TrecVid 2008 Event Annotation Guidelines

Version 1.6 April 28, 2008

1 Overview

TrecVid is a video event detection project. The TrecVid video data consists of surveillance footage from 5 camera views in the same airport over a period of 10 days. There are 2 hours per camera per day, making for a total corpus of 100 hours.

We will be annotating 10 required and 7 optional events over the 100 hours of data. The 100 hours is divided into 50 hours of Development Testing (Dev) and Evaluation (Eval) data.

The optional events will be annotated after the required events are at least partially completed. The Eval data will be annotated after the Dev data is completed.

We will divide up the 17 events into 4 sets, 2 required and 2 optional. Annotators will be assigned 1 event set and 1 video clip. Each video clip is 5 min 10 seconds in length.

The 10 seconds constitutes 5 seconds of overlap with both the previous and following video clip. After annotation is completed, we will review the overlapping portions with annotations to merge events which span multiple clips.

2 Event Annotation

The annotation task will be to tag the duration of the event. The video event annotation tool we will use, called ViPER, allows the user to watch the video and manually manipulate a line representing the duration of the event. This annotation will be saved as <startframe> and <endframe> for the event.

2.1 General Annotation Rules

The following rules apply to all events. Annotators must refer to these rules when deciding the taggability and extent of an event.

2.1.1 Reasonable Interpretation Rule

If according to a reasonable interpretation of the video, the event must have occurred, then it is a taggable event.

2.1.2 Occlusion Rules

Rule 1: If the annotator decides the event must have occurred but occlusion blocks the start time, the start time is then the start of the occlusion.

Rule 2: If the annotator decides the event must have occurred but occlusion blocks end time, the end time is then the end of the occlusion.

Rule 3: The occlusion can be the frame boundary (entering or exiting the frame), but a portion of the event must be determined to have occurred within the frame boundary according to the Reasonable Interpretation Rule.

2.2 Required Events

E05: PersonRuns

Description: Someone runs.

Start Time: The earliest time the subject is visibly running.

End Time: The latest time the subject is visibly running.

E06: CellToEar

Description: Someone puts a cell phone to his/her ear.

Start Time: When the subject starts to move the phone to his/her head.

End Time: When the phone reaches the head.

Comment: This event is intended to detect the movement of a phone to a subject's head. This event is not targeted at detecting a cell phone call that could potentially have several CellToEar sub-events. For instance, if a subject is already on a cell call and drops his/her arm momentarily to lift a bag, but then raises the arm again to continue the call, that is not a new CellToEar event.

This event is also not intended to detect the case when a subject is already on a cell call when s/he enter the frame (falling under the Occlusion Rule 3).

E08: ObjectPut

Description: Someone drops or puts down an object.

Start Time: The latest time the subject is known to have the object.

End Time: The earliest time the subject is known not to have the object

Comment: Humans are not considered objects. For instance, someone putting a baby into a stroller is not an ObjectPut event.

E14: PeopleMeet

Description: One or more people walk up to one or more other people, stop, and some communication occurs.

Start Time: The first communication between any member of one group to a member of the other group. The first communication between any member of the two groups.

End Time: The earliest time when the two groups are nearest to each other after the communication has occurred.

Comment: This is meant to cover a meeting event. If people meet, communicate for some time, and then get nearer to each other, the end time is the initial point when they are closest after communication has occurred.

E15: PeopleSplitUp

Description: From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame.

Start Time: The latest time when a group of people are nearest to each other.

End Time: The earliest time when at least one group member leaves the frame.

Comment: PeopleSplitUp and PeopleMeet should be considered independently. If a group is standing together communicating, then one or more people separate themselves, have a PeopleMeet event with a different group, then leave that group, and then leave the frame, there will be two PeopleSplitUp events. One PeopleSplitUp will have a longer extent, and both of their end times will be the person/people exiting the frame.

E16: Embrace

Description: Someone puts one or both arms at least part way around another person.

