2016 TRECVID MULTIMEDIA EVENT DETECTION EVALUATION PLAN

1 Introduction

This is the evaluation plan for Multimedia Event Detection and Retrieval (MED) track of the 2016 TRECVID evaluation. The multi-year goal of MED is to support the creation of event detection and retrieval technologies that will permit users to define their own complex events and to quickly and accurately search large collections of multimedia clips.

- This year we'll be using a subset of Yahoo's YFCC100M dataset to supplement the evaluation search set
- 10 new Ad-Hoc events will be added

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

A typical MED system (shown in Figure 1) is defined to have three separate modules: 1) metadata generation, 2) event query generation, and 3) event search. This year, two types of event query generation modules will be tested.

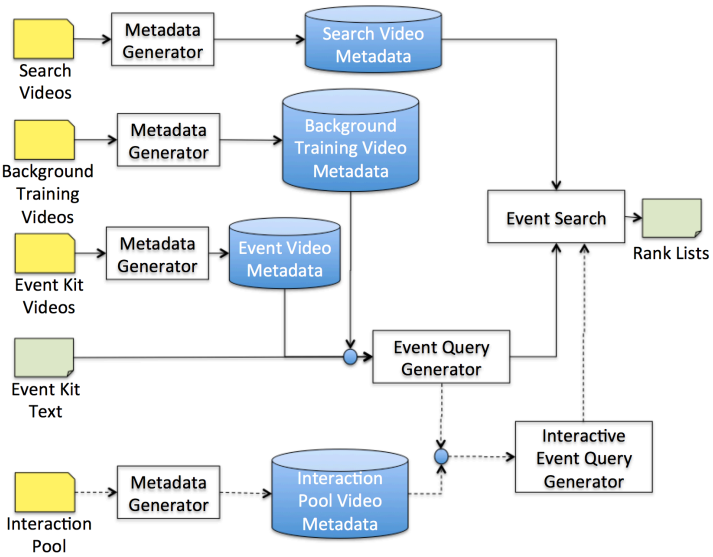

Figure 1: MED System Block Diagram: The MED processing modules are shown in blocks and data is depicted as icons. Folder icons represent a collection of videos while text-folders are text/csv documents. Dashed lines in the diagram show optional tasks. The "Event Search" of "Interactive Event Queries" is a separate is a separate run of the "Event Search" Module.

The event query generator and event search modules process each event independently. The modules must be run in the following order and are limited to only the inputs shown in the block diagram and described below:

1) The metadata generator extracts the video metadata for all videos in the input set. The metadata describes the content of the video that support the event search. Since this must be created prior to event query generation and event search, the metadata must be sufficiently rich such that it could be searched successfully for any event. The metadata generator is used to process 1) the search videos, 2) a common set of background training videos, and 3) each set of example event videos independently.
2) The event query generator uses the 1) event kit text, 2) background training video metadata, and 3) event video metadata to create an event query. The event query contains all information needed by the event search module.
3) The event search module searches the search video metadata for the event represented by the event query and creates a ranking for each search video.
4) Optionally, after the event query generator has finished execution, Interactive Event Query Generator builds a second query that makes use of a large pool of video data (the 20,000 video Interaction Pool) and a site-developed interface for a human enhance the query. The human can spend at most 15 minutes to refine the query using only the videos in the interaction pool. The augmented query is then used by the event search module (but not in tandem with initial event query) as a separate run.

NIST will provide all data inputs defining the event(s) to the MED system for testing and evaluation. **No Non-NIST provided data (video or annotations on the video) may be used as input to the Event Query Generator for a given event. Further, the use of the data as inputs and outputs of the modules of the system must be strictly observed**. For example, the search video metadata store is not to be used as an input to the event query generator, only the event search.

Participants may use any computer system architecture to run the metadata generator. However, participants must use ONLY a COTS standard personal computing platform or a single core of a cluster to execute the event query generator and the event search. The event searches are to be performed locally and the input and the output of the searches, i.e., the event/video rankings, are to be submitted to NIST as specified below for official scoring and analysis.

The MED evaluation is open to all that find the task of interest and who are willing to abide by the rules of the evaluation.

2 Data Resources

Internet multimedia data (i.e., video clips containing both audio and video) from the HAVIC data resources will be provided to registered MED participants. The HAVIC data resources, collected by the Linguistic Data Consortium (LDC), consist of publically available, user-generated content. The LDC will be the distribution point for the collection. See http://www.nist.gov/itl/iad/mig/MED16.cfm for data format, licensing, and acquisition instructions.

