A system for the semantic multi-modal analysis of news audio-visual content

Vasileios Mezaris, Spyros Gidaros, Georgios Th. Papadopoulos, Walter Kasper, Jörg Steffen, Roeland Ordelman, Marijn Huijbregts, Franciska de Jong, Ioannis Kompatsiaris, Michael G. Strintzis

***Abstract*—News related content is nowadays among the most popular types of content for users in everyday applications. Although the generation and distribution of news content has become commonplace, due to the availability of inexpensive media capturing devices and the development of media sharing services targeting both professional and user-generated news content, the automatic analysis and annotation that is required for supporting intelligent search and delivery of this content remains an open issue. In this paper, a complete architecture for knowledge-assisted multi-modal 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 novel 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 annotations.**

I. Introduction

Access to news related multimedia content, either amateur or professional, is nowadays a key element in business environments as well as everyday practice for individuals. The proliferation of broadband internet and the development of media sharing services over the World Wide Web have contributed to the shifting of traditional news content creators, such as news agencies and broadcasters, towards digital news manipulation and delivery schemes. At the same time, the availability of inexpensive media capturing devices has additionally triggered the creation and distribution of vast amounts of user-generated news audio-visual content, giving rise to citizen journalism. Several distribution channels, from generic ones (e.g. YouTube[1]) to dedicated citizen journalism services (e.g. YouReporter[2]) have been developed in the last few years as part of this evolution of the news distribution environment. Although the generation and distribution of news content has become commonplace, the automatic analysis and annotation that is required for supporting intelligent search and delivery of this content remains an open issue. In general, the cornerstone of the efficient manipulation of any type of multimedia material is the understanding of the semantics of it [1]; news related audio-visual content is no exception to this rule.

In response to the need for understanding the semantics of multimedia content in general, knowledge-assisted analysis has recently emerged as a promising category of techniques [2]. Knowledge-assisted analysis 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). Machine learning techniques are often used as part of knowledge-assisted analysis architectures, being suitable 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]. Other 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].

In this work, an architecture for the knowledge-assisted multi-modal analysis of news-related multimedia content is proposed. This initially employs state-of-the-art methods for the analysis of each individual modality (visual, audio, text) separately. Subsequently, a fusion technique that does not require training with the use of a manually annotated dataset is introduced for combining the individual modality analysis

V. Mezaris, S. Gidaros and I. Kompatsiaris are with the Informatics and Telematics Institute / Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, P.O.BOX 60361, Thermi 57001, Greece, {bmezaris, gidaros, ikom}@iti.gr.

G. Th. Papadopoulos and M. G. Strintzis are with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece, and also with the Informatics and Telematics Institute / Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, P.O.BOX 60361, Thermi 57001, Greece, {papad, michael}@iti.gr.

W. Kasper and J. Steffen are with DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrucken, Germany, {kasper,steffen}@dfki.de

R.Ordelman and F. de Jong are with the University of Twente, 7500 AE Enschede, The Netherlands, {ordelman, f.m.g.dejong}@ewi.utwente.nl.

M. Huijbregts is with the Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands, marijn.huijbregts@let.ru.nl; He was previously with the University of Twente, 7500 AE Enschede, The Netherlands.

This work was supported by the European Commission under contract FP6-027685 MESH. M. Huijbregts' work was partly carried out in the context of the project CHoral, which was funded under the NWO programme CATCH.

[1] http://www.youtube.com/

[2] http://www.youreporter.it/

results. This technique takes into account knowledge encoded in an appropriate ontology infrastructure, and its main novelty lies in that it explicitly takes into account the potential variability of the different uni-modal analysis techniques in terms of the decomposition of the audio-visual stream that they adopt, the fuzzy degrees of content-concept association that they produce, the concepts of the overall large-scale ontology that they consider, the varying semantic importance of each modality, and other factors.

The paper is organized as follows: related work on news multi-modal analysis is reviewed in section II. In section III the analysis problem that this work attempts to address is formulated and the overall architecture of the proposed approach is presented. The knowledge representation and the different uni-modal analysis techniques that are part of this architecture are outlined in sections IV and V, while the technique developed for combining the individual modality analysis results is presented in detail in section VI. Section VII reports on the experimental evaluation and comparison of the developed techniques, and conclusions are drawn in section VIII.

II. Related Work

Knowledge-assisted semantic multimedia analysis techniques can be classified, on the basis of the information that they exploit for analysis, to uni-modal and multi-modal ones. Uni-modal techniques exploit information that comes from a single modality of the content, e.g. they exploit only visual features for classification [6]. Multi-modal techniques, on the other hand, exploit information from multiple content modalities in an attempt to overcome the limitations and drawbacks of uni-modal ones. Applications of multi-modal techniques range from semantic multimedia analysis to audio-visual speech recognition [7], discourse processing in dialogue systems [8], and video retrieval [9].

In general, the multi-modal techniques can be broadly classified to those jointly processing low-level features that come from different modalities [10] [11], and those that combine the results of multiple uni-modal analysis techniques [12] [13]. Rank aggregation and other methods used primarily in retrieval applications to combine ranked lists of retrieval results [14] [15], can also be classified to the latter category. While it can be argued that each one of the two aforementioned classes of multi-modal techniques has its advantages and thus can be more or less suitable than the other for a given application, it is generally observed that techniques of the latter class are more suitable when a "deep" analysis of each modality is required (e.g. speech recognition and linguistic analysis of the transcripts, rather than mere classification of audio segments to a limited number of classes).

Regarding news content analysis in particular, there has been a number of approaches presented in the last few years. In some of them, the emphasis is on textual transcript processing; other modalities such as the visual one have limited contribution. For example, in [16], news video is segmented into shots and scenes using visual and audio analysis techniques; the semantic categorization of each resulting news segment is performed using only the results of natural language processing techniques on OCR-generated transcripts. In [17], the emphasis is again mostly on textual information processing, and the results of it together with limited visual analysis results (detected captions, faces, etc.) are fused for the purpose of visualization of large-scale news video collections, with the objective of facilitating browsing the collection and retrieving video clips. However, recent advances in visual information analysis and classification have made possible the extraction of rich semantic information from the visual modality as well; this should be exploited.

The number of supported classes of news content is another important factor when examining different news content analysis approaches. In [18], a two-layer classification scheme is introduced, where the second-layer classifier fuses the output of the individual first-layer classifiers, for building detectors for just two classes: anchor and commercial. In [11] the problem of fusing the results of different classifiers to eventually classify each news video segment to one of 5 classes (politics, society, health, sports, and finance) is treated as a Bayesian risk minimization problem. In [19], 10 news categories (i.e. Politics, Military, Sport, etc.) are defined, detectors are designed for processing textual and audio-visual information separately based on SVMs and GMMs, and a fusion strategy is used for deciding on the category membership of each news story. Although such methods highlight important aspects of news multi-modal analysis, the limited number of classes that they consider means that they either solve a very constrained problem (such as anchor or commercial detection) or that they result in a very broad classification of news content (i.e. to 5-10 classes). Acknowledging the need to consider a larger number of classes as well as multiple modalities, in [20] multimodal fusion is formulated as an optimization problem and generic methods for optimizing linear and non-linear combinations of modalities are discussed; again, however, testing of the developed techniques is reported on a rather limited number of classes.

Finally, the type of considered news content and the exact application that multi-modal fusion techniques support may vary among the relevant literature approaches. In [21], a generic approach to fusion is also proposed based on the use of conceptual graphs; however, the focus is on fusing TV program metadata such as program title and date, rather than semantic information coming from the analysis of the audio, visual etc. modalities. As a consequence, the developed formulation cannot handle uncertain input, e.g. the fuzzy degrees of content-concept association that individual modality analysis techniques such as visual classifiers typically produce. This technique has been used as part of a recommendation system for smart television [12]. In [22], the problem of consolidating information coming from various textual news sources on the Web is considered. The developed method can handle uncertain input (confidence levels for each individual analysis result), but employs simple majority voting for combining the results coming from the different news sources, rather than taking into account that the reliability of each source may differ. In [14], the problem of multi-modal fusion for retrieval is addressed and methods such as Borda Count and Borda Fuse

for combining ranked lists of retrieval results are discussed; however, these methods do not consider issues that are specific to multi-modal fusion for analysis, such as the existence of a different content decomposition for each modality.

III. Proposed Approach

A. Problem Formulation

The objective of analysis in this study is to associate each elementary temporal segment (e.g. video shot) of the audiovisual stream with one or more semantic concepts. Let us start by defining an ontology $\mathcal{O}$ that includes the set of concepts that are of interest to a given application domain and their hierarchy:

$$\mathcal{O} = \{C, \ \leq_C\} \tag{1}$$

where $C = \{c_k\}_{k=1}^{K}$ is the set of concepts and $\leq_C$ is a partial order on C called concept hierarchy or taxonomy. $C' \subset C$ is the set of top-level concepts of the ontology, i.e. the sibling concepts that define the coarsest possible classification of content according to $\mathcal{O}$. In any practical application, the employed ontology will normally include additional elements such as properties, concept relations in addition to those specifying the hierarchy, etc., as discussed in the following section. However, the above simplified ontology definition is sufficient at this point.

