DRAFT
2010 TRECVID MULTIMEDIA EVENT DETECTION EVALUATION PLAN

1 Overview

This document presents the evaluation plan for Multimedia Event Detection (MED) track for the TRECVID 2010 evaluation. The multi-year goal of MED is to assemble core detection technologies into a system that can quickly and accurately search a multimedia collection for user-defined events. An event for MED is "*an activity-centered happening that involves people engaged in process-driven actions with other people and/or objects at a specific place and time*".

A user searching for events in multimedia material may be interested in a wide variety of potential events. Since it is an intractable task to build special purpose detectors for each event a priori, a technology is needed that can take as input a definition of the event that a human can use to search a collection of multimedia clips. The MED evaluation series will define events via an **event kit** which consists of:

- An **event name** which is an mnemonic title for the event.
- An **event definition** which is a textual definition of the event.
- An **evidential description** which is a textual listing of the attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it is not a exhaustive list nor is it to be interpreted as required evidence.
- A set of **illustrative video examples** each containing an instance of the event. The examples are illustrative in the sense they help form the definition of the event but they do not demonstrate all the inherent variability or potential realizations.

The following topics are discussed below:

- Video source data
- The evaluation task
- Evaluation measures
- Evaluation Infrastructure
- Schedule

2 Video Source Data

A new collection of Internet multimedia (i.e., video clips containing both audio and video streams) will be provided to registered MED participants. The data, which was collected by the Linguistic Data Consortium, consists of publically available, user-generated content posted to the various Internet video hosting sites. Instances of the events were collected by specifically searching for target events using text-based Internet search engines. All included data has been reviewed for privacy and offensive material.

Video clips will be provided in MPEG-4 formatted files. The video will be encoded to the H.264 standard. The audio will be encoded using MPEG-4's Advanced Audio Coding (AAC) standard.

The video data collection will be divided into two data sets:

1. Development data consisting of 1746 total clips (~56 hours). The dev data set includes nominally 50 instances of each of the three MED '10 events and the rest of the clips are not on any of the three MED events.
2. Evaluation data consisting of 1742 total clips (~59 hours). The eval data set will include instances of the three events but the actual number of instances will not be release until sub evaluation.

The Linguistic Data Consortium will be the distribution point for the corpus. See the MED '10 web site, http://www.nist.gov/itl/iad/mig/med10.cfm, for licensing and acquisition instructions. The provided resources will include video clips, MED event annotations, and ancillary metadata for each clip.

Allowable side information (i.e., "contextual" information) will be provided in CSV (comma separated values) data tables.

3 Evaluation Task

The MED task is: given an Event Kit, find all clips that contain the event in a pre-indexed metadata store of the search corpus.

The MED task is a "multimedia" task in that systems will be expected to detect evidence of the event using either or both the audio and video streams of the clips. The events used for the MED '10 evaluation can be found on the MED '10 web site. Participant may implement systems for one or all of the specified events.

4 Evaluation Infrastructure

Systems will be evaluated on how well they can detect MED event instances in the evaluation corpus. The determination of correct detection will be at the clip level, i.e. systems will provide a response for each clip in the evaluation corpus. For testing purposes, each event will be considered independent.

System detection performance is measured as a tradeoff between two error types: missed detections (MD) and false alarms (FA). The two error types will be combined into a single error measure using the Normalized Detection Cost (NDC) model, which is a linear combination of the two errors. The NDC model distills the needs of an application profile into a set of predefined constant parameters that include the event priors and weights for each error type. The single operation point characterized by the NDC model is a small window into the performance of an event detection system. In addition to NDC measures, Detection Error Tradeoff (DET) curves [2] will be produced to graphically depict the tradeoff of the two error types over a wide range of operational points. The NDC model and the DET curve are related: the NDC model defines an optimal point along the DET curve.

The rest of this section defines the the input files to the systems, and the system output , followed by the two steps of the evaluation process: Decision Error Tradeoff (DET) curve production, and NDC computations.

