A bayesian network modeling approach for cross media analysis

Christina Lakka[a], Spiros Nikolopoulos[a,c], Christos Varytimidis[b], Ioannis Kompatsiaris[a]

[a]*Informatics and Telematics Institute, CERTH, 6th km Charilaou-Thermi Road, Thessaloniki - Greece, Tel: +30-2311/257.701-3, Fax: +30-2310/474.128*
[b]*School of Electrical and Computer Engineering, National Technical University of Athens - Greece*
[c]*School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK*

Abstract

Existing methods for the semantic analysis of multimedia, although effective for single-medium scenarios, are inherently flawed in cases where knowledge is spread over different media types. In this work we implement a cross media analysis scheme that takes advantage of both visual and textual information for detecting high-level concepts. The novel aspect of this scheme is the definition and use of a conceptual space where information originating from heterogeneous media types can be meaningfully combined and facilitate analysis decisions. More specifically, our contribution is on proposing a modeling approach for Bayesian Networks that defines this conceptual space and allows evidence originating from the domain knowledge, the application context and different content modalities to support or disproof a certain hypothesis. Using this scheme we have performed experiments on a set of 162 compound documents taken from the domain of car manufacturing industry and 118581 video shots taken from the TRECVID2010 competition. The obtained results have shown that the proposed modeling approach exploits the complementary effect of evidence extracted across different media and delivers performance improvements compared to the single-medium cases. Moreover, by comparing the performance of the proposed approach with an

Email addresses: lakka@iti.gr (Christina Lakka), nikolopo@iti.gr (Spiros Nikolopoulos), chrisvar@image.ntua.gr (Christos Varytimidis), ikom@iti.gr (Ioannis Kompatsiaris)

approach using Support Vector Machines(SVM), we have verified that in a cross media setting the use of generative rather than discriminative models are more suited, mainly due to their ability to smoothly incorporate explicit knowledge and learn from a few examples.

---

1. Introduction

The automatic extraction of semantic metadata from multimedia content has been recognized as a particularly valuable task for various applications of digital content consumption. Current literature has made considerable progress in this direction especially for single-medium scenarios. However, the methods proposed in the literature do not apply in cases where information is spread over different media types and unless considered simultaneously, its contribution cannot be fully exploited by the analysis process. Motivated by this, cross media analysis seeks to enhance semantic metadata extraction by exploiting information across media. Practically, the aim of such methods is to combine the evidence extracted from different media types and accumulate their effect in favor or against a certain hypothesis. These pieces of evidence can belong to different levels of granularity and used differently by the analysis mechanism. For instance, we can consider cross media analysis to be a general fusion problem that is carried out at different levels of abstraction, namely result-level [1], [2], [3], extraction-level [4], [5], [6] and feature-level [7], [8], [9].

In the result-level approach, information from each data source is initially extracted separately and, still separately, transformed into conceptual information. Though result-level approaches are closer to human cognition and more suited for exploiting explicit knowledge (i.e., knowledge that is explicitly provided by experts in the form or rules, ontologies or other formal languages for knowledge representation), their major drawback is that each extractor has to produce its own internal evidence in order to extract the conceptual information. In the extractor-level approach the conceptual information is not extracted separately from each modality but instead, the analysis mechanism takes into account evidence from other modalities. Information coming from one medium may assist the information extraction module of another medium, using as input the output of another extractor. However, in contrast to the result-level approaches where knowledge is incorporated into the conceptual space, in this case it can only be exploited as

part of a task specific mechanism. The feature-level approach consists in using all low-level features that can be extracted from each medium within the same analysis process. Initially, low-level features (e.g., text-tokens, named entities or image descriptors) are extracted separately from each modality and integrated into a common, concatenated representation. Subsequently, the common representation is used as input for the analysis process (i.e., classification, indexing, etc). Feature-level analysis aims at exploiting the joint existence of low-level features into the same resource, but it is rather difficult to incorporate explicit knowledge in this case.

Our work is motivated by the need to boost the efficiency of cross media analysis using the knowledge explicitly provided by domain experts (i.e., domain knowledge). This was the reason for developing a method that operates on the result-level of abstraction and allows domain knowledge to become part of the inference process. Our method combines the soft evidence (soft in the sense that a confidence degree is attached to every piece of evidence) collected from different media types, to support or disproof a certain hypothesis made about the semantic content of the analyzed resource. Soft evidence are obtained by applying single-medium analyzers on the low-level features of the different media types. Subsequently, these pieces of evidence are used to drive a probabilistic inference process that takes place in a Bayesian Network (BN). The structure and parameters of the BN are constructed by incorporating domain knowledge (expressed using ontologies) and application context (captured by conditional probabilities). We use the soft evidence to update the observable variables of the BN and verify or reject the examined hypothesis based on the posteriori probability of the remaining variables. Fig. 1 demonstrates the functional relations between the components of the proposed cross media analysis scheme.

The novelty introduced by the proposed method is that it manages to integrate into a common inference framework three types of information, a) information obtained from the analysis of heterogeneous content (i.e., the output of single-medium analyzers supplied as soft evidence), b) information about the domain that is provided explicitly, and c) contextual information that is learned from sample data. Our contribution is on proposing a modeling approach for the BN that results in a conceptual space of likelihood estimates. In this space the evidence originating from the domain knowledge, the application context and the different content modalities can be meaningfully combined and facilitate semantic metadata extraction.

We show using content from a real world application taken from the car

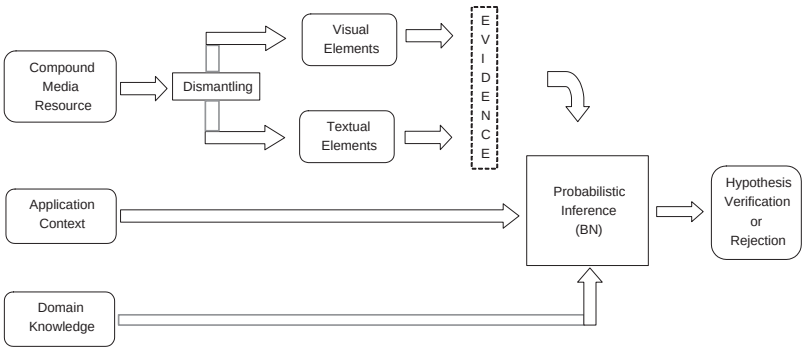


Figure 1: Cross media analysis scheme

manufacturing industry as well as from the TRECVID2010 competition, that performing cross media analysis using the proposed method leads to significant improvements compared to the cases where single-medium analyzers act separately. We also prove experimentally that, in a cross media setting, the generative models outperform the discriminative ones in fusing the extracted evidence, mainly due to their ability in efficiently handling prior knowledge and learning from a few examples.

The rest of the manuscript is organized as follows. Section 2 details the proposed approach for modeling the BN and determining the conceptual space. Section 3 describes the implemented cross media analysis scheme including details for the utilized single-medium analyzers, as well as the methodology adopted for integrating implicit and explicit knowledge in the BN. Experimental results are presented in Section 4. Section 5 reviews the related literature, while Section 6 concludes our paper and provides references to future work.

2. Modeling the Bayesian Network

In this section we describe the main contribution of our work, which is a generic approach for modeling BNs that can be used to define a conceptual space suitable for combining heterogeneous types of information. The types of information that are handled by this approach are: a) conceptual information shared amongst most individuals that determines the logical relations between concepts, such as sub-class, union, intersection, disjoint, etc (i.e., domain knowledge), b) information that qualitatively evaluates the co-existence of concepts, encoding for example how likely is for one concept to be present

when another concept is verified (i.e., application context), and c) information extracted from content analysis that encodes the support received from the analyzed low-level features in favor of a specific concept (single-medium evidence). Our approach relies on probabilities and probabilistic inference to define the common conceptual space.

More specifically, the explicitly provided domain knowledge is used to determine the structure of the BN and in this way enforce the logic rules of the domain during inference. The application context is approximated by the co-occurrence frequency between domain concepts, information that can be extracted using a sample of the population that is being modeled. The application context is encoded into the Conditional Probability Tables (CPTs) of the BN nodes, which influence the inference process when belief propagation takes place. However, the most critical point is how to incorporate the information received from content analysis. In order to do this, we treat the outcome of single-medium analyzers as soft evidence that are used to instantiate the nodes of a BN operating on a conceptual true-false space. The reason for selecting these states (i.e., true, false) to be the only possible states of all network nodes, was to establish a "lingua franca" between the heterogeneous types of information and facilitate the incorporation of domain knowledge in decision making. By adopting the proposed modeling approach the constructed BN does not operate on the low-level features of the content, which would constitute a typical application of the BN theory. Instead, it operates on the space determined by the probability estimates (that we call conceptual true-false space), obtained through the application of machine learning techniques on the low-level features (as described later in Section 3.2). In the following, we provide details on how the proposed modeling approach can be used to determine a BN for analyzing compound documents, but can be seamlessly applied to analyze any other multi-modal resource, or handle an arbitrary number of modalities.

Let us consider a set of compound documents D where each document is composed of its visual and textual part:

$$D_i = [T_i, V_i] \tag{1}$$

Let also t_i, v_i be the features extracted from T_i, V_i respectively. We consider the single-medium analyzers to be the functions $f_{c_j}(\cdot)$ and $g_{c_j}(\cdot)$ that outputs the probability of a given concept c_j being valid for a document either based on its textual or visual low-level features, respectively:

$$\begin{array}{l} f_{c_j}(T_i) = P(c_j = true|t_i), \quad \text{for the textual part of } D_i \\ g_{c_j}(V_i) = P(c_j = true|v_i), \quad \text{for the visual part of } D_i \end{array} \tag{2}$$

Thus, if we have a single-medium analyzer that is trained to detect all domain concepts $\forall c_j \in C$, it produces $|C|$ probabilities when applied on a document D_i. In order to construct a BN that operates on a conceptual true-false, for every concept c_j we create a discrete random variable with two states $r_z = \{true, false\}$. Then, we link these nodes based on their logical relations (as explained in Section 3.3.1) and learn the CPTs by applying the Expectation Maximization algorithm on sample data (as detailed in Section 3.3.1). We consider the output of the single-medium analyzer to formulate a new feature space y, determined from the probability estimates. We refer to this new feature space as conceptual true-false space. By applying the Bayes rule in feature space y we have for each concept c_j:

$$P_{c_j}(r_z|y) = \frac{P_{c_j}(y|r_z)P_{c_j}(r_z)}{P_{c_j}(y)}, \quad \forall c_j \in C \tag{3}$$

$P_{c_j}(r_z)$ represents our prior knowledge about c_j and in the conceptual true-false space we accept that $P_{c_j}(r_z = true)$ is equal to the frequency of appearance of c_j in the domain (i.e., how often appears in the training set). Respectively, we accept that $P_{c_j}(r_z = false) = 1 - P_{c_j}(r_z = true)$. $P_{c_j}(y)$ is a scale factor that guarantees that the posterior probabilities sum to one and equals:

$$P_{c_j}(y) = \sum_{r_z \in \{true, false\}} P_{c_j}(y|r_z)P_{c_j}(r_z) \tag{4}$$

$P_{c_j}(y|r_z)$ is the likelihood (or class conditional probability) of r_z with respect to y. $P_{c_j}(r_z|y)$ is the posterior probability of r_z after considering the analysis outcome and taking into consideration prior knowledge. In order to facilitate the analysis process we need to calculate the posterior probabilities for each independent piece of conceptual information (i.e., $\forall c_j \in C$), so we need to know $P_{c_j}(r_z = true|y)$. It is clear from eqs. (3) and (4) that in order to calculate this value, what we are missing is $P_{c_j}(y|r_z = true)$ and $P_{c_j}(y|r_z = false)$. Recalling that $f_{c_j}(\cdot)$ and $g_{c_j}(\cdot)$ provides us with a probability expressing how much support c_j receives from the textual or visual low-level features of the document respectively, we incorporate the

content analysis outcome into the decision process by instantiating the nodes of the BN as follows:

$$P_{c_j}(y|r_z = true) = \begin{cases} f_{c_j}(T_i), & \text{for textual evidence} \\ g_{c_j}(V_i), & \text{for visual evidence} \end{cases} \quad (5)$$

$$P_{c_j}(y|r_z = false) = \begin{cases} 1 - f_{c_j}(T_i), & \text{for textual evidence} \\ 1 - g_{c_j}(V_i), & \text{for visual evidence} \end{cases} \quad (6)$$

Thus, during the analysis process we inject, as explained above, the output of single-medium analyzers into the BN and perform probabilistic inference by propagating evidence beliefs. Eventually, the resulting posterior probability for the "true" state of the node corresponding to the concept that we want to detect, is considered to be the confidence degree for this concept.

