

Straw Man Proposal for the TRECVID 2008 Evaluation

Overview:

This paper presents a “straw man” proposal for evaluating event detection in surveillance video for TRECVID 2008. The goal of the evaluation will be to build and evaluate systems that can detect instances of a variety of observable events in the airport surveillance domain. The video source data to be used is a ~100-hour corpus of video surveillance data collected by the UK Home Office at the London Gatwick International Airport. This corpus will be divided temporally into development and evaluation subsets.

Two event detection tasks will be supported: a retrospective event detection task run with complete reference annotations and a “freestyle” experimental analysis track to permit participants to explore their own ideas with regard to the airport surveillance domain.

Because this is an initial effort, the evaluation will be run as more of an experimental test-bed than a formal bakeoff. By doing so, we propose two changes to the normal, common evaluation paradigm. First, the entire source video corpus will be released at the beginning of the project so that research can begin immediately. Participants will be on the “honor” system to keep the evaluation set as blind as possible. Second, two sets of events will be defined: a required set defined by NIST and the LDC whose descriptions and annotations will be released quickly for research to begin and an optional, secondary set of events nominated by participants. The development resources for nominated events will be released later in the year. These steps will hopefully encourage an acceleration of the research and knowledge sharing and will permit faster evolution of the evaluation paradigm.

The following topics are discussed below:

- Video source data
- Evaluation tasks
- Event definitions
- Tentative schedule

Video Source Data:

The source data will consist of 100 hours (10 days * 2 hours/day * 5 cameras), obtained from Gatwick Airport surveillance video data (courtesy of the UK Home Office). The corpus will be divided into development and evaluation subsets. In particular, the first 5 days of the corpus will be used as the development subset (devset), and the second 5 days of the corpus will be used as evaluation subset (evalset).

Developers may use the devset in any manner to build their systems including activities like sub-dividing it into internal test sets, jackknifed training, etc. During the summer months, NIST will conduct a dry-run evaluation using the devset as the video source.

While testing on the development data in a non-blind system test, the purpose of the dry run: to test the evaluation infrastructure, is most easily accomplished using the devset.

We plan to release the full corpus (devset + evalset) early in the evaluation cycle to give people the opportunity to preprocess the full corpus throughout the summer. The evaluation set must not be inspected or mined for information until after the evalset annotations are released. The evalset restriction applies to both evaluation tasks. However, participants can run feature extraction programs on the evalset to prepare for the formal evaluation.

During the formal evaluation, we propose that the system process both the devset and evalset (i.e., the entire video corpus) so that we can characterize system performance on the devset. When results are reported, both error rates will be reported as separate measurements.

Evaluation tasks:

This proposal includes the following evaluation tasks:

- Retrospective Event Detection: The task is to detect observations of events based on the event definition. Systems may process the full corpus using multiple passes prior to outputting a list of putative events observations.
- "Free-Style" Exercise. The purpose of this exercise is to support innovation and exploration of event detection in ways not anticipated by the above tasks. Freestyle participants must define tasks that are pertinent to the airport video surveillance domain and that can be implemented on this data set. Freestyle submissions must include rationale, clear definitions of the task, performance measures, reference annotations and a baseline system implementation.

Evaluation measures:

The discussion of metrics is deferred for a future version of the evaluation plan.

Events:

The TRECVID Event Detection evaluation is a pilot evaluation. One of the main goals of the evaluation is to explore how to define an event for the video domain. Initially, a video event is defined to be “an observable action or change of state in a video stream that would be important for airport security management”. Events may vary greatly in duration, from 2 frames to longer duration events that can exceed the bounds of the excerpt.

Events will be described through an “event description document”. The document will include a textual description of the event and a set of exemplar event occurrences (annotations). Each exemplar will indicate the source file and time coordinates of the event. Events used in the evaluations will be annotated over the full corpus.

Events will be considered independent events for the evaluation. Therefore systems will process the video stream separately for each event and not make use of any information gleaned from the set of events.

There will be two sets of events: Required events and Optional events. There is no implicit difference between the types of events included in the event sets. Rather, the event sets represent phased release of event definitions that (1) allow us to quickly develop a set of events and release their event descriptions, and (2) a second set to elicit community input for event definitions, requiring a delayed release of event descriptions. As the names suggest, all participants must run their systems on the required events whereas participants have the option to run their systems on the optional events.

The proposed schedule for event definitions and data release is as follows:

	Required Event Set	Optional Event Set
Event Selection	By LDC and NIST	Nominated by community input
Event Description Release	March 1	May 15
Development Annot. Release	June 1	July 1
Test Set Annot. Released	Oct. 1	Oct. 1
Participation	Required	Optional

Tentative Schedule:

The 2008 evaluation schedule for event detection includes the following milestones:

- Jan.--Feb.: Event detection planning & telecoms
- Feb.: Call for participation in TRECVID
- Mar. 1: Release of video data, required event definitions, and examples
- Mar. 30: Final evaluation plan & guidelines written
- Apr. 4: Call for participation in event detection
- Apr. 11: Deadline to commit
- May 1: Nominations for candidate events end
- May 15: Release of optional event definitions
- June 1: Release scoring tool
- June 1: Development annotations for required events released
- June 1: Dry Run test set specified
- July 1: Development annotations for nominated events released
- July: Dry run (systems run on Dev data)

Sept. 26: Obtain submissions for formal evaluation
Oct 1: Release of all annotations
Oct. 1: Distribute preliminary results
Oct. 10: Distribute final results
Oct. 27: Notebook papers due at NIST
November 17-18: Present results at TRECVID