Start Time: The latest time when subjects do not have physical contact prior to the embrace.

End Time: The earliest time when subjects do not have physical contact after an embrace.

Comment: This event is not intended to detect the case when subjects are already embracing when they enter the frame, and do not lose physical contact while in the frame (falling under the Occlusion Rule 3).

E18: Pointing

Description: Someone points

Start Time: The earliest time when the person has placed their finger/hand in the pointing position.

End Time: The earliest time when the person has changed the position of their finger/hand/arm to no longer be in a pointing position.

Comment: This does not begin when they raise their arm to point. There may be clear pointing events that do not involve raising one’s arm. For instance, a person could show another person an object, and point to the object while holding it. These pointing events are taggable, because it is only the pointing position itself that constitutes the event.

Pointing as part of a gesture in conversation is still pointing, so it is a taggable event.

E19: ElevatorNoEntry

Description: Elevator doors open with a person waiting in front of them, but the person does not get in before the doors close.

Start Time: The earliest time when the elevator doors are opening with person waiting in front of them.

End Time: The earliest time that the doors of the elevator are fully closed.

E20: OpposingFlow

Description: Someone moves through a door opposite to the normal flow of traffic [applies only where normal flow of traffic is defined. For TRECVid ‘08, this applies only to the doors in Camera 1]

Start Time: The earliest time when the person has begun to move through the door. If the person does not appear before they are already passing through the door, then Start Time is when the person appears.

End Time: When the person has fully passed through the doorway. Fully passed means that not only their body, but any objects they might be carrying, e.g., rolling luggage behind them, must have passed beyond the frame of the doorway.

Comment: Normal flow of traffic is currently defined for 1 set of doors only, the doors in Camera 1.

E21: TakePicture

Description: Someone takes a picture.

Start Time: The earliest time when a person holds a camera in a fixed position prior to activating it.

End Time: The earliest time when the camera moves away from a fixed position following the photograph.

Comment: This event does not distinguish between types of cameras, which may include cell phone cameras.

2.3 Optional Events

E01: DoorOpenClose

Description: Door opens and then closes

Start Time: A closed door begins to open.

End Time: An open door is closed.

E04: UseATM

Description: Someone inserts a card in the ATM and the event lasts until the person steps away from the ATM.

Start Time: The earliest time when the person has begun to insert the card into the ATM.

End Time: The earliest time when the person is moving away from the ATM.

Comment: The card isn't the deciding factor. The reasonable interpretation rule will cover this in cases where you only see a gesture moving towards the ATM.

E09: ObjectGet

Description: Someone picks up an object.

Start Time: The latest time the subject is known not to have the object.

End Time: The earliest time the subject is known to have the object.

Comment: This event does not apply to food, beverages, napkins, or eating utensils. Humans are not considered objects (see ObjectPut).

E10 VestAppears

Description: Someone in yellow/green safety vest appears.

Start Time: The earliest time the subject is visible.

End Time: The latest time the subject is visible.

E11: SitDown

Description: Someone sits down.

Start Time: The earliest time when the person has begun downward movement towards the object/location they will sit on.

End Time: The earliest time when the person is in a seated position on the object/location.

Comment: Seated means their body is in a resting position on the seating. They may settle down or switch position or be sitting on the edge of a chair, but they will still be considered already seated.

E12: StandUp

Description: Someone stands up.

Start Time: The earliest time when the person has begun upward movement off of the object or location at which they were seated.

End Time: The earliest time when the person is in a fully upright position.

E17: ObjectTransfer

Description: Someone transfers an object to another person when the object is always under the control of one of the people.

Start Time: The latest time when the transferor has possession of the object.

End Time: The earliest time when the object is in the possession of only the transferee.

Comment: This event is not meant to detect an event where one person puts down an object and another person picks it up. This will be covered by Events E08 and E09 (ObjectGet and ObjectPut). The object must be possession by someone at all points.

Humans are not considered objects (see ObjectPut).