



TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Multi-Camera Tracking for Employee Behavior Monitoring

Trần Quốc Lập – 20194443
August 15, 2023

ONE LOVE. ONE FUTURE

Introduction and Objective

Proposed Method and Evaluation

Application System Development

Conclusion and Future Work

INTRODUCTION & OBJECTIVE

Multi-camera tracking (MCT) aims to track people **across** cameras (keep identity).

MCT has **various applications**. In employee management, MCT can track and **monitor** employee behaviors.

Example: Amazon's worker surveillance

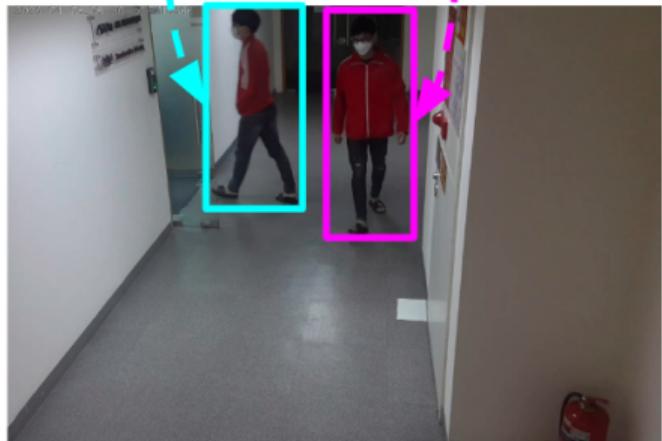
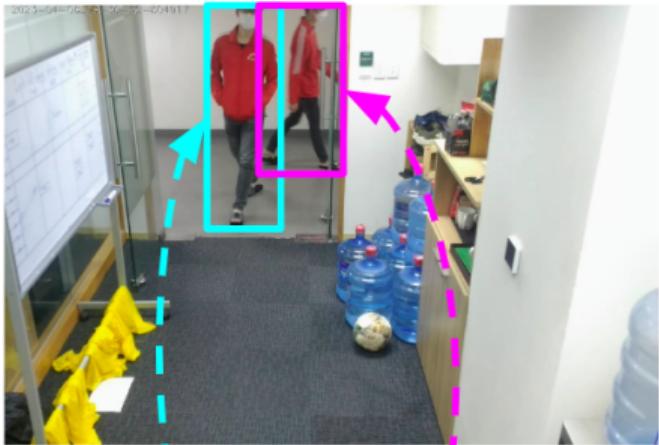


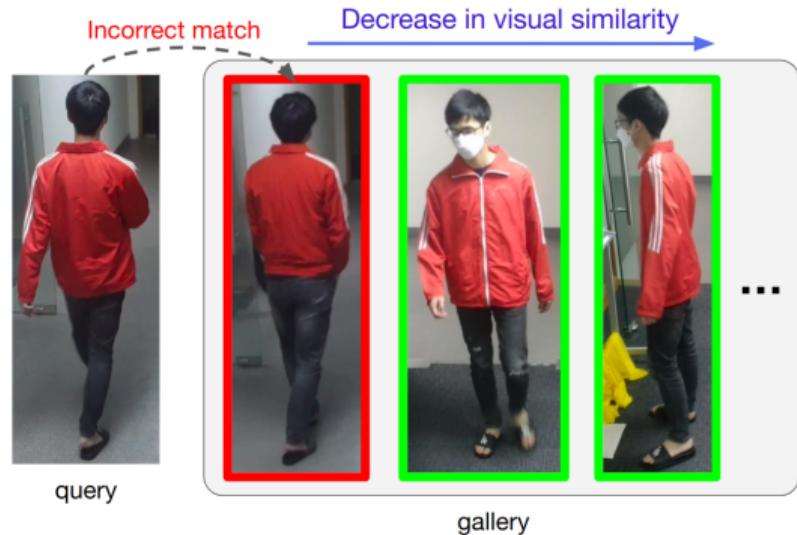


Figure 1: General pipeline for multi-camera tracking.

1. Person detection: Detect person in single frame (e.g. YOLOv3, YOLOv7).
2. Single camera tracking: Match boxes to form tracks (e.g. SORT, ByteTrack).
3. Multi-camera tracking: Match tracks across cameras.

To match tracks (keep ID) across cameras, popular methods compare people **visual appearance**, notably Re-ID.

Issue of Re-ID: easily matches **incorrectly** if people (e.g. employees) wear **uniforms**.



