

Multimedia Content Processing and Retrieval in the REVEAL THIS setting

Stelios Piperidis¹, Harris Papageorgiou¹, Katerina Pastra¹, Thomas Netousek², Eric Gaussier³, Tinne Tuytelaars⁴, Fabio Crestani⁵, Francis Bodson⁶, Chris Mellor⁷

Abstract— The explosion of multimedia digital content and the development of technologies that go beyond traditional broadcast and TV have rendered access to such content important for all end-users of these technologies. REVEAL THIS develops content processing technology able to semantically index, categorise and cross-link multiplatform, multimedia and multilingual digital content, providing the system user with search, retrieval, summarisation and translation functionalities.

Index Terms—audio-image-text analysis, cross-media linking and indexing, cross-media categorisation, cross-media summarisation, cross-lingual translation

I. INTRODUCTION

THE development of methods and tools for content-based organization and filtering of the large amount of multimedia information that reaches the user is a key issue for its effective consumption. Despite recent technological progress in the new media and the Internet, the key issue remains “how digital technology could *add value* to information channels and systems” [1].

REVEAL THIS aims at answering this question by tackling the following scientific and technological challenges:

- enrichment of multilingual multimedia content with semantic information like topics, speakers, actors, facts, categories
- establishment of semantic links between pieces of information presented in different media and languages
- development of cross-media categorization and summarization engines
- deployment of cross-language information retrieval and machine translation to allow users to search for and retrieve information according to their language preferences.

The REVEAL THIS project (www.reveal-this.org) is a thirty-month-STREP project funded by the FP6-IST programme of the European Commission, contract No FP6-IST-511689. It is designed and implemented by the REVEAL THIS consortium comprising Institute for Language and Speech Processing (Coordinator), SAIL LABS Technology AG, Xerox-The Document Company S.A.S, Katholieke Universiteit Leuven R&D, University of Strathclyde, BeTV SA and TVEyes UK Ltd.

¹ Institute for Language and Speech Processing, Athens, Greece, ²SAIL LABS Technology AG, ³Xerox - The Document Company S.A.S, ⁴Katholieke Universiteit Leuven R&D, ⁵University of Strathclyde, ⁶BeTV SA, ⁷TVEyes UK Ltd

Web, TV and/or Radio content is fed into the REVEAL THIS prototype, it is analysed, indexed, categorized, summarized and stored in an archive. This content can be searched and/or pushed to a user according to his/her interests. Novice and advanced computer users are targeted; they can both access the system through the web and perform simple or advanced searches respectively. Furthermore, mobile phone access to the system is possible through GPRS or Wireless Lan connection to the system’s mobile phone server. In this case, the system’s role is more proactive, in that it pushes information to the user according to the user’s profile. EU politics, news and travel data are handled by the system in English and Greek.

II. THE REVEAL THIS SYSTEM

As depicted in Figure 1, the REVEAL THIS system comprises a number of single-media and multimedia technologies that can be grouped, for presentation purposes, in the following subsystems: (i) Content Analysis & Indexing (CAIS), (ii) Cross-media Categorisation (CCS), (iii) Cross-media Summarisation (CSS), (iv) Cross-lingual Translation (CLTS), and (v) Cross-media Content Access and Retrieval.

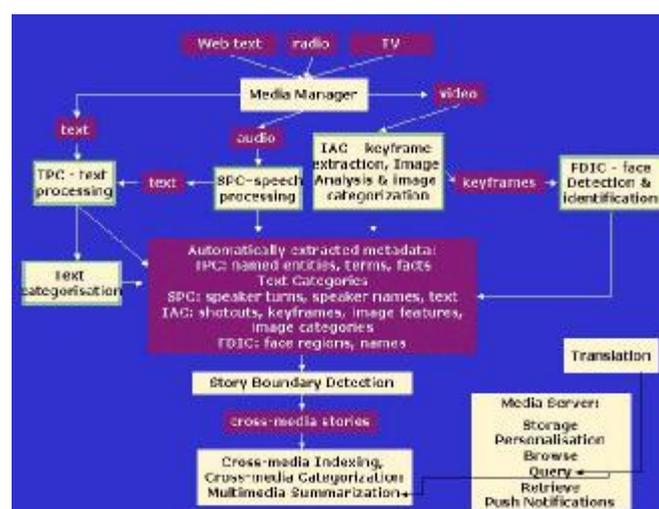


Figure 1 : REVEAL THIS system workflow

A. Cross-media Content Analysis

The CAIS subsystem consists of technologies and components for medium-specific analysis:

- Speech processing (SPC) – involving speech recognition, speaker identification and speaker turn detection

- Image analysis and categorization (IAC)- involving shot and keyframe extraction, low-level visual feature extraction, image categorisation [2]
- Face analysis (FDIC) – involving face recognition & identification [3]
- Text processing (TPC)- involving named entity, term & fact extraction, topic detection
- Cross-media Indexing (CMIC) – catering for the establishment of links between all above-mentioned metadata for a multimedia file using a modified TF-IDF and a Dempster-Shafer based approach [4]

The metadata/indices produced by the above components are aligned, synchronized, linked to the corresponding points of the source material (text, audio and video) and encoded in MPEG7. Information suggested by audio processing (speaker turns) and topic detection is taken into account to segment the audiovisual or audio files into segments, or what one could call “stories” i.e. thematic sections of the document. Categorisation, summarisation and translation of multimedia documents themselves make use of part of these metadata.