New for this year, the Yahoo Flickr Creative Commons 100M dataset (YFCC100M) is a large collection of images and video available on Yahoo! Flickr. All photos and videos listed in the collection are licensed under one of the Creative Commons copyright licenses. While the YFCC100M dataset contains both images and videos, only a subset of the videos will be used for this evaluation.

MED data resources will consist of 1) test and evaluation event kits, 2) two evaluation search video

collections from both the HAVIC and YFCC100M datasets, 3) two test search video collections, and 4) research resources. These are described in more detail in the following subsections.

2.1 Event Kits

There will be 30 events in MED16 divided into two sets, 20 pre-specified **(PS15)** and 10 ad-hoc **(AH16)**, events. The event names of the PS events are listed in Table 1. The AH16 events will be revealed according to the TRECVID master schedule.

Pre-Specified Events for MED16	
From MED12	**From MED13**
E021 - Bike trick	E031- Beekeeping
E022 - Cleaning an appliance	E032 – Wedding shower
E023 - Dog show	E033 – Non-motorized vehicle repair
E024 - Giving directions	E034 – Fixing a musical instrument
E025 - Marriage proposal	E035 – Horse riding competition
E026 - Renovating a home	E036 – Felling a tree
E027 - Rock climbing	E037 – Parking a vehicle
E028 - Town hall meeting	E038 – Playing fetch
E029 - Winning race without a vehicle	E039 – Tailgating
E030 - Working on a metal crafts project	E040 – Tuning a musical instrument

Table 1: MED16 Pre-Specified Evaluation Events

The event kit text as a whole defines the event. The event kit text consists of:

- Event name: A mnemonic title for the event
- Event definition: A textural definition of the event.
- Event explication: An expression of some event domain-specific knowledge needed by humans to understand the event definition.
- Evidential description: A textual listing of some attributes that are often indicative of an event instance. The evidential description provides a notion of some potential types of visual, acoustic, and temporal evidences indicating the event's existence. This is not an exhaustive list nor is it to be interpreted as required evidence.

Events kits may also be accompanied by a set of event video examples that illustrate some, but not all, possible realizations of the event. The evaluation will support event queries with a varying number of positive event video examples.

1. **0Ex**: No example video clips will accompany the event kits.
2. **10Ex**: 10 positive clips and up to 5 miss clips (i.e., non-positive clips) will accompany the event kits. This is the eventual goal of a MED system.
3. **100Ex**: 100 positive clips, and up to 50 miss clips will accompany the event kits.

To support training of an event, a common set of background training videos will be defined as input to the event query generator. Note that both the set of example event videos and the background training videos will be run through the metadata generator before the event query generation is run.

Participants are allowed to use the pre-specified event kits, event example videos, and the background training videos for testing their system development. The ground truth labels will be provided in "reference database" (defined in section 3.1) containing only markings for positive and miss clips.

2.2 Evaluation Search Videos

Participants will be provided a set of evaluation search videos. **MED16EvalFull**, which is comprised of the Progress collection and a subset of the YFCC100M collection consisting of approximately 200,000 videos, or **MED16EvalSub**, a 32,000 video subset of the MED16EvalFull collection. The Progress Search Set will be used for "blind" testing for MED12, MED13, MED14, MED15 and MED16 and therefore ground truth are not provided with this data.

Teams participating in the Pre-Specified Event task will have the choice of processing either MED16EvalFull, or the smaller MED16EvalSub set. Teams participating in the Ad-Hoc Event must process the full set, MED16EvalFull. [1]

Participants must adhere to the same rules as outlined for the test data and the following additional rules for any evaluation search video sets:

- Participants must not attempt to gain knowledge of an evaluation search set properties or content by manually inspecting the video, clip metadata, output of the processing, or statistics developed during the processing.
- The evaluation search sets are only to be used as part of an official evaluation submission to NIST MED evaluation.
- The evaluation search videos and their metadata stores must be deleted from systems (or made inaccessible) after each year's MED evaluation ends and re-copied and metadata generation applied the following year.

