Version 23 — May 29, 2012

TRECVID Multimedia Event Recounting (MER) Evaluation Plan

BACKGROUND

The goal of MER is to summarize the evidence in a specific video clip for it containing an instance of a specific event. The key is to focus on **content**. Each event kit explicitly defines an *event*. A clip that is *positive* for an event contains an *instance* of that event.

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.

Participation in MER is open to all current TRECVID participants.

The proposed evaluation of MER has two goals. The first is to evaluate whether the MER outputs by themselves allow human judges to identify which event is represented by a recounting. The second is to evaluate whether a recounting is sufficiently expressive of both the event and the clip it refers to such that judges can match each recounting to the clip from which it was derived.

This initial MER evaluation will have XML-tagged text-only output. In future years, TRECVID will have MER evaluations with text-only output and MER evaluations with text plus multimedia (graphics) output.

NIST will create a DTD that defines the MER output format, and a corresponding rendering tool to display valid MER output.

SYSTEM TASK

Given an event kit and a test video clip that contains an instance of the event, the MER system will produce a recounting summarizing the key evidence for the event in the clip. Evidence means observations of scene/context, persons, animals, objects, activities, text, non-linguistic audio, *and other evidence* supporting the detection of the event. Each observation will be associated with an indication of the system's confidence that the observation is correct.

Systems will produce an XML element for each observation, and that element will include attributes that give the following information. Note: the "id" and the "description" are required; the rest of the information is strongly encouraged but optional.

- id: a unique identifier that can be used in other XML elements to associate elements, e.g., to associate an object or person with an activity
- type: as explained below
- description: a concise textual statement of the observation
  (For example, if the *type* is *object*, the *description* might be *red Toyota Camry*)
- startTime: an offset into the clip
- endTime: an offset into the clip
- boundingBox: a bounding box defined by a time offset into the clip, the upper left (row, column) and lower right (row, column) coordinates to surround a visible piece of evidence at the time offset. The time offset should be between startTime and endTime
  (omit boundingBox for purely acoustic evidence)
- confidence: in the range 0.0 through 1.0, with 1.0 indicating highest confidence

In the XML element for an observation, the *type* attribute will have one of the following values.

- scene/context

  A "scene or context" is a descriptive set of information flowing from a physical environment. It could include things such as a cityscape, an agricultural farm, a natural setting, a park containing children's swings, or a broad activity such as a soccer game. Also included are unresolved groupings such as a crowd, a clump of trees, or a bunch of houses; or a sub-event— for example, lightning striking, a vehicle exploding, or a rock slowly tumbling down a hill.

- object

  "Object" is something inanimate that is visible in the clip. Examples include a tent, suitcase, building, or tree. It is possible for an object to be in motion.

- person

  "Person" means one human being.

- animal

  "Animal" means an animal, not a human.

- activity

  "Activity" is a person or animal doing something. Examples include a person running, putting up a tent, throwing a ball, playing basketball, talking, or hiding. Examples of an activity involving an animal include a dog fetching a stick or a cat chasing a mouse. Note that an activity involves a *living* actor.

- text

  "Text" is

  (1) any text visible in a clip (often referred to as "scene text"),

  (2) text overlaid on the clip (titles, closed-captions, etc.), or

  (3) understandable speech.
- non-linguistic audio

  "Non-linguistic audio" (also known as an acoustic event), is sound other than understandable speech. Examples include crash, gunshot, honk, laugh, sneeze, bark, or babble (as of a crowd).
- other

  "Other" is a place for site-defined additional useful information to understand the event, and is intended as an opportunity for MER developers to include evidence that does not fit into the categories of observations described in the preceding several paragraphs. Examples include camera motion, video quality, and relationships such as between a person and activity (i.e., a fisherman casting a line into the water). Developers who wish to explore relationships should note the potential usefulness of the *id* attribute in each observation.

For each clip, participants are to track and report separately: (1) the time required for evidence identification and exaction (including all preprocessing time required to ingest the clip), and (2) the time required for MER output generation.

MER-2012 EVENTS

The events used in the MER evaluation will be five of the pre-specified events used in the MED evaluation. These events are:

- cleaning an appliance
- renovating a home
- rock climbing
- town hall meeting
- working on a metal crafts project.

