Multimodal Queries to Access Multimedia Information Sources: First Steps*

Ángel Martínez[1], Sara Lana Serrano[1], José L. Martínez-Fernández[1,2], and Paloma Martínez[2]

[1] DAEDALUS, Data, Decisions And Language, S.A.
Avda. de la Albufera, 321, 28031 Madrid, Spain
{amartinez,slana,jmartinez}@daedalus.es
[2] Advanced Databases Group, Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911 Leganés, Spain
{joseluis.martinez,paloma.martinez}@uc3m.es

Abstract. This position paper deals with queries beyond text, mixing several multimedia contents: audio, video, image and text. Search approaches combining some of these formats have been studied, including *query by example* techniques in situations where only one format is considered. It is worth mentioning that most of these research works do not deal with text content. A new approach to allow users introducing multimodal queries and exploring multimedia repositories is proposed. For this purpose, different ranked result lists must be combined to produce the final results shown for a given query. The main goal of this proposal is to reduce the semantic gap between low level features and high level concepts in multimedia contents. The use of qualitative data giving more relevance to text content along with machine learning methods to combine results of monomodal retrieval systems is proposed. Although it is too soon to show experimentation results, a prototype implementing the approach is under development and evaluation.

Keywords: multimodal queries, search engine, query by example, results lists combination, multimedia retrieval, relevance feedback.

1 Introduction

All Internet users have seen a growth of multimedia content available on the web. Services such as YouTube, Google Image, Flickr, Yahoo Video and others, have experienced an enormous development in the last few years. Besides, Google has recently announced a new product, Google TV[1], where online streaming videos will be part of the content offered to the user. On the other hand, traditional search engines

* This work has been partially supported by the Spanish Center for Industry Technological Development (CDTI, Ministry of Industry, Tourism and Trade), through the project Buscamedia (CEN-20091026). Authors would like to thank all Buscamedia partners for their knowledge and contribution.

[1] http://www.google.com/tv/

F. Alvarez and C. Costa (Eds.): UCMEDIA 2010, LNICST 60, pp. 35–40, 2012.

have evolved to deal with multimedia resources although, at the time of writing, it is not possible to use an image or a video in a query in any of these services. They are based on a manual or semi-automatic content tagging, but no content based retrieval seems to be applied. The work presented in this paper defines an approach to fill this gap, in order to consider queries containing text and/or audio and/or video and/or images, i.e., multimodal queries [13].

Let's suppose a scenario where the user has more data to express a query than some words. Suppose you have an MP3 file of a song and the name of one of the singers but you want to know the title of the song. Or imagine that you have a film in AVI format, a photo of the film director and you want to obtain his name to look for his filmography and the links to trailers of those films, if available. In this context, somehow related multimodal data is being used to represent a query and not only text should be retrieved but also audio, video and image contents, or even combinations of them. Some works like [13],[5],[10] have considered multimodal queries, pointing out the fact that only in recent years the information retrieval community is moving from dealing with one media at a time to combining more than one media in a retrieval process. In contrast, if some of this research is studied in detail, it is possible to see that the term multimedia usually excludes textual content or it refers to metadata related to the content. There is no a real mix among documents, web pages, videos, images and audio files.

2 Accessing Multimedia Sources with Multimodal Queries

If Information Retrieval Systems are considered, there is a semantic gap between content representations stored in indexes and concepts represented by those contents. The main objective is to find the theme or subject of a content which could be referenced in a query. But there is an issue that cannot be left aside, the theme or subject must be specified from the point of view of the searcher. For example, in a picture of happy children playing with a ball in the shore with a blue sea at the background, some searchers can be interested in finding photos with blue background and some others could be interested in photos where smiling children appear. Of course, this would be the ideal behavior of a multimedia retrieval system. In practice, it is not still possible to take into account every facet of some content in order to support any user need. For the moment, specific domain and restricted applications can be carried out with acceptable results, as demonstrated by evaluation forums such as TREC[2] or CLEF[3].

In order to deal with the semantic content of multimedia data, different abstraction levels are adopted [8][6]. The first one is formed by low level features, such as color, texture, shape and others for images; motion and direction, among others, for video; tone, spectral rile off or energy entropy for audio; and words, lemmas or n-grams for text. The second abstraction level is constituted by semantic representations of the contents, which are obtained by processing low level features and group together all

[2] http://trec.nist.gov
[3] http://www.clef-campaign.org

contents referring to some concept. Finally, there is a third level where intelligent reasoning is considered to infer some new knowledge from the semantic content; as an example, if some content is related with the semantic concept *smile*, then it can be part of a query related to *happiness*.

