

# COMBINING MULTIMODAL AND TEMPORAL CONTEXTUAL INFORMATION FOR SEMANTIC VIDEO ANALYSIS

Georgios Th. Papadopoulos<sup>1,2</sup>, Vasileios Mezaris<sup>2</sup>, Ioannis Kompatsiaris<sup>2</sup> and Michael G. Strintzis<sup>1,2</sup>

<sup>1</sup>Information Proc. Lab., Electrical & Computer Eng. Dep., Aristotle Univ. of Thessaloniki, Greece

<sup>2</sup>Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

## ABSTRACT

In this paper, a graphical modeling-based approach to semantic video analysis is presented for jointly realizing modality fusion and temporal context exploitation. 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 every modality. Subsequently, an integrated Bayesian Network (BN) is introduced for simultaneously performing information fusion and temporal contextual knowledge exploitation, contrary to the usual practice of performing each task separately. The final outcome of the overall video analysis approach is the association of a semantic class with every shot. 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**— Semantic video analysis, modality fusion, temporal context, bayesian network

## 1. INTRODUCTION

The rapid advances in hardware technology have led to a tremendous increase in the total amount of video content generated and distributed everyday. As a consequence, 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 for tasks like search and organization of video content. More recently, the fundamental principle of processing the audio-visual information from a semantic-oriented perspective has been widely adopted, thus attempting to bridge the so called *semantic gap* [1] 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 in the literature, where the appropriate modality is selected depending on the application of concern [2]. 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. Hence, they efficiently capture the semantic information contained in the video content, which is embedded in multiple forms that are usually complementary to each other [3].

In addition to modality fusion, the use of context has been shown to further facilitate semantic video analysis. Among the available

contextual information types, temporal context is of increased importance in video analysis for modeling temporal relations between semantic elements or temporal variations of particular features. Xu et al. [4] introduce a HMM-based framework for modeling temporal contextual constraints in different semantic granularities. Additionally, Huang et al. [5] adopt a dynamic programming technique for searching for the most likely class transition path.

While a plethora of sophisticated methods have already been proposed with respect to modality fusion and temporal context modeling, however, the possibility of jointly performing both these tasks has not been examined. The latter would allow the exploitation of the possible correlations between the respective numerical data and consequently could further improve the recognition performance.

In this paper, a graphical modeling-based approach to semantic video analysis is presented for jointly realizing modality fusion and temporal context exploitation. Initially, the examined video sequence is segmented into shots. Following shot detection, appropriate color, motion and audio features are extracted for each shot. Then, HMMs are employed for performing an initial association of each shot with the semantic classes that are of interest separately for every modality. Subsequently, an integrated BN is introduced, which aims to simultaneously handle the problems of modality fusion and temporal context modeling, contrary to the usual practice of performing each task separately. The developed BN is provided with an appropriate network structure, capable of probabilistically learning the causal relationships and the complex interdependencies that are present among the audio-visual data. The utilized contextual information has the form of the temporal relations among the supported classes. The final outcome of the overall video analysis approach is the association of a semantic class with every shot.

The paper is organized as follows: Section 2 describes the video pre-processing steps. Section 3 outlines the single-modality shot-class association procedure. Section 4 details the proposed joint fusion and temporal context exploitation approach. Experimental results as well as comparative evaluation from the application of the proposed approach in the domain of news broadcast video are presented in Section 5, and conclusions are drawn in Section 6.

## 2. VIDEO PRE-PROCESSING

Objective of this work is the association of each shot  $s_i$ ,  $i = 1, \dots, I$ , of a video with one of the semantic classes of a set  $E = \{e_j\}_{1 \leq j \leq J}$  that are of interest in the given application domain. For shot detection, the algorithm of [6] is used, mainly due to its low computational complexity. Following temporal segmentation, appropriate motion, color and audio features are extracted for each resulting shot. In particular, the *accumulated motion energy fields*, introduced in [7], are utilized for providing local-level motion information to HMMs; thus,

The work presented in this paper was supported by the European Commission under contracts FP6-045547 VIDI-Video, FP6-027685 MESH, FP6-027026 K-Space and FP7-214306 JUMAS.

forming the shot's motion observation sequence. For the color and audio information processing, common techniques of the relevant literature are adopted; specifically, a set of global-level color histograms in the RGB color space and the widely used Mel Frequency Cepstral Coefficients (MFCC) are estimated for each shot, respectively. These features are used to form the corresponding shot's color and audio observation sequences.

