KNOWLEDGE-ASSISTED CROSS-MEDIA ANALYSIS OF AUDIO-VISUAL CONTENT IN THE NEWS DOMAIN

Vasileios Mezaris[1], *Spyros Gidaros*[1], *Georgios Th. Papadopoulos*[1], *Walter Kasper*[2],
Roeland Ordelman[3], *Franciska de Jong*[3], *Ioannis Kompatsiaris*[1]

[1]Informatics and Telematics Institute / Centre for Research and Technology Hellas,
1st Km Thermi-Panorama Rd, Thessaloniki 57001, Greece
[2]DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrucken, Germany
[3]University of Twente, 7500 AE Enschede, The Netherlands

ABSTRACT

In this paper, a complete architecture for knowledge-assisted cross-media analysis of News-related multimedia content is presented, along with its constituent components. The proposed analysis architecture employs state-of-the-art methods for the analysis of each individual modality (visual, audio, text) separately, and proposes a fusion technique based on the particular characteristics of News-related content for the combination of the individual modality analysis results. Experimental results on news broadcast video illustrate the usefulness of the proposed techniques in the automatic generation of semantic video annotations.

1. INTRODUCTION

Access to media content, either amateur or professional, is nowadays a key element in business environments as well as everyday practice for individuals. The widespread availability of inexpensive media capturing devices, the significant proliferation of broadband internet connections and the development of innovative media sharing services over the World Wide Web have contributed the most to the establishment of digital media as a necessary part of our lives. However, this increased significance of digital media as a means of communication has inevitably resulted in a tremendous increase in the amount of media material created every day. This presents new possibilities, particularly in the area of News manipulation and delivery, but also presents new and important challenges regarding the efficient organization, access and presentation of media material. The cornerstone of the efficient manipulation of media material is the understanding of the semantics of it, a goal that has long been identified as the “Holy grail” of content-based media analysis research [1].

Knowledge-assisted analysis has recently emerged as a promising approach towards the understanding of the semantics of multimedia content [2]. It refers to the coupling of traditional analysis techniques such as segmentation and feature extraction with prior knowledge for the domain of interest. The introduction of prior knowledge to the analysis task is a natural choice for countering the drawbacks of traditional approaches, which include the inability to extract sufficient semantic information about the multimedia content (e.g. semantic objects depicted and events presented, rather than lower-level audiovisual features) and the ambiguity of the extracted information (e.g. visual features may be very similar for radically different depicted objects and events).

Depending on the adopted knowledge acquisition and representation process, two types of approaches can be identified in the knowledge-assisted analysis literature: implicit, realized by machine learning methods, and explicit, realized by model-based approaches. The use of machine learning techniques has proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) [3], [4]. On the other hand, model-based analysis approaches make use of prior knowledge in the form of explicitly defined facts, models and rules, i.e. they provide a coherent semantic domain model to support inference [2], [5]. An approach combining the characteristics of both the above types is proposed in this work for the News domain, emphasizing on the cross-media aspects of the proposed knowledge-assisted analysis architecture.

2. ANALYSIS ARCHITECTURE

2.1. Knowledge Representation Overview

In a knowledge-assisted multimedia analysis system, such as the proposed, knowledge representation serves two main purposes: the representation of prior knowledge for the domain,

This work was supported by the European Commission under contract FP6-027685 MESH.

and the modelling of the analysis process and its results. To serve these goals, an ontology infrastructure has been built that comprises two main parts: a domain ontology, that represents the prior knowledge for the domain, and a multimedia ontology.

The developed domain ontology is based on an extension of the IPTC[1] tree for the News domain and includes a hierarchy or classes that range from rather abstract ones, such as "disaster and accident", to specific ones, such as "earthquake", "flood", etc. The latter are the least abstract classes to which an elementary news item can be associated; they are also referred to as subdomains and are denoted by D_l, $l = 1, \ldots, L$. In terms of visual analysis, these are at the same time the most abstract classes to which attempting to directly classify any piece of visual information based on its low-level visual properties would make sense. Consequently, in order to support efficient visual analysis, a set of even less abstract classes, i.e. local concepts C_j, $j = 1, \ldots, J$ describing possible spatial regions of an image rather than entire images, is also defined. Contextual information in the form of spatial relations between these concepts, as well as in the form of concept frequency of appearance, are also included in this ontology.

