TRECVid 2009 Event Annotation Guidelines

Version 1.0 June 2, 2009

1. Overview

TRECVid, the TREC Video Retrieval evaluation, is a laboratory-style evaluation that aims to model real world situations or significant component tasks involved in such situations. Its main goal is to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation. In addition to high-level feature extraction, search, and content-based copy detection, TRECVid 2009 includes the surveillance event detection task. The TRECVid 2009 surveillance video data consists of ~144 hours of footage from 5 camera views in London Gatwick airport (~100 hours training set, ~44 hours test set). We will be finding and tagging instances of 10 events in the video data.

We will divide up the 10 events into 4 sets, 3 sets of 3 events, and 1 set of 1 event (ElevatorNoEntry). The ElevatorNoEntry set will be completed by a single senior or lead annotator. All other events will be assigned in sets of 3 for each video clip. Video clips are about 5 minutes long.

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.

End Time Rule:

When the end time of an event is the time at which a person leaves the frame, tag the earliest frame when not only their body, but any objects they might be carrying, e.g., rolling luggage behind them, have passed out of the frame. If someone rolling luggage leaves the frame, and the luggage follows them, then tag the End time as the luggage leaving the frame. However, if the luggage remains partially in the frame, but the person is not visible, tag the End time as the person leaving the frame, as they may have left their luggage.

2.2. Required Event Descriptions

For most of these events, common-sense understanding of the event title should be enough to decide whether the Reasonable Interpretation rule applies.

A useful device to help access that intuition is to imagine telling a friend about what you saw, and then think about the words you used in your description. If you are trying to decide whether an event is PersonRuns, and you described it as “fast-walking”, then it is not running and it should not be tagged.

The Start Time and End Time descriptions are meant to provide consistency in the start and end time of annotation. They are not intended as rules to determine whether the event occurred. Common-sense judgment should trump these guidelines for taggability.

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.

Comment: Skipping, gliding, skateboarding are not PersonRuns events. As per the rules above, only running is taggable as PersonRuns.

E06: CellToEar

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

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

End Time: When the phone reaches the head.

Comments: This event is intended to detect the movement of a cell phone to a subject's head for the beginning of a phone call. If a person is already on the phone and drops his/her arm momentarily (e.g. 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 enters the frame, because both the Start and End Time of the Cell to Ear event occur outside of the frame (refer to Occlusion Rule 3).

A "Cell to face" event, if someone brings the phone to their mouth rather than their ear, is a taggable event, it is the same conceptual event of beginning a cell phone call by raising the phone to the head.

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

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

Rolling luggage is carried on the ground, so it can't be "put down" in the same sense as other items, but we would consider releasing the handle with the purpose of stopping the luggage and leaving it there, to be a taggable ObjectPut event.

Someone putting something somewhere on their person is not an ObjectPut. They aren't "putting down" the object. If they were wearing or holding the purse/backpack on their person, it is not taggable. If they put the object in a purse/backpack next to them or on the ground or something, that could be taggable.

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.

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

Comments: 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.

Please do not annotate PeopleMeet with the desk clerk(s) in Camera 3, where the black masking bar covers almost all of the desk, except for their hands. There is not enough visual evidence for it to fall under the Reasonable Interpretation rule.

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 split-off group member leaves the frame.

Comments: PeopleSplitUp and PeopleMeet should be considered independently. If a group is standing together communicating, one or more people separate themselves and join with a different group, then leave that group, and then leave the frame, there will be two PeopleSplitUp events, one for each group that was split from, both ending with the same leaving 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 (of any kind) after an embrace.

Comments: Piggyback rides, cheek kisses, putting hands on someone's shoulders, and picking up a child are NOT Embrace events. As per the rules above, only embraces are taggable as Embrace events. Embracing most commonly corresponds to “hugging”, although one may put their arm around someone’s shoulders or waist, which is an embrace but not a hug.

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 (refer to Occlusion Rule 3).

If two people have their arms around each other when they walk in the frame, then lose physical contact in the frame, that is an Embrace event, because the End Time of the embrace occurs in the frame.

E18: Pointing

Description: Someone points

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

Comments: This does not necessarily 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.

If a person engages in a long arm/hand Pointing gesture, with smaller hand/finger Pointing gestures in the middle, only 1 event, the extent of the long gesture, should be tagged.

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.

Comment: There are two sets of elevator doors. If an event fits the description, it is taggable regardless of which set of doors opens. This means the event is taggable both if the person chose to not enter the elevator, or if they simply didn't notice the other set of doors open.

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 '09 and 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.

Comments: If two or more people are simultaneously walking through the doorway, it should be tagged as 1 event. If they are not exactly simultaneous, each should be tagged as a separate event.

If a person stands in the doorway facing the normal flow of traffic but does not go through it, it is not a taggable event.

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.