2.3 Research Resources

The goal of MED is to perform a general ad-hoc search for an event using few exemplars. Teams will conduct research to achieve this goal. To this end, NIST/LDC will provide some data for the purposes of research:

1. Research video data with ground truth,
2. Research event kits for E001-E005 and E016-E020.

Participants may use the research resources for research and development of their system software components. These resources may be altered, amended or annotated in any way participants need to facilitate their research. However, none of these resources may be annotated in a public forum, (e.g., Amazon Mechanical Turk) unless the data is shared with all participants in MED before used by the team collecting the annotations. Most importantly, none of the research resources or team added annotations may be used as direct inputs to the MED system.

Teams, at their option, may download their own research data from the Internet, exclusive of the YFCC100M corpus, as long as they comply with all legal, licensing, and end-user agreements for the data, and use it to build/train systems. Downloaded videos MUST NOT be used to extend the pool of event videos for search in any way.

3 Evaluation Runs

[1] Unlike the PS evaluation, the reference annotations for the Ad-Hoc events will be produced via pooled assessment which is simpler when all teams process the same data set. Thus the full set is required.

A "run" for MED'16 is specified by the combination of Event Search Type, Search Set, and Event Exemplar set. The full set of runs is identified in the table below. Required runs are indicated by "*".

Event Search Type (Tasks)	Search Set	Event Exemplar Set	
		Required	Optional
Pre-Specified (Option of Search Set)	MED16EvalFull	10Ex	0Ex
	MED16EvalSub		100Ex
Ad-Hoc	MED16EvalFull	10Ex	
Interactive Ad-Hoc	MED16EvalFull	10Ex	

The required run will be designated as the team's 'primary' run during the submission process. Teams participating in the Interactive Ad-Hoc task must also participate in the Ad-Hoc task.

Teams will be allowed to submit an additional 4 Pre-Specified contrastive runs and 1 Ad-Hoc contrastive run.

After the Ad-Hoc submission deadline, NIST will decide how best to use the Ad-Hoc runs to develop the assessment pools.

3.1 Event Training Inputs

Event training inputs will be specified through a set of three Comma Separated Value (CSV)[2] tables which will be provided (and named) for each evaluation condition (e.g., 100EX, 10Ex, and 0EX)[3]:

- The Event Database - *_EventDB.csv files
  This two-column table defines the **EventID** and **EventName**.

- The Clip Database - *_ClipMD.csv files
  This five-column table contains the metadata for each clip in the collection. The fields are **ClipID**, **MEDIA_FILE, CODEC, MD5SUM, DURATION**.

- Judgment Database - *_JudgmentDB.csv
  This table specifies the exemplars to be associated with event kits. The three columns with data are ClipID, EventID, and INSTANCE_TYPE (either "positive" or "near_miss"). Note that 10Ex exemplars are also included in the 100Ex condition but the 100Ex videos cannot be included in the 10Ex condition.

3.2 System Outputs and Documentation

Teams are allowed to submit runs as specified above. For each submitted run to be considered complete and valid, participants must provide the information requested in this section. The requested information is a crucial element required for the research community to properly interpret the performance results.

[2] See Appendix C for the CSV file format specification.
[3] These tables are the authoritative sources for system inputs. Participants should avoid using directory listings of the collections as inputs because the tables are an experimental control mechanism.

3.2.1 System Description

The purpose of the system description document is to provide a list of the resources and techniques used to build the MED system and identify the computing resources and time required to process the test set.

Each submission must include a system description which describes:

- the algorithms used for each of the modules,
- the hardware components used and computation times of the metadata generation, event query generation, and event search modules
- the metadata store size

See Appendix B section B.1 for the template that covers the ***minimum*** requirements of the system description document.

3.2.2 Event Search Reporting

A MED system processes the metadata store detecting instances of each event independently. For each system submission, two output files must be created using Experiment Identifiers (**EXP-ID**) which describe the characteristics of the run (see Appendix B for the definition of *EXP-ID*):

1. <EXP-ID>.threshold.csv

   Each line will contain information pertaining to the processing of a single event. Events not processed should not be included in the file. The 3 fields in this file are as follows.

   - **EventID**: The Event ID processed - copied from the event database file.
   - **DetectionTPT**: A value indicating the number of hours used during the *event search phase* for the event.
   - **SEARCHMDTPT:** A value indicating the number of hours used to generate the metadata for the "Search Videos". The value MUST be identical for all events.

2. <EXP-ID>.detection.csv

   Each line will contain information pertaining to the processing of a single trial. The 2 fields in this file are as follows.