The developed multimedia ontology, on the other hand, is an expansion of the DOLCE IO pattern. The Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) is a core ontology that includes only the most reusable and widely applicable upper-level categories. To support its easy extension, modules that express generic conceptual patterns (termed Design Patterns) have also been defined for it. In this work, the DOLCE IO pattern was selected for developing the multimedia ontology and to this end it was further enriched with two additional properties, namely the 'hasDecomposition' and the 'refersTo' properties. The resulting information object is denoted with the term MMIO. The MMIO model combines the DOLCE IO pattern with the MPEG-7 standard for the representation of media content. More details on the developed multimedia ontology can be found in [6].

2.2. Multimedia Processing Overview

In general, the procedure of the analysis is as follows: the input to the analysis architecture is a multimedia news item; this is firstly decomposed to its individual modality components and each such is represented as a sub class of MMIO in the multimedia ontology. Each individual modality is then processed separately, i.e. is decomposed to its basic structural elements using modality-specific segmentation algorithms, is compactly represented using appropriate features, and is analyzed using suitable knowledge-assisted analysis techniques. This process results in the association of each subclass of MMIO with one or more subdomains of the domain ontology. Then, a cross-media analysis mechanism is invoked to combine the individual modality analysis results, to remove ambiguities and contradictory outputs, and produce a final semantic interpretation of the multimedia content.

3. SINGLE MODALITY ANALYSIS

3.1. Visual Analysis

The analysis of the visual information involves several processing steps that include basic ones such as shot decomposition and visual feature estimation, as well as knowledge-assisted analysis techniques that include global keyframe and region level classification as well as the fusion of these classification results to a single hypothesis set about the subdomain membership of the examined news item.

Preprocessing starts with temporal video decomposition to shots, which are the elementary video streams that can be associated with one of the subdomains D_l. The algorithm of [7] is used to this end. Following temporal video decomposition, a keyframe is identified for each shot and a rich set of MPEG-7 visual descriptors [8] that are necessary for its compact representation and subsequent processing is calculated for it. Descriptors are extracted both at the global image level and at the region level, after spatial segmentation is performed using the method of [9]. Currently, the Scalable Color, Homogeneous Texture and Edge Histogram descriptors are employed for global image classification, while Scalable Color, Homogeneous Texture, Region Shape and Edge Histogram are employed for region classification. As a final pre-processing stage, face detection is performed using a variant of the method of [10]; given a keyframe of the shot, the presence of one or more human faces is detected and their locations on the image grid are specified, allowing among others the evaluation of the area of the image that is taken by the face(s).

Following the preprocessing stage, a set of techniques aiming at the association of pieces of visual information with classes of the domain ontology is applied, starting with global image classification. In order to perform classification of the examined visual content into one of the subdomains defined in the ontology using global-image descriptions, a compound visual feature vector is initially formed from the previously specified MPEG-7 descriptors. Then, a Support Vector Machine (SVM) [11] structure is utilized to compute the class to which each piece of visual information belongs. This comprises L SVMs, one for every defined subdomain, each trained under the "one-against-all" approach. For the purpose of training, an appropriate training set of images manually classified to subdomains is assembled and is used. At the evaluation stage, each SVM returns for every image of unknown subdomain membership a numerical value in the range [0, 1]. This value denotes the degree of confidence to which the corresponding visual content is assigned to the subdomain rep-

[1] International Press Telecommunications Council, http://www.iptc.org /pages/index.php

resented by the particular SVM, and is computed from the signed distance of it from the corresponding SVM's separating hyperplane using a sigmoid function [12]. For each keyframe, the maximum of the L calculated degrees of membership indicates its classification based on global-level features, whereas all degrees of confidence, H_l, $l = 1, \dots, L$, constitute its subdomain hypothesis set.

Region-level classification follows, using the aforementioned SVM structure to compute an initial region-concept association for every visual content segment. Similarly to the previous case, an individual SVM is introduced for every local concept C_j of the employed ontology, in order to detect the corresponding association. Each SVM is again trained under the "one-against-all" approach. For that purpose, a training set of regions generated by means of automatic segmentation followed by manual region classification of the images is employed. As a result, for each region S_k, $k = 1, \dots, K$, the degree of confidence with which it is assigned to each of the local concepts C_j in the domain ontology is computed; the maximum of them indicates the classification of the examined region based on local-level features, whereas all degrees of confidence, h_j, $j = 1, \dots, J$, constitute its hypothesis set. The hypothesis sets for all regions of the keyframe are subsequently employed for inferring a new keyframe-subdomain association hypothesis set $\boldsymbol{H}_l$.