3. Cross media analysis scheme

In order to verify the benefits of the proposed modeling approach we use it to design a cross media analysis scheme that detects high-level concepts. High-level concept detection is usually the output of knowledge-related tasks and typically requires the synergy of information scattered in different places. The more the available information, the more easily is for the knowledge worker to infer the presence of a high-level concept. Independently of whether these pieces of information act cumulatively or complementary, they have an impact (i.e., positive or negative) on the confidence of the fact that a certain high-level concept is valid for the analyzed resource. In order to model this process we rely on the approach presented in Section 2 and implement a generative classifier based on BNs. The role of this classifier is to i) fuse the information extracted from different media types on the grounds of knowledge and context, ii) produce a confidence degree about the validity of a high-level concept in the analyzed resource, and iii) make a decision by applying a fixed threshold on this confidence degree. Since cross media analysis is mostly about simultaneously evaluating the appropriate evidence extracted across different media types, an important issue for making the aforementioned framework suitable for such purposes is the strategy by which evidence (and as a consequence their source modalities) are considered to be co-related.

In the following subsections we elaborate on the components that are used to implement the cross media analysis scheme for compound documents,

which are: a) a dismantling mechanism and a modality synchronization strategy for handling the compound media resources, b) the single-medium analysis techniques for extracting evidence using low level features, and c) the techniques used to construct and perform inference on a BN that is modeled as described in Section 2.

3.1. Compound documents dismantling & modality synchronization

Compound documents are multimedia documents that incorporate more that one media types in the same digital resource. OpenDocument, Microsoft Office's documents, PDF and web pages are indicative representation formats of such documents where visual and textual elements co-exist. A compound document may contain evidence for a high-level concept to be extracted across different media. However, it is not straightforward to know which media elements refer to the same high-level concept. Moreover, these documents carry additional information such as cross references or layout features (e.g., spatial proximity between a caption and an image frame) that have a major effect on the content essence. These features, although very important for human perception, are difficult for knowledge extraction algorithms to encode and exploit.

Document processing literature discusses several approaches to extract layout information from PDF, HTML and other structured documents, see [10] for an overview. Most of these approaches [11], [12] are based on manual or semi-automatically extracted templates that characterize each part of the document. However, the variety of layouts that a document editor is likely to use for expressing the intended meaning, makes it difficult for automated systems to consistently model them and make them available for analysis. This process is further hindered by the absence of a uniform document representation standard that could reduce the diversity of existing formats.

All the above, makes the employment of a dismantling and synchronization mechanism an important module of cross media analysis. This mechanism will be able to disassemble a compound document to its constituent parts and decide which of these parts should be considered simultaneously by the fusion process. For the purposes of our work, assuming a certain layout for the analyzed documents, we accept that a different topic is covered in each document page and disregard cases where more that one topics exist in the same page or a single topic extents to many pages. Thus, all media elements of the same document page are considered to be conceptually related. Given this assumption, we analyze a document on a per page basis

by fusing the output of single-medium analyzers that are independently applied on the media elements residing on the same page. Although such an assumption may seem inconsistent with a non-negligible number of cases, in this work we basically focus on how to effectively fuse cross media evidence on the grounds of knowledge and context, while existing approaches can be employed in cases where this assumption does not hold.

3.2. Single-medium analysis techniques

In this section we detail the techniques we have used to analyze the low-level features of a document and produce confidence degrees for the related concepts. It must be noted that although the framework described in Section 3 concerns the analysis of compound documents containing images and text, the part related to the methodology proposed for determining the common conceptual space, can be seamlessly applied to analyze any other type of multimedia resource (e.g., containing video or sound), provided that the corresponding single-medium analyzers are available. One such example is described in Section 4.6, where we use the proposed modeling approach to implement a video shot classification scheme.

3.2.1. Visual analysis

Visual evidence is extracted by applying concept detectors on the images contained in a document page. The method adopted for implementing the concept detectors is based on the Viola and Jones detection framework [13]. The functionality of this framework can be characterized by three key aspects, a) a scheme for image representation called *integral image*, that allows for very fast feature extraction, b) a method for constructing a classifier by selecting a small number of important features using AdaBoost [14], and c) a method for combining successively more complex classifiers in a cascade structure, which dramatically increases the speed of the detector by focusing attention on promising regions of the image.

In more detail, the visual information contained in an image is described by Haar-like features, introduced in [15] and depicted in Fig. 2. The values of these features are the differences between the sums of the white and black rectangular regions. In order to compute these sums efficiently, Viola and Jones make use of integral images. An integral image is an array corresponding to an image that contains in position (x, y) the sum of the intensity values of all pixels above and to the left of (x, y). For the Haar features that are rotated by 45^o, a rotated integral image is used, which accumulates the

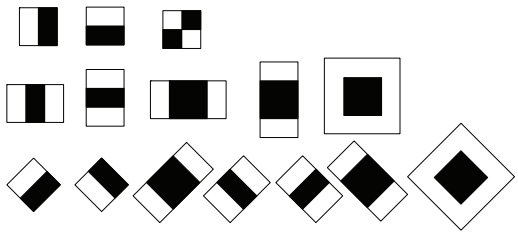


Figure 2: Haar-like features. The values of these features are the differences between the sums of the white and black rectangular regions.

values inside a triangle starting from point (x, y) and ending at the top of the image. The construction of an integral image requires a linear scan through the actual image and results in computing the feature responses in constant time. The efficient computation of the feature responses is essential, since all of them are computed at all positions and scales in an image, resulting in a very dense representation of approximately 100,000 feature responses for an image of size 20x20 pixels.

Then, the AdaBoost algorithm is used in order to train a classifier for an object category. AdaBoost creates a degenerate decision tree based on the responses of m Haar features that best describe the depicted concept. Classification time is reduced by using several low precision, fast classifiers connected in a cascade, instead of one high precision and slow classifier. In order to classify a sub-window of an image as positive (depicting the object), the sub-window has to be classified as positive by all the classifiers in the cascade, also called stages. If a sub-window is classified as negative (not depicting the object) by any single classifier, then it is rejected and not processed by the following stages, as depicted in Fig. 3. The detection task of finding the precise position and scale of the object is performed in a sliding window manner, checking every possible position and scale.

The output of the local concept detector is the exact position and scale at which a concept c_j was found in the analyzed image, as well as a confidence degree associated to every detection result. The confidence degree is extracted from the detectors inner structure, as depicted in Fig. 3. More specifically, the output of each classifier in the cascade is associated with a confidence degree S_i derived by a combination of the calculated decision thresholds. These thresholds are related to the partial or full detection of the concept of interest. The values extracted from all classifiers of the cascade, are then combined in a weighted sum to provide a confidence value for

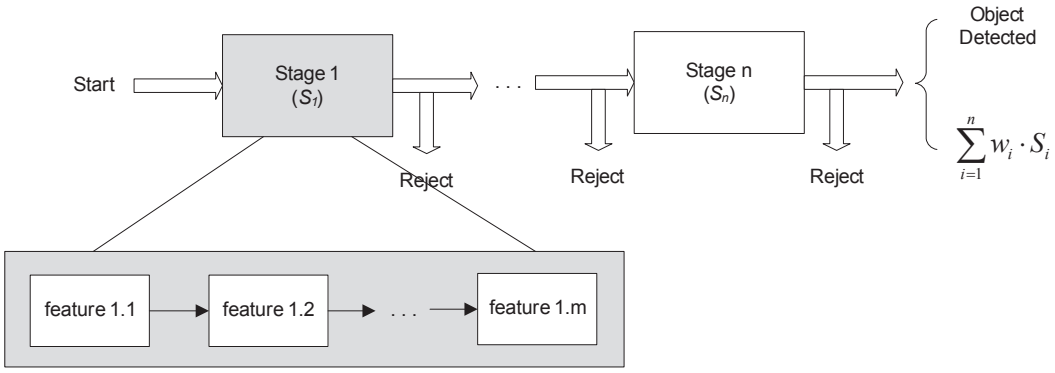


Figure 3: Confidence value derived from the cascade of classifiers.

each examined sub-window. The weights w_i applied to each stage output, emphasize the response of the last stages which are more discriminative than the initial low precision ones. The confidence value is then normalized in [0 1], based on the training set used to create the detector. For the purposes of our work we filter out cases with a very small confidence degree and we select the case with maximum confidence degree when multiple instances of the same concept are found on the same image.

3.2.2. Textual analysis

For obtaining textual evidence we need to estimate the semantic relatedness of a concept with the linguistic information contained in a document page. In order to do so, we should be able to measure the semantic relatedness between any two individual concepts and apply a page oriented summarization strategy, as detailed later in this section. Approximating human judgement and measuring the semantic relatedness between concepts has been a challenging task for many researchers. Most works in the literature make use of the WordNet lexical database [16] for achieving this objective.

WordNet can be viewed as a large graph where each node represents a real world concept and each link between nodes represents a relationship between the corresponding concepts. Every node consists of a set of words (synset), that linguistically describe the real world concept associated with the node, as well as a short description of this concept (gloss). Using the above, WordNet encodes a significantly large amount of knowledge and is able to facilitate a great number of methods determining the semantic relatedness between con-

cepts. Methods existing in the literature can be divided to the ones that use only the structure and content of WordNet to measure semantic relatedness [17], while others achieve this by also exploiting statistical data from large corpora, [18], [19], [20], [21], [22]. Another important characteristic of such methods is whether they are able to operate on all parts of speech [22], [20] or nouns only [17], [18], [19], [21]. For the purposes of our work we decided to employ a semantic relatedness measure that is based on context vectors and was originally presented by Patwardhan in [23]. The method introduced in this work relies on a different representation for WordNet glosses that is based on multidimensional vectors of co-occurrence counts. Its main advantage derives from its ability to combine the benefits of methods that use the knowledge from a large data corpus and the ones that rely solely on the strict definitions of WordNet (glosses).

In order to describe the method in more detail we need to determine the meaning of *word vectors* and *context vectors*. Every word in the word space has a corresponding word vector. The word vector corresponding to a given word is calculated as a vector of integers. The integers are the frequencies of occurrence of each word from the word space in the context. The context of a word is considered to be the words that appear close in the text with this word. Thus, each word in the word space represents a dimension of the vector of integers. Once the word vectors for all words in the word space are calculated, they are used to calculate the context vectors for every instance of a word. This is done by adding the word vectors of all words that appear in the context of this word.

