A cross media approach for compound document analysis

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

A cross media analysis scheme for the semantic interpretation of compound documents is presented. The proposed scheme is essentially a late-fusion mechanism that operates on top of single-media extractors output. Evidence extracted from heterogeneous sources are used to trigger probabilistic inference on a bayesian network that encodes domain knowledge and quantifies causality. 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

Multimedia data require cross media analysis in order to become fully comprehensible, since information carried by the different media (e.g. visual, audio, text) is important for the user. In this perspective, cross media/modal analysis is considered as an approach that seeks to enhance their interpretation by simultaneously exploiting evidence extracted from different media types that co-exist in the same digital source.

Although, this type of analysis can be viewed as a fusion problem with different levels of abstraction (e.g., feature level and result level [1]), additional problems exist, 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. These problems are of limited importance for typical fusion algorithms and highly depend on the nature of the analyzed resource.

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 investigate the existing cross media relations and develop a late-fusion scheme that uses probabilistic inference, incorporates domain knowledge and considers evidence extracted across media.

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 Mogalhães 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 resources. 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 approach to model high level concepts and perform semantic multimedia indexing using visual, audio and text cues.

The work presented in this paper is a late fusion method and its main contribution concentrates on investigating the combination of ontologies with probabilistic inference mechanisms for consistently fusion cross media evidence, extracted from compound documents.

2 Framework description

2.1 Cross Media Analysis Scheme

The proposed scheme implements a cross media classifier that fuses the detection results of single-media extractors, using domain knowledge and probabilistic inference. In order to develop this classifier, domain knowledge, expressed through ontologies, and contextual information, captured in conditional probabilities, are integrated into a decision model that bears the characteristics of a Bayesian Network (BN). A methodology for generating a BN out of an OWL ontology is employed for this purpose [8]. Evidence obtained by applying single-media extractors on different types of media elements, are used to trigger an inference algorithm grounded on message passing belief propagation [9].

Eventually, the cross media analysis scheme verifies or rejects a hypothesis made about the semantic content of the analyzed resource. In this way, our decision model benefits both from the logic-based approaches that are used to identify the relations between concepts, as well as the probabilistic approaches that are able to quantify and propagate the causality of these relations. In this context, the tasks carried out by our framework are a) adopting a methodology for handling the various types of media elements (i.e., text and images) existing under the same compound document, b) independently applying textual and visual analysis algorithms for extracting single-media information, c) fusing the intermediate results using a BN incorporating domain knowledge and eventually d) facilitating a cross media classifier that uses the existing visual and textual evidence to decide about the semantic content of the analyzed resource. Fig. 1 demonstrates the functional relations between the components of the proposed framework.

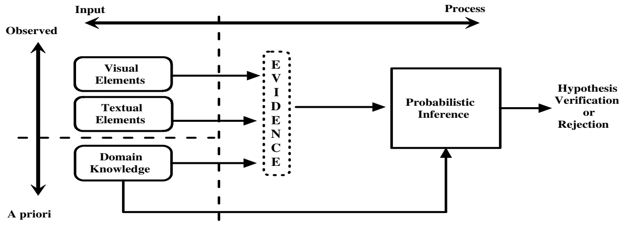


Figure 1. Cross media analysis framework

2.2 Compound Documents & elements synchronization

Typically, compound documents are multimedia documents that incorporate more that one types of media elements in the same digital resource, such as images and text. 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 (textual caption near to 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 and make them available for cross media 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 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. This is necessary for allowing the cross media analysis scheme to benefit from the document's layout information.

Motivated by the fact that this type of documents usually cover a different topic on each page, we consider all media elements of the same document page to be conceptually related. For this reason, the proposed framework analyzes a document on a per page basis, by fusing the output of single-media extractors, independently applied on the media elements residing on the same document page.

2.3 Single-media analysis modules

As already mentioned in Section 2.1, the probabilistic inference process is triggered by the visual and textual evidence obtained from single-media analysis modules. The purpose of this section is to provide some details on their functionality.

2.3.1 Visual analysis

The source of visual evidence for our framework is a global image classifier (interior/exterior) and four local-based concept detectors. The global classifier uses MPEG7 descriptors as visual features and a Support Vector Machine (SVM) for the classification process. The 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 can be shown 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

Domain knowledge will have to be elucidated and represented in machine understandable format in order to be exploitable by our framework. 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 [11]. If we let N_C to be the set of unary predicate symbols that are used to denote concepts, R to be the set of binary predicates that are used to denote relations between concepts and O the algebra defining the allowable operands for these sets, the part of experience that relates to the domain knowledge can be represented using N_C, R, O.