After global concept association has been performed using global and local-level information, a fusion mechanism is introduced for deciding upon the final keyframe - global concept association. This has the form of a weighted summation $G_l = \mu_l \cdot H_l + (1 - \mu_l) \cdot \boldsymbol{H}_l$. The subdomain with the highest G_l value constitutes the final semantic annotation of the respective video shot based of its visual information. A genetic algorithm is used for optimizing the weights μ for each subdomain, to account for the varying relevant importance of global and local information in different subdomains [6].

3.2. Audio Analysis

Audio analysis in the context of information understanding covers several themes. From a processing point of view, analysis of an audio file starts with the partitioning of the audio stream into segments such as speech, non-speech and music, and dividing the speech into speaker segments. In the context of automatic speech recognition (ASR) this step is referred to as speech activity detection (SAD) [13] and speaker diarization ("who speaks when") [14]. Class-specific audio analysis such as speaker dependent ASR or the classification of speaker characteristics is performed next. Using ASR to exploit the linguistic content that is available as spoken content in videos has proven to be helpful to bridge the semantic gap between media features and information understanding [15]. This is confirmed by the results of the TREC series of Workshops on Video Retrieval (TRECVID[2]). The TRECVID test collections contain not just video, but also ASR-generated transcripts of segments containing speech. Systems that do not exploit these transcripts typically do not perform as well as the systems that do incorporate speech features in their models [16].

ASR supports the conceptual querying of video content and the synchronization to any kind of textual resource that is accessible, including other annotations for audiovisual material [17]. The potential of ASR-based indexing has been demonstrated most successfully in the broadcast news domain [18]. Typically large vocabulary speaker independent continuous speech recognition (LVCSR) is deployed to this end. To support efficient recognition, it is crucial that the speech recognition system can adapt to the linguistic variations in the target collections, to reduce the number of out-of-vocabulary (OOV) words, i.e. words unknown to the recognition system. In the News domain, named entities (people, volcanoes, hurricanes, cities etc.) have a high change of being OOV unless they appear very frequently (e.g. "Bush", "Amsterdam") and are therefore explicitly added to the recognition vocabulary. To alleviate for this, words that are expected to occur on the basis of prior information (e.g. recent news reports) are dynamically included using a "rolling language model" approach [19]. The main processing stages of audio analysis are depicted in Fig.1.

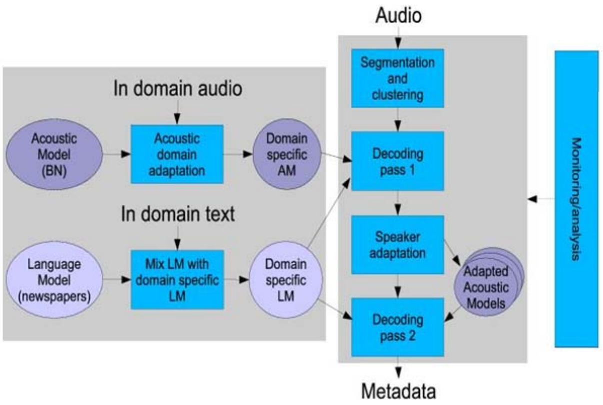


Fig. 1. Overview of the audio analysis chain

3.3. Linguistic Analysis

Textual information analysis of multimedia News-related material may be applicable to textual information coming from a number of different sources: textual annotations produced manually by the content creators, when such information is available; text extracted from the video frames by means of OCR techniques; and ASR transcripts produced by audio information analysis, as discussed above. In all three cases, textual information analysis will exploit for its application a suitable temporal decomposition, depending on the source of

[2] http://trecvid.nist.gov

textual information: for manual annotations, the temporal decomposition that has been manually defined for them; for text coming from OCR, all text extracted from a single keyframe will be analyzed together; finally, for ASR transcripts, it will be performed at the speaker level (i.e. exploiting the results of speaker diarization performed as part of the audio processing), independently processing each uninterrupted speech segment of a single speaker. In this work, we focus on exploiting the previously produced ASR transcripts to identify information such as locations and events of interest, according to the employed domain ontology.

Typical subtasks in linguistic analysis for information extraction involve

- Named Entity Recognition (NER) such as persons, organizations, locations, products and dates.
- Co-reference analysis to identify references to the same objects.
- Terminology extraction for identifying domain relevant vocabularies and relate them to the semantic concepts.