### 3. HMM-BASED CLASSIFICATION

HMMs are employed in this work for performing an initial shot-class association based on single-modality information, due to their suitability for modeling pattern recognition problems that exhibit an inherent temporality. In particular, a set of  $J$  HMMs is employed with respect to the motion modality, where an individual HMM is introduced for every defined semantic class  $e_j$ . Each HMM receives as input the motion observation sequence (described in Section 2) and at the evaluation stage returns a posterior probability, which represents the observation sequence's fitness to the particular model. This probability, denoted by  $h_{ij}^m$ , indicates the degree of confidence with which class  $e_j$  is associated with shot  $s_i$  based on motion information. Similarly to the motion analysis case, two individual sets of  $J$  HMMs each are employed with respect to the color and audio modalities. The corresponding posterior probabilities for shot-class association are denoted by  $h_{ij}^c$  and  $h_{ij}^a$ , respectively. It must be noted that a set of annotated video content, denoted by  $U_{tr}^1$ , is used for training the developed HMM structure. HMM implementation details can be found in [7].

### 4. 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 [8]. In this work, BNs, which are a particular type of graphical models, are employed. In particular, an integrated BN is proposed for jointly performing modality fusion and temporal context modeling. It is provided with an expandable network structure, which enables contextual knowledge acquisition and modeling in the form of the temporal relations among the supported high-level semantic classes, and is capable of incorporating information from different sources. To this end, a series of sub-network structures, which are appropriately integrated to the overall network, are defined.

#### 4.1. Modality Fusion

A BN structure is initially defined for performing the fusion of the computed single-modality analysis results, and a set of  $J$  such structures is introduced to this end, 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. 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 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

[8]. Hence, the set of values that variable  $CL_j$  can receive is chosen equal to  $\{cl_{j1}, cl_{j2}\} = \{True, False\}$ . On the other hand, a discretization step is applied to the estimated 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 variables, which 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: after a set of annotated video content, denoted by  $U_{tr}^2$ , is formed, and for which the initial shot-class association results are computed for each shot, the estimated posterior probabilities are grouped with respect to every possible class-modality combination, forming sets  $L_j^b = \{h_{nj}^b\}_{1 \leq n \leq N}$ , where  $b = \{a, c, m\} \equiv \{audio, color, motion\}$  is the modality used and  $N$  is the number of shots in  $U_{tr}^2$ . Then, the elements of the aforementioned sets are sorted in ascending order, and the resulting sets are denoted by  $\hat{L}_j^b$ . If  $Q$  denotes the number of possible values of every corresponding random variable, then:

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

where  $K = \lfloor \frac{N}{Q} \rfloor$ ,  $\hat{L}_j^b(o)$  denotes the  $o^{th}$  element of set  $\hat{L}_j^b$ , and  $b_{j1}, b_{j2}, \dots, b_{jQ}$  denote the values of variable  $B_j$  ( $B = \{A, C, M\}$ ).

Regarding the construction of the respective network structure, which is denoted by  $\mathbb{G}_j$ , variable  $CL_j$  is chosen to correspond 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. The direction of the arcs in the proposed network structure, which is illustrated in Fig. 1a), defines explicitly the causal relationships / conditional independence assumptions among the defined variables. In particular, it is assumed that the semantic class, to which a video shot belongs, fully determines the features observed with respect to every modality.

#### 4.2. Integration of Modality Fusion and Temporal Context Exploitation

Besides multi-modal information, contextual information can also contribute towards improved shot-class association performance. In this work temporal contextual information 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. Thus, information about the classes' occurrence order can serve as a set of constraints denoting their 'allowed' temporal succession. Advantageous characteristics of the presented method include: a) it encompasses a probabilistic approach for acquiring complex contextual knowledge after a training procedure is applied, and b) contextual constraints are applied within a restricted time interval, i.e. whole video structure parsing is not required for reaching good recognition results.

