UNIVERSITÉ Concordia UNIVERSITY

# ROBUST SCORING AND RANKING OF OBJECT TRACKING TECHNIQUES

Tarek Ghoniemy
Julien Valognes
Maria A. Amer

Department of Electrical and Computer Engineering

Montréal

# Motivation

- ❖ Object tracking is an active research area
- ❖ Numerous techniques are being continuously proposed

- ❖ To evaluate trackers: averaging of a performance measure over all test videos does not account for data dispersion nor for similarities between trackers

➔Our tracker ranking uses a robust estimator (MAD) to effectively quantify the statistical dispersion in the data

# Prior work

- Prior work reports average performance over all test data

- [1] address the statistical significance of data
  - However, they do not tackle similarities between trackers

- [1, 2, 3] run each tracker 15 times to obtain a better statistic, and average them to account for randomness

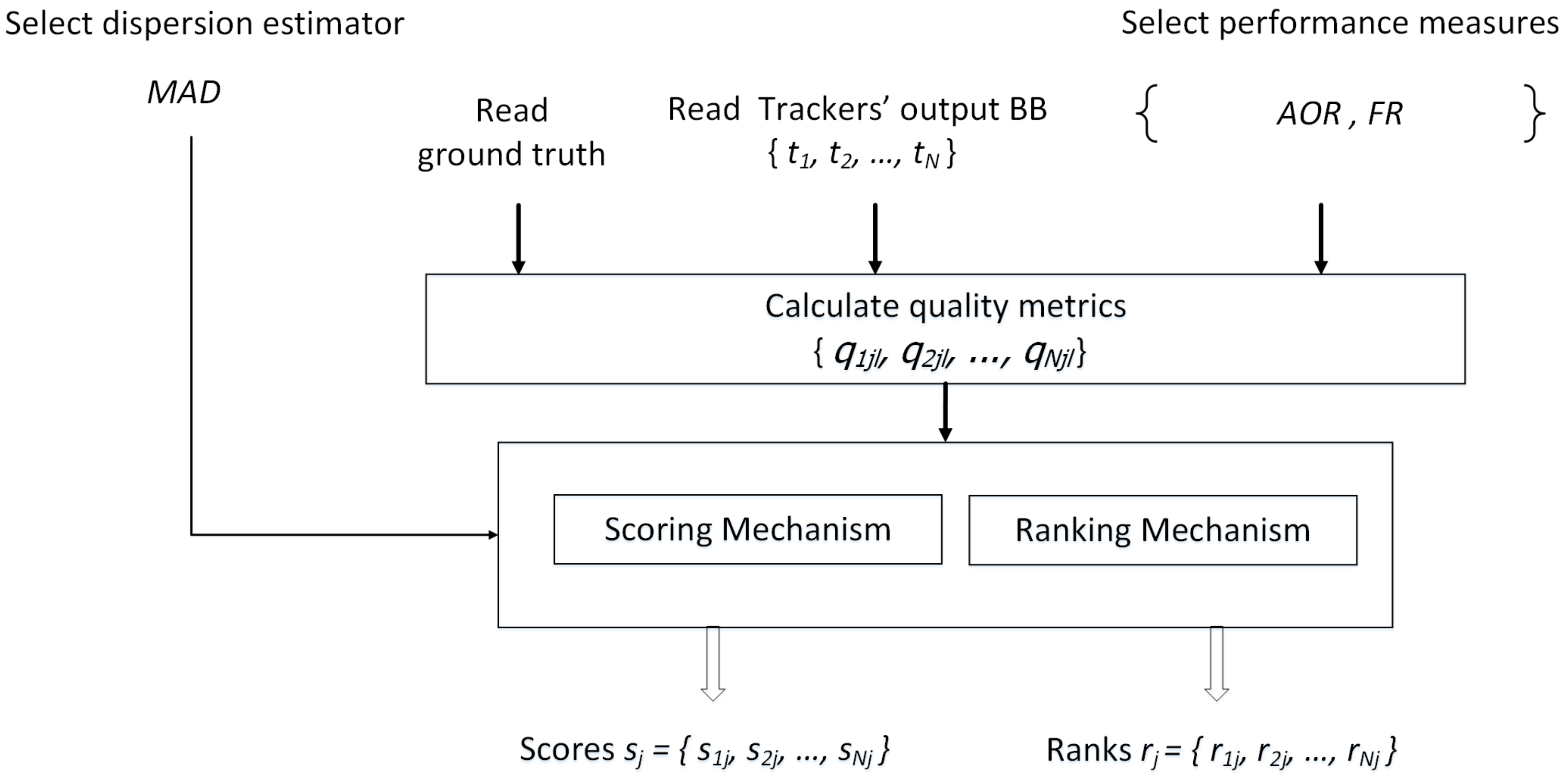
# Proposed method

- A good performance model should be robust against both outliers and deviations from model assumptions