4.1 System Inputs

Inputs to the system will be controlled by a Trial Index file. A Trial Index file is Comma

Separated Value (CSV) (see Appendix C for the CSV file format specification) file that specifies the detection trials a system must perform. Each line in the trial index contains a single detection trial. Each trial consists of the following three CSV fields:

- Field 1, Name "**TrialD**": a unique ID for the trial. It consists of the clip ID and event name.
- Field 2, Name "**ClipID**": the ID of the clip for which the system must provide a detection output.
- Field 3, Name "**Event**": the name of the event the system is expected to detect. The three values for MED '10 are: "*batting_in_run*", "*making_cake*", "*assembling_shelter*".

The trial index file will only contain the three CSV columns using the field names above.

4.2 System Outputs

Systems will record system outputs for each detection trial in a CSV formatted file. The system will generate the following fields for each detection trial and place them in a CSV record:

- Field 1, Name "**TrialID**": The trialID copied from the input trial index file.
- Field 2, Name "**Score**: A numeric score indicating how likely the event observation exists with more positive values indicating more likely observations.
- Field 3, Name "**Decision"**: A Boolean value ("y" or "n") indicating whether or not the event observation should be counted for the primary metric computation.

The decision scores and actual decisions permit performance assessment over a wide range of operating points. The decision scores provide the information needed to construct the DET curve. The actual decisions provide the mechanism for the system to indicate which putative observations to include in the NDC calculation: i.e., the putative decisions with a *true* actual decision.

Systems must ensure their decision scores values form a non-uniform density function so that the relative evidential strength between two putative terms is discernable. Second, the density function must be consistent across events for a single system so that event-averaged measures using decision scores are meaningful.

For the 2010 evaluation, the decision scores do not have to be consistent across events therefore a system may have a separate threshold for differentiating *true* and *false* actual decisions for each event.

Since developers may chose which events to build systems for, the generated CSV file for a system should only include TrialIDs for events for which the system was built. Please note that the evaluation code requires all TrialIDs of an attempted event to be present within the file.

4.3 Detection Error Tradeoff Curves

Graphical performance assessment uses a Detection Error Tradeoff (DET) curve that plots the system's missed detection probabilities (P_{Miss}) and false alarm probabilities (P_{FA}) that are a function of a detection threshold, Θ. This Θ is applied to the system's detection scores meaning the clips with decision scores above the Θ are 'declared' to be the set of detected instances. After Θ is applied, the following measurements are then computed separately for each event. The per-event formulas for P_{Miss} and P_{FA} are:

$$P_{Miss}(S,E_i,\Theta) = \frac{N_{Miss}(S,E_i,\Theta)}{N_{Targ}(E_i)}$$

$$P_{FA}(S,E_i,\Theta) = \frac{N_{FA}(S,E_i,\Theta)}{N_{NonTarg}(E_i)}$$

Where:

$$N_{Miss}(S,E_i,\Theta) = number\ of\ missed\ detections\ for\ system\ S, event\ E_i\ at\ decision\ score\ \Theta$$
$$N_{Target}(E_i) = number\ of\ clips\ containing\ event\ instances\ for\ event\ E_i$$
$$N_{NonTarg}(E_i) = number\ of\ clips\ that\ do\ not\ contain\ event\ instances\ for\ event\ E_i$$
$$N_{FA}(S,E_i,\Theta) = number\ of\ false\ alarms\ for\ event\ E_i at\ decision\ score\ \Theta$$

4.4 DCR Computations

The evaluation will use the Normalized Detection Cost (*NDC*) measure for evaluating system performance. *NDC* is a weighted linear combination of the system's probabilities of Missed Detection and False Alarm. The measure's derivation can be found in Appendix A and the final formula is summarized below. NIST will report an NDC for each event and not average them over events.

$$NDC(S,E_i,\Theta) = Cost_{Miss} \cdot P_{Miss}(S,E_i,\Theta) \cdot P_{Targ} + Cost_{FA} \cdot P_{FA}(S,E_i,\Theta) \cdot (1 - P_{Targ})$$