For the purposes of our work, we use OWL-DL [12] in order to express domain knowledge as a hierarchical structure $K_D = S(N_C, R, O)$ 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 for knowledge representation 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 probabilities 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 in order to facilitate probabilistic inference.

A Bayesian network is a directed acyclic graph $G = (V, A)$ whose nodes $v \in V$ represent variables and whose arcs $a \in A$ 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 [13] was applied on observation data for calculating the CPTs of all network nodes. Eventually the junction tree message passing belief propagation algorithm that was proposed by Lauritzen and Spiegelhalter [9], was utilized for performing evidence driven probabilistic inference.

3 Experimental Study

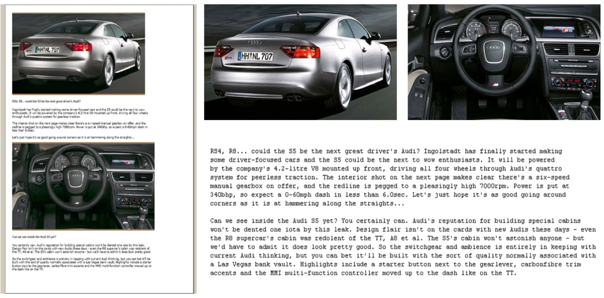
3.1 Experimental Platform

The domain selected for evaluating the previously described scheme was a competitor analysis scenario as realized in Centro Ricerche Fiat (CRF)[1]. The goal of a competitor analysis procedure is to constantly monitor the existent

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

competitors' products, analyze technological innovations, understand market trends and eventually try to anticipate customer needs. In a typical competitor analysis scenario the main role is played by the person responsible for data acquisition. The role of this person is to daily inspect a number of resources such as WWW pages, car exhibitions, car magazines, car brochures, leaflets 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 we have collected 54 pdf documents (containing $\approx$ 200 pages) 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[2]. All media elements extracted from the same page were collected on the same folder 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.



```
RS4, R8... could the S5 be the next great driver's Audi? Ingolstadt has finally started making
some driver-focused cars and the S5 could be the next to wow enthusiasts. It will be powered
by the company's 4.2-litre V8 mounted up front, driving all four wheels through Audi's quattro
system for peerless traction. The interior shot on the next page makes clear there's a six-speed
manual gearbox on offer, and the redline is pegged to a pleasingly high 7000rpm. Power is put at
340bhp, so expect a 0-60mph dash in less than 6.0sec. Let's just hope it's as good going around
corners as it is at hammering along the straights...

Can we see inside the Audi S5 yet? You certainly can. Audi's reputation for building special cabins
won't be dented one iota by this leak. Design flair isn't on the cards with new Audis these days - even
the R8 supercar's cabin was redolent of the TT, A8 et al. The S5's cabin won't astonish anyone - but
we'd have to admit it does look pretty good. So the switchgear and ambience is entirely in keeping with
current Audi thinking, but you can bet it'll be built with the sort of quality normally associated with
a Las Vegas bank vault. Highlights include a starter button next to the gearlever, carbonfibre trim
accents and the MMI multi-function controller moved up to the dash like on the TT.
```

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 and establishes the associations between existing concepts. 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 a new lightweight ontology that is mostly concerned with the concepts related to the ergonomic design of car components.

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 concepts (cabin, room, design etc.) and indicate which evidence provide support for which hypothesis. The constructed hierarchical structure K_D is depicted in Fig. 3.

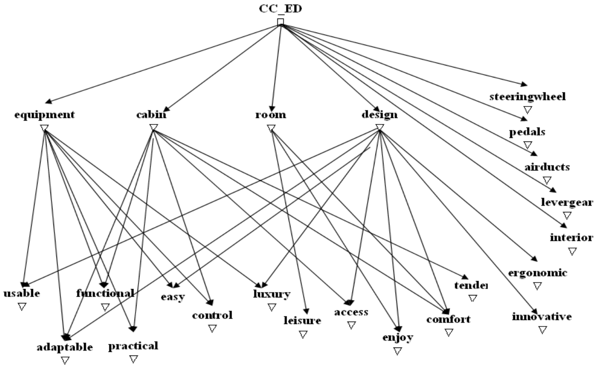


Figure 3. Domain concepts hierarchical structure

Two different annotation efforts were carried out for the purposes of our work. Since we have decided to consider the compound documents on a per page basis, the first annotation effort was to manually inspect and evaluate each document page as to whether it could be of interest for the ergonomic design of car components. The reason for constructing this type of ground truth was to allow measuring the performance of the proposed scheme.

