Accumulated Motion Energy Fields Estimation and Representation for Semantic Event Detection

G. Th. Papadopoulos[1,2], V. Mezaris[2], I. Kompatsiaris[2] and M. G. Strintzis[1,2]

[1]Information Processing Laboratory
Electrical & Computer Eng. Dep.
Aristotle University of Thessaloniki
Thessaloniki, GR-54124, Greece
strintzi@eng.auth.gr

[2]Informatics and Telematics Institute
Centre for Research and Technology Hellas
1st Km Thermi-Panorama Road
Thessaloniki, GR-57001, Greece
{papad, bmezaris, ikom}@iti.gr

ABSTRACT

In this paper, a motion-based approach for detecting high-level semantic events in video sequences is presented. Its main characteristic is its generic nature, i.e. it can be directly applied to any possible domain of concern without the need for domain-specific algorithmic modifications or adaptations. For realizing event detection, the video is initially segmented into shots and for every resulting shot appropriate motion features are extracted at fixed time intervals, thus forming a *motion observation sequence*. Then, Hidden Markov Models (HMMs) are employed for associating each shot with a semantic event based on its formed observation sequence. Regarding the motion feature extraction procedure, a new representation for providing local-level motion information to HMMs is presented, while motion characteristics from previous frames are also exploited. The latter is based on the observation that motion information from previous frames can provide valuable cues for interpreting the semantics present in a particular frame. Experimental results as well as comparative evaluation from the application of the proposed approach in the domains of tennis and news broadcast video are presented.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Concept learning; I.2.10 [**Vision and Scene Understanding**]: Video analysis; I.4.8 [**Scene Analysis**]: Motion

General Terms

Algorithms, Experimentation

Keywords

Hidden Markov Models, event detection, motion representation, accumulated motion fields

CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.
Copyright 2008 ACM 978-1-60558-070-8/08/07 ...$5.00.

1. INTRODUCTION

Given 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 in the development of sophisticated and user-friendly systems for skilful management of video sequences [3][6]. To this end, several approaches have been proposed in the literature regarding the tasks of indexing, searching, summarization and retrieval of video content [23][15].

Most recent approaches adopt the fundamental principle of shifting video manipulation techniques towards the processing of the visual content at a semantic level, thus attempting to bridge the so called *semantic gap* [22] and efficiently capture the underlying semantics of the content. Among these methodologies, approaches that exploit *a priori* knowledge have been particularly favored and have so far exhibited promising results. Prior knowledge, when used, guides low-level features extraction and facilitates high-level description derivation and semantic inference.

Depending on the adopted knowledge acquisition process, knowledge-assisted techniques are mainly divided into two categories, namely model-based and Machine Learning (ML)-based approaches. Model-based techniques make use of explicitly defined axioms, facts and rules, which are stored in appropriate knowledge structures such as ontologies and semantic nets. In [8], an ontology framework is proposed for detecting events in video sequences, based on the notion that complex events are constructed from simpler ones by operations such as sequencing, iteration and alternation. A large-scale concept ontology for multimedia (LSCOM) is designed in [17] to simultaneously cover a large semantic space and increase observability in diverse broadcast news video data sets. On the other hand, ML-based approaches utilize probabilistic methods for acquiring the appropriate implicit knowledge that will enable the mapping of the low-level audio-visual data to high-level semantic concepts and entities. In [28], a HMM-based framework is proposed, which models semantics in different levels of semantic granularity and supports the decomposition of complex analysis problems into simpler sub-problems. Moreover, in [20], Support Vector Machines (SVMs), which perform on top of specific feature detectors, are employed for detecting semantically meaningful events in broadcast video of multiple field sports. Although many methods have already been presented for realizing knowledge-assisted video analysis, most of them are only limited to domain specific applications, i.e. they exploit specific facts and characteristics that are only present in the examined domain, thus failing to effectively handle the problem of semantic video analysis at a more generic level.

In this paper, a motion-based approach for detecting high-level semantic events in video sequences, while making use of ML algorithms for implicit knowledge acquisition, is presented. On the contrary to the majority of the methods present in the relevant literature, its main characteristic is its generic nature, i.e. it can be directly applied to any possible domain of concern without the need for domain-specific algorithmic modifications or adaptations. In particular, only the high-level semantic events of concern need to be defined and a corresponding set of annotated video content needs to be provided for training purposes. The former represent semantically meaningful incidents that are of interest in a possible application case and have a temporal duration. For realizing event detection, the examined video sequence is initially segmented into shots. For every shot appropriate motion features are extracted at fixed time intervals, thus forming a *motion observation sequence*. Then, HMMs are employed for performing the association of each shot with one of the supported events based on its formed observation sequence. Prior to this, the provided set of annotated video content is used for training the utilized HMMs, i.e. for acquiring the appropriate implicit knowledge that will enable the mapping of the low-level audio-visual data to the defined high-level semantic events. Regarding the motion feature extraction procedure, unlike the majority of the approaches of the relevant literature, local-level analysis is supported for efficiently capturing the semantics of the visual medium. This is based on a new representation for providing local-level motion information to HMMs and is different in nature from the 'points of interest'-based local-level analysis commonly used for dedicated tasks like human action recognition [21]. Furthermore, motion characteristics from previous frames are additionally exploited. This is based on the observation that motion information from previous frames can provide valuable cues for interpreting the semantics present in a particular frame and results into an *accumulated motion field*. The final outcome of the overall video analysis framework outlined above is a semantic event associated with every shot, to which the examined video is decomposed. Experimental results as well as comparative evaluation from the application of the proposed approach in the domains of tennis and news broadcast video are presented.