Let us assume that I individual modality analysis tools exist. These tools may include visual video classification, linguistic analysis of speech transcripts, audio event detection etc. Each of these tools defines a decomposition D_i of a multimedia content item (i.e. creates an ordered set of temporal segments) and, considering all concepts of C or a subset of them, associates each segment of D_i with one or more concepts by estimating the corresponding "degrees of confidence". The values of the latter may be either binary $\{0, 1\}$ or (following normalization, if necessary) real in the range $[0, 1]$. Thus, application of the I aforementioned analysis tools to a multimedia content item will result to the definition of a set of content temporal decompositions

$$D = \{D_i\}_{i=1}^{I} \tag{2}$$

In the general case, each decomposition D_i is a different set of temporal segments, since modality-specific criteria are typically used for determining the latter; e.g. a meaningful elementary visual decomposition of video would probably be based on the results of visual shot change detection, while for automatic speech recognition (ASR) transcripts it would probably be based on audio classification or speaker diarization results instead. All the decompositions together define a temporal segment set S,

$$S = \{s_j\}_{j=1}^{J} \tag{3}$$

It is useful to observe that S, which contains all segments in D, is a set of temporal segments with no hierarchy, many of which may temporally overlap in full or in part (an example of this can be seen in Fig. 7). Each member of set S can be defined as a vector

$$s_j = [t_j^A, \ t_j^B, \ \{d_j(c_k)\}_{k=1}^{K}] \tag{4}$$

where t_j^A, t_j^B are the start- and end-time of the temporal segment and $d_j(c_k) \in [0, 1]$ is the degree with which the individual modality analysis tool that defined s_j associated it with concept c_k of the ontology after analysis of the relevant uni-modal information. In many cases, s_j would be expected to be a sparse vector (since $d_j(.)$ would normally be zero for the majority of concepts of the ontology) and therefore in practice may be represented more efficiently as a variable-length vector that includes only the non-zero values of $d_j(.)$, but the former representation is used in the sequel for notational simplicity.

The multi-modal analysis problem addressed in this work is, given the above set S of heterogeneous individual modality analysis results and the ontology $\mathcal{O}$, and using one of the decompositions of set D as a reference decomposition, to decide what is the most plausible annotation (or the ordered list of N most plausible annotations) for each temporal segment of the reference decomposition. It should be clarified that the term "reference decomposition" is used for denoting the decomposition that is used for associating the final multi-modal analysis results with the content; its selection is made by the user according to the specific user/application needs. For example, if a retrieval application requires the content to be indexed at the level of visual shots, this is the decomposition that should be used as reference decomposition during analysis, to ensure that multi-modal analysis results are indeed associated with every individual visual shot; if, on the contrary, indexing and retrieval e.g. at the speaker level (i.e. according to different speakers) is required, the corresponding decomposition should be used as the reference one during analysis. Evidently, the multi-modal analysis process can be repeated using each time a different reference decomposition, to allow for the multi-modal annotation for segments belonging to different decompositions (e.g. both visual shots and speaker segments), if this is required.

B. System Overview

An overview of the approach proposed in this work for addressing the multi-modal analysis problem discussed above is shown in Fig. 1. As can be seen in this figure, starting from the audiovisual content on the far left, different techniques for analyzing separately each individual modality (visual, audio, text) are executed in parallel, resulting in an extended set of uni-modal analysis results. These are represented with the use of a domain ontology and a multimedia ontology, that account for the domain knowledge (e.g. concepts) and the low-level properties of the content (e.g. decompositions), respectively. The independent processing of each modality allows the use of modality-specific techniques and criteria for identifying elementary temporal segments (e.g. visual shots, audio segments, etc.) and for estimating degrees of confidence for the association of each such temporal segment with the different possible concepts. Following the generation of the uni-modal analysis results, different possible associations between them (such as the overlapping of temporal segments, the relation of different concept annotations according to the concept hierarchy, etc.) are evaluated with the use of specific functions, and all these are combined in a two-stage process

for identifying the most plausible concept annotations for any given temporal segment: at the first stage, the overall influence of the various decompositions and the different concepts on the association of the given segment s_j (of the reference decomposition) with a top-level domain concept $c_k \in C'$ is evaluated. At the second stage, the above top-level concept annotation decision is propagated to the more specific (i.e. less abstract) concepts of C, to result in the selection of the most plausible specific concept annotation of s_j.

IV. Knowledge Representation

In a knowledge-assisted multimedia analysis system, such as the proposed one, knowledge representation serves two main purposes: the representation of prior knowledge for the domain, and the representation of the analysis 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[3] tree for the news domain and includes a hierarchy or classes that range from rather abstract ones, such as "disaster and accident" (i.e. the top-level concepts belonging to C'), to specific ones, such as "earthquake", "flood", etc. (Fig. 2). The latter classes are the least abstract ones with which an elementary news item can be associated. 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. region-level concepts $V = \{v_z\}_{z=1}^{Z}$ describing possible spatial regions of an image rather than entire images, is also defined. Examples of such region-level concepts include person, building, road, sky, flames, water, foliage, mountain, etc. Contextual information X in the form of concept frequency of appearance is also included in this ontology, extending the ontology definition of Eq. (1) as follows:

$$\mathcal{O} = \{C,\ \leq_C,\ V,\ X\} \tag{5}$$

The multimedia ontology, on the other hand, is a knowledge structure used for supporting the storage of information and of analysis results about the content (e.g. its different decompositions). Its development represents a choice concerning the practical implementation of the proposed system rather than the algorithmic aspects of it and therefore this ontology does not need to be discussed here; the interested reader is referred to [23] for a detailed presentation.

V. Single Modality Analysis Techniques

A. Visual Classification

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, such as global keyframe and region level classification and the fusion of these classification results to a single hypothesis set about the concept membership of each shot of the examined news item (Fig. 3).

Preprocessing starts with temporal video decomposition to shots, which are the elementary video streams that can be associated with one concept of the employed ontology. For shot segmentation the algorithm of [24] is employed, which works directly with frame histogram metrics computed over low resolution images extracted from the compressed video stream. Subsequently, a keyframe is identified for each shot and a rich set of MPEG-7 visual descriptors [25] is extracted for it, both at the global image level (Scalable Color, Homogeneous Texture and Edge Histogram descriptors) and at the region level (Scalable Color, Homogeneous Texture and Region Shape), following spatial segmentation to homogeneous regions using the method of [26]. As a final pre-processing stage, face detection is performed using a variant of the method of [27]; 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 preprocessing, 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 concepts 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) [28] structure is utilized to compute the class to which each piece of visual information belongs. This comprises L SVMs, one for every selected concept. It must be noted that the set of concepts for which visual classifiers are trained is typically a subset of $C - C'$, due to lack of sufficient training data for all concepts in $C - C'$ and also the fact that many of these concepts have no clear visual manifestation that would make the training of visual classifiers possible (e.g. concept "liberation"). Each SVM is trained under the "one-against-all" approach, using an appropriate training set of images that were manually classified to concepts. At the evaluation stage, each SVM returns for every image of unknown concept membership a numerical value in the range [0, 1]. This value denotes the degree of confidence with which the corresponding visual content is assigned to the concept represented by the particular SVM, and is computed from the signed distance of it from the corresponding SVM's separating hyperplane using a sigmoid function [29]. 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, $l = 1, \ldots, L$, constitute its concept hypothesis set.

Region-level classification follows, using a similar SVM structure to compute an initial region-concept association for every spatial region of the keyframe. As in the previous case, an individual SVM is introduced for every region-level concept v_z of the employed ontology, in order to detect the corresponding association. For training the SVMs, an appropriate training set (made of regions generated by automatic segmentation and

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

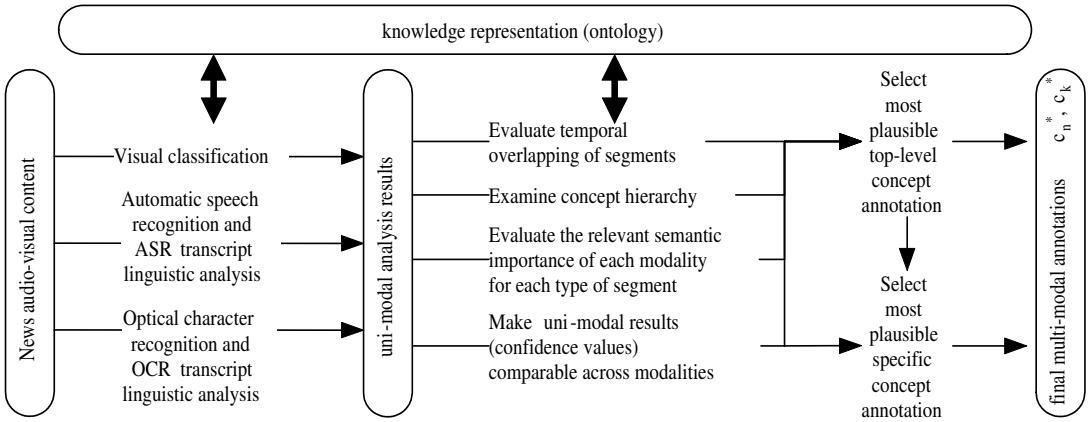

Fig. 1. Overview of the proposed approach for multi-modal analysis of news audio-visual content.

