A project proposal for
eNTERFACE'11

Multimodal Semantic Search and Summarization for Multimedia Repositories

1. Title and principal investigators

Title: **Multimodal Semantic Search for Multimedia Repositories**

Principal investigators: Milos Zelezny (UWB), Jerome Plumat (UCL-TELE), Vincent Nicolas (UCL-TELE), Olga Vybornova (UCL-TELE, ISA RAS), Ivan Smirnov (ISA RAS), Alexey Karpov (SPIIRAS)

2. Project objective: providing the rationale for the proposed project

This project aims to design and develop a multimedia and multimodal search system capable of dynamically generating documentaries based on search terms informed by users. This system will be multimedia-oriented in purpose and in concept, which means that its main resources will be media content (purpose) and these media content will be organized, searched and manipulated using multimedia as input and output (concept). Medicine will be the domain of experimentation and evaluation explored by this system in the context of the project. In these sub-domains there is a massive production of digital media that have not been stored, indexed and described for future applications. We will provide a multimedia search mechanism to assist medical professionals with efficient search, similar cases and clinical evolution of patients' conditions.

In the focus of development will be user-centric approach where the natural multimodal interaction and navigation capabilities, intelligent caching/ storing relying on easy multimedia archiving tools, as well as semantic search facilitating and personalizing fast access to multimedia database resources will provide effective relevance feedback of the system.

The main objectives of this project are:

- **Storage of medical data and tools allowing easy automatic archiving**. Storing media in the database should be automatic and should not give any supplementary charge to the user. So, the annotation and segmentation tools should be as easy to use as possible.

- **Construction of a database based on an ontology**. The ontology must be correctly defined in order to make the media retrieval possible. Moreover, this organization has to be related to the user parameters, to the parameters of the current query and to the tools used for the query. The user parameters, such as the preferences and the skills of the user will have an influence on the media retrieval and consequently the information contained in both the database annotation and the user parameters must be understood be the search engine. The input query may be done in several ways such as the use of natural language. The integration of all these tools is an important challenge.

- **Development of multimodal interaction for user's input queries and for navigation on the found results.** The system will accept both typed and vocal input, the conversational agent will conduct a dialog with the user helping the user to formulate a query in the optimal form acceptable for the system, and the user will be able to use gestures to navigate through the results found by the system.

The database will be organized by an ontology and will make the information retrieval easier even for other applications.

3. Background information: a brief review of the related literature, so as to let potential participants prepare themselves for the workshop

We will investigate a multimedia multimodal storage and retrieval system that aims to provide enough storage capacity, efficient data extraction and multimodal interaction. It is dedicated to medical image and text data. Multimodal interaction will make the system more usable and will simplify user's access to the data. It is not a secret that medical doctors are often not very proficient in computer works, and they do not know how data is organized and what would be the best format of query to obtain the desired result. For them it is preferable to have voice input so that they can formulate the query in natural language, where the system will transform the query in text form, help to adapt the query and narrow down the search to a specific field and organize the medical terms to match them with the MESH thesaurus concepts http://www.ncbi.nlm.nih.gov/pubmed/.

4. Detailed technical description:

Technical description

There are several crucial issues to be solved in order to meet the project Objectives:

1. The most difficult problem unsolved yet is integration of data from different formats – images and texts. How to match information extracted from the text query to the image segments and to extended texts from various other sources?

What are the possible solutions for this problem?

One of the mandatory points is to do multimodal and multimedia annotation – i.e. to assign certain labels for the image data and for the spoken and written language data, and these labels will classify media for retrieval, simplify taking the decision about fusion/non-fusion and they will enable correct crossmodal reference resolution.

Annotation is the assignment of meanings to segments in order to describe their contents. The system will support several annotation types. They go from a simplistic to a robust form of knowledge representation, giving more flexibility to different user profiles and being domain-independent. They can be assigned to segments and links, covering from simple to complex media contents. Due to the number of annotations supported, we are adopting three categories:

- *Structural Annotation*: describes physical and logical properties of the content (e.g. for a video: its defini-tion, frame rate, dimensions, format, etc), usually extracted from a low level media file processing and from media descriptors.
- *Linguistic Annotation*: has the pure format of text (words, sentences and narratives) and oriented to human readability, structured according to the grammar of the language. They could be referred to also as Textual Annotations or Lexical Annotation.
- *Semantic Annotation*: corresponds to the addition of semantic data or metadata to the content given one or more agreed ontology. Oriented to machine readability, adopts the structure of a graph composed of triples (subject + predicate + object).