Where:

$$\mathrm{E_i} = \text{the i}^{\text{th}}\text{ event}$$
$$Cost_{Miss} = 80\ a\ constant\ defining\ the\ cost\ of\ a\ missed\ detections.$$
$$Cost_{FA} = 1\text{: } a\ constant\ defining\ the\ cost\ of\ a\ false\ alarm.$$
$$P_{Target} = 0.001\text{: } a\ constant\ defining\ the\ a\ priori\ rate\ of\ event\ instances.$$

The measure's unit is in terms of cost per clip used. NDC has been normalized so that an *NDC*=0 indicates perfect performance and an *NDC*=1 is the cost of a system that provides no output, i.e. $P_{Miss}=1$ and $P_{FA}=0$.

Two versions of the NDC will be calculated for each system: the Actual NDC and the Minimum NDC.

4.4.1 Actual NDC

The Actual NDC is the primary evaluation metric. It is computed by counting clips with *true* actual decisions as clips the system declares to contain the event.

4.4.2 Minimum DCR

The Minimum NDC is a diagnostic metric. It is found by searching the DET curve for the Θ with the minimum NDC. The difference between the value of Minimum NDC and Actual NDC indicates the benefit a system could have gained by selecting a better threshold.

5 Submission of results

Submissions to NIST will be required only to allow NIST to perform a system-mediated improvements to the test set ground truth.

Submissions will be made via ftp according to the instructions in Appendix B. In addition to the system output, NIST requests a system description be supplied for each submission. This description should include: a description of the hardware used to process the data, computational resources (cpu runtime, memory footprint, etc.) and a description of the architecture and algorithms used in the system such as the features or reasoning process.

6 Schedule

Consult the main schedule on the TREVID 2010 web site http://www-nlpir.nist.gov/projects/tv2010/#schedule.

7 References

[1] Harold W. Kuhn, "The Hungarian Method for the assignment problem", *Naval Research Logistic Quarterly*, **2**:83-97, 1955.

[2] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., “The DET Curve in Assessment of Detection Task Performance”, Eurospeech 1997, pp 1895-1898.

Appendix A: Derivation of Normalized Detection Cost

Normalized Detection Cost (*NDC*) is a weighted linear combination of the system's Missed Detection and False Alarm probabilities. The constant parameters of NDC, which are specified below, represent both the richness of events in the source data and the relative detriment of particular clip detection errors to a hypothetical application.

The cost of a system begins with the cost of missing an event ($Cost_{Miss}$) and the cost of falsely detecting an event ($Cost_{FA}$). $N_{Miss}(S,E)$ is the number of missed detections for system *S*, event *E*. $N_{FA}(S,E)$ is the number of false alarms for the same system and event.

$$DetectionCost(S,E) = Cost_{Miss} \cdot N_{Miss}(S,E) + Cost_{FA} \cdot N_{FA}(S,E)$$

To facilitate comparisons across systems and test sets, we divide Detection Cost by the number of video clips $\mathrm{N}_{\mathrm{Trials}}$.

