

Scientific Report

ESF Travel Grant – 5112

Grantee: Prof. Burak Acar

This report summarizes the scope and the achievements of the 4-day visit to Rome, Italy, during the CLEF 2012 conference, hosted by Prof. Henning Müller, University of Applied Sciences Western Switzerland, Sierre (HES-SO), Switzerland.

Purpose of the Visit:

The 4-day visit during the CLEF 2012 conference was planned to discuss, finalize and propose, with an oral presentation, a new task for the imageCLEF workshop within the field of Content Based Case Retrieval in Medicine.

Description of the Work:

CLEF is a premium platform that gathers researchers in the field of multi-modal information retrieval. The relevant tasks, addressed so far by the CLEF community, has been mostly focused on textual data retrieval. Since 2004, a search and retrieval task in the medical domain has been pursued. As part of these efforts, a medical image and case retrieval task was defined. Yet, the aforementioned task aims at retrieving medical documents from the online literature, namely the publications in the PubMed database, using the visual information in the images published as well as the textual information gathered from image/figure captions. Despite this task definition being easily adaptable to existing retrieval tasks, it is recognized that it is far from having a real impact in the actual clinical workflow. During this visit, in close collaboration with Prof. Müller, I have worked on defining a new Content Based Liver Case Retrieval (CBCR-Liver) task within a specialized and structured database of past cases. A task proposition to the CLEF community was presented as a result of this work.

The novel part of CBCR-Liver task is that the retrieval units are real patients/cases that are described by multi-modal information at 4 levels: Patient, Study, Series and Pathologies in a hierarchy. The “Patient” level includes information about the demographics of the patient, as well his/her medical history relevant to a liver related diagnosis. The “Study” level includes information about a single application to a medical institution, as such, it involves the complaints, the current drugs used, the lab tests, etc. The “Series” level refers to a single radiological imaging, which is an abdominal CT scan of the liver. It involves the 3D DICOM CT images and image related annotations about the liver as a whole (size, vasculature, etc.). The “Pathology” level refers to a single finding (a lesion) marked in the liver, together with related image annotations. All data is structured and is kept in relational database that is dynamically linked to a novel ontology for the last two levels, namely for the “Series” and the “Pathology”. The ontology in question has been developed as part of an ongoing research project (CaReRa Project, PI: Prof. Burak Acar) in collaboration with medical institutions from Turkey and USA. The ontology is very comprehensive and is RadLex complaint. In short, the case descriptions are composed of textual,

Patient / Case:

- Demographics (Age, Gender)
- Regular drugs used (ATC codes)
- Past & Chronic diseases (ICD-10 codes)
- Surgeries (Turkish Ministry of Health codes)

Study:

- Complaints (Free text for future reference)
- Current drugs (ATC codes)
- Current diseases (ICD-10 codes)
- Physical Examination (Blood pressure; Pulse; Signs of Portal Hypertension such as Ascites, Oedema, Splenomegaly; Signs of Liver Dysfunction such as Jaundice)
- Pre-diagnosis (ICD-10 codes)
- Lab results (26 test results)

Series (with 3D liver mask):

- Phase of contrast enhanced CT data
- Liver surface, size and density characteristics
- Vasculature

Pathology (with 3D bounding box):

- Lesion Type (Hepatocellular carcinoma, metastases from multiple types of primary tumors, hemangioma, cyst, and focal nodular hyperplasia)
- Lesion Number, solitary or multiple (number of discrete lesions in the liver)
- Lesion Location (liver lobe or Couinaud hepatic segment)
- Lesion Size (major dimension in mm)
- Localization with respect to Gallbladder, Periphery, etc.
- Lesion Contour Properties (e.g. well defined, ill defined, smooth or lobulated)
- Lesion Shape
- Lesion Enhancement (Presence or Absence; temporal enhancement?)
- Lesion Enhancement Patterns (uniform, nonuniform, mosaic, hypervascular, etc)
- Lesion Internal Structure (solid, cystic, mixed)
- Lesion Special Features (calcification, scar, necrosis, gas)
- Density Properties (Water, soft tissue, fat, mixed)
- Septa Properties
- Wall Properties
- Lesion-Vasculature Relation

Figure 1 The multi-modal 4-level liver case description. The “Series” and the “Pathology” levels are linked to CaReRa\Liver ontology. All data is structured and is composed of textual (labels), numerical and visual information. Note that a series refers to a 3D abdominal CT scan.

numeric and visual information that need to be processed together for case retrieval. Part of the work that was carried out during this 4-day visit is related to defining a subset of this data that can be easily adopted and exploited by the CLEF community using their existing know-how and technologies. Figure 1 presents an overview of the data structure as used in the CaReRa project. Figure 2 presents the dataset that has been decided on for the task proposal.

