I-SEARCH – A Multimodal Search Engine based on Rich Unified Content Description (RUCoD)

Thomas Steiner
Google Germany GmbH
tomac@google.com

Lorenzo Sutton
Accademia Naz. di S. Cecilia
l.sutton@santacecilia.it

Sabine Spiller
EasternGraphics GmbH
sabine.spiller@easterngraphics.com

Marilena Lazzaro, Francesco Nucci, Vincenzo Croce
Engineering
{firstname.lastname}@eng.it

Alberto Massari, Antonio Camurri
University of Genova
alby@infomus.dist.unige.it, antonio.camurri@unige.it

Anne Verroust-Blondet, Laurent Joyeux
INRIA Rocquencourt
{anne.verroust, laurent.joyeux}@inria.fr

Jonas Etzold, Paul Grimm
Hochschule Fulda
{jonas.etzold, paul.grimm}@hs-fulda.de

Athanasios Mademlis, Sotiris Malassiotis, Petros Daras
CERTH/ITI
{mademlis, malasiot, daras}@iti.gr

Apostolos Axenopoulos, Dimitrios Tzovaras
CERTH/ITI
{axenop, tzovaras}@iti.gr

ABSTRACT
In this paper, we report on work around the I-SEARCH EU (FP7 ICT STREP) project whose objective is the development of a multimodal search engine. We present the project's objectives, and detail the achieved results, amongst which a Rich Unified Content Description format.

Categories and Subject Descriptors
H.3.4 [**Information Systems**]: Information Storage and Retrieval—*World Wide Web*; H.3.5 [**Online Information Services**]: Web-based services

Keywords
Multimodality, Rich Unified Content Description, IR

1. INTRODUCTION
1.1 Motivation
Since the beginning of the age of Web search engines in 1990, the search process is associated with a text input field. From the first search engine, Archie [7], to state-of-the-art search engines like WolframAlpha [2], this fundamental input paradigm has not changed. In a certain sense the search process has been revolutionized on mobile devices through the addition of voice input support like Apple's Siri [1] for iOS, Google's Voice Actions [3] for Android, and through Voice Search [4] for desktop computers. Support for the human voice as an input modality is mainly driven by shortcomings of (mobile) keyboards. One modality, text, is simply replaced by another, voice. However, what is still missing is a truly multimodal search engine. If the searched-for item is slow, sad, minor scale piano music, the best input modalities might be to just upload a short sample ("audio") and an unhappy smiley face or a sad body expressive gesture ("emotion"). When searching for the sound of Times Square, New York, the best input modalities might be the coordinates ("geolocation") of Times Square and a photo of a yellow cab ("image"). The outlined search scenarios are of very different nature, and even for human beings it is not easy to find *the* correct answer, let alone that such answer exists for each scenario. With I-SEARCH, we thus strive for a paradigm shift; away from textual keyword search, towards a more explorative multimodality-driven search experience.

1.2 Background
It is evident that for the outlined scenarios to work, a significant investment in describing the underlying media items is necessary. Therefore, in [6], we have first introduced the concept of so-called *content objects*, and second, a description format named *Rich Unified Content Description (RUCoD)*. Content objects are rich media presentations, enclosing different types of media, along with real-world information and user-related information. *RUCoD* provides a uniform descriptor for all types of content objects, irrespective of the underlying media and accompanying information.

1.3 Involved Partners and Paper Structure
The involved partners are CERTH/ITI (Greece), JCP-Consult (France), INRIA Rocquencourt (France), ATC (Greece), Engineering Ingegneria Informatica S.p.A. (Italy), Google (Ireland), University of Genoa (Italy), Exalead (France), University of Applied Sciences Fulda (Germany), Accademia Nazionale di Santa Cecilia (Italy), and EasternGraphics (Germany). In this paper, we give an overview on the I-SEARCH project so far. In Section 2, we outline the general objectives of I-SEARCH. Section 3 highlights significant achievements. We describe the details of our system in Section 4. Relevant related work is shown in Section 5. We conclude with an outlook on future work and perspectives of this EU project.

