Audio-visual classification of sports types

Rikke Gade[1], Mohamed Abou-Zleikha[2], Mads Græsbøll Christensen[2] and Thomas B. Moeslund[1]
[1]Visual Analysis of People Lab, Aalborg University, Denmark
[2]Audio Analysis Lab, Aalborg University, Denmark
{rg,moa,mgc,tbm}@create.aau.dk

Abstract

In this work we propose a method for classification of sports types from combined audio and visual features extracted from thermal video. From audio Mel Frequency Cepstral Coefficients (MFCC) are extracted, and PCA are applied to reduce the feature space to 10 dimensions. From the visual modality short trajectories are constructed to represent the motion of players. From these, four motion features are extracted and combined directly with audio features for classification. A k-nearest neighbour classifier is applied for classification of 180 1-minute video sequences from three sports types. Using 10-fold cross validation a correct classification rate of 96.11% is obtained with multimodal features, compared to 86.67% and 90.00% using only visual or audio features, respectively.

1. Introduction

Automatic analysis of sport is a large research area which has expanded lately, mainly due to the commercial interests in sports. Huge amounts of sports video are captured every day, for TV productions of popular sports leagues, for internal post-game analysis of performance and tactics or just for personal memories and sharing with friends.

Manual annotation of activities, tracks and events in huge amounts of videos are tedious, if not impossible. Automatised methods are therefore important if useful information and categorization should be extracted from these videos. The first step toward understanding the activities is often the recognition of sports types. This is also the focus of this paper with the goal of labeling large amounts of video based on the recognised sports types.

Due to the application of the system in public sports arenas, we choose to use privacy-preserving thermal video rather than traditional RGB video. However, the general methods proposed in this paper are not restricted to thermal imagery and could be applied on RGB video as well.

Previously we have worked towards sports type classification based only on the visual output of thermal cameras. One proposed method relied only on the detected positions of people [6] and another method was based on motion features extracted from tracklets [5]. Both methods proved good classification results. However, limitations exist when using only one modality. Visual features rely on the visibility of people, which can be obstructed by, e.g., occlusions. Fusing the information obtained in different modalities might add information able to solve ambiguous situations. For this reason, we will in this work investigate the effect of including audio features for classification of sports types.

1.1. Related work

Automatic event and highlight detection is a popular research topic used for several sports types. The applications of these methods are, e.g., replays during games and short summaries for news channels.

One example of this is analysis of TV productions from soccer matches. The survey by Oskouie et al. [14] discusses the topic of multimodal feature extraction and fusion for semantic mining of soccer video, using both visual, audio, and text features. The main events of interest for a summary here are goal, penalty, booking, shot on target and offside situation. In such applications, several manually operated cameras are often employed, and the chosen camera view can be used as a high level feature for recognising events. Furthermore, audio cues like applause of the spectators and sport commentator excitement are indicators of highlights. Likewise, speech recognition can be applied to detect keywords. Automatic highlight extraction can also be based on audio only, as proposed for baseball [15] and golf [21].

When broadcasting from sports events the focus will often switch between play, advertisements and studio sequences. Bai et al. [3] proposed a framework for classification of these segments based on audio. Subashini et al. [17] combined MFCC features from audio with colour histograms as visual features for classifying the categories of news, advertisement, sport, serial and movie. Using broadcast videos Xu et al. [22] proposed a multi-layer framework

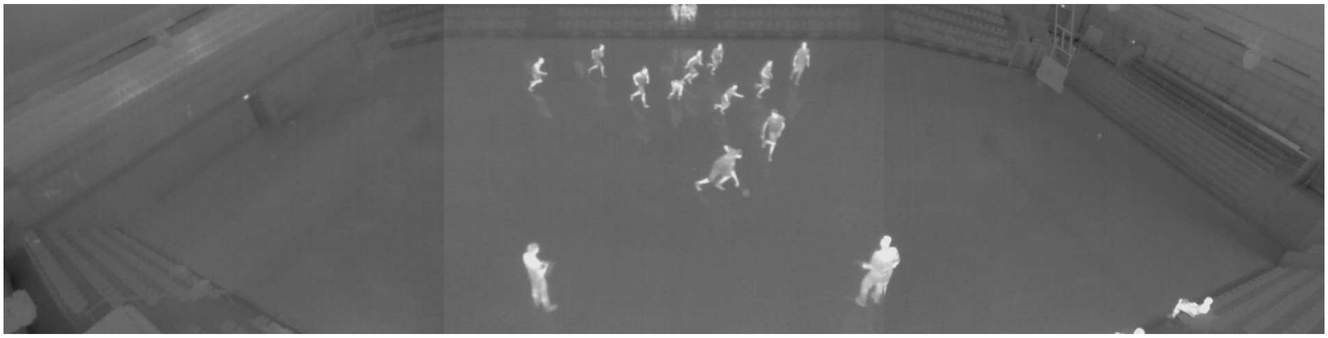


Figure 1. Example of input image.