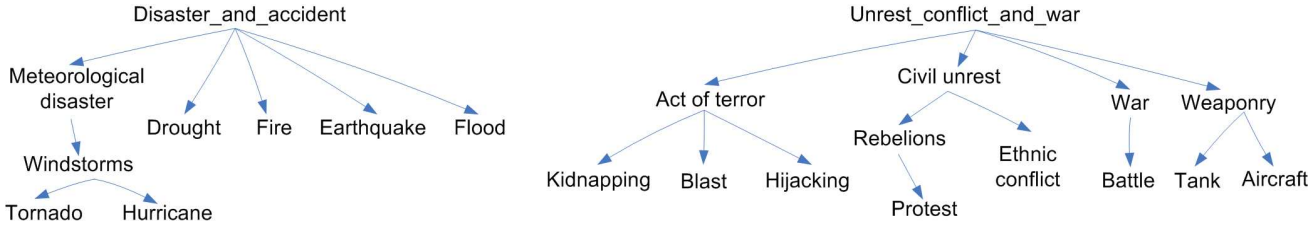

Fig. 2. Subset of concepts and their hierarchy in the employed ontology for news. Two of the 17 top-level concepts ("Disaster and accident", "Unrest, conflict and war") and a few of their sub-concepts are depicted.

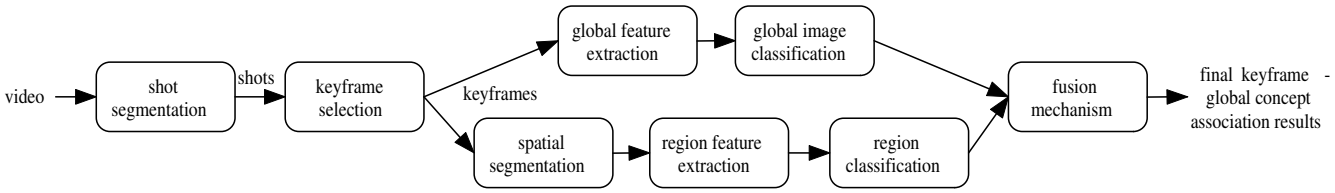

Fig. 3. Overview of the visual classification process.

manually assigned to region-level concepts) is employed. As a result, at the evaluation stage a degree of confidence is returned for each region r of unknown concept membership and each region-level concept v_z in the domain ontology. These results for all regions of the keyframe are subsequently employed for inferring a new keyframe-concept association hypothesis set H_l', as in [6].

Finally, a fusion mechanism in the form of a weighted summation $G_l = \lambda_l \cdot H_l + (1 - \lambda_l) \cdot H_l'$ is introduced for deciding upon the final keyframe - global concept association. The concept for which G_l is maximized is the most plausible annotation of the respective video shot based on visual information, while G_l, $l = 1, \ldots, L$, is the final visual classification hypothesis set. For optimizing the weights λ for each concept, a genetic algorithm is used, to account for the varying relevant importance of global and local information for the detection of different concepts [23]. Indicative examples of intermediate and final visual classification results are shown in Fig. 4.

B. Visual Analysis for Text Extraction

Besides the association of video shots with semantic classes (concepts) on the basis of the visual features of the corresponding keyframes, visual information, i.e. the keyframes, can also be used for extracting the text that is in some cases superimposed to them. In news content, this text typically encompasses in a very compact way semantic information such as person names or event summaries, some of which can be useful for analysis. To this end, text transcripts are generated by application of software developed on top of a commercial Optical Character Recognition (OCR) software development kit [4] to keyframes of the video. All keyframes extracted as discussed in the previous section are processed; the work flow of this processing involves (a) text regions detection on the keyframe, and (b) Optical Character Recognition, as depicted in Fig. 5(a). Both these tasks are perfomed using functions of the employed commercial software development kit. The resulting text transcripts subsequently undergo linguistic anal-

[4] ABBYY FineReader Engine 8.1

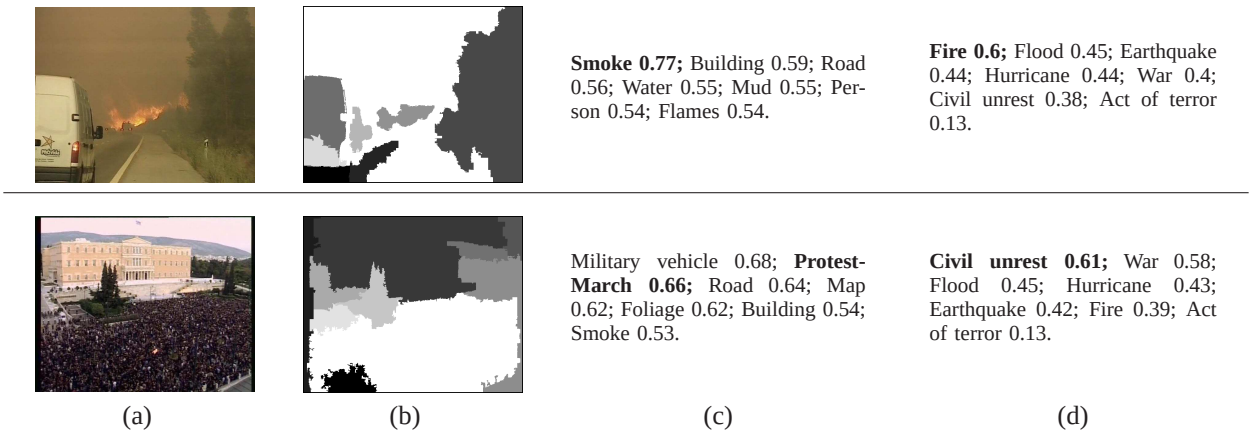

Fig. 4. Visual classification examples: (a) keyframe, (b) segmentation mask, (c) results of region classification for the spatial region shown in white in the mask (only a few region-level concepts, in descending order according to the estimated degree of confidence, are shown), (d) final keyframe classification results (in descending order according to the estimated degree of confidence), generated by combining the region-level classification results for all regions and the results of global classifiers. The concepts that are in agreement with the ground truth annotation are shown in bold. Taking into account all region-level classification results rather than the single highest-ranking region-level concept for every region, when estimating the final keyframe classification results, is motivated by the known imperfection of region classifiers (as seen in the second example).

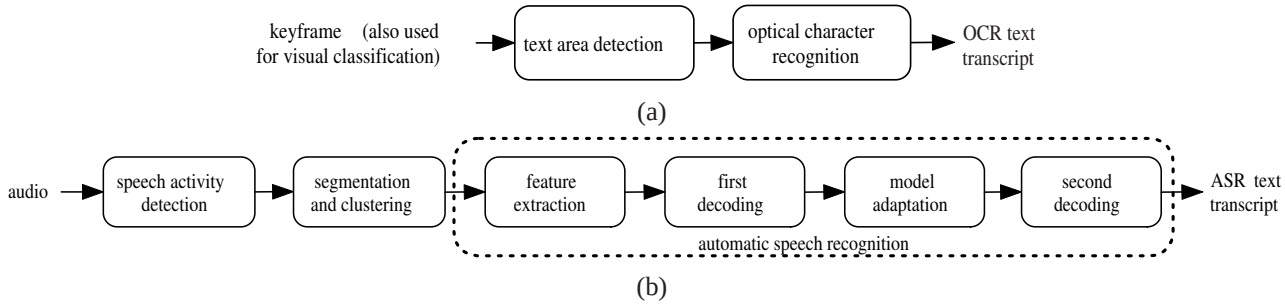

Fig. 5. Overview of (a) visual analysis for text extraction, (b) audio analysis. Both result in the generation of text transcripts.

ysis as discussed in section V-D.

C. Audio Analysis

The use of speech technology to exploit the linguistic content that is available as spoken content in videos has proven to be helpful in bridging the semantic gap between low-level media features and conceptual information needs [30] and its use has been advocated since many years. In this work, the SHoUT large vocabulary speech recognition system is used to this end.

The work flow of the system is depicted in Fig. 5(b). Processing of an audio file starts with speech activity detection (SAD) in order to filter out the audio parts that do not contain speech [31]. After SAD, speaker diarization is performed: the speech fragments are split into segments that only contain speech from one single speaker with constant audio conditions and each segment is labeled with a speaker ID following speaker clustering [32]. Subsequently, automatic speech recognition (ASR) is performed in four steps. First, features are extracted from the segmented audio and are normalized for speaker and audio variations. Next, a primary decoding pass is run. The output of this pass is used for adapting the acoustic model for each speaker cluster. Finally, the secondary decoding pass uses the adapted models for producing the final speech transcripts. For ASR decoding, a time synchronous Viterbi search is used, implemented using the token passing paradigm [33]. HMMs with three states and GMMs for its probability density functions are used to calculate acoustic likelihoods of context dependent phones. The employed decoder is described in more detail in [34].