Under the proposed approach, an individual BN structure is introduced, aiming to enable the acquisition of the appropriate implicit knowledge that will be utilized for performing temporal context exploitation. For that purpose, the respective BN takes 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, in order to decide upon the class that is eventually associated with shot  $s_i$ . For achieving this, an appropriate set of random variables needs to be defined, similarly to the case of the development of the BN used for modality fusion in Section 4.1. Specifically,

for the task of temporal context exploitation the following random variables are defined: a) a set of  $J$  variables ( $CL_j^i$ ), one for every defined class  $e_j$ ; these variables represent the classes that are eventually associated with shot  $s_i$ , and b) two sets of  $J \cdot TW$  variables ( $CL_j^{i-r}$  and  $CL_j^{i+r}$ ), which denote the shot-class associations of previous and subsequent shots, respectively;  $r \in [1, TW]$ , 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 their possible values, this is chosen equal to  $\{cl_{j1}^k, cl_{j2}^k\} = \{True, False\}$ . Additionally, the respective BN structure, denoted by  $\mathbb{G}_c$ , is illustrated in Fig. 1b). As can be seen from this figure, each random variable  $CL_j^k$  is defined to be conditionally dependent only on variables  $CL_j^{k-1}$ .

Having developed the structures  $\mathbb{G}_c$  and  $\mathbb{G}_j$  (Section 4.1), 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 a set of sub-structures integrated to the defined structure depicted in Fig. 1b), and encodes both cross-modal as well as temporal relations among the supported semantic classes.

Regarding the training process of the integrated BN, the set of all conditional probabilities among the defined conditionally-dependent random variables of  $\mathbb{G}$  are estimated from the set of annotated video content  $U_{tr}^2$  which was also used in Section 4.1 for input variable discretization. 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  $Z_i = \{a_j^k, c_j^k, m_j^k\}_{1 \leq j \leq J, i-TW \leq k \leq i+TW}$ , where  $a_j^k, c_j^k$  and  $m_j^k$  are the values of the variables  $A_j^k, C_j^k$  and  $M_j^k$  of  $\mathbb{G}$ , respectively. 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  $\mathbb{G}$ :  $P(CL_j^i = True | Z_i)$ , for  $1 \leq j \leq J$ . 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. EXPERIMENTAL RESULTS

The proposed approach was experimentally evaluated and compared with literature approaches using news broadcast videos. It should be emphasized here that application of the proposed approach 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 required. For the selected domain, the following semantic classes were defined: *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 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 32

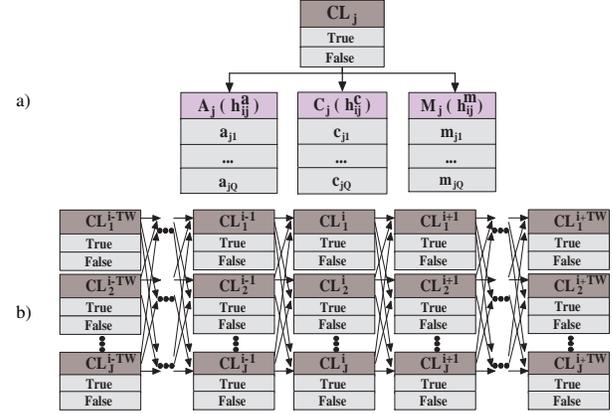


Fig. 1. BN structure for: a) modality fusion and b) temporal context.

videos of news broadcast from Deutsche Welle<sup>1</sup> were collected. After the temporal segmentation algorithm of [6] was applied, a corresponding set of 1188 shots was formed, which were manually annotated according to the respective domain class definitions. From the aforementioned videos, 8 of them (total of 338 shots; anchor: 70, reporting: 46, reportage: 174, graphics: 48) were used for training the developed HMM structure (training set  $U_{tr}^1$ ), 16 of them (total of 557 shots; anchor: 80, reporting: 71, reportage: 337, graphics: 69) were used for training the integrated BN (training set  $U_{tr}^2$ ), and the remaining 8 videos (total of 293 shots; anchor: 59, reporting: 28, reportage: 174, graphics: 32) were used for evaluation (test set  $U_{te}$ ).