In this work, the SProUT (Shallow Processing with Unification and Typed Feature Structures) platform is used as core engine for linguistic analysis [20], [21]. SProUT is equipped with a set of reusable Unicode-capable online processing components for basic linguistic operations, ranging from tokenization to reference matching. Since typed feature structures (TFS) are used as a uniform data structure for representing the input and output of each of these processing resources, they can be flexibly combined into a pipeline that produces several streams of linguistically annotated structures, which serve as an input for the shallow grammar interpreter, applied at the next stage of the cascade. The grammar formalism in SProUT is a blend of very efficient finite-state techniques and unification-based formalisms which are known to guarantee transparency and expressiveness. The backbone architecture of SProUT is depicted in Fig. 2.

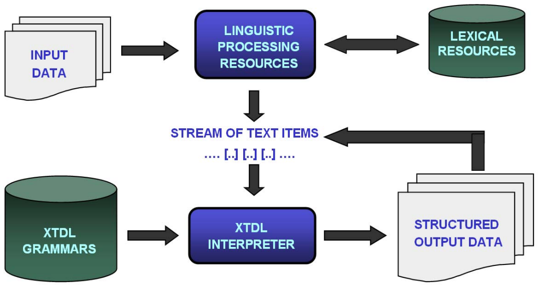


Fig. 2. Overview of the linguistic analysis chain

The analysis system was adapted to account for peculiarities of ASR results, the most important being:

- Linguistically non-well-formed ASR results invalidate many linguistic analysis patterns; only partial analyses are possible.
- Absence of internal segmentation marks, e.g. of sentence boundaries.

4. INFORMATION FUSION FOR CROSS-MEDIA ANALYSIS

4.1. Problem Formulation

In the previous section, several analysis techniques for individually treating a single modality (i.e. visual, audio, or textual information alone) were discussed. The motivation behind employing the above techniques, despite cross-media analysis being the overall objective, lies in the fact that the specific characteristic and needs of each modality (e.g. different decompositions, radically different features, etc.) and the varying kinds of semantic information that is modality can provide (e.g. region-level information for the visual modality, location names for the audio/text ones, etc.) necessitate different specialized analysis methodologies to be employed. However, this makes imperative the subsequent use of techniques that will integrate these results, i.e. remove ambiguities and contradictory outputs and produce a final semantic interpretation of the multimedia content.

Techniques presented in the literature for multimodal fusion, not necessarily for the purpose of knowledge-assisted analysis, include probabilistic approaches [22] and methods that treat information fusion as a structure fusion problem [23], [24]. Among the advantages of techniques belonging to the latter category, such as overlay, is that they rely on structures that represent the semantics of the information to be fused, to identify which portions of information are competing or contradictory. For these portions of information, they subsequently exploit rules for deciding on the prevalence of one source of them over the others; the remaining non-contradictory information coming from multiple sources is simply combined using the union operator. In a knowledge-assisted analysis setting, the knowledge structures that are employed for analysis as described in the previous sections can serve as the structures over which the unification and overlay of information can be realized. Furthermore, such rule-based techniques are computationally inexpensive, which is an important consideration when processing large volumes of data. The only concern regarding their applicability to the cross-media analysis of multimedia content lies on the employed rules for deciding on the prevalence of certain pieces of information over the others, which need to be adapted to the analysis task and the domain at hand. Based on these considerations, we adopt an approach similar in nature to unification and overlay that is adapted to realizing cross-media analysis of multimedia News content and term it “altered overlay”.

4.2. Altered Overlay

The overlay method, as originally defined in [24], is used for the fusion of information on the basis of its temporal priority. The need for adopting an altered version of the overlay method rather than the original one comes from the fact that in our application there is no distinction between more recent and less recent information, as is assumed by the original overlay method; all analysis results are generated by analysis of audio/visual/textual information that is presented to the multimedia consumer at the same time, only in different form. However, what can be used in our case to replace the temporal constraints upon which the original overlay method is based is the set of different degrees of confidence that exist for the different analysis results, as well as the different relevant confidence that we have on each modality, depending on the examined multimedia content. In this direction we proposed a technique in which the information that is not conflicting with other information will be inherited at the unified result and, as far as the conflicting information is concerned, rules and constraints are introduced to assist in deciding which is to be retained and which is to be discarded.