Thesis work: develop a MCT solution that can work well when people have **similar appearance**.

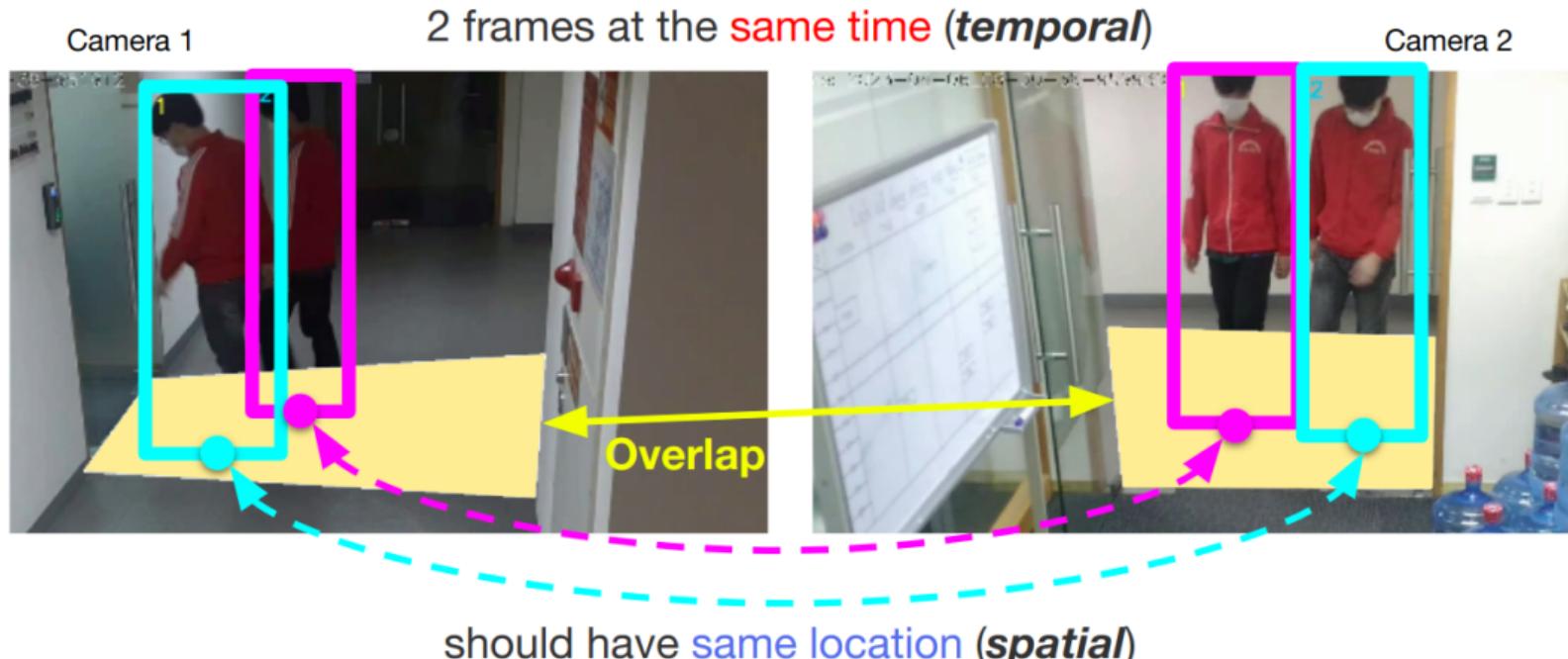


Figure 2: Proposed: MCT based on **Spatio-Temporal Association (STA)** with **overlapping** area.

Scope:

- ◊ Small overlap ($4m^2$) & synced cameras.
- ◊ Track employees wearing uniforms.
- ◊ Maximum 4 people.
- ◊ 3 cameras.
⇒ no public dataset
- ⇒ collect 36 videos, split into 3 sets that vary in moving direction and number of people gathering at the overlapping area: Easy (2 people), Medium (3 people), Hard (4 people).