- Our approach assigns
  1. a score to a tracker on a per-sequences basis subject to MAD
  2. a rank to a tracker depending on a MAD-based similarity

→ Our approach well highlights the **relative** performance of each tracker using **uncorrelated** objective measures

# Proposed method



Select dispersion estimator

Select performance measures

*MAD*

Read ground truth

Read Trackers' output BB $\{ t_1, t_2, \ldots, t_N \}$

$\{$ *AOR* , *FR* $\}$

Calculate quality metrics $\{ q_{1jl}, q_{2jl}, \ldots, q_{Njl} \}$

Scoring Mechanism

Ranking Mechanism

Scores $s_j = \{ s_{1j}, s_{2j}, \ldots, s_{Nj} \}$

Ranks $r_j = \{ r_{1j}, r_{2j}, \ldots, r_{Nj} \}$

# Proposed method

1. Calculate the quality data for all trackers according to a metric over all test sequences

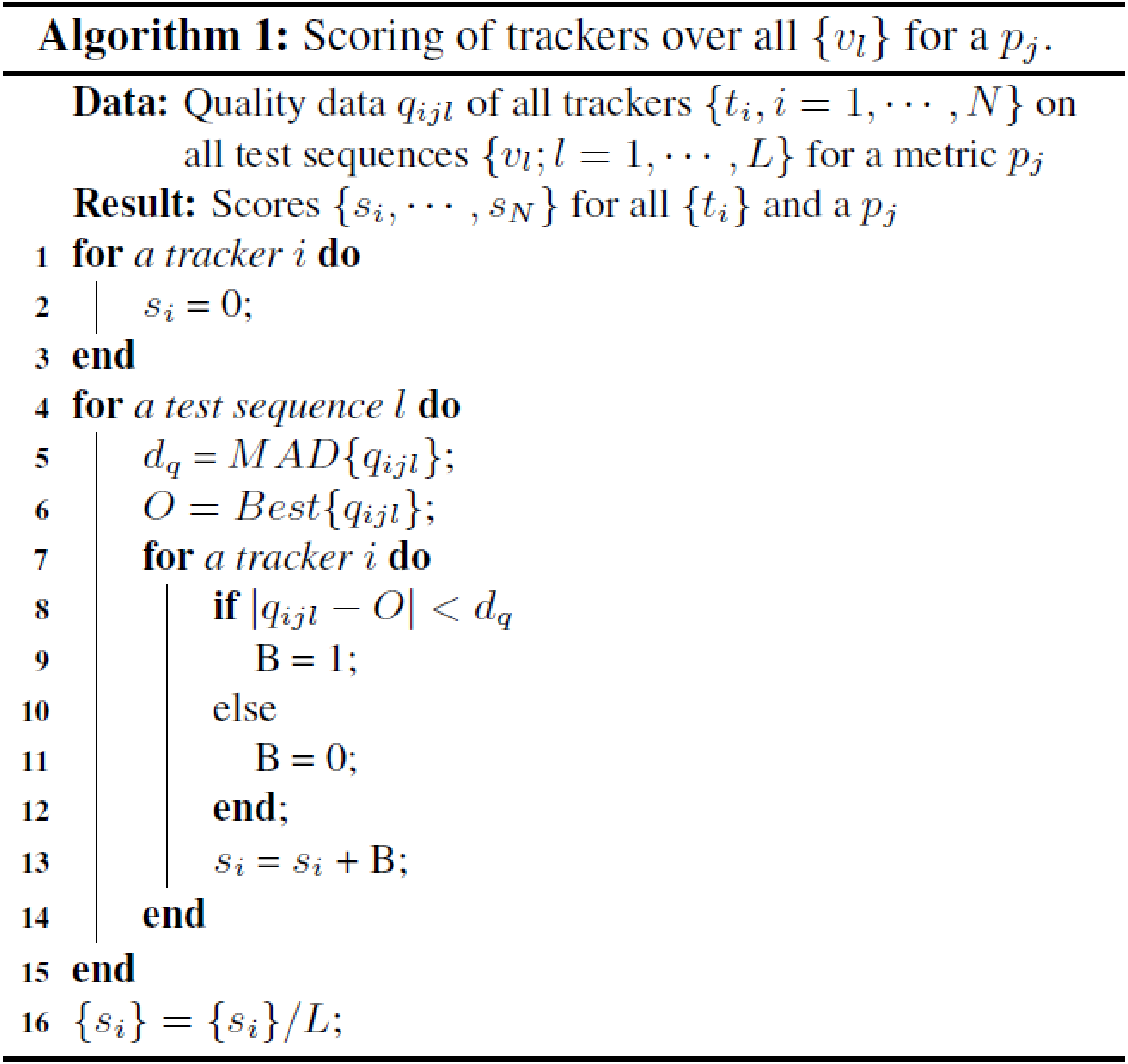
2. **Scoring**: apply MAD on the quality data for all trackers and assign a score to each tracker
   - Score: the tracker performs best and second best over all test sequences subject to the MAD method