The paper is organized as follows: Section 2 presents an overview of the relevant literature. Section 3 describes how HMMs are employed for realizing semantic event detection. The video pre-processing steps are described in Section 4. Section 5 outlines the proposed local-level motion representation and Section 6 details the methodology followed for incorporating motion characteristics from previous frames. Experimental results are presented in Section 7 and conclusions are drawn in Section 8.

2. RELATED WORK

HMMs have been widely used in speech recognition systems, due to their reported capability in modeling pattern recognition problems that exhibit an inherent temporality [19]. The wide variety of approaches where HMMs have been utilized include [25], where an HMM-based system is developed for recognizing distant-talking speech, [5], where a framework of online hierarchical transformation of HMM parameters is proposed for adaptive speech recognition, [24], where a weighted HMM and a subspace projection algorithm are proposed to address the discrimination and robustness issues for HMM-based speech recognition, and [29], where a scalable architecture for realizing real-time speech recognizers is presented.

Many recent research efforts in the field of semantic video analysis have also adopted the HMM theory as well in an attempt to benefit from the significant advantages and characteristics that HMMs present. In [11], a HMM-based system is proposed for performing joint scene classification and video temporal segmentation. In [9], a comparative study of three individual approaches for solving the problem of audio/visual mapping with the usage of HMMs is presented. Wang et al. proposes a multi-level framework to automatically recognize the genre of sports video [26]. Additionally, in [10], a HMM-based system is presented for categorizing a video sequence into one of a set of predefined sport classes.

An important topic in video analysis tasks is the detection of high-level semantic events, i.e. incidents that are of increased semantical importance and which can be used for realizing efficient and effective video indexing, searching and retrieval. Thus, increased research activity has been devoted in the development of appropriate systems for accurate and robust detection of semantic events in videos for several application cases. In [4], a HMM-based system is developed for extracting highlights from baseball game videos. An approach that supports the detection of events such as 'foul' and 'shot at the basket' in basketball videos is presented in [16]. Additionally, Gaussian Mixture Hidden Markov Models (GMHMMs) are used in [14] for identifying traffic events. In [28], a HMM-based framework is presented, where events are detected under the fundamental principle of decomposing a complex analysis problem into simpler sub-problems and automatically integrating those subproblems for recognition. Moreover, HMMs are used in [27] for identifying the events 'play' and 'break' in soccer videos. Additional approaches include the works of [2] and [12], where layered HMMs and a priori domain specific information is integrated into HMMs, respectively.

Despite the plurality of the proposed approaches and the significant results that have already been presented, the majority of the developed algorithms are domain specific and the need for a generic approach arises as a challenging research issue. Moreover, motion analysis is mainly limited to global or camera motion level, which is not always adequate. In the following sections the individual steps of the proposed method which aims to overcome the aforementioned limitations are presented in detail.

3. HIDDEN MARKOV MODELS

HMMs constitute a powerful statistical tool for solving problems that have an inherent temporality, i.e. they consist of a process that unfolds in time [19][7]. The fundamental idea is that every process is made of a set of internal states and every state generates an observation when the process lies in that state. Thus, the sequential transition of the process among its constituent states generates a corresponding observation sequence. The latter is characteristic for every different process. It must be noted that a HMM requires a set of suitable training data for adjusting its internal structure, i.e. for efficiently modeling the process with which it is associated. At the evaluation stage, a HMM, which receives as input a possible observation sequence, estimates a posterior probability, which denotes the fitness of the input sequence to that model.

Under the proposed approach, HMMs are employed for detecting high-level semantic events in video sequences. In accordance to the HMM theory, each event corresponds to a process that is to be modeled by an individual HMM and the features extracted from the video stream constitute the respective observation sequences. More specifically, the first step in the development of the proposed video analysis framework is the definition of a set of high-level semantic events, denoted by $E = \{e_j,\ j = 1, ..J\}$. The latter represent semantically meaningful incidents that are of interest in a possible application case and have a temporal duration. A set of annotated video content, denoted by U_{tr}, is used for training the utilized HMMs, while a similar set, denoted by U_{te}, is formed for the subsequent evaluation stage.

4. VIDEO PRE-PROCESSING

At the signal level, the examined video sequence is initially segmented into shots, which constitute the elementary image sequences of video. For shot detection the algorithm of [13] is used, mainly due to its low computational complexity. Output of the segmentation algorithm is a set of shots, denoted by $S = \{s_i, i = 1, ...I\}$, to which the examined video is decomposed. Under the proposed approach each shot will be associated with one of the supported events, e_j, on the basis of its semantic contents.

After the examined video is segmented into shots, a set of frames are selected at equally spaced time intervals for each shot s_i starting with the first frame of it. The time interval between two sequentially selected frames, i.e. the temporal sampling frequency, is denoted by SF_t. Then, a dense motion field is estimated for every selected frame. The optical flow estimation algorithm of [18] was used for computing this dense motion field, since satisfactory results can be obtained by its application in a variety of motion estimation cases. From the computed dense motion field a corresponding motion energy field is calculated, according to the following equation:

$$K(b, c, t) = \|\overrightarrow{V(b, c, t)}\| \tag{1}$$

where $\overrightarrow{V(b, c, t)}$ is the estimated dense motion field, $\|.\|$ denotes the norm of a vector, and $K(b, c, t)$ is the resulting motion energy field. Variables b, c get values in the ranges $[1, V_{dim}]$ and [1, H_{dim}] respectively, where V_{dim} and H_{dim} are the motion field vertical and horizontal dimensions, whereas 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 [28]. Then, low-pass filtering is performed to the computed field for denoising and removing intense motion discontinuities. The resulting low-passed motion energy field, $M(b, c, t)$, is of high dimensionality, which decelerates the video processing, while motion information at this level of detail is not always required for the analysis purposes. Thus, it is consequently down-sampled, according to the following equations:

$$R(x, y, t) = M(\frac{2x-1}{2} \cdot VS_{step}, \frac{2y-1}{2} \cdot HS_{step}, t) \tag{2}$$

$$x = 1, ...D\ , \quad y = 1, ...D \tag{3}$$

$$VS_{step} = \frac{V_{dim}}{D}\ , \quad HS_{step} = \frac{H_{dim}}{D} \tag{4}$$

where $R(x, y, t)$ is the estimated down-sampled motion energy field and HS_{step}, VS_{step} are the corresponding horizontal and vertical spatial sampling frequencies. As can be seen from Eq. 3, the dimensions of the down-sampled field are predetermined and set equal to D. Since the aforementioned down-sampled motion energy fields, $R(x, y, t)$, are estimated for all the selected frames of each shot s_i, they are in turn utilized to compute the respective *motion observation sequence*. The latter will be used in order to associate the respective shot with a particular event e_j, as will be described in the sequel.

5. POLYNOMIAL APPROXIMATION

As already described in Section 2, the majority of the methods present in the relevant literature are focusing only at global- or camera-level motion processing approaches [10] [16]. Nevertheless, local-level analysis of the motion signal can provide significant cues which, if suitably exploited, can facilitate in efficiently capturing the underlying semantics of the examined video. Thus, a new representation for providing local-level motion information to HMMs is presented.

According to the HMM theory [19], 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, a compact and discriminative representation of motion features is required. For that purpose, the down-sampled motion energy field, $R(x, y, t)$, estimated for every selected frame (as described in Section 4), and which actually represents a motion energy distribution surface, is approximated by a 2D polynomial function, of the following form:

$$f(p, q) = \sum_{k,l} a_{kl} \cdot ((p - p_0)^k \cdot (q - q_0)^l)\ , \tag{5}$$

$$0 \leq k, l \leq T \ \ and \ \ 0 \leq k + l \leq T \tag{6}$$

where T is the order of the function, a_{kl} its coefficients and p_0, q_0 are defined as $p_0 = q_0 = \frac{D}{2}$. The approximation was performed using the least-squares method. In Fig. 1, indicative motion energy field approximation results are illustrated for tennis broadcast videos, showing the selected frame (first column), the estimated dense motion field (second column), the resulting motion energy field (third column) and its corresponding polynomial approximation, $\widehat{R}(x, y, t)$, (column 4). As can be seen from this figure, the polynomial approximation efficiently captures the most dominant motion characteristics.

The polynomial coefficients are calculated for every selected frame and are used to form an observation vector. These observation vectors are in turn utilized to form a respective shot observation sequence, namely the *motion observation sequence*, as described in Section 1. Then, a set of J HMMs is employed, where an individual HMM is introduced for every defined event e_j, in order to perform the association of the examined shot, s_i, with the defined events, e_j, based on motion information. More specifically, each HMM receives as input the aforementioned motion observation sequence and at the evaluation stage returns a *posterior* probability, which represents the observation sequence's fitness to the particular model. This probability indicates the degree of confidence, denoted by h_{ij}, with which event e_j is associated with shot s_i. The pairs of all supported events and their respective degrees of confidence computed for shot s_i, comprise the shot's hypothesis set H_i, where $H_i = \{h_{ij},\ j = 1, ...J\}$.

The proposed approximation of motion energy distribution approach, although quite simple, accomplishes to provide a very compact motion representation, since it estimates a low-dimensionality observation vector, while achieving to efficiently capture the most dominant motion characteristics of the examined frame. Despite its sometimes rough approximation, the polynomial coefficients accomplish to encompass even relatively small motion energy changes. Thus, polynomial coefficients, as will be shown in the experimentations part as well, constitute an effective motion representation approach for HMMs, since the estimated observation vectors are of low-dimensionality and simultaneously they are made of adequately statistical independent quantities, which facilitate HMMs to efficiently adjust their internal structure to these data.

6. ACCUMULATED MOTION ENERGY FIELD COMPUTATION

As described in the previous section, motion energy information and especially local-level motion energy distribution information from sequentially selected frames can provide valuable cues which can significantly facilitate the detection of high-level semantic events in videos. Nevertheless, motion energy distribution at a particular frame may not always provide adequate amount of information for discovering the underlying semantics of the examined video sequence, since different events may present similar motion patterns over a period of time. This fact generally hinders the identification of the correct event through the examination of motion features at distinct sequentially selected frames. In this section, an approach is presented for overcoming this problem, i.e. the problem of distinguishing between events that may present similar motion patterns over a period of time during their occurrence. Specifically, the fundamental idea of the proposed method is the incorporation of motion energy distribution information from previous frames for efficiently capturing the semantics present in a particular frame. The latter results into an accumulated motion energy field.

