Towards the Creation of a Unified Framework for Multimodal Search and Retrieval

Apostolos Axenopoulos, Petros Daras, and Dimitrios Tzovaras

Centre for Research and Technology Hellas, Informatics and Telematics Institute,
6th Km Charilaou-Thermi Road Rd., 57001, Thermi, Thessaloniki, Greece
{axenop,daras,tzovaras}@iti.gr

Abstract. In this paper, a novel framework for search and retrieval of multimodal content is introduced as part of the EU-funded project I-SEARCH. The main objective of I-SEARCH is to create a unified framework for multimodal content search, i.e. to retrieve content of any media type (text, 2D images, video, audio and 3D) by using as query any of the above media, along with real-world information, expressive and social cues. The outcome will be a highly user-centric search engine, able to deliver to the end-users only the content of interest, satisfying their information needs and preferences, which is expected to significantly improve end-user's experience. The paper will present the concept of I-SEARCH, as well as its major scientific advances.

Keywords: Multimodal Content Search and Retrieval, user-centric search engine, RUCoD.

1 Introduction

Multimedia content, which is available over the Internet, is increasing at a rate faster than the respective increase of computational power and storage capabilities. Due to the widespread availability of digital recording devices, improved modeling tools, advanced scanning mechanisms as well as display and rendering devices, even over mobile environments, users are more and more empowered to live a more immersive and unforgettable experience with last-generation digital media, through experiencing audiovisual content. It is therefore now possible for users to rapidly move from a mainly textual-based to a media-based "embodied" Internet, where rich audiovisual content (images, sound, videos), 3D representations (avatars) and reconstructions, virtual and mixed reality worlds, serious games, life-logging applications, multimodal yet affective utterances (gestures, facial expressions, eye movements, etc.) become a reality.

This growth of popularity of media is not accompanied by the rapid development of media search technologies. The most popular media services in the Web are typically limited to textual search [1, 2]. However, the last years, significant efforts have been devoted, mainly by the European research community, for achieving

F. Alvarez and C. Costa (Eds.): UCMEDIA 2010, LNICST 60, pp. 26–34, 2012.

content-based search of images [6, 8, 9, 10], video [7, 11, 12, 13, 14] and 3D models [3, 15, 16, 17, 18]. Same endeavors are also lately noticed by the big players in these fields (Google image, Google SketchUp [4]).

Despite the significant achievements in multimedia search technologies, the existing solutions still lack several important features, which could guarantee high-quality search services and improved end-user experience. These features are listed below:

- A unified framework for multimodal content search and retrieval: this will enable users express their queries in any form most suitable for them, retrieve content in various forms providing the user with a complete view of the retrieved information and interact with the content using the most suitable modality for the particular user and under the specific context each time.
- Sophisticated mechanisms for interaction with content: these will exploit at best the social and collaborative behavior of users interacting with the content, which will enable them to better express what they want to retrieve.
- Efficient presentation of the retrieved results: this will optimally present to the user the most relevant results according to the query and the user preferences.

Towards this direction, the I-SEARCH project [5] aims to provide a novel unified framework for multimedia and multimodal content indexing, search and retrieval. The I-SEARCH framework will be able to handle specific types of multimedia (text, 2D image, sketch, video, 3D objects, audio and combination of the above) and support multimodal interaction means (gestures, face expressions, eye movements) along with real world information (GPS, temperature, time, weather sensors, RFID objects,), which can be used as queries and retrieve any available relevant content of any of the aforementioned types and from any end-user access device. Furthermore I-SEARCH will be able to integrate even non-verbal yet implicit, emotional cues, and social descriptors, in order to better express what the user wants to retrieve.

The proposed search engine is expected to be highly user-centric in the sense that only the content of interest will be delivered to the end-users, satisfying their information needs and preferences, which is expected to dramatically improve end-user experience. Furthermore, I-SEARCH introduces the use of advanced visual analytic technologies for search results presentation in order to facilitate their fast and easy interpretation and also to support optimal results presentation under various contexts (i.e. user profile, end-user terminal, available network bandwidth, interaction modality preference, etc.).