In order to measure the semantic relatedness between two concepts the method of [23] represents each concept in WordNet by a *gloss vector*. A *gloss vector* is essentially a context vector formed by considering a WordNet gloss as the context. More specifically, having created the word vectors for all words in the word space, the gloss vector for a WordNet concept is created by adding the word vectors of all words contained in its gloss. For example, consider the gloss of *lamp - an artificial source of visual illumination*. The gloss vector for lamp would be formed by adding the word vectors of *artificial*, *source*, *visible* and *illumination*.

Eventually, the semantic relatedness between two concepts is defined as the cosine of the angle between the corresponding normalized gloss vectors:

$$SemanticRelatedness(c_1, c_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|v_1|v_2|} \tag{7}$$

where c_1, c_2 are the two given concepts, and $\overrightarrow{v_1}$, $\overrightarrow{v_2}$ are the gloss vectors corresponding to the concepts. The motive behind our choice of relying on context vectors over the other existing measures for semantic relatedness is threefold. Context vectors are able to: i) exploit information both from a large data corpora as well as from the WordNet descriptions (glosses), ii) handle all different parts of speech possing no limitations on the amount of linguistic information contained in a document page that can be used to derive an overall degree of semantic relatedness with the query concept, iii) produce values normalized to [0,1], which is crucial for our analysis given the probabilistic standpoint of our framework.

After having defined a method for measuring the semantic relatedness between any two individual concepts, we need a methodology for extracting the overall semantic relatedness between a concept and the linguistic information contained in a document page. In order to do so, we use the previously described approach to measure the semantic relatedness between the word expressing the concept of interest and all words contained in a document page. In this way we get as many semantic relatedness values as the number of words contained in the document page. Subsequently, we only consider the words with semantic relatedness above the 64% of the maximum semantic relatedness value of all words in this page. This percentage was found to yield optimal performance in a series of preliminary experiments. By averaging between the selected values we get a number between [0,1] that indicates the semantic relatedness of the query word with the linguistic information contained in a document page. This number is used as the confidence degree of this concept for the examined document page.

3.3. Constructing, training and performing inference on the BN

In Section 2 we have described how to model a Bayesian Network so as to allow the combination of heterogeneous types of information. In this Section we describe how the explicit knowledge expressed in an ontology is used to determine the network structure, as well as how the contextual knowledge that is implicit in the training data is extracted using a learning algorithm. Moreover, we refer to the algorithm employed for performing evidence-driven probabilistic inference using as evidence the output of single-medium analyzers.

3.3.1. Integrating explicit & implicit knowledge into a BN

Ontologies have emerged as a very powerful tool able to express knowledge in different levels of granularity, handle the diversity of content essence, and govern its semantics [24]. The general knowledge about a specific domain can be expressed by a structure K_D that associates the domain concepts and relations using the allowable operators. An algorithm able to traverse this structure can answer questions like whether two concepts are related, or what type of relation associates these concepts. On the other hand, BNs are directed acyclic graphs whose nodes represent random variables and whose arcs encode the conditional dependency between them. In order to integrate the knowledge expressed explicitly by an ontology into a BN, we need to determine a set of rules for mapping ontological elements (i.e., concepts and relations) to graph elements (i.e., nodes and arcs). The structural translation rules described in [25] were adopted for determining the structure of the BN out of an OWL ontology [26]. Specifically, all concepts are directly translated into network nodes (i.e., discrete variables). Then, an arc is drawn between two nodes in the network if the corresponding two concepts are related by a "rdfs:subClassOf" relation in K_D. Additionally, the methodology of [25] describes also how to create specific subnetworks for modeling the ontology constraints between concept nodes (i.e., owl:disjointWith, owl:unionOf, owl:intersectionOf, owl:unionOf, owl:complementOf, owl:equivalentClass). However, these properties are not exploited in this work since no such need arise from the domain ontology.

In order to tackle the problem that an arc in a BN does not necessarily imply causality, we have followed the methodology proposed in [25] where an arc is drawn between two concepts if they are related with the "rdfs:subClassOf" relation and according to the relation direction (i.e., from superclass to subclass). By systematically applying all structural translation rules we determine the structure of a BN, based on the explicit knowledge expressed in an OWL ontology. Apart from ontologies, other representation structures capable of reflecting human experience exists, such as causal maps [27]. However, the use of ontologies was advocated by their wide acceptance and appeal in representing knowledge for various domains.

Concerning the knowledge implicit in the data it is essentially translated into the prior and conditional probabilities associated with each node in the BN. More specifically, the CPTs of all network nodes are learned by applying the Expectation Maximization (EM) algorithm on a set of compound

documents annotated with concept labels. EM [28] estimates the prior and conditional probabilities for each node by iterating between an expectation (E) and a maximization (M) step. During the expectation step the expected values for all data are calculated using the underlying BN and applying regular Bayes inference. The maximization step finds the maximum likelihood of a BN given the now extended data. The importance of using accurate and meaningful data for estimating the probabilistic information of the BN has been well studied in the literature [29]. Typical sources for probabilistic information are (statistical) data, literature (e.g., medical handbooks or journals where probabilistic information is given about medical disorders and symptoms), and human experts. In our work we have decided to follow the (statistical) data approach and apply the EM algorithm on training data. The reason for adopting this approach rather than setting the necessary probabilities manually, was on the one hand to avoid the need for human intervention when switching to a different domain and on the other hand to avoid bias on the initial conditions of the network. However, when using (statistical) data as the source of probabilistic information, special care will have to be taken so as no bias is introduced as the result of the data collection strategy. Such bias is likely to affect the performance of the resulting BN and can not be easily detected once the BN has been constructed. The (statistical) data used in our work have been collected so as to form a balanced (i.e., ensuring approximately proportional presence of each concept in both train and test set) sample of the population that is being modeled. Thus, the probability of introducing bias to the resulting BN was minimized.

3.3.2. Evidence-driven probabilistic inference

In order to perform evidence driven probabilistic inference on the constructed BN we rely on message passing algorithms. Pearl [30] introduced a message passing mechanism where messages are exchanged between parent and child nodes carrying the information required to update their beliefs. Although intuitively consistent, the message passing algorithm proposed by Pearl suffer from scalability issues due to the excessive number of messages that need to be exchanged over the network. In order to overcome this deficiency, Lauritzen and Spiegelhalter [31] exploit a range of local representations for the network joint probability distribution so as to reduce the number of messages that need to be exchanged. The junction tree [32], that was used to conduct our experiments, is an algorithm that takes advantage of such local representations and to the best of our knowledge is the most

efficient and scalable belief propagation algorithm.

4. Experimental Study

The goal of our experimental study was to evaluate our proposed methodology in three different aspects: i) how much improvement is achieved by the employment of the proposed cross-media analysis scheme compared to single-medium solutions, ii) whether the choice of a generative over a discriminative model is more suited for fusing evidence coming from heterogeneous sources, and iii) whether the additional cost of engineering an ontology for expressing domain knowledge, actually pays off in terms of efficiency when compared with less costly approaches like using a simplified BN or learning its structure from data using the K2 algorithm [33]. Finally, we have verified the efficiency of our framework to more general application, by using the proposed modeling approach to implement a video shot classification scheme.

4.1. Testbed

The domain selected for performing our experimental study concerns forecasting the launch of competitors' models, as defined in cooperation with Centro Ricerche Fiat (CRF)[1]. The goal of a competitor analysis department is to constantly monitor the existent competitors' products, understand market trends and try to anticipate customer needs. The information needed to achieve that, is scattered throughout the Internet (i.e., blogs and forums), and covered by a long tail of international and national automotive magazines. In a typical scenario the main role is played by the person responsible for data acquisition that has the responsibility of daily inspecting a number of resources such as WWW pages, car exhibitions, car magazines, etc, that are likely to publish material of potential interest. The collected information is subsequently used in the *set-up* stage of new vehicles (i.e., the development stage where a first assessment of the future vehicle's features is carried out). This process is of great value to many companies because it contributes to keeping new products design up to date.

One of the tasks defined by the experts was to be able to automatically evaluate a document with respect to its interest for the *car components ergonomic design*. The fact that most of the collected documents use both

[1] http://www.crf.it/

visual and textual descriptions, motivated the construction of a cross media classifier recognizing compound resources that are valid for the high-level concept *car components ergonomic design*.

For the purposes of our evaluation a dataset of 162 pdf documents (containing 1453 pages) was collected, that are primarily advertising brochures describing the characteristics of new car models. Each pdf document was dismantled into its visual and textual constituent parts using the xpdf library[2]. All media elements extracted from the same page were kept together so as not to lose any conceptual relations originating from the document's layout. The linguistic information was gathered in a single text file while the visual representations were extracted to independent image files as depicted in Fig. 4.

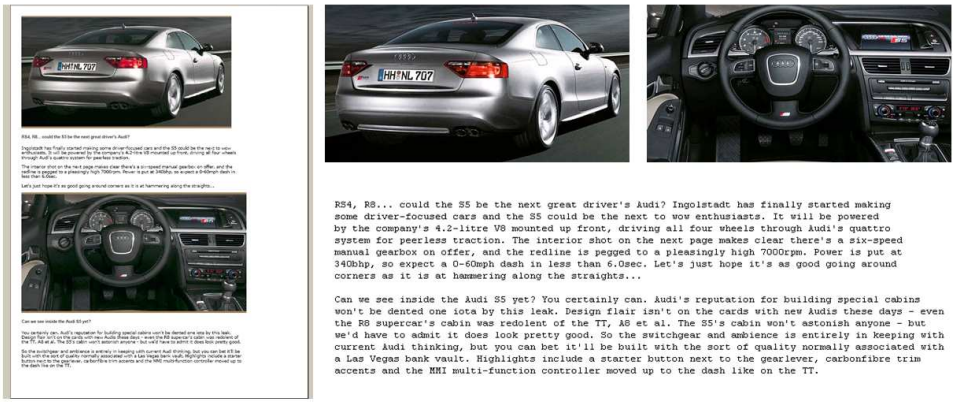


Figure 4: Dismantling a pdf document to its constituent parts

Two different manual annotation efforts were carried out for the purposes of our work. Since we have decided to consider the pdf documents on a per page basis, the first annotation effort was to manually inspect each of the 1453 document pages and record in an annotation file whether they are valid for the high level concept *car components ergonomic design*. The second annotation effort involved going through all 1453 document pages and marking for each page which of the ontology concepts are present or not. The result of this annotation process was a set of concept labels for each of the 1453 document pages, suitable for measuring the co-occurrence between

[2]http://www.foolabs.com/xpdf/

any two concepts of the domain. These sets of concept labels were used to learn the CPTs of the BN nodes. In cases where the total amount of training samples is large and their manual annotation with concept labels becomes tedious, alternative sources of annotation can be used, such as the social networks. The rapid growth of social media, that emerged as the result of users willingness to communicate, socialize, collaborate and share content, has resulted in the generation of a tremendous volume of user contributed data. These data have been made available on the Web, usually along with an indication of their meaning (i.e., tags). Although these data usually come with a high level of noise, they exhibit noise reduction properties given that they encode the collective knowledge of multiple users. Thus, the CPTs can be learned by exploiting the abundant availability of tagged images in social networks, which can be used to acquire the information about the co-occurrence of concepts in a domain. Out of the 162 documents, 149 (928 pages) were used for training I^{train} the BN (see Section 3.3) and 13 (525 pages) were used for testing I^{test}.