In particular, as described in Section 4, for every selected frame, an individual low-passed motion energy distribution field, $M(b, c, t)$, is calculated (Eq. 2). Then, while taking into account the previously mentioned observations and considerations, for each selected frame an *accumulated* motion energy distribution field is formed, according to the following equation:

$$M_{acc}(b, c, t, \tau) = \frac{\sum_0^\tau w(\tau) \cdot M(b, c, t-\tau)}{\sum_0^\tau w(\tau)},\ \tau = 0, 1, \dots, \quad (7)$$

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

$$w(\tau) = \frac{1}{v^{f\cdot\tau}},\ v > 1 \quad . \quad (8)$$

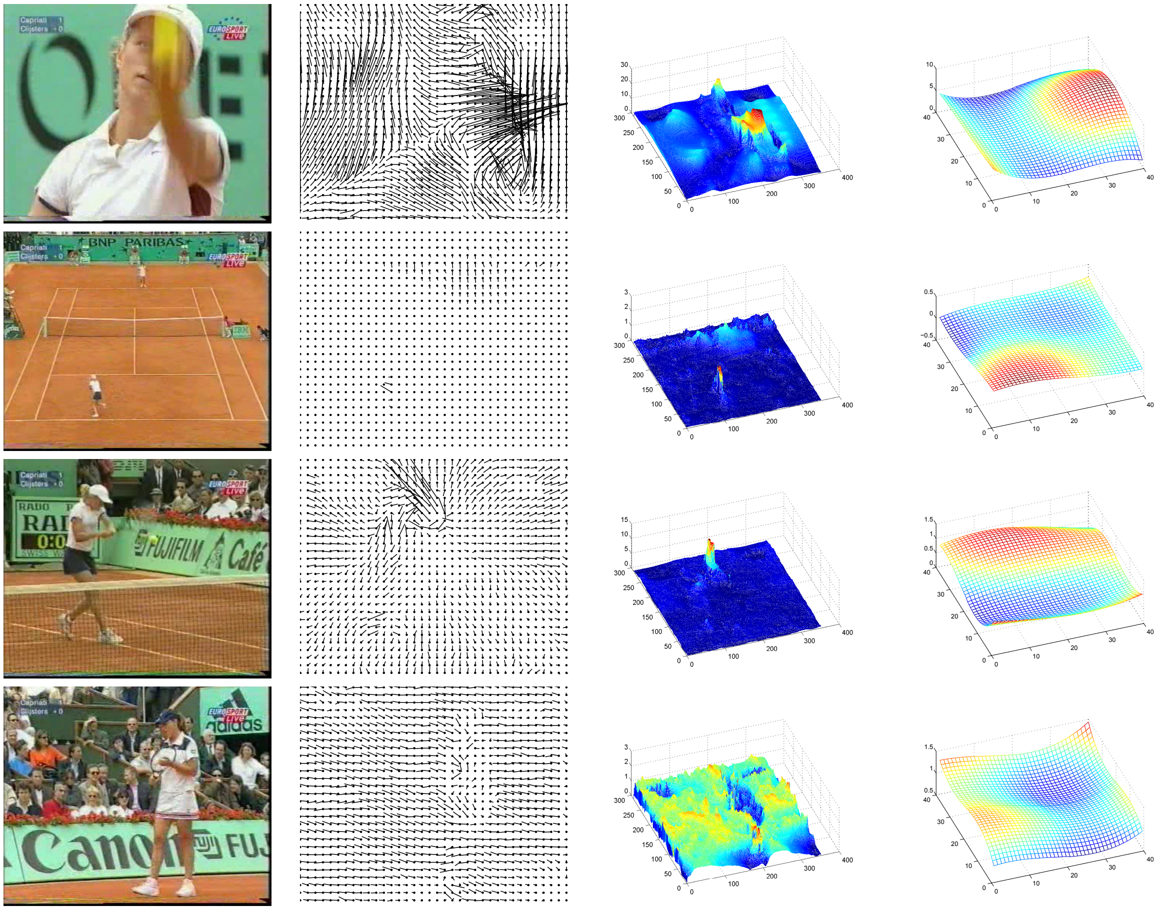
As can be seen from Eq. 8, the accumulated motion energy distribution field takes into account motion information from previous frames and, in particular, it gradually adds decreasing importance to motion information from distant frames to the currently examining one. The respective down-sampled accumulated motion energy field is denoted by $R_{acc}(x, y, t, \tau)$ and is calculated similarly to Eq. 2-4 using $M_{acc}(b, c, t, \tau)$ instead of $M(b, c, t)$.

In order to better demonstrate the usefulness and the efficiency of the proposed approach, an example of computing the accumulated motion energy fields for two individual events of the tennis broadcast domain is illustrated in Fig. 2. The first frame (first row) corresponds to a moment in time when the event break occurs. Specifically, the player is walking along the court holding her racket after a point has been gained. The second frame (second row) corresponds to a moment when the event serve occurs and particularly when the player is starting her attempt to serve while standing-up. The corresponding motion energy fields are presented in the second column. From a careful observation, it can be seen that the two fields present similar motion characteristics, since in both cases the player's silhouette is formed, which may lead to the incorrect identification of the corresponding events. In columns 3 and 4, the estimated accumulated fields are illustrated when motion characteristics from the previous ($M_{acc}(b, c, t, \tau),\ \tau = 1$) and the previous two frames ($M_{acc}(b, c, t, \tau),\ \tau = 2$) are taken into account, respectively. It can be seen that the two resulting accumulated motion fields, especially for the second case, present significant dissimilarities with respect to motion energy distribution, which can facilitate in the discrimination between the two events.

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 5 is followed for providing motion information to the respective HMM structure and realizing event detection 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, $R(x, y, t)$. In Fig. 3, indicative results of energy field polynomial approximations, while exploiting motion characteristics from previous frames, are presented. As can be seen from this figure, where the same frames as in Fig. 1 are used, the utilized polynomial function captures and efficiently models the motion energy distribution dissimilarities when no information ($\widehat{R}_{acc}(x, y, t, \tau),\ for\ \tau = 0$) and information from previous frames ($\widehat{R}_{acc}(x, y, t, \tau),\ for\ \tau = 2$) is exploited, respectively.