Output of the audio analysis process is a temporal decomposition of the audio stream to speaker segments, and a textual transcript for each such segment.

D. 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 (section V-B); and ASR transcripts produced by audio 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 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, the SProUT platform (*Shallow Processing with Unification and Typed Feature Structures*) is used as core annotation and information extraction engine. SProUT combines finite state techniques with unification of *typed feature structures* (TFS) [35]. The TFS framework provides a powerful device for representing and propagating information. Rules are expressed by regular expressions over input TFSs that get instantiated by the analysis. The reader is referred to [36], [37] for more details on SProUT.

Output of linguistic analysis, regardless of the source of the input, is a set of content-concept associations using the concepts of set C of the employed ontology (section IV), and additional information in the form of locations, person names and other attributes. Linguistic analysis is applied separately to the information coming from each of the possible input sources (i.e. ASR, OCR, etc.), not only because of differences in the content decompositions and in the way that linguistic analysis needs to process the different inputs, but also because the output of linguistic analysis for each information source needs to be treated differently when combining the individual modality analysis results, as discussed in the following section. Indicative linguistic analysis results for ASR and OCR transcripts are shown in Fig. 6.

VI. Generalized Concept Overlay for Multi-Modal Analysis

After having processed the individual modalities separately, the objective is to combine their results, i.e. to remove ambiguities and contradictory outputs and produce a final semantic interpretation of the multimedia content. A simple, yet crude solution to the combination of individual modality analysis results without using a manually annotated dataset for training would be to disregard the concept hierarchy $\leq_C$ of the ontology, identify all segments of S that temporally overlap in full or in part with the examined temporal segment s_j of the reference decomposition D_i, aggregate the corresponding degrees $d_j(.)$ and select as most plausible annotation the concept c_k for which $d_j(c_k)$ is maximized. This simple approach, however, presents several important drawbacks. Firstly, ignoring the concept hierarchy means that we choose not to consider the semantic similarity or dissimilarity of the different possible annotations; consequently, all possible annotations are treated as contradictory, although this may not be the case (e.g. one may simply be a sub-concept of the other). Secondly, we treat the temporal overlapping of the segments of S as a binary variable, whereas the degree of this overlapping could in fact be useful for determining the significance of an annotation coming from segment s_m for the analysis of the reference temporal segment s_j. Thirdly, we ignore the fact that the semantic importance of all modalities is not necessarily equal and may even vary with respect to the type of content; in news video semantic analysis, for example, the visual and audio modalities carry different weights when examining a studio shot and when examining an external reporting shot. Finally, we overlook that values $d_j(.)$ generated by different analysis tools are not directly comparable in the general case.

To alleviate the identified drawbacks of the aforementioned simplistic approach, we propose a method that is somewhat related to the overlay technique, proposed in [8] for the fusion of structured information on the basis of its temporal priority. In our approach however the decision criterion cannot be the temporal priority of concept detection, since the multimedia content is decomposed to segments (elementary temporal units) instead of being treated as a single item whose annotation may evolve in time. The order of execution of the different uni-modal analysis techniques is clearly not relevant. Instead, the aforementioned considerations about the temporal overlapping of segments, semantic importance of the modalities, etc. have to be taken into account.

Starting with the quantification of the temporal overlapping of the segments of S, we define function $\tau : S^2 \rightarrow [0, 1]$ such that

$$\tau(s_j, s_m) = \begin{cases} \frac{\min(t_j^B, t_m^B) - \max(t_j^A, t_m^A)}{t_j^B - t_j^A} & \text{if } \Gamma > 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where s_j is the reference segment and

$$\Gamma = (t_j^B - t_m^A)(t_m^B - t_j^A) \quad (7)$$

The meaning of function τ is illustrated with an example in Fig. 7.

In order to take advantage of the concept hierarchy, we define function $\phi : C^2 \rightarrow [0, 1]$ such that

$$\phi(c_k, c_n) = \begin{cases} 1, & \text{if } c_n{=}c_k \text{ or } c_n \text{ is a sub-concept of } c_k \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Note that $\leq_C$ is used for evaluating if one concept is a sub-concept of another and that, by definition, sub-concepts are not limited to immediate children of c_k.

In order to take into account the varying semantic importance of the different modalities with respect to the type of content, we define a domain-specific partitioning W of the reference decomposition D_i to a set of disjoint types of segments,

$$W = \{W_q\}_{q=1}^{Q} \quad (9)$$

In the experiments reported in this work, the decomposition of the visual modality to shots was used as the reference decomposition, and three content types (W1: Studio shots; W2: External reporting with a dominant face on the video; W3: External reporting with no dominant face on the video) were defined. Partitioning W is used for defining $\mu : (W, D) \rightarrow [0, 1]$, a domain-specific function such that $\mu(s_j, s_m)$, where $s_j \in W_q$ and $s_m \in D_i$, indicates the relevant semantic importance of the modality corresponding to decomposition D_i for the analysis of segments of type W_q. An example of function $\mu(s_j, s_m)$ defined for News video is illustrated in Fig. 8.

Finally, in order to account for values $d_j(.)$ generated by different analysis tools not being directly comparable, we define a set of tool- and domain-specific functions $\xi_i,\ i = 1, \ldots, I$,

AS NIGHT FELL OVER BAGHDAD ON MONDAY COALITION WARPLANES CARRIED OUT A NEW WAVE OF AIR ATTACKS	War 1.0	Location: Bagdad, Iraq; Day: Monday
FIRES IN PORTUGAL	Fire 1.0	Location: Portugal
(a)	(b)	(c)

Fig. 6. Linguistic analysis examples: (a) text transcripts (the first one coming from ASR and the second from OCR), (b) content-concept associations using the concepts of set C, (c) additional information in the form of locations etc.

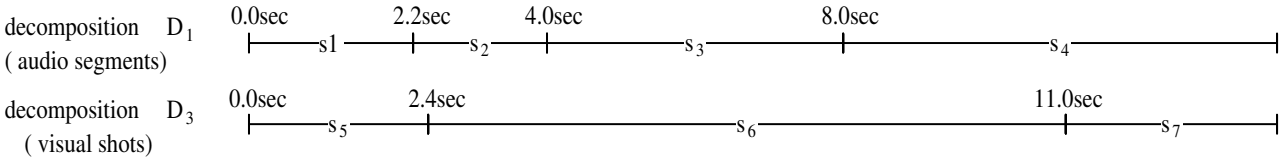

Fig. 7. An example illustrating the use of function τ. For the depicted decompositions, $\tau(s_3, s_6) = \frac{8-4}{8-4} = 1$, i.e. in performing multi-modal annotation of s_3, the visual analysis results of s_6 would be taken into account with a temporal weight of 1 (since the only visual shot temporally overlapping with s_3 is s_6). On the contrary, $\tau(s_6, s_3) = \frac{8-4}{11-2.4} = 0.47 < 1$, since s_3 is not the only audio segment temporally overlapping with s_6. Thus, in performing multi-modal annotation of s_6, the audio analysis results of s_3 would be taken into account with a temporal weight of 0.47 and using this weight they would be combined (or would compete) with audio analysis results coming from s_2 and s_4 that also temporally overlap with s_6; the sum of temporal weights for s_2, s_3 and s_4 would be equal to 1.

Partitioning W of reference decomposition D_3.			
Decomposition D_1 (ASR)	a_3	a_1	a_3
Decomposition D_2 (OCR)	1	1	1
Decomposition D_3 (Visual Classification)	0	a_2	a_3
	(W_1)	(W_2)	(W_3)

Fig. 8. Example of function $\mu(s_j, s_m)$ defined for News video, where $0 < a_2 < a_1 < 1$ and $0 < a_3 < 1$, indicating the relevant semantic importance of the modality corresponding to decomposition D_i for the analysis of segments of type W_q. According to this example, when performing the multi-modal analysis of a studio shot (column W_1), visual classification results are not taken into account, while ASR linguistic analysis results have lower importance that OCR linguistic analysis results; similar knowledge is encoded for shots of types W_2 and W_3, as discussed in more detail in the experimental evaluation section.

one for each modality, that attempt to make values $\xi(d_j(.))$ comparable across modalities. This can be done by enforcing them to have common statistics (e.g. the same mean value, or the same distribution such as a uniform one) over a reasonably large dataset. It must be noted that in this process no ground truth annotation is required for the employed dataset. In the sequel, the index to ξ will be omitted for notational simplicity; the use of function ξ that corresponds to the tool which generated its argument value $d_j(.)$ will be implied.