To obtain the event kits and associated video clips, participants will need to complete the MED12 *Evaluation License* and send it to the Linguistic Data Consortium (LDC). See the Data Resources and Data Licensing sections of the MED 2012 web page

http://nist.gov/itl/iad/mig/med12.cfm

DATA RESOURCES

Three data sets will be provided for the 2012 MER evaluation track – each containing clips from the MER event set listed above. These three MER data sets are as follows.

1. **MER Development Test Set** – This dataset is limited to 6 video clips from each of the 5 events in the MER event set, and is provided to support research and a dry run of the evaluation pipeline. There will be exactly 30 video clips in this dataset.
2. **MER Evaluation Test Set** – This dataset is limited to 6 video clips from each of the 5 events in the MER event set, and is provided to support the evaluation specified below. There will be exactly 30 video clips in this data set.
3. **MER Progress Test Set** – This dataset is defined for MER participants who also participate in MED. NIST will select exactly 30 positive video clips for evaluation (6 video clips from each of the 5 MER events).

MER participants are to generate a recounting for each of the 30 clips in the **MER Evaluation Test SET**.

MER participants that are also participating in MED (pre-specified) must additionally generate a recounting for all MED (pre-specified event) clips that their MED system identifies as being above their MED system's decision threshold for being positive, for all of the five MER events.

DRY RUN

All participants will be required to participate in a dry run exercise using the **MER Development Test SET** to ensure that both the system outputs are being generated as expected and are parsable by the evaluation pipeline. This exercise will also provide insight into how the recounting will be rendered for the judges in the formal evaluation.

EVALUATION

MER outputs from the **MER Evaluation Test SET** and from the **MER Progress Test SET** will be evaluated by judges (panels of experienced video analysts and possibly Linguistic Data Consortium staff). The two corpora, and each system, will be judged separately. The judges will perform two tasks.

First, without seeing the clips, the judges will attempt to identify which of the five events is represented by each MER output.

Next, for each MER event and each system separately, the judges will be provided with the six positive clips along with the output from a system and will attempt to match each recounting with the clip from which it was derived.

NIST will assess the MER outputs by analyzing how accurately the judges are able to perform the two tasks and the time required for the judging.

METRICS

Metrics for distinguishing one event from another, using only MER output:

The *system performance* metric for this subtask is the fraction of the judgments that correctly identified which of the five events is represented by each MER output, averaged across the *events* and judges.

The *event difficulty* (or confusability) metric for this subtask is the fraction of these judgments that were correct, averaged across the *systems* and judges.

In addition, we will compute the fraction of the judgments that were correct for each combination of system and event, averaged across only the judges.

Metrics for distinguishing which clip is described, using MER output plus the clips:

The *system performance* metric for this subtask is the fraction of the matches, of recountings to the clips from which they were derived, that were correct, averaged across the *events* and judges.

The *event difficulty* metric for this subtask is the fraction of the matches (of recountings to clips) that were correct, averaged across the *systems* and judges. This metric reflects the difficulty or confusability of the clips that were chosen for the event.

In addition, we will compute the fraction of the matches that were correct for each combination of system and event, averaged across only the judges.

Metrics for Judgment Time:

NIST will report the time required for judges to try to match MER outputs to events and to match MER outputs to clips.

Metric for System Speed:

As mentioned, participants will report the time required for their system to process each clip and generate the recounting. NIST will report that information as part of the TRECVID MER results.

SCHEDULE

Date (*all dates 2012*)	Milestone
April 13	Final design published in the TRECVID Guidelines on the web
April 30	Explicit expression of interest due, via email to NIST. Participants will not be added after April 30th
May 18	Rendering tool and MER DTD available
June 5	MER Progress Set Test Data available
June 15	Initial prototype of Evaluation tools available
June 15	MER Development Test Set available
July 11	MER Evaluation Test Set available
July 15	Dry Run results due at NIST
July 15	Finalized Evaluation tools available
August 7	Dry Run results available to participants
September 18	MER submissions due at NIST * MER Evaluation Test Set * MER Progress Test Set
October 17	Evaluation results available to participants
November 26-28	TRECVID Workshop

INPUTS/OUTPUTS

Input data formats will be as in existing HAVIC data.

MER output data formats will be ASCII XML text. NIST will provide a rendering tool and a MER DTD schema to be used to specify and validate system output.