4.2. High level concept detection using the cross media analysis scheme

For conducting our experiments we have engineered three ontologies (one for each of the evaluated cases: textual-only, visual-only and cross media) that are mostly concerned with concepts related to the ergonomic design of car components. All three ontologies were engineered based on the knowledge acquired by going through a sufficient number of related documents and getting acquaintance with the domain details. These ontologies were used to determine the structure of three different BNs (one for each evaluation case). In all cases, the node modeling the high-level concept *car components ergonomic design* was the root node of the constructed BN. For learning the CPTs of the BN nodes, the Expectation Maximization algorithm was applied on I^{train}. Depending on the concepts included in the employed ontology, only the annotations referring to these concepts were included in the corresponding training set.

After constructing the BNs the analysis process runs as follows. Depending on the examined case (textual-only, visual-only, or cross media) the single-medium analyzers are applied on the constituent parts of a document page. Their probabilistic output is injected into the BN nodes as described in Section 2. This triggers an inference process that progressively modifies the posterior probabilities of all connected nodes in the network using message passing belief propagation. When the process is completed the posterior

probability of the root node modeling the high-level concept *car components ergonomic design* (represented with the *CA_ED* symbol in all figures), is compared against a fixed threshold. If the threshold is exceeded the detector decides positively, otherwise the document page is considered as not being relevant with the ergonomic design of car components. An illustration of this procedure for the cross-media case is depicted in Fig. 5. For measuring the efficiency of the high-level concept detector we have used recall versus precision curves. The threshold value of Fig. 5 is uniformly scaled between [0,1] for conducting the experiments in all cases.

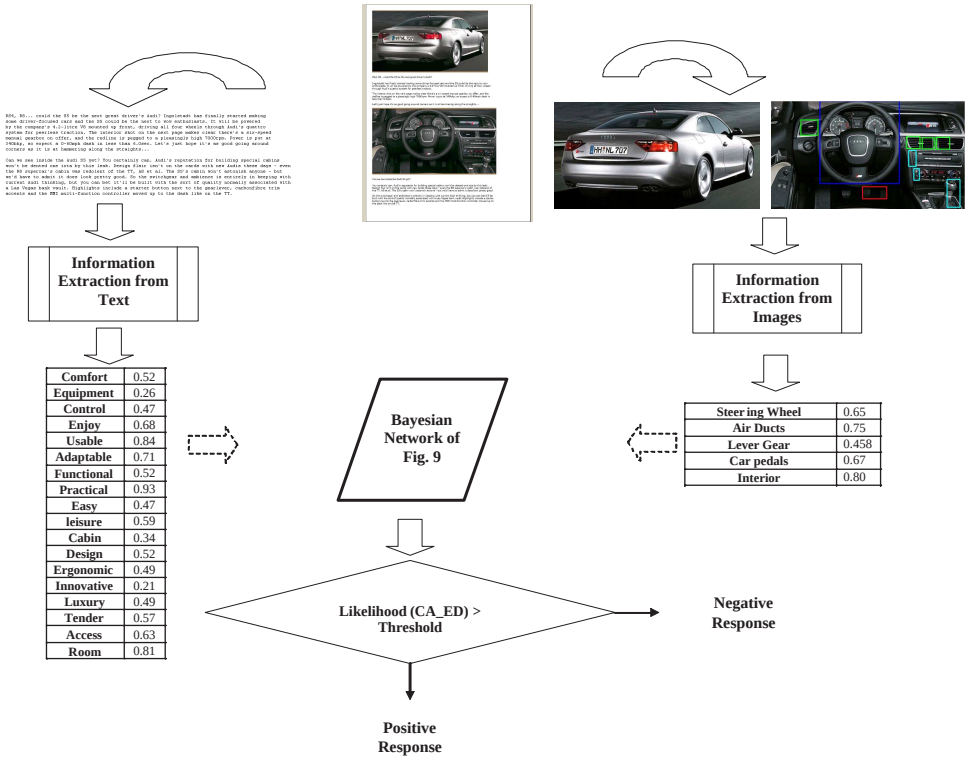
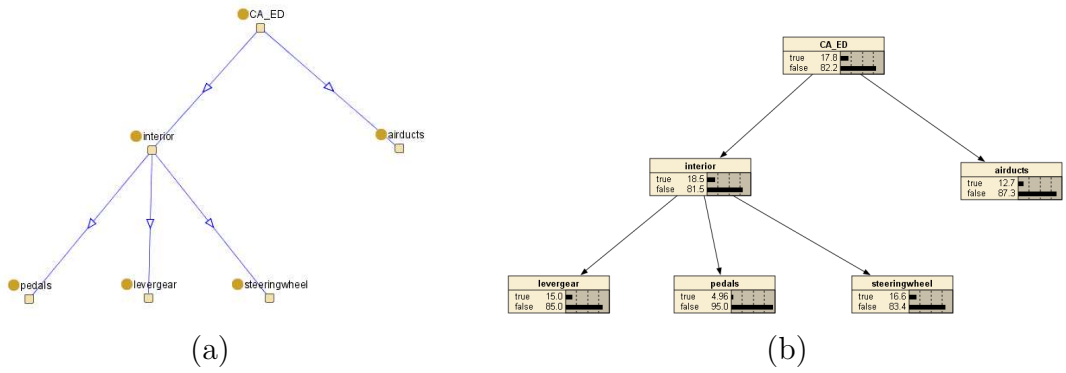


Figure 5: Inference process illustration for the cross media setting

4.3. Single vs Cross media analysis

In the case of visual-only analysis, the general knowledge about the specific domain was expressed by the ontology depicted in Fig. 6(a). This ontology associates five visual concepts, namely *air ducts, steering wheels, gear*

levers, car pedals and interior with the high-level concept *car components ergonomic design*. The trained BN used for this setting is depicted in Fig. 6(b). Five detectors trained to identify the five concepts of the domain ontology were implemented using the method of Section 3.2.1. These detectors were trained using an independent dataset of 3230 images depicting car interiors that was strongly annotated at region-detail. Each of these detectors was attached to the corresponding BN node of Fig. 6(b) and was used to trigger the process of probabilistic inference. By applying these five detectors on every image contained in a document page and using their output to instantiate the network nodes, we are able to decide about the existence of the high-level concept *car components ergonomic design* in a document page, based solely on the information depicted on the images of this page. The obtained results are depicted in Fig 11.



(a) (b)

Figure 6: Experimental setting using only visual concepts, a) Domain ontology for document analysis using only visual evidence, b) Bayesian Network for visual analysis

In the case of textual-only analysis, we used eighteen different concepts, namely *access, cabin, design, leisure, comfort, easy, enjoy, luxury, room, tender, ergonomic, equipment, innovative, usable, practical, functional, adaptable and control* for obtaining the textual evidence. Using these eighteen concepts we constructed the ontology of Fig. 7 that encodes the associations between the textual concepts and the high-level concept of *car component ergonomic design*. The trained BN used in this setting is depicted in Fig. 8. The confidence degrees that are used to instantiate the BN nodes are obtained by applying the textual analysis method described in Section 3.2.2 for each

document page and using the above linguistic descriptions as query words. As in the previous case this setting allows us to decide about the existence of the high-level concept *car components ergonomic design* in a document page, based solely on the information included in the textual descriptions of this page. The precision versus recall curve obtained from textual-only analysis is depicted in Fig 11.

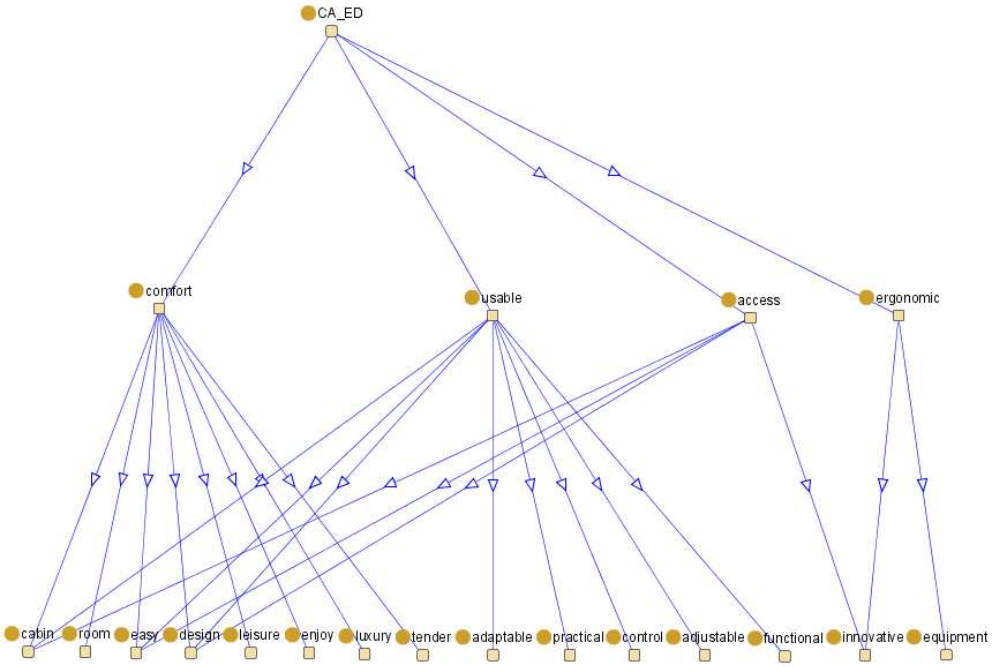


Figure 7: Domain ontology for document analysis using only textual concepts

For the case of cross-media analysis, both textual and visual concepts were used for the construction of the ontology depicted in Fig. 9. This ontology expresses the domain knowledge across media and reflects the cross-relations between textual and visual concepts. The trained BN that was used for performing inference in this setting is depicted in Fig. 10. The confidence degrees obtained by applying the aforementioned textual and visual single-medium analyzers on the constituent parts of a document page, are used to instantiate the BN nodes and perform inference using evidence across media. The results achieved by the high level concept detector in this setting are depicted in Fig 11.