3. **Ranking**: calculate the mean values of the data and apply the MAD to assign a rank to each tracker

# Scoring

- A tracker scores best (or second best) when it achieves best (or second best) average performance among all trackers **for a** test sequence
- Scoring does not only select numerically first and second best measures, but accounts for groups of best and second best measures
  - **for each** test sequence, define a MAD deviation threshold which evaluates a set's close affiliation to either a best score or a second best score

$$d_q = Median(|q_{i,j,l} - Median(\{q_{i,j,l}\})|)$$



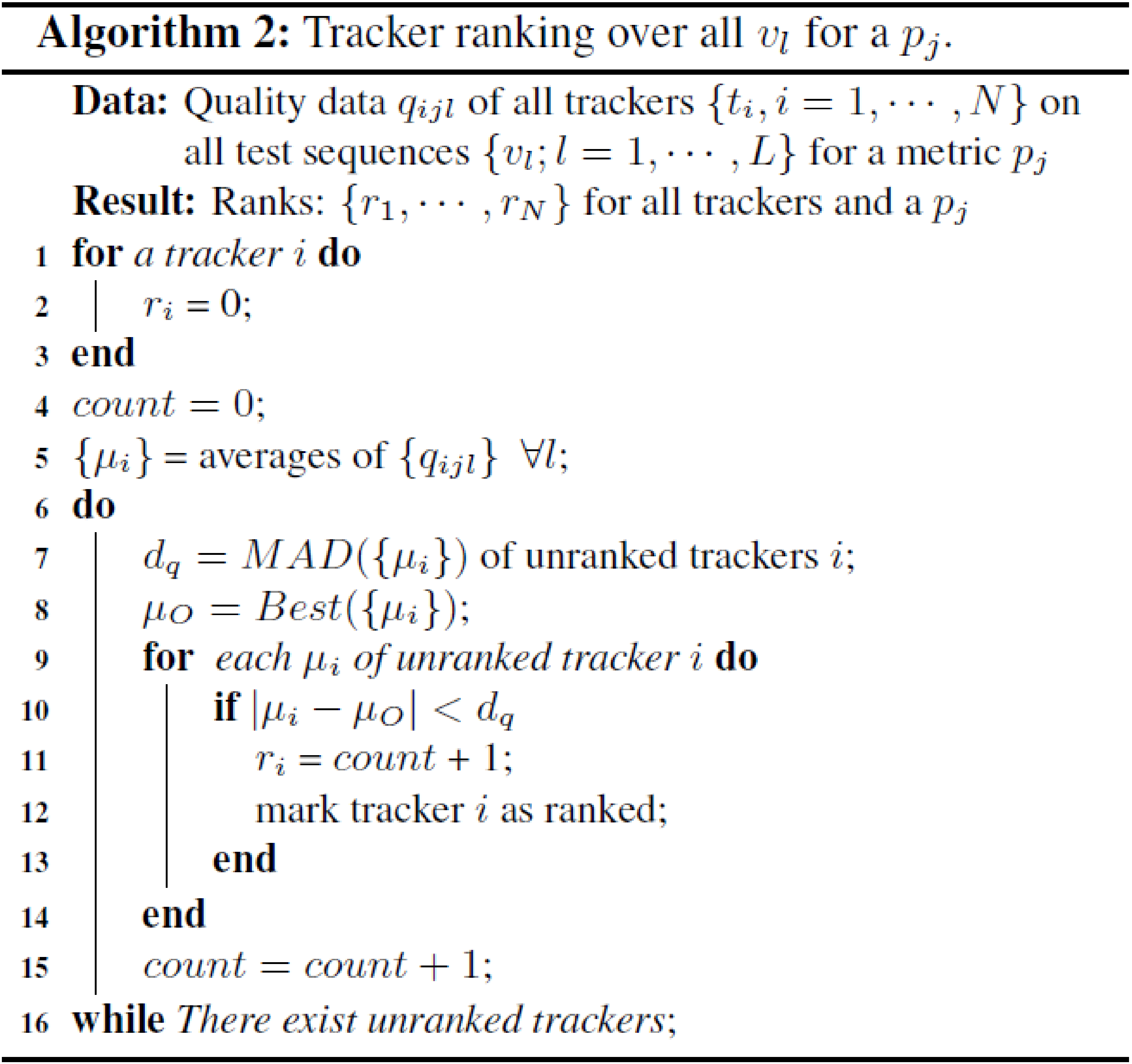
**Algorithm 1:** Scoring of trackers over all $\{v_l\}$ for a $p_j$.

```
Data: Quality data $q_{ijl}$ of all trackers $\{t_i, i = 1, \cdots, N\}$ on
      all test sequences $\{v_l; l = 1, \cdots, L\}$ for a metric $p_j$
Result: Scores $\{s_i, \cdots, s_N\}$ for all $\{t_i\}$ and a $p_j$
for a tracker $i$ do
    $s_i = 0$;
end
for a test sequence $l$ do
    $d_q = MAD\{q_{ijl}\}$;
    $O = Best\{q_{ijl}\}$;
    for a tracker $i$ do
        if $|q_{ijl} - O| < d_q$
            B = 1;
        else
            B = 0;
        end;
        $s_i = s_i$ + B;
    end
end
$\{s_i\} = \{s_i\}/L$;
```

