Introduction for VIRAT Video Dataset Release 2.0
(Doc version 1.0, 2011 Sep 30th)

1. Introduction

This document describes the following contents for VIRAT Data Release 2.0: **data, evaluation criteria, annotation standards,** and **activity types.**

For data download, please refer to the following document:
docs/VIRAT_Video_Dataset_Download_Instruction_Release2.pdf

2. Dataset and Evaluation Criteria

2.1. Scenes

Release 2.0 includes videos recorded from total 11 scenes, captured by stationary HD cameras (1080p or 720p). There may be very slight jitter in videos due to wind. Videos are encoded in H.264.

The following 11 scenes are included in this release:
VIRAT_S_0000
VIRAT_S_0001
VIRAT_S_0002
VIRAT_S_0100
VIRAT_S_0101
VIRAT_S_0102
VIRAT_S_0400
VIRAT_S_0401
VIRAT_S_0500
VIRAT_S_0502
VIRAT_S_0503

There are multiple video clips from each scene, and each clip will contain zero or more instances of activities from 12 categories (see Section 3.2 for complete list).

2.2. Filename formats

All the filenames are formatted as follows:
VIRAT_S_XXYYZZ_KK_SSSSSS_TTTTTT.mp4

Above, each symbols after the string 'VIRAT_S' are numerics as follows:

```
XX: collection group ID
YY: scene ID
ZZ: sequence ID
KK: segment ID (within sequence)
SSSSSS: starting seconds in %06d format. E.g., 1 min 2 sec is 000062.
TTTTTT: ending seconds in %06d format.
```

Intuitively, participants can identify videos from a same scene by comparing the first four digits XXYY. All the rest of the digits encode the time of the day each video clip is captured, and may not be useful for this competition.

3. Annotation Standards

There are a total 12 different types of events annotated. Some video clips are fully annotated with 12 event types (event type 01-12), while other clips are annotated with 6 event types (event type 01-06) only. The list of clips with each annotation mode can be found in:
`docs/README_annotations_evaluations.xls`

3.1. Object Annotations

Note that, due to the nature of Internet Mechanical Turk annotations, annotations may not satisfy the guidelines outlined below.

Objects mean 'people', 'vehicle', and arbitrary 'objects' such as bags being loaded into vehicles. Only the visible part of objects are labeled, and are minimally (mostly not) extrapolated beyond occlusion. For example, if the upper body of a person is the only visible part, then, only the upper body should be labeled as 'person'. Moving objects are supposed to be represented as a single object or a single track. Note that, due to limitations of MT-based annotations, sometimes, tracks fragment into multiple segments.

Every annotated object has duration information which consists of starting frame number and the duration, which equals (ending frame number-starting frame number + 1).

Bounding box around the objects should be 'whole' and 'tight'. By 'tight', we mean that bounding boxes should be as tight as possible and should not extrapolate beyond the objects being labeled. For example, minimal background should be part of bounding boxes around a person or a vehicle. On the other hand, 'whole' means that all related parts are captured in the bounding boxes (e.g. all the visible limbs of people should be in the bounding box, not just the person's torso).

Static objects, such as parked vehicles which are not involved in any activities or locations which people interact with (such as parking spots), are annotated as well. The annotations of these activity-free stationary objects throughout video clip is optional, and do not always exist. Moving objects which are not involved in the six or twelve types of activities considered have optional bounding boxes and may not exist (although most cases, they do). For moving objects that are involved in considered activities, they always exist.

A vehicle is defined as a wheeled or tracked motorized device used to transport cargo (either human or nonhuman). For vehicles, there are three sub-classes: **car, bike, and vehicle**. 'Car' includes any passenger vehicle such as sedan/truck/van etc. 'Bike' includes any bi-wheel vehicles such as bicycle and motor-bikes. 'Vehicle' includes other vehicles, not belonging to car or bike, such as construction vehicles or lawn-mowers. 'Vehicles' may still indicate 'car' or 'bike', but, it is supposed to be less-specific.

3.2. Event (Activity) Annotations

Events are annotated and represented as the set of objects being involved and the temporal interval of interest. For example, a label of a 'person entering a vehicle' should consist of the following information: (1) a reference to the bounding box of a person, (2) a reference to the bounding box of a vehicle, and (3) the time interval for the event. In some cases, the reference for some small objects may be missing for some frames or entirely due to annotation difficulty, e.g., objects are too small.

There are total 12 different types of event in Release 2.0. The precise definitions of each are described below. For event sentences (classes) enlisted below, the underlined words correspond to the objects that needs to be annotated with bounding boxes during the duration of events. Bounding boxes for as many frames as possible during the event duration should be marked. If some objects are invisible, bounding boxes for those frames are allowed to be missed.