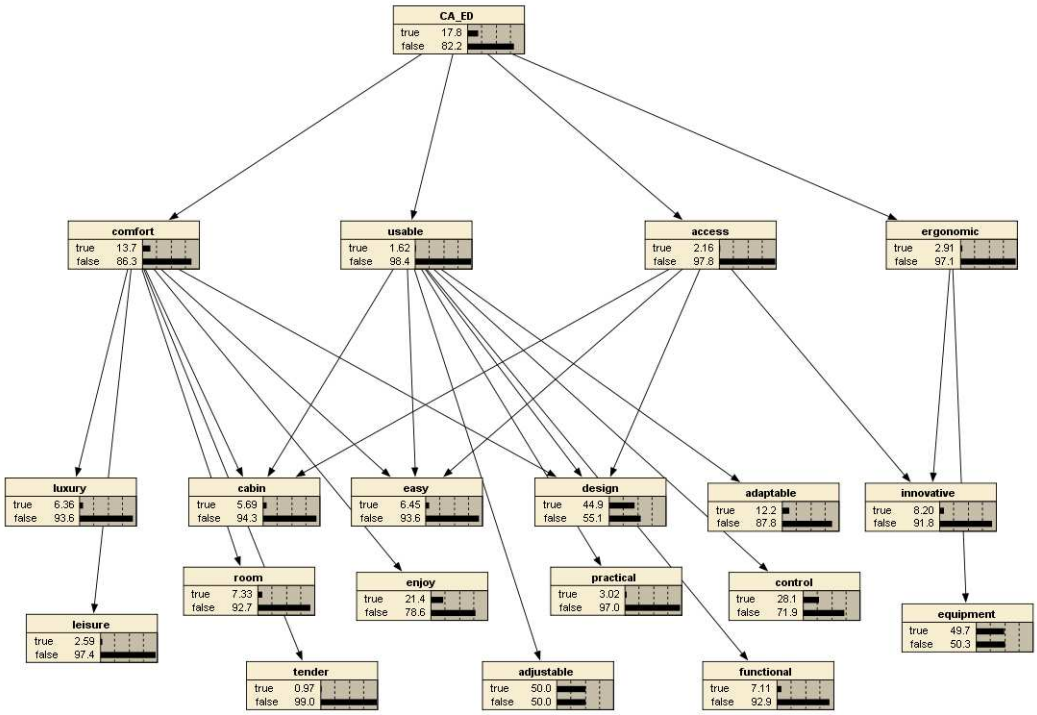


Figure 8: Bayesian Network for textual-only analysis

It is clear from the comparative diagram of Fig. 11 that the configuration of the framework using evidence across media, outperforms the cases where evidence originates exclusively from one media type. We can see that textual analysis performs significantly better than visual analysis mainly due to the increased number of evidence that have been used in this setting. However, when the textual and visual evidence are combined in the cross media setting, the high-level concept detector manages to further improve its efficiency for most of the applied threshold values. In conclusion, with this experiment we verify that there are cases where the evidence existing across different media types carry complementary information, that can only be translated into facts when considered in a synergetic fashion.

4.4. Generative vs Discriminative model

The second goal of our experimental study was to investigate the superiority of generative models like BNs over discriminative models like Support Vector Machines (SVMs) [34], to more efficiently incorporate and benefit from

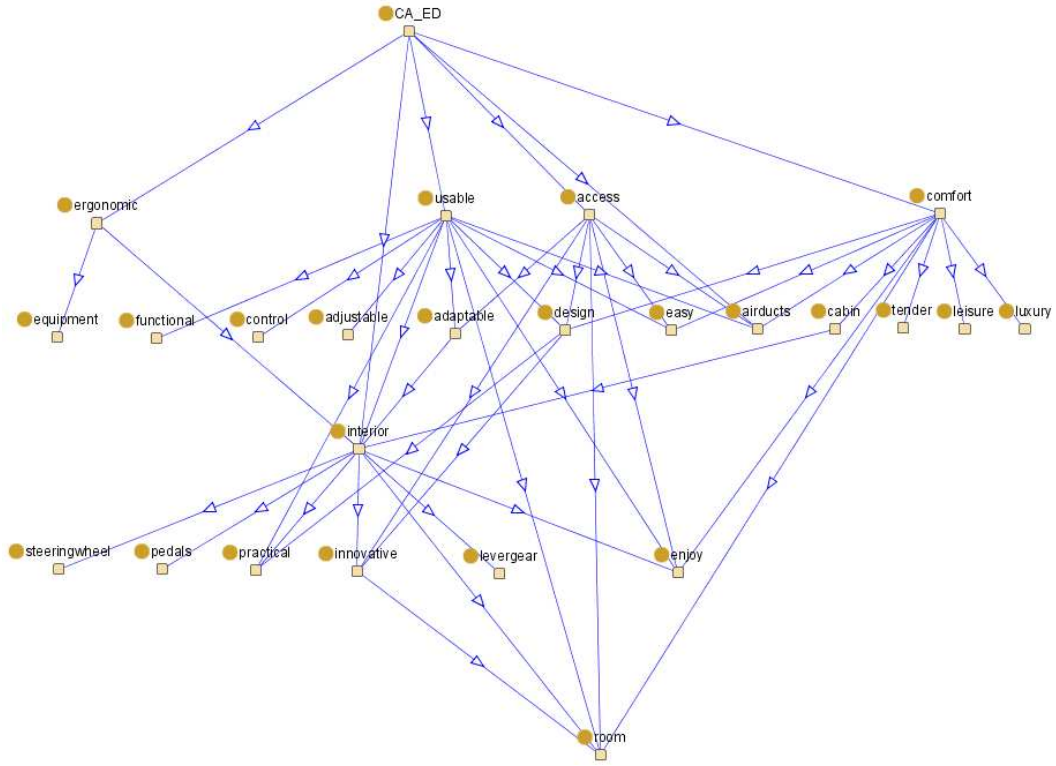


Figure 9: Domain ontology for document analysis using both visual and textual concepts

explicit knowledge. The motive behind using BNs in our work was their ability to smoothly incorporate explicit knowledge through their parameters and structure, as well as to learn efficient models from small training sets. This is in contrast to the approaches based on SVMs, since there is no straightforward way to incorporate explicit knowledge in these cases, as it can only be done at the level of the kernel. Moreover, when relying on SVMs, robust models can only be learned when there is a significant number of training samples available.

In order to verify the above, we compared our generative classifier based on BNs with a discriminative classifier implemented using SVMs. The feature space for training the SVM models was determined by concatenating the confidence degrees generated from the single-medium analyzers, resulting in a 23-dimensional feature vector for each document page. The SVMlight library [35] was employed for learning an one-class classifier recognizing the

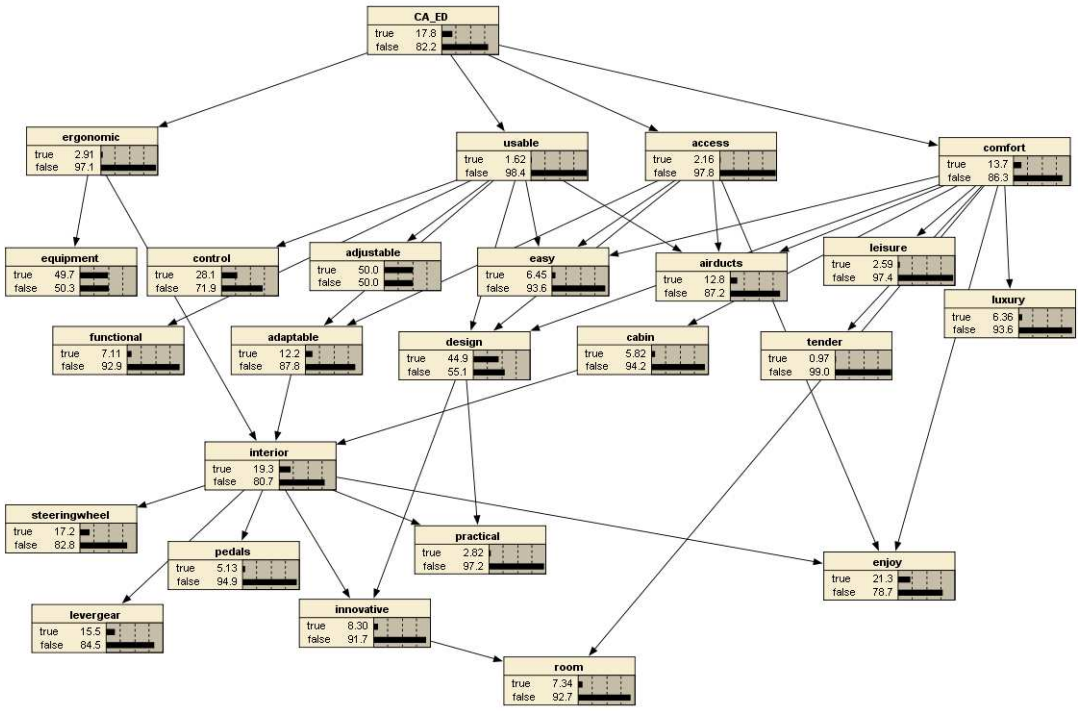


Figure 10: Bayesian Network for cross media analysis

concept *car components ergonomic design*, using the same train/test split as in the case of BNs. A polynomial kernel function was used for learning the SVM models. Since the one class SVM models are known to be rather sensitive on the ratio between positive and negative examples, we have tried 4 different ratios (i.e., 1/1, 1/2, 1/3 and 1/4) in order to optimally tune the classifier. Using the full train set the positive/negative ratio is approximately 1/4. The bar diagrams of Fig. 12(a) shows the F-measure scores achieved by the SVM-based classifiers using all four positive/negative ratios, as well as the score achieved by the BN classifier for the optimal threshold value. We can see that the BN classifier outperforms all SVM-based classifiers with the smallest improvement being $\approx 3\%$ (1/3 case) and the largest being $\approx 12\%$ (1/4 case).

Moreover, in order to verify that, in contrast to SVMs, BNs are able to learn efficient models even from just few examples, we performed several experiments by reducing the number of samples included in the train/test

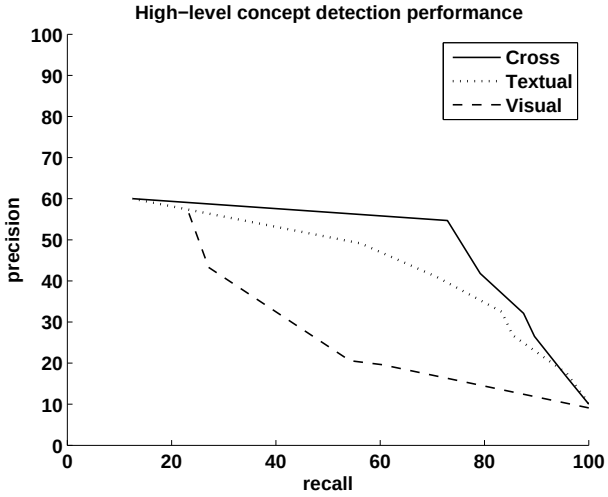
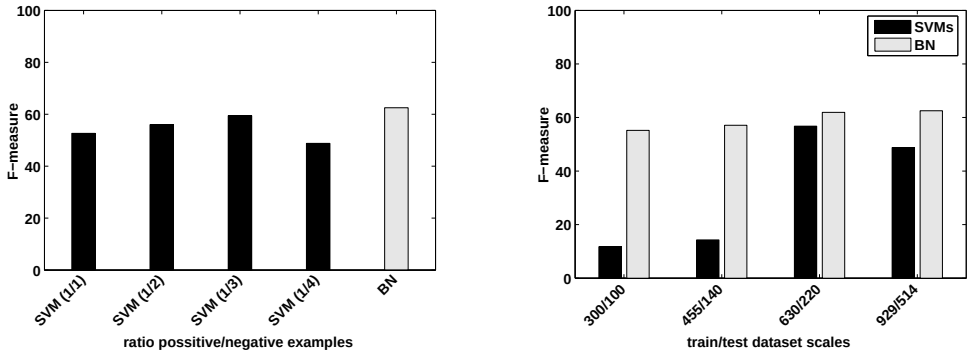


Figure 11: Cross vs single media analysis performance

datasets. Fig.12(b) shows the F-measure scores achieved using both approaches for four different scales of the train/test datasets. For this experiment the SVM-based classifiers were trained using all positive and negative samples included in each of the different dataset scales. It is clear that the models learned using BNs manage to deliver good performance even when trained with a particularly small number of samples. This is not the case for the models learned using SVMs, where the number of training samples needs to grow at approximately 600 in order to deliver good results. Both experiments verify the superiority of generative models in handling prior knowledge more efficiently and learning from a few examples. This attribute is particularly useful in cross media analysis since the cost of manual annotation in a cross media fashion is even higher from the single-medium cases, making the generation of a significant number of training examples very expensive.