After video temporal segmentation to shots, for every resulting shot appropriate audio, motion and color features were extracted, as described in Section 2. The estimated features were provided as input to the developed HMM structure for performing an initial shot-class association separately for every utilized modality (Section 3). Subsequently, the integrated BN described in Section 4 was used for realizing joint information fusion and temporal context exploitation. The value of variable  $Q$  in Eq. (1), which determines the number of discrete values of each random variable in every BN sub-network used for information fusion, was set equal to 11. Significantly lower values led into coarse discretization, which resulted into poor recognition performance. On the other hand, significantly greater values led to increased network complexity and resulted into under-training occurrences. The developed BN was trained using the Expectation Maximization (EM) approach, which was experimentally shown to lead to better classification results than other learning methods like frequency counting and gradient descend-based. Moreover, probability propagation was realized using a junction tree mechanism [9].

In Table 1, quantitative class association results are given in the form of the calculated confusion matrices, when solely single-modality information is used, as well as when the introduced integrated BN is applied for  $TW = 1, 2$  and 3. Additionally, the value of the overall classification accuracy, i.e. the percentage of the video shots that are correctly classified, is also given for every case. For calculating these results, it has been considered that  $\arg \max_j (h_{ij}^a)$ ,  $\arg \max_j (h_{ij}^c)$ ,  $\arg \max_j (h_{ij}^m)$  and  $\arg \max_j (h_{ij}^f)$ , indicate the class  $e_j$  that is associated with shot  $s_i$  after every respective algorithmic step.

From the presented results, it can be seen that the integrated BN

<sup>1</sup><http://www.dw-world.de/>

**Table 1.** Semantic class association results ( $e_1$ : anchor,  $e_2$ : reporting,  $e_3$ : reportage and  $e_4$ : graphics).

Method	Actual Class	Associated Class				Overall Accuracy
		$e_1$	$e_2$	$e_3$	$e_4$	
Audio	$e_1$	72.88%	27.12%	0.00%	0.00%	59.73%
	$e_2$	0.00%	67.86%	28.57%	3.57%	
	$e_3$	2.87%	31.61%	62.64%	2.87%	
	$e_4$	28.13%	18.75%	40.63%	12.50%	
Color	$e_1$	54.24%	16.95%	28.81%	0.00%	72.70%
	$e_2$	0.00%	14.29%	85.71%	0.00%	
	$e_3$	0.57%	12.07%	87.36%	0.00%	
	$e_4$	0.00%	0.00%	21.88%	78.13%	
Motion	$e_1$	83.05%	8.47%	8.47%	0.00%	80.55%
	$e_2$	0.00%	<b>75.00%</b>	25.00%	0.00%	
	$e_3$	1.15%	13.22%	83.91%	1.72%	
	$e_4$	3.13%	3.13%	31.25%	62.50%	
Fusion and Context TW=1	$e_1$	<b>93.22%</b>	3.39%	3.39%	0.00%	90.10%
	$e_2$	7.14%	64.29%	28.57%	0.00%	
	$e_3$	1.72%	4.02%	93.68%	0.57%	
	$e_4$	3.13%	3.13%	6.25%	87.50%	
Fusion and Context TW=2	$e_1$	<b>93.22%</b>	3.39%	3.39%	0.00%	90.78%
	$e_2$	0.00%	64.29%	35.71%	0.00%	
	$e_3$	0.57%	3.45%	94.25%	1.72%	
	$e_4$	3.13%	0.00%	6.25%	90.63%	
Fusion and Context TW=3	$e_1$	<b>93.22%</b>	3.39%	3.39%	0.00%	<b>91.47%</b>
	$e_2$	0.00%	67.86%	32.14%	0.00%	
	$e_3$	0.57%	4.60%	93.68%	1.15%	
	$e_4$	3.13%	0.00%	0.00%	<b>96.88%</b>	
Method of [5]	$e_1$	59.32%	0.00%	38.98%	1.69%	81.23%
	$e_2$	0.00%	14.29%	85.71%	0.00%	
	$e_3$	0.00%	0.57%	<b>99.43%</b>	0.00%	
	$e_4$	0.00%	0.00%	18.75%	81.25%	
Method of [4]	$e_1$	77.97%	0.00%	22.03%	0.00%	82.25%
	$e_2$	3.57%	21.43%	75.00%	0.00%	
	$e_3$	0.00%	6.90%	93.10%	0.00%	
	$e_4$	3.13%	0.00%	12.50%	84.38%	