In the following, a description of the I-SEARCH concept and the project’s main objectives is initially provided, followed by the major scientific advances proposed by I-SEARCH, such as the Rich Unified Content Description (RUCoD), Multimodal Annotation Propagation, Multimodal Interaction and Visualization.

2 I-SEARCH Objectives and Conceptual Architecture

The aim of the I-SEARCH project is the development of the first search engine able to handle a wide range of specific types of multimedia and multimodal content (text, 2D image, sketch, video, 3D objects, audio and combination of the above), which can be used as queries and retrieve any available relevant content of any of the aforementioned types.

2.1 I-SEARCH Objectives

Towards the realization of the first multimodal search engine, I-SEARCH introduces the concept of Rich Unified Content Description (RUCoD). RUCoD will consist of a multi-layered structure (from low-level to high-level descriptors), which will integrate content's geometrical, topological, temporal, multisensory and multimodal information and meta-tags connected with the intrinsic properties of the content (static features such as shape, colour, texture, dimension, etc.), dynamic properties (temporal descriptors, how it behaves, in which activities it is normally used, who uses it, etc.), non-verbal expressive and emotional descriptors, social descriptors (how content is related to users, social/collaborative use of the content), content descriptors as for the behavior of the humans included in the (visual and vocal) content, descriptors for users' behavior as for how content is intended to be elaborated and manipulated, individually or socially. Further, novel multimodal annotation propagation algorithms will be developed, i.e. annotating information of one form/modality using information describing other forms/modalities of the same object.

Furthermore, the development of intelligent content interaction mechanisms is proposed, so that only the content of interest will be delivered to the users. This will be achieved by providing users with natural and expressive and multimodal interfaces as well as through personal and social-based relevance (including recommendation-based) feedback mechanisms. In this context, social and collaborative behavior of users interacting with the content will be exploited at best, which will help users to better express what they want to retrieve.

Finally, I-SEARCH will provide a novel way for presentation of the multimodal data retrieved by the search engine, by utilizing visual analytics technologies. The unified content representation will be exploited for generating advanced reasoning algorithms to drive the visualization of the I-SEARCH RUCoD-compliant content. Visual Analytics will provide an analytical process for presenting the search results in the optimal way to aid the user in finding the result that optimally matches the query in a fast and efficient way.

2.2 I-SEARCH Architecture

The overall conceptual architecture of I-SEARCH is shown in the figure below:

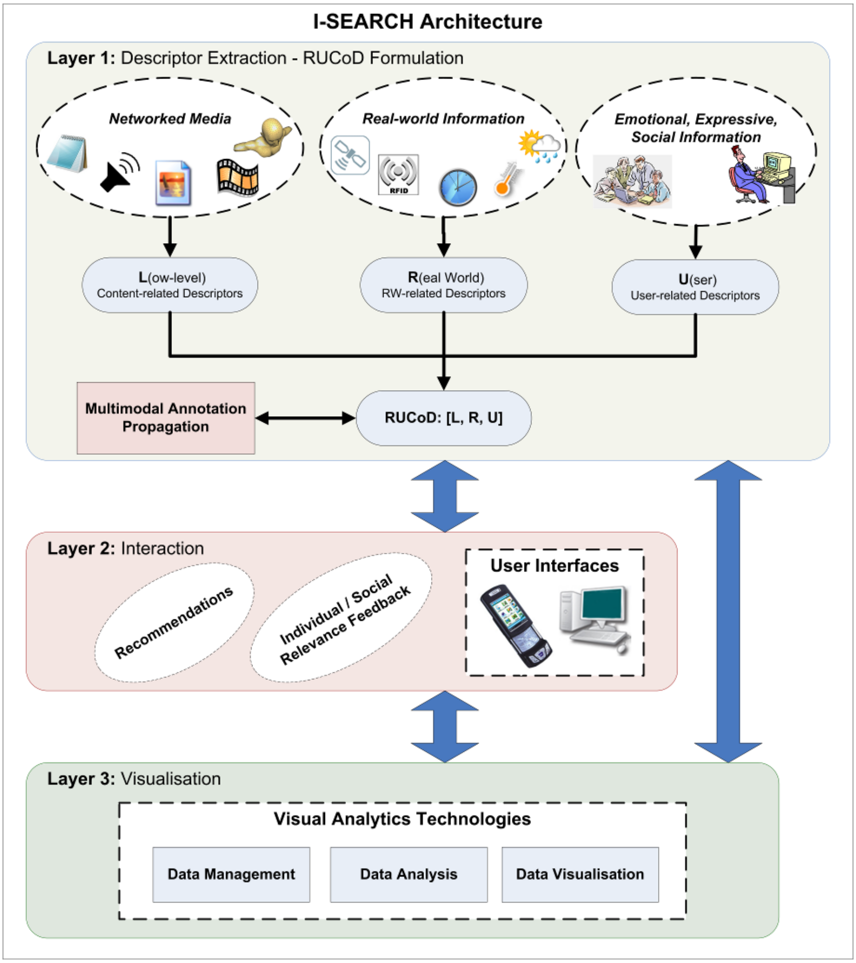
I-SEARCH Architecture
Layer 1: Descriptor Extraction - RUCoD Formulation
Networked Media
Real-world Information
Emotional, Expressive, Social Information
RFID
L(ow-level) Content-related Descriptors
R(eal World) RW-related Descriptors
U(ser) User-related Descriptors
Multimodal Annotation Propagation
RUCoD: [L, R, U]
Layer 2: Interaction
Recommendations
Individual / Social Relevance Feedback
User Interfaces
Layer 3: Visualisation
Visual Analytics Technologies
Data Management
Data Analysis
Data Visualisation

Fig. 1. The I-SEARCH Architecture

The three distinct layers of the I-SEARCH conceptual architecture are analyzed in the sequel.

Layer 1 (Descriptor Extraction – RUCoD Formulation): This layer includes all the descriptor extraction mechanisms, which lead to the novel RUCoD descriptor. Three main types of descriptors constitute the unified RUCoD descriptor:

- L(ow-level), content-related descriptors: L-descriptors are directly extracted from the networked media (text, audio, image, video and 3D), by utilizing low-level feature extraction mechanisms.
- R(eal world) - related descriptors: R-descriptors refer to the real world information captured from various sensors integrated in the environment. Such sensors include GPS, temperature, time, weather sensors, RFID objects, etc.

- U(ser), user-related descriptors: U-descriptors include non-verbal expressive, emotional and social descriptors. They are called user-related because they describe the user behavior associated with the content.

Layer 2 (Interaction): This layer involves the novel sophisticated mechanisms for interaction with content. It consists of the following three modules:

- Recommendations module: it deals with the feedback added by experts that are the most appreciated in a community upon a define topic.
- Relevance Feedback module: relevance feedback captures the user satisfaction upon retrieval of results and can be either individual or social.
- User interfaces, available for several types of end-user devices.

Layer 3 ((Visualization): This layer offers the mechanisms for efficient presentation of the retrieved results. It consists of Visual Analytics technologies, which provide an efficient way of presenting the retrieved data with respect to:

- Data management.
- Data analysis.
- Data visualization.

3 I-SEARCH Innovative Components

I-SEARCH aims to provide new insight into the nature of next generation search engines for audiovisual content. Its main innovative aspects are analyzed below.

3.1 A Rich Unified Content Description (RUCoD)

I-SEARCH will handle several types of multimedia content, along with real-world and user-related information. In order to describe all this information in a uniform way, the concept of the Rich Unified Content Description (RUCoD) is introduced. Figure 2 presents the generic RUCoD representation of any multimedia, real world and user-related information supported by the I-SEARCH system.