4.5. Cases with missing or noisy domain knowledge

It is evident that our framework benefits from the existence of knowledge about the domain. However, there can be cases where such knowledge is either noisy or missing (i.e., the list of domain concepts is known but the relations between them are not). In such cases, our framework can be applied using either a trivial structure for the BN, or using a BN the structure of which is determined from sample data. In order to evaluate the performance of our framework when domain knowledge is noisy, we have considered the



(a) (b)

Figure 12: a) Comparing generative with discriminative models using different ratios for the positive/negative examples b) Comparing generative with discriminative models using different scales for the train/test datasets

following two approaches for determining the structure of the BN. The first approach assumes the most trivial structure for the BN and initiates our framework using a naive BN. The naive BN is the simplest classifier based on Bayes' rule, it assumes that all variables are independent from each other and all nodes are directly connected to the root node. The second approach is based on methods that are able to derive the structure of the BN from sample data. One such method is [36] where prior knowledge, provided in the form of a temporal BN called *prior network*, is combined with sample data in order to learn one or more BNs that are much closer to the actual structure of the domain than the initial *prior network*. A similar method is the well-established, score-based Cooper's K2 algorithm [33] which attempts to recover the underlying distribution of nodes in the form of a Directed Acyclic Graph (DAG), without making any assumptions about their structure. For the purposes of our work we have decided to employ the K2 algorithm in order to evaluate the performance of a BN, the structure of which is determined without using any prior information about the relations between the domain concepts.

More specifically, the K2 algorithm takes as input the number and ordering of nodes ($n = 24$ in our case), an upper bound for the parents of its node and a set of training data, which in our case correspond to the concept

label annotations described in Section 4.1. The set of nodes includes the 23 visual and textual concepts as well as the high level concept *car components ergonomic design*. The ordering of the nodes was determined based on the frequency of appearance (in descending order) of the corresponding concepts in the training data. In order to avoid networks with high complexity we have set the upper bound of parent nodes to be four. The BN generated using the K2 algorithm is depicted in Fig. 13. In Fig. 14, we compare the performance achieved by a BN constructed based on the cross media domain ontology as described in Section 3.3.1, against the performance of a naive BN and the performance of a BN, the structure of which is determined using the K2 algorithm. In all cases, inference was performed as described in Section 3.3 and the curves were drawn by modifying the threshold value between [0,1].

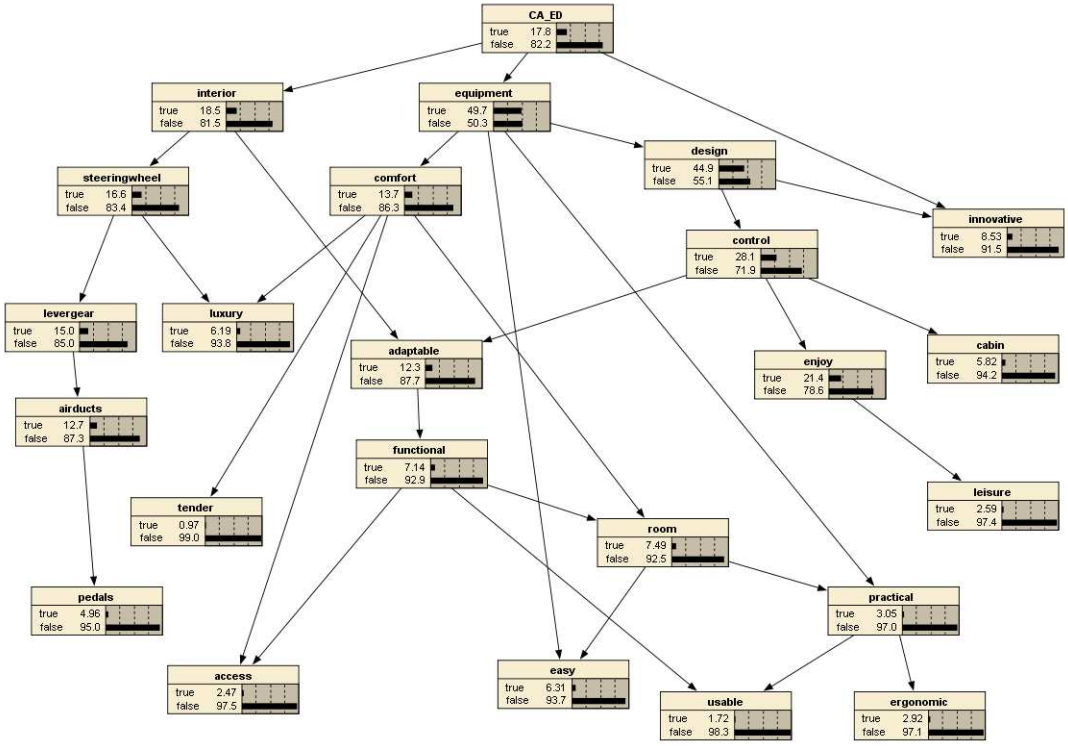


Figure 13: Bayesian Network derived from sample data using the K2 algorithm

It is clear from the results that the incorporation of explicit knowledge is particularly useful when combining information from heterogenous sources.

We can see that the BN using the ontology, clearly outperforms the naive and K2 algorithm-based approaches. This is attributed to the fact that the domain ontology manages to capture the underlying cross-modal relations and boost the classification performance. Moreover, the fact that the naive BN approach achieves better results from K2, further advocates the need for incorporating explicit knowledge (even as a simple two level hierarchy) when combining information from heterogeneous sources.

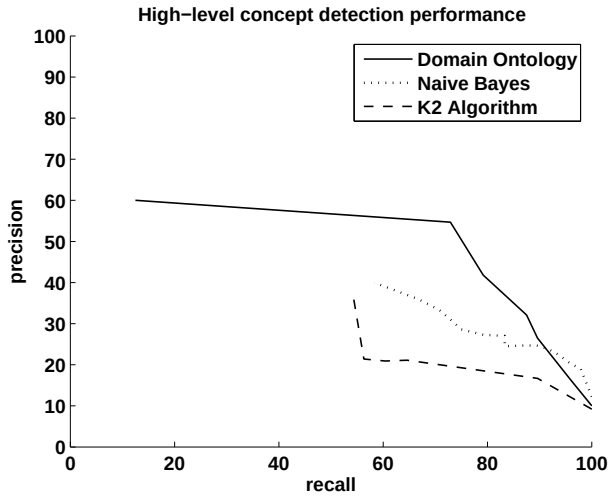


Figure 14: Comparative diagram for the different methods used to determine the BN structure

4.6. Video shot classification

In order to verify the efficiency of our BN modeling approach to more general applications, we have used it to implement an ontology-based classifier for video shots. For building and evaluating this classifier we have relied on the TRECVID2010 development dataset IACC.1.tv10.training[3], that has been provided by TRECVID organizers to facilitate training in various different tasks of 2010 competition. The dataset is composed of 118581 shots annotated with 130 concepts[4]. The reason for choosing this dataset over the datasets used in the previous years, was that 2010 was the first year where

[3]http://www-nlpir.nist.gov/projects/tv2010/tv2010.html#IACC.1.tv10.training
[4]http://www-nlpir.nist.gov/projects/tv2010/TV10-concepts-130_UPDATED.xlsx

the organizers provided an ontology with the relations between 104 of the 130 available concepts. The availability of such ontologies is an important motivation for employing the proposed modeling approach, since the incorporation of domain knowledge in the analysis process is one of its great advantages. In order to facilitate training and testing we have split the 118581 shots to 59291 training T^{train} and 59290 testing T^{test} shots.

4.6.1. Engineering the ontology and building the BN

By examining the ontology relations provided with the dataset, we observed that there were 9 concepts, namely *Person, Outdoor, Indoor, Vegetation, Vehicle, Politics, Animal, Sports, Science_Technology*, that acted as super-classes of all other concepts in the ontology. Based on this fact, and given that the goal of our approach is to infer the presence of a high-level concept by accumulating the effect of the existing evidence, we consider these 9 concepts to be the root concepts or our ontologies. Then, we implement a multi-class classifier for these concepts using cross media analysis. Out of the remaining 95 concepts, 45 we chosen as textual concepts based on the availability of Automatic Speech Recognition (ASR) transcripts for a relatively high number of the shots annotated with these concepts. This selection strategy was motivated by the need to ensure that there will be sufficient textual information to extract evidence for the textual concepts. The remaining 50 concepts were considered as visual. When considering the textual-only or visual-only analysis case, the root concepts are only supported by the 45 textual or the 50 visual concepts, respectively. In the cross media analysis case all available concepts are used. The output of the multi-class video-shot classifier is a confidence degree for each of the 9 root concepts. Crisp decisions can be taken by applying a threshold on these confidence degrees. Having engineered the ontologies for the three analysis cases (i.e., textual-only, visual-only and cross media), we used the methodology described in Sections 2 and 3.3.1 to construct the corresponding BNs. The CPTs were learned by applying the EM algorithm on the concept labels of the shots included in T^{train} and probabilistic inference was performed as described in Section 3.3.2.

4.6.2. Modality synchronization

Each of the shots included in the TRECVID2010 development datasaet consists of its key-frame (i.e., an image) and the ASR transcripts of the spoken dialogs within the shot time-frame. In this case we consider that a

conceptual relations exists between the key-frame and the ASR transcript of a shot. Thus, classification is performed for every shot by combining the visual and textual evidence extracted from the corresponding key-frame and ASR transcript, respectively.

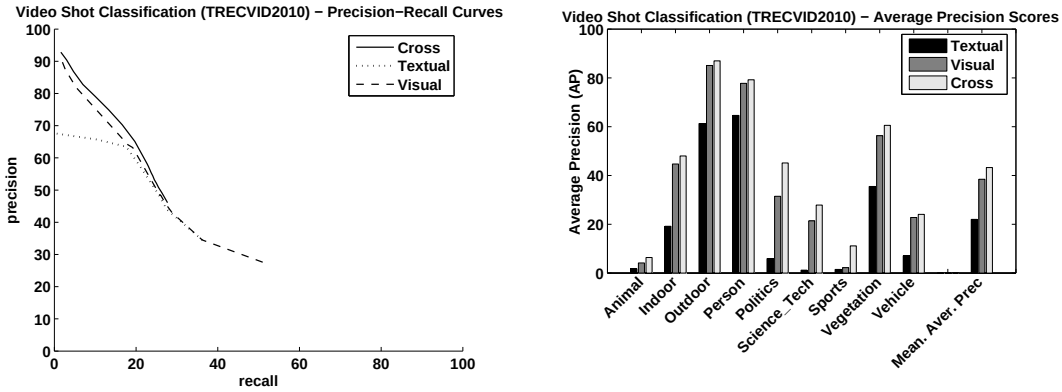
4.6.3. Single-medium analysis

For extracting the likelihood estimates of the textual concepts we have employed the textual analysis approach described in Section 3.2.2. In this case, the values of semantic relatedness are estimated between the textual concept and every word included in the ASR transcript of the analyzed shot. By averaging the semantic relatedness values as described in Section 3.2.2 we obtain a likelihood estimate per textual concept, for each shot.