The number of events per category per clips and useful statistics can be found from:
`docs/README_annotations_evaluations.xls`

3.2.1. Person loading an Object to a Vehicle

Description: An object moving from a person to a vehicle. The act of 'carrying' should not be included in this event.

Annotation: 'Person', 'Object' (optional), and 'Vehicle' should be annotated.

Start: The event begins immediately when the cargo to be loaded is “extended” toward the vehicle (i.e., before one's posture changes from one of 'carrying', to one of 'loading.').

End: The event ends after the cargo is placed in the vehicle and person-cargo contact is lost. In the event of an occlusion, it ends when the loss of contact is visible.

3.2.2. Person Unloading an Object from a Vehicle

Description: An object moving from a vehicle to a person.

Annotation: 'Person', 'Object' (optional), and 'Vehicle' should be annotated.

Start: The event begins immediately when the cargo begins to move. If the start of the event is occluded, it begins when the cargo movement is first visible.

End: The event ends after the cargo is released. If a person, while holding the cargo, begins to walk away from the vehicle, the event ends (at which time the person is 'carrying'). The event also ends if the vehicle drives away while the person is still in contact with the cargo; after the vehicle has been in motion for more than 2 seconds, the person is 'carrying'.

3.2.3. Person Opening a Vehicle Trunk

Description: A person opening a trunk. A trunk is defined as a container specifically designed to store nonhuman cargo on a vehicle. A trunk need not have a lid (i.e., the back of a pickup truck is a trunk), and it need not open from above (i.e., the back of a van, which opens via double doors, is also a trunk).

Annotation: 'Person', and 'Vehicle' should be annotated with bounding boxes for as many frames as possible during the event duration. The bbox annotation of 'Trunk' is optional.

Start: The event begins when the trunk starts to move.

End: The event ends after the trunk has stopped moving.

3.2.4. Person Closing a Vehicle Trunk

Description: A person closing a trunk.

Annotation: 'Person', and 'Vehicle' should be annotated with bounding boxes for as many frames as possible during the event duration. The bbox annotation of 'Trunk' is optional.

Start: The event begins when the trunk starts to move.

End: The event ends after the trunk has stopped moving.

3.2.5. Person getting into a Vehicle

Description: A person getting into, or mounting (e.g., a motorcycle), a vehicle.

Annotation: 'Person', and 'Vehicle' should be annotated.

Start: The event begins when the vehicle's door moves, or, if there is no door, 2 s before ½ of the person's body is inside the vehicle.

End: The event ends when the person is in the vehicle. If the vehicle has a door, the event ends after the door is shut. If not, it ends when the person is in the seated position, or has been inside the vehicle for 2 seconds (whichever comes first).

3.2.6. Person getting out of a Vehicle

Description: A person getting out of, or dismounting, a vehicle.

Annotation: 'Person', and 'Vehicle' should be annotated.

Start: The event begins when the vehicle's door moves. If the vehicle does not have a door, it begins 2 s before ½ of the person's body is outside the vehicle.

End: The event ends when standing, walking, or running begins.

3.2.7. Person gesturing

Description: A person gesturing. Gesturing is defined as a movement, usually of the body or limbs, which expresses or emphasizes an idea, sentiment, or attitude. Examples of gesturing include pointing, waving, and sign language.

Annotation: 'Person' should be annotated.

Start: The event begins when the gesture is evident. For example, when waving, the gesture when the waver begins to raise their arm into the “waving position.”

End: The event ends when the motion ends

3.2.8. Person digging (Note: not existing in Release 2.0)

Description: A person digging. Digging may or may not involve the use of a tool (i.e., digging with one's hands is still considered 'digging'; hands are the tool).

Annotation: 'Person' should be annotated.

Start: The event begins when the tool makes contact with the ground.

End: The event ends 5 s after the tool has been removed from the ground, or immediately if the digging tool is dropped.

3.2.9. Person Carrying an Object

Description: A person carrying an object. The object may be carried in either hand, with both hands, or on one's back. Object annotation by bboxes are optional and subject to the difficulty.

Annotation: 'Person', and 'Object' (optional) are annotated.

Start: The event begins when the person who will carry the object, makes contact with the object. If someone is carrying an object that is initially occluded, the event begins when the object is visible.

End: The event ends when the person is no longer supporting the object against gravity, and contact with the object is broken. In the event of an occlusion, it ends when the loss of contact is visible.

3.2.10. Person running

Description: A person running for more than 2s.

