ESTIMATION AND REPRESENTATION OF ACCUMULATED MOTION CHARACTERISTICS FOR SEMANTIC EVENT DETECTION

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

[1]Information Processing Lab., Electrical & Computer Eng. Dep., Aristotle Univ. of Thessaloniki, Greece
[2]Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

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 examined video sequence is initially segmented into shots and for every resulting shot appropriate motion features are extracted. Then, Hidden Markov Models (HMMs) are employed for performing the association of each shot with one of the high-level semantic events that are of interest in any given domain. 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. Experimental results as well as comparative evaluation from the application of the proposed approach in the domain of news broadcast video are presented.

Index Terms— knowledge-assisted video analysis, HMMs, event detection, motion representation

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. To this end, several approaches have been proposed in the literature regarding the tasks of indexing, searching, summarization and retrieval of video content [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, thus attempting to bridge the so called *semantic gap* [2]. Among the video analysis methodologies of the latter category, approaches that exploit *a priori* knowledge have been particularly favored and have so far exhibited promising results.

Knowledge-assisted video analysis techniques have been dominated by the usage of Machine Learning (ML) algorithms. 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 [3], a Hidden Markov Model (HMM)-based system is proposed for performing joint scene classification and video temporal segmentation. Additionally, in [4], Support Vector Machines (SVMs) 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 a single 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, 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 interest need to be defined for any given domain and a corresponding set of annotated video content needs to be provided for training purposes. For realizing event detection, the examined video sequence 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, 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. Regarding the motion feature extraction procedure, unlike the majority of the approaches of the relevant literature that are mainly limited to global or camera motion, local-level analysis is supported for efficiently capturing the semantics of the visual medium. More specifically, a new representation for providing local-level motion information to HMMs is presented, while motion characteristics from previous frames are also exploited. The fundamental idea of the latter is that incorporating motion information from previous frames can significantly facilitate in capturing the semantics present in a particular frame and results into an *accumulated motion field*.

The paper is organized as follows: Section 2 describes how HMMs are employed for realizing semantic event detection. The video pre-processing steps are described in Section 3. Section 4 outlines the proposed local-level motion representation and Section 5 details the methodology followed for incorporating motion characteristics from previous frames. Experimental results and comparative evaluation from the application of the proposed approach in the news broadcast domain are presented in Section 6, and conclusions are drawn in Section 7.

2. HIDDEN MARKOV MODELS

Under the proposed approach, HMMs are employed for detecting high-level semantic events in video sequences. In accordance to

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

the HMM theory [5], 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 application of the proposed video analysis method 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.

3. VIDEO PRE-PROCESSING

Prior to event detection, the examined video sequence is segmented into shots, which constitute the elementary image sequences of video. This results in a set of shots, denoted by $S = \{s_i,\ i = 1, ...I\}$, to which the examined video is decomposed.

Following shot segmentation, 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, making use of an optical flow estimation algorithm. 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. Then, low-pass filtering is applied 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.

4. POLYNOMIAL APPROXIMATION

In this section, the motion information processing for associating each video shot with the supported high-level semantic events is detailed. According to the HMM theory [5], 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 3), and which actually represents a motion energy distribution surface, is approximated by a 2D polynomial function, of the following form:

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

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

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 is performed using the least-squares method.

The polynomial coefficients are calculated for every selected frame and are used to form an observation vector. These observation vectors are utilized to form the motion observation sequence. 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 shot-event association. Each HMM receives as input the aforementioned motion observation sequence and at the evaluation stage returns a posterior probability, which 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 $H_i = \{h_{ij},\ j = 1, ...J\}$. HMM implementation details are discussed in the experimental results section.

5. ACCUMULATED MOTION ENERGY FIELD COMPUTATION

As described in Section 1, motion characteristics at particular frames may not always provide an 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 during their occurrence. In this section, an approach is presented for overcoming this problem.

In particular, as described in Section 3, for every selected frame, an individual low-passed motion energy distribution field, $M(b, c, t)$, is calculated (Eq. (2)). Then, for each selected frame an accumulated motion energy distribution field is formed, according to the following equation:

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

where t is the current frame, τ denotes previously selected frames and $w(\tau)$ is a time-dependent normalization factor. The latter is modeled by the following time descending function:

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

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 previous frames to the currently examined one. The respective down-sampled accumulated motion energy

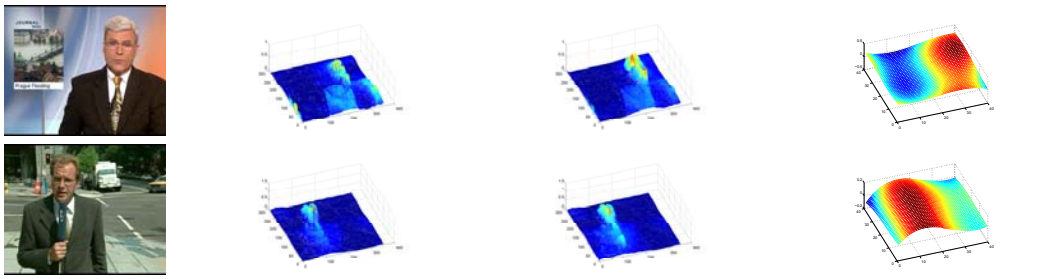

Fig. 1. Examples of accumulated motion energy field estimation and polynomial approximation for the anchor (1st row) and reporting (2nd row) events in a news video.

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 $M(b,c,t)$.

