Version 10 — July 31, 2013

TRECVID 2013 Multimedia Event Recounting (MER) Evaluation Plan

The Multimedia Event Recounting evaluation analyzes the recounting output by the MED system. The purpose of recounting is to enable a user to rapidly and accurately identify events in the clips returned from the MED system. Participation in the MER evaluation is open to all MED participants in TRECVID 2013 whose system always produces a recounting for each clip that their MED system deems positive.

BACKGROUND

Each event in this evaluation

- is a complex activity occurring at a specific place and time;
- involves people interacting with other people and/or objects;
- consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity; and
- is directly observable.

The following terminology will be helpful. Each event kit explicitly defines an *event*. A clip that is *positive* for an event contains an *instance* of that event. The recounting summarizes the *important evidence* that the clip contains the event. Ideally, the recounted evidence should be the evidence that the MED system used to detect the event. The recounting can also state that no evidence for the event was found. It is important for the system to *locate* the important pieces of evidence — both temporally (where in the clip) and spatially (where in the frame) to help the user quickly locate the event within a clip. Evaluation of the recountings will focus on the ability of a human judge to rapidly and accurately decide which clips contain an actual instance of the event of interest.

The Event Kit text for 2013 will consist of

- Event name: a mnemonic title for the event
- Event definition: a textual definition of the event
- Evidence: a textual statement of evidence that indicates the existence of the event.

RECOUNTING Format

MED Systems will produce an output recounting file for each event and each clip. This file will be in XML format. The recounting is a list of observed evidence, where each piece of evidence on the list includes:

- id: a unique identifier
- Description: a concise textual summary of the piece of evidence. This would ideally be an English sentence or phrase meaningful to the user, but could alternatively be a code word
- Confidence: a score indicates how confident the system is in its detection and localization of this piece of evidence (range from least confident at 0.0 through most confident at 1.0).
- Importance: a score that indicates how important this evidence was in detecting the event (range from least important at 0.0 through most important at 1.0).
- Sources: a list of one or more sources of this piece of information or evidence, drawn from the following four possible values.
    - video (not involving OCR)
    - visible_text (via OCR)
    - speech (transcribed via ASR)
    - non_speech_audio (sounds not involving ASR textual transcription)
- Presentation Order: an ordinal number indicating the order in which the MER Workstation should present the evidence to the user (i.e., 1, 2, 3, . . .).
- Pointer List: A list of pointers for each piece of evidence, to spatially and temporally locate the evidence within the clip. Each element on the list includes:
    - Start Time: an offset into the clip (either a time offset or a frame number) of the start of the piece of evidence
    - End Time: an offset into the clip to the end of the piece of evidence
    - Start Visual Bounding Box: defined at Start Time, the upper left (row, column), and lower right (row, column) pixel coordinates (in the video frame). The upper left corner of the video frame has coordinates 0:0 and the lower right corner of the video frame has coordinates that are the size of the video frame in pixels.
    - End Visual Bounding Box: defined at End Time, and the upper left (row, column) and lower right (row, column) pixel coordinates (in the video frame)

    Note: NIST will refer to the sub-clips that are defined by this list as "snippets."

NIST will provide a DTD for this data format. The Start Visual Bounding Box and End Visual Bounding Box should be completely omitted for a snippet that is purely audio; otherwise both are required for each pointer in the Pointer List. A bounding box is allowed to enclose the entire video frame. The DTD will allow the recounting to optionally state the source(s) of evidence for each snippet. A schema file to be used to validate the recountings is provided as part of the submission checker (see the submission instructions).

A recounting will be maximally useful if it can help a user to more rapidly and accurately find the clips of interest in a collection of clips. Therefore, the recounting should (1) include only the key pieces of evidence (omitting anything unimportant), (2) textually state each such piece of evidence accurately and concisely, (3) maximally narrow down the time span where each of those pieces of evidence occurs in the clip, and (4) maximally narrow down the bounding boxes that identify where the piece of evidence occurs in the frame.

