EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Joint modality fusion and temporal context exploitation for semantic video analysis

Georgios Th Papadopoulos[1,2*], Vasileios Mezaris[1], Ioannis Kompatsiaris[1] and Michael G Strintzis[1,2]

Abstract

In this paper, a multi-modal context-aware approach to semantic video analysis is presented. Overall, the examined video sequence is initially segmented into shots and for every resulting shot appropriate color, motion and audio features are extracted. Then, Hidden Markov Models (HMMs) are employed for performing an initial association of each shot with the semantic classes that are of interest separately for each modality. Subsequently, a graphical modeling-based approach is proposed for jointly performing modality fusion and temporal context exploitation. Novelties of this work include the combined use of contextual information and multi-modal fusion, and the development of a new representation for providing motion distribution information to HMMs. Specifically, an integrated Bayesian Network is introduced for simultaneously performing information fusion of the individual modality analysis results and exploitation of temporal context, contrary to the usual practice of performing each task separately. Contextual information is in the form of temporal relations among the supported classes. Additionally, a new computationally efficient method for providing motion energy distribution-related information to HMMs, which supports the incorporation of motion characteristics from previous frames to the currently examined one, is presented. The final outcome of this overall video analysis framework is the association of a semantic class with every shot. Experimental results as well as comparative evaluation from the application of the proposed approach to four datasets belonging to the domains of tennis, news and volleyball broadcast video are presented.

Keywords: Video analysis, multi-modal analysis, temporal context, motion energy, Hidden Markov Models, Bayesian Network

1. Introduction

Due to the continuously increasing amount of video content generated everyday and the richness of the available means for sharing and distributing it, the need for efficient and advanced methodologies regarding video manipulation emerges as a challenging and imperative issue. As a consequence, intense research efforts have concentrated on the development of sophisticated techniques for effective management of video sequences [1]. More recently, the fundamental principle of shifting video manipulation techniques towards the processing of the visual content at a semantic level has been widely adopted. Semantic video analysis is the cornerstone of such intelligent video manipulation endeavors, attempting to bridge the so called *semantic gap* [2] and efficiently capture the underlying semantics of the content.

An important issue in the process of semantic video analysis is the number of modalities which are utilized. A series of single-modality based approaches have been proposed, where the appropriate modality is selected depending on the specific application or analysis methodology followed [3,4]. On the other hand, approaches that make use of two or more modalities in a collaborative fashion exploit the possible correlations and interdependencies between their respective data [5]. Hence, they capture more efficiently the semantic information contained in the video, since the semantics of the latter are typically embedded in multiple forms that are complementary to each other [6]. Thus, modality fusion generally enables the detection of more complex and

* Correspondence: papad@iti.gr
[1]CERTH/Informatics and Telematics Institute 6th Km. Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece
Full list of author information is available at the end of the article

higher-level semantic concepts and facilitates the effective generation of more accurate semantic descriptions.

In addition to modality fusion, the use of context has been shown to further facilitate semantic video analysis [7]. In particular, contextual information has been widely used for overcoming ambiguities in the audio-visual data or for solving conflicts in the estimated analysis results. For that purpose, a series of diverse contextual information sources have been utilized [8,9]. Among the available contextual information types, temporal context is of particular importance in video analysis. This is used for modeling temporal relations between semantic elements or temporal variations of particular features [10].

In this paper, a multi-modal context-aware approach to semantic video analysis is presented. Objective of this work is the association of each video shot with one of the semantic classes that are of interest in the given application domain. Novelties include the development of: (i) a graphical modeling-based approach for jointly realizing multi-modal fusion and temporal context exploitation, and (ii) a new representation for providing motion distribution information to Hidden Markov Models (HMMs). More specifically, for multi-modal fusion and temporal context exploitation an integrated Bayesian Network (BN) is proposed that incorporates the following key characteristics:

(a) **It simultaneously handles the problems of modality fusion and temporal context modeling**, taking advantage of all possible correlations between the respective data. This is a sharp contradistinction to the usual practice of performing each task separately.
(b) **It encompasses a probabilistic approach for acquiring and modeling complex contextual knowledge** about the long-term temporal patterns followed by the semantic classes. This goes beyond common practices that e.g. are limited to only learning pairwise temporal relations between the classes.
(c) **Contextual constraints are applied within a restricted time interval**, contrary to most of the methods in the literature that rely on the application of a time evolving procedure (e.g. HMMs, dynamic programming techniques, etc.) to the whole video sequence. The latter set of methods are usually prone to cumulative errors or are significantly affected by the presence of noise in the data.

All the above characteristics enable the developed BN to outperform other generative and discriminative learning methods. Concerning motion information processing, a new representation for providing motion energy distribution-related information to HMMs is presented that:

(a) **Supports the combined use of motion characteristics from the current and previous frames**, in order to efficiently handle cases of semantic classes that present similar motion patterns over a period of time.
(b) **Adopts a fine-grained motion representation**, rather than being limited to e.g. dominant global motion.
(c) **Presents recognition rates comparable to those of the best performing methods of the literature, while exhibiting computational complexity much lower than them** and similar to that of considerably simpler and less well-performing techniques.

An overview of the proposed video semantic analysis approach is illustrated in Figure 1.

The paper is organized as follows: Section 2 presents an overview of the relevant literature. Section 3 describes the proposed new representation for providing motion information to HMMs, while Section 4 outlines the respective audio and color information processing. Section 5 details the proposed new joint fusion and temporal context exploitation framework. Experimental results as well as comparative evaluation from the application of the proposed approach to four datasets belonging to the domains of tennis, news and volleyball broadcast video are presented in Section 6, and conclusions are drawn in Section 7.

2. Related work

2.1. Machine learning for video analysis

The usage of Machine Learning (ML) algorithms constitutes a robust methodology for modeling the complex relationships and interdependencies between the low-level audio-visual data and the perceptually higher-level semantic concepts. Among the algorithms of the latter category, HMMs and BNs have been used extensively for video analysis tasks. In particular, HMMs have been distinguished due to their suitability for modeling pattern recognition problems that exhibit an inherent temporality [11]. Among others, they have been used for performing video temporal segmentation, semantic event detection, highlight extraction and video structure analysis (e.g. [12-14]). On the other hand, BNs constitute an efficient methodology for learning causal relationships and an effective representation for combining prior knowledge and data [15]. Additionally, their ability to handle situations of missing data has also been reported [16]. BNs have been utilized in video analysis tasks such as semantic concept detection, video segmentation and event detection (e.g. [17,18]), to name a few.

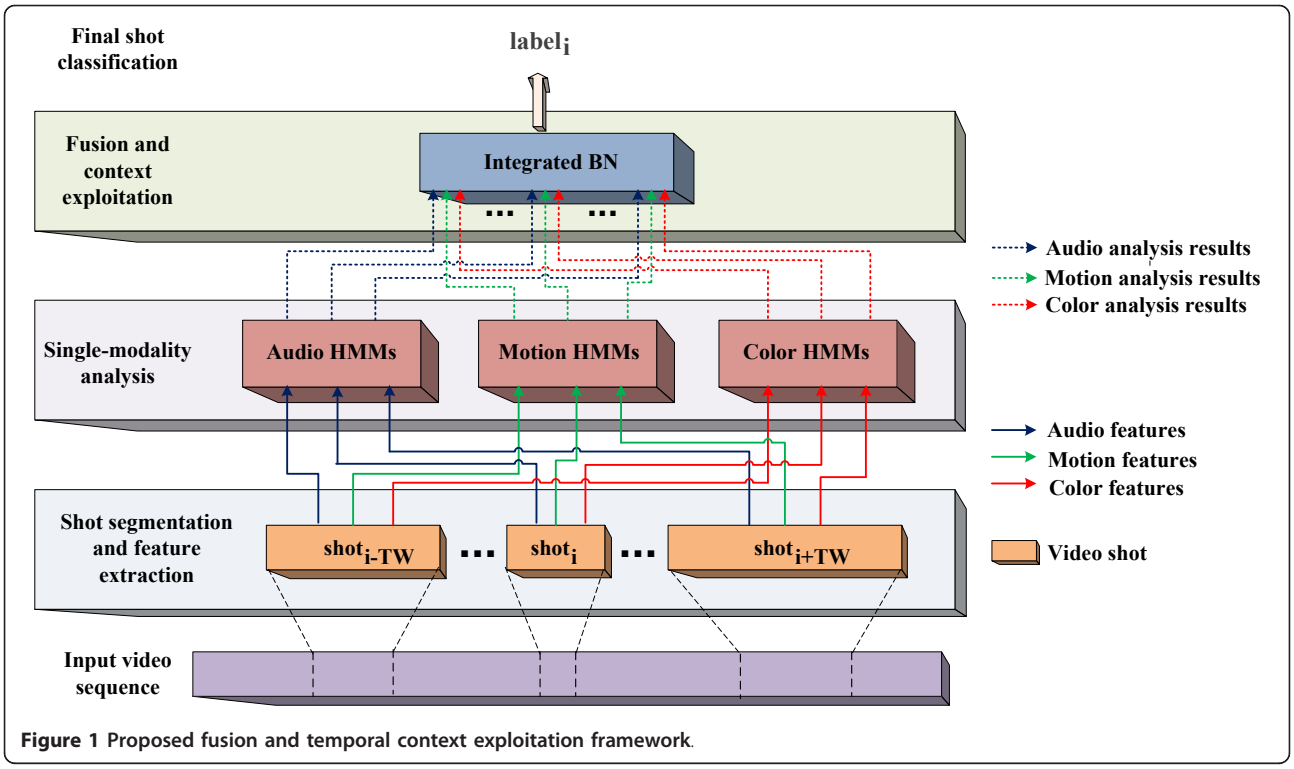

Figure 1 Proposed fusion and temporal context exploitation framework.

A review of machine learning-based methods for various video processing tasks can be found in [19]. Machine learning and other approaches specifically for modality fusion and temporal context exploitation towards semantic video analysis are discussed in the sequel.

2.2. Modality fusion and temporal context exploitation

Modality fusion aims at exploiting the correlations between data coming from different modalities to improve single-modality analysis results [6]. Bruno et al. introduce the notion of the multimodal dissimilarity spaces for facilitating the retrieval of video documents [20]. Additionally, a subspace-based multimedia data mining framework is presented for semantic video analysis in [21], which makes use of audio-visual information. Hoi et al. propose a multimodal-multilevel ranking scheme for performing large-scale video retrieval [22]. Tjondronegoro et al. [23] propose a hybrid approach, which integrates statistics and domain knowledge into logical rule-based models, for highlight extraction in sports video based on audio-visual features. Moreover, Xu et al. [24] incorporate web-casting text in sports video analysis using a text-video alignment framework.

On the other hand, contextual knowledge, and specifically temporal-related contextual information, has been widely used in semantic video manipulation tasks, in order to overcome possible audio-visual information ambiguities. In [25], temporal consistency is defined with respect to semantic concepts and its implications to video analysis and retrieval are investigated. Additionally, Xu et al. [26] introduce a HMM-based framework for modeling temporal contextual constraints in different semantic granularities. Dynamic programming techniques are used for obtaining the maximum likelihood semantic interpretation of the video sequence in [27]. Moreover, Kongwah [28] utilizes story-level contextual cues for facilitating multimodal retrieval, while Hsu et al. [29] model video stories, in order to leverage the recurrent patterns and to improve the video search performance.

While a plethora of advanced methods have already been proposed for either modality fusion or temporal context modeling, the possibility of jointly performing these two tasks has not been examined. The latter would allow the exploitation of all possible correlations and interdependencies between the respective data and consequently could further improve the recognition performance.