2. PROJECT GOALS

With the I-SEARCH project, we aim for the creation of a multimodal search engine that allows for both multimodal in- and output. Supported input modalities are *audio*, *video*, *rhythm*, *image*, *3D object*, *sketch*, *emotion*, *social signals*, *geolocation*, and *text*. Each modality can be combined with all other modalities. The graphical user interface (GUI) of I-SEARCH is not tied to a specific class of devices, but rather dynamically adapts to the particular device constraints like varying screen sizes of desktop and mobile devices like cell phones and tablets. An important part of I-SEARCH is a *Rich Unified Content Description (RUCoD)* format that consists of a multi-layered structure that describes low and high level features of content and hence allows this content to be searched in a consistent way by querying *RUCoD* features. Through the increasing availability of location-aware capture devices such as digital cameras with GPS receivers, produced content contains exploitable real-world information that form part of *RUCoD* descriptions.

3. PROJECT RESULTS

3.1 Rich Unified Content Description

In order to describe content objects consistently, a *Rich Unified Content Description (RUCoD)* format was developed. The format is specified in form of XML schemas and available on the project website[1]. The description format has been introduced in full detail in [6], Listing 1 illustrates *RUCoD* with an example.

3.2 Graphical User Interface

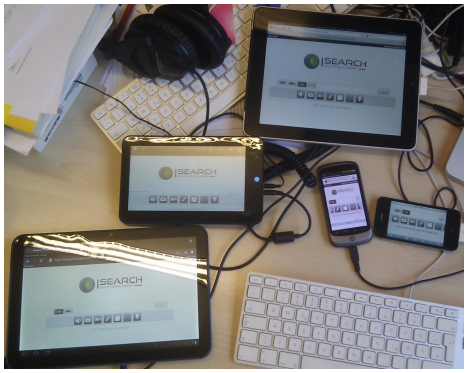
The I-SEARCH graphical user interface (GUI) is implemented with the objective of sharing one common code base for all possible input devices (Subfigure 1b shows mobile devices of different screen sizes and operating systems). It uses a JavaScript-based component called *UIIFace* [8], which enables the user to interact with I-SEARCH via a wide range of modern input modalities like touch, gestures, or speech. The GUI also provides a WebSocket-based collaborative search tool called *CoFind* [8] that enables users to search collaboratively via a shared results basket, and to exchange messages throughout the search process. A third component called *pTag* [8] produces personalized tag recommendations to create, tag, and filter search queries and results.

3.3 Video and Image

The video mining component produces a video summary as a set of recurrent image patches to give a visual representation of the video to the user. These patches can be used to refine search and/or to navigate more easily in videos or images. For this purpose, we use a technique of Letessier *et al.* [13] consisting of a weighted and adaptive sampling strategy aiming to select the most relevant query regions from a set of images. The images are the video key frames and a new clustering method is introduced that returns a set of suggested object-based visual queries. The image search component performs approximate vector search on either local or global image descriptors to speed up response time on large scale databases.

[1] *RUCoD* XML schemas: `http://www.isearch-project.eu/isearch/RUCoD/RUCoD_Descriptors.xsd` and `http://www.isearch-project.eu/isearch/RUCoD/RUCoD.xsd`.

(a) Multimodal query consisting of *geolocation*, *video*, *emotion*, and *sketch* (in progress).



(b) Running on some mobile devices with different screen sizes and operating systems.

(c) Treemap results visualization showing different clusters of images.

Figure 1: I-SEARCH graphical user interface.

```
<RUCoD>
  <Header>
    <ContentObjectType>
      Multimedia Collection
    </ContentObjectType>
    <ContentObjectName xml:lang="en-US">
      AM General Hummer
    </ContentObjectName>
    <ContentObjectID/>
    <ContentObjectVersion>1</ContentObjectVersion>
    <ContentObjectCreationInformation>
      <Creator>
        <Name>CoFetch Script</Name>
      </Creator>
    </ContentObjectCreationInformation>
    <Tags>
      <MetaTag name="UserTag" type="xsd:string">
        Hummer
      </MetaTag>
    </Tags>
    <ContentObjectTypes>
      <MultimediaContent type="Object3d">
        <FreeText>Not Mirza's model.</FreeText>
        <MediaName>2001 Hummer H1</MediaName>
        <MetaTag name="UserTag" type="xsd:string">
          Hummer
        </MetaTag>
        <MediaLocator>
          <MediaUri>
            http://sketchup.google.com/[...]
          </MediaUri>
          <MediaPreview>
            http://sketchup.google.com/[...]
          </MediaPreview>
        </MediaLocator>
        <MediaCreationInformation>
          <Author>
            <Name>ZXT</Name>
          </Author>
          <Licensing>
            Google 3D Warehouse License
          </Licensing>
        </MediaCreationInformation>
        <Size>1840928</Size>
      </MultimediaContent>
      <RealWorldInfo>
        <MetadataUri filetype="rwml">
          AM_General_Hummer.rwml
        </MetadataUri>
      </RealWorldInfo>
    </ContentObjectTypes>
  </Header>
</RUCoD>
```

Listing 1: Sample *RUCoD* snippet (namespace declarations and some details removed for legibility reasons).