The second annotation effort aimed at the generation of a sufficient number of observations for training the BN. 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.

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

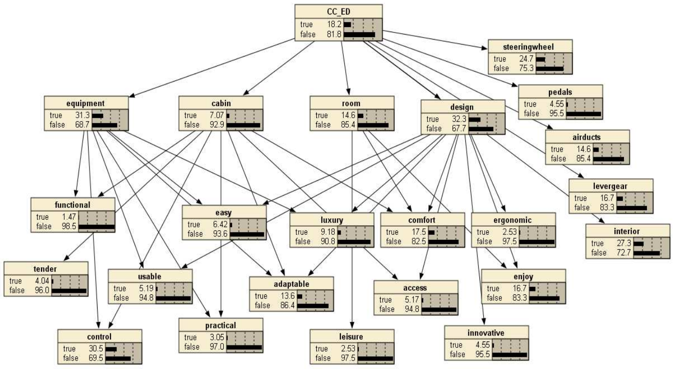


Figure 4. Bayesian Network

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 identifying 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, as described in Section 2.3.2. 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, except 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 output of the classifier. 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. As evidence are injected into the network this degree modifies according to the causality relations that have been learned 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.

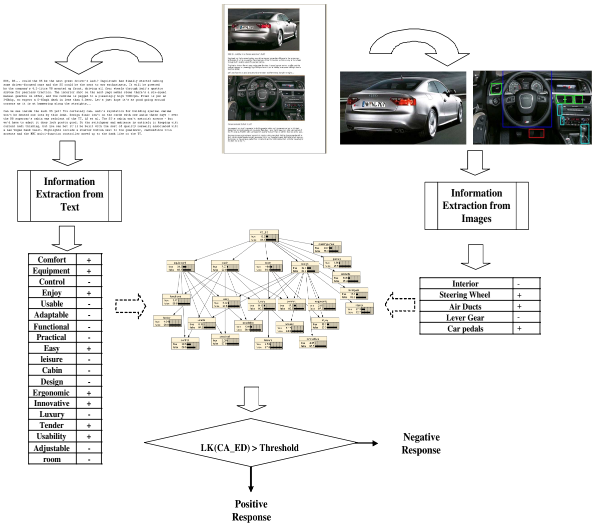


Figure 5. Inference process illustration

3.3 Results

Recall versus precision curves were utilized for evaluating the performance of the classifier against the manually constructed ground truth. 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.

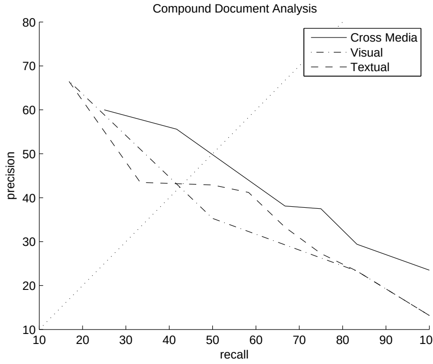


Figure 6. Performance evaluation curves

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.

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 this evidence is derived solely from textual or visual resources. One important drawback of the aforementioned scheme that was partially reflected in the evaluation results, is the large amount of observations required to train the BN and learn the 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 sufficient large amount of reliable annotations could hinder the adoption of such schemes.

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, in terms of mean value and deviation, followed by each single media/modality extractor output. For instance, if the output of two different classifiers that follow the same probability distribution but exhibit considerably different values for their deviation, are used to perform probabilistic inference within the same BN, it is likely that the fused decisions will be dominated by the suggestions of the classifier with the highest deviation. 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," vol. 7, pp. 1–10, Feb. 2000.

[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 Int. Conf. on Adv. in Intelligent Systems - Theory and Applications*, Nov. 2004.

[9] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," pp. 415–448, 1990.

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

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