2.3. Motion representation for HMM-based analysis

A prerequisite for the application of any modality fusion or context exploitation technique is the appropriate and effective exploitation of the content low-level properties, such as color, motion, etc., in order to facilitate the derivation of a first set of high-level semantic descriptions. In video analysis, the focus is on motion representation

and exploitation, since the motion signal bears a significant portion of the semantic information that is present in a video sequence. Particularly for use together with HMMs, which have been widely used in semantic video analysis tasks, a plurality of motion representations have been proposed. You et al. [30] utilize global motion characteristics for realizing video genre classification and event analysis. In [26], a set of motion filters are employed for estimating the frame dominant motion in an attempt to detect semantic events in various sports videos. Additionally, Huang et al. consider the first four dominant motions and simple statistics of the motion vectors in the frame, for performing scene classification [12]. In [31], particular camera motion types are used for the analysis of football video. Moreover, Gibert et al. estimate the principal motion direction of every frame [32], while Xie et al. calculate the motion intensity at frame level [27], for realizing sport video classification and structural analysis of soccer video, respectively. Common characteristic of all the above methods is that they rely on the extraction of coarse-grained motion features, which may perform sufficiently well in certain cases. On the other hand, in [33] a more elaborate motion representation is proposed, making use of higher-order statistics for providing local-level motion information to HMMs. This accomplishes increased recognition performance, at the expense of high computational complexity.

Although several motion representations have been proposed for use together with HMMs, the development of a fine-grained representation combining increased recognition rates with low computational complexity remains a significant challenge. Additionally, most of the already proposed methods make use of motion features extracted at individual frames, which is insufficient when considering video semantic classes that present similar motion patterns over a period of time. Hence, the potential of incorporating motion characteristics from previous frames to the currently examined one needs also to be investigated.

3. Motion-based analysis

HMMs are employed in this work for performing an initial association of each shot s_i, $i = 1, ..., I$, of the examined video with one of the semantic classes of a set $E = \{e_j\}_{1 \le j \le J}$ based on motion information, as is typically the case in the relevant literature. Thus, each semantic class e_j corresponds to a process that is to be modeled by an individual HMM, and the features extracted for every shot s_i constitute the respective observation sequence [11]. For shot detection, the algorithm of [34] is used, mainly due to its low computational complexity.

According to the HMM theory [11], the set of sequential observation vectors that constitute an observation sequence need to be of fixed length and simultaneously of low-dimensionality. The latter constraint ensures the avoidance of HMM under-training occurrences. Thus, compact and discriminative representations of motion features are required. Among the approaches that have already been proposed (Section 2.3), simple motion representations such as frame dominant motion (e.g. [12,27,32]) have been shown to perform sufficiently well, when considering semantic classes that present quite distinct motion patterns. When considering classes with more complex motion characteristics, such approaches have been shown to be significantly outperformed by methods exploiting fine-grained motion representations (e.g. [33]). However, the latter is achieved at the expense of increased computational complexity. Taking into account the aforementioned considerations, a new method for motion information processing is proposed in this section. The proposed method makes use of fine-grained motion features, similarly to [33] to achieve superior performance, while having computational requirements that match those of much simpler and less well-performing approaches.

3.1. Motion pre-processing

For extracting the motion features, a set of frames is selected for each shot s_i. This selection is performed using a constant temporal sampling frequency, denoted by SF_m, and starting from the first frame. The choice of starting the selection procedure from the first frame of each shot is made for simplicity purposes and in order to maintain the computational complexity of the proposed approach low. Then, a dense motion field is computed for every selected frame making use of the optical flow estimation algorithm of [35]. Consequently, a motion energy field is calculated, according to the following equation:

$$M(u, v, t) = ||\vec{V}(u, v, t)|| \quad (1)$$

Where $\vec{V}(u, v, t)$ is the estimated dense motion field, $||.||$ denotes the norm of a vector and $M(u, v, t)$ is the resulting motion energy field. Variables u and v get values in the ranges $[1, V_{\text{dim}}]$ and $[1, H_{\text{dim}}]$ respectively, where V_{dim} and H_{dim} are the motion field vertical and horizontal dimensions (same as the corresponding frame dimensions in pixels). Variable t denotes the temporal order of the selected frames. The choice of transforming the motion vector field to an energy field is justified by the observation that often the latter provides more appropriate information for motion-based recognition problems [26,33]. The estimated motion energy field $M(u, v, t)$ is of high dimensionality. This decelerates the video processing, while motion information at this level of detail is not always required for analysis purposes.

Thus, it is consequently down-sampled, according to the following equations:

$$R(x,y,t) = M\left(\frac{2x-1}{2}\cdot V_s, \frac{2y-1}{2}\cdot H_s, t\right)$$
$$x = 1,\ldots,D, \quad y = 1,\ldots,D, \quad V_s = \lfloor\frac{V_{\dim}}{D}\rfloor, \quad H_s = \lfloor\frac{H_{\dim}}{D}\rfloor \tag{2}$$

where $R(x, y, t)$ is the estimated down-sampled motion energy field of predetermined dimensions and H_s, V_s are the corresponding horizontal and vertical spatial sampling frequencies.

3.2. Polynomial approximation

The computed down-sampled motion energy field $R(x, y, t)$, which is estimated for every selected frame, actually represents a motion energy distribution surface and is approximated by a 2D polynomial function of the following form:

$$\phi(\mu,\nu) = \sum_{\gamma,\delta}\beta_{\gamma\delta}\cdot(\mu-\mu_0)^{\gamma}\cdot(\nu-\nu_0)^{\delta}, \quad 0\le\gamma,\delta\le T \quad \text{and} \quad 0\le\gamma+\delta\le T \tag{3}$$

where T is the order of the function, $\beta_{\gamma\delta}$ its coefficients and μ_0, ν_0 are defined as $\mu_0 = \nu_0 = \frac{D}{2}$. The approximation is performed using the least-squares method.

The polynomial coefficients, which are calculated for every selected frame, are used to form an observation vector. The observation vectors computed for each shot s_i are utilized to form an observation sequence, namely the shot's motion observation sequence. This observation sequence is denoted by OS_i^m, where superscript m stands for motion. Then, a set of J HMMs can be directly employed, where an individual HMM is introduced for every defined semantic class e_j, in order to perform the shot-class association based on motion information. Every HMM receives as input the aforementioned motion observation sequence OS_i^m for each shot s_i and at the evaluation stage returns a posterior probability, denoted by $h_{ij}^m = P(e_j|OS_i^m)$. This probability, which represents the observation sequence's fitness to the particular HMM, indicates the degree of confidence with which class e_j is associated with shot s_i based on motion information. HMM implementation details are discussed in the experimental results section.

3.3. Accumulated motion energy field computation

Motion characteristics at a single frame may not always provide an adequate amount of information for discovering the underlying semantics of the examined video sequence, since different classes may present similar motion patterns over a period of time. This fact generally hinders the identification of the correct semantic class through the examination of motion features at distinct sequentially selected frames. To overcome this problem, the motion representation described in the previous subsection is appropriately extended to incorporate motion energy distribution information from previous frames as well. This results in the generation of an accumulated motion energy field.

Starting from the calculated motion energy fields $M(u, v, t)$ (Equation (2)), for each selected frame an accumulated motion energy distribution field is formed according to the following equation:

$$M_{\text{acc}}(u,v,t,\tau) = \frac{\sum_0^{\tau} w(\tau)\cdot M(u,v,t-\tau)}{\sum_0^{\tau} w(\tau)}, \quad \tau = 0,1,\ldots, \tag{4}$$

where t is the current frame, τ denotes previously selected frames and $w(\tau)$ is a time-dependent normalization factor that receives different values for every previous frame. Among other possible realizations, the normalization factor $w(\tau)$ is modeled by the following time descending function:

$$w(\tau) = \frac{1}{\eta^{\zeta\cdot\tau}}, \quad \eta > 1,\ \zeta > 0. \tag{5}$$

As can be seen from Equation (5), the accumulated motion energy distribution field takes into account motion information from previous frames. In particular, it gradually adds motion information from previous frames to the currently examined one with decreasing importance. The respective down-sampled accumulated motion energy field is denoted by $R_{\text{acc}}(x, y, t, \tau)$ and is calculated similarly to Equation (2) using $M_{\text{acc}}(u, v, t, \tau)$ instead of $M(u, v, t)$. An example of computing the accumulated motion energy fields for two tennis shots, belonging to the break and serve class respectively, is illustrated in Figure 2. As can be seen from this example, the incorporation of motion information from previous frames ($\tau = 1, 2$) causes the resulting $M_{\text{acc}}(u, v, t, \tau)$ fields to present significant dissimilarities with respect to the motion energy distribution, compared to the case when no motion information from previous frames ($\tau = 0$) is taken into account. These dissimilarities are more intense for the second case ($\tau = 2$) and they can facilitate towards the discrimination between these two semantic classes.

During the estimation of the $M_{\text{acc}}(u, v, t, \tau)$ fields, motion energy values from neighboring frames at the same position are accumulated, as described above. These values may originate from object motion, camera motion or both. Inevitably, when intense camera motion is present, it will superimpose any possible movement of the objects in the scene. For example, during a rally event in a volleyball video, sudden and extensive camera motion is observed, when the ball is transferred from one side of the court to the other. This camera motion supersedes any action of the players during that period.

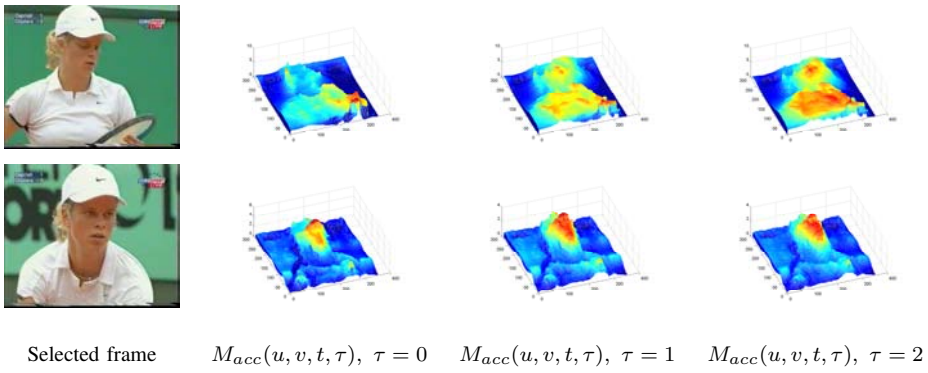

Figure 2 Examples of $M_{acc}(u, v, t, \tau)$ estimation for the break (1st row) and serve (2nd row) semantic classes in a tennis video.

Under the proposed approach, the presence of camera motion is considered to be part of the motion pattern of the respective semantic class. In other words, for the aforementioned example it is considered that the motion pattern of the rally event comprises relatively small player movements that are periodically interrupted by intense camera motions (i.e. when a team's offence incident occurs). The latter consideration constitutes the typical case in the literature [12,26,27].

Since the down-sampled accumulated motion energy field, $R_{acc}(x, y, t, \tau)$, is computed for every selected frame, a procedure similar to the one described in Section 3.2 is followed for providing motion information to the respective HMM structure and realizing shot-class association based on motion features. The difference is that now the accumulated energy fields, $R_{acc}(x, y, t, \tau)$, are used during the polynomial approximation process, instead of the motion energy fields, $R(x, y, t)$.

3.4. Discussion

In the authors' previous work [33], motion field estimation by means of optical flow was initially performed for all frames of each video shot. Then, the kurtosis of the optical flow motion estimates at each pixel was calculated for identifying which motion values originate from true motion rather than measurement noise. For the pixels where only true motion was observed, energy distribution-related information, as well as a complementary set of features that highlight particular spatial attributes of the motion signal, were extracted. For modeling the energy distribution-related information, the polynomial approximation method also described in Section 3.2 was followed. Although this local-level representation of the motion signal was shown to significantly outperform previous approaches that relied mainly on global- or camera-level representations, this was accomplished at the expense of increased computational complexity. The latter was caused by: (a) the need to process all frames of every shot, and (b) the need to calculate higher-order statistics from them and compute additional features.

