2008 TRECVid Event Detection Evaluation Plan

1. Overview

This document presents the evaluation plan for 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.

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
- Evaluation Infrastructure
- Event definitions
- Schedule

2. 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 Lingusistic Data Consortium will provide event annotations for the entire corpus according to the milestones listed in the schedule.

The 100-hour 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 such as 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 is 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.

Allowable side information (i.e., "contextual" information) will include resources posted on the TRECVid Event Detection website as well as any annotations constructed by developers based on the devset. Participants may share devset annotations. No annotation of the evalset is permitted prior to the evaluation submission deadline.

3. **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 an additional 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.

4. Evaluation Infrastructure

Systems will be evaluated on how well they can detect event occurrences in the evaluation corpus. The determination of correct detection will be based solely on the temporal similarity between the annotated reference event observations and the system-detected event observations.

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 Detection Cost Rate (DCR) model, which is a linear combination of the two errors. The DCR model distills the needs of a hypothetical application into a set of predefined constant parameters that include the event priors and weights for each error type. While the chosen constants have been motivated by

discussions with the research and user communities, the single operation point characterized by the DCR model is a small window into the performance of an event detection system. In addition to DCR measures, Detection Error Tradeoff (DET) curves will be produced to graphically depict the tradeoff of the two error types over a wide range of operational points. The DCR model and the DET curve are related: the DCR model defines an optimal point along the DET curve.

The rest of this section defines the system output, followed by the three steps of the evaluation process: temporal alignment, Decision Error Tradeoff (DET) curve production, and DCR computations.

4.1. System Outputs

Systems will record observations of events in a VIPER-formatted XML file as described in the "TRECVid 2008 Event Detection: ViPER XML Representation of Events" document. Each event observation generated by a system will include the following items:

- Start frame: The frame number indicating the beginning of the observation (the first frame in the video source file is frame #1.)
- End frame: The frame number indicating the last frame of the observation.
- Decision score: A numeric score indicating how likely the event observation exists with more positive values indicating more likely observations.
- Actual Decision: A Boolean value 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. In order to construct a fuller DET curve, a system must over-generate putative observations far beyond the optimal point for the system's best DCR value. The actual decisions provide the mechanism for the system to indicate which putative observations to include in the DCR calculation: i.e., the putative decisions with a *true* actual decision.

Systems must ensure their decision scores have the following two characteristics: first, the values must 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.

Since the decision scores are consistent across events, the system must use a single threshold for differentiating *true* and *false* actual decisions.

4.2. Event Alignment

Event observations can occur at any time and for any duration. Therefore, In order to compare the output of a system to the reference annotations, an optimal one-to-one mapping is needed between the system and reference observations. The mapping is

required because there is no pre-defined segmentation in the streaming video. The alignment will be performed using the Hungarian Solution to the Bipartite Graph [1] matching problem by modeling event observations as nodes in the bipartite graph. The system observations are represented as one set of nodes, and the reference observations are represented as a second set of nodes. The kernel formulas below assume the mapping is performed for a single event (E_j) at a time.

$$Kernel(O_{s,i}) = -1 \;;; The\ kernel\ score\ for\ false\ alarms$$
$$Kernel(O_{r,j}) = 0 \;;;\; The\ kernel\ score\ for\ missed\ detections$$
$$Kernel(O_{s,i}, O_{r,j}) = \begin{cases} \emptyset \quad if\ Mid(O_{s,i},) > End(O_{r,i}) - \Delta_T \\ \emptyset \quad if\ Mid(O_{s,i},) < Beg(O_{r,i}) + \Delta_T \\ 1 + \mathrm{E}_T * TimeCongru(O_{s,i}, O_{r,j}) + \mathrm{E}_{DS} * DecScoreCongru(O_{s,i}) \end{cases}$$
$$TimeCongru(O_{s,i}, O_{r,j}) = \frac{Min\left(End(O_{r,j}), End(O_{s,i})\right) - Max\left(Beg(O_{r,j}), Beg(O_{s,i})\right)}{Max\left(\frac{1}{25}, Dur(O_{r,j})\right)}$$
$$DecScoreCondgru(O_{s,i}) = \frac{Dec(O_{s,i}) - MinDec(s)}{RangeDec(s)}$$

Where:

$O_{s,i} = The\ i^{th}\ observation\ of\ the\ event\ for\ the\ System\ s$
$O_{r,j} = The\ j^{th}\ reference\ observation\ of\ the\ event$
$Beg(\ \ ) = The\ beginning\ of\ the\ observation's\ time\ span$
$Mid(\ \ ) = The\ midpoint\ of\ the\ observation's\ time\ span$
$End(\ \ ) = The\ end\ of\ the\ observation's\ time\ span$
$Dec(O_{s,i}) = The\ decision\ score\ of\ observation\ O_{s,i}$
$MinDec(s) = The\ minimum\ decision\ score\ for\ system\ s$
$RangeDec(s) = The\ range\ of\ decision\ scores\ for\ system\ s$
$\Delta_T = \boldsymbol{TBD} \;;; a\ constant\ differentiating\ the\ mappable\ and\ un-mappable\ observations$
$\mathrm{E}_T = 1e-8 \;;; a\ constant\ to\ weight\ time\ congruence$
$\mathrm{E}_D = 1e-6 \;;; a\ constant\ to\ weight\ decision\ score\ congruences$