$$\begin{aligned} DetectionCost(S,E) &= \frac{\mathrm{Cost}_{\mathrm{Miss}} \cdot N_{Miss}(S,E) + Cost_{FA} \cdot N_{FA}(S,E)}{\mathrm{N}_{\mathrm{Trials}}} \\ &= \mathrm{Cost}_{\mathrm{Miss}} \cdot \frac{\mathrm{N}_{\mathrm{Miss}}(\mathrm{S,E})}{\mathrm{N}_{\mathrm{Trials}}} + \mathrm{Cost}_{\mathrm{FA}} \cdot \frac{\mathrm{N}_{\mathrm{FA}}(\mathrm{S,E})}{\mathrm{T}_{\mathrm{Trials}}} \\ &= \mathrm{Cost}_{\mathrm{Miss}} \cdot \frac{\mathrm{N}_{\mathrm{Miss}}(\mathrm{S,E})}{\mathrm{N}_{\mathrm{Targ}}(\mathrm{E})} \cdot \frac{\mathrm{N}_{\mathrm{Targ}}(\mathrm{E})}{\mathrm{N}_{\mathrm{Trials}}} + \mathrm{Cost}_{\mathrm{FA}} \cdot \frac{\mathrm{N}_{\mathrm{FA}}(\mathrm{S,E})}{\mathrm{N}_{\mathrm{NonTargTrials}}} \cdot \frac{\mathrm{N}_{\mathrm{NonTargTrials}}(\mathrm{S,E})}{\mathrm{N}_{\mathrm{Trials}}} \\ &= Cost_{Miss} \cdot P_{Miss}(S,E) \cdot P_{Target}(E) + Cost_{FA} \cdot R_{FA}(S,E) \cdot (1 - P_{Target}(E)) \end{aligned}$$

$P_{Target}(E)$ is the probability of a clip containing the event. This value is dependent on the event but providing this prior to a system for each event changes the definition of an event – it includes the event definition and the prior. Instead, we replace the event-dependent prior with a single, global prior, P_{Target}, that in combination with the $Cost_{Miss}$ and $Cost_{FA}$ reflects the characteristics of an application profile. Since the evaluation corpus will have an engineered richness, the single prior is warranted. The modified formula becomes:

$$DetectionCost(S,E) = \; Cost_{Miss} \cdot P_{Miss}(\mathrm{S,E}) \cdot P_{Target} + Cost_{FA} \cdot P_{FA}(\mathrm{S,E}) \cdot (1 - P_{FA}(\mathrm{S,E}))$$

The range of the DCR_{Sys} measure is [0,∞). To ground the costs, a second normalization scales the cost to be 0 for perfect performance and 1 to be the cost of a system that provides no output (either providing no output, $P_{Miss} = 1$ and $P_{FA} = 0$, or declaring every clip to be an instance). The resulting formula is the Normalized Detection Cost of a system (*NDC*).

$$NormDectectionCost(S,E) = \frac{DetectionCostRate(S,E)}{MINIMUM(Cost_{Miss} \cdot P_{Target}, Cost_{FA} \cdot (1 - P_{Target})\,)}$$

Appendix B: Submission Instructions

The packaging and file naming conventions for MED ’10 relies on **Experiment Identifiers** (EXP-ID) to organize and identify the files for each evaluation condition and link the system inputs to system outputs. Since EXP-IDs may be used in multiple contexts, some fields contain default values. The following section describes the EXP-IDs to be used for the development dataset (devset) and evaluation dataset (evalset).

The following EBNF describes the EXP-ID structure:

EXP-ID ::= <TEAM>_2010_MED_<DATA>_<SYSID>_<VERSION>

where,

<TEAM> ::= your TRECVID Team Name

<DATA> ::= either “DEV” or “EVAL”

<SYSID> ::= a site-specified string (that does not contain underscores) designating the system used.

The SYSID string must be present. It is to begin with p- for a primary system (i.e., your best system) or with c- for any contrastive systems. For example, this string could be p-baseline or c-contrast. This field is intended to differentiate between runs for the same evaluation condition. Therefore, a different SYSID should be created for runs where any changes were made to a system.

<VERSION> ::= 1..n (with values greater than 1 indicating multiple runs of the same experiment/system)

In order to facilitate transmission to NIST and subsequent scoring, submissions must be made using the following protocol, consisting of three steps: (1) preparing a system description, (2) packaging system outputs and system descriptions, and (3) transmitting the data to NIST.

B.1 System Descriptions

Documenting each system is vital to interpreting evaluation results. As such, each submitted system, (determined by unique experiment identifiers), must be accompanied by a system description with the following information:

Section 1. Experiment Identifier(s)

List all the experiment IDs for which system outputs were submitted. Experiment IDs are described in further detail above.

Section 2. System Description

A brief technical description of your system; if a contrastive test, contrast with the primary system description.

