VIRAT Video Dataset Release 1.0 Evaluation

Introduction (doc version 1.2, 2010 Feb 10th)

Table of Contents

1. Introduction 2
2. Dataset and Evaluation Criteria 2
   - 2.1. Scenes, Videos, Evaluations 2
   - 2.2. Filename formats 2
   - 2.3. Annotation formats 3
   - 2.4. Event Ground Truth 4
   - 2.5. Sample Software 4
3. Evaluation Metrics 5
   - 3.1. Definitions: 5
     - 3.1.1. Activity 5
     - 3.1.2. Detection 5
     - 3.1.3. Activity Matching Criterion 5
   - 3.2. Activity-level metrics for Competition 5
     - 3.2.6. Official Scoring Software 6
4. Annotation Standards 9
   - 4.1. Object Annotations 9
   - 4.2. Event (Activity) Annotations 10
5. Contact Information 12

1. Introduction

This document describes the following contents for VIRAT Data Release 1.0: **data, evaluation criteria, annotation standards,** and **activity types.**

2. Dataset and Evaluation Criteria

2.1. Scenes, Videos, Evaluations

Release 1.0 includes videos recorded from total 6 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.

Each video clip will contain 1~20 instances of activities from 6 categories: (1) person loading an object to a vehicle, (2) person unloading an object from a vehicle, (3) person opening a vehicle trunk, (4) person closing a vehicle trunk, (5) person getting into a vehicle, and (6) person getting out of a vehicle.

For **training**, subset of videos from 3 scenes will be released. Training datasets will contain both videos and annotations. Approximate homographies of image to world mappings are provided as part of dataset to aid data users who need geometry information for tracking.

For **testing**, additional videos from 3 scenes in training datasets, and more videos from three additional scenes will be released. Testing dataset will contain only videos, and participants will submit recognition results in a specified format (TBA) by due date which will be approximately 2 weeks after the testing data release. Details of formatting will be announced soon.

There **are two evaluation modes for testing datasets**:
- For 3 overlapping scenes in test datasets w.r.t. training sets, participants can use any scene-specific knowledge during their testing.
- For 3 new scenes in test datasets, participants should run their algorithms as scene-independent recognition modules, i.e., participants should only rely on training dataset to run their algorithms.

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.

2.3. Annotation formats

For every video clip, there are two annotation files in whitespace-delimited: (1) object annotations, and (2) event annotations.

As an example, for a video file named 'VIRAT_S_000002.mp4', corresponding annotation files will be named as follows: `VIRAT_S_000002.viratdata.objects.txt, VIRAT_S_000002.viratdata.events.txt.` Note that, sample video file VIRAT_S_000002_smallsize.mp4/avi is included with lower resolution than the original data, just to reduce the size of the package. In the training dataset, original files with full HD quality will be included.

2.3.1 Object annotation format

Every line (row) in object annotation file indicates information about an annotated object at a specific video frame.

For total 8 columns, each column corresponds to the following information:

```
1) Object ID: each object has unique ID and IDs may not be consecutive, i.e, skip.
2) Duration of object: the total length of the object with Object ID
3) Frame number: frame number in video with zero-base.
4) X_lt: left-top x coordinate of bbox on image, with coordinate center at image
left top (in pixels).
5) Y_lt: left-top x coordinate of bbox on image
6) Width: width of bbox
7) Height: height of bbox
8) Object Type: type of object (Unknown=0, person=1, car=2, other vehicle=3,  other
object=4, bike=5)
```

2.3.2 Event annotation format

Each line (row) in event annotation file indicates information about duration of event and involved objects. Note that each event has a single fixed bounding box for the entire duration.

```
1) Event ID: each event is associated with an ID (separately counted from Object ID)
2) Event Type: unknown=0, loading=1, unloading=2, opening_trunk=3, closing_trunk=4,
getting_into_vehicle = 5, getting_out_of_vehicle = 6.
3) Event length: duration of event
4) Event start frame: base zero
5) Event end frame
6) X_lt: bbox X_lt
7) Y_lt: bbox Y_lt
8) Width: bbox width
9) Height: bbox height
10) Number of objects involved: total number of objects involved
```

After the above 10 fixed columns, there will be variable number of columns where each column corresponds to each existing object. If the value==1, it indicates that the particular object is involved. Note that 3rd additional column==1 in event annotation corresponds to the 3rd object ID (in sorted order), i.e., it does not mean that object ID==3 is involved. It may be object ID==5.