The aim of the approach proposed in this work was to overcome the aforementioned limitations in terms of computational complexity, while also attempting to maintain increased recognition performance. For achieving this, the polynomial approximation that can model motion information was directly applied to the accumulated motion energy fields $M_{acc}(u, v, t, \tau)$. These were estimated for only a limited number of frames, i.e. those selected at a constant temporal sampling frequency (SF_m). This choice alleviates both the need for processing all frames of each shot and the need for computationally expensive statistical and other features calculations. The resulting method is shown by experimentation to be comparable with simpler motion representation approaches [12,27,32] in terms of computational complexity, while maintaining a recognition performance similar to that of [33].

4. Color- and audio-based analysis

For the color and audio information processing, common techniques from the relevant literature are adopted. In particular, a set of global-level color histograms of F_c-bins in the RGB color space [36] is estimated at equally spaced time intervals for each shot s_i, starting from the first frame; the corresponding temporal sampling frequency is denoted by SF_c. The aforementioned set of color histograms are normalized in the interval [-1, 1] and subsequently they are utilized to form a corresponding observation sequence, namely the color observation sequence which is denoted by OS_i^c. Similarly to the motion analysis case, a set of J HMMs is employed, in order to realize the association of the examined shot s_i with the defined classes e_j based solely on color information. At the evaluation stage each HMM returns a posterior probability, which is denoted by $h_{ij}^c = P(e_j|OS_i^c)$

and indicates the degree of confidence with which class e_j is associated with shot s_i. On the other hand, the widely used Mel Frequency Cepstral Coefficients (MFCC) are utilized for the audio information processing [37]. In the relative literature, apart from the MFCC coefficients, other features that highlight particular attributes of the audio signal have also been used for HMM-based audio analysis (like standard deviation of zero crossing rate [12], pitch period [38], short-time energy [39], etc.). However, the selection of these individual features is in principle performed heuristically and the efficiency of each of them has only been demonstrated in specific application cases. On the contrary, the MFCC coefficients provide a more complete representation of the audio characteristics and their efficiency has been proven in numerous and diverse application domains [40-44]. Taking into account the aforementioned facts, while also considering that this work aims at adopting common techniques of the literature for realizing generic audio-based shot classification, only the MFCC coefficients are considered in the proposed analysis framework. More specifically, F_a MFCC coefficients are estimated at a sampling rate of SF_a, while for their extraction a sliding window of length F_w is used. The set of MFCC coefficients calculated for shot s_i serves as the shot's audio observation sequence, denoted by OS_i^a. Similarly to the motion and color analysis cases, a set of J HMMs is introduced. The estimated posterior probability, denoted by $h_{ij}^a = P(e_j|OS_i^a)$, indicates this time the degree of confidence with which class e_j is associated with shot s_i based solely on audio information. It must be noted that a set of annotated video content, denoted by U_{tr}^1, is used for training the developed HMM structure. Using this, the constructed HMMs acquire the appropriate implicit knowledge that will enable the mapping of the low-level audio-visual data to the defined high-level semantic classes separately for every modality.

5. Joint modality fusion and temporal context exploitation

Graphical models constitute an efficient methodology for learning and representing complex probabilistic relationships among a set of random variables [45]. BNs are a specific type of graphical models that are particularly suitable for learning causal relationships [15]. To this end, BNs are employed in this work for probabilistically learning the complex relationships and interdependencies that are present among the audio-visual data. Additionally, their ability of learning causal relationships is exploited for acquiring and modeling temporal contextual information. In particular, an integrated BN is proposed for jointly performing modality fusion and temporal context exploitation. Key part of the latter is the definition of an appropriate and expandable network structure. The developed structure enables contextual knowledge acquisition in the form of temporal relations among the supported high-level semantic classes and incorporation of information from different sources. For that purpose, a series of sub-network structures, which are integrated to the overall network, are defined. The individual components of the developed framework are detailed in the sequel.

5.1. Modality fusion

A BN structure is initially defined for performing the fusion of the computed single-modality analysis results. Subsequently, a set of J such structures is introduced, one for every defined class e_j. The first step in the development of any BN is the identification and definition of the random variables that are of interest for the given application. For the task of modality fusion the following random variables are defined: (a) variable CL_j, which corresponds to the semantic class e_j with which the particular BN structure is associated, and (b) variables A_j, C_j and M_j, where an individual variable is introduced for every considered modality. More specifically, random variable CL_j denotes the fact of assigning class e_j to the examined shot s_i. Additionally, variables A_j, C_j and M_j represent the initial shot-class association results computed for shot s_i from every separate modality processing for the particular class e_j, i.e. the values of the estimated posterior probabilities h_{ij}^a, $h_{ij}^c$and h_{ij}^m (Sections 3 and 4). Subsequently, the space of every introduced random variable, i.e. the set of possible values that it can receive, needs to be defined. In the presented work, discrete BNs are employed, i.e. each random variable can receive only a finite number of mutually exclusive and exhaustive values. This choice is based on the fact that discrete space BNs are less prone to under-training occurrences compared to the continuous space ones [16]. Hence, the set of values that variable CL_j can receive is chosen equal to $\{cl_{j1}, cl_{j2}\} = \{True, False\}$, where $True$ denotes the assignment of class e_j to shot s_i and $False$ the opposite. On the other hand, a discretization step is applied to the estimated posterior probabilities h_{ij}^a, h_{ij}^c and h_{ij}^m for defining the spaces of variables A_j, C_j and M_j, respectively. The aim of the selected discretization procedure is to compute a close to uniform discrete distribution for each of the aforementioned random variables. This was experimentally shown to better facilitate the BN inference, compared to discretization with constant step or other common discrete distributions like gaussian and poisson.

The discretization is defined as follows: a set of annotated video content, denoted by U_{tr}^2, is initially formed and the single-modality shot-class association results are

computed for each shot. Then, the estimated posterior probabilities are grouped with respect to every possible class-modality combination. This results in the formulation of sets $L_j^b = \{h_{nj}^b\}_{1\le n\le N}$, where $b \in \{a, c, m\} \equiv \{audio, color, motion\}$ is the modality used and N is the number of shots in U_{tr}^2. Consequently, the elements of the aforementioned sets are sorted in ascending order, and the resulting sets are denoted by $\acute{L}_j^b$. If Q denotes the number of possible values of every corresponding random variable, these are defined according to the following equations:

$$B_j = \begin{cases} b_{j1} \text{ if } h_{ij}^b \in [0, \acute{L}_j^b(K)) \\ b_{jq} \text{ if } h_{ij}^b \in [\acute{L}_j^b(K\cdot(q-1)), \acute{L}_j^b(K\cdot q)), \quad q \in [2, Q-1] \\ b_{jQ} \text{ if } h_{ij}^b \in [\acute{L}_j^b(K\cdot(Q-1)), 1] \end{cases} \tag{6}$$

where $K = \lfloor \frac{N}{Q} \rfloor$, $\acute{L}_j^b(0)$ denotes the oth element of the ascending sorted set $\acute{L}_j^b$, and b_{j1}, b_{j2}, ..., b_{jQ} denote the values of variable B_j ($B \in \{A, C, M\}$). From the above equations, it can be seen that although the number of possible values for all random variables B_j is equal to Q, the corresponding posterior probability ranges with which they are associated are generally different.

The next step in the development of this BN structure is to define a Directed Acyclic Graph (DAG), which represents the causality relations among the introduced random variables. In particular, it is assumed that each of the variables A_j, C_j and M_j is conditionally independent of the remaining ones given CL_j. In other words, it is considered that the semantic class, to which a video shot belongs, fully determines the features observed with respect to every modality. This assumption is typically the case in the relevant literature [17,46] and it is formalized as follows:

$$Ip(z, Z_j - z | CL_j), \quad z \in Z_j \text{ and } Z_j = \{A_j, C_j, M_j\}, \tag{7}$$

where $Ip(.)$ stands for statistical independence. Based on this assumption, the following condition derives, with respect to the conditional probability distribution of the defined random variables:

$$P(a_j, c_j, m_j | cl_j) = P(a_j | cl_j) \cdot P(c_j | cl_j) \cdot P(m_j | cl_j), \tag{8}$$

where $P(.)$ denotes the probability distribution of a random variable, and a_j, c_j, m_j and cl_j denote values of the variables A_j, C_j, M_j and CL_j, respectively. The corresponding DAG, denoted by $\mathbb{G}_j$, that incorporates the conditional independence assumptions expressed by Equation (7) is illustrated in Figure 3a. As can be seen from this figure, variable CL_j corresponds to the parent node of $\mathbb{G}_j$, while variables A_j, C_j and M_j are associated with children nodes of the former. It must be noted that the direction of the arcs in $\mathbb{G}_j$ defines explicitly the causal relationships among the defined variables.

From the casual DAG depicted in Figure 3a and the conditional independence assumption stated in Equation (8), the conditional probability $P(cl_j|a_j, c_j, m_j)$ can be estimated. This represents the probability of assigning class e_j to shot s_i given the initial single-modality shot-class association results and it can be calculated as follows:

$$P(cl_j | a_j, c_j, m_j) = \frac{P(a_j, c_j, m_j | cl_j) \cdot P(cl_j)}{P(a_j, c_j, m_j)} = \frac{P(a_j|cl_j) \cdot P(c_j|cl_j) \cdot P(m_j|cl_j) \cdot P(cl_j)}{P(a_j, c_j, m_j)} \tag{9}$$

From the above equation, it can be seen that the proposed BN-based fusion mechanism accomplishes to adaptively learn the impact of every utilized modality on the detection of each supported semantic class. In particular, it adds variable significance to every single-modality analysis value (i.e. values a_j, c_j and m_j) by calculating the conditional probabilities $P(a_j|cl_j)$, $P(c_j|cl_j)$ and $P(m_j|cl_j)$ during training, instead of determining a unique impact factor for every modality.

5.2. Temporal context exploitation

Besides multi-modal information, contextual information can also contribute towards improved shot-class association performance. In this work, temporal contextual information in the form of temporal relations among the different semantic classes is exploited. This choice is based on the observation that often classes of a particular domain tend to occur according to a specific order in time. For example, a shot belonging to the

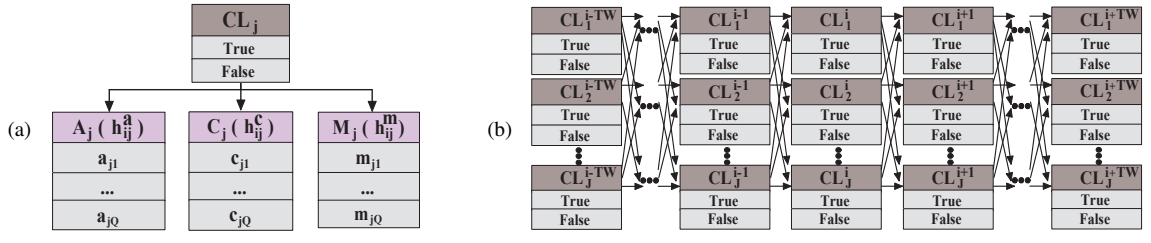

Figure 3 Developed DAG $\mathbb{G}_j$ for modality fusion (a) and $\mathbb{G}_c$ for temporal context modeling (b).

class 'rally' in a tennis domain video is more likely to be followed by a shot depicting a 'break' incident, rather than a 'serve' one. Thus, information about the classes' occurrence order can serve as a set of constraints denoting their 'allowed' temporal succession. Since BNs constitute a robust solution to probabilistically learning causality relationships, as described in the beginning of Section 5, another BN structure is developed for acquiring and modeling this type of contextual information. Although other methods that utilize the same type of temporal contextual information have already been proposed, the presented method includes several novelties and advantageous characteristics: (a) it encompasses a probabilistic approach for automatically acquiring and representing complex contextual information after a training procedure is applied, instead of defining a set of heuristic rules that accommodate to a particular application case [47], and (b) contextual constraints are applied within a restricted time interval, i.e. whole video sequence structure parsing is not required for reaching good recognition results, as opposed to e.g. the approaches of [12,26].

