Compound document analysis by fusing evidence across media

Spiros Nikolopoulos Christina Lakka
Ioannis Kompatsiaris
Informatics and Telematics Institute, CERTH, 6th km Charilaou-Thermi Road, Thessaloniki - Greece
{nikolopo, lakka, ikom}@iti.gr

Christos Varytimidis Konstantinos Rapantzikos
Yannis Avrithis
School of Electrical and Computer Engineering, National Technical University of Athens - Greece
{chrisvar, rap, iavr}@image.ntua.gr

Abstract

In this paper a cross media analysis scheme for the semantic interpretation of compound documents is presented. It is essentially a late-fusion mechanism that operates on top of single-media extractors output and it's main novelty relies on using the evidence extracted from heterogeneous media sources to perform probabilistic inference on a bayesian network that incorporates knowledge about the domain. Experiments performed on a set of 54 compound documents showed that the proposed scheme is able to exploit the existing cross media relations and achieve performance improvements.

1 Introduction

Cross media analysis is an approach that seeks to enhance multimedia interpretation by simultaneously exploiting evidence extracted across media. This type of analysis is motivated by the fact that information carried by different communication channels (e.g., visual, textual, audio) is important for a user to fully comprehend the intended meaning. Although, such an approach can be viewed as a fusion problem with different levels of abstraction (e.g., feature level and result level [1]), additional issues exist. These issues are of limited interest for typical fusion algorithms and usually depend on the nature of the analyzed resource, such as how to decide which modalities to correlate (e.g., based on spatial proximity, temporal co-occurrence) or how to incorporate cross media features into the analysis process.

In this paper, we focus on compound documents which are multimedia resources represented in various format types including Portable Document Format, Microsoft Word Document, Open Document Format, Microsoft PowerPoint Presentation, that are being massively generated and exchanged as a result of several knowledge management activities. Our goal is to develop a late-fusion scheme that uses probabilistic inference and domain knowledge to meaningful combine the effect of cross media evidence.

Related work in this field can be divided in methods that try to exploit cross media relations at the feature level and the ones operating on the result level. In the first case, mathematical transformations are employed to determine a space where features extracted from heterogeneous sources (visual, audio, text) can be optimally combined. The work presented by Magalhães and Rüger [2] is an indicative example of this category where information theory and a maximum entropy model are utilized to integrate heterogeneous data into a unique feature space. In [3] Wu et al. work on the same direction by introducing a method that initially finds statistical independent modalities from raw features and subsequently applies super-kernel fusion to determine their optimal combination. The linear correlation model is used in [4] by Li et al. for investigating different cross modal associations and a new method, that treats features from different modalities as independent subsets, is presented.

In the second case, sophisticated late-fusion mechanisms are employed for improving the accuracy of multimedia understanding as in [5], where Naphade and Huang propose a semantic video indexing, filtering and retrieval scheme. It is based on generic models representing semantic concepts and uses bayesian networks for fusing features extracted from heterogeneous sources. In the same direction Snoek et al. [6] use an authoring driven approach for semantic multimedia indexing, by combining the notions of content, style and context. Text, visual and audio features are concatenated into a single feature vector and the best analysis path is learned, on a per concept basis. Adams et al. in [7] use

Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) and Support Vector Machines (SVMs) for modeling the so-called atomic semantic concepts based on their visual, audio and text features. Subsequently, bayesian networks are used in a late-fusion setting to model high level concepts and perform semantic multimedia indexing using visual, audio and text cues.

The work presented in this paper is essentially a late fusion method and its main contribution concentrates on investigating the combination of ontologies with probabilistic inference mechanisms for consistently fusing the evidence extracted across different media types.

2 Framework description

2.1 Cross media analysis scheme

The proposed scheme can be viewed as a cross media one-class classifier of compound media resources. It uses domain knowledge and probabilistic inference for fusing the detection results of single-media extractors and decides positively if the resulting confidence degree is relatively high. Ontologies are used to express domain knowledge and conditional probabilities are employed to capture contextual information, in terms of quantifiable causality relations. All the above are integrated into a Bayesian Network (BN) using the methodology of [8]. During the analysis phase, visual and textual evidence obtained by applying single-media extractors on different types of media elements, are used to perform inference on the BN. Eventually, the proposed scheme verifies or rejects a hypothesis made about the semantic content of the analyzed resource based on the fused output of the BN. In this way, we benefit both from ontologies, that are used to identify the relations between concepts, and probabilistic methods, that are able to quantify and propagate the causality of these relations.