2.4. Event Ground Truth

Ground truth bbox of an event is defined around the history of moving bboxes of the involved person during the duration of an event. In detail, from the moving bboxes of an involved person, the minimum and maximum spatial span of the bboxes are computed as: `xmax, xmin, ymin, ymax.` Then, the event bounding box is computed by expanding it to include surrounding pixels both horizontally (factor of 3), and vertically (factor of 1.5) as follows:

```
cx = (xmin + xmax) / 2.0
cy = (ymin + ymax) / 2.0

xmin_event = cx - 3.0*(cx-xmin_)
ymin_event = cy - 1.5*(cy-ymin_)
xmax_event = cx + 3.0*(xmax_-cx)
ymax_event = cy + 1.5*(ymax_-cy)
```

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.

Ground truth duration of an event is precisely defined by the starting and ending frame numbers described in Section 2.3.22.3.2 NOTE that any ground truth duration is at least 2.0 seconds long.

2.5. 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_viratdata1.m`'. By running the sample software, it will store event-specific frames with involved objects in the 'images' folder, at the quarter size of the original image. Sample videos and annotations can be found in other folders.

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.

3. Evaluation Metrics

3.1. Definitions:

3.1.1. Activity

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.

3.1.2. 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.

3.1.3. Activity Matching Criterion

An activity or detection is defined as a tuple of (label, track). Given an activity pair of {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 20%, 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 temporal intersection ratios should be above 20% 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.

3.2. Activity-level metrics for Competition

Activity 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, Pd, False Alarm Rate, F-scores, weighted aggregate F-score. **NOTE that judges may define their own weights on their disposal to incorporate multiple metrics.**

3.2.1. **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 3.2.2.

3.2.2. **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'.

3.2.3. **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.*

3.2.4. **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:

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

3.2.5. **Weighted Aggregate F-score**: a weighted aggregate F-score will be used as a measure for judge decisions. 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.

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

3.2.6. Official Scoring Software

Official scoring software will be distributed by the end of Feb. 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.

3.3. Correct Detections and False Alarms

3.3.1. Correct Detections

3.3.1.1. 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.

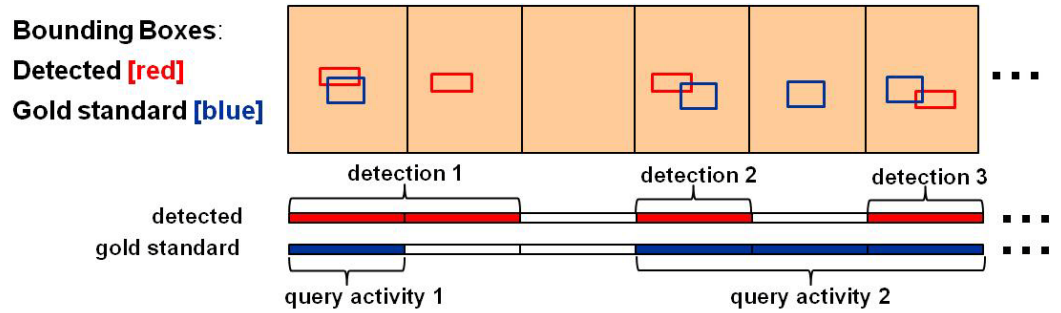


Figure 1. Examples of multiple correct detections (detection 2 & 3) on one query activity (activity 2), assuming that detections 2 & 3 have valid temporal overlap with the query activity 2.

3.3.1.2. 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.

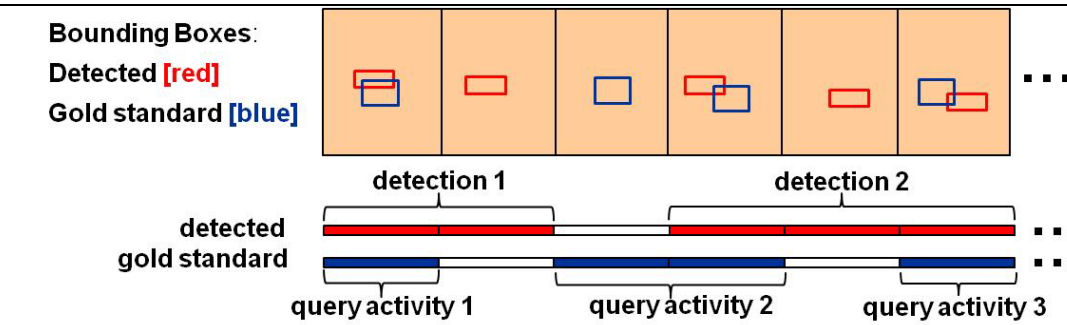


Figure 2. Examples of one detection (detection 2) counted as two correct detections against two ground truth activities (query activities 2 & 3), assuming that they satisfy spatio-temporal matching criteria.

3.3.1.3. Detections and ground truths can both occur concurrently, and detections will be scored against all concurrent ground truths, following the policies in 3.3.1.1 and 0. Examples are shown in Figure 3.