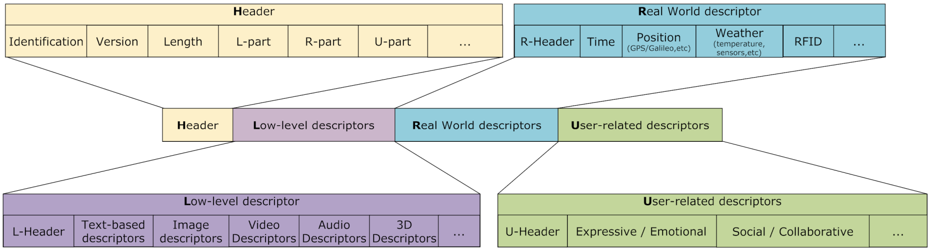

Fig. 2. The RUCoD Format

The major advantage of RUCoD is that it enables to easily link information from different media, e.g., a music piece can be linked with video, text documents, and possible further related material, a 3D object can be linked with text, 2D images, video and so on. In this way, search and retrieval of multimedia content is treated in a unified manner, i.e. users will be able to use as query any of the RUCoD-supported media types (text, 2D image, sketch, video, 3D objects and audio) and retrieve any available relevant content of any of the aforementioned types.

Beyond the intrinsic properties of the multimedia documents, RUCoD will also support descriptors related to expressive/emotional and social cues, multisensory and real world information. Whereas traditional content-based retrieval systems only extract and take into account either low-level features of the media data or descriptors of user's preferences, the search engine featured in I-SEARCH will also consider the multisensory information (user's gesture, mood, time, location, etc.) associated with the content.

3.2 Multimodal Annotation Propagation

Multimodal annotation propagation deals with annotating information of one form using information describing other forms of the object. More specifically, the automatic multimodal annotation propagation will be applied to non-available content and to non-available description of the RUCoD, which means that if some of the types of content describing an object are not available, or the user hasn't added metadata, the system will learn from past history/use of the object itself or similarity with other objects and will propagate this data to the empty cells of the RUCoD. Figure 3 shows an example of multimodal annotation propagation (the user adds metadata to an image described in RUCoD and the system propagates this information to annotate also all other types of information contained in the database that are linked to the specific image).

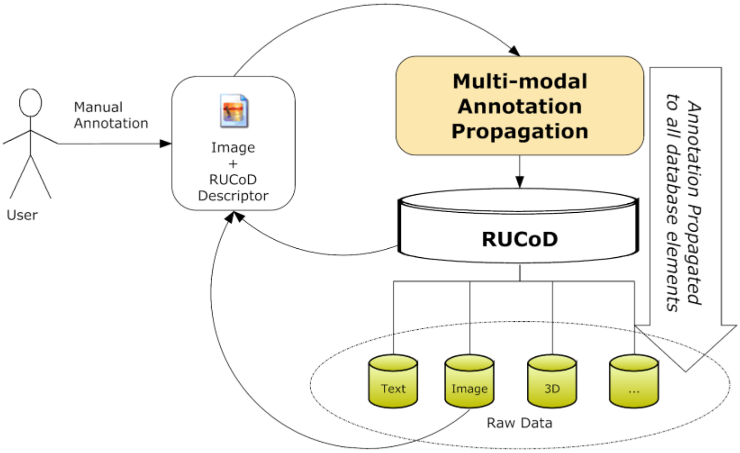
Manual Annotation
User
Image + RUCoD Descriptor
Multi-modal Annotation Propagation
RUCoD
Annotation Propagated to all database elements
Text
Image
3D
...
Raw Data

Fig. 3. Multimodal annotation propagation example

3.3 Multimodal Interaction

Multimodal interaction refers to novel, sophisticated mechanisms and interfaces for embodied interaction with content, with a main focus on multimodality (including non-verbal, full-body), context-awareness, and emotional/expressive interaction. Natural and expressive interfaces will be based on analysis of user's movement and gesture.

A preliminary block diagram of the interaction module of I-SEARCH is shown in Figure 4. The interoperability of the module with the remaining I-SEARCH components is assured through the representation of the high-, mid- and low-level interaction descriptors in RUCoD. The interaction manager module caters for interfacing the I-SEARCH interaction module with the networked media descriptor extraction and the visualisation modules of I-SEARCH.