The tasks carried out by our framework are a) adopting a dismantling methodology for handling the compound media resources, b) independently applying textual and visual analysis on the dismantled elements for extracting single-media evidence, c) fusing the single media evidence by performing probabilistic inference on a BN that incorporates domain knowledge and eventually d) decide about the semantic content of the analyzed resource based on the likelihood of the BN root node. Fig. 1 demonstrates the functional relations between the components of the proposed framework.

2.2 Compound documents dismantling & elements synchronization

Typically, compound documents are multimedia documents that incorporate more that one types of media elements in the same digital resource. OpenDocument, Microsoft Office's documents, PDF, web pages are indicative representation formats of such documents where visual and textual elements co-exist. Apart from visual and textual information, these documents carry additional features that originate from the document layout, such as the spatial proximity of two information elements (e.g., caption near an image frame) and 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. Specifically, the wide 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 cross media analysis.

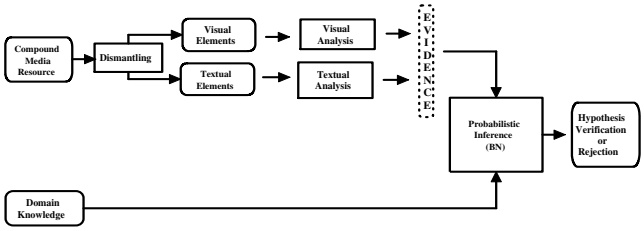

Figure 1. Cross media analysis framework

All the above, makes the existence of a dismantling mechanism an important prerequisite for cross media analysis. This mechanism will be able to disassemble a compound document to its constituent parts and decide which of these elements should be considered simultaneously by the fusion process. Assuming a certain layout for the analyzed documents we admit that a different topic is covered in each document page and disregard cases where more that one topics exist in the same page. 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-media extractors that are independently applied on the media elements residing on the same page.

2.3 Single-media analysis modules

The process of probabilistic inference is triggered by the visual and textual evidence obtained from single-media extractors. The purpose of this section is to provide some implementation details on their functionality. At this point it is important to note that although the described framework involves only images and text, it could be seamlessly used to fuse evidence originating from other types of media (e.g., video, audio), provided that the corresponding single media extractors are available.

2.3.1 Visual analysis

The source of visual evidence is a global image classifier and four local-based concept detectors. The reason for employing both global and local concept detectors was to exploit visual evidence extracted in different levels of granularity. The global classifier is based on the MPEG-7 EdgeHistogram descriptor [9], and uses Support Vector Machines (SVM) for learning the classification model. The local concept detectors are based on the Viola and Jones detection framework [10] that uses Haar-like features to represent the visual information and the AdaBoost algorithm to train the detector. By using integral images, each Haar-like feature is computed in constant time, which results in an extremely fast overall detection process, despite the fact that for each concept, a detector scans images in every possible position and scale. The AdaBoost training algorithm selects the Haar-features from a pool of 100,000 features that best describe the depicted concept. Computational time is further reduced by the use of several low precision, fast detectors connected in a cascade, instead of one high precision and slow detector. The output of the global classifier is a binary value indicating the absence or presence of a concept, while the local classifiers also output the exact location and scale of the detected concepts, as demonstrated in Fig. 5.

2.3.2 Textual analysis

For obtaining textual evidence, custom modules are employed to analyze textual descriptions. The functionality of these modules consists of finding references to a specific concept, based on a look-up table containing different linguistic expressions of this concept, as well as derivatives, synonyms, etc. Regular expressions are used to facilitate this functionality and provides the cross media analysis scheme with a binary value indicating the absence or presence of a concept.

2.4 Domain knowledge & probabilistic inference

2.4.1 Ontologies

Ontologies have emerged as a powerful tool able to express knowledge in different levels of granularity, handle the diversity of content essence and govern its semantics [11]. For the purposes of our work, we use OWL-DL in order to express domain knowledge as a hierarchical structure K_D that associates domain concepts. Apart from ontologies, other representation structures capable of equivalently reflecting human experience exist (e.g., conceptual graphs). However, the use of ontologies was advocated by their wide acceptance and appeal to the area of knowledge engineering [11].

2.4.2 Bayesian Networks

The ability of Bayes' theorem to compute the posterior probability of a hypothesis by relating the conditional and prior of two random variables and essentially update or revise beliefs in light of new evidence, was the reason for considering the use of bayesian networks for fusing cross media evidence.