B. Cross-media categorisation

The categorization subsystem considers documents containing not only text or images but a combination of different types of media (text, image, speech, video). A multiple-view fusion method is adopted, which builds 'on top' of two single-media categorizers, a textual and an image categorizer, without the need to re-train them. Data annotated manually for both textual and image categories is used for training the cross-media categorizer. In that set, dependencies between single-media category systems are exploited in order to refine the categorization decisions made [5].

C. Cross-media summarisation

The cross-media summarisation subsystem (CSS) determines and presents the most salient parts according to the users' profiles and interests by fusing video, audio and textual metadata. It comprises three major components: the textual-based summarization (TS), the visual-based summarization (VS), and the cross-media summarization components, aiming at fusing the two analyses and creating a self-contained object. Building on the MEAD development platform [6], the TS component *extracts* the top-ranked sentences of a story: for each sentence, a salience score is computed as a weighted sum of summary-worthy features.

The VS component comprises the *scene segmentation, scene clustering & labelling* modules [6]. The scene segmenter segments the video sequence into scenes. Scene boundaries are detected as local minima in the visual coherence function with each scene corresponding, ideally, to a story of the video. Scene clustering caters for simple applications that need a few indicative images. Keyframes of the scene are clustered into larger parts, from which a prototypical image is chosen. Clustering is repeated iteratively to acquire a hierarchical cluster tree. The prototypes of these clusters can be seen as representative images of the scene. Going a step further, scene labelling is

invoked, for creating structured views of a file; currently, it is news programmes that can be browsed in such a way, allowing the user to watch all “anchor”, “interview”, and “reportage” segments. The module labels all shots of a file accordingly, by exploiting a belief propagation network. Finally, the CSS brings all these pieces of information together, providing visualisation interfaces (SMIL/HTML+TIME) that enable the user to preview multimedia objects effectively, before downloading them.

D. Cross-lingual Translation

The CLTS subsystem allows users to query documents written in different languages, to categorise content expressed in different languages and to preview language specific summaries. A bilingual lexicon extraction module is used to generate lexical equivalences for query translation purposes, but also to replace keywords in a target language, in case a document is linguistically not well formed (e.g. output from a speech recognizer) and thus, not effectively translated. Last, a statistical machine translation module is responsible for providing translations of the textual part of the summaries produced by the Cross-Media Summarization Subsystem.

E. Usability Evaluation

Apart from technical evaluation of the system components (cf [2], [3], [5], [7]), the integrated REVEAL THIS prototype is currently being evaluated by prospective users following a task-based approach. A pool of about 30 users for each application (pull and push) has been created. Based on typical search sessions of these users, appropriate search tasks have been created for the users to undertake using the REVEAL THIS prototype. Feedback from the evaluation will guide the final system refinements. The technology developed is envisaged to contribute to a content management platform that can be *used by content providers*, to add value to their content, and *directly by end users*, for accessing multimedia information.

REFERENCES

- [1] K. Pastra and S. Piperidis, "Video Search: New Challenges in the Pervasive Digital Video Era", Journal of Virtual Reality and Broadcasting, in press
- [2] F. Perronnin, C. Dance, G. Csuska, M. Bressan, "Adapted Vocabularies for Generic Visual Categorization", European Conference on Computer Vision (ECCV), Graz, Austria, 2006
- [3] M. De Smet, R. Fransens, L. Van Gool, "A generalised EM approach for 3D model based face recognition under occlusions" , in Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), New York, USA, 2006
- [4] M. Yakici and F. Crestani, "Cross-media Indexing in the Reveal This prototype", in Proceedings of the LREC workshop on "Crossing media for improved information access", Genoa, Italy, 2006
- [5] J.Renders, E.Gaussier, C.Goutte, F.Pacull, G.Csuska, "Categorization in multiple category systems", Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, USA, 2006
- [6] B. Georgantopoulos, T. Goedeme, S. Lounis, H. Papageorgiou, T. Tuytelaars, L. Van Gool, "Cross-media summarization in a retrieval setting", in Proceedings of the LREC 2006 workshop on "Crossing media for improved information access", Genoa, Italy, 2006
- [7] M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, K. Yamada, P. Langlais and A. Mauser , "Translating with Non-contiguous Phrases", In Proceedings of HLT/EMNLP, Vancouver, Canada, 2005