Under the proposed approach, an appropriate BN structure is constructed for supporting the acquisition and the subsequent enforcement of temporal contextual constraints. This structure enables the BN inference to take into account shot-class association related information for every shot s_i, as well as for all its neighboring shots that lie within a certain time window, for deciding upon the class that is eventually associated with shot s_i. For achieving this, an appropriate set of random variables is defined, similarly to the case of the development of the BN structure used for modality fusion in Section 5.1. Specifically, the following random variables are defined: (a) a set of J variables, one for every defined class e_j, and which are denoted by CL_j^i; these variables represent the classes that are eventually associated with shot s_i, after the temporal context exploitation procedure is performed, and (b) two sets of $J \cdot TW$ variables denoted by CL_j^{i-r} and CL_j^{i+r}, which denote the shot-class associations of previous and subsequent shots, respectively; $r\varepsilon[1, TW]$, where TW denotes the length of the aforementioned time window, i.e. the number of previous and following shots, whose shot-class association results will be taken into account for reaching the final class assignment decision for shot s_i. All together the aforementioned variables will be denoted by CL_j^k, where $i - TW \leq k \leq i + TW$. Regarding the set of possible values for each of the aforementioned random variables, this is chosen equal to $\{cl_{j1}^k, cl_{j2}^k\} = \{True, False\}$, where $True$ denotes the association of class e_j with the corresponding shot and $False$ the opposite.

The next step in the development of this BN structure is the identification of the causality relations among the defined random variables and the construction of the respective DAG, which represents these relations. For identifying the causality relations, the definition of causation based on the concept of manipulation is adopted [15]. The latter states that for a given pair of random variables, namely X and Y, variable X has a causal influence on Y if a manipulation of the values of X leads to a change in the probability distribution of Y. Making use of the aforementioned definition of causation, it can be easily observed that each defined variable CL_j^i has a causal influence on every following variable CL_j^{i+1}, $\forall j$. This can be better demonstrated by the following example: suppose that for a given volleyball game video, it is known that a particular shot belongs to the class 'serve'. Then, the subsequent shot is more likely to depict a 'rally' instance rather than a 'replay' one. Additionally, from the extension of the aforementioned example, it can be inferred that any variable $CL_j^{i_1}$ has a causal influence on variable $CL_j^{i_2}$ for $i_1 < i_2$. However, for constructing a causal DAG, only the direct causal relations among the corresponding random variables must be defined [15]. To this end, only the causal relations between variables $CL_j^{i_1}$ and $CL_j^{i_2}$, $\forall j$, and for $i_2 = i_1 + 1$, are included in the developed DAG, since any other variable $CL_j^{i_1}$ is correlated with $CL_j^{i_2}$, where $i_1' + 1 < i_2'$, transitively through variables $CL_j^{i_3}$, for $i_1' < i_3' < i_2'$. Taking into account all the aforementioned considerations, the causal DAG $\mathbb{G}_c$illustrated in Figure 3b is defined. Regarding the definition of the causality relations, it can be observed that the following three conditions are satisfied for $\mathbb{G}_c$: (a) there are no hidden common causes among the defined variables, (b) there are no causal feedback loops, and (c) selection bias is not present, as demonstrated by the aforementioned example. As a consequence, the causal Markov assumption is warranted to hold. Additionally, a BN can be constructed from the causal DAG $\mathbb{G}_c$ and the joint probability distribution of its random variables satisfies the Markov condition with $\mathbb{G}_c$[15].

5.3. Integration of modality fusion and temporal context exploitation

Having developed the causal DAGs $\mathbb{G}_c$, used for temporal context exploitation, and $\mathbb{G}_j$, utilized for modality fusion, the next step is to construct an integrated BN structure for jointly performing modality fusion and temporal context exploitation. This is achieved by replacing each of the nodes that correspond to variables CL_j^k in $\mathbb{G}_c$ with the appropriate $\mathbb{G}_j$, using j as selection

criterion and maintaining that the parent node of $\mathbb{G}_j$ takes the position of the respective node in $\mathbb{G}_c$. Thus, the resulting overall BN structure, denoted by $\mathbb{G}$, comprises of a set of sub-structures integrated to the DAG depicted in Figure 3b. This overall structure encodes both cross-modal as well as temporal relations among the supported semantic classes. Moreover, for the integrated causal DAG $\mathbb{G}$, the causal Markov assumption is warranted to hold, as described above. To this end, the joint probability distribution of the random variables that are included in $\mathbb{G}$, which is denoted by P_{joint} and satisfies the Markov condition with $\mathbb{G}$, can be defined. The latter condition states that every random variable X that corresponds to a node in $\mathbb{G}$ is conditionally independent of the set of all variables that correspond to its nondescendent nodes, given the set of all variables that correspond to its parent nodes [15]. For a given node X, the set of its nondescendent nodes comprises all nodes with which X is not connected through a path in $\mathbb{G}$, starting from X. Hence, the Markov condition is formalized as follows:

$$Ip(X, ND_X|PA_X), \tag{10}$$

where ND_X denotes the set of variables that correspond to the nondescendent nodes of X and PA_X the set of variables that correspond to its parent nodes. Based on the condition stated in Equation (10), P_{joint} is equal to the product of the conditional probability distributions of the random variables in $\mathbb{G}$ given the variables that correspond to the parent nodes of the former, and is represented by the following equations:

$$\begin{aligned} &P_{\text{joint}}\left(\{a_j^k, c_j^k, m_j^k, cl_j^k\}_{1\le j\le J}^{i-TW\le k\le i+TW}\right) = P_1 \cdot P_2 \cdot P_3 \\ &P_1 = \prod_{j=1}^{J}\prod_{k=i-TW}^{i+TW} P(a_j^k|cl_j^k)\cdot P(c_j^k|cl_j^k)\cdot P(m_j^k|cl_j^k) \\ &P_2 = \prod_{j=1}^{J}\prod_{\acute{k}=i-TW+1}^{i+TW} P(cl_j^{\acute{k}}|cl_1^{\acute{k}-1},\ldots,cl_j^{\acute{k}-1}),\ P_3 = \prod_{j=1}^{J} P(cl_j^{i-TW}), \end{aligned} \tag{11}$$

where a_j^k, c_j^k and m_j^k are the values of the variables A_j^k, C_j^k and M_j^k, respectively. The pair $(\mathbb{G}, P_{\text{joint}})$, which satisfies the Markov condition as already described, constitutes the developed integrated BN.

Regarding the training process of the integrated BN, the set of all conditional probabilities among the defined conditionally-dependent random variables of $\mathbb{G}$, which are also reported in Equation (11), are estimated. For this purpose, the set of annotated video content U_{tr}^2, which was also used in Section 5.1 for input variable discretization, is utilized. At the evaluation stage, the integrated BN receives as input the single-modality shot-class association results of all shots that lie within the time window TW defined for shot s_i, i.e. the set of values $W_i = \{a_j^k, c_j^k, m_j^k\}_{1\le j\le J}^{i-TW\le k\le i+TW}$ defined in Equation (11). These constitute the so called evidence data that a BN requires for performing inference. Then, the BN estimates the following set of posterior probabilities (degrees of belief), making use of all the pre-computed conditional probabilities and the defined local independencies among the random variables of $\mathbb{G}: P(CL_j^i = True|W_i)$, for $1 \le j \le J$. Each of these probabilities indicates the degree of confidence, denoted by h_{ij}^f, with which class e_j is associated with shot s_i.

5.4. Discussion

Dynamic Bayesian Networks (DBNs), and in particular HMMs, have been widely used in semantic video analysis tasks due to their suitability for modeling pattern recognition problems that exhibit an inherent temporality (Section 2.1). Regardless of the considered analysis task, significant weaknesses that HMMs present have been highlighted in the literature. In particular: (a) Standard HMMs have been shown not to be adequately efficient in modeling long-term temporal dependencies in the data that they receive as input [48]. This is mainly due to their state transition distribution, which obeys the Markov assumption, i.e. the current state that a HMM lies in depends only on its previous state. (b) HMMs rely on the Viterbi algorithm during the decoding procedure, i.e. during the estimation of the most likely sequence of states that generates the observed data. The resulting Viterbi sequence usually represents only a small fraction of the total probability mass, with many other state sequences potentially having nearly equal likelihoods [49]. As a consequence, the Viterbi alignment is rather sensitive to the presence of noise in the input data, i.e. it may be easily misguided.

In order to overcome the limitations imposed by the traditional HMM theory, a series of improvements and modifications have been proposed. Among the most widely adopted ones is the concept of Hierarchical HMMs (H-HMMs) [50]. These make use of HMMs at different levels, in order to model data at different time scales; hence, aiming at efficiently capturing and modeling long-term relations in the input data. However, this results in a significant increase of the parameter space, and as a consequence H-HMMs suffer from the problem of overfitting and require large amounts of data for training [48]. To this end, Layered HMMs (L-HMMs) have been proposed [51] for increasing the robustness to overfitting occurrences, by reducing the size of the parameter space. L-HMMs can be considered as a variant of H-HMMs, where each layer of HMMs is trained independently and the inferential results from

each layer serve as training data for the layer above. Although L-HMMs are advantageous in terms of robustness to under-training occurrences compared to H-HMMs, this attribute is accompanied by reduced efficiency in modeling long-term temporal relationships in the data. While both H-HMMs and L-HMMs have been experimentally shown to generally outperform the traditional HMMs, maintaining that the requirements concerning their application are met, their efficiency still depends heavily on the corresponding generalized Viterbi algorithm; hence, they do not fully overcome the limitations of standard HMMs.

Regarding the integrated BN developed in this work, on the other hand, a fixed time window of predetermined length is used with respect to each shot s_i. This window denotes the number of previous and following shots whose shot-class association results (coming from all considered modalities) are taken into account for reaching the final class assignment decision for shot s_i. Hence, the resulting BN is capable of modeling complex and long-term temporal relationships among the supported semantic classes in a time interval equal to the defined time window, as can be seen from term P_2 in Equation (11). This advantageous characteristic significantly differentiates the proposed BN from HMM-based approaches (including both H-HMMs and L-HMMs). The latter take into account information about only the previous state ω_{t-1} for estimating the current state ω_t of the examined stochastic process [11]. Furthermore, the final class association decision is reached independently for each shot s_i, while taking into account the evidence data W_i defined for it rather than being dependent upon the final class association decision reached for shot s_{i-1}. More specifically, the set of posterior probabilities $P(CL_j^i = True|W_i)$ (for $1 \leq j \leq J$), which are estimated after performing the proposed BN inference for shot s_i (as described in Section 5.3), are computed without being affected by the calculation of the respective probabilities $P(CL_j^{i-1} = True|W_{i-1})$ estimated for shot s_{i-1}. To this end, the detrimental effects caused by the presence of noise in the input data are reduced, since evidence over a series of consecutive shots are examined in order to decide on the final class assignment for shot s_i. At the same time propagation of errors caused by noise to following shots (e.g. shots s_{i+1}, s_{i+2}, etc.) is prevented. On the other hand, HMM-based systems rely on the fundamental principle that for estimating the current state ω_t of the system information about only its previous state ω_{t-1} is considered; thus, rendering the HMM decoding procedure rather sensitive to the presence of noise and likely to be misguided. Taking into account the aforementioned considerations, the proposed integrated BN is expected to outperform other similar HMM-based approaches of the literature, as will be experimentally shown in Section 6.

6. Experimental results

The proposed approach was experimentally evaluated and compared with literature approaches using videos of the tennis, news and volleyball broadcast domains. The selection of these application domains is made mainly due to the following characteristics that the videos of the aforementioned categories present: (a) a set of meaningful high-level semantic classes, whose detection often requires the use of multi-modal information, is present in such videos, and (b) videos belonging to these domains present relatively well-defined temporal structure, i.e. the semantic classes that they contain tend to occur according to a particular order in time. In addition, the semantic analysis of such videos remains a challenging problem, which makes them suitable for the evaluation and comparison of relevant techniques. It should be emphasized here that application of the proposed method to any other domain, where an appropriate set of semantic classes that tend to occur according to particular temporal patterns can be defined, is straightforward, i.e. no domain-specific algorithmic modifications or adaptations are needed. In particular, only a set of manually annotated video content is required by the employed HMMs and BNs for parameter learning.