7. EXPERIMENTAL RESULTS

In this section experimental results from the application of the proposed method, as well as comparative evaluation with other approaches in the literature, are presented. Although the method is generic, i.e. it can be directly ap-



Selected frame Motion field Motion energy field Polynomial approximation

Figure 1: Examples of motion energy field approximation with polynomial function

plied to any possible domain of concern without the need for domain-specific algorithmic modifications or adaptations (as described in Section 1), particular domains need to be selected for experimentation; to this end, the domains of tennis and news broadcast video are utilized in this work. For the selected domains, the corresponding sets of high-level semantic events that are of interest and a brief description of their respective definitions are given in the sequel.

Tennis domain:

rally when the actual game is played

serve is defined as 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 servis hit

replay when a particular incident of increased importance is broadcasted again, usually in slow motion

break when a break in the game occurs, i.e. the actual game is interrupted for example after a point is gained, and the camera may show the players resting or the audience

News domain:

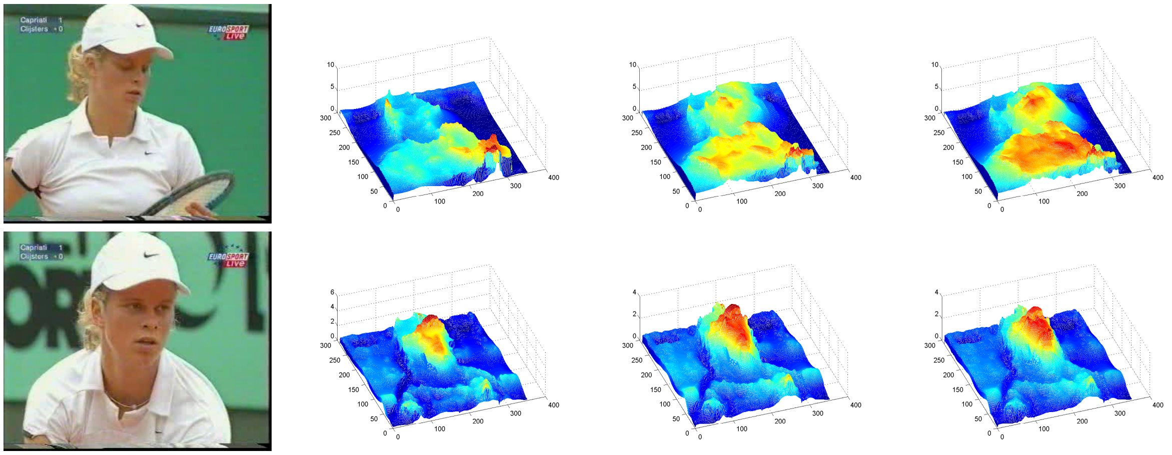
anchor when the anchor person announces the news in a studio environment

reporting when live-reporting takes place or a speech\ interview is broadcasted

reportage comprises of the displayed scenes, either indoors or outdoors, relevant to every broadcasted news item

graphics when any kind of graphics is depicted in the video sequence, including news start\end signals, maps, tables or text scenes

For experimentation in the tennis domain, a set of 12 videos showing professional tennis games from various international tournaments was collected. After the temporal segmentation algorithm described in Section 4 was applied, a corresponding set of 1449 shots was formed, which were manually annotated according to the respective tennis domain event definitions. From the aforementioned videos, 4 of them (499 shots) were used for training the developed HMMs structure (training set U_{tr}, described in Section 3) and the remaining 8 (950 shots) were used for evaluation (testing set U_{te}). A similar procedure was followed for the news domain; 24 videos



Selected frame $M_{acc}(b,c,t,\tau),\ for\ \tau=0$ $M_{acc}(b,c,t,\tau),\ for\ \tau=1$ $M_{acc}(b,c,t,\tau),\ for\ \tau=2$

Figure 2: Examples of accumulated motion energy field estimation

of news broadcast from Deutsche Welle[1] were collected and the respective sets U_{tr} and U_{te} comprising 342 and 582 shots respectively, were formed.

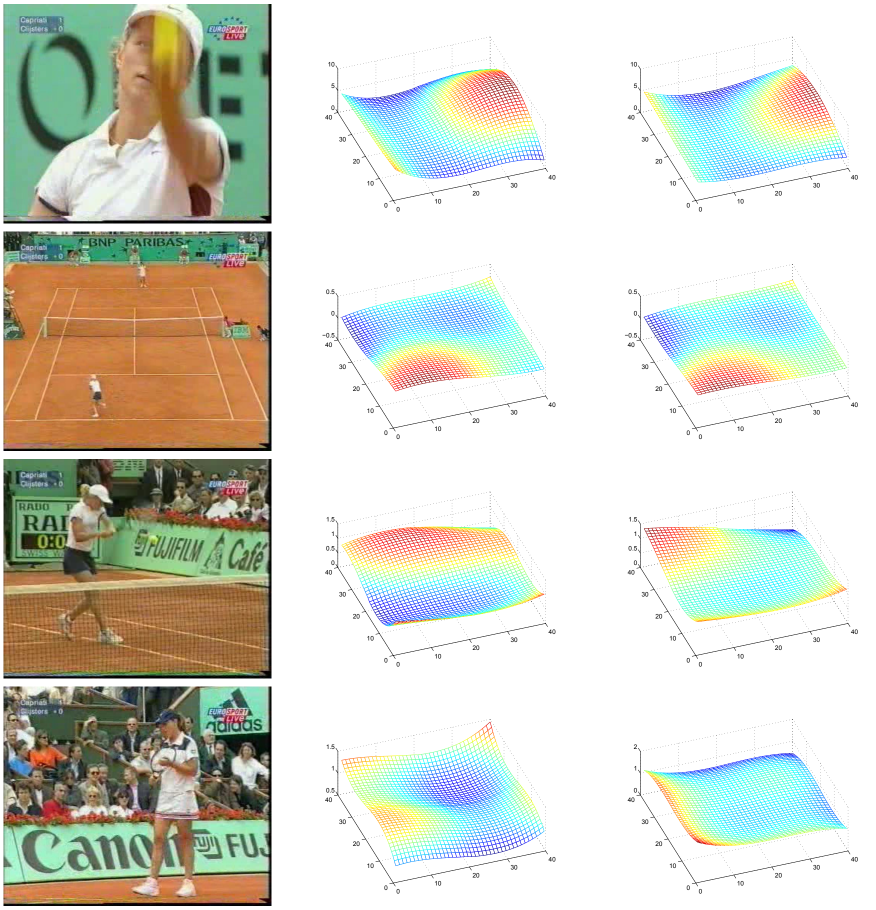
For every shot a set of frames was subsequently selected at equally spaced time intervals, as described in Section 4. The value of the temporal sampling frequency, SF_t, was set to 125ms based on experimentation. It has been observed that small variations around this value resulted into negligible changes in the overall detection performance. Then, for every selected frame the respective accumulated low-passed down-sampled motion energy field, $R_{acc}(x, y, t, \tau)$, was estimated and subsequently approximated by a 2D polynomial function, as described in Sections 6 and 5, respectively. A third order polynomial function was used for the approximation procedure, according to Eq. 5, since it produced the most accurate approximation results. The value of the parameter D in Eq. 3-4, which is used to define the horizontal, HS_{step}, and vertical, VS_{step}, spatial sampling frequencies, was set equal to 40. This value was shown to represent a good compromise between the need for time efficiency and effective polynomial approximation. Significantly lower values were shown to result into the generation of very few samples that could not be utilized for robust polynomial approximation. Additionally, the values of parameters v and f that define the time descending function in Eq. 8 were set equal to 3 and 0.5, respectively, after experimentation. The estimated polynomial coefficients were used to form the motion observation sequence for every shot, which was in turn provided as input to the developed HMMs structure in order to associate the shot with one of the supported events, as described in Section 5.