Although some of the previous research works deal with more than one format for the data, ([6] considers audio and video features to index the collection, [7], [11] deals with images and texts) it is difficult to find initiatives where a combination of formats is used to define a query. Most of the research works on multimedia information fusion deal with classifiers and combinations of them in different ways. [3] describes a combination of linear classifiers in order to reinforce semantics in the retrieval process. [2] successfully applies heterogeneous late fusion of independent retrieval models to retrieve data from a collection of annotated images.

Some evaluation tasks where queries are specified through text and image are defined in [7], but it is not the common situation. On the other hand, commercial search engines do not allow using other than text to define a query nor a combination of text, video, image and audio.

2.1 An Approach to Multimodal Retrieval

This section describes an innovative approach proposed in this research work to process multimodal queries in multimedia environments, see Fig. 1. The aim is to perform different search processes, in a somehow independent way, mixing obtained results for each process in a unique ranked list. The process to mix partial result lists is driven by semantics, which is obtained from all available formats (shapes recognition in images, concepts detection in video [6], etc.)

According to [9] the data fusion taken into account in this proposal could be considered into the ranking level model category, where several ranked streams are used to produce a consensus ranking, but also into the decision level model category, where the composing retrieval systems only provide a final decision (not a ranked list of probable results).

Seven different modules can be distinguished, which are described below:

- *Query by example*: This block includes common query by example techniques for each media. For media contents other than text, these algorithms will find similar contents for the different media, obtaining one result list for each. In the case of audio indexing, *query by humming* is also considered. Text queries consisting on a complete document would go through this path.
- *Text Search*: It covers all issues related to text based queries including: *free text* queries, formed by some words; *question answering*, providing accurate answers to questions; and *metadata queries*, asking about structured data defined in a semantic resource, including ontologies. The final goal is to provide relevant results for any kind of text-based search expression written by a user, but applying the right text search procedures depending on the content and structure of the query. A unique result list is obtained as a response to a text-based query.

- *Media Indexing*: Techniques for content-based indexing of image, audio and video data are part of this piece. [1] and [8] describe some of the methods to build indexes for each type of media. The output for this module is formed by three indexes, one for video, one for audio and another one for image contents.

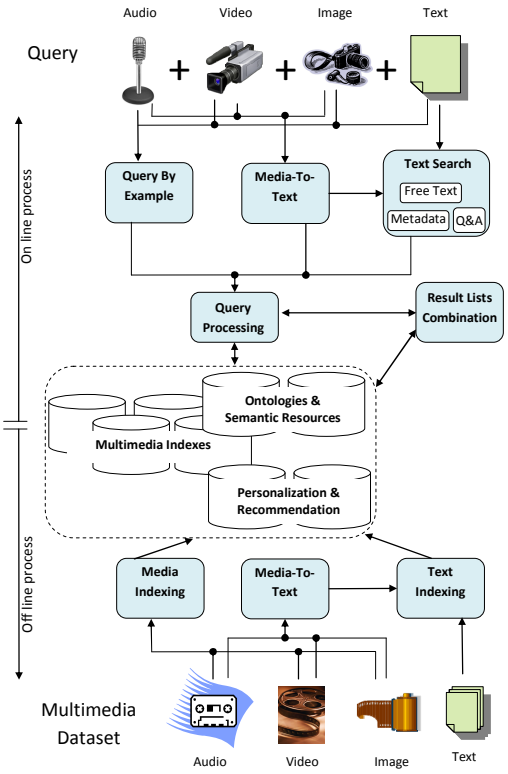

Fig. 1. Graphical representation of the proposed approach to multimodal query expression and retrieval

- *Text Indexing*: It covers all issues related to text based information retrieval, including the ability to provide semantic annotations for texts, employing semantic resources such as ontologies, thesaurus and so on. The output is a text index that will be used by the text search process.
- *Media-To-Text*: This module is in charge of obtaining textual representations for media contents, to make possible a text search process using these transformations. Of course, depending on the type of media, specific transformation methods are applied. Some of these methods have been described in [8]. Some of them have low quality ratios (as an example, only 60%-70% of the words are correctly recognized by ASR systems) which will be transmitted along the proposed retrieval process. One of the goals of this research work is to apply natural language processing tools as well as semantic web resources, such as ontologies, to reduce the effect of transformation errors.