   1. **TrialID**: The Trial ID processed - copied from the input trial index file.
   2. **Rank**: The rank of the video for the given event search with values between 1 (most likely video to contain the event) to N (least likely video to contain the event)

4 Evaluation Measures

System output will be evaluated by how well the system retrieves and detects MED events in evaluation search video metadata and by the computing resources used to do so. The determination of correct detection will be at the clip level, i.e. systems will provide a response for each clip/event pair in the evaluation search video set. Participants must process each event independently in order to ensure each event will be scored independently.

For each event, a MED system produces a threshold value and an event confidence score between 0.0

and 1.0 for each video in the search set. This score will be used to rank the videos, in descending order. Using ground truth, the rank of each true-positive event video will be computed by NIST forming a rank vector, **rank**(tp) where tp = 1 to P_E and P_E is the total number of positives for the event. Videos with identical scores will be ranked randomly.

Precision and Recall can be computed for each position in the rank vector. Recall, **Recall**(tp) is the index of the rank vector tp divided by the total number of positive videos P_E. Precision, **Prec**(tp), is the index of the rank vector, tp, divided by the rank of that positive, **rank**(tp).

4.1 Primary Measures

Retrieval Metric for the Pre-Specified Event task: Mean Average Precision, MAP

For a set of events (i.e., PS16), the mean average precision (*MAP*) score will be computed as:

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP\,(q)$$

where Q denotes the number of different events and *AP*(q) is the average precision for event q.

$$AP(q) = \frac{1}{P_E}\sum_{tp=1}^{P_E} Prec(tp) = \frac{1}{P_E}\sum_{tp=1}^{P_E} \frac{tp}{rank(tp)}$$

where P_E denotes the number positives of event q.

Retrieval Metric for the Ah-Hoc Event task: Mean Inferred Average Precision, InfAP

The Ad-Hoc systems will be evaluated using the Mean Inferred Average Precision metric as described by Yilmaz et al. in "Estimating Average Precision with Incomplete and Imperfect Judgments".

InfAP will also be implemented on the Pre-Specified Event submissions as a contrastive measure.

Metadata Generation Processing Speed

NIST will report the **real-time** factor to complete all steps necessary to build the metadata store.

Real-time factor is the Total Processing Time (TPT) for the process (as reported in the system description) divided by the number of hours of video in the test collection. TPT is the wall clock time (in hours) used during a computation phase, including I/O, from start to finish. The processing time for parallelized sub-steps adds to TPT as a single step. The processing time for metadata "shared" across sites, e.g., speech transcription, person tracking, etc. (including time used to incorporated data into the metadata store) adds to TPT as a single-step.

Event Query Generation Processing Speed

NIST will report the *real-time* factor for each event processed during the event query generation phase.

Event Search Processing Speed

NIST will report the *real-time* for each event processed during event search.

4.2 Evaluation Tools and Command Line Example

NIST will use the Detection EVAluation (DEVA) tools within the NIST Framework for Detection Evaluation (F4DE) toolkit and other tools to score the evaluation submissions. Usage instructions will be posted on the MED 16 web site.

5 Result Submission Instructions

See Appendix B

6 Schedule

For TRECVID related schedule information please consult the main schedule on the TREVID 2016 web site.

7 References

"Estimating Average Precision with Incomplete and Imperfect Judgments", Emine Yilmaz and Javed A. Aslam. Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management (CIKM). November, 2006.

Appendix B: Submission Instructions

The packaging and file naming conventions for MED16 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.

The following EBNF describes the EXP-ID structure:

EXP-ID::=<TEAM>_ MED16_ <SEARCH>_ <EVENTSET>_<EKTYPE>_<SMGHW>_<SYS>_<VERSION>
where

<TEAM> ::= your Short TRECVID Team Name. "+" or "_" characters must be removed.

<SEARCH> ::= either "MED16EvalFull" or "MED16EvalSub"

<EVENTSET> :== either "PS", "AH", "iAH" for Pre-Specified, Ad-Hoc, and Interactive Ad-Hoc tasks respectively as defined in Section 2.1.

<EKTYPE> :== either "100Ex", "10Ex", "0Ex"

<SMGHW> ::= either "SML", "MED", "LRG" indicating the "class" of hardware used to process the search collection. The value to select is the closest match to the following categories:
"SML" :== 100 CPU cores + 1,000 GPU cores
"MED" :== 1,000 CPU cores + 10,000 GPU cores
"LRG" :== 3,000 CPU cores + 30,000 GPU board

<SYS> ::= a site-specified string (that contains letters, dashes, and numbers) designating the system used.

