A UNIFIED FRAMEWORK FOR SEMANTIC EVENT DETECTION

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

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

ABSTRACT

In this poster, a generic multi-modal context-aware framework for detecting high-level semantic events in video sequences is presented. Initially, Hidden Markov Models (HMMs) are employed for performing an initial association of the examined video with the events of interest separately for every modality. Then, an integrated Bayesian Network (BN) is introduced for simultaneously performing information fusion and contextual knowledge exploitation.

1. INTRODUCTION

During the recent years, intense research efforts have concentrated in the development of sophisticated and user-friendly systems for skilful management of video sequences. Most of them have adopted the fundamental principle of shifting video manipulation techniques towards the processing of the visual content at a semantic level. Moreover, the usage of multi-modal as well as contextual information has emerged as a common practice for overcoming the ambiguity that is inherent in the visual medium.

In this poster, a generic multi-modal context-aware framework for detecting high-level semantic events in video sequences, making use of Machine Learning algorithms for implicit knowledge acquisition, is presented. Initially, HMMs are employed for performing an initial association of the examined video with the events of concern separately for every utilized modality. Then, an integrated BN is introduced for simultaneously performing information fusion of the individual modality analysis results and contextual knowledge exploitation.

2. OBJECTIVE OF WORK

The objective of the proposed approach is the detection of a set of predefined semantic events, denoted by $E = \{e_j,\ j = 1, ..J\}$, for a particular domain. The latter represent semantically meaningful incidents that are of interest in a possible application case and have a temporal duration. The accurate and efficient detection of them can facilitate tasks like video indexing, search and retrieval with respect to semantic criteria [1].

3. VIDEO PRE-PROCESSING

At the signal level, the examined video sequence is initially segmented into shots, denoted by $S = \{s_i, i = 1, ...I\}$, which constitute the elementary image sequences of video. For every shot, a global-level color histogram is calculated at equally spaced time intervals. Similarly, a set of dense motion fields are estimated with respect to the motion modality. Moreover, the widely used Mel Frequency Cepstral Coefficients (MFCCs) are utilized for the audio information processing.

4. HMM-BASED ANALYSIS

4.1. Color- and Audio-based Analysis

After a set of color histograms is estimated for each shot, as described in Section 3, they are utilized to form the corresponding shot's *color observation sequence*. The latter is provided as input to a HMM structure, which performs the association of every shot with the supported events based solely on color information. In particular, an individual degree of confidence, h^C_{ij}, is calculated for denoting the degree with which shot s_i is associated with every event e_j . With respect to the audio information processing, the computed MFCCs are used to form the shot's *audio observation sequence*. Similarly to the color analysis case, a degree of confidence, h^A_{ij}, is calculated to indicate the corresponding association.

4.2. Motion-based Analysis

4.2.1. Polynomial Approximation

For every estimated dense motion field, which is computed as described in Section 3, a corresponding motion energy field, $M(b, c, t)$, is calculated. The latter, which actually represents a motion energy distribution surface, is approximated by a 2D polynomial function, of the following form:

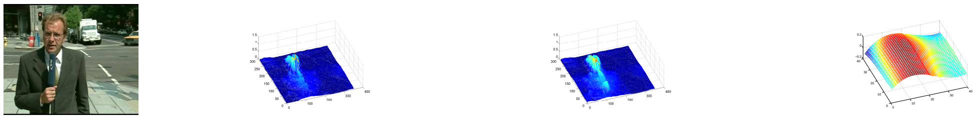
$$f(p,q) = \sum_{k,l} a_{kl} \cdot ((p-p_0)^k \cdot (q-q_0)^l) \ , \qquad 0 \le k,l \le T \ \ and \ \ 0 \le k+l \le T \tag{1}$$

The approximation is performed using the least-squares method. Subsequently, the estimated polynomial coefficients, a_{kl}, are used to form the respective shot's *motion observation sequence*. The latter is provided as input to the developed HMM structure for performing the association of each shot with the supported events based solely on motion information [2]. Similarly to the color and audio analysis cases, a degree of confidence, h^M_{ij}, is calculated to indicate the corresponding association.

4.2.2. Accumulated Motion Energy Field Computation

In order to overcome the problem of distinguishing between events that may present similar motion patterns over a period of time during their occurrence, an *accumulated motion energy field* is estimated with respect to every computed $M(b, c, t)$ [2], according to the following equation:

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.



