, A. Cavallaro, S. Prince, and D. Alexander, Eds. London, UK: BMVA Press, Sep. 2009, pp. 124.1–124.11.

[11] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-temporal Features," in *Proc. of the 14th Int. Conf. on Computer Communications and Networks*, ser. ICCCN '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 65–72.

[12] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *In Proc. of Fourth Alvey Vision Conf.*, 1988, pp. 147–151.

[13] I. Laptev, "On Space-Time Interest Points," *Int. J. Comput. Vision*, vol. 64, no. 2-3, pp. 107–123, Sep. 2005.

[14] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice," *arXiv preprint arXiv:1405.4506*, 2014.