Using the above definitions, a two-stage process can be defined for combining all the individual modality analysis results. At the first stage, the overall influence of the various decompositions and the different concepts $c_n \in C$ on the association of a segment s_j (of the reference decomposition) with a top-level domain concept $c_k \in C'$ is defined as follows:

$$\psi(s_j, c_k) = \sum_{n=1}^{K} \left[\phi(c_k, c_n) \cdot \left(\sum_{m=1}^{J} \tau(s_j, s_m) \cdot \mu(s_j, s_m) \cdot \xi(d_m(c_n)) \right) \right] \quad (10)$$

Then,

$$k^* = \arg\max_k \left(\psi(s_j, c_k) \right) \quad (11)$$

indicates the single most plausible top-level concept annotation c_{k^*} of segment s_j. In case the application under consideration allows for more than one top-level concept to be assigned to a single segment, several strategies for retaining the x most plausible top-level concepts by examining the values of $\psi(s_j, c_k)$ for all k can be defined, according to the specific application needs.

At the second stage, in order to generate a more specific annotation of segment s_j, the above top-level concept annotation decision has to be propagated to the more specific (i.e. less abstract) concepts of C. This is performed by evaluating which sub-concept of c_{k^*} contributed the most to its selection in the previous processing step (similarly to Eq. (8), not being limited to immediate children of c_{k^*}). In particular, for every c_n that does not belong to C' and for which $\phi(c_{k^*}, c_n) = 1$ the following value is calculated:

$$\rho(s_j, c_n) = \sum_{m=1}^{J} \tau(s_j, s_m) \cdot \mu(s_j, s_m) \cdot \xi(d_m(c_n)) \quad (12)$$

Then,

$$n^* = \arg\max_n \left(\rho(s_j, c_n)\right) \quad (13)$$

indicates the single most plausible specific concept annotation c_{n^*} of segment s_j. Again, more than one such concepts could also be assigned to s_j by examining the values of $\rho(s_j, c_n)$, if desired.

A couple of examples of the above two-stage process for assigning concept annotations to a visual shot are shown in Fig. 9. For the first one (top row of the figure), the shot's actual subject is "war in Iraq" and the keyframe is shown on the left side of the figure. The degrees of confidence with which a concept is associated with this shot on the basis of visual and audio information (taking into account all audio segments that temporally overlap in full or in part with the shot) are shown next to each concept in parenthesis and in brackets, respectively. The solid arrows "(a)" indicate the first stage of the Generalized Concept Overlay: all the evidence (i.e. degrees of confidence) coming from the analysis of the different modalities independently, are taken into account according to Eq. (10) for estimating a score associating the visual shot with each of the considered top-level domain concepts. These scores are shown next to the two such top-level concepts visible in this figure. The highest of these scores, in this example equal to 0.67 and corresponding to the "unrest, conflict and war" concept, is selected as dictated by Eq. (11). Subsequently, at the second stage of the Generalized Concept Overlay, the decision made on the top-level concept annotation is propagated to the more specific concepts that contributed to this decision, i.e. the sub-concepts of "unrest, conflict and war". This is illustrated by the dashed arrows "(b)". As a result of this, a new score is calculated for each of these sub-concepts according to Eq. (12) (these scores are not shown in this figure for readability purposes), and the largest of these scores indicates the single most plausible specific concept annotation of the shot, which in this example is "war". This result is in agreement with both visual and audio information analysis, as well as with the actual subject of the shot as identified during its manual annotation. In the second example of the same figure, the same process is shown for a "windstorms" shot. In this case, the visual and audio information analysis results are not in agreement. ASR linguistic analysis has identified the correct annotation; visual classification does not support the "Windstorms" concept (no such visual classifier has been trained), and identifies "war" as the most plausible annotation and "hurricane" as the second most plausible one. Combining these results and particularly taking into account that both "hurricane" and "windstorms" provide strong evidence in favor of the "disaster and accident" top-level concept, the correct annotation is identified.

The motivation behind the Generalized Concept Overlay is that it is difficult to directly combine the results of different analysis tools for determining the least abstract concept that should be used to annotate a temporal segment, considering that each individual modality analysis tool defines its own temporal content decomposition, takes into account its own subset of concepts (as also shown in the second example of Fig. 9), and has its own overall importance for analysis. Instead, taking advantage of the concept hierarchy and the fact that the results of concept detection at any level of this hierarchy can be directly propagated to the higher levels of it, we chose to make a decision on the classification of each temporal segment to the top-level concepts first, where all analysis results can be taken into account, and then at a second stage to follow an inverse process in order to make the final classification decision considering the less abstract concepts as well. A significant advantage of the proposed approach over learning-based ones (e.g. based on Bayesian Networks, Supervised Rank Aggregation approaches [14], etc.) is that no training is required for combining the individual modality analysis results. As shown in Eqs. (10) and (12), the proposed approach is based on evaluating functions ϕ, τ, μ and ξ, whose parameters are not determined from annotated training samples. Only classification of the content to one of the defined segment types (in our experiments, W_1 to W_3) is needed, which is independent of the concepts in C and can be realized by one or more generic classifiers (e.g. a studio/non-studio visual classifier). In contrast to this, taking into account all the above peculiarities of content (e.g. different decompositions etc.) and that the number of concepts in C may be in the order of hundreds of thousands, it is evident that a learning-based approach would require a very large amount of training data that is not generally available.

VII. Experimental Evaluation

A. Dataset and System Setup

The proposed news semantic multi-modal analysis system was experimentally evaluated on a test dataset of 91 short broadcast news videos from Deutsche Welle[5], having a total duration of approximately 4 hours. These were selected from a pool of 30 hours of video, on the basis of their relevance with the two top-level concepts depicted in Fig. 2, that were chosen for experimentation purposes. About 81% of the videos of the test dataset (74 out of 91) included audio, while very few videos included some frames with captions or other text that could be extracted by OCR techniques. Some of the videos were composed of an anchor shot followed by several external reporting shots; others included more than one sequences of anchor plus external reporting shots, while some others had no anchor shots at all. Shot segmentation of the test dataset, as discussed in section V-A, resulted in a total of 4576 shots. For enabling objective evaluation of the automatic analysis results, each shot was manually annotated with one concept of the ontology. In addition to the shot-level manual annotations, the annotators were asked to associate each entire video with a single concept of the ontology, corresponding to the temporally most dominant topic of the video. Manual annotation of each piece of visual information was performed by two annotators separately and, in case disagreement was observed in their annotations, these were reviewed by a third one.

Three uni-modal analysis methods, discussed in section V, were employed as the basis for multi-modal analysis: automatic speech recognition (ASR) and linguistic analysis of

[5]http://www.dw-world.de/

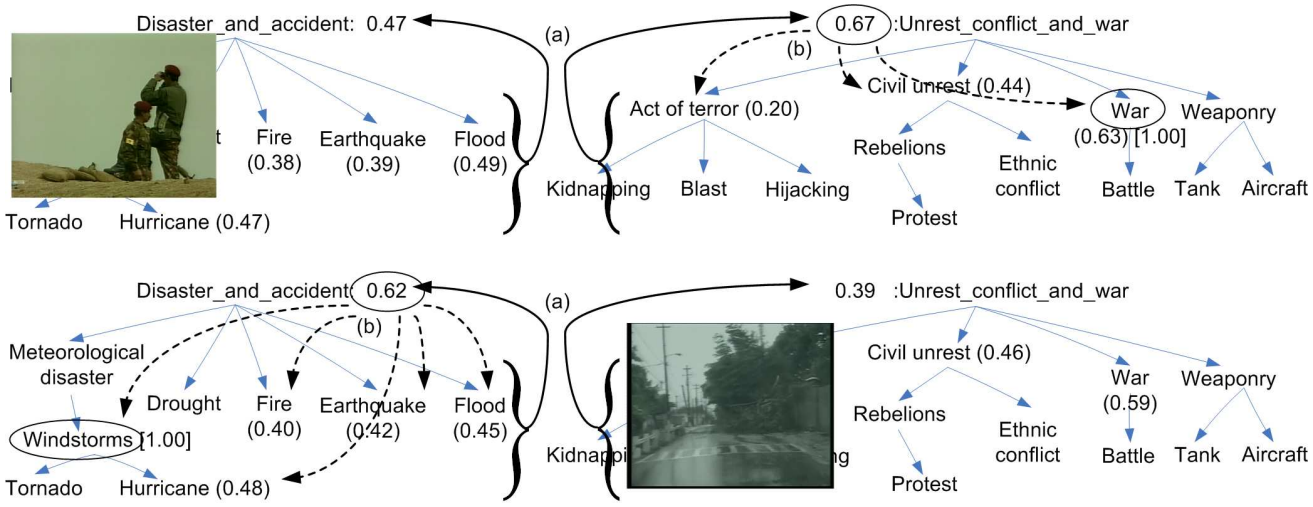

Fig. 9. Examples of the two-stage process for combining all the individual modality analysis results that is part of the Generalized Concept Overlay.

the ASR transcripts, resulting to decomposition D_1; linguistic analysis of optical character recognition (OCR) transcripts (D_2), and visual classification based on a combination of global and local features (D_3). For training the visual classifiers, a separate training set of Deutsche Welle videos was employed and visual classifiers were trained for the first 7 of the concepts of Table I. These concepts were selected on the basis of their frequency in the training and testing datasets. For less frequent concepts, such as the remaining ones of Table I, no visual classifiers were trained; therefore, these could be associated with the multimedia content only by means of linguistic analysis of ASR and OCR transcripts, which was not restricted to a subset of the concepts in C. The audio and linguistic analysis modules were developed with the use of other suitable corpora, not related to the employed test dataset of Deutsche Welle videos.

