2022 OpenFAD Evaluation Plan (Open Fine-grained Activity Detection)

Date: 2022-07-29

Yooyoung Lee, Lukas Diduch, Jonathan Fiscus, Jeffrey Byrne*

National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899

*Visym Labs, Cambridge MA, 02140

Contact: fad-te@list.nist.gov

Disclaimers

Certain commercial equipment, instruments, software, or materials are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, NIST, or the U.S. Government.

Table of Contents

1 Introduction

This document describes the evaluation for the NIST 2022 Open Fine-grained Activity Detection (OpenFAD) challenge. The evaluation plan covers task definitions, task conditions, file formats for system inputs and outputs, evaluation metrics and protocols for participating in OpenFAD challenges (see details at https://openfad.nist.gov).

OpenFAD is an activity classification and detection evaluation to measure how well systems can automatically classify or temporally detect fine-grained activities collected from the Consented Activities of People (CAP) dataset using handheld devices (https://visym.com/collector). Fine-grained activity detection is an emerging research area. Previous workshops, such as ActivityNet at CVPR'21 and Behavior at ICCV'21, have focused on temporal localization and activity classification, however fine-grained classes have been limited to sports video analysis (e.g. DeeperAction workshop at ICCV'21). The CAP data is focused on fine-grained activities related to coarse-grain activities in the MEVA data and evaluated in the ActEV Evaluations. The hypothesis we intend to test experimentally is whether coarse-grained activity detection can be improved with better modeling of fine-grained activities.

The OpenFAD challenge provides a set of data (e.g., training, validation and test sets) to participants to train and run a system on their own hardware platform and submit their system outputs to a web-based leaderboard for scoring and displaying results. Data resources are available for download at https://visym.github.io/cap. Scorer code is available for download at https://github.com/usnistgov/FadScorer.

Any questions or comments concerning the OpenFAD should be sent to fad-te@list.nist.gov.

2 Tasks

The OpenFAD evaluation focuses on the two tasks: Activity Classification and Temporal Activity Detection. The primary goal of the challenges is to understand system behavior in classification and detection of coarse-grained (super-class) vs fine-grained (sub-class) activities.

2.1 Task Definition

2.1.1 Activity Classification

The Activity Classification (AC) task is to assign a single activity class label to each video clip from a set of predefined classes and provide a confidence score.

The system will be presented with minimally 4-second trimmed video clips, each clip containing a single activity that is one of the predefined activity classes. The AC system must output predicted activity classes and confidence scores with higher numbers indicating the clip is more likely to contain that activity. The confidence score can be any real number in the range [0, 1] and a single confidence score is required per clip. The primary metric for measuring classification performance will be Mean Average Precision (mAP) (see Section 6.1.1).

2.1.2 Temporal Activity Detection

The Temporal Activity Detection (TAD) task is to automatically detect and temporally localize each activity instance in untrimmed video.

The system will be presented nominally 45-second untrimmed video clips, where no activity instances overlap during the same temporal duration and each clip can contain more than one activity occurrence. The total number of instances per clip and the activity classes present in the clip is unknown to the TAD system. The system must output a predicted activity class from the set of class labels in the activity classification task, a start frame, end frame and real valued confidence score for each detected activity in the clip. The primary metric for measuring detection performance will be Mean Average Precision (mAP) (see Section 6.2.1).

2.2 Protocol and Rules

The challenge participants agree not to probe the test videos via manual/human means such as looking at the videos or annotating videos to produce the activity type and timing information from prior, during and after the evaluation. The participants are NOT allowed to use the test set for purposes of training, modeling, or tuning their algorithms. All machine learning or statistical analysis algorithms must complete training, model selection, and tuning prior to running the OpenFAD test data.

Each participant is allowed to submit system output for evaluation once per 24 hour period.

All clips must be processed independently of each other within a given task and across all tasks, meaning content extracted from the video clip data must not affect the processing of another clip.

While participants may report their own results, participants may not make advertising claims about their standing in the evaluation, regardless of rank, or winning the evaluation, or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113)14 shall be respected: NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.

At the conclusion of the evaluation, NIST may generate a report summarizing the system results for conditions of interest. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.