Let us start by considering what kind of semantic metadata we can have from individual modality analysis to build upon, and what are the major categories of multimedia news content where the various modalities carry different weights. The semantic metadata we can have from previous analysis steps can be grouped in 2 main categories:

- Results of classification of the multimedia content to a subdomain of those defined in the knowledge structures (e.g. earthquake, flood, etc.), i.e. degrees of confidence for subdomain-multimedia content association. These are produced both by visual analysis and by linguistic analysis of audio transcripts and any associated textual information.

- Details of an event (e.g. for disaster events, location, number of victims, etc.), that are represented by various concepts and properties in the employed knowledge structures. These can only come from linguistic analysis of either audio transcripts or any associated textual information. Visual analysis can only indirectly support this by extracting textual information from the visual medium (e.g. legends and super-titles on the video frames) by means of OCR techniques.

Consequently, visual, audio and text analysis compete only for the part of analysis related to the classification of the multimedia content to a subdomain of those defined in the domain ontology; for the rest of the analysis results, only audio and text analysis compete.

The multimedia news content, on the other hand, considering audiovisual streams rather than mostly-textual news, can be classified to the following major categories on the basis of the different weights that in each case the various modalities carry:

- Studio shots. Here, linguistic analysis of audio transcripts is expected to be most reliable, due to the controlled environment in which audio information is captured. Visual analysis, on the other hand, is of no usefulness (besides OCR techniques), in the absence of visual information that can give hints on the semantics of the news item presented by the anchorperson. Consequently, in the case of studio shots the altered overlay operator discards the results of visual analysis. In order for the overall cross-media analysis system to decide on whether a given shot falls into this category, a properly trained global image classifier is employed; due to the characteristic and non-varying visual properties of the studio environment, such a classifier can achieve very high correct classification rates.

- External reporting with a dominant face on the video. In this case, linguistic analysis of audio transcripts is again expected to be most reliable, despite audio being captured in a not-fully-controlled environment. Visual information analysis is expected to be of limited use, again due to the nature of visual information: if a face is dominant in the video (covering at least 20% of the image grid), then there is relatively limited background visual information whose analysis could result in a reliable classification of the shot to one of the considered subdomains; even more importantly, though, from a video production point of view, this is a clear indication that the face conveys in this case the visually most important information. Consequently, the altered overlay operator again discards the results of visual analysis, in this case providing that (contradictory or not) results of linguistic analysis of audio transcripts exist. In case linguistic analysis fails to produce any results (e.g. in the absence of any useful cues in the speech part, despite the type of multimedia content), the visual analysis results are retained, since they are generally assumed in this case to be less reliable than audio analysis results but they are not coincidental, as is the case with visual analysis of studio shots. In order for the overall cross-media analysis system to decide on whether a given shot falls into this category, the face detection technique outlined in section 3 is employed; despite the imperfection of any such technique, very high correct face detection rates can be achieved when the face dominance criterion (at least 20% of the image grid) is satisfied.

- External reporting with no dominant face on the video but with speech voiceover. In this case, both semantically important visual information and semantically important audio information is available; there is little justification in considering the one or the other to be more reliable. Consequently, the analysis results of the two modalities are considered to be of equal relevant weight and the degrees of confidence for each individual analysis result are directly comparable; altered overlay calculates the sum of the degrees of confidence for a given content-subdomain association over both modalities, and retains the content-subdomain association for which this sum is maximized. Prior to the sum calculation, the degrees of confidence are modified by application of a histogram-equalization-like function that is modality-specific and is learned from an appropriately large training set, with the objective to make the distribution of degrees of confidence comparable over the different modalities, thus justify the choice of summation for combining them. In order for the overall cross-media analysis system to decide on whether a given shot falls into this category, the face detection and global studio classification techniques used in the previous two cases are employed.

- External reporting with no dominant face on the video and no speech. In this last case, the visual modality clearly carries the most semantics; audio, if any, may include noise or music. Only visual analysis results are generated in this case at the individual modality analysis stage and the altered overlay simply retains them.

In the previous cases, we have not examined the use of textual information that may be extracted by means of OCR from the video (e.g. legends and super-titles) or may accompany the multimedia information in the form of a short manual annotation. In general, OCR techniques are mature techniques that have high success rates. Consequently, the result of linguistic analysis on them is expected to be of high accuracy; the linguistic analysis of manual annotations is expected to be an even more reliable source of semantic information. In our altered overlay scheme, consequently, the results of text analysis (meaning here pure text rather than ASR transcripts), when available, supersede those of any other modality. It should be noted here that the availability of manual annotations depends on the content creator and the overall information usage scenario and, consequently, cannot be taken as granted.