combining visual features, audio features, and text features. The resulting high-level semantics can be used for event and highlight detection, video editing, and tactical analysis.

For sports type classification most work is based on visual features. These features can represent the specific court, such as dominating colours of the image [13], which can also be combined with motion features [7, 18], or combined with dominant grey level, cut rate, and motion rate [16]. Also relying on the appearance of the court are methods based on edge direction, intensity, or ratio [10, 23]. More generic approaches have also been proposed, such as SURF features, which were tested on 14 different sports types [12]. Wilson et al. [20] proposed a Hidden Markov Model framework based on motion features and recognition of core events. Based only on the detection and tracking of people, Lee and Hoff [11] classified individual trajectories from two different sports types.

Wang et al. [19] combined low-level visual and audio features for classification of sports types. Two cinematic features were chosen; camera motion, and dominant colours. For audio features MFCC were extracted and Hidden Markov Models were applied for classification of three sports types plus a news category.

The work presented in this paper is intended for application in multi-purpose indoor arenas during everyday activities as well as match days. Therefore, rather than relying on professionally produced broadcast videos, this work will focus on video captured by static thermal cameras. Hence, the camera view will not contain any information, no camera motion will be observed, and features representing the appearance of the court are not related to the specific activity. Furthermore, as the activities can be captured at training sessions as well as games, it is assumed that the audio is caused by players rather than spectators.

2. Data aquisition

The observed multi-purpose arena has a standard court size of 20×40 meters, which is sufficient for handball, indoor soccer, basketball, volleyball, and badminton, as well as a large number of other sports types which are not restricted to specific court dimensions.

For video capturing we have chosen to use thermal cameras due to privacy issues in the public arenas. However, this technology still have some limitations in terms of resolution and field-of-view. As a compromise between wide field-of-view and sufficient resolution we use cameras of type AXIS Q1922, which has a resolution of 640×480 pixels and a field-of-view of 57°. In order to capture the full court area three cameras are combined as illustrated in figure 2. The cameras are manually adjusted to have adjacent field-of-views after rectification of the images, which allows for direct stitching of the images for each frame. An example of the resulting image is shown in figure 1. An initial calibration of the camera system is performed to calculate a mapping between image and world coordinates, following the procedure described in [6].

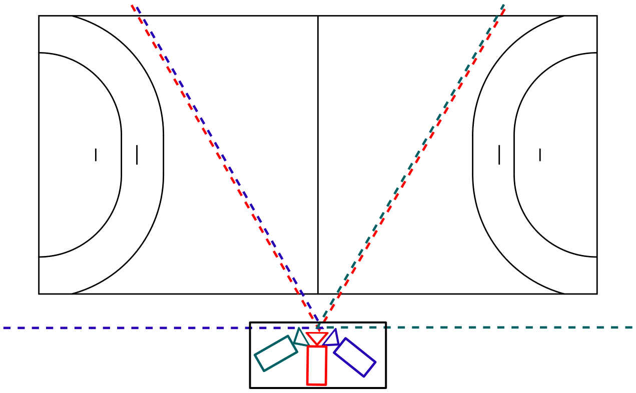


Figure 2. Ilustration of the camera views.

The cameras are mounted in one box at a height of approx. 10 meters at the centre of the longest edge of the court, as illustrated in figure 2. Audio is included in the video of the centre camera, thus synchronised in time with the visual input.