Due to the fact that the annotations provided by TRECVID are at the global level of the image and not at the level of regions, as required by the technique of Section 3.2.1, we have employed a different method for visual analysis. In this case, the visual representation of the images was extracted by applying the feature extraction technique described in [37]. More specifically, a set of interest points was detected in every image by applying the Harris-Laplace point detector on intensity channel [38]. For each of the identified interest points a 128-dimensional SIFT descriptor was computed using the version described by Lowe [39]. Then, a Visual Word Vocabulary (Codebook) [40] was created by using the K-Means algorithm to cluster in 500 clusters, approximately 3 million SIFT descriptors that were sub-sampled from a total amount of ≈200 million SIFT descriptors, extracted from ≈120 thousand training images. The Codebook allows the SIFT descriptors of all interest points to be vector quantized against the set of Visual Words and create a histogram of 500 dimensions. Finally, additional histograms were extracted from specific parts of the image. Using a 2x2 subdivision of the image, one histogram was extracted for each image quarter. Similarly, using a 1x3 subdivision consisting of three horizontal bars, one histogram was extracted for each bar. In the end all histograms were concatenated to form a 4000-dimensional visual representation of the image. After obtaining the visual representation of the images, Support Vector Machines (SVMs) [34] were used for generating the concept detection models. The 59291 key-frames included in T^{train} were used for training the concept detection models. Tuning arguments included the selection of Gaussian radial basis kernel and the use of cross validation for selecting the kernel parameters.

4.6.4. Video-shot classification results

The performance of our video-shot classifier was evaluated on T^{test}, for the cases of visual-only, textual-only and cross media analysis. In Fig. 15 we report results for the 9 root concepts. Fig. 15(a) depicts the precision-recall curves achieved by each analysis case. The curves are obtained by uniformly scaling the decision threshold between [0,1] and averaging between all root concepts. As expected the video-shot classifier incorporating evidence across media outperforms the classifiers that incorporate only textual or only visual information. In contrast to the analysis results on compound documents reported in Fig. 11, in this case the video-shot classifier based on visual analysis performs better than the classifier relying on textual analysis. This can be attributed to the low quality of ASR transcripts or the complete absence of transcripts for a non-negligible amount of shots.



(a) (b)

Figure 15: Cross vs single media analysis performance using TRECVID2010 dataset a) Precision-recall curves obtained by uniformly scaling the decision threshold between [0,1] and averaging between all root concepts, b) Average precision scores for the 9 root concepts

In Fig. 15(b) we report the Average Precision (AP) scores for the 9 root concepts, since this is the metric used by the TRECVID organizers. We can see that the improvement in performance achieved by the cross-media classifier is consistent across all root concepts and in certain cases by a significant amount, as in the case of *Sports*. Our experimental results show that the superiority of the cross media classifier over its single-medium counterparts is

evident in all experimental settings, advocating the efficiency of the proposed approach for modeling the BN.

In order to compare our work with existing state-of-the art methods, we have relied on the evaluation results released by the organizers of TRECVID2010 for the task of *Semantic Indexing*. In the context of this task all submissions were evaluated for a set of 30 concepts[5], subset of the total set of 130 concepts. In order to facilitate the comparison of our work with the methods participated in the competition, we have employed a modified version of our video shot classifier. This version works in a similar way with the previous case, with the additional functionality that likelihood estimates are also given for the root nodes of the BN, providing useful evidence for the existence of their child nodes. In this way we manage to obtain inferred confidence degrees for 26 of the concepts that have been used for evaluation. No confidence degrees were obtained for the concepts *Doorway*, *Explosion Fire*, *Hand*, *Telephones*, since they were not included in the ontology provided by the organizers. Fig. 16 compares the Average Precision achieved by our framework against the top-scoring run and the average performance among all 101 runs, submitted for the *Semantic Indexing* task [41].

It is important to note that the performance figures depicted in Fig. 16 are not directly comparable due to the following reasons. The dataset used for training and testing are not identical, since we have trained our classifier using half portion of the development dataset and evaluated its performance using the other half. On the contrary, the methods submitted for the *Semantic Indexing* competition used the full development dataset for training and evaluated their performance using an independent test set. Moreover, the performance scores provided by the organizers refer to the Inferred Average Precision [42] which is an approximation of Average Precision when the available annotations are incomplete. The figures provided for our framework refer to Average Precision since we had complete annotations for our test set. Despite the above, it is clear that our method compares favorably with the performance achieved by the state-of-the-art methods. Among the 26 evaluated concepts our method outperforms the top-scoring methods in

[5]Airplane flying, Animal, Asian_People, Bicycling, Boat-ship, Bus, Car_Racing, Cheering, Cityscape, Classroom, Dancing, Dark-skinned_People, Demo or protest, Doorway, Explosion_Fire, Female-Human-Face-Closeup, Flowers, Ground_Vehicles Hand, Mountain, Nighttime, Old_People, Running, Singing, Sitting_Down, Swimming, Telephones, Throwing, Vehicle, Walking

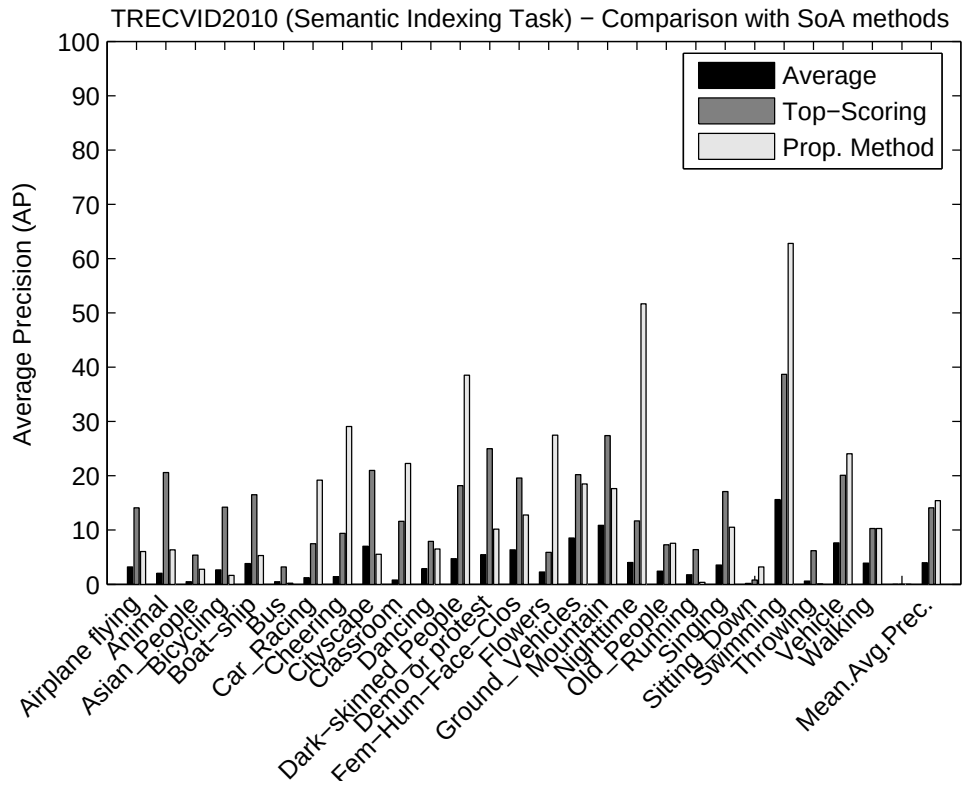


Figure 16: Comparison of our framework for 26 concepts against the top-scoring method and the average performance among all 101 runs, submitted for TRECVID2010 *Semantic Indexing* task.

11 and surpass the average performance score in 21 cases. The Mean Average Precision achieved by our framework (15.4%) is improved by 1.3% compared to the Mean Average Precision of the top-scoring methods (14.1%) and by 11.4% compared to the average performance scores (4%).

5. Related Work

In the research field of multimedia analysis, indexing and retrieval, various methods have been proposed for fusing the evidence extracted from different media sources. Statistical methods are widely used for multimodal integration [43], where the query object is classified based on the distribution of patterns in the space spanned by pattern features. The most frequently

encountered methods are Bayesian Networks that assign a pattern to the class which has the maximum estimated posterior probability, and Hidden Markov Models (HMM) that assign a pattern to a class based on a sequential model of state and transition probabilities.

In this context, our study can be considered to share similar objectives with various works in this field. Within the scope of probabilistic inference, Hospedales and Vijayakumar [44] implement a multisensory detection, verification and tracking mechanism by inferring the association between observations. More specifically, in order to solve the who-said-what problem they present a principle probabilistic approach, where Bayesian inference is used for combining multiple sensing modalities. The proposed model is claimed to be sufficient for robust multitarget tracking and data association in audiovisual sequences. In [45] Choi et al. present three classifier fusion methods and evaluate their efficacies on raw data sets. They use class-specific Bayesian fusion, joint optimization of the fusion process and individual classifiers, and employ dynamic fusion for combining the posterior probabilities from individual classifiers. The results of the proposed approaches are generally better than the majority voting and the naive Bayes fusion approaches, and significantly reduce the overall diagnostic error in automotive systems. Compared to Bayesian Networks, Hidden Markov Models are capable not only to integrate multimodal features but also to include sequential features. In [46] the MFHMM (Multistream Fused Hidden Markov Model) is presented as a generalization of a two-stream fused HMM [47] for integrating coupled audio and visual features. MFHMM is used for linking the multiple HMMs and is claimed by the authors to be an optimal solution according to the maximum entropy principle and the maximum mutual information criterion. In [48] the authors rely on SVMs and present a late fusion scheme where the unimodal features are initially used to learn separate concept classifiers. Then the output of these classifiers are concatenated to determine a new feature space and learn an SVM-based integrated concept classifier.

Recently, semi-supervised graph-based methods have also attracted the interest of researchers for narrowing the semantic gap between the low- and high-level features. Hoi et al. [49] present multi-modal fusion through graphs in addition with a multilevel graph-based ranking scheme for content-based video retrieval. They present the semi-supervised ranking (SSR) method to exploit both labeled and unlabeled data effectively and further explore a multilevel ranking solution to solve the scalability problem of SSR. The proposed multilevel ranking scheme achieves good performance for large scale

applications and also provides a solution to the overfitting problem. In the same direction, Wang et al.[50] present the OMG-SSL method, optimized multigraph-based semi-supervised learning, as an efficient video annotation scheme. The proposed approach is equivalent to fusing multiple graphs and then conducting semi-supervised learning on the fused graph. According to the results, the OMG-SSL method improves the learning performance and can be easily extended through utilizing more graphs. The work in [51] proposes a fusion framework in which classification models are build for each data source independently. Then, using a hierarchical taxonomy of concepts, a Conditional Random Field (CRF) based fusion strategy is designed. According to the fusion scheme described in this work, a graph is defined over the hierarchical taxonomy (i.e., a tree over categories) where its node represents a category. The scores from different unimodal classifiers referring to the same category are concatenated in a feature vector, which serves as the observation of the corresponding node. This work is very similar with our approach from the perspective of integrating explicit knowledge into the analysis process. However, in this case the scores obtained from the unimodal classifiers are concatenated to form the observation vector for each node. The advantage of our approach over this work is that we use the space of likelihood estimates as a "lingua franca" between the heterogeneous types of information, removing the need to homogenize the output of unimodal classifiers. A semi-supervised approach is employed in [52] where the authors propose to facilitate the learning process by integrating both visual and linguistic information, as well as unlabeled multi-modal data. Their approach is based on co-training which is a semi-supervised learning algorithm that requires two distinct "views" of the training data. Co-training first learns a separate classifier for each view using labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data. Compared to our work the aforementioned approach is unable to exploit the prior information derived from the co-occurrence of concepts, as well as the knowledge derived from the domain.

