2008 TRECVid Event Detection Straw Man Evaluation Plan

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. By doing so, we propose two changes to the typical evaluation paradigm. First, the entire source video corpus will be released early so that research can begin immediately. Participants will be on the honor system to keep the evaluation set blind. 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, (event definitions and annotations), 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
- Evaluation measures
- Event definitions
- 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 the 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 will release the full corpus (devset + evalset) early in the evaluation cycle to give people the opportunity to preprocess the full corpus throughout the year. 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 each. 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. The primary condition for this task will be single-camera input (i.e., the camera views are processed independently). Multiple-camera input may *optionally* be run as a contrastive condition.

- "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 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:
We propose to use the Average Normalized Detection Cost Rate (*ANDCR*) as the primary metric for evaluating system performance. *ANDCR* is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per unit time). The measure's derivation can be found in Appendix A.

ANDCR calculates performance over an ensemble of events by averaging the Missed Detection probabilities for all events with at least one true event occurrence ($N_{EventsNZ}$) and averaging the False Alarm Rates over all events separately. By separating the two averages, the measure can incorporate events with no true occurrences while remaining defined.

$$ANDCR(S) = \frac{\sum_i^{N_{EventsNZ}} P_{Miss}(S, E_i)}{N_{EventsNZ}} + Beta * \frac{\sum_j^{N_{Events}} R_{FA}(S, E_j)}{N_{Events}}$$

Where:

$$\begin{aligned}
\mathrm{N_{EventsNZ}} &= \text{number of event with } N_{Targ}(E_i) > 0 \\
N_{Events} &= n\text{umber of events} \\
N_{Miss}(S, E_i) &= number\ of\ missed\ detections\ for\ system\ S, event\ E_i \\
N_{Target}(E_i) &= number\ of\ event\ observations\ for\ event\ E_i \\
N_{FA}(S, E_j) &= number\ of\ false\ alarms\ for\ event\ E_j \\
P_{Miss}(S, E_i) &= \frac{N_{Miss}(S, E_i)}{N_{Targ}(E_i)} \\
R_{FA}(S, E_j) &= \frac{N_{FA}(S, E_j)}{T_{Source}} \\
Cost_{Miss} &= TBD\colon a\ constant\ defining\ the\ cost\ of\ a\ missed\ detections. \\
Cost_{FA} &= TBD\colon a\ constant\ defining\ the\ cost\ of\ a\ false\ alarm. \\
R_{Targ} &= TBD\colon a\ constant\ defining\ the\ a\ priori\ rate\ of\ event\ observations. \\
\mathrm{Beta} &= \frac{\mathrm{Cost_{FA}}}{\mathrm{Cost_{Miss}} * \mathrm{R_{Target}}}
\end{aligned}$$

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

Side information:

Allowable side information (i.e., "contextual" information) will include resources posted on the TRECVid Event Detection website, plus any annotation of the devset. Participants may share devset annotations. However, no annotation of the evalset should occur prior to the evaluation submission deadline.

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 values of the event. Events used in the evaluation will be annotated over the full corpus.

Events will be considered independent events for the evaluation. Therefore, systems will treat each event independently, for example, using a separately trained model for each event.

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. Systems should output detection results for all events in the required set. For the optional set, systems can output detection results for some or all of the events.

Submission of results:
Submissions will be made via ftp according to the instructions in Appendix B. In addition to the system output, a system description is also required for each condition. This description must include a description of the hardware used to process the data, and a detailed description of the architecture and algorithms used in the system.

Schedule:

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

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

Appendix A: Derivation of Average Normalized Detection Cost

Average Normalized Detection Cost Rate (*ANDCR*) is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per unit time). The measure's unit is in terms of Cost per Unit Time and is derived as follows.

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 convert Detection Cost to a rate by dividing by the length of the source data. Typically, we make this conversions to percentages by dividing by the count of discrete units for which systems make decisions. In a streaming environment, there are no discrete units, therefore normalizing by unit time is a more appropriate normalization. Note also that the measure of Type I error, R_{FA}, is commonly used in surveillance-style applications.

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

$R_{Target}(E)$ is the rate of occurrences for 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, R_{Target}, that in combination with the $Cost_{Miss}$ and $Cost_{FA}$ reflects the surrogate application. While the events for the evaluation have not been selected yet, we expect the them to have similar numbers of occurrences: neither too frequent or too rare. Therefore, the single prior is warranted. The modified formula becomes:

$$DetectionCostRate(S,E) = Cost_{Miss} \cdot P_{Miss}(\mathrm{S,E}) \cdot R_{Target} + Cost_{FA} \cdot R_{FA}(\mathrm{S,E})$$

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 (therefore $P_{Miss} = 1$ and $P_{FA} = 0$). The resulting formula is the Normalized Detection Cost Rate of a system (*NDCR*).

$$
\begin{aligned}
NormDectectionCostRate(S,E) &= \frac{DetectionCostRate(S,E)}{Cost_{Miss} \cdot R_{Target}} \\
&= \frac{\left(Cost_{Miss} \cdot P_{Miss}(S,E) \cdot R_{Target} + Cost_{FA} \cdot R_{FA}(S,E)\right)}{Cost_{Miss} \cdot R_{Target}} \\
&= P_{Miss}(S,E) + \frac{Cost_{FA} \cdot R_{FA}(S,E)}{Cost_{Miss} \cdot R_{Target}} \\
&= \mathrm{P}_{\mathrm{Miss}}(S,E) + Beta \cdot R_{FA}(S,E)
\end{aligned}
$$

Where:

$$\mathrm{Beta} = \frac{\mathrm{Cost}_{\mathrm{FA}}}{\mathrm{Cost}_{\mathrm{Miss}} \cdot \mathrm{R}_{\mathrm{Target}}}$$

Beta is separated out because it is composed of constant values that define the parameters of the surrogate application.

To calculate performance over an ensemble of events, we define Average Normalized Detection Cost Rate (*ANDCR_{Sys}*) by averaging the Missed Detection probabilities for all events with at least one true event occurrence (*N_{EventsNZ}*) and averaging the False Alarm Rates overall events. By separating the two averages, the measure can incorporate events with no true occurrences while remaining defined.

$$ANDCR(S) = \frac{\sum_{i}^{N_{EventsNZ}} P_{Miss}(S,E_i)}{N_{EventsNZ}} + Beta * \frac{\sum_{j}^{N_{Events}} R_{FA}(S,E_j)}{N_{Events}}$$

Where:

$$
\begin{aligned}
\mathrm{N}_{\mathrm{EventsNZ}} &= \text{number of event with } N_{Targ}(E_i) > 0 \\
N_{Events} &= n\text{umber of events} \\
P_{Miss}(S,E_i) &= \frac{N_{Miss}(S,E_i)}{N_{Targ}(E_i)} \\
R_{FA}(S,E_j) &= \frac{N_{FA}(S,E_j)}{T_{Source}} \\
\mathrm{Beta} &= \frac{\mathrm{Cost}_{\mathrm{FA}}}{\mathrm{Cost}_{\mathrm{Miss}} * \mathrm{R}_{\mathrm{Target}}} \text{ ;; defined a priori}
\end{aligned}
$$

Appendix B: Submission Instructions

This appendix will be filled out at a later date.