Annotation: 'Person' should be annotated.

Start: When a person is visibly running.

End: The event will end 2 s after the person is no longer running. If transitioning to Standing, Walking or Sitting the event will end after after Standing, Walking or Sitting.

3.2.11. Person entering a facility

Description: A person entering a facility

Annotation: 'Person' should be annotated.

Start: The event begins 2 s before that person crosses the facility's threshold.

End: The event ends after the person has completely disappeared from view.

3.2.12. Person exiting a facility

Description: A person exiting a facility

Annotation: 'Person' should be annotated.

Start: The event begins as soon as the person is visible.

End: The event ends 2 seconds after the person is completely out of the facility.

3.3. Annotation formats

Please refer to:
`docs/README_format_release2.txt`

Examples of computed ground truth are shown below where both the person and vehicle bounding boxes are shown in different colors, and the event bboxes are marked by thick red bbox.

3.4. Sample Software

Sample Matlab scripts to draw event annotations on videos and save annotation images can be found in the 'software' folder of 'sample dataset'. Main file is '`test_draw_viratdata2.m`'.

The software may need Matlab versions equal or newer than 2010a. The purpose of the software is to provide more specific ideas about the annotation file formats and to demonstrate the quality of samples. There will be no individual support to modify the software for different systems and supported video formats. The software has been tested and runs successfully with Windows 7 and Matlab 2010b. For older versions of Matlabs, 'VideoReader' object in source code may be replaced to 'mmreader' to work.

4. Evaluation methodologies (recommended for shared comparisons)

There **are two evaluation modes for testing datasets**:

- Scene-independent learning and recognition
- Scene-adapted learning and recognition

For scene-independent learning and recognition, event detectors are trained on scenes which are not included in the test scenes. This means 11-fold cross-validation where each fold consists of clips from one scene each.

For scene-adapted learning and recognition, clips from the same scene may be used during the training process, and applied to the clips from the same scene (but, these test clips are not used during the training process). For shared folds, we provide 5-fold sets where the folds can be found from the enclosed file:
`docs/README_annotations_evaluations.xls`

Because some clips are annotated with all 12 events and others are annotated only with 6 event types, there are different folds for different event types for two sets of events, either for event 01-06, or event 07-12.

4.1. Important Definitions for shared comparisons (Recommended):

4.1.1. Event

Any observed human motion related to vehicles (a person exiting a vehicle, a person closing a vehicle trunk etc.) excluding riding a bicycle or motorcycle.

4.1.2. Detection

A detection T is a sequence of frames F, each of which is attributed with a framenumber TS and a location L within the geographic area within F. The location L is a bounding box with four attributes {x_lt, y_lt, w, h} where x_lt and y_lt are x and y coordinates of the left top of the bounding box (left top of the image is origin), and w and h are bounding box size in x and y coordinate direction.

4.1.3. Event Matching Criterion

An event detection is defined as a tuple of (label, track). Given a pair or events {detection A, ground truth B}, A matches B if:

a) (spatial match) for every bounding box pairs per frame, if the two intersection ratios are both above 10%, that particular frame detection of A is regarded as a match for ground truth frame detection of B. Two intersections ratios are computed by the number of intersected pixels divided by both bounding boxes from A & B. For example, one may have correct detections for frames 101, 104, 105, and etc.
b) (temporal match) Both-way temporal intersection ratios should be above 10% to be regarded as a temporal match. Temporal intersections are the number of frames which satisfies (a) spatial match criteria. Two temporal intersection ratios are computed by the size of temporal intersection divided by both durations of A & B.
c) (label match) the activity labels associated with A and B are the same.

4.2. Event-level metrics for Competition

Event metrics are defined over a set of frames of video clips. Any activity said to occur must take place in its entirety within the timespan of the frames and the spatial bounds.

The performance of an algorithm can be measured using the following metrics, which include: Precision (primary), Pd (primary), False Alarm Rate (optional), F-scores (optional), weighted aggregate F-score (optional).

The number of detections to be counted should follow the policies outlined in Sec. 4.3 for fair comparison and to avoid over-counting.

4.2.1. **Precision**: Precision is the ratio TP/D, where D is the total number of detections (correct and incorrect); and TP is the number of correct detections, identical to the definition in 4.2.2.

4.2.2. **Probability of Detection (Pd):** a Pd is the ratio TP/T for every category, where T is the number of ground-truth activities in archive, and TP is the number of correctly detected activities matched to a member of T according to the activity-matching criterion. Pd is identical to 'recall'.

