

MEVA Annotation JSON File Format

Version: 0.1
Date: 2019-05-30
ActEV Team
NIST

TABLE OF CONTENTS	
Background	3
Overview	3
1. Annotation Output Files	3
1.1. File Index	3
1.2. Activity Index	4
1.3. Activity Instance Annotation File	4
References	9
Disclaimer	10

Background

The ActEV Independent evaluations in 2019 will use the Multiview Extended Video with Activities ([MEVA
http://mevadata.org](http://mevadata.org)) test video dataset for the activities as defined in the annotation guide [1]. NIST is encouraging the crowd sourcing of annotations to greatly enhance the amount of annotated MEVA training data for the activities and share them with the ActEV research community. This document describes the requirements a “sharer” must adhere to and the common data files for exchanging data. By now, you should be familiar with:

- The MEVA video data
- Reviewed the annotation exemplars
- Reviewed the annotation guidelines for both the tracked object types and the activity definitions.

To annotate a video, one must temporally localize each activity instance (in terms of the start/end the performed activity) and spatially annotate bounding boxes of object(s) associated with the activity instance.

Annotators can use any of the of video annotation tools used by the video analytics community, some of the commonly used ones are described in [2][3][4][5][6][7][8][9]. Since most video annotation tools have their own output format, annotation sharers are responsible for translating their annotations into a common format, where all the data can be aggregated, compared and merged to create a training dataset.

The common annotation data format is a set of three, JSON-formatted data files that describe the video files annotated, the activities annotated within the videos, and activity instance annotations themselves.

By convention, the names of three files are:

<code>file-index.json</code> <code>activity-index.json</code> <code>activities.json</code>
--

Section 1 describes each of the three data file formats and the content they hold.

1. Annotation Files

Annotations are represented by three file types. (1) The **file-index** defines the list of video files annotated and metadata about the video files including the annotated frames. (2) The **activity-index** enumerates the list of annotated activities and metadata about the activities including the annotated objects. The textual description of the activities and objects are contained in the “ActEV Annotation Definitions for MEVA Data”[1]. (3) the **activities** file contains the annotations for each instance of an activity.

Each triplet of files forms a coherent set of annotations meaning a triplet could represent a single file of a single activity or multiple files for multiple activities.

These definitions will be used throughout the document:

- “**Activity Instance**”: an observed instance of activity. It could be visibly present within a single camera view or across multiple camera views (if an annotation regime supports multiview annotation which is optional).

- **“Frame State Signal”**: a signal-based representation of a given variable’s “state” at frame X which continues until the “state” changes. The data structure is a dictionary with keys being a frame number and the value being the state. The value can be either a simple data type or another dictionary.
- **“<value>”**: a dictionary key that is a name, e.g., a file name, that must be unique within the given dictionary

1.1. FILE INDEX

The file index JSON is a two-level dictionary with first being indexed by the video source file’s name and the second level representing metadata about the file. An example, along with an explanation of the fields is included below.

```
{
  "2018-03-07.16-50-00.16-55-00.bus.G475.avi": {
    "framerate": 30,
    "filename":
"MEVA/video/KF1/2018-03-07/16/2018-03-07.16-50-00.16-55-00.bus.G475.avi",
    "selected": {
      "1": 1,
      "20941": 0
    }
  },
  "2018-03-07.16-50-00.16-55-00.admin.G329.avi": {
    "framerate": 30,
    "filename":
"MEVA/video/KF1/2018-03-07/16/2018-03-07.16-50-00.16-55-00.admin.G329.avi",
    "selected": {
      "11": 1,
      "201": 0,
      "300": 1,
      "20656": 0
    }
  }
}
```

- **<file>**: A string for the name of the video file without a path and including the extension.
 - **framerate**: number of frames per second of video
 - **filename**: the relative path and name of the file relative to the distributed video directories.
 - **selected**: a Frame State Signal with keys representing a frame number and the values being 1 (for annotated frames) and 0 (for unannotated frames) within the file <file>.
 - **<framenum>**: 1 or 0, indicating whether or not the activity will be annotated for the given frame. Note that records are only added here when the value changes. For example in the above sample, frames 1 through 20940 in file “VIRAT_S_000000.mp4” are selected for annotation. The default signal value is 0 (not-selected), and the frame index begins at 1, so for file “VIRAT_S_000001.mp4”, frames 1 through 10 are not selected. Also note that the signal must be turned off at some point after it’s been turned on.

1.2. ACTIVITY INDEX

The activity index JSON file lists the activities annotated for ALL files enumerated in the file index. The file is a two-level dictionary with first being indexed by the activity name (as defined in the MEVA Annotation spec and the second level representing metadata about the activity. It is assumed that if an activity is present in this file, then the activity annotation files below have been annotated for that activity (even if no instances of the activity is found.) An example, along with an explanation of the fields is included below.

```
{
  "Closing": {
    "objectTypes": [
      "Door",
      "Person",
      "Vehicle"
    ]
  },
  "Closing_Trunk": {
    "objectTypes": [
      "Person",
      "Vehicle"
    ]
  },
  "Entering": {
    "objectTypes": [
      "Door",
      "Person",
      "Vehicle"
    ]
  },
  "Exiting": {
    "objectTypes": [
      "Door",
      "Person",
      "Vehicle"
    ]
  },
  "Loading": {
    "objectTypes": [
      "Person",
      "Vehicle",
      "Prop"
    ]
  }
}
```

- <activity>: A string for the name of the activity as defined in the MEVA Annotation Spec.
 - objectTypes: An array of strings enumerating the set of objects annotated for each instance with respect to the given activity.