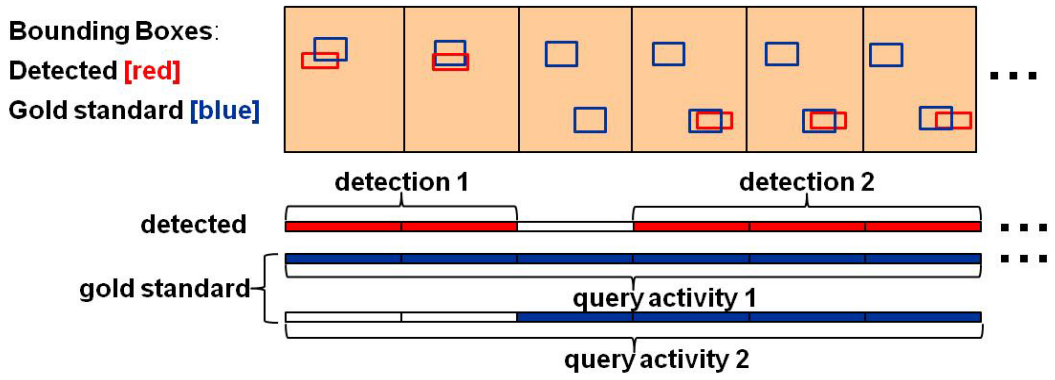
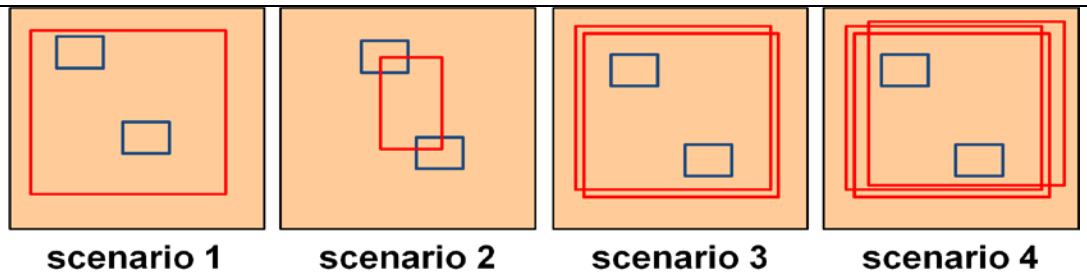


Figure 3. Where there are concurrent events, the bounding box provided with each detection must overlap the annotated bounding boxes. In the example above, both activities are detected, assuming they satisfy spatial and temporal match criteria.

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.

3.3.2. False Alarms

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

3.3.2.1. Examples of false alarms:

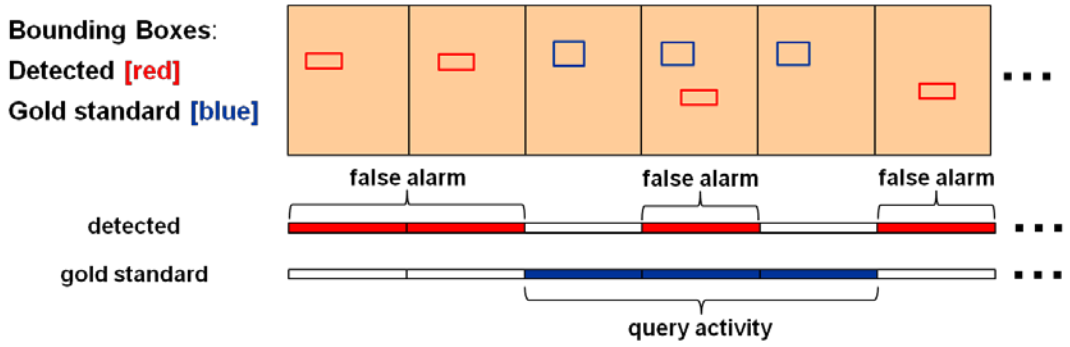


Figure 5. Examples of false alarms.

4. Annotation Standards

4.1. Object Annotations

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 upper body of a person is the only visible part, then, only the upper body should be labeled as 'person'.

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. For example, 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 such as parking spots, which people interact with are annotated separately as stationary objects. 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 types of activities considered for this workshop have optional bounding boxes and they may not exist. 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.

4.2. 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 6 different types of activities, for this competition. 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.

4.2.1. 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', 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.

4.2.2. Person Unloading an Object from a Vehicle

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

Annotation: 'Person', 'Object', 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'.

4.2.3. 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.

4.2.4. 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.

4.2.5. 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).

4.2.6. 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.

5. Contact Information

For any question regarding VIRAT Video Dataset, please send an e-mail to: contact@viratddata.org

6. 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.