Patient / Case:

- Age & Gender
- Hepatitis status
- History of malignancy

Study:

- Blood tests:
 - Metabolic Panel: BUN, Creat, Sodium, Potassium, bicarbonate
 - Liver Function Tests: Bilirubin (direct and indirect), AST, ALT, Alk Phos, Protein (total), Albumin
 - CBC (Complete Blood Count): RBC, WBC, Platelets
 - AFP (alpha fetoprotein)
 - CEA (carcinoembryonic antigen)

Series (with 3D liver mask):

- Phase of contrast enhanced CT data

Pathology (with 3D bounding box):

- (Incomplete) CaReRa\Liver ontology descriptions

Figure 2 The dataset proposed for the imageCLEF workshop. It is a subset of the CaReRa dataset

A critical component of the proposition is the task definition itself. Unlike existing tasks and/or approaches, the retrieval units here are the real (anonymized) cases. And the essence of the whole task is to realize a clinical experience sharing (CES) platform. CES refers to making clinical experiences available to a large community through tools that enable search and retrieval in a specialized database of past cases, where this collection of past cases represents the collective clinical experience knowledge-base. It is envisioned that the users of such a system will be able to use the case they are required to give a diagnostic decision about, as the query case to retrieve *relevant* past cases that would help them in decision making through comparative diagnosis. The definition of *relevancy* in terms of cases is critical and much of the work during the visit has focused on related discussions. It is concluded that two cases are relevant to each other if they aid the diagnosis mutually, by being positive (cases with the same diagnosis but potentially different characteristics) or negative (cases with different diagnosis but with similar characteristics) examples. Consequently, the performance of the system can only be measured using manually generated case relevancy matrices on the training and test cases. The feasibility of this approach is discussed and it is concluded that, with the aid of our collaborators on this task (Stanford University, USA, Istanbul University, TR, Yeditepe University, TR, TU Munich,

DE and HES-SO, CH), this can be achieved and would be a significant contribution to the community. It should be noted that relevant case retrieval is not a CAD (Computer Aided Diagnosis) approach, though a CAD approach is also of interest and can be pursued with this dataset. Consequently, through discussions, it is decided that a second task, where the goal would be to retrieve cases with the same diagnosis as that of the query case, is also defined for the imageCLEF workshop. The evaluation of this second task is straight forward. Furthermore, inclusion of this second task is expected to increase the interest in terms of participants.

Following the presentation of the proposed tasks and receiving the comments of the CLEF community, a roadmap regarding the data collection and gold standard generation (the case relevancy matrices) is formed. The proposition will be defended by Prof. Henning Müller and should it be officially accepted, we will proceed according to the aforementioned roadmap.

Main Results:

The main results of this 4-day visit can be summarized as follows:

- The strategies for presenting the medical case retrieval task for clinical experience sharing (CES) within a real clinical workflow, to the CLEF community have been discussed. The focus has been on strategies that would enable easy adaptation of the existing know-how in content based search and retrieval to this new domain and paradigm.
- The extent of the multi-modal medical case data has been discussed and concluded. The focus has been on the feasibility in terms of data collection and gold standard generation.
- Two sub-tasks, the “Relevant Case Retrieval” and the “Diagnostic Classification”, have been defined.
- A roadmap has been formed. It describes the steps towards data collection, gold standard generation and fund raising.

Future Collaborations:

The task propositions are jointly proposed by Bogazici University, TR, Stanford University, USA, Istanbul University, TR, Yeditepe University, TR, TU Munich, DE and HES-SO, CH. This group of collaborators will form the seed of a consortium. By means of the presentation of the task at CLEF 2012, we envision to enlarge this collaboration network. Acceptance of the proposed tasks for imageCLEF 2013 would boost the process. We are also in touch with some other institutions in Europe and USA, promoting these ideas. It is envisioned that a multi-national, primarily European collaboration network will be formed as a result of these initiatives.

Projected Publications and Impact:

Should our proposal be accepted for imageCLEF, we expect to have a series of publications annually based on the dataset we offer for the task. The publications are expected to be in multi-modal information retrieval, information fusion, machine learning applications, evaluation strategies and user interfaces. In the future, we envision that the proposed dataset would also enable research in data mining to discover correlations between multi-modal (and seemingly unrelated) data. With proper extension of the dataset,

research on radiogenomics may well be started, which is of great interest to the above mentioned institutions.

The impact of this research would in

- Improved accuracy in diagnosis through CES, esp. by means of making the collective clinical experience available to junior MDs,
- Medical education through providing students a platform to assess the subtle differences between cases that are critical in diagnosis.
- Personalized medicine via retrieving cases that are relevant to a query case by means of serving as a model.

21.09.2012

Prof. Burak Acar

TU Munich, Institute of Informatics, CAMP

Garching bei Munchen, Germany

A handwritten signature in black ink, appearing to read 'Burak', is centered below the contact information. The signature is stylized and written in a cursive-like font.