Regarding the HMM structure implementation details, fully connected first order HMMs, i.e. HMMs allowing all possible hidden state transitions, were utilized for performing the mapping of the low-level motion features to the high-level semantic events. 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 HMMs was considered as a free variable. The developed HMM structure was realized using the software libraries of [1].

In Table 1, quantitative event detection results are given in the form of the calculated confusion matrices when the accumulated motion energy fields, $R_{acc}(x, y, t, \tau)$, are used during the approximation step for $\tau = 0, 1, 2$ and 3, respectively. Note that $\tau = 0$ corresponds to the case where no motion information from previous frames is exploited. Additionally, the value of the overall detection accuracy is also given. The latter is defined as the percentage of the video shots that are assigned the correct event. It must be noted that shots containing commercials were not taken into account during the evaluation, while it has been regarded that $\arg\max_j(h_{ij})$ denotes the event e_j that is eventually associated with shot s_i.

From the observation of the presented results (first method in Table 1), it can be seen that generally the proposed polynomial approximation approach for providing motion information to HMMs is beneficial, since an overall detection accuracy of 74.01% and 77.22% is reached for the tennis and the news domain, respectively. This verifies the claim that the approximation of the motion energy field with a polynomial function achieves to efficiently capture the most dominant motion characteristics and models them suitably in a form that can be utilized effectively by HMMs. More specifically, in the tennis domain the event rally is recognized correctly at a high rate (93.15%), since it presents a representative and distinguishable motion pattern. Additionally, events serve and break also exhibit satisfactory results (69.57% and 66.95%, respectively). Regarding the recognition of replay, it presents a relatively low recognition rate (38.46%) and is mainly confused with serve and break. The latter is justified by the observation that replays are actually important incidents during the game that are broadcasted again usually in a close-up view and in slow-motion. Thus, it is expected to present similar local motion characteristics

[1] http://www.dw-world.de/



Selected frame $\widehat{R}_{acc}(x,y,t,\tau),\ for\ \tau=0$ $\widehat{R}_{acc}(x,y,t,\tau),\ for\ \tau=2$

Figure 3: Examples of motion energy field approximation with polynomial function without and with exploitation of motion information from previous frames

with the events serve and break. On the other hand, for the news domain events anchor, reportage and graphics are correctly identified at high recognition rates (79.55%, 79.44% and 87.50%, respectively). With respect to the event reporting, although it exhibits satisfactory results (60.98%), it tends to be confused with anchor and reportage. The latter is caused by the fact that speech or interview occurrences may present similar motion patterns with anchor speaking or reportage scenes, respectively.

Table 1 also depicts the impact of the exploitation of the introduced accumulated motion energy fields for different values of τ on the performance of the proposed algorithm. As can be seen from the presented results, the event detection performance generally increases for both domains when the accumulated motion energy fields, $R_{acc}(x, y, t, \tau)$, are used for small values of τ ($\tau = 1, 2$), compared to the case where no motion information from previous frames is utilized during the motion energy fields computation, i.e. when $\tau = 0$. Specifically, a maximum increase of 3.08% in the overall event detection accuracy is observed when $\tau = 1$ for the tennis domain, and a corresponding increase of 2.85% is presented when $\tau = 2$ for the news domain. For the aforementioned cases, some events of each domain are particularly favored by the utilization of the introduced accumulated motion energy fields. The above results justify the claim that incorporating information from previous frames during the motion energy field computation can facilitate in distinguishing between events that present similar motion patterns over a period of time during their occurrence. On the other hand, from the above table it can be seen that when the value of τ is further increased, the overall performance improvement decreases and for the particular case of $\tau = 3$ for the tennis domain it diminishes. This is mainly due to the fact that when taking into account information from many previous frames the estimated accumulated motion fields for each frame tend to become very similar. Thus, polynomial coefficients tend to have also very similar values and hence HMMs cannot observe a characteristic sequence of features that unfolds in time.