A Bayesian network is a directed acyclic graph whose nodes represent variables and whose arcs encode the conditional dependencies between them. Hence, a bayesian network can be used to facilitate three dimensions of perception: a) provide the means to store and utilize domain knowledge K_D, an operation that is served by the network structure and prior probabilities, b) organize and make accessible information concerning the amount of influence between evidence and hypotheses, which is supported by the Conditional Probability Tables (CPTs) and c) allow the propagation of evidence beliefs using message passing algorithms, an action facilitated by the Bayes' theorem. For the purposes of our work we employed a methodology similar to [8] for determining the structure of a bayesian network out of an OWL ontology. Concerning the network parameters, Expectation Maximization was applied on observation data for calculating the CPTs of all network nodes. Eventually the junction tree algorithm [12] was employed for performing message passing belief propagation.

3 Experimental Study

3.1 Experimental platform

The domain selected for evaluating our framework was a competitor analysis scenario as realized in Centro Ricerche Fiat (CRF) (http://www.crf.it/). The goal of a competitor analysis department is to constantly monitor the existent competitors' products, understand market trends and try to anticipate customer needs. In a typical scenario the main role is played by the person responsible for data acquisition. It's role is to daily inspect a number of resources such as WWW pages, car exhibitions, car magazines, etc, that are likely to publish material of potential interest. Most of these multimedia documents use both visual and textual descriptions. The focus of our analysis was to evaluate these documents with respect to their interest for the *car components ergonomic design*. Therefore, we worked under the assumption that a document will be worth considering by the competitor analysis department if it contains information talking about the design and ergonomic features of car components. This fact motivated the construction of a classifier recognizing this type of content by evaluating evidence extracted across media.

For the purposes of our evaluation a dataset of 54 pdf

documents (containing $\approx$ 200 pages) was provided by CRF, 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 xpdf library (http://www.foolabs.com/xpdf/). All media elements extracted from the same page were kept together so as not to loose any conceptual relations originating from the document's layout. The textual descriptions were gathered in a single txt file while the visual representations were extracted to independent image files as depicted in Fig. 2.

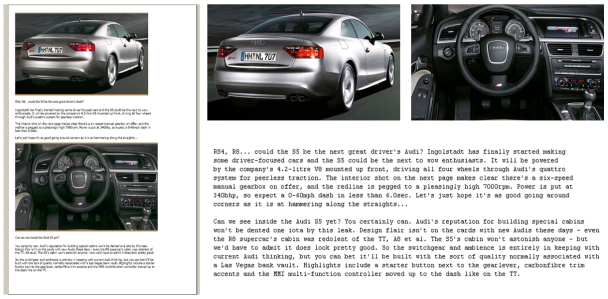

Figure 2. Dismantling a document to its constituent parts

CRF has developed a domain ontology that describes the various processes carried out within the competitor analysis department. However, due to its multi-functional purpose this ontology was deemed inappropriate for our goals, since what we are interested in are the concrete associations between the topic of *car components ergonomic design* and any visual or textual cues that could potentially lead us to the conclusion that a document page talks about this topic. For this reason, we have developed in co-operation with CRF a new lightweight ontology that is mostly concerned with the concepts related to the ergonomic design of car components, as show in Fig 3.

The ontology design process involved going through a sufficient number of documents and identifying which keywords and images are usually present when the page subject is concerned with *car components ergonomic design*. The purpose of this ontology was to establish qualitative associations between the identified ontology concepts and indicate which evidence provide support for which hypothesis.

Two different annotation efforts were carried out by human subjects 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 $\approx$ 200 document pages, and record in an annotation file whether it is of interest for the *car components ergonomic design* or not. This file was used as ground truth for measuring the performance of our framework.

The second annotation effort aimed at generating a sufficient number of observations for training the BN (i.e., estimate the dependencies between nodes). This annotation task involved going through each document page and marking in an annotation file all "mentions", textual or visual, of the concepts belonging to the lightweight ontology. The result of this annotation process was a file with $\approx$ 200 entries, reflecting the frequency of co-occurrence between domain concepts. This file was utilized for estimating the prior probabilities and calculating the CPTs of the BN depicted in Fig. 4, generated from the lightweight ontology according to Section 2.4.

