D−META Grand Challenge – Call for papers

The D−META Grand Challenge aims to set up the basis for comparison, analysis, and further improvement of multimodal data annotations and multimodal interactive systems. The main goal of this Grand Challenge is to foster research and development in multimodal communication and to further elaborate algorithms and techniques for building various multimodal applications. Held by two coupled pillars, method benchmarking and annotation evaluation, the D−META challenge envisions a starting point for transparent and publicly available application and annotation evaluation on multimodal data sets.

Main features

Paper dead line: **31-Jul-2012**.

Summary paper with all the resutls published at IEEE/ACM **ICMI'2012 Proceedings**.

Scope

The D−META challenged is organized by tasks. Each tasks to be solved has one (or more) associated data set. In order to evaluate the performance on each task, a Golden Standard will be provided. Such standard consists of (partial) annotations and sets the ground truth over the data set concerning the targetted task.

Authors may submit their methods/benchmarks for applications and/or their systems/comparisons for annotations and evaluation. We expect papers covering areas such as: (i) applications of an algorithm to a data set(s) to solve precise tasks, (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 organise the next challenge.

In the following, the tasks are outlined, further in the text, they are detailed.

AVRGR Recognize gestures addressed to the robot by means of the vision and the audio.

AVSR Detect, localize and track multiple speakers using audio-visual information.

CEP Estimate the level of engagement in a video-mediated communication.

AVCGR Recognize conversational gestures in first encouter dialogues.

AVFGR Recognize feedback gestures in first encouter dialogues.

Important dates

The schedule has the following important dates:

19-Mar-2012	Data set annotation is released
31-Jul-2012	Paper deadline
24-Aug-2012	Author notification
14-Sep-2012	Camera-ready
Oct-2012	Work presented at D−META'12

Format

The papers should be formatted following the ACM/IEEE format as in the ICMI 2012 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. A summary paper collecting the results presented by the authors will be published together with the ICMI proceedings.

Detailed tasks

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

AVGR - Audio Visual Robot Gesture Recognition

The aim of this task is to recognize gestures from audio and video. In the Ravel data set [1, 3], the "Robot Gesture" scenario was conceived for this particular purpose. Indeed, several actors/actresses performed gestures such as "point", "yes", "not", ... The task is the isolated recognition of these robot gestures.

Evaluation metric In order to have a common way to evaluate different methods, the "confusion matrix" should be used when targetting this task. A confusion matrix for N classes is an $N \times N$ matrix with its ij-th element (i-th row, j-column) means: how many instances os class i the method recognized as class j. The confusion matrix for all the gestures should be provided. Obviously, testing and training subsets should be different. We reccomend a leave-one-out strategy for evaluation.

AVSR - Audio Visual Speaker Recognition

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 [3]. See the data set web site[1] 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 speaer is computed. If it is not closer than a threshold τ_{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 occasionaly in figures also) for values of τ_{loc} in the range 1cm to 50cm.

CEP - Conversational Engagement Prediction

The aim of this task is to estimate the level of engagement of participants in a group video-mediated communication (conversing, responding, following, no interest). The dataset for this task consists of several audio and video recordings as could be recorded from a potential home teleconference system (see [4][2]). 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 and video 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.

[1] http://ravel.humavips.eu

[2] http://medusa.fit.vutbr.cz/TA2/TA2/

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. 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 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. The participants can choose to ignore the provided voice activity annotations as such submissions will be evaluated separately from those submissions using them.

AVCGR - Audio Visual Conversational Gesture Recognition

The aim of the task is to recognize the interlocutor's gestures in conversational interactions. The data has been collected within the NOMCO project [2], and concerns first encounter dialogues in Finnish and Swedish languages.

Evaluation metric To evaluate the results, 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. Thus, the evaluation is similar to the AVRGR task.

AVFGR - Audio Visual Feedback Gesture Recognition

The aim of the task is to classify the interlocutor's gestures into those that are related to feedback giving in conversational interactions, and those that are other type of gestures (e.g. scratching an itchy arm). The data used is collected within the NOMCO project [2] and concerns first encounter dialogues in Finnish and Swedish languages.

Evaluation metric In order to evaluate the performance of the methods targetting this task, the confusion matrix and the f1-score should be provided.

References

[1] The HUMAVIPS project. http://humavips.eu/.

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

[3] Xavier Alameda-Pineda, Jordi Sanchez-Riera, Vojtech Franc, Johannes Wienke, Jan Cech, Kaustubh Kulkarni, Antoine Deleforge, and Radu P. Horaud. The ravel data set. In *IEEE/ACM ICMI 2011 Workshop on Multimodal Corpora*, Alicante, Spain, November 2011.

[4] Michal Hradi, Shahram Eivazi, and Roman Bednak. Voice activity detection in video mediated communication from gaze. In *Accepted to ETRA'12*, page 6, 2012.