The decomposition of the visual modality to shots was chosen for serving as the reference decomposition, and based on this three types of content were defined as follows: W_1: Studio shots; W_2: External reporting with a dominant face on the video; W_3: External reporting with no dominant face on the video. A reliable studio/non-studio visual classifier and a face detector [38] were employed for automatically assigning each shot to one of these three types. Based on partitioning W, function μ was heuristically defined as:

$$\mu(s_j, s_m) = \begin{cases} 1, & \text{if } s_m \in D_2 \\ a_1, & \text{if } s_m \in D_1 \text{ and } s_j \in W_2) \\ a_2, & \text{if } s_m \in D_3 \text{ and } s_j \in W_2) \\ 0, & \text{if } (s_m \in D_3 \text{ and } s_j \in W_1) \\ a_3, & \text{otherwise} \end{cases} \quad (14)$$

where $0 < a_2 < a_1 < 1$ and $0 < a_3 < 1$. Function μ (also illustrated in Fig. 8) essentially encodes commonsense knowledge about news analysis, such as that audio information is more important than visual information when considering studio shots, etc. For experimentation, values $a_1 = 0.7$ and $a_2 = a_3 = 0.5$ were chosen.

Functions ξ_i were defined as $\xi_i(d_j(.)) = d_j(.)$ for $i = 1, 2$ (i.e. for the ASR and OCR linguistic analysis results), whereas for the visual classification results, ξ_3 was defined such that values $\xi_3(d_j(.))$ had a uniform distribution in $[0, 1]$ over a validation dataset.

B. Experimental Results

In Table II, results on the entire test dataset are presented for each of the employed uni-modal analysis techniques as well as for the Generalized Concept Overlay of section VI and two variants of it, illustrating the effect of modeling functions $\mu(s_j, s_m)$ and $\tau(s_j, s_m)$ as constants. Comparison with our earlier work on multi-modal analysis of news content [39] and with the unsupervised Borda Count and Borda Fuse methods [14] is also presented in this table. In [39], a multi-modal analysis approach that neither exploited the concept hierarchy nor took into account the variability of concept subsets considered by the individual modality analysis tools was proposed; only the concepts belonging to the intersection of the latter subsets were considered for combining the individual modality analysis results. The unsupervised Borda Count and Borda Fuse methods [14], [40] on the other hand consider all concepts of the employed ontology. They both treat the results of each uni-modal analysis technique as a ranked list, thus taking into account the rank of every concept in the list (i.e. first, second, etc.) rather than the actual values $d_j(.)$. This can be perceived as imposing a normalization of the uni-modal analysis results that is different than that of functions ξ_i used in Eqs. (10) and (12). They then fuse the ranked lists produced by the different uni-modal analysis tools. The rank of each result serves as the sole criterion in Borda Count, which averages the ranks of a given concept over all ranked lists. In Borda Fuse, the rank and the weight of each modality according to the type of the examined segment (i.e. the values of function μ used in Eqs. (10) and (12)) are employed. Using the latter, Borda Fuse calculates a weighted average of the ranks of a given concept over all ranked lists. In both methods, the concept for which the estimated average

TABLE I
EXAMINED CONCEPTS

Identifier	1	2	3	4	5	6	7
Concept name	Earthquake	Fire	Flood	Hurricane	War	Act of terror	Civil unrest
Identifier	8	9	10		11	12	13
Concept name	Windstorms	Riots	Massacre		Demonstration	Rebellions	Genocide

rank indicates that this concept appears higher than all other concepts in the fused list is selected as the final outcome of fusion. It can be seen in Table II that the proposed Generalized Concept Overlay approach outperforms the former approaches, achieving a higher correct annotation rate for the specific concepts c_{n^*} extracted by multi-modal analysis, and a higher or equal correct annotation rate for the top-level concepts c_{k^*}. The complete Generalized Concept Overlay also outperforms simpler variants of it that model certain functions as constants (i.e. they consider $\mu(s_j, s_m) = const$ and $\tau(s_j, s_m) = const$, respectively). It should be noted that the proposed approach does not require training with the use of a manually annotated dataset for combining the individual modality analysis results; thus, it may be particularly suitable to the large-scale semantic classification problem where training is difficult.

In Table III, similar results on a subset of the test dataset are presented; this subset comprises the 692 shots (out of the 4576 in total) for which at least two of the single modality analysis tools have returned an analysis result (e.g. shots for which at least one partially overlapping, in terms of time, audio segment exists and has been assigned to a semantic class by means of ASR and linguistic analysis). The motivation behind presenting results for this subset of the dataset is to illustrate more clearly the effect of different approaches in the way the different uni-modal analysis results are combined. It can be seen in Table III that, for this subset of the dataset, the majority of the results have been produced by visual classification and by linguistic analysis of ASR transcripts; due to the nature of the employed dataset (it is not rich in text that could be extracted by means of OCR), OCR linguistic analysis results are scarce. Concerning the multi-modal analysis techniques, it can be seen that the proposed approach significantly outperforms, in terms of the specific concepts c_{n^*} extracted by multi-modal analysis, our earlier work [39] (Chi Square=24.05, df=1, p$<$ .05) and the Borda Count method (Chi Square=22.0, df=1, p$<$ .05). The impact of function $\tau(s_j, s_m)$, in comparison to defining $\tau(s_j, s_m) = const$, is also shown to be significant (Chi Square=4.96, df=1, p$<$ .05). Less pronounced differences (thus also of lower statistical significance) in favor of the proposed approach are observed when comparing with the Borda Fuse method and when considering the annotation rates for the top-level concepts c_{k^*}. In evaluating the statistical significance of annotation performance differences in the above pairwise comparisons of approaches, the null hypothesis was defined as the annotation performance being the same for both approaches in each pair.

Corresponding confusion matrices for the the 692 shots and the 13 most frequent concepts of the dataset (in the order they are listed in Table I) are shown in Fig. 10. For visualization purposes, only the shots that were actually annotated with a concept are taken into account in each of these confusion matrices (thus, the "no result" outcomes of each analysis method were ignored when calculating the corresponding percentages, ranging from 0% to 100%). This was necessary for effectively visualizing e.g. the OCR linguistic analysis results that are scarce; consequently, the colors in Fig. 10 are not comparable between Fig. 10(b) and (c), and between any of these two and any of the remaining confusion matrices of the same figure. It can be seen in Fig. 10 that visual classification is most susceptible to annotation errors; ASR linguistic analysis is considerably more reliable overall but still consistently confuses between certain pairs of concepts (e.g. 3:Flood and 4:Hurricane; 5:War and 7:Civil Unrest); OCR linguistic analysis is very reliable. The Borda Count method as well as the method of [39] are shown to be moderately successful in combing the different uni-modal analysis results, since they are strongly affected by errors coming primarily from visual analysis. The Borda Fuse method and the proposed one are shown to be more successful, with the Borda Fuse method being affected a bit more by errors coming from ASR linguistic analysis (e.g. consistently confusing between concepts 5:War and 7:Civil Unrest), while the proposed approach is shown to handle better some of the errors coming from ASR linguistic analysis at the expense of being somewhat more sensitive to erroneous visual analysis results.

In order to examine how visual classification accuracy, which can clearly vary significantly depending on the choice of visual classifiers, the available training data, etc., affects the overall multi-modal analysis, an experiment was carried out where only subsets of the previously trained classifiers rather than all of them were considered. In particular, the 7 visual classifiers were ordered according to the prevalence of their corresponding concepts in the test set, in ascending order, and experiments were carried out by excluding the first of them, the first two, the first three, etc. In the last experiment of this series, only one visual classifier that corresponds to the single most prevalent concept in our test set was considered. The results are presented in Fig. 11, indicating that when the number of visual classifiers is reduced and consequently lower correct annotation rates are achieved by visual classification, the proposed multi-modal analysis approach succeeds to compensate this loss to a significant extent by exploiting the results of the other modalities, providing that such results exist. Still, it is evident from the same figure that visual classification does contribute to the final outcome of multi-modal analysis; this contribution is small, in the portion of the dataset for which other modality analysis results exist, and far more significant when considering the entire dataset.