# Sequence-pooled Ranking

Given the quality data of all sequences and all trackers for a metric

1. Calculate the mean quality (score) for each tracker by averaging over all test sequences
2. Mark all trackers as unranked and each tracker keeps contributing to the ranking process until it is assigned a rank
3. Find the best mean value among all unranked trackers
4. Assign a first rank to any tracker that has a mean quality closest to that best within the MAD in the first round
   - Mark that tracker as ranked
5. Repeat the same process but only for the rest of the unranked trackers



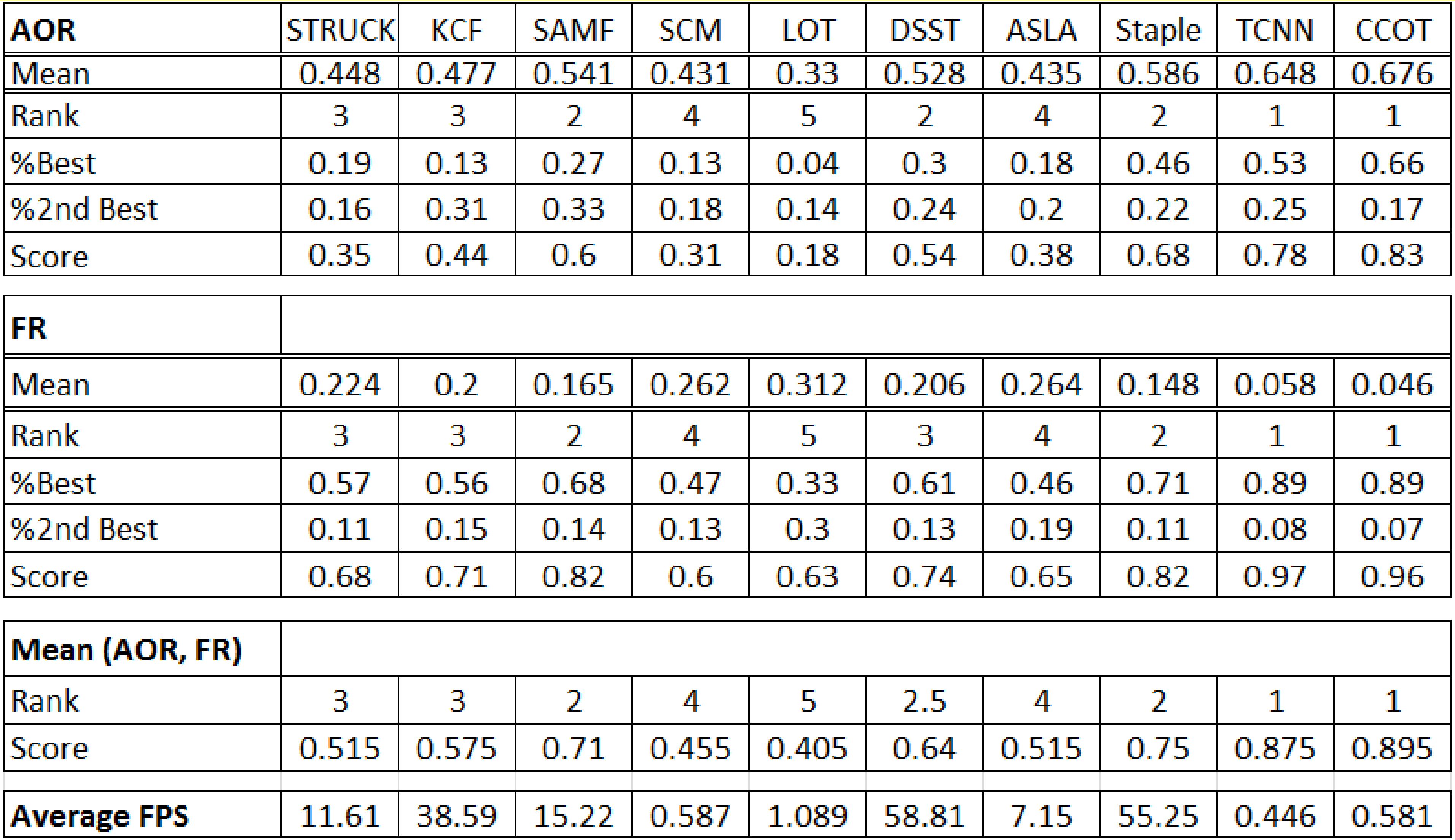
**Algorithm 2:** Tracker ranking over all $v_l$ for a $p_j$.

```
Data: Quality data $q_{ijl}$ of all trackers $\{t_i, i = 1, \cdots, N\}$ on
      all test sequences $\{v_l; l = 1, \cdots, L\}$ for a metric $p_j$
Result: Ranks: $\{r_1, \cdots, r_N\}$ for all trackers and a $p_j$
for a tracker $i$ do
    $r_i = 0$;
end
$count = 0$;
$\{\mu_i\}$ = averages of $\{q_{ijl}\}$ $\forall l$;
do
    $d_q = MAD(\{\mu_i\})$ of unranked trackers $i$;
    $\mu_O = Best(\{\mu_i\})$;
    for each $\mu_i$ of unranked tracker $i$ do
        if $|\mu_i - \mu_O| < d_q$
            $r_i = count + 1$;
            mark tracker $i$ as ranked;
        end
    end
    $count = count + 1$;
while There exist unranked trackers;
```