3 Data Resources

The Consented Activities of People (CAP) dataset will be used for OpenFAD evaluations. The CAP data contains a new annotated dataset of fine-grained and coarse-grained activity classes of consented people, curated using the Visym Collector platform. The dataset was collected with the following goals:

- **Atomic.** Activities are atomic at ≤ 3 seconds and visually grounded (e.g. activities should be unambiguously determined from the pixels).

- **Non-overlapping.** All activities are performed independently, and no activities are performed jointly or simultaneously overlapping with other activities (e.g. a subject will not simultaneously perform the "person uses cell phone" activity while performing the "person removes hat" activity).
- **Non-occluded.** All activities are visible to the camera and not fully occluded by the subject or other objects. Activities may be partially occluded by the body of the subject, but are still visible in the clip.
- **Person centered.** All activities are collected from handheld mobile devices at a fixed security perspective (e.g. looking down on a scene from above) and include a single consented person as the primary subject. Non-consented subjects have their faces blurred in-app. Subjects are tasked with performing specific atomic activities, person/object or person/person interactions.
- **Fine-grained class vs Coarse-grained class.** All activities are selected so that there are subtle differences between classes where the activity representation and discrimination is critical for performance, rather than the scene context or object detection. For example, the dataset was collected with the coarse-class "person dresses" and fine-class that includes "person puts on shirt" and "person puts on jacket", which are visually distinct activity classes where the motion pattern and object interactions are important for representation. Furthermore, the dataset was provided with activity classes that are closely related such as "person talks on cell phone" vs. "person scratches head" and "person hugs person" vs. "person kisses cheek of person" where subtle discrimination between activity classes is necessary. A list of the fine-grained activities in the dataset is available along with the coarse-grained activity classes and a one-to-one mapping from fine to coarse classes.
- **Around the house.** In order to maximize engagement with the distributed collector community, the collection involves objects, locations and activities that most collectors have easy access to and can easily perform without practice.
- **Ethical.** All videos are collected with informed consent for how the videos submitted by collectors will be shared and used.
- **Worldwide**. Videos are collected from over 780 collectors in 33 countries.

The CAP dataset is annotated with bounding boxes around the primary actor, and temporal start/end frames for each activity instance. An interactive visualization and video summary is available for review on the dataset distribution site.

The CAP dataset is split into activity classification and temporal activity detection subsets. The CAP activity classification subset is a set of trimmed clips of subjects performing exactly one activity. We leveraged the Visym Collector platform to task subjects around the world to collect and annotate videos of consented people performing activities. Each clip is temporally centered on the activity, each clip is three seconds long on average and at least one person is present in each clip.

The CAP activity detection subset is a set of untrimmed clips of subjects performing one or more activities in a sequence. We instruct the collectors on the Visym Collector platform to choose from a list of activities to perform in a "scenario". For example, one scenario may be "Come inside from the cold" which involves a sequence of activities such as opening doors, taking off a hat or taking off a jacket. The subject performs the selected activities in any order they choose, then the collector records the scenario on their mobile device using the Collector app, and edits the video in-app to annotate the chosen activities. Each activity detection video contains a variable number of activities performed by one or more people at different times in the video.

The CAP training and validation dataset is available for public download in the form of a single archive file containing videos and metadata at the CAP dataset distribution site. Performers may choose a training and validation set containing videos that are temporally clipped around each activity, or temporally padded to a

minimum duration of 4 seconds, as is best suited for the performer system. The CAP dataset site includes a dataset explorer, dataset summary and activity definitions as described in https://visym.github.io/cap.

The training and validation dataset is licensed with Creative Commons Attribution 4.0 International (CC BY 4.0) license. All subjects with personally identifiable information visible (PII) in the videos have consented for the use of their PII for the purposes of visual AI research.

3.1 Training/Validation Set

The OpenFAD training/validation set is delivered as a single gzipped tarball for each task which is unpacked to the following contents and subdirectory structure:

README.md	A helpful documentation file in markdown format
cap_activity_classification_ref.csv	The main reference file for the activity classification task (AC task only)
cap_temporal_activity_detection_ref.csv	The main reference file for the temporal activity detection task (TAD task only)
train_val.json	The recommended split of video_file_id into training/validation subsets
annotations/	A subdirectory with dense spatiotemporal annotations and metadata in vipy json format
videos/	A subdirectory containing video clips organized by <activity_id>/<video_file_id>.mp4

The following constitutes the reference file format for the AC task.