A good balance of representativeness and performance is the use of ontologies to annotate segments. Ontology is used to describe a domain of knowledge, which is composed of taxonomy of concepts from a certain domain of knowledge, semantic relationships between these concepts, and instances of these concepts that are representations of scenarios under the modelled domain. Concepts are more representative than tagging because they are well positioned in the domain, but they are less efficient than tagging because there is a cost of exploring the graph of meanings related to them. Comparing with transcription and description, ontologies are less representative than full transcription, but more explicit, computational friendly, and enable derivation of implicit knowledge through automated inference. Ontologies are seen as a helpful tool facilitating efficient search because ontologies define relationships between concepts, impose restrictions on these relationships allowing or forbidding certain ones under certain circumstances, as well as ontologies allow to disambiguate concepts, thus making search more relevant.

The supported media demands the implementation of the following types of segmentation: spatial (can delimit a static region in an image, video frame or 3D model), temporal (can delimit a sequential region in an audio, video or 3D scene) and spatio-temporal (a mix of the two previous types of segmentation where each spatial segment also has time property. Once the segment is created, it can be annotated using one or several techniques. The annotation phase is important to allow an efficient localization of segments based on their intrinsic meanings. To cover different expectations, a segment can be indexed using all supported annotations types: 1) property (more related to media files characteristics than to its content. It is collected during the initial processing, where embedded algorithms are executed automatically, or can be informed manually by the user, if necessary. Examples of properties are: dimensions, resolution, size, format, volume, duration, etc.), 2) Tagging (it is the assignment of keywords to the media content. Each keyword represents a simple word that identifies the content in the segment or in the links between segments. This method is simple, efficient and widely used, but it lacks representativeness), 3) transcription (complete textual descriptions of images, video scenes or audios. This is a very precise method, but very complex from the computational point of view) and 3) AdHoc (AdHoc annotations do not have commitment to be accurate in terms of content meaning. They could represent opinions, comments, external links, references, etc. AdHoc also does not have any priority in the searching mechanism and it is retrieved when the related media is already available for the user, appearing as an additional or complementary information because AdHoc annotations are informal, free-text, and can lead to erroneous decisions) and 5) domain concepts (a good balance of representativeness and performance is the use of ontologies to annotate segments)

2. Another important aspect in this research is multimodal interaction.

In this project the user will be allowed to enter a search query using several modalities. The aim of this task is to develop an input query module allowing using keyboard, graphical user interface, pen gestures or spoken input. The most familiar modality for the search queries is the keywords entry through the keyboard. The users are required to enter keywords to edit boxes and select preferred modality. An auto-completion and auto-correction of grammar and medical terminology are usable features integrated to this modality.

Another input modality for user query entry will be spoken input. Spoken input will be transformed in text form and further processed to formulate the query in the optimal system-understandable form. For natural language queries we will explore the ways of prompting optimal query formulation for the user. The user will have a possibility to refine the query by selecting the semantic meaning of the query. What does the user want to find? It can be information about the disease, information about any particular organ or part of the body, a method of the disease treatment, comparison of different methods, influence of some factor on the disease going, precedents on such disease treatment in clinical practice, details of the method – how long and under which conditions the method can be used, interconnection with other methods, etc. The databases for search will depend on the category of the query.

Asking for the user's feedback and narrowing the search in this way, it is possible to obtain more reliable and accurate results. Reformulating the user's query in the form of a template where the user has to fill the gaps with desired terms from the medical MeSH thesaurus, we can make the search in multimedia resources on the key terms associated with MeSH. Example of a template query: *"If it is efficient to treat disease X with medicine Y?"*

The aim of queries reformulation in the form of templates is to avoid "noisy" redundant components in queries that can be often seen in the queries of medical doctors, like "There is a patient with a difficult case, and I would like to know if it is possible to treat his disease with this or that medicine?"

The modality of a query can be chosen in the user profile and type of installation of the module. System feedback will be given by an embodied conversational agent (ECA) as a modality of the intelligent user interface. The intelligent agent interacts with the environment through an animation model of the physical body. ECA is ideal for this setting because the richer communication style makes interacting with the agent enjoyable. ECA is represented graphically in the form of 2D or 3D animation, for example as a human or cartoon animal. The embodied agent aims to unite gesture, facial expression and speech to face-to-face communication with the user and grow a powerful interaction between the user and the search system. The user gets better orientation in search query as well as personal interactive navigation.