5. EXPERIMENTAL RESULTS

To experimentally evaluate the above techniques in News-related multimedia content, a data set of Deutche Welle (DW) broadcast videos related to different kinds of natural disasters was assembled. A partial view of subdomains and region-level concepts of the domain ontology that are related to the assembled content is shown in Fig. 3.

Following the training of any involved machine learning methods (e.g. visual classifiers) with the use of a training subset of the assembled data set, the proposed knowledge-assisted cross-media analysis architecture was applied to the remaining videos of the data set. Indicative results at various stages of the analysis process are presented in Fig. 4.

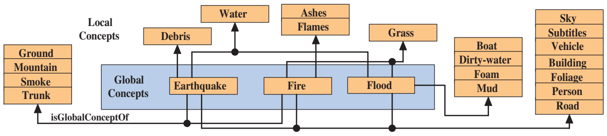


Fig. 3. A partial example of subdomains and region-level concepts of the ontology.

In the second column of Fig. 4, results of visual classification of shots (corresponding keyframes are shown in the first column of the same figure) using global features alone, are presented for various shots belonging to 3 different subdomains (Fire, Flood, Earthquake). Higher values (degrees of confidence) next to each subdomain label indicate better confidence (i.e. for the first shot, “fire” is identified as the most likely subdomain membership for it). Classification results are presented here both for shots whose visual analysis may be able to reveal its semantics and for others which cannot be classified to one of the defined subdomains using visual features alone, i.e. studio shots; results for the latter are coincidental at this stage.

In the third column of Fig. 4, results of visual classification using both global and local features are presented; these are the final results of visual analysis. From these results it is clear that, although coincidental results may change to worse (such as the first studio shot presented in that figure, which was originally classified correctly to “fire” and subsequently was misclassified to “flood”), in general the results have improved; for example, one shot showing debris and clearly belonging to the earthquake subdomain (last row of the figure), that was previously misclassified to “fire” is now correctly classified.

In the fourth column of Fig. 4, the textual transcripts that ASR has produced by analysis of the audio stream and the results of linguistic analysis upon the ASR transcripts are shown. Evidently, linguistic analysis is capable of revealing for some shots not only the type of event (flood, fire etc.) but also additional information such as location names (e.g.“lisbon”) and other data. Rich annotations can be generated this way, providing that the information is included in the ASR transcripts; however, as can be seen in Fig. 4, in several cases no useful audio information is associated with a shot and thus ASR transcript analysis is not able to provide meaningful results.

Combining the results in columns 3 and 4 using the altered overlay technique proposed in the previous section, we arrive to cross-media analysis results; these are presented in the fifth column of Fig. 4. These include rich semantic annotations, providing that the necessary information was included in the multimedia stream, correct any erroneous results of individual modality analysis (e.g. of visual classification), and provide semantic annotations even in the absence of related information in one of the examined modalities (e.g. studio shots providing no characteristic visual cues; shots with no audio information).

6. CONCLUSIONS

In this paper the techniques that we have used for cross-media analysis of News-related multimedia content were presented. Various individual components realizing knowledge-assisted analysis of singe-modality content were outlined and an architecture integrating them in a knowledge-assisted cross-media analysis framework was proposed. An altered overlay technique, used for realizing the fusion of individual modality analysis results was presented in detail, being a key component of the proposed analysis architecture. Experimental results have illustrated the usefulness of the proposed approach. Future work includes the use of additional analysis tools and the refinement of the overall analysis methodology by allowing further interaction between the individual modality analysis techniques at intermediate processing stages.

7. REFERENCES

[1] S.-F. Chang, “The holy grail of content-based media analysis,” *IEEE Multimedia*, vol. 9, no. 2, pp. 6–10, Apr.-Jun. 2002.

[2] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V.K. Papastathis, and M.G. Strintzis, “Knowledge-Assisted Semantic Video Object Detection,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1210–1224, October 2005.

[3] J. Assfalg, M. Berlini, A. Del Bimbo, W. Nunziat, and P. Pala, “Soccer highlights detection and recognition using hmms,” in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, July 2005, pp. 825–828.