- *Query Processing*: The right combination of the different media provided in the query is a key issue before running information retrieval processes on isolated media. For example, the following situations should be distinguished: if a given text must be found on an audio transcript or if it is the title of an audio file.
- *Results list combination*: According to previous component descriptions, several retrieval results lists are obtained, depending on the media used to represent the query. This final module is in charge of producing a unique result list, either using elements of different lists to build an entry for the final list or merging partial results lists in a final list or both. This module can also take into account ranking conditions, in order to pay more attention to results coming from one list or from another (for example, in some situations results coming from the video retrieval list could be more relevant than those coming from the text-based one). It is also possible to take into account personalization and recommendation data to consider users' preferences and information about previous experiences and interactions with other users. There exist a huge previous research work in the area, as can be read in [4] and [12].

3 Future Work

This paper identifies a shortage in current multimedia retrieval methods: it is difficult to find approaches dealing with all media: text, audio, video and image. Most of the previous research work seems coming from the video and audio retrieval communities and, if text is considered, it is supposed to come from transcripts or OCR processes. For this reason, a framework to consider all types of media is defined, based on the combination of partial result lists obtained by retrieval processes on isolated media. Besides, if multimodal queries are provided, the ability to integrate the different formats in queries on isolated media is also taken into account; semantic resources are applied for this purpose.

Of course, there is a lot of work to do, an initial implementation of the approach is under development, which must be tested under somehow standard evaluation frameworks, such as the ones developed in TREC or CLEF forums.

References

1. Bashir, F.I., Khanvilkar, S., Schonfeld, D., Khokhar, A.: Multimedia Systems: Content Based Indexing and Retrieval. In: Chen, W. (ed.) The Electrical Engineering Handbook, sec. 4, ch. 6. Academic Press (2004)
2. Escalante, H.J., Hérnandez, C.A., Sucar, L.E., Montes, M.: Late fusion of heterogeneous methods for multimedia image retrieval. In: Proceeding of the 1st ACM International Conference on Multimedia information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, pp. 172–179. ACM, New York (2008)
3. Joshi, D., Naphade, M., Natsev, A.: Semantics reinforcement and fusion learning for multimedia streams. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007, Amsterdam, The Netherlands, July 09-11, pp. 309–316. ACM, New York (2007)

4. Martínez-Santiago, F.: El problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: cálculo de la relevancia documental en dos pasos. Doctoral Thesis, UNED (2004)
5. Mittal, A.: An Overview of Multimedia Content-Based Retrieval Strategies, Informatica. International Journal of Computing and Informatics 30(3), 347–356 (2006)
6. Naphade, M.R., Kristjansson, T., Frey, B., Huang, T.S.: Probabilistic Multimedia Objects Multijects: A novel Approach to Indexing and Retrieval in Multimedia Systems. In: Proc. IEEE International Conference on Image Processing, vol. 3, pp. 536–540 (1998)
7. Nowak, S., Dunker, P.: Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsikrika, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 94–109. Springer, Heidelberg (2010), http://www.clef-campaign.org/2009/working_notes/Overview_VCDT.pdf
8. PetaMedia: State of the art report. PetaMedia Deliverable D 5.1 (2008)
9. Poh, N., Kittler, J.: Multimodal Information Fusion, Multimodal Signal Processing: Theory and Applications for Human-Computer Interaction. In: Thiran, J.-P., Bourlard, H., Marques, F. (eds.) to appear in Elsevier Science (2009) ISBN-13: 978-0-12-374825-6
10. Olsson, J.S., Oard, D.W.: Combining Speech Retrieval Results with Generalized Additive Models. In: Proceedings of ACL 2008: HLT, Association for Computational Linguistics, pp. 461–469 (2008)
11. Tollari, S., Detyniecki, M., Marsala, C., Fakeri-Tabrizi, A., Amini, M., Gallinari, P.: Exploiting Visual Concepts to Improve Text-Based Image Retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 701–705. Springer, Heidelberg (2009)
12. Wiguna, W., Fernández-Tébar, J., García-Serrano, A.: Using a Fuzzy Model for Combining Search Results from Different Information Sources to Build a Metasearch Engine. In: Computational Intelligence, Theory and Applications, pp. 325–334 (2006), doi:10.1007/3-540-34783-6_34
13. Yan, R.: Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval. PhD thesis, Carnegie Mellon University (2006)