ClassView [53] is the method presented by Fan et al. for performing video indexing and retrieval. The authors use a hierarchical, semantics sensitive classifier for bridging the semantic gap between low- and high-level features, while the expectation maximization algorithm is used to determine the feature subspace and the classification rule. The domain-dependent concept hierarchy of video contents in the database, similar to our work, determines

the hierarchical structure of the semantics-sensitive video classifier. The proposed scheme turns out to be effective and closer to the human-level video retrieval. Wei et al. [54] fuse multimodal cues hierarchically via a cross-reference method. The authors present CR-Reranking for inferring the most relevant shots, achieving high accuracy. First the initial search results are clustered in diverse feature spaces, then the clusters are ranked by their relevance to the query and finally all the clusters are hierarchically fused via the cross-reference strategy. Finally, Lim et al. [55] combine generative with discriminative models in a sequential manner. Generative models that incorporate explicit knowledge are constructed using a small set of training samples. Subsequently, these generative models are used to classify new samples and augment the existing set with new training samples. In this way the authors manage to generate a set of training samples, sufficiently large to learn a robust discriminative classifier. Thus, the incorporation of explicit knowledge is not so much intended to facilitate the classification process by enforcing certain rules, but to indirectly improve the classification performance of the discriminative classifier by offering more training samples. Compared to [55] the advantage of our work is that explicit knowledge is made part of the inference process and directly influence the classification performance.

6. Conclusions

In this manuscript we have proposed a modeling approach for the BN that determines a conceptual space. This space allows machine learning techniques and probabilistic inference frameworks to be effectively combined for the purpose of semantic multimedia analysis. We have used the proposed conceptual space to combine evidence originating from different multimedia types and perform cross media analysis of compound documents and video shots. Our experiments have verified that there are cases where the information contained in a multi-modal resource can only be extracted if evidence are considered across media. Moreover, it has been proven that information coming from the domain knowledge is particularly useful, especially when dealing with heterogeneous types of content. Interesting were the results showing that when performing cross media analysis at the result-level, the generative models are more suited for incorporating explicit knowledge and outperform the discriminative models that luck a straightforward way to benefit from such knowledge. One important requirement of the presented scheme is that it needs a deep modeling of the analysis context (in terms of

engineering the domain ontology and producing cross media annotations), which makes the approach appropriate for cases where this effort is justified by the added value in the application. Our plans for future work include the use of the proposed modeling approach for combining information from more media types (i.e., images, text, sound, sensor data).

Acknowledgment

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

[1] M. R. Naphade, T. S. Huang, A probabilistic framework for semantic video indexing, filtering, and retrieval, IEEE Transactions on Multimedia 3 (1) (2001) 141–151.

[2] F. Souvannavong, B. Merialdo, B. Huet, Multi-modal classifier fusion for video shot content retrieval, in: In Proceedings of the 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005), 2005.

[3] V. Mihajlovic, M. Petkovic, W. Jonker, H. Blanken, Multimodal content-based video retrieval, in: H. Blanken, A. de Vries, H. Blok, L. Feng (Eds.), Multimedia Retrieval, Data-Centric Systems and Applications, Springer Verlag, Berlin, 2007, pp. 271–294.

[4] G. Snoek, M. Worring, J. Geusebroek, D. Koelma, F. Seinstra, A. Smeulders, The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing, IEEE Transactions on Pattern. Anal. and Mach. Intel., 28 (10) (2006) 1678–1689.

[5] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, J. R. Smith, Semantic indexing of multimedia content using visual, audio, and text cues, EURASIP Journal on Applied Signal Processing 2003 (2) (2003) 170–185.

[6] M. Možina, C. Giuliano, I. Bratko, Arguments extracted from text in argument based machine learning, in: Proceedings of 1st Asia Conference on Intelligent Information and Database Systems, 2009.

[7] J. Magalhaes, S. Rüger, Information-theoretic semantic multimedia indexing, in: CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, 2007, pp. 619–626.

[8] Y. Wu, E. Y. Chang, K. C.-C. Chang, J. R. Smith, Optimal multimodal fusion for multimedia data analysis, in: MULTIMEDIA '04, ACM, New York, USA, 2004, pp. 572–579.

[9] D. Li, N. Dimitrova, M. Li, I. K. Sethi, Multimedia content processing through cross-modal association, in: MULTIMEDIA '03, ACM, New York, USA, 2003, pp. 604–611.

[10] A. Laender, B. Ribeiro-Neto, A. Silva, J. Teixeira, A brief survey of web data extraction tools, in: SIGMOD Record, Vol. 31, 2002.

[11] A. Arasu, A. H. Garcia-Molina, Extracting structured data from web pages, in: ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, 2003.

[12] B. Rosenfeld, R. Feldman, J. Aumann, Structural extraction from visual layout of documents, in: ACM Conference on Information and Knowledge Management (CIKM), 2002.

[13] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1 (2001) 511.

[14] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: Proceedings of the Second European Conference on Computational Learning Theory, Springer-Verlag, London, UK, 1995, pp. 23–37.

[15] C. P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: ICCV '98: Proceedings of the Sixth International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA, 1998, p. 555.

[16] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database (Language, Speech, and Communication), The MIT Press.

[17] C. Leacock, M. Chodorow, Combining local context with wordnet similarity for word sense identification, in: C. Fellbaum (Ed.), WordNet: A Lexical Reference System and its Application, The MIT Press, 1998.

[18] J. J. Jiang, D. W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: International Conference Research on Computational Linguistics, 1997.

[19] D. Lin, An information-theoretic definition of similarity, in: ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 296–304.

[20] G. Hirst, D. S. Onge., Lexical chains as representations of context for the detection and correction of malapropisms, in: C. Fellbaum (Ed.), WordNet: A Lexical Reference System and its Application, The MIT Press, 1998.

[21] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: IJCAI, 1995, pp. 448–453.

[22] S. Banerjee, Extended gloss overlaps as a measure of semantic relatedness, in: In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, 2003, pp. 805–810.

[23] S. Patwardhan, Incorporating dictionary and corpus information into a context vector measure of semantic relatedness, Master's thesis (August 2003).

[24] J. Cardoso, The semantic web vision: Where are we?, IEEE Intelligent Systems 22 (5) (2007) 84–88.

[25] Z. Ding, Y. Peng, R. Pan, A bayesian approach to uncertainty modeling in owl ontology, in: Proc. of International Conference on Advances in Intelligent Systems - Theory and Applications, 2004.

[26] D. L. McGuinness, F. van Harmelen, OWL web ontology language overview, W3C recommendation, W3C, http://www.w3.org/TR/2004/REC-owl-features-20040210/ (Feb. 2004).

[27] S. Nadkarni, P. P. Shenoy, A causal mapping approach to constructing bayesian networks, Decision Support Systems 38 (2) (2004) 259 – 281. doi:DOI: 10.1016/S0167-9236(03)00095-2.

[28] G. J. McLachlan, T. Krishnan, The EM algorithm and extensions, 2nd Edition, John Wiley and Sons, 1997.

[29] M. Druzdzel, L. van der Gaag, Building probabilistic networks: "where do the numbers come from?" guest editors' introduction, Knowledge and Data Engineering, IEEE Transactions on 12 (4) (2000) 481 –486.

[30] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[31] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, in: Readings in uncertain reasoning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, pp. 415–448.

[32] F. V. Jensen, F. Jensen, Optimal junction trees, in: C. M. Kaufmann (Ed.), Proc. of the 10th Conf. on Uncertainty in Artif. Intel., San Mateo, 1994.

[33] G. F. Cooper, T. Dietterich, A bayesian method for the induction of probabilistic networks from data, in: Machine Learning, 1992, pp. 309–347.

[34] B. Scholkopf, A. Smola, R. Williamson, P. Bartlett, New support vector algorithms, Neural Networks 22 (2000) 1083–1121.

[35] T. Joachims, Making large-scale support vector machine learning practical, in: Advances in kernel methods: support vector learning, MIT Press, Cambridge, MA, USA, 1999, pp. 169–184.

[36] D. Heckerman, D. Geiger, D. M. Chickering, Learning bayesian networks: The combination of knowledge and statistical data, Machine Learning 20 (1995) 197–243, 10.1007/BF00994016.

[37] K. E. van de Sande, T. Gevers, C. G. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Transactions on

Pattern Analysis and Machine Intelligence 32 (2010) 1582–1596. doi:http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.154.

[38] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, Int. J. Comput. Vision 73 (2) (2007) 213–238. doi:http://dx.doi.org/10.1007/s11263-006-9794-4.

[39] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110. doi:http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

[40] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA, 2003, p. 1470.

[41] Trec video retrieval evaluation notebook papers and slides (Nov. 2010). URL `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`

[42] E. Yilmaz, J. A. Aslam, Estimating average precision when judgments are incomplete, Knowl. Inf. Syst. 16 (2) (2008) 173–211. doi:http://dx.doi.org/10.1007/s10115-007-0101-7.

[43] C. G. Snoek, M. Worring, Multimodal video indexing: A review of the state-of-the-art, Multimedia Tools and Applications 25 (2003) 5–35.

[44] T. M. Hospedales, S. Vijayakumar, Structure inference for bayesian multisensory scene understanding, IEEE Transanctions on Pattern Analysis and Machine Intelligence 30 (12) (2008) 2140–2157.

[45] K. Choi, S. Singh, A. Kodali, K. Pattipati, J. Sheppard, S. Namburu, S. Chigusa, D. Prokhorov, L. Qiao, Novel classifier fusion approaches for fault diagnosis in automotive systems, Instrumentation and Measurement, IEEE Transactions on 58 (3) (2009) 602–611.

[46] Z. Zeng, J. Tu, B. Pianfetti, T. Huang, Audiovisual affective expression recognition through multistream fused hmm, IEEE Transactions on Multimedia 10 (4) (2008) 570–577. doi:10.1109/TMM.2008.921737.

[47] H. Pan, S. Levinson, T. Huang, Z.-P. Liang, A fused hidden markov model with application to bimodal speech processing, IEEE Transactions on Signal Processing 52 (3) (2004) 573–581.

[48] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, Early versus late fusion in semantic video analysis, in: MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, New York, NY, USA, 2005, pp. 399–402. doi:http://doi.acm.org/10.1145/1101149.1101236.

[49] S. C. Hoi, M. R. Lyu, A multi-modal and multi-level ranking scheme for large-scale video retrieval, IEEE Transactions on Multimedia 10 (4) (2008) 607–619.

[50] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Transactions on Circuits and Systems for Video Technology 19 (5) (2009) 733–746.

[51] Z. Wang, M. Zhao, Y. Song, S. Kumar, B. Li, Youtubecat: Learning to categorize wild web videos, 2010, pp. 879 –886. doi:10.1109/CVPR.2010.5540125.

[52] S. Gupta, J. Kim, K. Grauman, R. J. Mooney, Watch, listen & learn: Co-training on captioned images and videos, in: ECML/PKDD (1), 2008, pp. 457–472.

[53] J. Fan, A. Elmagarmid, X. Zhu, W. Aref, L. Wu, Classview: hierarchical video shot classification, indexing, and accessing, IEEE Transactions on Multimedia 6 (1) (2004) 70–86.

[54] S. Wei, Y. Zhao, Z. Zhu, N. Liu, Multimodal fusion for video search reranking, IEEE Transactions on Knowledge and Data Engineering 99 (PrePrints).

[55] S. H. Lim, L.-L. Wang, G. DeJong, Integrating prior domain knowledge into discriminative learning using automatic model construction and phantom examples, Pattern Recogn. 42 (12) (2009) 3231–3240. doi:http://dx.doi.org/10.1016/j.patcog.2008.12.012.