# Results: Set-up

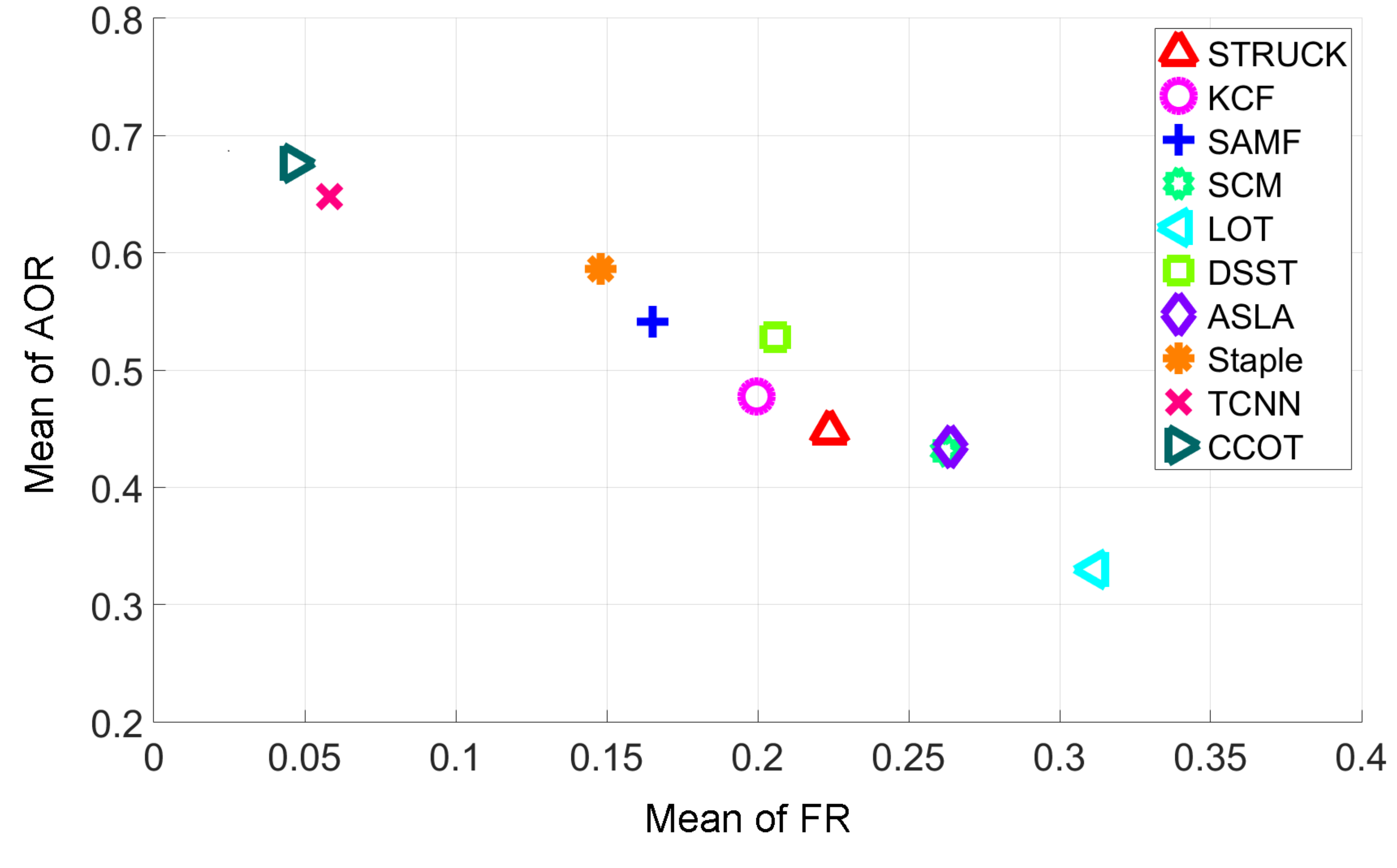
- ✓ We tested 10 trackers:
  - – some have similar quality
  - – others have fully different performance
- ✓ We used 100 video sequences from OTB benchmark
  - – include 11 tracking challenges illumination, occlusion, etc.
- ✓ To minimize the effect of randomness in the algorithms, we ran each tracker 5 times on each video sequence
- ✓ We used uncorrelated metrics
  - – accuracy (overlap ratio AOR) and
  - – robustness (failure rate FR)