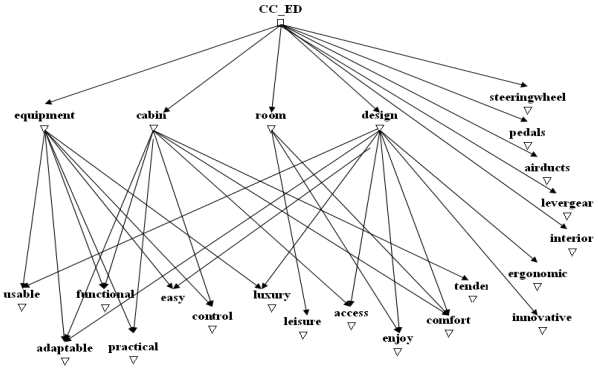

Figure 3. Domain ontology concepts

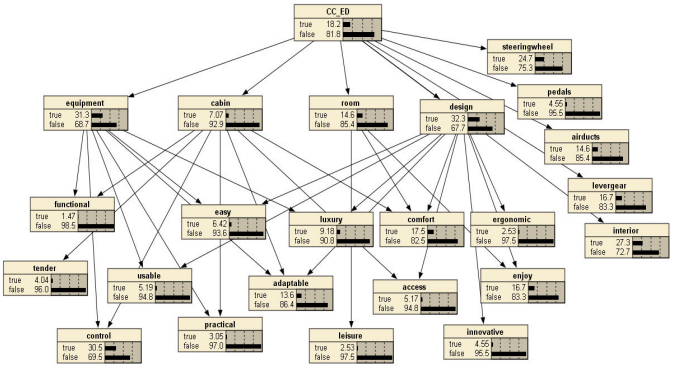

Figure 4. Bayesian Network

3.2 Experiment design

The goal of our experimental study was to investigate whether evidence gathered across media can actually improve the performance of a compound document analysis scheme, compared to the cases where evidence are derived solely, from textual or visual elements.

In this context, four local classifiers trained to detect different car components, namely air ducts, steering wheels, gear levers and car pedals and one global classifier iden-

tifying the depicted car view, namely interior or exterior, served as the information extraction modules for visual evidence. The global classifier was trained on a set of 3500 images that was manually annotated, while the region classifiers were trained on a dataset of 690 images of car interiors that were also manually annotated in a region based manner. On the other hand, for textual evidence, 18 custom modules were employed to analyze the textual descriptions of each page using Table 1 as look-up table, see Section 2.3.2.

Table 1. Look-up table used for textual analysis

Concept	Linguistic Expression
access	access, accessibility, accessible, accessing
cabin	cabin, compartment
design	design, designed, designing
leisure	leisure, relaxation, rest
comfort	comfort, comfortable, comfortableness
easy	easy, easiness, ease
enjoy	enjoy, enjoyable, pleasure, joy
luxury	luxury, luxurious, luxe, de luxe, deluxe
room	room, space, spacious
tender	tender, pulpy, peaty
ergonomic	ergonomic, ergonomics, bioengineering , biotechnology
equipment	equipment, instrument
innovative	innovative, innovation, innovational, modern
usable	usable, usability, useable, utilizable
adaptable	adaptable, adaptive, adaption
control	control, controlling
functional	function, functional, functioning
practical	practical

In this way, a single-media information extraction module, producing binary output (i.e., presence or absence), was attached to each network node of Fig. 4, expect of course the *CC_ED* node which is the one modeling the concept of *car components ergonomic design* and determines the output of the cross media classifier. The analysis process involves applying all aforementioned information extraction modules on the constituent parts of a document page and according to their output, update the value of the corresponding network nodes. Upon nodes update an inference process is triggered that progressively modifies nodes likelihood, using message passing belief propagation. Eventually, the likelihood of *CA_ED* node is compared against a predefined threshold that determines the decision of our framework. An illustration of this procedure is depicted in Fig. 5.

The cross media classifier of Fig. 5 is capable of producing an output independently of the amount and origin of the evidence injected into the network. When no evidence are injected, the confidence degree of the fact that the analyzed page is concerned with *car components ergonomic design*, is equal to the frequency of appearance of such pages in the training set (i.e., prior probability of *CC_ED* node). As evidence are injected into the network this degree modifies according to the causality relations that have been learned