1.3. ACTIVITY INSTANCE ANNOTATION FILE

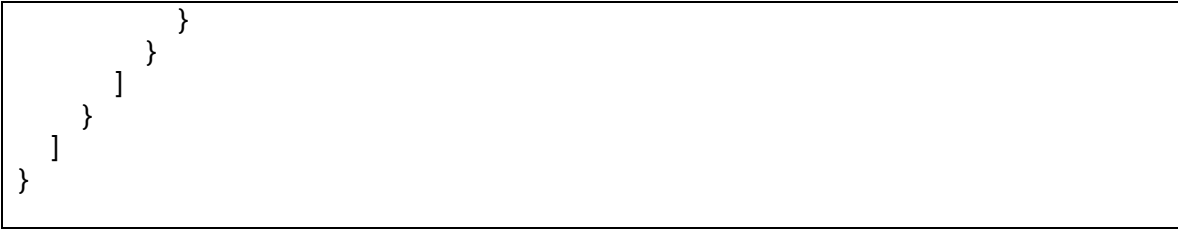
The activities JSON file lists the activities annotated for ALL files enumerated in the file index. The file is a two-level dictionary with first being indexed by the activity name (as defined in the MEVA Annotation spec and the second level representing metadata about the activity. It is assumed that if an activity is present in this file, then the activity annotation files below have been annotated for that activity (even if no instances of the activity is found.) An example, along with an explanation of the fields is included below.

The annotation output file should be a JSON file that includes a list of videos annotated, along with a collection of activity instance records with spatio-temporal localization information (depending on the annotation). An activity detection annotation output file is included inline below, followed by a description of each field.

The annotation output is shown for videos annotated, with a collection of activity instance records with temporal localization information only. An activity detection annotation output file is included inline below, followed by a description of each field.

```
{
  "activities": [
    {
      "activity": "Talking",
      "activityID": 1,
      "localization": {
        "VIRAT_S_000000.mp4": {
          "1": 1,
          "20": 0,
          "100": 1,
          "112": 0,
        }
      },
      "objects": [
        {
          "objectType": "person",
          "objectID": 1,
          "localization": {
            "VIRAT_S_000000.mp4": {
              "1": { "boundingBox": { "x": 10, "y": 30, "w": 50, "h":
20 } }
              "20": {}
              "100": { "boundingBox": { "x": 10, "y": 30, "w": 50, "h":
20 } }
              "104": { "boundingBox": { "x": 60, "y": 60, "w": 50, "h":
20 } }
              "108": { "boundingBox": { "x": 30, "y": 90, "w": 50, "h":
20 } }

              "112": {}
            }
          }
        }
      ]
    }
  ]
}
```



- activities: An array of annotated activity instances. Each instance is a dictionary with the following fields:
 - activity: The name (e.g. “Talking”) from the MEVA Annotation Spec. [1]
 - activityID: a unique, numeric identifier for the activity instance. The value must be unique within the list of activity detections for all video source files processed (i.e. within a single activities JSON file)
 - localization: The temporal localization of the activity instance encoded as dictionary of Frame State Signals indexed by the video file id(s) for which the activity instance is witnessed. Each Frame State Signal (for a video) has keys representing a frame number and the value being 1 (the activity instance is present) and 0 (otherwise) within the given file. Multiple Frame State Signals can be used to represent an activity instance being present in multiple video views. In this case, frame numbers are relative with respect to the video file.
 - objects: An array of objects annotated with respect to the activity instance. Each unique object is represented by the following dictionary:
 - objectType: A string identifying the objects type as one of the track types defined in the MEVA Annotation Spec.
 - objectID: unique, numeric identifier for the object. The value must be unique within a single activities JSON file.
 - Localization: The temporal-spatial localization of the object encoded as dictionary of Frame State Signals indexed by the video file id for which the object is witnessed. Each Frame State Signal (for a given video) has keys representing a frame number and the value is a dictionary describing the spatial localization of the object. The spatial dictionary has 1 key ‘boundingBox’ which is itself a dictionary described as a pixel ‘x’, ‘y’, ‘w’, and ‘h’ for the the x-position, y-position, width and height respectively. The (0,0) (x,y) position is the top left pixel.

REFERENCES

1. [ActEV Annotation Definitions for MEVA Data](#)
2. [Kitware annotation tool](#)
3. [The VGG Image Annotator](#)

4. [Scalabel \(used for annotation of Berkeley DeepDrive project\)](#)
5. [VATIC - Video Annotation Tool](#)
6. [BeaverDam](#)
7. [VoTT: Visual Object Tagging Tool](#)
8. [Computer Vision Annotation Tool \(CVAT\)](#)
9. [Efficient Annotation of Segmentation Datasets with Polygon-RNN++](#)

DISCLAIMER

Certain commercial equipment, instruments, software, or materials are identified in this evaluation plan to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.