6.1. Datasets

For experimentation in the domain of tennis broadcast video, four semantic classes of interest were defined, coinciding with four high-level semantic events that typically dominate a broadcasted game. These are: (a) rally: when the actual game is played, (b) serve: is the event starting at the time that the player is hitting the ball to the ground, while he is preparing to serve, and finishes at the time the player performs the serve hit, (c) replay: when a particular incident of increased importance is broadcasted again, usually in slow motion, and (d) break: when a break in the game occurs, i.e. the actual game is interrupted and the camera may show the players resting, the audience, etc. For the news domain, the following classes were defined: (a) anchor: when the anchor person announces the news in a studio environment, (b) reporting: when live-reporting takes place or a speech/interview is broadcasted, (c) reportage: comprises of the displayed scenes, either indoors or outdoors, relevant to every broadcasted news item, and (d) graphics: when any kind of graphics is depicted in the video sequence, including news start/end signals, maps, tables or text scenes. Finally, for experimentation in the domain of volleyball broadcast video, two sets of semantic classes were defined. The first one comprises the

same semantic classes defined for the tennis domain (volleyball-I), while for the second set (volleyball-II) the following nine classes are defined: rally, ace, serve, serve preparation, replay, player celebration, tracking single player, face close-up and tracking multiple players. The semantic classes defined for the volleyball-II domain are generally sub-classes of the corresponding ones defined for the volleyball-I domain.

Following the definition of the semantic classes of interest, an appropriate set of videos was collected for every selected domain. Each video was temporally segmented using the algorithm of [34] and every resulting shot was manually annotated according to the respective class definitions. Then, the aforementioned videos were used to form the following content sets for each domain: training set U_{tr}^1 (used for training the developed HMM structure), training set U_{tr}^2 (utilized for training the integrated BN) and test set U_{te} (used for evaluation). Detailed descriptions of these datasets, which constitute extensions of the datasets used in [33], are given in Table 1. Additionally, the annotations and features for each dataset are publicly available [a].

Due to the large quantity and significant diversity of the real-life videos that were collected for each domain, the risk of over-training (i.e., of classifier over-fitting) was considered to be low in our experiments. This assumption is reinforced by the fact that the proposed approach achieves high recognition rates on 4 datasets of diverse nature and varying complexity, while also outperforming other common techniques of the literature, as shown in the following sections. Based on this, only typical methodologies for avoiding over-fitting occurrences and maintaining high generalization ability were considered in this work (e.g. selecting appropriate training algorithms for the employed ML models, as outlined in the sequel; setting not too strict termination criteria during training; etc.). However, for use or evaluation of the proposed techniques on smaller, rather specific or less diverse datasets (e.g. datasets generated under significantly constrained environmental conditions), exploiting more sophisticated techniques such as cross-validation (rather than employing fixed training/evaluation sets) can be envisaged, similarly to e.g. [14].

6.2. Implementation details

For the initial shot-class association (Sections 3 and 4), the value of the temporal sampling frequency SF_m used for motion feature extraction was set equal to 125 ms. Considering that the frame rate of the utilized videos is equal to 25 fps (Table 1), the aforementioned value of

Table 1 Datasets used for experimentation

Domain	Content used	Formed sets		
		U_{tr}^1	U_{tr}^2	U_{te}
Tennis e_1:rally, e_2:serve, e_3:replay, e_4:break	16 videos (352 × 288, 25 fps) of professional tennis games from various international tournaments	437 shots e_1:167 e_2:44 e_3:27 e_4:199	754 shots e_1:258 e_2:85 e_3:41 e_4:370	424 shots e_1:138 e_2:52 e_3:23 e_4:211
News e_1:anchor, e_2:reporting, e_3:reportage, e_4:graphics	32 videos (352 × 288, 25 fps) of news broadcast from Deutsche Welle[1]	338 shots e_1:70 e_2:46 e_3:174 e_4:48	557 shots e_1:80 e_2:71 e_3:337 e_4:69	293 shots e_1:59 e_2:28 e_3:174 e_4:32
Volleyball-I e_1:rally, e_2:serve, e_3:replay, e_4:break	20 videos (352 × 264, 25 fps) of Volleyball broadcast from the Beijing 2008 men's olympic tournament	262 shots e_1:67 e_2:42 e_3:27 e_4:126	562 shots e_1:129 e_2:94 e_3:69 e_4:270	532 shots e_1:151 e_2:74 e_3:71 e_4:236
Volleyball-II E_1:rally, e_2:ace, e_3:serve, e_4:serve preparation, e_5:replay, e_6:player celebration, e_7:tracking single player, e_8:face close-up, e_9: tracking multiple players	Same with Volleyball-I videos. The videos forming test set U_{te} are the same with the ones used for evaluation in the volleyball-I domain. Difference in the total number of considered shots is due to the more extended set of semantic classes used for performing manual video annotation.	422 shots e_1:96 e_2:18 e_3:50 e_4:24 e_5:41 e_6:78 e_7:49 e_8:23 e_9:43	452 shots e_1:90 e_2:20 e_3:45 e_4:32 e_5:55 e_6:99 e_7:34 e_8:17 e_9:60	538 shots e_1:122 e_2:17 e_3:60 e_4:19 e_5:71 e_6:94 e_7:57 e_8:21 e_9:77

the sampling frequency means that the processing of approximately 8 frames per second is required by the proposed approach, i.e. every third frame of each shot is selected. A third order polynomial function was used, according to Equation (3), and the value of parameter D in Equation (2), which is used to define the horizontal and vertical spatial sampling frequencies (H_s and V_s, respectively) was set equal to 40, similarly to [33]. Parameters η and ζ that define the time descending function in Equation (5) were set equal to 3 and 0.5, respectively. In parallel to motion feature extraction, color histograms of F_c = 16 bins were calculated at a temporal sampling frequency of SF_c = 125 ms (Section 4). With respect to the audio information processing, F_a = 12 MFCC coefficients were estimated at a sampling rate of SF_a = 20 ms, while for their extraction a sliding window of length F_w = 30 ms was used. The value of SF_a is different than that of SF_m (used for motion feature extraction) due to the nature of the audio information and its MFCC representation, which require that MFCC coefficients are calculated at a relatively high rate and in temporal windows of correspondingly short duration [52]. The values of the aforementioned parameters were selected after experimentation. It was observed that small deviations from these values resulted into negligible variations in the overall classification performance.

Regarding the HMM structure of Sections 3 and 4, fully connected first order HMMs, i.e. HMMs allowing all possible hidden state transitions, were utilized for performing the mapping of the single-modality low-level features to the high-level semantic classes. For every hidden state the observations were modeled as a mixture of Gaussians (a single Gaussian was used for every state). The employed Gaussian Mixture Models (GMMs) were set to have full covariance matrices for exploiting all possible correlations between the elements of each observation. Additionally, the Baum-Welch (or Forward-Backward) algorithm was used for training, while the Viterbi algorithm was utilized during the evaluation. Furthermore, the number of hidden states of the HMM models for every separate modality was considered as a free variable. The developed HMM structure was realized using the software libraries of [53].

After shot-class association based on single-modality information is performed separately for every utilized modality, the integrated BN described in Section 5 was used for realizing joint modality fusion and temporal context exploitation. The value of variable Q in Equation (6), which determines the number of possible values of random variables A_j, C_j and M_j in the $\mathbb{G}_j$ BN substructure, was set equal to 9, 11, 7 and 10, for the tennis, news, volleyball-I and volleyball-II domains, respectively. These values led to the best overall inferential results, as will be discussed in detail in Section 6.4.1. The developed BN was trained using the Expectation Maximization (EM) approach, while probability propagation was realized using a junction tree mechanism [54].

6.3. Motion analysis results

In this section experimental results from the application of the proposed motion-based shot-class association approach are presented. In Table 2, quantitative class association results are given in the form of the calculated recognition rates when the accumulated motion energy fields, $R_{\text{acc}}(x, y, t, \tau)$, are used during the approximation step for τ = 0, 1, 2 and 3, respectively, for all selected domains. The class recognition rate is defined as the percentage of the video shots that belong to the examined class and are correctly associated with it. Additionally, the values of the overall classification accuracy and the average precision are also given. The overall classification accuracy is set equal to the percentage of all shots that are associated with the correct semantic class. On the other hand, the average precision is defined equal to the weighted sum of the estimated precision values of every supported class, using the classes' frequency of appearance as weight; the precision value of a given class is equal to the percentage of the shots that are associated with it and they truly belong to it. It has been regarded that $\arg\max_j(h_{ij}^m)$ indicates the class e_j that is associated with shot s_i. Moreover, the frame processing rate, which is defined as the number of video frames that are processed per second (fps) on average, is also given. The latter metric is introduced for approximating the computational complexity of the proposed method; a frame rate of 25 fps would indicate real-time processing for the videos used. It must be noted that all experiments were conducted using a PC with Intel Quad Core processor at 2.4 GHz and a total of 3 GB RAM.

From the presented results, it can be seen that the proposed approach achieves high values for both overall classification accuracy and average precision for τ = 0 in all selected domains, while most of the supported classes exhibit increased recognition rates. It is also shown that the class association performance generally increases when the $R_{\text{acc}}(x, y, t, \tau)$ are used for small values of τ, compared to the case where no motion information from previous frames is utilized, i.e. when τ = 0. Specifically, a maximum increase, up to 5.46% in the news domain, is observed in the overall class association accuracy when τ = 1. On the other hand, it can be seen that when the value of τ is further increased (τ = 2, 3), the overall performance improvement decreases. This is mainly due to the fact that when taking into account information from many previous frames the estimated

Table 2 Semantic class association results based on motion information.

Domain	Measure		R_{acc} for τ = 0	R_{acc} for τ = 1	R_{acc} for τ = 2	R_{acc} for τ = 3	Method of [33]	Method of [12]	Method of [32]	Method of [27]	Method of [26]
Tennis	Class recognition rate (%)	e_1	**99.28**	97.83	98.55	97.83	94.93	**98.55**	89.13	92.03	97.83
		e_2	75.00	**78.85**	71.16	73.08	61.54	**73.08**	34.62	34.62	46.15
		e_3	34.78	**43.48**	39.13	34.78	**52.17**	34.78	21.74	**52.17**	47.83
		e_4	60.19	**68.25**	64.93	65.88	**75.83**	57.82	54.50	40.76	63.98
	Overall accuracy (%)		73.35	**77.83**	75.24	75.47	**79.01**	71.70	61.56	57.31	71.93
	Average precision (%)		77.33	**80.54**	79.88	79.03	**81.33**	69.88	68.57	70.25	70.78
	Frame processing rate (fps)		5.44	5.34	5.33	5.32	1.09	4.89	5.29	5.51	0.91
News	Class recognition rate (%)	e_1	84.75	83.05	**86.44**	83.05	**94.92**	86.44	77.97	83.05	69.49
		e_2	67.86	75.00	75.00	**78.57**	**71.43**	57.14	21.43	42.86	25.00
		e_3	76.44	**83.91**	81.03	81.03	**85.63**	66.67	58.62	48.85	68.97
		e_4	56.25	**62.50**	56.25	56.25	**62.50**	56.25	53.13	50.00	**62.50**
	Overall accuracy (%)		75.09	**80.55**	78.84	78.50	**83.62**	68.60	58.36	55.29	64.16
	Average precision (%)		81.02	**84.09**	81.94	82.46	**85.83**	67.02	57.44	68.52	63.09
	Frame processing rate (fps)		5.60	5.49	5.49	5.48	1.03	5.10	5.54	5.70	0.98
Volleyball-I	Class recognition rate (%)	e_1	**90.73**	**90.73**	87.42	87.42	94.70	87.42	23.18	72.19	**98.01**
		e_2	70.27	81.08	**82.43**	78.38	**85.14**	71.62	45.95	79.73	78.38
		e_3	53.52	64.79	**77.46**	76.06	**59.15**	54.93	42.25	47.89	32.39
		e_4	**89.83**	**89.83**	83.05	83.47	**88.98**	75.42	50.42	31.78	77.97
	Overall accuracy (%)		82.52	**85.53**	83.46	82.89	**86.09**	75.56	40.98	52.07	77.63
	Average precision (%)		83.85	**87.10**	85.57	84.84	**87.33**	77.83	48.51	62.51	76.59
	Frame processing rate (fps)		6.04	6.03	6.02	5.96	1.02	5.50	5.94	6.14	0.91
Volleyball-II	Class recognition rate (%)	e_1	84.43	88.52	**91.80**	86.07	94.26	**95.08**	24.59	48.36	90.16
		e_2	**88.24**	**88.24**	82.35	82.35	58.82	29.41	**70.59**	17.65	23.53
		e_3	76.67	**83.33**	**83.33**	71.67	**85.00**	66.67	60.00	28.33	80.00
		e_4	47.37	**57.89**	47.37	**52.63**	52.63	26.32	15.79	36.84	21.05
		e_5	74.65	**78.87**	69.01	77.46	**60.56**	47.89	22.54	12.68	25.35
		e_6	63.83	54.26	62.77	**65.96**	**82.98**	72.34	6.38	42.55	50.00
		e_7	52.63	**63.16**	59.65	54.39	**57.89**	19.30	12.28	22.81	24.56
		e_8	42.86	**47.62**	42.86	42.86	**52.38**	19.05	9.52	19.05	23.81
		e_9	51.95	**62.34**	57.14	59.74	**48.05**	25.97	7.79	7.79	33.77
	Overall accuracy (%)		67.84	**71.56**	70.63	69.70	**72.11**	56.32	21.93	29.37	51.30
	Average precision (%)		70.03	**75.66**	71.20	71.70	**75.22**	56.94	30.50	37.40	51.85
	Frame processing rate (fps)		6.04	6.03	6.02	5.96	1.02	5.50	5.94	6.14	0.91