The kernel function for observation comparisons, $Kernel(O_{s,i}, O_{r,j})$, has two levels. The first level, indicated by the $\emptyset$ values, differentiates potentially mappable observation pairs from non-mappable observation pairs. The second level takes into account the temporal congruence of the system and reference event observations and the observation's detection score in relation to the system's range of detection score. The decision scores are taken into account to facilitate the DET curve generation. By giving more weight to higher confidence score observations, realignment can be avoided during DET curve production.

4.3. Detection Error Tradeoff Curves

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

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

$$R_{FA}(S, E_i, \Theta) = \frac{N_{FA}(S, E_i, \Theta)}{T_{Source}}$$

Where

$N_{Miss}(S, E_i, \Theta)$
$= number\ of\ missed\ detections\ for\ system\ S, event\ E_i\ at\ decision\ score\ \Theta$

$N_{Target}(E_i) = number\ of\ event\ observations\ for\ event\ E_i$

$N_{FA}(S, E_i, \Theta) = number\ of\ false\ alarms\ for\ event\ E_i at\ decision\ score\ \Theta$

$T_{Source} = The\ total\ duration\ of\ the\ video\ segments\ in\ hours$

The formulas to compute averages over all events are defined as:

$$P_{Miss}(S, \Theta) = \frac{\sum_i^{N_{EventsNZ}} P_{Miss}(S, E_i, \Theta)}{N_{EventsNZ}}$$

$$R_{FA}(S, \Theta) = \frac{\sum_j^{N_{Events}} R_{FA}(S, E_j, \Theta)}{N_{Events}}$$

$\mathrm{N_{EventsNZ}} = \text{number of event with } N_{Targ}(E_i) > 0$

$N_{Events} = n\text{umber of events}$

PMiss(S, Θ) is not defined for all events because $N_{Targ}(E_i)$ may be 0. Therefore *PMiss(S, Θ)* is calculated over the set of events with true occurrences. This enables the evaluation of a system on events that do not exist in the test corpus.

4.4. DCR Computations

The evaluation will use the Average Normalized Detection Cost Rate (*ANDCR*) measure 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 and the final formula is summarized below.

$$ANDCR(S, \Theta) = P_{Miss}(S, \Theta) + Beta * R_{FA}(S, \Theta)$$

Where:

$$\mathrm{Beta} = \frac{\mathrm{Cost_{FA}}}{\mathrm{Cost_{Miss}} * \mathrm{R_{Target}}}$$

$Cost_{Miss} = \boldsymbol{TBD}: a\ constant\ defining\ the\ cost\ of\ a\ missed\ detections.$

$Cost_{FA} = \boldsymbol{TBD}: a\ constant\ defining\ the\ cost\ of\ a\ false\ alarm.$

$$R_{Targ} = \boldsymbol{TBD}: a\ constant\ defining\ the\ a\ priori\ rate\ of\ event\ observations.$$

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. $P_{Miss}=1$ and $P_{FA}=0$.

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

4.4.1. Actual ANDCR

The Actual ANDCR is the primary evaluation metric. It is computed by restricting the putative observations to those with *true* actual decisions.

4.4.2. Minimum ANDCR

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

5. Events

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 temporal coordinates of the event.

Events will be considered to be independent for the evaluation. Therefore, systems may build separately trained models 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. 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.

6. 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.

7. 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.--Mar.: Event detection planning & telecons
Feb.: Call for participation in TRECVid
Mar. 10: 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

8. 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 Average Normalized Detection Cost Rate

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 ANDCR formula includes three predefined parameters that represent both the richness of events in the source data and the relative detriment of particular error types for a hypothetical application. This surrogate application provides a view of system performance that is consistent across systems[1].

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.

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

$R_{Target}(E)$ 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 characteristics of the surrogate application. While the events for the evaluation have not been selected yet, we expect the them to have similar numbers of occurrences[2]: neither too frequent or too rare. Therefore, the single prior is warranted. The modified formula becomes:

[1] Provided developers tune their systems using the defined parameters.
[2] For instance within 2-3 orders of magnitude.

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

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*).

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

Where:

$$\mathrm{Beta} = \frac{\mathrm{Cost_{FA}}}{\mathrm{Cost_{Miss}} \cdot \mathrm{R_{Target}}}$$

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(S)*) 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:

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

The measure's unit is in terms of Cost per Unit Time and is derived as follows.

Appendix B: Submission Instructions

This appendix will be filled out at a later date.