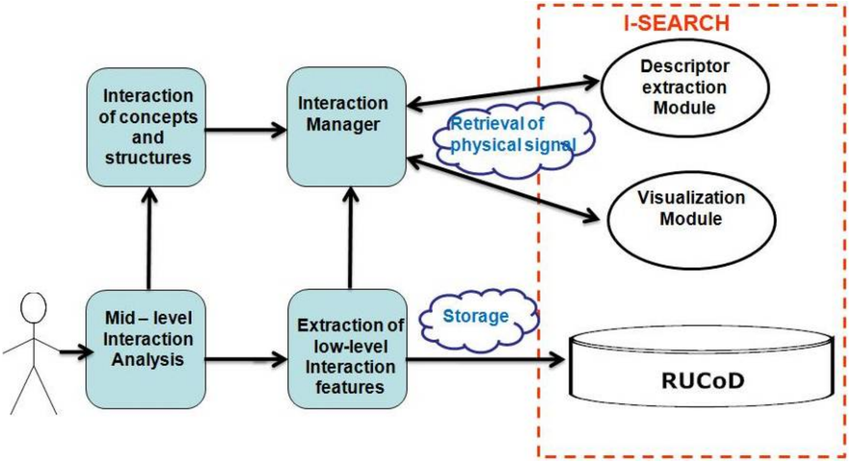
I-SEARCH
Interaction of concepts and structures
Interaction Manager
Descriptor extraction Module
Retrieval of physical signal
Visualization Module
Mid – level Interaction Analysis
Extraction of low-level Interaction features
Storage
RUCoD

Fig. 4. Preliminary block diagram of the interaction module of I-SEARCH

3.4 Visualization

I-SEARCH introduces the use of visual analytics technologies for enhancing the presentation layer of search engines. The visual technologies will be built upon a framework that will provide analytic reasoning methods of search results. Novel data representation and transformation mechanisms will convert the RUCoD format of search results to structured forms that will enable the visualization. Finally, adaptive visual presentation mechanisms will be provided which will support the presentation of search results under various contexts utilizing various information visualization technologies. Figure 5 shows an example of the visual analytics approach that will be followed in I-SEARCH for results presentation. As shown, the search results are analyzed by the reasoning module in order to extract relevant knowledge that is then used for the transformation of the results to the appropriate format and the subsequent presentation of the results using the most appropriate visualization method.