4.2.3. **False Alarm Rate (FAR)**: a FAR per activity type is the ratio FP/NORM, where FP is the number of false positives whose detected activities do not match a member of T, and NORM is a normalizing factor based on the number of frames so that FP/NORM is in units of *activities per minute.*

4.2.4. **F-score**: F-score is computed as the harmonic mean of Pd and Precision. It captures summary capability of detectors based on Pd and Precision. The F-score will be computed as follows:

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{Pd} + \frac{1}{\text{Precision}}\right)}$$

4.2.5. **Weighted Aggregate F-score**: a weighted aggregate F-score can be additional used. This score will capture the overall performance of developed detectors across categories. Weights across all categories will sum to one, and set to be proportional to the number of samples.

$$\text{Weighted Aggregate F - score} = \frac{1}{w_i \times \sum_{i=1}^{n} \frac{1}{F_i}}$$

4.2.6. Official Scoring Software

Official scoring software for Release 2.0 format is planned to be distributed in the future. Note that the existing scoring software for Release 1.0 does not work with Release 2.0 data. Please check data webpage (viratdata.org) for updates.

Users will be able to compute the quality of their detections results. Correct detections, false alarms, and all the competition metrics will be computed automatically given user detection results formatted in the specified format, which will be described with the software distribution documentation.

4.3. Correct Detections and False Alarms

4.3.1. Correct Detections

4.3.1.1. An element of T may be matched by multiple elements of D; this counts as a single hit for T but eliminates the matching elements of D from being counted as false alarms. Examples are shown in Figure 1.

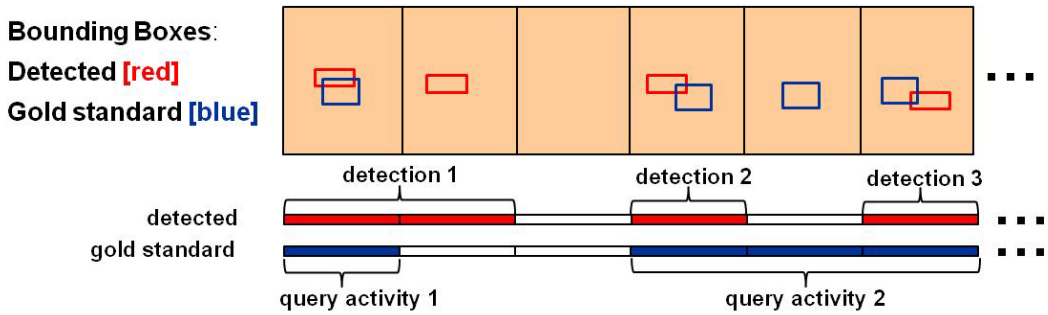


Figure 1. Examples of multiple correct detections (detection 2 & 3) on one query activity (activity 2), assuming that detections 2 & 3 have valid temporal overlap with the query activity 2.

4.3.1.2. An element of D may match multiple elements of T. In such cases, a detection can contribute towards multiple correct detections. Examples are shown in Figure 2.

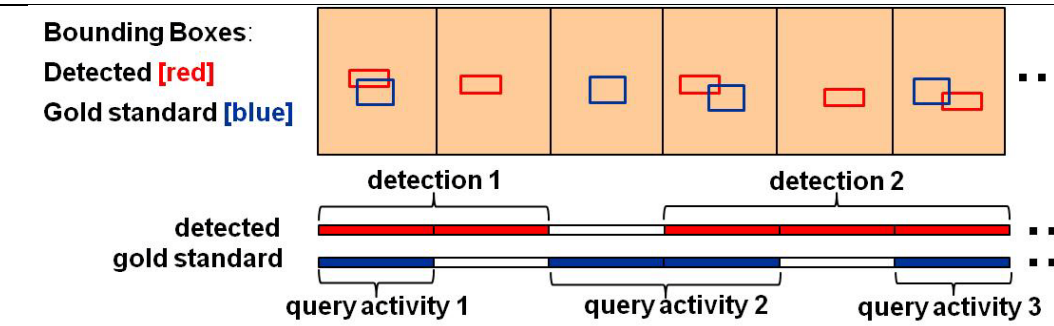


Figure 2. Examples of one detection (detection 2) counted as two correct detections against two ground truth activities (query activities 2 & 3), assuming that they satisfy spatio-temporal matching criteria.

4.3.1.3. Multiple detections and ground truths can both occur concurrently, and detections will be scored against all concurrent ground truths, following the policies in 4.3.1.1 and 4.3.1.2. Examples are shown in Figure 3.