List all events processed on a single line as follows:

Events_Processed: *Event1 Event2 Event3 …*

Section 3. Training:

A list of resources used for training and development.

Section 4. References:

A list of all pertinent references.

B.2 Packaging Submissions

All system output submissions must be formatted according to the following directory structure:

```
output/<EXP-ID>/<EXP-ID>.txt
output/<EXP-ID>/<EXP-ID>.csv
```

where,

EXP-ID is the experiment identifier as described in section B.1,
<EXP-ID>.txt is the system description file as specified above (section B.2),
<EXP-ID>.csv is the CSV-formatted system output file

B.3 Transmitting Submissions

To prepare your submission, first create the previously described file/directory structure. This structure may contain the output of multiple experiments, although you are free to submit one experiment at a time if you prefer. The following instructions assume that you are using the UNIX operating system. If you do not have access to UNIX utilities or ftp, please contact NIST to make alternate arrangements.

First, change directory to the parent directory of your “output/” directory. Next, type the following command:

```
tar -cvf - ./output | gzip > <SITE>_<SUB-NUM>.tgz
```

where,

<SITE> is the ID for your site
<SUB-NUM> is an integer *1 to n*, where 1 identifies your first submission, 2 your second, etc.

This command creates a single tar/gzip file containing all of your results. Next, ftp to jaguar.ncsl.nist.gov giving the username 'anonymous' and (if requested) your e-mail address as the password. After you are logged in, issue the following set of commands, (the prompt will be 'ftp>'):

```
ftp> cd incoming
ftp> binary
ftp> put <SITE>_<SUB-NUM>.tgz
ftp> quit
```

Note that because the “incoming” ftp directory (where you just ftp’d your submission) is write protected, you will not be able to overwrite any existing file by the same name (you will get an error message if you try), and you will not be able to list the incoming directory (i.e., with the “ls” or “dir” commands). Please note whether you get any error messages from the ftp process when you execute the ftp commands stated above and report them to NIST.

The last thing you need to do is send an e-mail message to jfiscus@nist.gov, brian.antonishek@nist.gov and martial@nist.gov to notify NIST of your submission. The following information should be included in your email:

- the name of your submission file,
- the file size,
- a listing of each of your submitted experiment IDs.

Please submit your files in time for us to deal with any transmission errors that might occur well before the due date if possible. Note that submissions received after the stated due dates for any reason will be marked late.

Appendix C: Comma Separated Value File Format Specifications

The MED evaluation infrastructure uses Comma Separated Value (CSV) formatted files with an initial field header line as the data interchange format for all textual data. The EBNF structure the infrastructure uses is as follows:

```
CSVFILE :== <HEADER> <DATA>*

      <HEADER> :== <VALUE> {“,” <VALUE> }* <NEWLINE>
      <DATA> :== <VALUE> {“,” <VALUE> }* <NEWLINE>
      <VALUE> :== <DOUBLEQUOTE><TEXT_STRING><DOUBLEQUOTE>
```

The first data record in the files is a header line. The header lines are required by the evaluation infrastructure and the field names for the trial index file and the system output file are dictated by Sections 4.1 and 4.2.

Each header and data record in the table is one line of the text file. Each field value is delimited by double quotes and is separated from the next value with a comma.

An example trial index is:

```
"TrialID","ClipID","Event"
"72.assembling_shelter","72","assembling_shelter"
"72.batting_in_run","72","batting_in_run"
"72.making_cake","72","making_cake"
"285.assembling_shelter","285","assembling_shelter"
"285.batting_in_run","285","batting_in_run"
"285.making_cake","285","making_cake"
```

An example system output file is:

```
"TrialID","Score","Decision"
"72.assembling_shelter","0.062712","n"
"72.batting_in_run","0.978791","y"
"72.making_cake","0.115392","n"
"285.assembling_shelter","0.801007","y"
"285.batting_in_run","0.861036","y"
"285.making_cake","0.120700","n"
```