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 4 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, $R(x,y,t)$. An example of computing the accumulated motion energy fields, as well as the corresponding polynomial approximations ($\widehat{R}_{acc}(x,y,t,\tau)$), for two individual events of the news broadcast domain is illustrated in Fig. 1.

6. 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 applied to any possible domain of concern without the need for domain-specific algorithmic modifications or adaptations, a domain needs to be selected for experimentation; to this end, the domain of news broadcast video is utilized in this work. For the selected domain, the corresponding set of high-level semantic events that are of interest comprises the following events: *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) and *graphics* (when any kind of graphics is depicted in the video sequence, including news start\end signals, maps, tables or text scenes).

Then, a set of 24 videos of news broadcast from Deutsche Welle[1] was collected. After the temporal segmentation algorithm described in Section 3 was applied, a corresponding set of 924 shots was formed, which were manually annotated according to the event definitions already described. From the aforementioned videos 8 of them (342 shots) were used for training the developed HMM structure (training set U_{tr}, described in Section 2) and the remaining 16 (582 shots) were used for evaluation (testing set U_{te}).

At the signal level, after video temporal segmentation to shots, for every resulting shot a set of frames was selected at equally spaced time intervals, as described in Section 3. The value of the temporal sampling frequency, SF_t, was set to 125ms based on experimentation. It has been observed that small deviations from 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 4 and 5. 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. 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 HMM structure in order to associate the shot with one of the supported events, as described in Section 4.

Regarding the HMM structure implementation details, fully connected first order HMMs were utilized. For every hidden state the observations were modeled as a mixture of Gaussians, which were set to have full covariance matrices. Additionally, the Baum-Welch (or Forward-Backward) algorithm was used for training, while the Viterbi algorithm was utilized during the evaluation. The number of hidden states of the HMMs was considered as a free variable.

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. Additionally, the value of the overall detection accuracy is also given, which is defined as the percentage of the video shots that are assigned the correct event. It has been regarded that $\arg\max_j(h_{ij})$ indicates the event e_j that is associated with shot s_i.

From the results presented 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 77.22% is reached. Additionally, all the supported events are correctly identified at high recognition rates. 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 when the accu-

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

Table 1. Event detection results based on motion information

Method	Actual Event	Detected Event: Anchor	Reporting	Reportage	Graphics	Overall Accuracy
$R_{acc}(x,y,t,\tau)$ for $\tau=0$	Anchor	**79.55**%	20.45%	0.00%	0.00%	
	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%	77.22%
$R_{acc}(x,y,t,\tau)$ for $\tau=1$	Anchor	77.27%	22.73%	0.00%	0.00%	
	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%	79.72%
$R_{acc}(x,y,t,\tau)$ for $\tau=2$	Anchor	75.00%	25.00%	0.00%	0.00%	
	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**%	**80.07**%
$R_{acc}(x,y,t,\tau)$ for $\tau=3$	Anchor	**79.55**%	20.45%	0.00%	0.00%	
	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**%	79.36%
Method of [6]	Anchor	18.18%	4.55%	0.00%	77.27%	
	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%	61.21%
Method of [7]	Anchor	52.27%	6.82%	0.00%	40.91%	
	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%	59.07%

mulated motion energy fields, $R_{acc}(x,y,t,\tau)$, are used for small values of τ (an increase of 2.50% and 2.85% in the overall event detection accuracy is observed when $\tau=1$ and $\tau=2$, respectively), compared to the case where no motion information from previous frames is utilized during the motion energy fields computation, i.e. when $\tau=0$. 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 (an increase of 2.14% is observed when $\tau=3$). 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 1, the performance of the proposed method is compared with the motion representation approaches for providing motion information to HMM-based systems presented in [6] and [7]. Specifically, Gibert et al. estimates the principal motion direction of every frame [6], while Xie et al. calculates the motion intensity at frame level [7]. From the presented results, it can be easily observed that the proposed approach ($R_{acc}(x,y,t,\tau)$, for $\tau=2$) outperforms the aforementioned algorithms for all supported events as well as in overall detection accuracy. This verifies that local-level analysis of the motion signal can lead to increased event detection performance.

7. CONCLUSIONS

In this paper, a motion-based approach for detecting high-level semantic events in video sequences was presented. The proposed algorithm is generic and it is based on a new representation for providing local-level motion information to HMMs, while motion characteristics from previous frames are also exploited. Future work includes the investigation of corresponding algorithms for color/audio signal processing that will allow the integration of the proposed motion-based approach in a multi-modal event detection scheme.

8. REFERENCES

[1] S.F. Chang, "The holy grail of content-based media analysis," *Multimedia, IEEE*, vol. 9, no. 2, pp. 6–10, 2002.

[2] AWM Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *PAMI, IEEE Trans. on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] J. Huang, Z. Liu, and Y. Wang, "Joint scene classification and segmentation based on hidden Markov model," *Multimedia, IEEE Trans. on*, vol. 7, no. 3, pp. 538–550, 2005.

[4] D.A. Sadlier and N.E. OConnor, "Event Detection in Field Sports Video Using Audio–Visual Features and a Support Vector Machine," *IEEE TCSVT*, vol. 15, no. 10, pp. 1225, 2005.

[5] LR Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[6] X. Gibert, H. Li, and D. Doermann, "Sports video classification using HMMS," *ICME, IEEE Int. Conf. on*, vol. 2, 2003.

[7] L. Xie, P. Xu, S.F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pat. Recog. Let.*, vol. 25, no. 7, pp. 767–775, 2004.