# Results

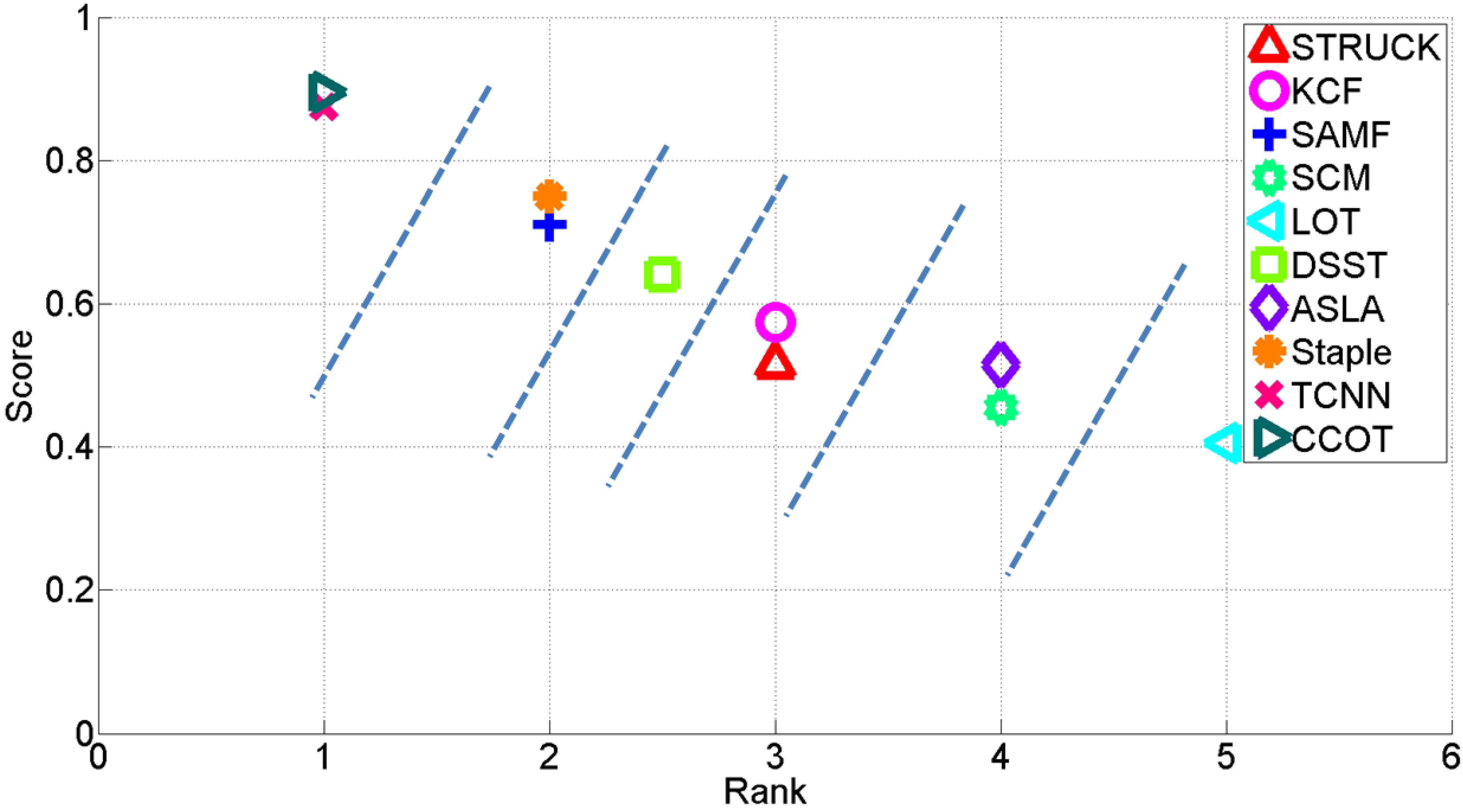


| **AOR** | STRUCK | KCF | SAMF | SCM | LOT | DSST | ASLA | Staple | TCNN | CCOT |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 0.448 | 0.477 | 0.541 | 0.431 | 0.33 | 0.528 | 0.435 | 0.586 | 0.648 | 0.676 |
| Rank | 3 | 3 | 2 | 4 | 5 | 2 | 4 | 2 | 1 | 1 |
| %Best | 0.19 | 0.13 | 0.27 | 0.13 | 0.04 | 0.3 | 0.18 | 0.46 | 0.53 | 0.66 |
| %2nd Best | 0.16 | 0.31 | 0.33 | 0.18 | 0.14 | 0.24 | 0.2 | 0.22 | 0.25 | 0.17 |
| Score | 0.35 | 0.44 | 0.6 | 0.31 | 0.18 | 0.54 | 0.38 | 0.68 | 0.78 | 0.83 |
| | | | | | | | | | | |
| **FR** | | | | | | | | | | |
| Mean | 0.224 | 0.2 | 0.165 | 0.262 | 0.312 | 0.206 | 0.264 | 0.148 | 0.058 | 0.046 |
| Rank | 3 | 3 | 2 | 4 | 5 | 3 | 4 | 2 | 1 | 1 |
| %Best | 0.57 | 0.56 | 0.68 | 0.47 | 0.33 | 0.61 | 0.46 | 0.71 | 0.89 | 0.89 |
| %2nd Best | 0.11 | 0.15 | 0.14 | 0.13 | 0.3 | 0.13 | 0.19 | 0.11 | 0.08 | 0.07 |
| Score | 0.68 | 0.71 | 0.82 | 0.6 | 0.63 | 0.74 | 0.65 | 0.82 | 0.97 | 0.96 |
| | | | | | | | | | | |
| **Mean (AOR, FR)** | | | | | | | | | | |
| Rank | 3 | 3 | 2 | 4 | 5 | 2.5 | 4 | 2 | 1 | 1 |
| Score | 0.515 | 0.575 | 0.71 | 0.455 | 0.405 | 0.64 | 0.515 | 0.75 | 0.875 | 0.895 |
| | | | | | | | | | | |
| **Average FPS** | 11.61 | 38.59 | 15.22 | 0.587 | 1.089 | 58.81 | 7.15 | 55.25 | 0.446 | 0.581 |

