Proc. 4th International Workshop on Semantic Media Adaptation and Personalization (SMAP09), San Sebastian, Spain, December 2009, pp. 27-32.

Generalized concept overlay for semantic multi-modal analysis of audio-visual content

Vasileios Mezaris, Spyros Gidaros, Ioannis Kompatsiaris
Informatics and Telematics Institute / Centre for Research and Technology Hellas
6th Km Charilaou-Thermi Road, Thermi 57001, Greece
Email: {bmezaris, gidaros, ikom}@iti.gr

Abstract—**In this work, the problem of performing multi-modal analysis of audio-visual streams by effectively combining the results of multiple uni-modal analysis techniques is addressed. A non-learning-based approach is proposed to this end, that 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 concepts of an ontology that they consider, the varying semantic importance of each modality, and other factors. Preliminary results from the application of the proposed approach to broadcast News content reveal its effectiveness.**

Keywords-**Video analysis; Semantic multi-modal analysis;**

I. Introduction

Multimedia content is nowadays a key element of our everyday lives. The widespread availability of inexpensive capturing devices, the proliferation of broadband internet connections and the development of innovative media sharing services over the World Wide Web have contributed the most to this. However, the tremendous increase in the amount of multimedia content created every day raises new and important challenges regarding the efficient organization, access and presentation of it. The cornerstone of the efficient manipulation of such 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].

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 [2]. 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. Multi-modal techniques can be broadly classified to those jointly processing low-level features that come from different modalities [3] [4], and those that combine the results of multiple uni-modal analysis techniques [5] [6]. 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). This necessitates the use of different modality-specific analysis techniques, prior to jointly considering the multiple modalities.

In this work, we address the problem of multi-modal analysis of audio-visual streams by combining the results of multiple uni-modal techniques, and we propose a formulation that takes into account the potential variability of the different uni-modal analysis techniques in terms of the decomposition of the stream that they adopt, the concepts of an ontology that they consider, the varying semantic importance of each modality, etc. The remainder of the paper is organized as follows: In section II, our multi-modal analysis problem is formulated. In section III, the proposed generalized concept overlay approach is presented. In section IV experimental results from the application of the proposed approach in the News domain are presented and conclusions are drawn in section V.

II. Problem formulation

The objective of analysis in this study is to associate each elementary temporal segment of the audiovisual steam 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. [7]. However, the above simplified ontology definition is sufficient for our study, since we focus on exploiting only the concepts and their hierarchy for a specific task of multi-modal analysis, namely for combining the results of multiple uni-modal analysis techniques.

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 its own elementary segments of interest in a multimedia content item and, considering all concepts of C or a subset of them, associates each elementary segment 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 $[0, 1]$. The application of the 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 ASR transcripts it would probably be based on audio classification or speaker diarization results instead. All the decompositions together define a temporal segment set

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

It is useful to observe that S is a set of temporal segments with no hierarchy, where multiple segments may temporally overlap in full or in part. 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. An overview of the proposed approach is shown in Fig. 1.

III. Generalized concept overlay

A simple, yet crude non-learning-based solution to the analysis problem formulated above 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

$$\tau(s_j, s_m) = \begin{cases} \frac{\min(t_j^B, t_m^B) - \max(t_j^A, t_m^A)}{t_j^B - t_j^A} & \text{if } \Gamma > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where s_j is the reference segment and

$$\Gamma = (t_j^B - t_m^A)(t_m^B - t_j^A) \tag{6}$$

In order to take advantage of the concept hierarchy, we

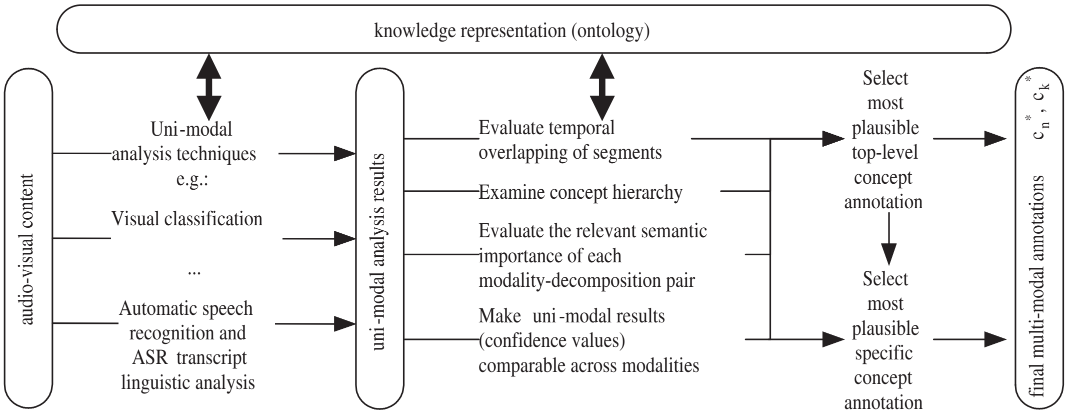
knowledge representation (ontology)
audio-visual content
Uni-modal analysis techniques e.g.:
Visual classification
...
Automatic speech recognition and ASR transcript linguistic analysis
uni-modal analysis results
Evaluate temporal overlapping of segments
Examine concept hierarchy
Evaluate the relevant semantic importance of each modality-decomposition pair
Make uni-modal results (confidence values) comparable across modalities
Select most plausible top-level concept annotation
Select most plausible specific concept annotation
c_n^*, c_k^*
final multi-modal annotations