DATA TO BE PROCESSED AND JUDGED IN THE 2013 MER EVALUATION

MER participants should generate a recounting for each clip that their MED system deems to be positive in the FullSys-PROGAll-PS13-100Ex condition (this is the "Required" column in the table in section 3 of the MED Evaluation Plan). MER will *not* be evaluated for any other conditions. See the MED Evaluation Plan for other details.

NIST will select a subset of the recountings for evaluation. All submissions will be judged on their recountings for the selected set of (clip, event) pairs.

EVALUATION

The system's recountings will be evaluated by a panel of judges. NIST will create a MER Workstation and provide it to participants and judges. Using this workstation, the judge will be instructed to study the event kit text (not the example videos) and then to assess the recounting by:

1. Reading the entire list of textual descriptions
2. Viewing/hearing ***all*** the snippets defined by the spatiotemporal pointers
3. On the basis of the recounting, classifying the clip as one of the following:
    - The clip *contains an instance* of the event
    - The clip *does not* contain an instance of the event
    - I do not know because the *recounting* does not allow me to tell whether the clip contains an instance of the event
    - I do not know because the *event kit* does not allow me to tell whether the clip contains an instance of the event

The event kit text will be available to the judge for reference.

The MER Workstation will *not* display the source(s) of the information — neither for each piece of evidence, nor for the snippet(s) associated with each piece of information. The stated sources of information will be used for post-hoc understanding of the system.

For each submission and each event, NIST will measure the following characteristics of the recountings:

- Percent Recounting Review Time: percentage of clip time the judges took to perform steps 1–3. above

$$Percent\ Recounting\ Review\ Time = \frac{Total\ time\ needed\ to\ perform\ steps\ 1-3}{Total\ duration\ of\ clips\ to\ be\ assessed}$$

- Accuracy: the degree to which the judges' assessments (step 3 above) agree with the MED ground truth.

$$Accuracy = \frac{Number\ of\ correctly\ labeled\ clips}{Number\ of\ clips\ to\ be\ assessed}$$

- Precision of the observation text: the mean of the judges' scores on the following question, which will be asked for each observation, across judges for each event, for each system.

How well does the text of this observation describe the snippet(s)?

▪ *A: Excellent*	*(4 points)*
▪ *B: Good*	*(3 points)*
▪ *C: Fair*	*(2 points)*
▪ *D: Poor*	*(1 point)*
▪ *F: Fails*	*(zero points)*

MER submissions whose recountings enable the judges to assess recountings the most rapidly and accurately will be considered the best. NIST will be evaluating the recountings, not evaluating the judges. We are interested in recountings that state the evidence in a way that human readers find easily understandable.

The metric about the precision of the observation text is intended to provide the system developers with useful feedback about their recountings.

SOFTWARE RESOURCES

NIST will provide a MER Workstation to be used in judging. The MER Workstation will be implemented in Ruby on Rails and will run over the internet using a typical web browser. Participants are encouraged to improve the MER Workstation and to provide feedback to NIST. The best suggestions will be integrated into the final MER Workstation used by the judges for the evaluation.

DRY RUN

All participants will be required to participate in the MED dry run and generate recountings. The purpose of the dry run is to ensure both that the system outputs are being generated as expected and that they can be processed by the evaluation pipeline. This exercise will also provide the evaluation team with insight into how the recounting will be used on the MER Workstation for the judges in the formal evaluation. No evaluation of the recountings in the dry run submission will occur, and the only feedback that will be provided to performers will relate to any problems with data formatting, conformance to the DTD or Schema, and so forth.

Teams are encouraged to use the MER Workstation provided with their own judges to assess their performance.

For the MER dry run, systems should run the MED dry run and generate a recounting for each clip that their MED system deems to be positive. See the MED Evaluation Plan for other details.