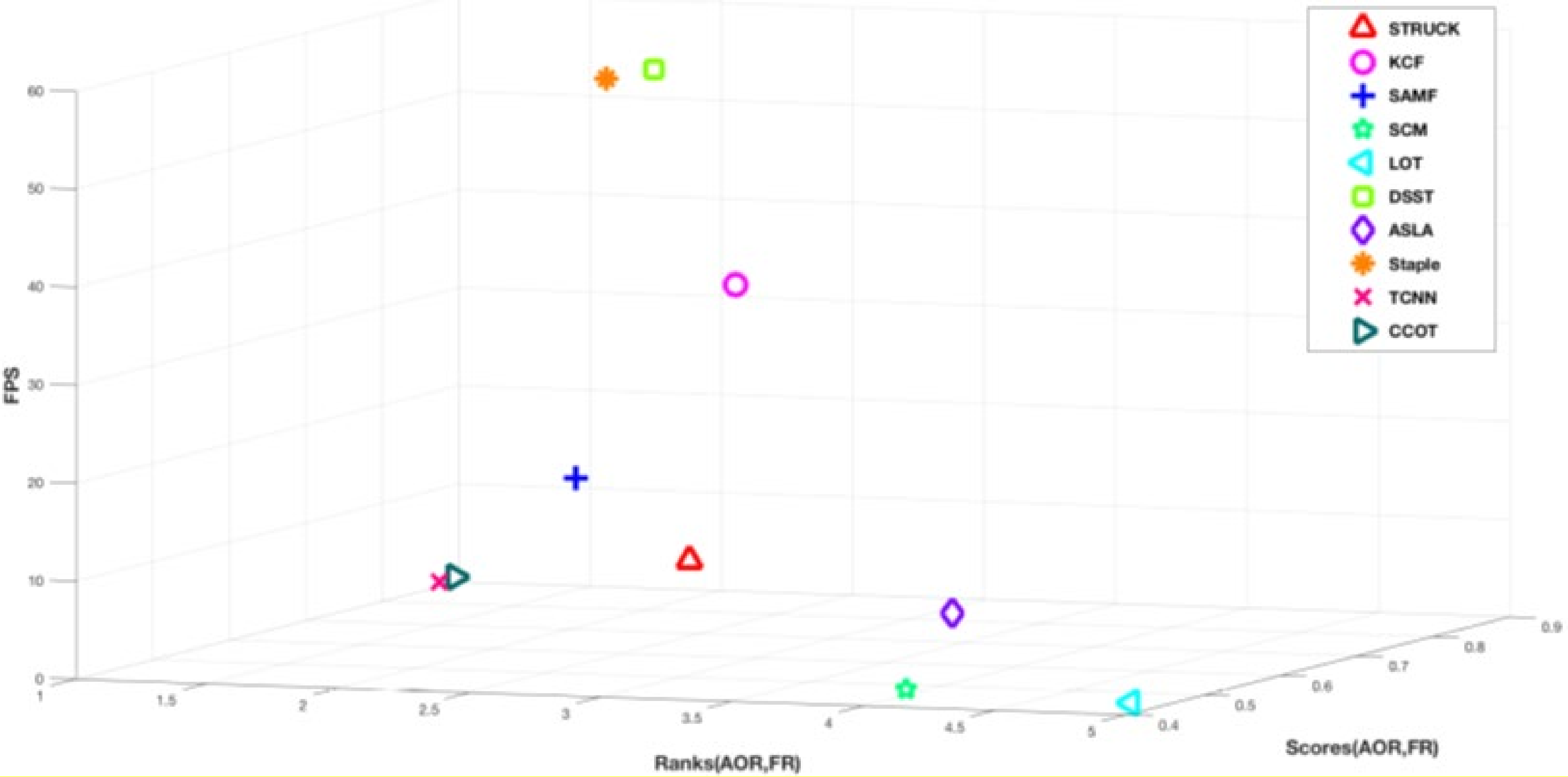
# Results:
# Averaging: trackers are **not** well grouped



# Results:
# Rank-score: trackers are better grouped



# Results: Score, Rank, and FPS



# Conclusion

- Evaluation using uncorrelated performance measures coupled with the MAD to account for both tracker outliers and similarities
  - We tested *interquartile range* and *median maximal distance*, but their scores did not well discriminate trackers as the MAD did
- Different than related work, our approach
  - evaluates a tracker's performance relative to the performance of **all** tested trackers
  - accounts for data dispersion and better categorizes trackers than widely-used averaging
- Our approach performs more discriminative ranking as the number of trackers to evaluate increases

# References

[1] M. Kristan et al., “The visual object tracking VOT2013 challenge results,” in Proc. IEEE Int. Conf. Computer Vision Workshops, Sydney, Dec. 2013, pp. 98–111.

[2] M. Kristan et al., “The visual object tracking VOT2014 challenge results,” in European Conf. Computer Vision Workshops, Zurich, Switzerland, Sept. 2014, pp. 191–217, Springer.

[3] M. Kristan et al., “The visual object tracking VOT2015 challenge results,” in Proc. IEEE Int. Conf. Computer Vision Workshops, Santiago, Dec. 2015, pp. 564–586.