In Table 2, the performance of the proposed method is compared with the motion representation approaches for providing motion information to HMM-based systems presented in [10] and [27]. Specifically, Gilbert et al. makes use of the available motion vectors for estimating the principal motion direction of every frame [10]. Additionally, Xie et al. calculates the motion intensity at frame level [27]. From the presented results, it can be easily observed that the proposed approach outperforms the aforementioned algorithms for most of the supported events as well as in overall detection accuracy for both supported domains. More specifically, only the method proposed by Xie [27] presents a higher recognition rate for the event replay (61.54%) in the tennis domain. This is due to replays exhibiting characteristic global camera motion patterns (like zoom-ins and zoom-outs), which were better represented by motion intensity estimation at frame level. The presented results verify that local-level analysis of the motion signal can lead to increased event detection performance.

8. CONCLUSIONS

In this paper, a motion-based approach for detecting high-level semantic events in video sequences was presented. The proposed algorithm is generic, i.e. it can be directly applied to any possible domain of concern without the need for domain-specific algorithmic modifications or adaptations. It is based on a new representation for providing local-level motion information to HMMs, while motion characteristics from previous frames are also exploited. Experimental results in the domains of tennis and news broadcast video demonstrated the efficiency of the proposed approach. Future work includes the examination of more sophisticated motion representation schemes for motion-based recognition applications and the investigation of corresponding algorithms for color/audio signal processing that will allow the integration of the proposed motion-based approach in a multimodal event detection scheme.

9. ACKNOWLEDGMENTS

The work presented in this paper was supported by the European Commission under contracts FP6-001765 aceMedia, FP6-027685 MESH, FP6-045480 CHORUS and FP6-027026 K-Space, and by the GSRT under project DELTIO.

10. REFERENCES

[1] http://htk.eng.cam.ac.uk/. *Hidden Markov Model Toolkit, HTK.*

[2] M. Barnard and J. Odobez. Sports Event Recognition Using Layered HMMS. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1150–1153, 2005.

[3] S. Chang. The holy grail of content-based media analysis. *Multimedia, IEEE*, 9(2):6–10, 2002.

[4] C. Cheng and C. Hsu. Fusion of audio and motion information on HMM-based highlight extraction for baseball games. *Multimedia, IEEE Transactions on*, 8(3):585–599, 2006.

[5] J. Chien. Online hierarchical transformation of hidden markov models for speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 7(6):656–667, Nov 1999.

[6] N. Dimitrova, H. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *Multimedia, IEEE*, 9(3):42–55, 2002.

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

[8] A. Francois, R. Nevatia, J. Hobbs, R. Bolles, and J. Smith. VERL: an ontology framework for representing and annotating video events. *IEEE MultiMedia Magazine*, 12(4):76–86, 2005.

[9] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. Kakumanu, and O. Garcia. Audio/visual mapping with cross-modal hidden Markov models. *Multimedia, IEEE Transactions on*, 7(2):243–252, 2005.

[10] X. Gibert, H. Li, and D. Doermann. Sports video classification using HMMS. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 2, 2003.

[11] J. Huang, Z. Liu, and Y. Wang. Joint scene classification and segmentation based on hidden Markov model. *Multimedia, IEEE Transactions on*, 7(3):538–550, 2005.

[12] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. HMM based structuring of tennis videos using visual and audio cues. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 3, 2003.

[13] V. Kobla, D. Doermann, and K. Lin. Archiving, indexing, and retrieval of video in the compressed domain. *Proc. of the SPIE Conference on Multimedia Storage and Archiving Systems*, 2916:78–89, 1996.

[14] X. Li and F. Porikli. A hidden Markov model framework for traffic event detection using video features. *Image Processing, 2004. ICIP'04. 2004 International Conference on*, 5, 2004.

[15] Z. Li, G. Schuster, and A. Katsaggelos. MINMAX optimal video summarization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(10):1245–1256, 2005.

[16] S. Liu, M. Xu, H. Yi, L. Chia, and D. Rajan. Multimodal Semantic Analysis and Annotation for Basketball Video. *EURASIP Journal on Applied Signal Processing*, 2006, 2006.

[17] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia Magazine*, 13(3):86–91, 2006.

[18] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck. Determination of Optical Flow and its Discontinuities using Non-Linear Diffusion. *Computer vision-ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2-6, 1994: Proceedings*, 1994.

[19] L. Rabiner. A tutorial on hidden Markov models and selected applications inspeech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[20] D. Sadlier and N. O'Connor. Event Detection in Field Sports Video Using Audio–Visual Features and a Support Vector Machine. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1225, 2005.

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 3, 2004.

[22] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[23] C. Snoek, M. Worring, D. Koelma, and A. Smeulders. A Learned Lexicon-Driven Paradigm for Interactive