Numbers in bold indicate the best performance among the considered methods, according to a given measure.

$R_{acc}(x, y, t, \tau)$ fields for each frame tend to become very similar. Thus, polynomial coefficients tend to also have very similar values and hence HMMs cannot observe a characteristic sequence of features that unfolds in time for every supported semantic class. The above results demonstrate that the proposed accumulated motion energy fields can lead to improved shot-class association performance.

The performance of the proposed method is compared with the motion representation approaches for providing motion information to HMM-based systems presented in [12,26,27,32], as well as with the authors' previous work [33] (as described in Section 3.4). Specifically, Huang et al. consider the first four dominant motion vectors and their appearance frequencies, along with the mean and the standard deviation of motion vectors in the frame [12]. Additionally, Gibert et al. make use of the available motion vectors for estimating the principal motion direction of every frame [32]. On the other hand, Xie et al. calculate the motion intensity at frame level [27], while Xu et al. estimate the energy redistribution for every frame and subsequently a set of motion filters are applied for detecting the observed dominant motions [26]. From the presented results, it is shown that the proposed approach outperforms the algorithms of [12,26,27,32] for most of the supported classes as well as in overall classification performance in all selected domains. On the other hand, it can also be seen that the performance of the proposed approach is comparable with the one attained by the application of the method of [33] (note that the results for the method of [33] and other works that are reported in Table 2 may be somewhat different from those reported in [33], in absolute numbers; this is due to the datasets used in [33] being different than those used here). In particular, the method of [33] presents higher overall classification accuracy and average precision in the ranges [0.55, 3.07%] and [0.23, 1.74%], respectively, in the selected domains. However, it is shown that the proposed method performs faster than the method of [33] by a factor in the range [4.90, 5.91], while its time performance is also comparable or better than that of [12,26,27,32] that were implemented; all the latter methods exhibit considerably lower overall classification performance in all domains. Thus, the proposed motion-based shot-class association approach achieves to combine increased recognition performance with relatively low computational complexity, compared to the relevant literature. It must be noted that the approximation of the methods' computational complexity by the introduced frame processing rate metric is performed due to the inevitable difficulty in defining the computational complexity in a closed form for most cases (e.g. the computational complexity of the method of [33] depends heavily on the type of the videos and the kinds of the motion patterns that they contain).

6.3.1. Effect of the degree of the polynomial function

In order to investigate the effect of the introduced polynomial function's degree on the overall motion-based shot-class association performance (Section 3), the latter was evaluated when parameter T (Equation (3)) receives values ranging from 1 to 6. Additionally, the accumulated motion energy fields, $R_{acc}(x, y, t, \tau)$, are used for $\tau = 1$ in all selected domains. Values of parameter T greater than 6 resulted in significantly decreased recognition performance. The corresponding shot-class association results are illustrated in Figure 4, where it can be seen that the use of a 3rd order polynomial function leads to the best overall performance in all domains. It must be noted that for the cases of the 5th and 6th order polynomial function, Principal Component Analysis (PCA) was used for reducing the dimensionality of the observation vectors and overcoming HMM under-training occurrences. The target dimension of the PCA output was set equal to the dimension of the observation vector that is generated when using a 4th order polynomial function (i.e. the highest value of T for which HMM under-training occurrences were not observed).

6.4. Overall analysis results

In this section experimental results of the overall developed framework are presented. In order to demonstrate and comparatively evaluate the efficiency of the proposed integrated BN, the following experiments were conducted:

(1) application of the developed BN
(2) application of a variant of the proposed approach, where a SVM-based classifier is used instead of the developed BN
(3-4) application of the methods of [12] and [26]
(5-6) application of the methods of [12] and [26], using the low-level features of Sections 3 and 4

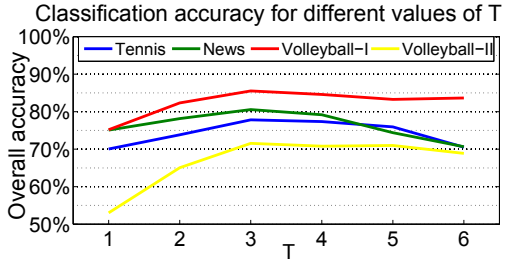

Figure 4 Motion-based semantic class association results for different values of parameter *T*.

instead of the ones originally proposed in [12] and [26].

Experiment 1 demonstrates the shot-class association performance obtained by the application of the proposed integrated BN, which jointly performs modality fusion and temporal context exploitation. Experiment 2 is conducted in order to comparatively evaluate the effectiveness of the developed BN, which constitutes a generative classifier, against a discriminative one. Discriminative classifiers are easier to be developed, while they are generally considered to outperform generative ones [55], when sufficient amount of training data is available. To this end, a variant of the proposed approach is implemented, where a SVM-based classifier is used instead of the developed BN. In particular, an individual SVM is introduced for every defined class e_j to detect the corresponding instances and is trained under the 'one-against-all' approach. Each SVM, which receives as input the same set of posterior probabilities with the developed BN (i.e. the evidence data W_i defined in Section 5.3), returns at the evaluation stage for every shot s_i a numerical value in the range [0, 1]. This value denotes the degree of confidence with which the corresponding shot is assigned to the class associated with the particular SVM (similarly to the h_{ij}^f value also defined in Section 5.3). Implementation details regarding the developed SVM-based classifier can be found in [9]. In all cases, it has been considered that $\arg\max_j(h_{ij}^a)$, $\arg\max_j(h_{ij}^m)$, $\arg\max_j(h_{ij}^m)$ and $\arg\max_j(h_{ij}^f)$ indicate the class e_j that is associated with shot s_i after every respective algorithmic step. The performance of the developed BN is also compared with the HMM-based video analysis approaches presented in [12] and [26] (experiments 3 and 4). Specifically, Huang et al. [12] propose a 'class transition penalty' approach, where HMMs are initially employed for detecting the semantic classes of concern using multi-modal information and a product fusion operator. Subsequently, a dynamic programming technique is adopted for searching for the most likely class transition path. On the other hand, Xu et al. [26] present a HMM-based framework capable of modeling temporal contextual constraints in different semantic granularities, while multistream HMMs are used for modality fusion. It must be noted that apart from the motion and color features proposed in [26] (observed dominant motions and mean RGB values, respectively), audio information is also used for the purpose of comparison in experiment 4. In particular, the MFCC coefficients (described in Section 4) are also provided as input to the employed multistream HMMs. Additionally, in order to compensate the effect of the different approaches originally using different color, motion and audio features, in experiments 5 and 6 the methods of [12] and [26] receive as input the same video low-level features utilized by the proposed method and described in Sections 3 and 4. Hence, the latter two experiments will facilitate in better demonstrating the effectiveness of the proposed BN, compared to other similar approaches that perform the modality fusion and temporal context exploitation procedures separately. It must be highlighted at this point that the method of [26] actually constitutes a particular type of L-HMMs, namely a composite HMM with 3 layers.

Results of experiments 1 and 2, which are affected by parameter TW (Equation 11), were carried out for TW between 1 and 6. In Figure 5, the results for TW = 1, 2 and 3 are reported in detail, in terms of the difference in classification accuracy compared to the best single-modality analysis result for each domain. The latter are depicted in parentheses. From these results, it can be seen that the proposed integrated BN achieves a significant increase (up to 15.80% in the volleyball-II domain) in the overall classification accuracy for all selected domains for TW = 1, while the recognition rates of most of the supported classes are substantially enhanced. Additionally, it can also be seen that further increase of the value of parameter TW (TW = 2, 3) leads to a corresponding increase of the overall classification accuracy. Among the classes that are particularly favored by the application of the proposed integrated BN are those that present significant variations in their video low-level features, while also having quite well-defined temporal context. Such classes are break and graphics in the tennis and news domain, respectively. In particular, shots belonging to the class break usually depict significantly different types of scenes (e.g. the players resting or the audience), while also having quite well-defined temporal context (video shots belonging to the class break are often successive and usually interrupted by shots depicting a serve hit). Similarly, shots belonging to the class graphics differ significantly in terms of their low-level audio-visual features (due to the different graphical environments that are presented during a news broadcast, like news start/end signals, weather maps, sport tables, etc.), while they also present characteristic temporal relations. It was observed that values of parameter TW greater than 3 (i.e. TW = 4, 5, 6) were experimentally shown to result into marginal changes in the overall classification performance (i.e. changes in the overall accuracy smaller than 0.10%) and negligible variations in the classes' recognition rates; these results are not included in Figure 5 for brevity. All the above results demonstrate the potential of reaching increased shot-class association results by jointly