3.4 Audio and Emotions

I-SEARCH includes the extraction of expressive and emotional information conveyed by a user to build a query, and the possibility to build queries resulting from a social verbal or non-verbal interaction among a group of users. The I-SEARCH platform includes algorithms for the analysis of non-verbal expressive and emotional behavior expressed by full body gestures, for the analysis of the social behavior in a group of users (e.g., synchronization, leadership), and methods to extract real-world data.

3.5 3D Objects

The 3D object descriptor extractor is the component for extracting low level features from 3D objects and is invoked during the content analytics process. More specifically, it takes as input a 3D object and returns a fragment of low level descriptors fully compliant with the *RUCoD* format.

3.6 Visualization

I-SEARCH uses sophisticated information visualization techniques that support not only querying information, but also browsing techniques for effectively locating relevant information. The presentation of search results is guided by analytic processes such as clustering and dimensionality reduction that are performed after the retrieval process and intend to discover relations among the data. This additional information is subsequently used to present the results to the user by means of modern information visualization techniques such as treemaps, an example of such can be seen in Subfigure 1c. The visualization interface is able to seamlessly mix results from multiple modalities

3.7 Orchestration

Content enrichment is an articulated process requiring the orchestration of different workflow fragments. In this context, a so-called *Content Analytics Controller (CAC)* was developed, which is the component in charge of orchestrating the content analytics process for content object enrichment via low level description extraction. Orchestration relies on content object media and related info, handled by a *RUCoD* authoring tool.

3.8 Content Providers

The first content provider in the I-SEARCH projects holds an important Italian ethnomusicology archive. The partner makes available all of its digital content to the project as well as its expertise for the development of requirements and use cases related to music. The second content provider is a software vendor for the furniture industry with a big catalogue of individually customizable pieces of furniture. Both partners are also actively involved in user testing and the overall content collection effort for the project via deployed Web services that return their results in the *RUCoD* format.

4. SYSTEM DEMONSTRATION

With I-SEARCH being in its second year, there is now some basic functionality in place. We maintain a bleeding-edge demonstration server[2], and have recorded a screencast[3] that shows some of the interaction patterns. The GUI runs on both mobile and desktop devices, and adapts dynamically to the available screen real estate, which, especially on mobile devices, can be a challenge. Supported input modalities at this point are *audio*, *video*, *rhythm*, *image*, *3D object*, *sketch*, *emotion*, *geolocation*, and *text*. For *emotion*, an innovative emotion slider open source solution [14] was adapted to our needs. The GUI supports *drag and drop* user interactions and we aim for supporting low level device access for audio and video uploads. For *3D objects*, we support Web GL powered 3D views of models. *Text* can be entered via

[2]Demonstration: `http://isearch.ai.fh-erfurt.de/`

[3]Screencast: `http://youtu.be/-chzjEDcMXU`

speech input based on the WAMI toolkit [10], or via keyboard. First results can be seen upon submitting a query, and the visualization component allows to switch back and forth between different views.

5. RELATED WORK

We start covering related work with a differentiation of terms. *Multimodal search* can be used in two senses; (i), in the sense of multimodal result output based on unimodal query input, and (ii), in the sense of multimodal result output *and* multimodal query input. We follow the second definition, i.e., require the query input interface to allow for multimodality.