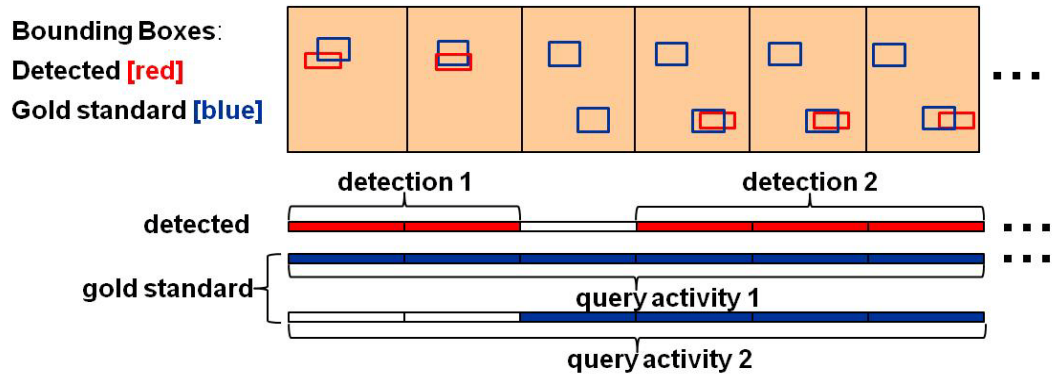
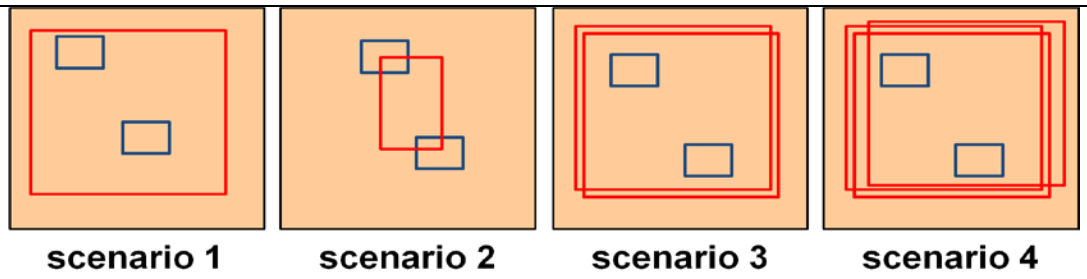


Figure 3. Where there are concurrent events, the bounding box provided with each detection must overlap the annotated bounding boxes. In the example above, both activities are detected, assuming they satisfy spatial and temporal match criteria.

When detection bounding boxes overlap with multiple annotated ground truth activities, the policies shown in Figure 4 are applied (assuming they all satisfy spatio-temporal match criteria):



Detected: Red
Ground Truth: Blue

Figure 4.

- Scenario 1 - single detection that encloses two annotated query events: both detections count towards P_D (provided they both satisfy spatio-temporal match criteria)
- Scenario 2 - single detection that overlaps two annotated query events: both detections count towards P_D (same provision)
- Scenario 3 – two detections enclose two annotated query events: both detections count towards P_D (same provision)
- Scenario 4 – three detections enclosing two annotated query events: two detections count towards P_D (same provision); no penalty in FAR for third detection.

4.3.2. False Alarms

Detections that do not match any existing ground truth are counted towards false alarms.

4.3.2.1. Examples of false alarms:

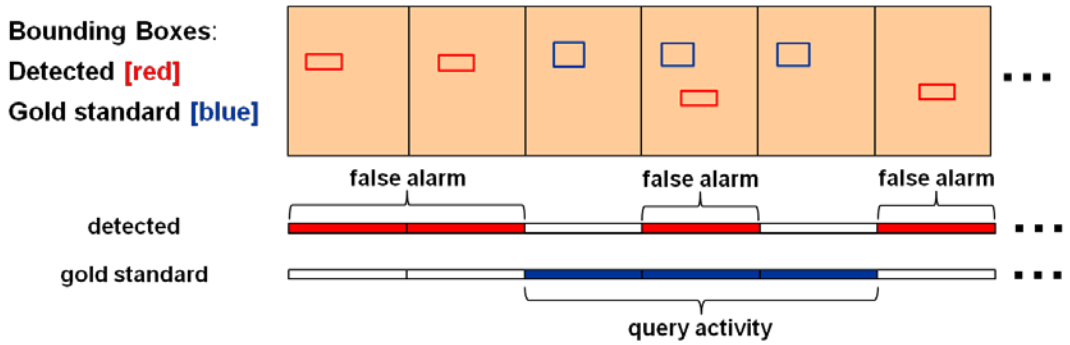


Figure 5. Examples of false alarms.

5. Disclosure

The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.