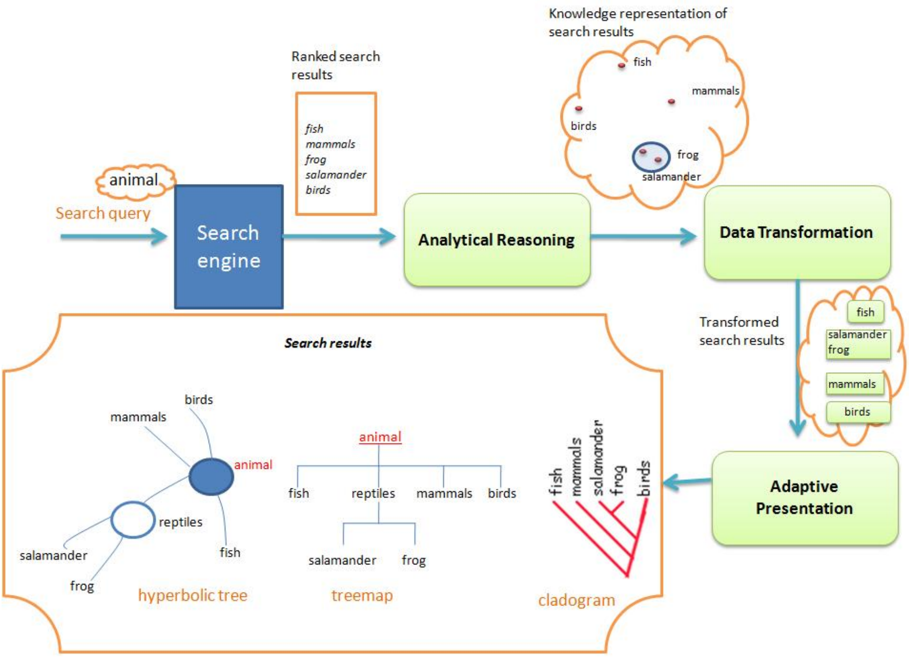

Fig. 5. The Visual Analytics Approach

4 Conclusions

In this paper, a novel framework to multimedia search engines has been introduced. Through the realization of the I-SEARCH project, numerous cutting-edge technologies are going to revolutionize the way multimedia search over the Internet is performed. After an overview of the conceptual architecture of I-SEARCH, an analysis of its main innovative components followed. These include specification of a Rich Unified Content Description (RUCoD), Multimodal Annotation Propagation, Multimodal Interaction mechanisms and novel results presentation capabilities. The I-SEARCH framework will be able to handle at the same time several types of multimedia (text, 2D image, sketch, video, 3D objects, audio and combination of the above) and multimodal content (gestures, face expressions, eye movements) along with real world information (GPS, temperature, time, weather sensors, RFID objects, etc.). Finally, the search engine will be dynamically adapted to end-user's device, which will vary from a simple mobile phone to a high-performance PC.

References

1. Flickr Photo Sharing, `http://www.flickr.com/`
2. Internet Archive, `http://www.archive.org/`
3. VICTORY project official website, `http://www.victory-eu.org:8080/victory`

4. Google SketchUp, `http://sketchup.google.com/`
5. I-SEARCH project official website, `http://www.isearch-project.eu/`
6. The Pharos Project, `http://www.pharosproject.net/`
7. The VITALAS Project, `http://vitalas.ercim.org/`
8. Worring, M., Gevers, T.: Interactive retrieval of color images. Int. J. Image Graph 1(3), 387–414 (2001)
9. Kokare, M., Biswas, P.K., Chatterji, B.N.: Texture image retrieval using new rotated complex wavelet filters. IEEE Transactions on Systems, Man, and Cybernetics, Part B 35(6), 1168–1178 (2005)
10. Attalla, E., Siy, P.: Robust shape similarity retrieval based on contour segmentation polygonal multiresolution and elastic matching. Pattern Recognition 38(12), 2229–2241 (2005)
11. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int. Journal of Computer Vision 60(2) (2004)
12. Joly, A., Frélicot, C., Buisson, O.: Feature statistical retrieval applied to content-based copy identification. In: Proc. of Int. Conf. on Image Processing (2004)
13. Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujemaa, N., Stentiford, F.: Video Copy Detection: a Comparative Study. In: Proc. of the ACM International Conference on Image and Video Retrieval (2007)
14. Laptev, I., Lindeberg, T.: Local Descriptors for Spatio-Temporal Recognition. In: Proc. of ECCV (2004)
15. Zarpalas, D., Daras, P., Axenopoulos, A., Tzovaras, D., Strintzis, M.G.: 3D Model Search and Retrieval Using the Spherical Trace Transform., EURASIP Journal on Advances in Signal Processing, Article ID 239110 (2007)
16. Papadakis, P., Pratikakis, I., Perantonis, S., Theoharis, T.: Efficient 3D shape matching and retrieval using a concrete radialized spherical projection representation. Pattern Recognition 40(9), 2437–2452 (2007)
17. Daras, P., Axenopoulos, A.: A 3D Shape Retrieval Framework Supporting Multimodal Queries. International Journal of Computer Vision (July 2009), doi:10.1007/s11263-009-0277-2
18. Ohbuchi, R., Osada, K., Furuya, T., Banno, T.: Salient local visual features for shape-based 3D model retrieval. In: Proc. of the IEEE International Conference on Shape Modeling and Applications (SMI 2008), pp. 93–102 (2008)