Table 1: Event detection results when $R_{acc}(x, y, t, \tau)$ is used

Tennis domain						
Method	Actual Event	Detected Event				Overall Accuracy
		Rally	Serve	Replay	Break	
$R_{acc}(x, y, t, \tau)$ for $\tau = 0$	Rally	93.15%	0.00%	0.00%	6.85%	74.01%
	Serve	0.00%	69.57%	0.00%	30.43%	
	Replay	0.00%	30.77%	**38.46**%	30.77%	
	Break	9.32%	14.41%	9.32%	66.95%	
$R_{acc}(x, y, t, \tau)$ for $\tau = 1$	Rally	94.52%	1.37%	0.00%	4.11%	**77.09**%
	Serve	0.00%	**73.91**%	4.35%	21.74%	
	Replay	0.00%	38.46%	**38.46**%	23.08%	
	Break	9.32%	11.02%	8.47%	**71.19**%	
$R_{acc}(x, y, t, \tau)$ for $\tau = 2$	Rally	**95.89**%	1.37%	0.00%	2.74%	75.77%
	Serve	0.00%	65.22%	4.35%	30.43%	
	Replay	0.00%	38.46%	23.08%	38.46%	
	Break	8.47%	11.86%	8.47%	**71.19**%	
$R_{acc}(x, y, t, \tau)$ for $\tau = 3$	Rally	91.78%	2.74%	0.00%	5.48%	74.01%
	Serve	0.00%	65.22%	4.35%	30.43%	
	Replay	0.00%	38.46%	30.77%	30.77%	
	Break	9.32%	12.71%	8.47%	69.49%	
News domain						
Method	Actual Event	Detected Event				Overall Accuracy
		Anchor	Reporting	Reportage	Graphics	
$R_{acc}(x, y, t, \tau)$ for $\tau = 0$	Anchor	**79.55**%	20.45%	0.00%	0.00%	77.22%
	Reporting	14.63%	60.98%	24.39%	0.00%	
	Reportage	1.67%	15.56%	79.44%	3.33%	
	Graphics	12.50%	0.00%	0.00%	87.50%	
$R_{acc}(x, y, t, \tau)$ for $\tau = 1$	Anchor	77.27%	22.73%	0.00%	0.00%	79.72%
	Reporting	4.88%	**70.73**%	24.39%	0.00%	
	Reportage	1.67%	12.22%	81.67%	4.44%	
	Graphics	6.25%	6.25%	0.00%	87.50%	
$R_{acc}(x, y, t, \tau)$ for $\tau = 2$	Anchor	75.00%	25.00%	0.00%	0.00%	**80.07**%
	Reporting	4.88%	**70.73**%	24.39%	0.00%	
	Reportage	1.11%	12.22%	**82.22**%	4.44%	
	Graphics	0.00%	6.25%	0.00%	**93.75**%	
$R_{acc}(x, y, t, \tau)$ for $\tau = 3$	Anchor	**79.55**%	20.45%	0.00%	0.00%	79.36%
	Reporting	9.76%	65.85%	24.39%	0.00%	
	Reportage	1.67%	12.22%	81.11%	5.00%	
	Graphics	6.25%	0.00%	0.00%	**93.75**%	

Video Retrieval. *Multimedia, IEEE Trans. on*, 9(2):280–292, 2007.

[24] K. Su and C. Lee. Speech recognition using weighted hmm and subspace projection approaches. *Speech and Audio Processing, IEEE Transactions on*, 2(1):69–79, Jan 1994.

[25] T. Takiguchi, S. Nakamura, and K. Shikano. Hmm-separation-based speech recognition for a distant moving speaker. *Speech and Audio Processing, IEEE Transactions on*, 9(2):127–140, Feb 2001.

[26] J. Wang, C. Xu, and E. Chng. Automatic Sports Video Genre Classification using Pseudo-2D-HMM. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)-Volume 04*, pages 778–781, 2006.

[27] L. Xie, P. Xu, S. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognition Letters*, 25(7):767–775, 2004.

[28] G. Xu, Y. Ma, H. Zhang, and S. Yang. An HMM-based framework for video semantic analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(11):1422–1433, 2005.

[29] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga. Scalable architecture for word hmm-based speech recognition. *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, 3:III–417–20 Vol.3, 23-26 May 2004.

Table 2: Comparative event detection results

Tennis domain						
Method	Actual Event	Detected Event				Overall Accuracy
		Rally	Serve	Replay	Break	
$R_{acc}(x,y,t,\tau)$ for $\tau=1$	Rally	**94.52**%	1.37%	0.00%	4.11%	**77.09**%
	Serve	0.00%	**73.91**%	4.35%	21.74%	
	Replay	0.00%	38.46%	38.46%	23.08%	
	Break	9.32%	11.02%	8.47%	**71.19**%	
Method of [10]	Rally	89.04%	9.59%	1.37%	0.00%	60.79%
	Serve	56.52%	21.74%	8.70%	13.04%	
	Replay	15.38%	30.77%	23.08%	30.77%	
	Break	27.97%	14.41%	2.54%	55.08%	
Method of [27]	Rally	93.15%	6.85%	0.00%	0.00%	55.95%
	Serve	8.70%	30.43%	26.09%	34.78%	
	Replay	7.69%	15.38%	**61.54**%	15.38%	
	Break	14.41%	16.10%	32.20%	37.29%	

News domain						
Method	Actual Event	Detected Event				Overall Accuracy
		Anchor	Reporting	Reportage	Graphics	
$R_{acc}(x,y,t,\tau)$ for $\tau=2$	Anchor	**75.00**%	25.00%	0.00%	0.00%	**80.07**%
	Reporting	4.88%	**70.73**%	24.39%	0.00%	
	Reportage	1.11%	12.22%	**82.22**%	4.44%	
	Graphics	0.00%	6.25%	0.00%	**93.75**%	
Method of [10]	Anchor	18.18%	4.55%	0.00%	77.27%	61.21%
	Reporting	7.32%	17.07%	43.90%	31.71%	
	Reportage	1.67%	8.89%	80.00%	9.44%	
	Graphics	12.50%	6.25%	0.00%	81.25%	
Method of [27]	Anchor	52.27%	6.82%	0.00%	40.91%	59.07%
	Reporting	9.76%	39.02%	29.27%	21.95%	
	Reportage	6.11%	23.33%	63.89%	6.67%	
	Graphics	6.25%	18.75%	0.00%	75.00%	