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

Fig. 1. Example of accumulated motion energy field estimation and polynomial approximation for the reporting event in a news video.

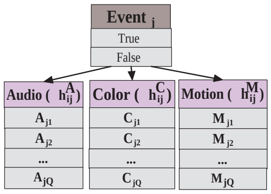


Fig. 2. Developed BN for modality fusion.

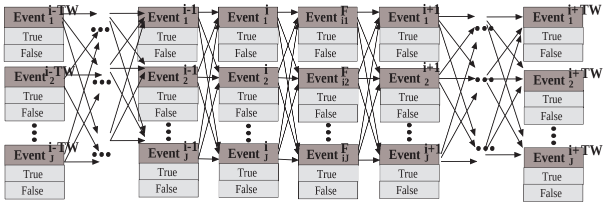


Fig. 3. Integrated BN for joint modality fusion and temporal context modeling.

Table 1. Event detection results.

Actual Event	Detected Event			
	Anchor	Reporting	Reportage	Graphics
Anchor	77.97%	10.17%	11.86%	0.00%
Reporting	0.00%	60.71%	39.29%	0.00%
Reportage	4.60%	1.15%	94.25%	0.00%
Graphics	9.38%	0.00%	2.30%	88.32%
Overall accuracy				86.01%

$$M_{acc}(b,c,t,\tau) = \frac{\sum_0^\tau w(\tau)\cdot M(b,c,t-\tau)}{\sum_0^\tau w(\tau)},\ \tau = 0,1,\dots\ , \qquad (2)$$

where $w(\tau)$ is modeled by the following time descending function:

$$w(\tau) = \frac{1}{v^{f\cdot\tau}},\ v > 1\ . \qquad (3)$$

Following their extraction, a procedure similar to the one described in Section 4.2.1 is followed for providing motion information to the respective HMM structure, where now the computed $M_{acc}(b,c,t,\tau)$ are used during the polynomial approximation process, instead of the $M(b,c,t)$. In Fig. 1, an indicative example of energy field polynomial approximation ($\widehat{M}_{acc}(x,y,t,\tau)$) is presented.

5. FUSION AND CONTEXT EXPLOITATION

5.1. Information Fusion

Under the proposed approach, BNs are employed for fusing the computed single-modality analysis results. In particular, a set of J BNs is introduced, one for every defined event e_j. In Fig. 2 the network structure of every utilized BN is illustrated. This network topology defines explicitly the causal relationships between the respective variables, i.e. the event that is depicted in a shot determines the features observed with respect to every modality. Every BN estimates a degree of belief for the parent node, which constitutes a quantitative indication of the association between each shot s_i and the respective event e_j based on multi-modal information.

5.2. Context Exploitation

In order to overcome the inherent ambiguity of the visual medium, an integrated BN model is introduced for acquiring and exploiting the appropriate contextual information, i.e. the supported events' temporal occurrence order. Specifically, the developed BN, whose topology is illustrated in Fig. 3, receives as input the shot-event associations based on multi-modal information of every shot s_i (Section 5.1), as well as of all its neighboring shots than lie within a certain time window. It must be noted that all network nodes expect $Event^F_{ij}$ correspond to the appropriate parent nodes of the BNs that have been developed for performing information fusion (Section 5.1). At the evaluation stage, the integrated BN estimates a degree of belief for every $Event^F_{ij}$ node, denoted by h^F_{ij}, which indicates the degree of confidence with which event e_j is eventually associated with shot s_i.

6. EXPERIMENTAL RESULTS AND CONCLUSIONS

The proposed framework was tested on videos belonging to the news broadcast domain [2]. The results presented in Table 1 demonstrate the efficiency of the proposed approach.

7. REFERENCES

[1] G. Th. Papadopoulos, V. Mezaris, I. Kompatsiaris and M. G. Strintzis, *Accumulated Motion Energy Fields Estimation and Representation for Semantic Event Detection*, in Proc. of CIVR, Niagara Falls, Canada, July 2008.

[2] G. Th. Papadopoulos, V. Mezaris, I. Kompatsiaris and M. G. Strintzis, *Estimation and Representation of Accumulated Motion Characteristics for Semantic Event Detection*, in Proc. of IEEE ICIP-MIR 2008), San Diego, USA, October 2008.