Resources needed:
The open-source software that will be used to meet the project objectives:
- Yasmim multimedia archiving tool (developed in UCL-TELE)
- MedicalStudio (developed in UCL-TELE)
- ITK - Insight Segmentation and Registration Toolkit (http://www.itk.org/**)**
- VTK – Visualization Toolkit (http://www.vtk.org/)
- HTK – speech recognition toolkit (http://htk.eng.cam.ac.uk/)
- Protégé (http://protege.stanford.edu/)

Staff needed to fulfil the task:
- a specialist in medical image processing,
- a specialist in multimedia semantic search
- a specialist in speech processing
- a specialist in ontologies and multimodal data integration
- a computer scientist for the overall system integration

5. Work plan and implementation schedule: a tentative timetable detailing the work to be done during the workshop

WP1 – (Pre-workshop preparation): collecting of the example multimedia database. Additional data collection during the workshop if necessary.

Task 1: to collect a database of medical images of X-rays, ultrasound, mammography, tomography, etc, that shall be segmented and annotated.

Task 2: to collect texts – scientific papers, electronic resources, encyclopedia descriptions, doctors' comments, related to various medical sub-fields.

WP2 – Intelligent multimedia search, distributed multimedia storage, indexation and retrieval (the 1std and 2nd week)

Task 2.1. Integration of natural language search engine with multimedia archiving system

Task 2.2. Classification of media for retrieval

Task 2.3. Evaluation and improvement of classification decision-making

WP3 - Multimodal interaction (the 2nd and 3rd week)

Task 3.1. Input query using different interaction modalities

Task 3.2. Navigation through retrieved results using different interaction modalities

WP4 – Overall system integration (the 4th week)

Task 4.1. Multimedia search and multimodal interaction components integration

Task 4.2. Demonstration

6. Benefits of the research: expected outcomes of the project

Since the task of multimedia search is very versatile, the proper solution of at least one problems listed in the Objectives, will be considered a good result of the project. However the main anticipated outcome is seen as development of a reliable working component for multimedia data retrieval and multimodal access to this data We are going to integrate all the software necessary for efficient multimedia multimodal search and make it functioning as a whole coordinated system.

7. Profile of team:

Milos Zelezny (UWB) – multimodal interaction, ECA
Jerome Plumat (UCL-TELE) – image processing
Vincent Nicolas (UCL-TELE) – medical image processing
Olga Vybornova (UCL-TELE, ISA RAS) – ontologies and multimodal data integration
Ivan Smirnov (ISA RAS) – intelligent multimedia search
Alexey Karpov (SPIIRAS) – speech processing

More team members needed:

- a specialist in English speech processing
- a computer scientist for the overall system integration

8. References

[1] Mendonça, H.: Multi-purpose and Extensible Framework for Multimedia Content Description . PhD thesis, Université catholique de Louvain, Belgium (2010)

[2] Bulterman, D.C.A.: User-centered control within multimedia presentations. Multimedia Syst. 12(4-5), 423–438 (2007)

[3] Schenk, S., Saathoff, C., Staab, S., Scherp, A.: Semaplorer - interactive semantic exploration of data and media based on a federated cloud infrastructure. J. Web Sem. 7(4), 298–304 (2009)

[4] Schrijver, D.D., Neve, W.D., Wolf, K.D., Sutter, R.D., de Walle, R.V.: An optimized mpeg-21 bsdl framework for the adaptation of scalable bitstreams. J. Visual Communication and Image Representation 18(3), 217–239 (2007)

[5] Sofokleous, A.A., Angelides, M.C.: Dcaf: An mpeg-21 dynamic content adaptation framework. Multimedia Tools Appl. 40(2), 151–182 (2008)

[6] Karpov, A., Ronzhin, A., Kipyatkova I.S., Ronzhin A.L., Akarun L.: Multimodal Human Computer Interaction with MIDAS Intelligent Infokiosk. ICPR 2010: 3862-3865

[7] Sargin, M.E., Aran, O., Karpov, A., Ofli F., Yasinnik Y., Wilson S., Erzin E., Yemez Y., A. Tekalp M.: Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis. ICME 2006: 893-896

[8] Mendonça H., Vybornova O., Lawson J. Y. L., Macq B., Vanderdonckt J.: High Level Data Fusion on a Multimodal Interactive Application Platform // submitted to the Eleventh International Conference on Multimodal Interfaces and the Sixth Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009), November 2-6, 2009, Cambridge, MA, USA.