`cap_activity_classification_ref.csv`

video_file_id	(string) The globally unique ID of the video clip (e.g., F5F8CFF4-3A54-498B-8CAA-07E76DF9FF63_0)
frame_rate	(float) The frame rate in frames per second of the video clip (e.g. 24.0)
activity_id	(string) The ID of the activity class defining the activity category

The following constitutes the reference file format for the TAD task.

`cap_temporal_activity_detection_ref.csv`

video_file_id	(string) The globally unique ID of the video clip (e.g., F5F8CFF4-3A54-498B-8CAA-07E76DF9FF63_0.mp4)
frame_rate	(float) The frame rate in frames per second of the video clip (e.g. 24.0)
activity_id	(string) The ID of the activity class defining the activity category
start_frame	(int) The starting frame where the target activity occurs (inclusive), relative to the video frame_rate
end_frame	(int) The ending frame where the target activity occurs (inclusive), relative to the video frame_rate

3.2 Test Set

The OpenFAD test set is delivered as a single gzipped tarball which is unpacked to the following contents and subdirectory structure:

README.md	A helpful documentation file in markdown format
cap_activity_classification_index.csv	The system input file for the activity classification task
cap_temporal_activity_detection_index.csv	The system input file for the temporal activity detection task
videos/	A flat subdirectory containing video clips organized by <video_file_id>.mp4

The CAP test set contains trimmed and untrimmed videos for evaluation of the activity classification and temporal activity detection tasks. The ground truth annotations for the CAP test set is sequestered for evaluation and will not be shared with the challenge performers.

The CAP test set will be available as a single archive file. The archive file download URL will be available to challenge performers once they have registered on the challenge site, and accepted a license agreement. The license agreement states that the performers will not (i) redistribute the videos, (ii) share the videos publicly (iii) use any portion of the test set for training or (iv) attempt to annotate or analyze the test set for any other purpose than to generate a system output file for scoring.

The CAP test set for the activity classification task contains trimmed videos that are temporally padded to a minimum duration of 4 seconds centered on an activity. A performer may select only the central frames of a clip for testing, as best suited to the performer system.

4 System Input

For a given task, a system's input is the task index file, called `cap_<task_id>_index.csv`. Given an index file, each row specifies a test trial. Taking the corresponding video(s) as input(s), systems perform classification and detection tasks.

The following format constitutes the index file for the system input:

```
cap_temporal_activity_detection_index.csv
cap_activity_classification_index.csv
```

video_file_id	The video file ID to process (e.g., F5F8CFF4-3A54-498B-8CAA-07E76DF9FF63_0). Videos are distributed with the flat relative directory structure videos/<video_file_id>.MP4
frame_rate	The frame rate in frames per second of the video file e.g., 24.0

5 System Output

In this section, the types of system outputs are defined. The OpenFADScore package [TBD link] contains a submission checker that validates the submission in both the syntactic and semantic levels. Participants should check their submission prior to sending them to NIST. NIST will reject submissions that do not pass validation. Consult the OpenFADScore documentation for validation instructions [TBD link].

5.1 System Output File

The system output file must be a CSV file. The filename for the output file is **user defined** and must use the following convention; **user-defined string** that identifies the submission with **no spaces or special characters besides '_-.'** (e.g., `` `sub_ac_method-4.csv` ``).

The system output CSV file for the AC task must follow the format below:

video_file_id	(string) The ID of the video clip, e.g., 000001
activity_id	(string) e.g., 'person_scratches_head'
confidence_score	(float) in range [0,1], larger is more confident

The system output CSV file for the TAD task must follow the format below:

video_file_id	(string) The ID of the video clip, eval defined
activity_id	(string) The ID of the activity detected, eval defined

start_frame	(int) The starting frame where the target activity occurs relative to the frame_rate of the video clip
end_frame	(int) The ending frame where the target activity occurs relative to the frame_rate of the video clip
confidence_score	(float) in range [0,1], larger is more confident

5.2 Validation/Submission

5.2.1 Validation

The video_file_id column in the system output `[submission-file-name].csv` must be consistent with the video_file_id in the `cap_<task_id>_index.csv` file. The row order may change, but the number of the files and file names from the system output must match to the index file.