Another experiment was subsequently carried out, making the assumption that each entire video (rather than each in-

TABLE II
MULTI-MODAL ANALYSIS RESULTS IN THE NEWS DOMAIN - ENTIRE DATASET

Process	c_{n^*} correct	c_{n^*} incorrect	c_{n^*} no result	c_{k^*} correct	c_{k^*} incorrect	c_{k^*} no result
Visual classification	43.4%	56.6%	0%	81.6%	18.4%	0%
ASR linguistic analysis	8.5%	3.2%	88.3%	11.6%	2.9%	85.5%
OCR linguistic analysis	0.4%	0%	99.6%	0.6%	0.2%	99.2%
Generalized Concept Overlay with $\mu(s_j, s_m) = const$	46.5%	53.5%	0%	82.8%	17.2%	0%
Generalized Concept Overlay with $\tau(s_j, s_m) = const$	46.2%	53.8%	0%	**82.9%**	17.1%	0%
Generalized Concept Overlay	**47.1%**	52.9%	0%	**82.9%**	17.1%	0%
Multi-modal analysis method of [39]	45.1%	54.9%	0%	82.6%	17.4%	0%
Borda Count method [14]	45.2%	54.8%	0%	82.8%	17.2%	0%
Borda Fuse method [14]	46.9%	53.1%	0%	**82.9%**	17.1%	0%

TABLE III
MULTI-MODAL ANALYSIS RESULTS IN THE NEWS DOMAIN - DATASET RESTRICTED TO SHOTS FOR WHICH MORE THAT ONE SINGLE MODALITY ANALYSIS RESULTS EXIST

Process	c_{n^*} correct	c_{n^*} incorrect	c_{n^*} no result	c_{k^*} correct	c_{k^*} incorrect	c_{k^*} no result
Visual classification	35.4%	64.6%	0%	73.1%	26.9%	0%
ASR linguistic analysis	56.5%	39.6%	3.9%	76.9%	19.2%	3.9%
OCR linguistic analysis	2.3%	0%	97.7%	3.8%	0%	96.2%
Generalized Concept Overlay with $\mu(s_j, s_m) = const$	56.4%	43.6%	0%	80.8%	19.2%	0%
Generalized Concept Overlay with $\tau(s_j, s_m) = const$	54.2%	45.8%	0%	**81.8%**	18.2%	0%
Generalized Concept Overlay	**60.1%**	39.9%	0%	81.2%	18.8%	0%
Multi-modal analysis method of [39]	47%	53%	0%	79.8%	20.2%	0%
Borda Count method [14]	47.5%	52.5%	0%	80.9%	19.1%	0%
Borda Fuse method [14]	58.8%	41.2%	0%	81.6%	18.4%	0%

dividual shot) is about a single subject, thus all shots of it can and should be associated with a single concept of the employed ontology. The motivation behind this experiment was to test the influence of the selected content decomposition to the performance of multi-modal analysis, and in particular the possible improvement of analysis results when considering larger story-telling units ("scenes") rather than visual shots as the elementary pieces of video information; taking the whole video as a scene is clearly the extreme case. In this experiment, the manually-generated video-level annotations discussed at the beginning of this section were used as ground truth annotation for evaluation purposes, in place of the shot-level ones. The Generalized Concept Overlay technique was adapted to this scenario by being applied at the shot level, as in all previous experiments, and its results being subsequently evaluated by a simple voting mechanism which selected the single most dominant concept across all shots as the final annotation for the entire video. As a result, the correct classification rates of the Generalized Concept Overlay rose to 75.3% and 93.1% for c_{n^*} and c_{k^*} respectively on the entire dataset, showing a significant increase compared to the results of Table II.

Finally, it should be noted that besides the annotation of each shot or other temporal segment with one concept of the ontology expressing the thematic categorization of the news item, the result of multi-modal analysis can also include additional semantic information such as location names, person names, etc. These are extracted as part of the linguistic analysis of ASR and OCR transcripts. Although elaborate techniques for combining such additional information can be envisaged (e.g. similar to the one presented in this work for the thematic categorization results), in practice a simple unification approach was adopted in our experiments; more specifically, the additional information coming from ASR and OCR analysis was accumulated and in case of contradictory information, the OCR results prevailed. As a result, over one third of the shots in our dataset was automatically annotated with information that is in addition to c_n^*, c_k^*; out of this, approximately 55% concerned location names, 22% person names, and 6% dates. The evaluation of the correctness of these results is beyond the scope of this work, since the focus is on the thematic categorization results discussed above, but these clearly indicate the added value of using multiple specialized individual modality analysis tools in a multi-modal analysis scheme, rather than attempting to jointly process at a single stage all low-level features that come from the different modalities.

VIII. CONCLUSIONS

The detailed analysis of the results in the previous section, where the corresponding confusion matrices where presented, revealed that multi-modal analysis using the proposed Generalized Concept Overlay approach succeeds in improving the results of any of the employed uni-modal analysis methods. Nevertheless, it is evident that the breadth of this improvement is greatly dependent upon the individual modality analysis results that serve as input to multi-modal analysis. These, in turn, depend not only on the performance of the employed single-modality analysis methods but also (and maybe even to a greater degree) on the specifics of the content itself, i.e. whether it contains audio or not, whether news-related legends are typically projected or not on the screen by the news agency producing the content or by the broadcaster, etc. In the case of the employed Deutsche Welle dataset, it was shown that although ASR and OCR linguistic analysis can provide valuable and very accurate information about

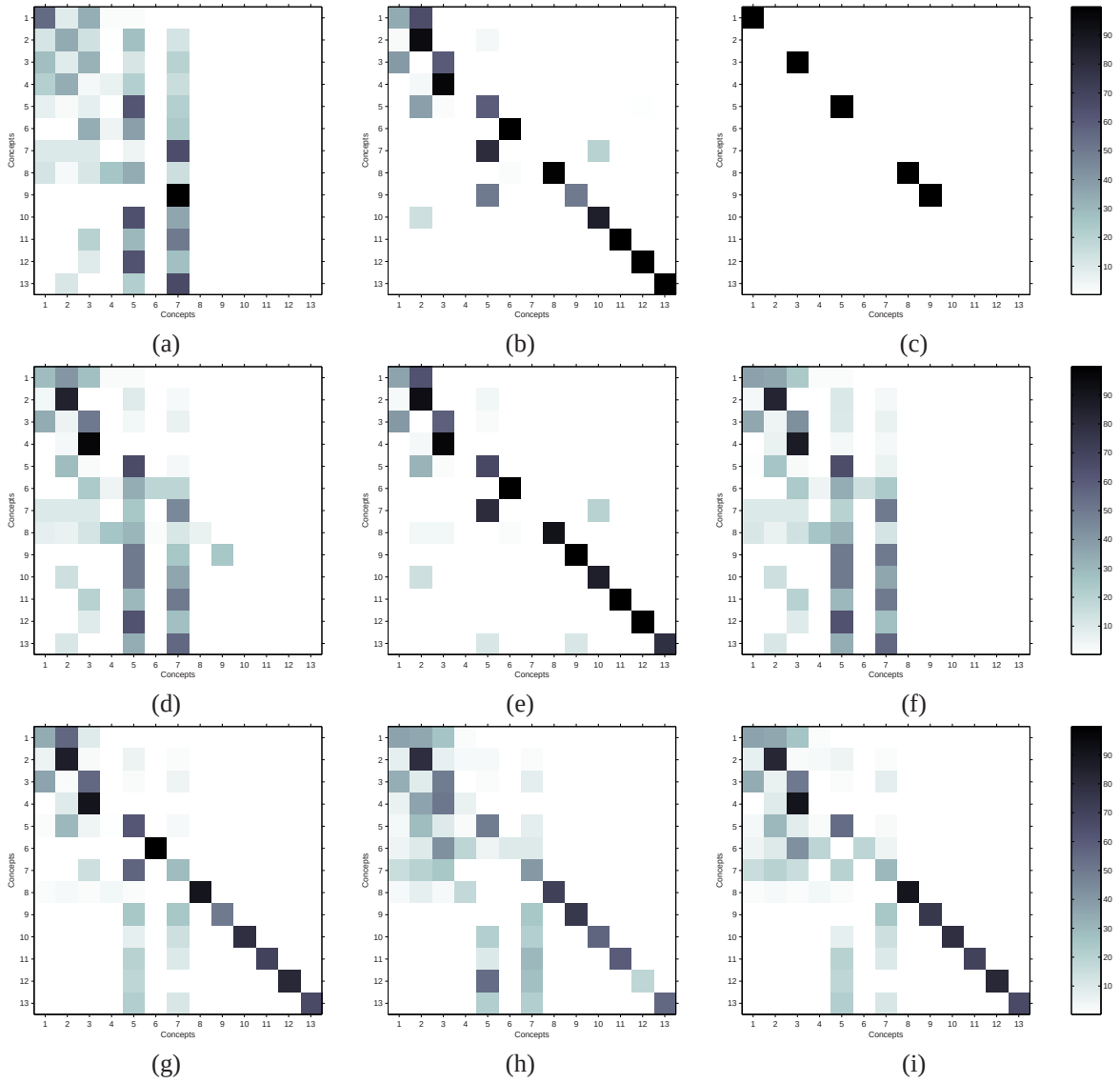

Fig. 10. Confusion Matrices for the 13 concepts of Table I - dataset restricted to shots for which more that one single modality analysis results exist. (a) Visual Classification, (b) ASR linguistic analysis, (c) OCR linguistic analysis, (d) Borda Count method [14], (e) Borda Fuse method [14], (f) method of [39], (g) Generalized Concept Overlay with $\mu(s_j, s_m) = const$, (h) Generalized Concept Overlay with $\tau(s_j, s_m) = const$, (i) Generalized Concept Overlay.