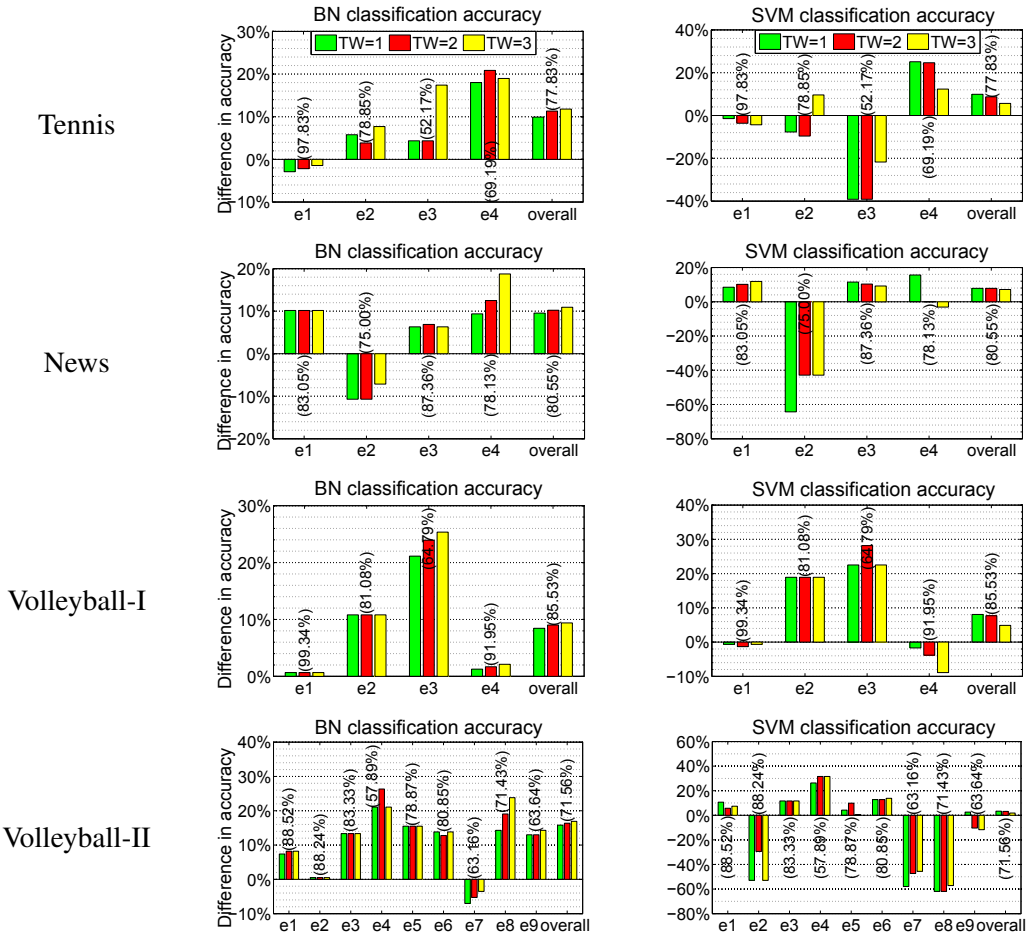

Figure 5 Impact of the value of parameter *TW* (time window *TW* = 1, 2 and 3) on the results of joint modality fusion and temporal context exploitation, when using the developed BN (first column of sub-figures) or an SVM classifier (second column). In all sub-figures, the vertical bars indicate the difference in classification accuracy compared to the best single-modality analysis result for each domain; the latter are given in parentheses.

performing modality fusion and temporal context exploitation.

Considering the corresponding SVM results (experiment 2), it is shown in Figure 5 that a significant increase (up to 9.91% in the tennis domain) in the overall classification accuracy can also be obtained for TW = 1 compared to the best single-modality analysis result, when a SVM-based classifier is used instead of the developed BN for all domains. This is lower or equal to the corresponding results of the BN for TW = 1, with the highest difference of approximately 12.46% being observed in the volleyball-II domain, i.e. the domain with the highest number of supported semantic classes. Additionally, two important observations can be made: (a) the overall performance improvement decreases when parameter TW receives greater values (TW = 2, 3), as opposed to the results of experiment 1, and (b) not all supported classes are favored (e.g. reporting exhibits a dramatic decrease of 64.29% in its recognition rate for TW = 1 in the news domain). These observations suggest that the methodology proposed in this work for representing and learning the joint probability distribution $P_{\text{joint}}(\{a_j^k, c_j^k, m_j^k, cl_j^k\}_{1\le j\le J}^{i-TW\le k\le i+TW})$ (Section 5.3) is advantageous compared to directly modeling the probability distributions $P(CL_j^i = True|W_i)$, as the employed SVM-based classifier does. This observation can be considered as an extension of the findings presented by Adams et al. in [17], where BNs and SVMs were experimentally shown to be equally efficient for the task of modality fusion.

In Table 3, quantitative class association results are given for experiments 1-6, as well as from every separate modality processing, in the form of the calculated recognition rates for all selected domains. The values of the overall classification accuracy and the average precision

Table 3 Semantic class association results using multiple modalities and temporal context.

Domain	Measure		Audio	Motion	Color	Integrated BN for TW = 3 (1)	SVM for TW = 1 (2)	Method of [12] (3)	Method of [26] (4)	Method of [12] using proposed features (5)	Method of [26] using proposed features (6)
Tennis	Class recognition rate (%)	e_1	80.43	**97.83**	92.03	96.38	96.38	92.75	84.78	93.48	88.41
		e_2	13.46	78.85	76.92	**86.54**	71.15	53.85	28.85	67.31	57.69
		e_3	17.39	43.48	52.17	**69.57**	13.04	21.74	17.39	21.74	52.17
		e_4	40.76	68.25	69.19	88.15	94.31	88.15	**97.63**	88.63	90.52
	Overall accuracy (%)		49.06	77.83	76.65	**89.62**	87.74	81.84	80.66	83.96	83.73
	Average precision (%)		54.98	80.54	85.02	**90.75**	87.41	82.22	81.96	84.38	83.86
News	Class recognition rate (%)	e_1	72.88	83.05	54.24	**93.22**	91.53	52.54	54.24	59.32	77.97
		e_2	67.86	**75.00**	14.29	67.86	10.71	7.14	21.43	14.29	21.43
		e_3	62.64	83.91	87.36	93.68	98.85	**99.43**	97.13	**99.43**	93.10
		e_4	12.50	62.50	78.13	**96.88**	93.75	81.25	81.25	81.25	84.38
	Overall accuracy (%)		59.73	80.55	72.70	**91.47**	88.40	79.18	79.52	81.23	82.25
	Average precision (%)		71.24	84.09	76.67	**91.58**	86.03	81.51	80.88	83.75	81.50
Volleyball-I	Class recognition rate (%)	e_1	68.87	90.73	99.34	**100.00**	98.68	80.13	90.73	97.35	98.01
		e_2	64.86	81.08	58.11	91.89	**100.00**	68.92	83.78	85.14	89.19
		e_3	12.68	64.79	59.15	**90.14**	87.32	15.49	21.13	33.80	35.21
		e_4	63.56	89.83	91.95	94.07	90.25	**97.88**	96.19	94.49	94.92
	Overall accuracy (%)		58.46	85.53	84.96	**94.92**	93.61	77.82	82.89	85.90	87.03
	Average precision (%)		59.90	87.10	84.81	**95.14**	94.24	81.78	85.30	85.38	87.04
Volleyball-II	Class recognition rate (%)	e_1	74.59	88.52	81.15	96.72	**99.18**	**99.18**	96.72	**99.18**	**99.18**
		e_2	41.18	**88.24**	58.82	**88.24**	35.29	23.53	23.53	64.71	23.53
		e_3	61.67	83.33	81.67	**96.67**	95.00	**96.67**	90.00	95.00	90.00
		e_4	36.84	57.89	47.37	78.95	**84.21**	68.42	57.89	78.95	52.63
		e_5	19.72	78.87	76.06	**94.37**	83.10	90.14	78.87	85.92	76.06
		e_6	79.79	54.26	80.85	94.68	93.62	93.62	91.49	**96.81**	**96.81**
		e_7	29.82	**63.16**	35.09	59.65	5.26	33.33	29.82	28.07	42.11
		e_8	57.14	47.62	71.43	**95.24**	9.52	57.14	19.05	61.90	28.57
		e_9	22.08	62.34	63.64	**77.92**	66.23	55.84	46.75	75.32	55.84
	Overall accuracy (%)		51.49	71.56	70.82	**88.48**	74.91	78.44	71.75	82.34	75.65
	Average precision (%)		57.18	75.66	73.75	**88.86**	78.74	77.86	66.19	82.49	72.65

Numbers in bold indicate the best performance among the considered methods, according to a given measure.

are also given for every case. It must be noted that a time-performance measure (similar to the average frame processing rate defined in Section 6.3) is not included in Table 3. This is due to the fact that the execution of any of the modality fusion and temporal context exploitation methods reported in experiments 1-6 represents a very small fraction (less than 2%) of the overall video processing time. The latter essentially corresponds to the generation of the respective single-modality analysis results. Following the discussion on Figure 5, only the best results of experiments 1 and 2 are reported here, i.e. using $TW = 3$ for the BN and $TW = 1$ for the SVM-based classifier. It can be seen that the proposed BN outperforms the SVM-based approach as well as the methods of [12] and [26] for most of the supported classes as well as in overall classification performance. Additionally, it is also advantageous compared to the case where the methods of [12] and [26] utilize the video low-level features described in Sections 3 and 4 (experiments 5 and 6). This is mainly due to: (a) the more sophisticated modality fusion mechanism developed in this work, compared to the heuristic assignment of weights to every modality in [26] and the assumption of statistical independence between the features of different modalities in [12], (b) the more complex temporal relations that are modeled by the developed integrated BN, compared to the methods of [26] and [12] that rely on class transition probability learning, and (c) the fact that the proposed method performs jointly modality fusion and temporal context exploitation; hence, taking advantage of all possible correlations between the respective numerical data. It must be emphasized here that these results verify the theoretic analysis given in Section 5.4, which indicated that the proposed integrated BN was expected to outperform other similar HMM-based approaches, e.g. [26].

In order to investigate whether the employed datasets are sufficiently large for the differences in performance observed in Table 3 to be statistically significant, a statistical significance test is used. This takes into account the overall shot classification accuracy in each selected domain and uses the chi-square measure [56] together with the following null hypothesis: "there is no significant difference in the total number of correctly classified shots between the results obtained after the application of the BN and the results obtained after the application of another similar approach of the literature". The latter is the hypothesis that is to be rejected if the test is passed. The test revealed that all performance differences observed in Table 3 between the proposed approach and the methods of [26] and [12] (using either their original features or the low-level features proposed in Sections 3 and 4) are statistically significant. In particular, the lowest chi-square values calculated for the tennis, news, volleyball-I and volleyball-II domains according to the aforementioned pairwise method comparisons are as follows: (Chi-square = 10.09, df = 1, $P < 0.05$), (Chi-square = 17.06, df = 1, $P < 0.05$), (Chi-square = 29.34, df = 1, $P < 0.05$) and (Chi-square = 13.95, df = 1, $P < 0.05$). Regarding the comparison with the SVM-based method (experiment 2), the difference in performance is statistically significant for the challenging volleyball-II domain (Chi-square = 6.96, df = 1, $P < 0.05$). For the other three datasets that involve only 4 classes, less pronounced performance differences (thus also of lower statistical significance) are observed between the proposed approach and the SVM one. However, it should be noted that: (a) despite the small difference in overall performance, the SVM classifier often introduces a dramatic decrease in the recognition rate of some of the supported semantic classes, as discussed earlier in this section, and (b) the SVM classifier, as applied in this work, constitutes a variation of the proposed approach, i.e. its performance is also boosted by jointly realizing modality fusion and temporal context exploitation, as opposed to the literature works of [26] and [12].

6.4.1. Effect of discretization

In order to examine the effect of the proposed discretization procedure on the performance of the developed integrated BN, the latter was evaluated for different values of parameter Q (Equation (6)). This parameter determines the number of possible values of random variables A_j, C_j and M_j in the $\mathbb{G}_j$ BN sub-structure. Results when parameter Q receives values in the interval [3,15] are illustrated in Figure 6. It can be seen that the developed BN tends to exhibit relatively decreased recognition performance, when parameter Q receives low values ($Q\varepsilon[3,6]$) for $TW = 1, 2, 3$ in all domains. This is mainly due to the fact that low values of Q led to coarse discretization, which resulted to decreased shot-class association performance. Additionally, when Q receives values ranging approximately from 7 to 11, the proposed approach presents relatively small variations in its recognition performance, which is close to its maximum overall shot-class association accuracy for any value of TW and in all domains. On the other hand, values greater than 11 led to increased network complexity and resulted to under-training/overfitting occurrences; hence, leading to a corresponding gradual decrease in the overall shot-class association performance.