To validate your system output locally please use the command-line command as shown in Appendix A.

5.2.2 Submission

System output submission to NIST for subsequent scoring must be made through the web-platform using the submission instructions described on the webpage (https://openfad.nist.gov/help/submissions). To prepare your submission, you will first make **.tar.gz** (or **.tgz**) file of your system output CSV file via the UNIX command `'tar zcvf [submission_name].tgz [submission_file_name].csv'` and then upload the system output tar file under a new or existing 'System' label. This system label is a longitudinal tracking mechanism that will allow you to track improvements to your specific technology over time.

Please submit your files in time for us to deal with any transmission errors that might occur well before the due date if possible. Note that submissions received after the stated due dates for any reason will be marked late.

6 Metrics Definition

In this section, two types of metrics are used in evaluating system performance: activity classification and temporal activity detection metrics.

6.1 Activity Classification (AC) metrics

The activity classifiers aim to predict an activity class in a given video clip with a high confidence; therefore, a classification performance is computed using two attributes per processed video: the activity class and the confidence score.

To estimate the measures below, the confusion matrix is defined as:

- True Positives (TP): correctly predicted a target class for a ground-truth target class at a confidence score threshold
- False Positives (FP): falsely predicted as a target class for a ground-truth non-target class at a confidence score threshold

6.1.1 Mean Average Precision (mAP)

For the AC task, the interpolated mean average precision (mAP) [1][2] is used as a primary metric. For multi-class classification, precision addresses the question of how many of the activities that the classifier predicted to be true positives are actually true positives. Recall addresses the question of how many of all the actual true positive activities (ground truth positives) are found to be true positives by the classifier.

The activity classifier performance is estimated using the activity class labels and the confidence score. The confidence score can be taken into account in the precision and recall by considering as true positives scores that are larger than or equal to a confidence score threshold τ. The precision (*P*) and recall (*R*) can be thus defined as a function of the confidence score threshold:

$$P(\tau) = \frac{\#TP(\tau)}{\#TP(\tau) + \#FP(\tau)} \quad (1)$$

$$R(\tau) = \frac{\#TP(\tau)}{\#ground\ truth} \quad (2)$$

where both precision and recall are in the range [0, 1], with 0 being the worst and 1 being perfect. The notation $\#TP(\tau)$ is a number of the true positives at threshold τ. The notation $\#FP(\tau)$ is a number of the false positives at threshold τ. The precision-recall curve is a plot of precision as a function of recall, so the area under the precision-recall curve (AP) can be defined using an approximate discrete form:

$$AP = \sum_{k=0}^{K} (R(k) - R(k+1))\, P_{interp}(R(k)) \quad (3)$$

where $P_{interp}(R(k))$ is the interpolated precision for recall value $R(k)$ at rank k, such that:

$$\tau(k),\ k = 1, 2, \ldots, K \ and\ \tau(i) > \tau(j)\ for\ i > j \quad (4)$$

The curve is interpolated to avoid a zig-zag behavior and the detailed interpolation procedure can be found in Padilla et. al [1]. The AP value will be calculated for each activity class independently and then averaged over all activity classes to yield mean $AP\ (mAP)$:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad for\ N\ activity\ classes \quad (5)$$

6.2 Temporal Activity Detection (TAD) Metrics

Temporal activity detection predicts the temporal location and activity type of multiple activity instances in a video clip. Each activity instance is defined by an activity type, begin/end time and a confidence score. The temporal segment detected by a system output is compared to the ground-truth activity instance spans using multiple intersection over union (IoU) thresholds to determine if a detection is correct. The detection is represented by three attributes: an activity class, a temporal IoU, and a confidence score.

To estimate system performance based on the measures below, the confusion matrix is defined as:

- True Positives (TP): correctly predicted a target class for a ground-truth target class and temporal segment has an IoU threshold above a specified temporal IoU threshold.