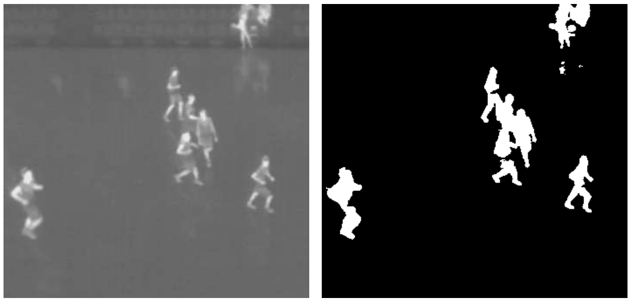
3. Visual feature extraction

For use in multi-purpose arenas, the visual features must be extracted from activities performed by athletes rather than information about the surroundings. Therefore, people must be automatically detected in each frame. After that, tracking of each individual will provide information about their motion and locations, which can finally be used for extracting relevant motion features. The following subsections will discuss each of these steps.

3.1. Detection

Detection of people using computer vision is generally a difficult task, due to the large variations in appearance and pose. The use of thermal imaging reduces the problem of appearance changes, as the images represent only temperature, which varies very little between people. Furthermore, the human temperature is most often different from the surrounding temperatures. Foreground extraction is thereby possible by simple thresholding of pixel intensities. However, since the cameras have automatic gain adjustment, the level of pixel values changes dynamically, and so must the threshold value. An automatic threshold method based on the entropy of the image is therefore applied [9].

The result of thresholding is a binary image which, ideally, represents humans in white and anything else in black. However, the mapping between white objects and positions of individual people are not always trivial. Non-human warm objects might be observed and represent noise, and the separation of people can be difficult in cases of occlusion. Figure 3 shows an example of the thermal and resulting binary image in a case of occlusions between people.



(a) (b)

Figure 3. Example of thresholding in cases of occlusion between people. (a) thermal input, (b) binary output.

With the purpose of having each white object representing a single person, the binary objects are processed through split and merge procedures as described in [4].

3.2. Tracking

After detecting people, we need to construct temporal trajectories to analyse the motion. Tracking individual people through heavy occlusions and erratic motion, which are often seen in sports, is very challenging and no robust solutions exist yet. Re-identification of targets after occlusions are necessary to produce long trajectories, but due to the sparse appearance information in thermal images, the solutions for re-identification here are very limited. The identity related to each trajectory is not of interest in this work, so instead we aim for constructing short trajectories, also called tracklets, with reliable motion information.

We use the well-known approach of the Kalman filter [8]. This method is one of the predict-match-update schemes, which predicts the next position of the object from the previous state (described by, e.g., position and velocity), then updates the estimate when a (probably noisy) measurement is obtained. Using Kalman filtering for multi-target tracking can be done by assigning a new Kalman filter for each new target, however, it implies some reasoning for assigning each detection to the right tracker. This is here determined by the shortest Euclidean distance within a given threshold. If a detection is not assigned to a tracker, a new Kalman filter is started. Likewise, if no detections are assigned to a tracker in n consecutive frames, the Kalman filter is terminated. n is experimentally set to 10 frames.

To be independent of the image perspective we transform the detected positions of people in the image into world coordinates before tracking. This is done by applying a homography matrix, calculated during initialisation. Terminating the tracks with no possibility of re-identification later will naturally lead to more split trajectories. But for the purposes of this work it is preferable to have short reliable tracklets instead of trying to resolve complex situations with a higher probability of false tracks.

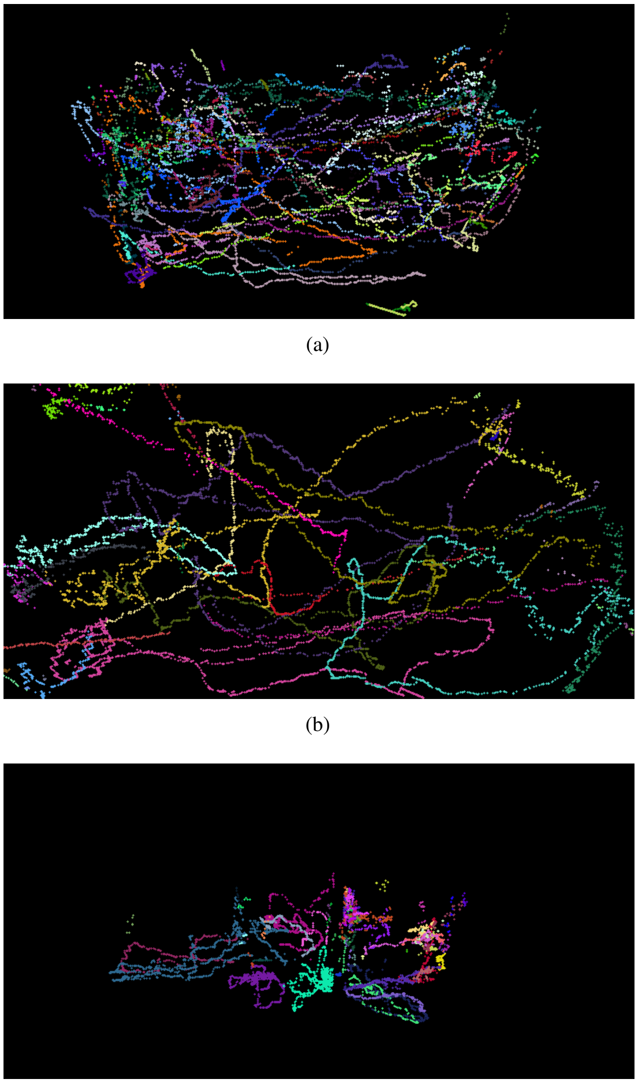
Figure 4 presents visualisations of typical tracklets during one minute of basketball, soccer, and volleyball. Each tracklet is assigned a random colour.