Figure 1. Overview of the proposed approach.

define function $\phi : C^2 \rightarrow [0,1]$ such that

$$\phi(c_k,c_n)=\begin{cases} 1, & \text{if } c_n{=}c_k \text{ or } c_n \text{ is a sub-concept of } c_k \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

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,

$$W=\{W_q\}_{q=1}^{Q} \tag{8}$$

This 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 $s_m \in D_l$, indicates the relevant semantic importance of the modality corresponding to decomposition D_l for the analysis of segments of type W_q.

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 $\xi_i,\ i,1.\ldots,I$, one for each modality, that attempt to make values $\xi(d_j(.))$ comparable across modalities by enforcing them to have common statistics over a reasonably large dataset. In the sequel, the index to ξ will be omitted for notational simplicity; the use of the function ξ that was defined for the modality from which its argument value $d_j(.)$ comes 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:

$$\psi(s_j,c_k)=\sum_{n=1}^{K}\left[\phi(c_k,c_n)\cdot\left(\sum_{m=1}^{J}\tau(s_j,s_m)\cdot\mu(s_j,s_m)\cdot\xi(d_m(c_n))\right)\right] \tag{9}$$

Then,

$$k^*=\arg\max_k\big(\psi(s_j,c_k)\big) \tag{10}$$

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 concepts 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. As can be seen in Eq. (9), this first stage of the multi-modal analysis process jointly takes into account the concept hierarchy (expressed by $\phi(c_k,c_n)$), allowing all concepts that are sub-concepts of c_k to contribute to its selection; the temporal overlapping of segments (function $\tau(s_j,s_m)$); the varying semantic importance of the different modalities (function $\mu(s_j,s_m)$); and the need to make values $d_j(.)$ generated by different analysis tools comparable (function $\xi(d_m(c_n))$).

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. (7), 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:

$$\rho(s_j,c_n)=\sum_{m=1}^{J}\tau(s_j,s_m)\cdot\mu(s_j,s_m)\cdot\xi(d_m(c_n)) \tag{11}$$

Then,

$$n^* = \arg\max_n \big(\rho(s_j, c_n)\big) \tag{12}$$

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. Similarly to the first stage of the proposed multi-modal analysis process, the temporal overlapping of segments, the varying semantic importance of the different modalities and the need to make confidence values generated by different analysis tools comparable are also taken into account in Eq. (11).

The motivation behind the above two-stage process is that, considering that each individual modality analysis tool defines its own temporal content decomposition, takes into account its own subset of concepts, and has its own overall importance for analysis, it is difficult to combine the results of different analysis tools for directly determining the least abstract concept that should be used to annotate a temporal segment of the content. 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 [9], and others) is that no training is required for combining the individual modality analysis results. 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 or thousands, it is evident that a learning-based approach would require a very large amount of training data that is not generally available.

IV. Experimental results

One of the possible applications of the proposed generalized concept overlay approach is in the domain of News audio-visual content analysis. In the News domain, the "deep" analysis of audio-visual information (e.g. the linguistic analysis of ASR and OCR transcripts) can provide valuable semantically-rich metadata about the content (e.g. person or location names), which would otherwise be impossible to extract. Thus, the application of different modality-specific analysis techniques prior to jointly considering the multiple modalities is most appropriate.

For News analysis, an ontology based on an extension of the IPTC[1] tree for news categorization was employed in this work. The 17 top-level IPTC categories served as the top-level concepts in C', while more than 1000 concepts overall where included in C. A small subset of the concepts and their hierarchy are depicted in Fig. 2.

Three uni-modal analysis methods were employed as the basis for multi-modal analysis: automatic speech recognition (ASR) and linguistic analysis of 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). Visual classification was based on a limited number of visual classifiers (7; each for one different concept of the ontology). The OCR transcripts were generated by application of commercial OCR software to video keyframes. The ASR, linguistic analysis and visual classification methods employed are outlined in [10].

The decomposition of the visual modality to shots was chosen for serving as the reference decomposition, and based on this four 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 but with speech voiceover; W_4: External reporting with no speech. A reliable studio/non-studio visual classifier and a face detector [11] were employed, along with the aforementioned uni-modal semantic analysis techniques, for automatically assigning each shot to one of these four types. Based on partitioning W, function μ was heuristically defined as:

$$\mu(s_j, s_m) = \begin{cases} 1, & \text{if } s_m \in D_2 \\ a_1, & \text{if } s_m \in D_1 \text{ and } s_j \in W_2) \\ a_2, & \text{if } s_m \in D_3 \text{ and } s_j \in W_2) \\ 0, & \text{if } (s_m \in D_3 \text{ and } s_j \in W_1) \\ & \text{or } (s_m \in D_1 \text{ and } s_j \in W_4) \\ a_3, & \text{otherwise} \end{cases} \tag{13}$$

where $0 < a_2 < a_1 < 1$ and $0 < a_3 < 1$. For experimentation, values $a_1 = 0.7$ and $a_2 = a_3 = 0.5$ were used.