Submission instructions for the Dry Run

The Dry Run submission should be packaged up like the main evaluation submission.

SUBMISSION INSTRUCTIONS FOR THE MAIN EVALUATION

MER submissions should be packaged with MED submissions according to the following directory structure

output/<EXPID>/MER/<EVENTID>/<TRIALID>.mer.xml

output/<EXPID>/<MEDFILES>

Where:

EXPID: *refer to the MED13 Evaluation Plan Appendix B*

EVENTID: The event kit ID, e.g., E022, E026

TRIALID: <CLIPID>.<EVENTID>, e.g., 012345.E022

MEDFILES: MED13 submission files, *refer to the MED13 Evaluation Plan Appendix B.2*

MER participants whose MED system will process the PROGAll dataset should only include an MER submission for experiments that meet the following criteria (*refer to the MED13 Evaluation Plan Appendix B*).

<SYS> = "FullSys"

<SEARCH> = either "MED13DRYRUN" or "PROGAll"

<EVENTSET> = "PS"

<EKTYPE> = "100Ex"

MER participants whose MED system is processing PROGSub *instead of* PROGAll should include a MER submission for experiments that meet the following criteria.

<SYS> = "FullSys"

<SEARCH> = either "MED13DRYRUN" or "PROGSub"

<EVENTSET> = "PS"

<EKTYPE> = "100Ex"

The structure and contents of the MER directory will be checked against the corresponding MED submission. Specifically, MER output should only be present for trials with a positive MED decision.

MER submissions must conform to this structure in order to pass validation. Submissions that do not pass validation will be rejected.

Validating the Submission:

MER submission validation is performed as a part of the TV13MED-SubmissionChecker. Instructions on how to use the validator are included in Appendix B.3 of the MED13 Evaluation Plan.

Each mer.xml file will be validated using the MER13 schema file, which can be viewed on the MER13 webpage.

Transmitting Submissions:

Participants should follow the Transmitting Submissions instructions in Appendix B.4 in the MED13 evaluation plan when sending submissions.

For the MER submission primer you will need the primer files from the MED13 page (MED13DRYRUN_Files-v2.tar.bz2 from http://www.nist.gov/itl/iad/mig/med13.cfm) and the MER example submission from the MER13 page (MED13_testTEAM-MER_MED13DRYRUN_PS_1.tar.bz2 listed as "MER example submission" from http://www.nist.gov/itl/iad/mig/mer13.cfm). As well as the F4DE toolkit (at least 3.0.0) installed.

Extract the MED13DRYRUN_Files-v2.tar.bz2 archive. Then run the command

```
TV13MED-SubmissionChecker  -V MED13_testTEAM-MER_MED13DRYRUN_PS_1.tar.bz2
   -T MED13DRYRUN_Files/MED13DRYRUN_20130501_TrialIndex.csv
```

For more information on the TV13MED-SubmissionChecker see the TRECVID MED13 Scoring Primer document
(available in F4DE at F4DE/DEVA/doc/TRECVid-MER13-ScoringPrimer.html).

SCHEDULE

Date *(all dates 2013)*	**Milestone**
April 1	Final design published in the TRECVID Guidelines on the web, and MER DTD available
May 1	Initial draft version of MER Workstation available to participants.
June 5	2013 MED Progress Set disk drives available (for new participants)
July 1	MED/MER dry run begins
July 31	MED/MER dry run ends (MER dry run submissions due at NIST)
Sept 10	MED/MER participants submit Pre-specified Event runs (including recountings)
Sept 13	NIST releases Adjudicated MED results for the Pre-specified Events
Sept 27	NIST releases pre-adjudication MER results on the Pre-specified Events
Oct 2	NIST releases post-adjudication MER results on the Pre-specified Events
Oct 4	TRECVID speaker proposals due at NIST by noon (Gaithersburg time)
Nov 20 – 22	TRECVID workshop at NIST in Gaithersburg, MD