3.3. Motion features

The tracklets contain information on the observed movements. To be able to classify sports types, we aim to extract features with different characteristics for each activity.

For constructing a robust system a number of requirements are considered when choosing the features: The features should be invariant to the size and direction of the court, the position of players on the court, and the direction of play. Furthermore, the features must be robust to noisy detections and tracking errors. From these criteria we choose the following four features:

Lifespan [frames] is measured in number of frames before the tracklet is terminated. This feature implicitly represents the complexity of the sequence; the lifespan of each tracklet will be shorter when the scene is highly



(a)

(b)

(c)

Figure 4. Tracklets presented in a top-down view of the world plane from a typical 1-minute period of (a) basketball, (b) soccer, and (c) volleyball.

occluded:

$$ls = n_{end} - n_{start} \tag{1}$$

where n is the frame number.

Total distance [m] represents the total distance travelled, measured as the sum of frame-to-frame distances in world coordinates:

$$td = \sum_{i=0}^{ls-1} d(i, i+1) \tag{2}$$

where d is the Euclidean distance function.

Distance span [m] is measured as the maximum distance between any two points of the trajectory. This feature is a measure of how far the player move around at the court:

$$ds = \max(d(i,j)), \quad 0 < i < ls, \; 0 < j < ls \tag{3}$$

Mean speed [m/s] is measured as a mean value during the lifespan of the tracklet:

$$ms = \frac{td \cdot n_{seq}}{ls \cdot t} \tag{4}$$

where t is the duration of the video sequence in seconds, and n_{seq} the duration of the sequence in number of frames.

Each feature is calculated for each of the unknown number of tracklets produced for each video sequence. The mean value of all tracklets is then calculated and used as the feature for the given video sequence.

We test all combinations of the features described above, from using a single feature to using all four. We find that the best results are obtained when using all four features, indicating that none of them are redundant or misleading. The features are combined with equal weighting.

4. Audio Feature Extraction

In order to classify the sports type using the audio information, a perceptual time-frequency representation of audio is used. This representation is the Mel Frequency Cepstral Coefficients (MFCC). MFCC features are considered one of the main features those are used for audio signal processing applications such as speech and speaker recognition, and emotion recognition in music and speech.

To extract the MFCC features, the signal is pre-emphasised to give the audio signal an energy boost in the high frequencies, since the spectrum in lower frequencies usually contains more energy than the high frequency ones. Then the signal is divided into successive overlapping frames. For each frame, the short-time Fourier transform is calculated. Since the human perception of frequencies is logarithmic, a mel-filter banking and a mapping to the logarithmic scale are performed. Finally, an inverse discrete Fourier transform is applied to get the the MFCC features. Figure 5 illustrates the procedure of audio feature extraction.

The HTK toolbox [1] is used to extract the features. The size of the analysis window is 25 ms and the overlap window size is 10 ms. 25 MFCC features in addition to the log of the energy is extracted. To catch the dynamic characteristics of the audio signal, the first and second derivatives are also computed for the extracted features. As a result, a 78*600 feature matrix per sample is extracted. These features are summarised by estimating the mean $(m_1, ..., m_{78})$