An interesting multimodal search engine was developed in the scope of the PHAROS project [5]. With the initial query being keyword-based, content-based or a combination of these, the search engine allows for refinement in form of facets, like location, that can be considered modalities. I-SEARCH develops this concept one step further by supporting multimodality from the beginning. In [9], Rahn Frederick discusses the importance of multimodality in search-driven on-device portals, i.e., handset-resident mobile applications, often preloaded, that enhance the discovery and consumption of endorsed mobile content, services, and applications. Consumers can navigate on-device portals by searching with text, voice, and camera images. Rahn Frederick's article is relevant, as it is specifically focused on mobile devices, albeit the scope of I-SEARCH is broader in the sense of also covering desktop devices. In a W3C Note [12], Larson *et al.* describe a multimodal interaction framework, and identify the major components for multimodal systems. The multimodal interaction framework is not an architecture *per se*, but rather a level of abstraction above an architecture and identifies the markup languages used to describe information required by components and for data flows among components. With Mudra [11], Hoste *et al.* present a unified multimodal interaction framework supporting the integrated processing of low level data streams as well as high level semantic inferences. Their architecture is designed to support a growing set of input modalities as well as to enable the integration of existing or novel multimodal fusion engines. Input fusion engines combine and interpret data from multiple input modalities in a parallel or sequential way. I-SEARCH is a search engine that captures modalities sequentially, however, processes them in parallel.

6. FUTURE WORK AND CONCLUSION

The efforts in the coming months will focus on integrating the different components. Interesting challenges lie ahead with the presentation of results and result refinements. In order to test the search engine, a set of use cases has been compiled that covers a broad range of modalities, and combinations of such. We will evaluate those use cases and test the results in user studies involving customers of the industry partners in the project.

In this paper, we have introduced and motivated the I-SEARCH project and have shown the involved components from the different project partners. We have then presented first results, provided a system demonstration, and positioned our project in relation to related work in the field. I-SEARCH is now in a decisive phase of the project, where the components function in isolation, however, need to be integrated in order to work well in orchestration with the entire I-SEARCH framework. The coming months will be fully dedicated to the integration efforts and we are optimistic to successfully evaluate the set of use cases in the project's final year.

7. ACKNOWLEDGMENTS

This work was partially supported by the European Commission under Grant No. 248296 FP7 I-SEARCH project.

8. REFERENCES

[1] Apple iPhone 4S – Ask Siri to help you get things done. Avail. at `http://www.apple.com/iphone/features/siri.html`.

[2] WolframAlpha. Avail. at `http://www.wolframalpha.com/`.

[3] Google Voice Actions for Android, 2011. Avail. at `http://www.google.com/mobile/voice-actions/`.

[4] Google Voice Search – Inside Google Search, 2011. Avail. at `http://www.google.com/insidesearch/voicesearch.html`.

[5] A. Bozzon, M. Brambilla, Fraternali, et al. PHAROS: an audiovisual search platform. In *Proceedings of the 32nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '09, pages 841–841, New York, NY, USA, 2009. ACM.

[6] P. Daras, A. Axenopoulos, V. Darlagiannis, et al. Introducing a Unified Framework for Content Object Description. *Int. Journal of Multimedia Intelligence and Security (IJMIS)*. Accepted for publication. Avail. at `http://www.lsi.upc.edu/~tsteiner/papers/2010/rucod-specification-ijmis2010.pdf`, 2010.

[7] A. Emtage, B. Heelan, and J. P. Deutsch. Archie, 1990. Avail. at `http://archie.icm.edu.pl/archie-adv_eng.html`.

[8] J. Etzold, A. Brousseau, P. Grimm, and T. Steiner. Context-aware Querying for Multimodal Search Engines. In *18th Int. Conf. on MultiMedia Modeling (MMM 2012), Klagenfurt, Austria*, January 4-6, 2012. http://www.lsi.upc.edu/ tsteiner/papers/2012/context-aware-querying-mmm2012.pdf.

[9] G. R. Frederick. Just Say "Britney Spears": Multi-Modal Search and On-Device Portals, Mar. 2009. Avail. at `http://java.sun.com/developer/technicalArticles/javame/odp/multimodal-odp/`.

[10] A. Gruenstein, I. McGraw, and I. Badr. The WAMI Toolkit for Developing, Deploying, and Evaluating Web-accessible Multimodal Interfaces. In *ICMI*, pages 141–148, 2008.

[11] L. Hoste, B. Dumas, and B. Signer. Mudra: A Unified Multimodal Interaction Framework. 2011. Avail. at `http://wise.vub.ac.be/sites/default/files/publications/ICMI2011.pdf`.

[12] D. R. James A. Larson, T.V. Raman. W3C Multimodal Interaction Framework. Technical report, May 2003. Avail. at `http://www.w3.org/TR/mmi-framework/`.

[13] P. Letessier, O. Buisson, and A. Joly. Consistent visual words mining with adaptive sampling. In *ICMR*, Trento, Italy, 2011.

[14] G. Little. Smiley Slider. Avail. at `http://glittle.org/smiley-slider/`.