[4] L. Zhang, F.Z. Lin, and B. Zhang, “Support vector machine learning for image retrieval,” in *Proc. IEEE Int. Conf. on Image Processing (ICIP01)*, Thessaloniki, Greece, October 2001.

[5] L. Hollink, S. Little, and J. Hunter, “Evaluating the application of semantic inferencing rules to image annotation,” in *Proc. 3rd Int. Conf. on Knowledge Capture (K-CAP05)*, Banff, Canada, October 2005.

[6] G.Th. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, “Ontology-Driven Semantic Video Analysis Using Visual Information Objects,” in *Proc. Int. Conf. on Semantics and Digital Media Technologies (SAMT07)*, Genova, Dec. 2007.

[7] J. Bescos, “Real-time Shot Change Detection over On-line MPEG-2 Video,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 475–484, April 2004.

[8] T. Sikora, “The MPEG-7 Visual standard for content description - an overview,” *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, vol. 11, no. 6, pp. 696–702, June 2001.

[9] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, “Still image segmentation tools for object-based multimedia applications,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 4, pp. 701–725, June 2004.

[10] N. Dallal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR05))*, San Diego, CA, USA, June 2005.

[11] V.N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[12] D.M.J. Tax and R.P.W. Duin, “Using two-class classifiers for multiclass classification,” in *Proc. 16th Int. Conf. on Pattern Recognition (ICPR02)*, Quebec City, Canada, August 2002, pp. 124–127.

[13] M.A.H. Huijbregts, C. Wooters, and R.J.F. Ordelman, “Filtering the Unknown: Speech Activity Detection in Heterogeneous Video Collections,” in *Interspeech 2007*, 2007.

[14] D.A. van Leeuwen and M.A.H. Huijbregts, “The AMI Speaker Diarization system for NIST RT06s meeting data,” in *RT2006, Springer LNCS*, 2007, vol. 4299, pp. 371–384.

[15] F.M.G. de Jong, T. Westerveld, and A.P. de Vries, “Multimedia search without visual analysis: the value of linguistic and contextual information,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 365–371, 2007.

[16] A.F. Smeaton, P. Over, and W. Kraaij, “Evaluation Campaigns and TRECVid,” in *8th ACM Int. Workshop on Multimedia Information Retrieval (MIR06)*, Santa Barbara, CA, 2006.

[17] F.M.G. de Jong, R.J.F. Ordelman, and M.A.H. Huijbregts, “Automated speech and audio analysis for semantic access to multimedia,” in *Int. Conf. on Semantic and Digital Media Technologies, SAMT 2006*, Athens, Greece, December 2006.

[18] J.S. Garofolo, C.G.P. Auzanne, and E.M. Voorhees, “The TREC SDR Track: A Success Story,” in *Eighth Text Retrieval Conference*, Washington, USA, 2000, pp. 107–129.

[19] C. Auzanne, J.S. Garofolo, J.G. Fiscus, and W.M. isher, “Automatic Language Model Adaptation for Spoken Document Retrieval,” in *Proceedings of RIAO 2000*, 2000, pp. 132–141.

[20] M. Becker, W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schafer, and F. Xu, “Sprout- shallow processing with unification and typed feature structures,” in *Proc. Int. Conf. on NLP*, Mumbai, India, 2002.

[21] W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schafer, and F. Xu, “Shallow processing with unification and typed feature structures - foundations and applications,” *Kunstliche Intelligenz*, vol. 2004, no. 1, pp. 17–23.

[22] S. Nakamura, “Statistical Multimodal Integration for Audio - Visual Speech Processing,” *IEEE Trans. on Neural Networks*, vol. 13, no. 4, July 2002.

[23] M. Johnston, P.R. Cohen, D. McGee, S.L. Oviatt, J.A. Pitman, and I. Smith, “Unification based multimodal integration,” in *Proc. of the 35th ACL*, Madrid, Spain, 1997, pp. 281–288.

[24] J. Alexander and T. Becker, “Overlay as the basic operation for discourse processing in the multimodal dialogue system,” in *Workshop Notes of the IJCAI-01 Workshop ok Knowledge and reasoning in practical dialogue Systems*, Seattle, Washington, August 2001.