achieves an increase of 9.55% (for  $TW = 1$ ) in the overall classification accuracy, compared to the best single-modality results, while the recognition rates of most of the supported classes are significantly improved. Additionally, it is also shown that further increase of the value of parameter  $TW$  ( $TW = 2, 3$ ) leads to a corresponding increase of the overall classification accuracy (10.23% and 10.92%, respectively). Moreover, it can be observed that it is the class graphics that presents the highest performance improvement (exhibiting an increase of 18.75% for  $TW = 3$  compared to the best single-modality results). This is due to the fact that shots belonging to the class graphics exhibit significant variations in the video low-level 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 (video shots belonging to the class graphics are usually successive). Hence, modality fusion and temporal context exploitation can contribute towards improving the recognition rate of this class. Greater values of parameter  $TW$  were shown to result into negligible variations in the overall classification performance.

In order to comparatively investigate the performance of the developed framework, the HMM-based video analysis approaches presented in [4] and [5] were implemented and evaluated. Specifically, Xu et al. [4] present a HMM-based framework capable of modeling temporal contextual constraints in different semantic granularities, while multistream HMMs are used for modality fusion. On the other hand, Huang et al. [5] propose a ‘class transition penalty’ approach, where HMMs are initially employed for detecting the semantic classes of concern by means of multi-modal information,

and subsequently a dynamic programming technique is adopted for searching for the most likely class transition path. It must be noted that the same video low-level features utilized by the proposed approach (Section 2) were also provided as input to the methods of [4] and [5]. According to the results presented in Table 1, it can be seen that the introduced framework outperforms the aforementioned approaches for most of the supported classes as well as in overall classification accuracy. This is mainly due to: a) the more sophisticated modality fusion mechanism developed, compared to the arbitrary assignment of weights to every modality in [4], and the assumption of statistical independence between the features of different modalities in [5], and b) the more complex temporal relations that are modeled by the proposed BN, compared to the methods of [4] and [5] that rely on class transition probability learning. Moreover, the proposed method performs jointly modality fusion and temporal context exploitation; hence, taking advantage of the possible correlations between the respective numerical data.

## 6. CONCLUSIONS

In this paper, a generic graphical modeling-based approach to semantic video analysis, capable of jointly performing modality fusion and temporal context exploitation, was presented. Experimental results in the domain of news broadcast video demonstrated the efficiency of the proposed method. Future work includes the investigation of alternative algorithms for acquiring contextual knowledge and their integration in the proposed framework, like the notion of semantic concept co-occurrence in a video shot, which can be realized by introducing an appropriate detector for every concept and subsequently associating each of them with children nodes in the developed BN structure  $\mathbb{G}_j$ .

## 7. REFERENCES

- [1] A.W.M. Smeulders et al., “Content-based image retrieval at the end of the early years,” *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 22, no. 12, pp. 1349–1380, 2000.
- [2] B.T. Truong, C. Dorai, and S. Venkatesh, “Automatic Genre Identification for Content-based Video Categorization,” *Int. Conf. on Pattern Recognition*, vol. 15, pp. 230–233, 2000.
- [3] C.G.M. Snoek and M. Worring, “Multimodal Video Indexing: A Review of the State-of-the-art,” *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.
- [4] G. Xu, Y.F. Ma, H.J. Zhang, and S.Q. Yang, “An HMM-based framework for video semantic analysis,” *CSVT, IEEE Trans. on*, vol. 15, no. 11, pp. 1422–1433, 2005.
- [5] 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.
- [6] V. Kobla, D. Doermann, and K. Lin, “Archiving, indexing, and retrieval of video in the compressed domain,” *SPIE Conf. on Multimedia Storage and Archiving Systems*, pp. 78–89, 1996.
- [7] G.Th. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, “Accumulated motion energy fields estimation and representation for semantic event detection,” *Int. Conf. on Content-based Image and Video Retrieval*, pp. 221–230, 2008.
- [8] D. Heckerman, “A tutorial on learning with Bayesian networks,” *Learning in Graphical Models*, pp. 301–354, 1998.
- [9] F. V. Jensen and F. Jensen, “Optimal junction trees,” *Proc. of Conf. on Uncertainty in Artificial Intelligence*, 1994.