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.

Results on a dataset of 91 short broadcast news videos from Deutsche Welle[2] belonging to the two top-level IPTC categories of Fig. 2, and comparison with our earlier work on this topic [10] and the unsupervised Borda Count method (one of the methods discussed in [9]) are shown in Table I. For this evaluation, only the portion of the dataset for which results were returned by more than one uni-modal analysis tools was considered. In [10], a multi-modal analysis approach that neither exploited the concept hierarchy

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

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

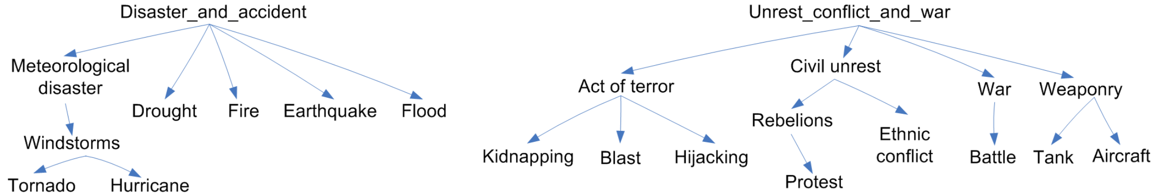

Figure 2. Subset of concepts and their hierarchy in the employed ontology for News. Two of the 17 top-level concepts ("Disaster and accident", "Unrest, conflict and war") and a few of their sub-concepts are depicted.

Table I
MULTI-MODAL ANALYSIS RESULTS IN THE NEWS DOMAIN

Method	c_{n^*} correct annotation rate	c_{k^*} correct annotation rate
Method of [10]	47%	79.8%
Borda Count [9]	47.5%	80.9%
Proposed approach	60.1%	81.2%

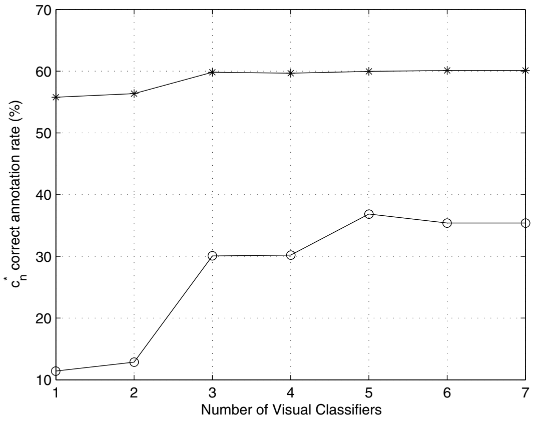

Figure 3. Results of visual classification ("-o-") and of Generalized Concept Overlay ("-*-") when the number of visual classifiers varies.

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 method on the other hand considers all concepts of the employed ontology; it treats the results of each uni-modal analysis technique as a ranked list, and works by fusing these lists. It can be seen in Table I that the proposed approach outperforms the former methods, achieving higher correct annotation rates for both the top-level concepts c_{k^*} and the more specific ones c_{n^*} extracted by multi-modal analysis.

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 proposed generalized concept overlay, 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. 3, indicating that when the number of visual classifiers is reduced and consequently lower correct annotation rates are achieved by visual classification, the proposed generalized concept overlay approach succeeds to compensate this loss to a significant extent, by exploiting the results of the other modalities.

V. CONCLUSIONS

In this paper a Generalized Concept Overlay approach was proposed for effectively combining the results of multiple uni-modal analysis techniques to effect multi-modal semantic analysis of audio-visual streams. Preliminary experiments on broadcast News content reveal its effectiveness and its superiority over two previous approaches of the literature. The proposed approach is also applicable to other domains beyond the News one, since it was developed based on a domain-independent problem formulation.

ACKNOWLEDGMENT

This work was supported by the European Commission under contract FP6-027685 MESH. The authors would like to thank the MESH partners who provided the ASR and linguistic analysis results used in the experiments.

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] 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.

Proc. 4th International Workshop on Semantic Media Adaptation and Personalization (SMAP09), San Sebastian, Spain, December 2009, pp. 27-32.

[3] 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.

[4] W.-N. Lie and C.-K. Su, “News video classification based on multi-modal information fusion,” in *Proc. IEEE Int. Conf. on Image Processing (ICIP05)*, September 2005, pp. 1213–1216.

[5] 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.

[6] 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.

[7] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis, “Ontology-Driven Semantic Video Analysis Using Visual Information Objects,” in *Proc. Second Int. Conf. on Semantics and Digital Media Technologies (SAMT07), Springer LNCS, vol. 4816*, Genova, Italy, December 2007, pp. 56–69.

[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] 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*. New York, NY, USA: ACM, 2007, pp. 481–490.

[10] 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.

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