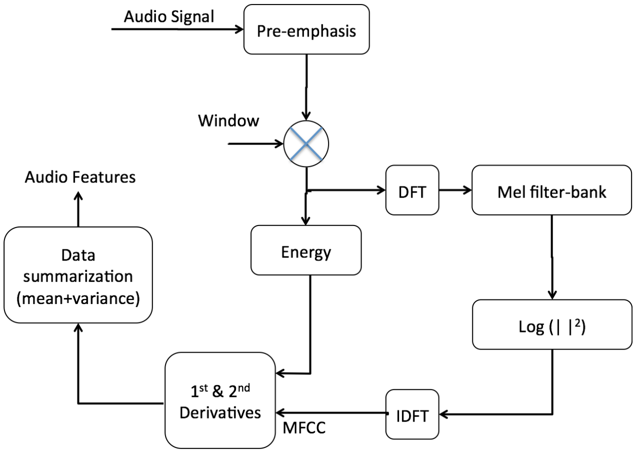


Figure 5. Procedure for audio feature extraction.

and the variance $(c_1, ..., c_{78})$ in the matrix. The result is a 78*2 feature matrix which is then converted to a vector of 156 features by applying a vectorisation operation. Finally, PCA is performed to reduce the feature space from 156 to 10 features.

5. Experiments

For experiments we captured video of three different sports types from one multi-purpose indoor arena. By capturing all data from the same arena, visual and audio features related to the specific arena are constant and classification can only be based on features related to the observed activity.

One continuous hour of video, including audio, is extracted from each of the three sports types; Basketball, soccer, and volleyball. The videos are then divided into 1-minute sequences, which results in a total of 180 sequences to classify. During the one hour of video for each sport no manual selection has been performed, meaning that non-play segments like breaks and player substitutions are included in the data.

The experiments are run as 10-fold cross validation; using one 10th of the data for test and the remaining part for training, then repeating the process 10 times with a new data subset for test each time.

Classification is tested in WEKA [2], which has several standard classifiers implemented. We test two decision tree classifiers (J48 and LMT), one naive Bayes classifier (NaiveBayes), two logistic regression models (Logistic and SimpleLogistic), and a lazy classifier (IBk/kNN). The best overall classification are obtained using the kNN classifier. The value of k is a compromise between using a large value to obtain a reliable estimate, and still have all k-nearest neighbours very close to the sample **x**. We test values of k from 1-15 and choose $k = 9$ for the best performance and being an odd number to avoid ties.

5.1. Results

Table 1 compares the classification results using visual features only, audio features only, and combined audio-visual features.

	Visual	Audio	Combined
Correct classification	86.67%	90.00%	**96.11%**
Precision	0.875	0.900	**0.962**
Recall	0.867	0.900	**0.961**

Table 1. Classification results of 180 1-minute video sequences from three sports types using a 10-fold cross validation.

Tables 2, 3, and 4 present the confusion matrices for classification of the three sports types using visual, audio, and combined audio-visual features, respectively.

Classified as Truth	Soccer	Basketball	Volleyball
Soccer	**53**	5	2
Basketball	1	**46**	13
Volleyball	0	3	**57**

Table 2. Confusion matrix, visual features.

Classified as Truth	Soccer	Basketball	Volleyball
Soccer	**56**	1	3
Basketball	4	**53**	3
Volleyball	3	4	**53**

Table 3. Confusion matrix, audio features.

Classified as Truth	Soccer	Basketball	Volleyball
Soccer	**58**	1	1
Basketball	1	**56**	3
Volleyball	0	1	**59**

Table 4. Confusion matrix, combined features.

From the confusion matrices it can be observed that for visual features, the largest misclassification happens for basketball videos, classified as volleyball. For all other methods only very few samples are misclassified, and the errors seem equally distributed between sports types. Using combined features, only seven samples are misclassified of a total of 180 sequences, resulting in a high correct classification rate of 96.11%. Our previous work on classification of sports types from only visual data obtained classification rates of 89.64% [6] and 94.5% [5]. These results are directly comparable in terms of image types and arena. However, different datasets were used for each work.

6. Conclusion

In this work we have shown the benefits of combining audio and visual features for sports type classification. As most video formats can include audio, no extra work are required in terms of data capturing, and the methods used are robust to noise and variability in play. MFCC features are extracted from the audio and combined with visual motion features. Classifying 180 1-minute video sequences representing three different sports types, captured in the same sports arena, we reach a correct classification rate of 96.11%, which is significantly higher than using visual or audio features alone.

The use of thermal cameras here are chosen for privacy reasons. However, the methods apply to other types of videos as well, only the detection step would need to be replaced in order to use RGB video.