the semantics of the content, treating the video at the shot level results in relatively few shots being annotated by these components with anything other than "no result". This is consistent with the nature of broadcast news, where one of the prevailing journalistic rules in preparing the presentation of news can be summarized as "let the images tell their own story". Consequently, the exact type of the incident in question (e.g. a "fire") is not verbally repeated in every visual shot; it is more often announced by an anchorperson during a "studio" shot, followed by several shots where the visual modality prevails and few, if any, semantics are conveyed by speech or legends on the screen. This is the reason why when larger story-telling units are considered as the elementary pieces of news information (e.g. as in our last experiment, where the entire video was treated as a single story-telling unit), considerable increase in the correct semantic annotation rates can be achieved. On the other hand, though, linguistic analysis of ASR and OCR transcripts is invaluable in extracting additional semantic metadata such as location names, person names, etc., which are beyond the reach of any visual analysis technique unless considering very restricted application scenarios (e.g. involving a limited number of people that appear on the video and for which appropriate face recognition models can be trained, etc.). These conclusions provide the guidelines for the use of the analysis techniques presented in this work as well as of other similar techniques in real-life multimedia news management applications.

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. Papastathis, and M. 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.

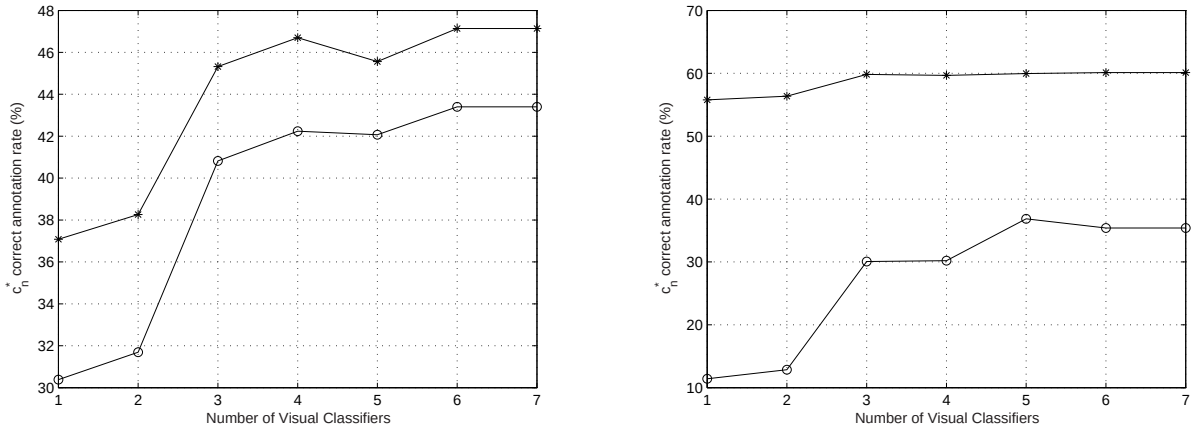

Fig. 11. Results of visual classification ("-o-") and of Generalized Concept Overlay ("-*-") when the number of visual classifiers (and, consequently, the performance of visual classification) varies, for the datasets used in (a) Table II, (b) Table III.

[3] J. Assfalg, M. Berlini, A. D. 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. 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. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Combining Global and Local Information for Knowledge-Assisted Image Analysis and Classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[7] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive Multimodal Fusion by Uncertainty Compensation With Application to Audiovisual Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, March 2009.

[8] J. Alexander and T. Becker, "Overlay as the basic operation for discourse processing in the multimodal dialogue system," in *Proc. IJCAI-01 Workshop on Knowledge and reasoning in practical dialogue Systems*, Seattle, Washington, August 2001.

[9] J. Zhang, "A Novel Video Searching Model Based on Ontology Inference and Multimodal Information Fusion," in *Int. Symposium on Computer Science and Computational Technology*, Shanghai, December 2008, pp. 489–492.

[10] W. H. Lin and A. G. Hauptmann, "News video classification using SVM-based multimodal classifiers and combination strategies," in *Proc. ACM Multimedia (MM'02)*, 2002, pp. 323–326.

[11] W.-N. Lie and C.-K. Su, "News video classification based on multimodal information fusion," in *Proc. IEEE Int. Conf. on Image Processing (ICIP05*, September 2005, pp. 1213–1216.

[12] C. Laudy and J.-G. Ganascia, "Information fusion in a TV program recommendation system," in *Proc. Int. Conf. on Information Fusion*, June 2008, pp. 1–8.

[13] W. Wahlster, "Fusion and Fission of Speech, Gestures, and Facial Expressions," in *Proc. Int. Workshop on Man-Machine Symbiotic Systems*, Kyoto, Japan, 2002, pp. 213–225.

[14] Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li, "Supervised rank aggregation," in *Proc. 16th Int. Conf. on World Wide Web*, 2007, pp. 481–490.

[15] L. Kennedy, S.-F. Chang, and A. Natsev, "Query-Adaptive Fusion for Multimodal Search," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 567–588, April 2008.

[16] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual, audio and text analysis for news video," in *IEEE Int. Conf. on Image Processing (ICIP 2000)*, Vancouver, Canada, September 2000, pp. 520–523.

[17] H. Luo, J. Fan, S. Satoh, J. Yang, and W. Ribarsky, "Integrating multi-modal content analysis and hyperbolic visualization for large-scale news video retrieval and exploration," *Signal Processing Image Communication*, vol. 23, no. 7, pp. 538–553, August 2008.

[18] M.-Y. Chen and A. Hauptmann, "Multi-modal classification in digital news libraries," in *Proc. Joint ACM/IEEE Conference on Digital Libraries*, Tucson, Arisona, USA, June 2004, pp. 212–213.

[19] P. Wang, R. Cai, and S.-Q. Yang, "A hybrid approach to news video classification with multimodal features," in *Proc. Fourth Int. Conf. on Information, Communications and Signal Processing (ICICS-PCM 03)*, Singapore, December 2003, pp. 787–791.

[20] Y. Wu, C.-Y. Lin, E. Chang, and J. Smith, "Multimodal information fusion for video concept detection," in *Proc. IEEE Int. Conf. on Image Processing*, vol. 4, Singapore, Oct. 2004, pp. 2391–2394.

[21] C. Laudy, J.-G. Ganascia, and C. Sedogbo, "High-level fusion based on conceptual graphs," in *Proc. Int. Conf. on Information Fusion*, Quebec, Canada, July 2007, pp. 1–8.

[22] Y. Lv, W. Ng, J. Lee, B. Sun, and D. Yeung, "Information extraction based on information fusion from multiple news sources from the web," in *IEEE Int. Conf. on Systems, Man and Cybernetics*, Singapore, Octomber 2008, pp. 1471–1476.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] D. Tax and R. 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.

[30] F. de Jong, T. Westerveld, and A. 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.

[31] M. Huijbregts, C. Wooters, and R. Ordelman, "Filtering the unknown: Speech activity detection in heterogeneous video collections," in *proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[32] D. van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," in *Machine Learning for Multimodal*

Interaction (MLMI), ser. Lecture Notes in Computer Science, vol. 4299. Berlin: Springer Verlag, October 2007, pp. 371–384.

[33] S. J. Young, N. Russell, and J. Thornton, "Token passing: A simple conceptual model for connected speech recognition systems," Cambridge University Engineering Dept, Tech. Rep., 1989.

[34] M. Huijbregts, R. Ordelman, and F. de Jong, "Fast N-Gram language model look-ahead for decoders with static pronunciation prefix trees," in *proceedings of Interspeech*, Brisbane, Australia, September 2008.

[35] B. Carpenter, *The Logic of Typed Feature Structures*. Cambridge: CUP, 1992.

[36] W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu, "Shallow processing with unification and typed feature structures — foundations and applications," *Künstliche Intelligenz*, vol. 1, pp. 17–23, 2004.

[37] W. Drozdzynski, H.-U. Krieger, J. Piskorski, and U. Schäfer, "Sprout - a general-purpose nlp framework integrating finite-state and unification-based grammar formalisms," in *Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing*. Helsinki: Springer, 9 2005.

[38] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. Journal of Computer Vision*, vol. 2, no. 57, pp. 137–154, 2004.

[39] V. Mezaris, S. Gidaros, G. T. Papadopoulos, W. Kasper, R. Ordelman, F. de Jong, and I. Kompatsiaris, "Knowledge-assisted cross-media analysis of audio-visual content in the News domain," in *Proc. Int. Workshop on Content-Based Multimedia Indexing (CBMI08)*, London, UK, June 2008, pp. 280–287.

[40] J. A. Aslam and M. Montague, "Models for Metasearch," in *Proc. 24th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, USA, September 2001, pp. 276–284.