7. Conclusions

In this paper, a multi-modal context-aware framework for semantic video analysis was presented. The core functionality of the proposed approach relies on the introduced integrated BN, which is developed for

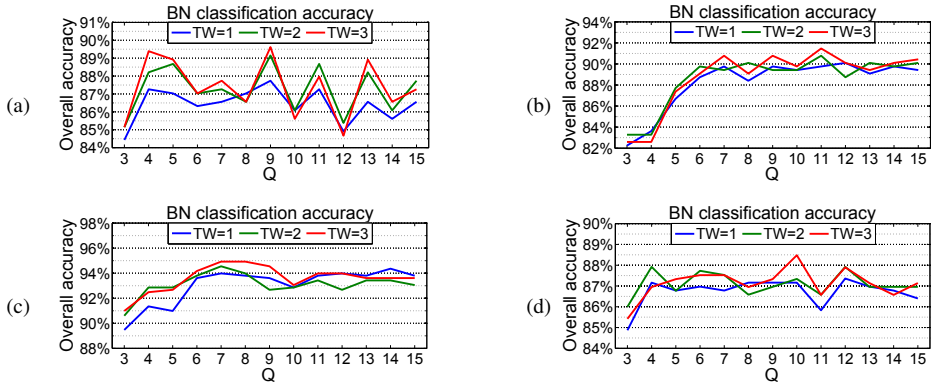

Figure 6 BN shot classification results for different values of parameter *Q* in the (a) tennis, (b) news, (c) volleyball-I and (d) volleyball-II domain.

performing joint modality fusion and temporal context exploitation. With respect to the utilized motion features, a new representation for providing motion energy distribution-related information to HMMs is described, where motion characteristics from previous frames are exploited. Experimental results in the domains of tennis, news and volleyball broadcast video demonstrated the efficiency of the proposed approaches. Future work includes the examination of additional shot-class association schemes as well as the investigation of alternative algorithms for acquiring and modeling contextual information, and their integration in the proposed framework.

Endnotes

[a] http://mklab.iti.gr/project/svaef.

Acknowledgements

The work presented in this paper was supported by the European Commission under contracts FP7-248984 GLOCAL and FP7-249008 CHORUS +.

Author details

[1]CERTH/Informatics and Telematics Institute 6th Km. Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece [2]Electrical and Computer Engineering Department of Aristotle University of Thessaloniki, Thessaloniki, Greece

Competing interests

The authors declare that they have no competing interests.

Received: 3 November 2010 Accepted: 13 October 2011
Published: 13 October 2011

References

1. A Hanjalic, R Lienhart, W Ma, J Smith, The holy grail of multimedia information retrieval: so close or yet so far away? Proc IEEE. **96**(4), 541–547 (2008)
2. A Smeaton, Techniques used and open challenges to the analysis, indexing and retrieval of digital video. Inf Syst. **32**(4), 545–559 (2007). doi:10.1016/j.is.2006.09.001
3. W Zhu, C Toklu, S Liou, Automatic news video segmentation and categorization based on closed-captioned text. IEEE International Conference on Multimedia and Expo (ICME) 829–832 (2001)
4. HL Wang, L-F Cheong, Taxonomy of directing semantics for film shot classification. IEEE Trans Circuits Syst Video Technol. **19**(10), 1529–1542 (2009)
5. C Snoek, M Worring, Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools Appl. **25**(1), 5–35 (2005)
6. Y Wang, Z Liu, J Huang, Multimedia content analysis-using both audio and visual clues. Signal Process Mag IEEE. **17**(6), 12–36 (2000). doi:10.1109/79.888862
7. J Luo, M Boutell, C Brown, Pictures are not taken in a vacuum. Signal Process Mag IEEE. **23**(2), 101–114 (2006)
8. D Vallet, P Castells, M Fernandez, P Mylonas, Y Avrithis, Personalized content retrieval in context using ontological knowledge. IEEE Trans Circuits Syst Video Technol. **17**(3), 336 (2007)
9. GT Papadopoulos, V Mezaris, I Kompatsiaris, MG Strintzis, Combining global and local information for knowledge-assisted image analysis and classification. EURASIP J Adv Signal Process. **2007**(2) (2007)
10. D Byrne, P Wilkins, G Jones, A Smeaton, NO?'? Connor, Measuring the impact of temporal context on video retrieval. in *Proceedings of International Conference on Content-Based Image and Video Retrieval* 299–308 (2008)
11. L Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. **77**(2), 257–286 (1989). doi:10.1109/5.18626
12. J Huang, Z Liu, Y Wang, Joint scene classification and segmentation based on hidden Markov model. IEEE Trans Multimedia. **7**(3), 538–550 (2005)
13. J Zhou, X-P Zhang, An ica mixture hidden markov model for video content analysis. IEEE Trans Circuits Syst Video Technol. **18**(11), 1576–1586 (2008)
14. X Gao, Y Yang, D Tao, X Li, Discriminative optical flow tensor for video semantic analysis. Comput Vis Image Underst. **113**(3), 372–383 (2009). doi:10.1016/j.cviu.2008.08.007
15. R Neapolitan, *Learning Bayesian Networks* (Prentice Hall Upper Saddle River, NJ, 2003)
16. D Heckerman, A tutorial on learning with Bayesian networks, *Learning in graphical models* (MIT Press Cambridge, MA, 1998)
17. W Adams, G Iyengar, C Lin, M Naphade, C Neti, H Nock, J Smith, Semantic indexing of multimedia content using visual, audio, and text cues. EURASIP J Appl Signal Process. **2**, 170–185 (2003)
18. M-H Hung, C-H Hsieh, Event detection of broadcast baseball videos. IEEE Trans Circuits Syst Video Technol. **18**(12), 1713–1726 (2008)
19. Y Gong, W Xu, *Machine Learning for Multimedia Content Analysis* (Springer, New York, 2007)
20. E Bruno, N Moenne-Loccoz, S Marchand-Maillet, Design of multimodal dissimilarity spaces for retrieval of video documents. IEEE Trans Pattern Anal Mach Intell. **30**(9), 1520–1533 (2008)
21. M Shyu, Z Xie, M Chen, S Chen, Video semantic event/concept detection using a subspace-based multimedia data mining framework. IEEE Trans Multimedia. **10**(2), 252–259 (2008)
22. S Hoi, M Lyu, A multimodal and multilevel ranking scheme for large-scale video retrieval. IEEE Trans Multimedia. **10**(4), 607–619 (2008)

23. D Tjondronegoro, Y Chen, Knowledge-discounted event detection in sports video. IEEE Trans Syst Man Cybern Part A Syst Hum. **40**(5), 1009–1024 (2010)
24. C Xu, J Wang, L Lu, Y Zhang, A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Trans Multimedia. **10**(3), 421–436 (2008)
25. J Yang, A Hauptmann, Exploring temporal consistency for video analysis and retrieval. in *Proceedings of ACM International Workshop on Multimedia Information Retrieval* 33–42 (2006)
26. G Xu, Y Ma, H Zhang, S Yang, An HMM-based framework for video semantic analysis. IEEE Trans Circuits Syst Video Technol. **15**(11), 1422–1433 (2005)
27. L Xie, P Xu, S Chang, A Divakaran, H Sun, Structure analysis of soccer video with domain knowledge and hidden Markov models. Pattern Recognit Lett. **25**(7), 767–775 (2004). doi:10.1016/j.patrec.2004.01.005
28. K Wan, Exploiting story-level context to improve video search. IEEE International Conference on Multimedia and Expo (ICME) 289–292 (April 2008)
29. W Hsu, L Kennedy, S Chang, Video search reranking through random walk over document-level context graph. in *Proceedings of International Conference on Multimedia* 971–980 (2007)
30. J You, G Liu, A Perkis, A semantic framework for video genre classification and event analysis. Signal Process Image Commun. **25**(4), 287–302 (2010). doi:10.1016/j.image.2010.02.001
31. Y Ding, G Fan, Sports video mining via multichannel segmental hidden Markov models. IEEE Trans Multimedia. **11**(7), 1301 (2009)
32. X Gibert, H Li, D Doermann, Sports video classification using HMMs. IEEE Int Conf Multimedia Expo (ICME). **2**, 345–348 (2003)
33. GT Papadopoulos, A Briassouli, V Mezaris, I Kompatsiaris, MG Strintzis, Statistical motion information extraction and representation for semantic video analysis. IEEE Transactions Circuits Syst Video Technol. **19**(10), 1513–1528 (2009)
34. V Kobla, D Doermann, K Lin, Archiving, indexing, and retrieval of video in the compressed domain. in *Proceedings of SPIE Conference on Multimedia Storage Archiving Systems*. **2916**, 78–89 (1996)
35. B Lucas, T Kanade, An iterative image registration technique with an application to stereo vision. in *International Joint Conference on Artifical Intelligence*. **3**, 674–679 (1981)
36. M Geetha, S Palanivel, HMM Based Automatic Video Classification Using Static and Dynamic Features. in *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA)* 277–281 (2007)
37. Z Xiong, R Radhakrishnan, A Divakaran, T Huang, Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. IEEE International Conference on Multimedia and Expo (ICME). **3** (2003)
38. C Cheng, C Hsu, Fusion of audio and motion information on HMM-based highlight extraction for baseball games. IEEE Trans Multimedia. **8**(3), 585–599 (2006)
39. M Kolekar, S Sengupta, A hierarchical framework for generic sports video classification. Comput Vis ACCV. **3852**, 633–642 (2006). doi:10.1007/11612704_63
40. S Ikbal, T Faruquie, HMM based event detection in audio conversation. in *Proceedings of IEEE International Conference on Multimedia and Expo, IEEE* 1497–1500 (2008)
41. B Zhang, W Dou, L Chen, Audio content-based highlight detection using adaptive Hidden Markov Model. International Conference on Intelligent Systems Design and Applications (2006)
42. D Zhang, D Gatica-Perez, S Bengio, I McCowan, Semi-supervised adapted hmms for unusual event detection. IEEE Comput Soc Conf Comput Vis Pattern Recognit. **1**, 611–618 (2005)
43. J Wang, E Chng, C Xu, H Lu, Q Tian, Generation of personalized music sports video using multimodal cues. IEEE Trans Multimedia. **9**(3), 576–588 (2007)
44. M Xu, L Chia, J Jin, Affective content analysis in comedy and horror videos by audio emotional event detection. in *IEEE International Conference on Multimedia and Expo* 4 (2005)
45. C Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006)
46. M Petkovic, V Mihajlovic, W Jonker, S Djordjevic-Kajan, Multi-modal extraction of highlights from TV formula 1 programs. in *IEEE International Conference on Multimedia and Expo (ICME)* (2002)
47. B Liang, S Lao, W Zhang, G Jones, AF Smeaton, *Video Semantic Content Analysis Framework Based on Ontology Combined MPEG-7, vol. 4918/2008*, (Springer, Berlin/Heidelberg, 2008), pp. 237–250
48. M Barnard, J Odobez, Sports Event Recognition Using Layered HMMs. IEEE International Conference on Multimedia and Expo (ICME) 1150–1153 (2005)
49. M Brand, Voice puppetry, in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, (ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1999), pp. 21–28
50. S Fine, Y Singer, N Tishby, The hierarchical hidden Markov model: analysis and applications. Mach Learn. **32**(1), 41–62 (1998). doi:10.1023/A:1007469218079
51. N Oliver, E Horvitz, A Garg, Layered representations for human activity recognition. in *Fourth IEEE International Conference on Multimodal Interfaces*. **3**(8) (2002)
52. S Davis, P Mermelstein, Comparison of parametric representations for monosylabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process. **28**(10), 357–366 (1980)
53. *Hidden Markov Model Toolkit, HTK* http://htk.eng.cam.ac.uk/
54. FV Jensen, F Jensen, Optimal junction trees. in *Proceedings of Conference on Uncertainty in Artificial Intelligence* (1994)
55. A Ng, M Jordan, On discriminative versus generative classifiers: a comparison of logistic regression and naive bayes. Adv Neural Inf Process Syst. **2**, 841–848 (2002)
56. P Greenwood, M Nikulin, *A guide to chi-squared testing*, (Wiley-Interscience, 1996)

doi:10.1186/1687-6180-2011-89
Cite this article as: Papadopoulos *et al.*: **Joint modality fusion and temporal context exploitation for semantic video analysis.** *EURASIP Journal on Advances in Signal Processing* 2011 **2011**:89.