- False Positives (FP): an incorrect prediction of a ground-truth temporal segment as a non-target activity class above a specified temporal IoU threshold.
- Missed Detections (MD): an undetected ground-truth temporal segment above a specified temporal IoU threshold. MD is not taken into account for computing mAP performance measure.

Due to annotation error/ambiguity, we utilize a no-score region (NR) that is a system instance under the NR reference duration not scored.

6.2.1 Mean Average Precision (mAP) @ IoU

For the TAD task, the mAP at IoU is used as the primary metric. The TAD performance measure is computed independently for each activity class, by accessing the temporal IoU between the reference and system output instances along with a confidence level of predicted system output. For example, if the IoU value of the reference and system instances are greater than or equal to a given IoU threshold (e.g., IoU >= 0.2), it is scored as correct detection. If multiple instances of system output have the same IoUs with a reference instance, an instance with a higher confidence score is selected as correct detection while the others are counted as incorrect detection. The no-score region (NR) duration of the reference is neither panelized nor scored for system instances that meet the IoU criteria. A pictorial depiction of the system performance computation is illustrated in Figure 1.

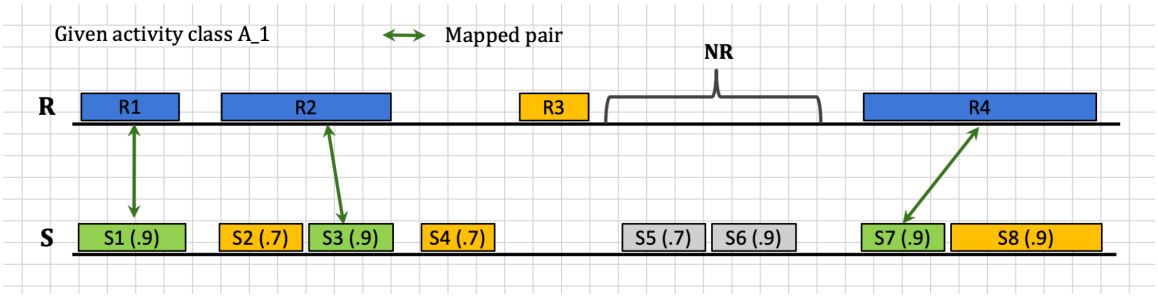

Figure 1. Depiction of system performance measures using IoU and confidence score
(In the system output (*S*), the first number indicates *instance id* and the second indicates *confidence score*. For example, S1 (.9) represents the instance S1 with corresponding confidence score 0.9. Green indicates TP instances, red for FP instances, yellow for MD instances, and gray indicates instances that are not scored.)

The confidence score is taken into account in the precision and recall, as illustrated in the equations (1) and (2), for the TAD task. The $AP@IoU$ value can be calculated for each activity class independently and averaged over all activity classes to yield:

$$mAP@IoU = \frac{1}{N}\sum_{i=1}^{N} AP@IoU_i \quad for\ N\ activity\ classes \tag{6}$$

The IoU thresholds in this evaluation are between 0.2 and 0.7 with a step size of 0.1.

Appendix A SCORER USAGE

The OpenFADScorer includes the performance measure calculation of mAP (see Section 6) using system's input and output for activity classification (AC) and temporal activity detection (TAD) tasks. The following examples demonstrate how to use the validation and scoring their system output using the OpenFADScorer tool:

Validation

```
# validate classification system output

fad-scorer validate-ac-hyp -r [reference.csv] -y [system_output].csv

# validate detection system output

fad-scorer validate-tad-hyp -r [reference.csv] -y [system_output].csv
```

Scoring

```
fad-scorer score-ac -r [reference.csv] -y [system_output.csv] -o [output_dir]

fad-scorer score-tad -r [reference.csv] -y [system_output.csv] -o [output_dir]
```

REFERENCES

[1] R . Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit," *Electronics*, vol. 10, no. 3, p. 279, Jan. 2021, doi: 10.3390/electronics10030279

[2] G. Salton and M.J. Mcgill, "Introduction to modern information retrieval", McGraw-Hill Inc, 1986