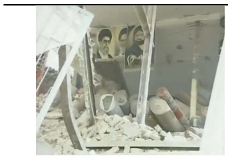
	Earthquake: 0.46 Fire: **0.58** Flood: 0.55	Earthquake: 0.39 Fire: 0.36 Flood: **0.42**	...IN THE HIGHWAY ONE HUNDRED DOLLARS MORE THAN FORTY YEARS AGO THE **LISBON** IS SURROUNDED BY THE PLANE'S FLIGHT GOING TO TRY TO GET THROUGH THAT IN FACT THE HIGHWAY HAS BEEN OFFICIALLY CLOSED WILD**FIRE**S IN FACT...	event type: fire location: lisbon
	Earthquake: 0.30 Fire: **0.87** Flood: 0.23	Earthquake: 0.22 Fire: **0.63** Flood: 0.20	...IN THE HIGHWAY ONE HUNDRED DOLLARS MORE THAN FORTY YEARS AGO THE **LISBON** IS SURROUNDED BY THE PLANE'S FLIGHT GOING TO TRY TO GET THROUGH THAT IN FACT THE HIGHWAY HAS BEEN OFFICIALLY CLOSED WILD**FIRE**S IN FACT...	event type: fire location: lisbon
	Earthquake: 0.49 Fire: **0.69** Flood: 0.31	Earthquake: 0.32 Fire: **0.55** Flood: 0.25	BUT WHILE	event type: fire
	Earthquake: 0.41 Fire: **0.60** Flood: 0.48	Earthquake: 0.32 Fire: 0.38 Flood: **0.45**	...AND WHAT'S GOING TO DO SOMETHING ABOUT THE EVACUATION OF WOUNDED INSTRUCTIONS TO LEAVE THEIR HOMES SOME THIRTY THOUSAND PEOPLE NOW LIVING IN EMERGENCY SHELTERS BY THE WAY THE **FLOOD**WATERS TO RECEDE...	event type: flood
	Earthquake: **0.52** Fire: 0.47 Flood: 0.50	Earthquake: 0.41 Fire: 0.29 Flood: **0.46**	...THAT DOESN'T APPLY TO THE WAR IS HEATING UP AT THE TIME IS RIGHT THERE WITH THE KEY THING THAT IS THERE ANY CONFIDENCE IN THE LAST TIME BACK IN PARTICULAR AREAS DESPITE THE EFFORTS THE VILLAGES HAVE BEEN EVACUATED BUT ON THE OTHER SIDE OF THE GROUP THAT HAS BEEN **FLOOD**ED WITH SEVERAL DAYS...	event type: flood
	Earthquake: 0.44 Fire: 0.49 Flood: **0.54**	Earthquake: 0.34 Fire: 0.32 Flood: **0.49**	-	event type: flood
	Earthquake: 0.34 Fire: **0.69** Flood: 0.43	Earthquake: 0.33 Fire: **0.43** Flood: 0.41	...BUT WHILE THE **BERLIN** INTERNATIONAL RESCUE WORKERS SAY THEY HAVE LITTLE HOPE OF FINDING ANYONE ALIVE AND WELL ALL OF THE **MASSIVE** **EARTHQUAKE** LIKE THE ONE THING THAT YOU KNOW MORE THAN ONE THOUSAND PEOPLE WERE BELIEVED TO HAVE GOTTEN SIX POINTS BACK TO WORK...	event type: earthquake degree: massive location: berlin
	Earthquake: **0.74** Fire: 0.49 Flood: 0.26	Earthquake: **0.65** Fire: 0.29 Flood: 0.24	...THERE ARE GOING TO GO TO THE POINT THAT YOU WERE THINKING OF DOING ANYONE ELSE ON THE ROAD WAS BUILDING WAS BUILT FROM THE BRINK OF THE POLICE TO DO IT IN THE MIDDLE CLASS AND WHAT WE'RE GOING TO DO THAT WITH PEOPLE THAT...	event type: earthquake
	Earthquake: 0.57 Fire: **0.58** Flood: 0.34	Earthquake: **0.42** Fire: 0.40 Flood: 0.31	HOW IMPORTANT ARE THESE BUILDINGS AND WHAT IT'S AN AMERICAN IN EAST JERUSALEM AND THEN	event type: earthquake

Fig. 4. Indicative analysis results: a keyframe of each shot (column 1), results of visual classification using global features alone (column 2), results of final visual classification using global and local features (column 3), results of audio analysis (ASR transcripts) and in bold & underlined the terms that linguistic analysis has found in the transcripts to correspond to distinct classes and properties of the knowledge structures, such as event types and locations (column 4), final knowledge-assisted cross-media analysis results (column 5).