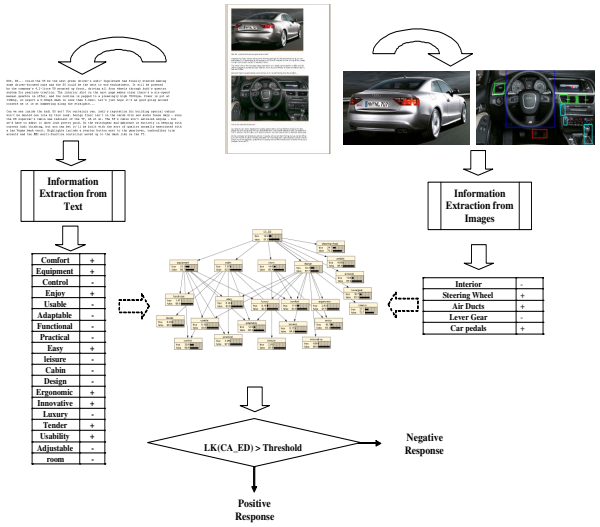

Figure 5. Inference process illustration

from the BN. This property allows us to evaluate the performance of the cross media classifier using evidence extracted only from text, only from images, or both.

3.3 Results

Recall versus precision curves were utilized for evaluating the performance of our framework in testing all 200 document pages. The threshold value of Fig. 5 was uniformly scaled between [0,1] for drawing the evaluation curves depicted in Fig. 6. The $x = y$ line (dotted line) has been also drawn to give an indication of the point where a balanced tradeoff between recall and precision is obtained.

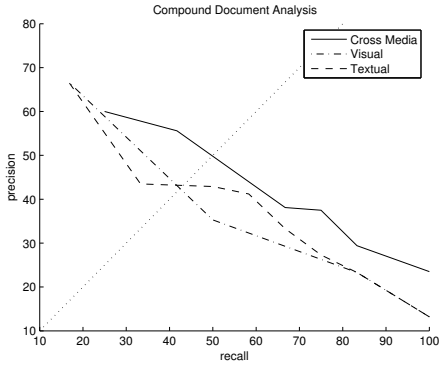

Figure 6. Performance evaluation curves

By inspecting the evaluation curves of Fig. 5 one can verify that the configuration of the framework using cross media evidence sufficiently outperforms the cases where evidence originate exclusively from one media type. This

leads us to the conclusion that a BN incorporating domain knowledge and performing evidence-triggered probabilistic inference, is an efficient late-fusion mechanism in terms of exploiting the existing cross media relations. This is mainly due to the ability of proposed framework to exploit domain knowledge in a bayesian setting and for this reason, meaningful evaluate the co-existence of evidence regardless of their origin.

3.4 Discussion & future work

In this paper we show how an evidence driven probabilistic inference framework that incorporates domain knowledge, can be used to facilitate cross media analysis of compound documents. Experiments showed that the proposed scheme performs optimally when provided with cross media evidence, compared to the cases where these evidence are derived solely from textual or visual elements. One important drawback of the aforementioned scheme is that it needs a deep modeling of the context that requires a sufficiently large amount of observations required for training the BN and learning the true causality relations. Taking into consideration that cross media annotation is an even more tedious and difficult task than single media annotation, the time and effort required to generate a sufficiently large amount of reliable annotations could hinder the adoption of such schemes in production systems.

Eventually, as future work, the incorporation of non-crisp single media information extraction modules could greatly boost the efficiency of the aforementioned scheme. The fact that all evidence are injected into the network as hard evidence (i.e., confidence equal to 100%) essentially disregards the inherent capability of BN to meaningful handle uncertainty. However, in this case, special care should be given on the type of probability distribution followed by each single media/modality extractor output. Investigating normalization schemes that could alleviate the effect of such cases is also included within our plans for future research.

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.

References

[1] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," in *Computing and Information Systems*, vol. 7. University of Paisley, Feb 2000, pp. 1–10.

[2] J. Magalhaes and S. Rüger, "Information-theoretic semantic multimedia indexing," in *CIVR '07*. New York, USA: ACM, 2007, pp. 619–626.

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

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

[5] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 141–151, 2001.

[6] G. Snoek, M. Worring, J. Geusebroek, D. Koelma, F. Seinstra, and A. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Transactions on Pattern. Anal. and Mach. Intel.,*, vol. 28, no. 10, pp. 1678–1689, Oct. 2006.

[7] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 170–185, 2003.

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

[9] B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada, "Colour and texture descriptors," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, vol. 11, no. 6, pp. 703–715, Jun 2001.

[10] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR (1)*, 2001, pp. 511–518.

[11] J. Cardoso, "The semantic web vision: Where are we?" *IEEE Intelligent Systems*, vol. 22, no. 5, pp. 84–88, 2007.

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