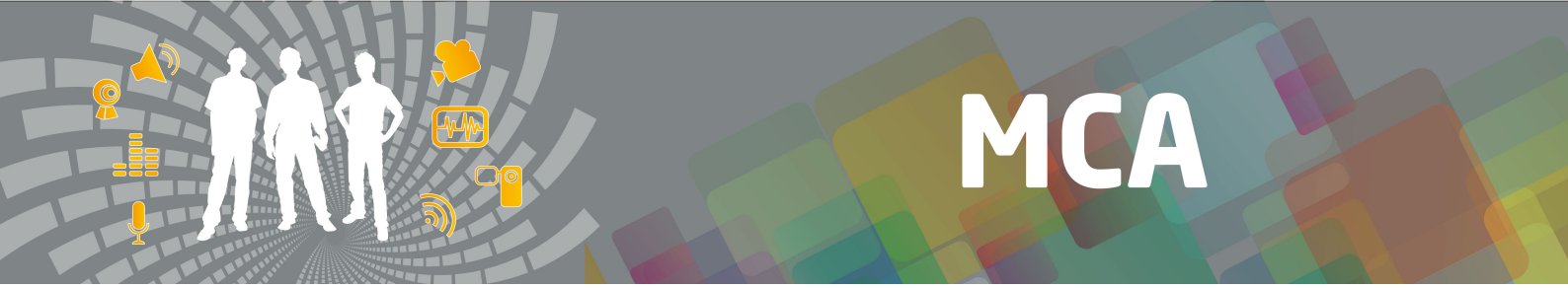

Multimodal Conversational Analytics
MCA Grand Challenge – Call for papers

The MCA challenge aims to bring together researchers from across disciplines related to multimodal conversational analytics. The challenge follows the D-META Challenge organized at the ICMI 2012 in Santa Monica. The D-META challenge had two coupled pillars, method benchmarking and annotation evaluation, and its starting point was set in transparent and publicly available annotations, tasks, and evaluations on shared multimodal data sets. This second challenge continues the same tradition.

Paper dead line: **15-Jul-2013**.
Accepted papers will be part of the IEEE/ACM **ICMI'2013 Proceedings**.

Scope

The scope of the MCA challenge concerns the multimodal analysis of primary cues and qualities of conversations. It proposes to set up the basis for comparison, analysis, and further improvement of multimodal data annotations and multimodal interactive systems which are important in building various multimodal applications. Such machine learning-based challenges do not exist in the Multimodal Interaction community, and by focusing on the elaboration of algorithms and techniques on shared data sets, we aim to foster the research and development of multimodal interactive systems.
The challenge is organized by tasks, which concern three different aspects of multimodal conversational analytics:

MCA-Engagement Level of engagement in video-mediated group conversation.

MCA-Gestures Feedback and conversational gestures in first encounter dialogues.

MCA-Layout Physical layout in short multi-speaker conversations.

We invite papers dealing with these tasks and covering especially areas such as (i) applications of inference algorithms to a data set(s), (ii) benchmark of several algorithms using the same data set(s), (iii) extensions of the annotation scheme with new relevant features, (iv) applications of the data to an automatic system, (v) discussions on ecologically valid data sets and (vi) position papers of how to organize the next challenge.

Important dates

The schedule has the following important dates:

15-Jul-2012	Paper deadline
1-Sep-2012	Author notification
1-Oct-2012	Camera-ready
Dec-2012	Work presented at MCA'13

Format

The papers should be formatted following the ACM/IEEE format as in the ICMI 2013 proceedings. No more than six pages long, the manuscripts should contain the motivation, a brief description of the benchmarked methods, and an extensive discussion of the obtained results. No description of the data sets is needed, but the citation to the reference papers. Accepted papers will be part of the IEEE/ACM ICMI Proceedings.

Detailed tasks

The tasks outlined before are explained in detail in the following.

MCA-Engagement

Description

The aim of this task is to estimate the level of conversational engagement of participants in a group video-mediated communication. The dataset for this task consists of several auditory, video and gaze recordings from a potential home teleconference system (see [1, 6]). Each recording captures interaction between a group of co-located participants and one remote participant, involved in activities ranging from casual conversation to simple social games. The audio-visual recordings are accompanied by gaze recordings of the remote participant, manually-annotated head positions and voice activity annotations. The experiments will be done for the remote participant for whom gaze data is available.

Evaluation metric

A ground truth annotation for training part of the dataset will be made available to the participants. Ground truth for testing part of the dataset will be released after the challenge. In the submission, participants are expected to provide a short description of their system, and its outputs for short intervals of the testing data in a defined simple format for evaluation. The official metric used in the evaluations will be a weighted classification cost reflecting the similarities between the different levels of engagement. The weights will be made public together with the training data. Additionally, DET (Detection Error Tradeoff) curves and confusion matrices will be generated. A baseline two-class engagement recognition achieved EER of 74% [5].

MCA-Gestures

Description

The challenge has two subtasks which aim at promoting studies in the detection and interpretation of the relevant gesturing in the context of conversational interactions. The first subtask is to recognize the interlocutors gestures in general so as to distinguish those that have a communicative function (e.g. giving feedback) from those that are other type of gestures (e.g. scratching an itchy arm). The second subtask is to classify the communicative gestures further, and to use relevant features for the recognition of feedback giving features. The data used for this challenge concerns first encounter dialogues in Finnish and Swedish languages, and is collected within the NOMCO project [2, 7].

Evaluation metric

In order to evaluate the performance of the methods targeting this task, the confusion matrix and the f1-score should be provided. In the detection of a gesture, a frame-based counting will be applied. The recognition method should output one of the annotated classes or no class for each frame. This will be compared to the ground truth in order to build a confusion matrix.

MCA-Layout

Description

The aim of the task is to detect, localize and track speakers from audio-visual sequences. The data used are some scenarios of the interaction part in the Ravel data set [4]. See the data set web site [3] for more information on which sequences to use.

Evaluation metric

In order to evaluate the results, the (euclidean) distance matrix between the detected speakers and the ground-truth speakers should be computed. Each ground-truth speaker should be associated at most to one detected speaker. The assignment procedure is as follows. For each detected speaker its closest ground-truth speaker is

computed. If it is not closer than a threshold loc it is marked as false positive, otherwise the detected speaker is assigned to the ground-truth speaker. Then, for each ground-truth speaker the number of detected clusters are assigned to it is checked. If there is none, it is marked as missing detection. Otherwise, the closest detected speaker becomes the true positive and the remaining ones become false positives. Recall, precision and accuracy values should be shown in tables (and occasionally in figures also) for different values of loc in the range 1cm 50cm.

References

[1] Brno social interaction dataset. http://medusa.fit.vutbr.cz/TA2/TA2/.

[2] The NOMCO project. http://sskkii.gu.se/nomco/.

[3] The ravel data set. http://ravel.humavips.eu.

[4] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Cech, K. Kulkarni, A. Deleforge, and R. P. Horaud. Ravel: An annotated corpus for training robots with audiovisual abilities. *Journal on Multimodal User Interfaces*, 2012.

[5] R. Bednarik, S. Eivazi, and M. Hradis. Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification gaze. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Article 10*, 2012.

[6] M. Hradi, S. Eivazi, and R. Bednak. Voice activity detection in video mediated communication from gaze. In *Proceedings of ETRA'12*, 2012.

[7] C. Navarretta, E. Ahlsn, J. Allwood, K. Jokinen, and P. Paggio. Feedback in nordic first-encounters: a comparative study. In *Proceedings of LREC 2012*, May 2012.