The SYSID string must be present. It is to begin with "p-" for the one and only primary system (i.e., your single best system) or with "c-" for any contrastive systems. It is then followed by an identifier for the system (only alphanumerical characters allowed, no spaces). 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 used 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.

Section 3 Metadata Generation System Hardware Description and Runtime Computation

Describe the hardware setup(s) to perform the metadata generation phase and report the number of CPU and GPU cores used to determine the <SMGHW> for the system.

A hardware setup is the aggregate of all computational components used to perform this phase. Examples of a system might be: a 16-node, Dual Quad Core 2.26 GHz Intel Xeon, 24GB RAM per node, with a 10TB Data Server.

Section 4 Event Search Hardware Description

Describes the computing hardware, including the number of CPU cores, used for executing the event search(s).

The hardware setup is the aggregate of all computational components used to perform this phase. This hardware platform must be limited to a COTS standard personal computing platform.

Section 5 Training data and knowledge sources

Lists the resources used for system development, and runtime knowledge sources beyond the provided MED corpora.

Section 6 References

A list of pertinent references.

B.2 Packaging Submissions

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

output/<EXPID>/<EXPID>.txt
output/<EXPID>/<EXPID>.detection.csv
output/<EXPID>/<EXPID>.threshold.csv

where,

EXPID is the experiment identifier as described in Section B.1,
<EXPID>.txt is the system description file as specified above (Section B.1),
<EXPID>.detection.csv is the CSV-formatted system output file containing the detection scores for each TrialID (see Section 3.2.2).
<EXPID>.threshold.csv is the CSV-formatted system output file containing the detection thresholds and processing speed measurements (see Section 3.2.2)

Multiple EXPIDs may be present under the "output/" directory.

B.3 Validating the Submission

The F4DE distribution contains a submission checker that validates the submission both at a syntactic and semantic level. Participants should check their submission prior to sending it to NIST. NIST will reject submissions that do not pass validation. The TRECVID MED16 Scoring Primer document (DEVA/doc/TRECVid-MED16-ScoringPrimer.html within the F4DE release) contains instructions for how to use the validator. NIST will use the following command line to validate MED submission files.

%TV16MED-SubmissionChecker --TrialIndex <DATA>_TrialIndex.csv \
MED16_testTEAM_MED16EvalFull_PS _2.tar.bz2

B.4 Transmitting Submissions

To prepare your submission, first create the previously described file/directory structure. 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 > MED16_<TEAM>_<SEARCH>_<EVENTSET>_<SUBNUM>.tgz

where,

<TEAM>, <SEARCH>, and <EVENTSET> are the same as defined above.

<SUBNUM> is an integer *1 to n*, where 1 identifies your first submission, 2 your second, etc.

For example: MED16_testTEAM_MED16EvalFull_PS_2.tgz

Important: only the latest submission will be used for scoring, but a submission file can contain multiple EXPIDs.

This command creates a single tar/gzip file containing all of your results. After shipment to NIST (as described in the next step), NIST will validate your submission with a syntactic and semantic validator.

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> epsv4 off
ftp> passive auto
ftp> put MED16_<TEAM>_<SEARCH>_<EVENTSET>_<SUBNUM>.tgz
ftp> quit

Note: the "epsv4 off' is designed to bypass a limitation of the FTP server, and should return: "EPSV/EPRT on IPv4 off.". If your prompt returns something different it is likely that your system does not support this feature, and therefore it is not needed.

Note: the "passive auto" should return "Passive mode: on; fallback to active mode: on".

Note that because the "incoming" ftp directory (where you just ftp-ed 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 med_poc@nist.gov to notify NIST of your submission. The following information should be included in your email:

- the name of your submission file,
- the md5 of the file
- 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 (*_TrialIndex.csv”):

```
"TrialID","ClipID","EventID"
"72.P001","72","P001"
"72.P002","72","P002"
"72.P003","72","P003"
"285.P001","285","P001"
"285.P002","285",”P002"
"285.P003","285","P003"
```

An example system output file is (“.detection.csv):

```
"TrialID","Rank"
"72.P001","1"
"72.P002","3"
"72.P003","2"
"285.P001","3"
"285.P002","2"
"285.P003","1"
```

An example threshold system output file is (*.threshold.csv):

```
"EventID",”DetectionTPT”,”SEARCHMDTPT
```

```
"P001","0.54",”5923.3”
"P002","0.74",”5923.3”
```