In future work we will extend our dataset and perform tests on a larger number of sports types. Furthermore, it would be interesting to research the possibilities of classification of intensity levels for each activity, e.g., warm-up, practise, or competition level.

References

[1] HTK Toolkits, Cambridge, U.K. http://htk.eng.cam.ac. 4

[2] WEKA, University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/. 5

[3] L. Bai, S.-Y. Lao, H.-X. Liao, and J.-Y. Chen. Audio classification and segmentation for sports video structure extraction using support vector machine. In *International Conference on Machine Learning and Cybernetics*, pages 3303–3307, Aug 2006. 1

[4] R. Gade, A. Jørgensen, and T. B. Moeslund. Long-term occupancy analysis using graph-based optimisation in thermal imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[5] R. Gade and T. B. Moeslund. Classification of sports types from tracklets, 2014. KDD Workshop on Large-Scale Sports Analytics. 1, 5

[6] R. Gade and T. B. Moeslund. Classification of sports types using thermal imagery. In T. B. Moeslund, G. Thomas, and A. Hilton, editors, *Computer Vision in Sports*, Advances in Computer Vision and Pattern Recognition, pages 209–227. Springer International Publishing, 2014. 1, 2, 5

[7] X. Gibert, H. Li, and D. Doermann. Sports video classification using HMMS. In *International Conference on Multimedia and Expo (ICME)*, 2003. 2

[8] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 3

[9] J. Kapur, P. Sahoo, and A. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, 29(3):273 – 285, 1985. 3

[10] C. Krishna Mohan and B. Yegnanarayana. Classification of sport videos using edge-based features and autoassociative neural network models. *Signal, Image and Video Processing*, 4:61–73, 2010. 2

[11] J. Y. Lee and W. Hoff. Activity identification utilizing data mining techniques. In *IEEE Workshop on Motion and Video Computing (WMVC)*, 2007. 2

[12] L. Li, N. Zhang, L.-Y. Duan, Q. Huang, J. Du, and L. Guan. Automatic sports genre categorization and view-type classification over large-scale dataset. In *17th ACM international conference on Multimedia (MM)*, 2009. 2

[13] P. Mutchima and P. Sanguansat. TF-RNF: A novel term weighting scheme for sports video classification. In *IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, 2012. 2

[14] P. Oskouie, S. Alipour, and A.-M. Eftekhari-Moghadam. Multimodal feature extraction and fusion for semantic mining of soccer video: a survey. *Artificial Intelligence Review*, 42(2):173–210, 2014. 1

[15] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, pages 105–115, New York, NY, USA, 2000. ACM. 1

[16] M. Sigari, S. Sureshjani, and H. Soltanian-Zadeh. Sport video classification using an ensemble classifier. In *7th Iranian Machine Vision and Image Processing (MVIP)*, 2011. 2

[17] K. Subashini, S. Palanivel, and V. Ramaligam. Audio-video based segmentation and classification using SVM. In *Third International Conference on Computing Communication Networking Technologies (ICCCNT)*, pages 1–6, July 2012. 1

[18] D.-H. Wang, Q. Tian, S. Gao, and W.-K. Sung. News sports video shot classification with sports play field and motion features. In *International Conference on Image Processing (ICIP)*, 2004. 2

[19] J. Wang, C. Xu, and E. Chng. Automatic sports video genre classification using Pseudo-2D-HMM. In *18th International Conference on Pattern Recognition (ICPR)*, 2006. 2

[20] S. Wilson, C. Mohan, and K. Murthy. Event-based sports videos classification using HMM framework. In T. B. Moeslund, G. Thomas, and A. Hilton, editors, *Computer Vision in Sports*, Advances in Computer Vision and Pattern Recognition, pages 229–244. Springer International Publishing, 2014. 2

[21] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang. Effective and efficient sports highlights extraction using the minimum description length criterion in selecting GMM structures [audio classification]. In *IEEE International Conference on Multimedia and Expo*, volume 3, pages 1947–1950 Vol.3, June 2004. 1

[22] C. Xu, J. Cheng, Y. Zhang, Y. Zhang, and H. Lu. Sports video analysis: Semantics extraction, editorial content creation and adaptation. *Journal of Multimedia*, 4(2):69–79, 2009. 1

[23] Y. Yuan and C. Wan. The application of edge feature in automatic sports genre classification. In *IEEE Conference on Cybernetics and Intelligent Systems*, 2004. 2