Figure 3: Camera setup and FOV

PROPOSED METHOD & EVALUATION

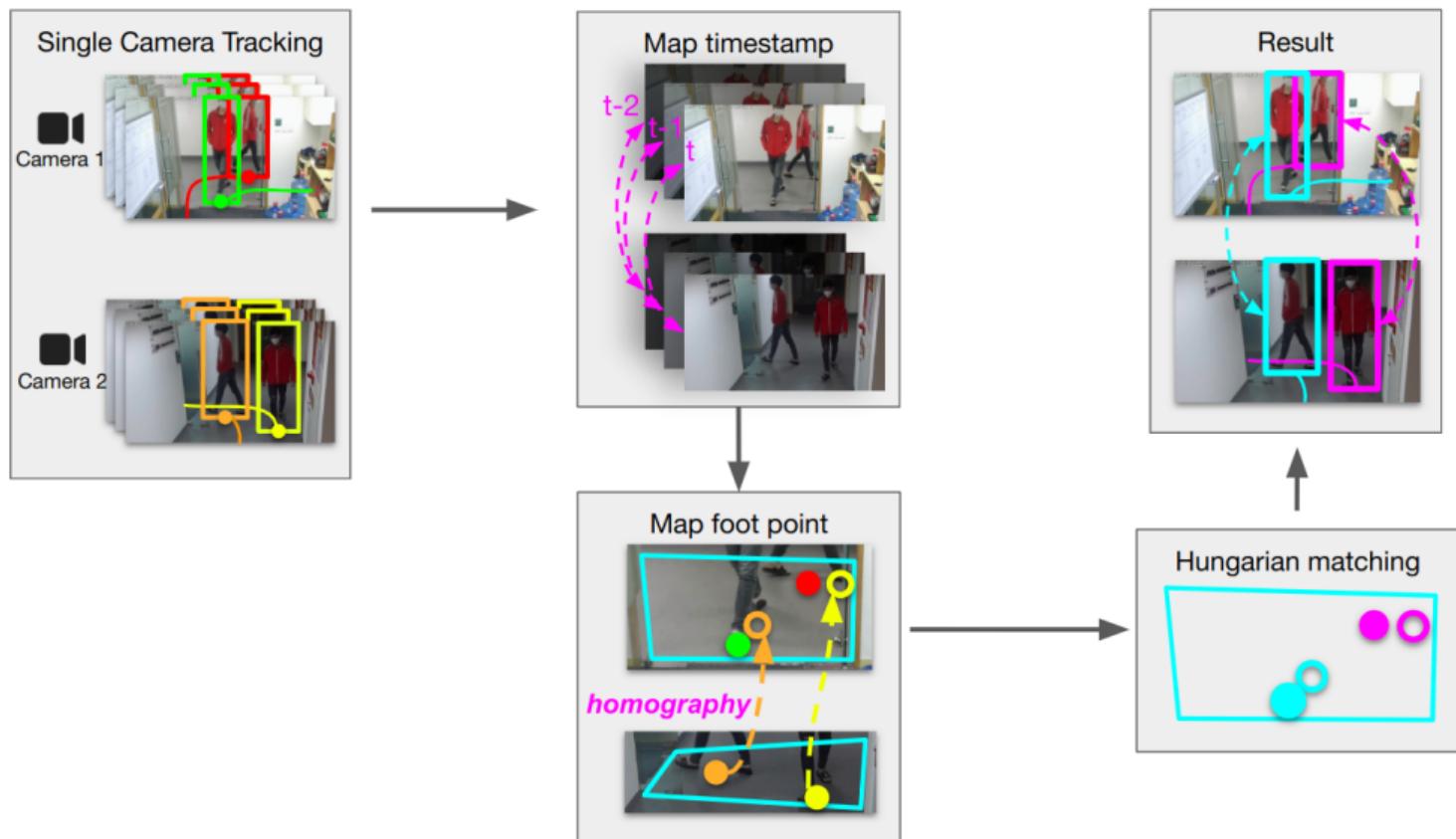


Figure 4: Overview of the proposed STA method.

Evaluation metric:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

| Video set | #TP | #FP | #FN | F1 |
|-----------|------|-----|-----|-------|
| Easy | 511 | 7 | 7 | 0.986 |
| Medium | 662 | 46 | 22 | 0.951 |
| Hard | 966 | 144 | 41 | 0.913 |
| Total | 2139 | 197 | 70 | 0.941 |

Table 1: **TP**: correctly matched pair. **FP**: incorrectly matched pair. **FN**: missed pair.

- ◊ **Promising** but decreases with complexity.
- ◊ **#FP >> #FN**, especially with **hard** and **medium** sets.



Figure 5: With Re-ID, all individuals wearing uniform are considered the same person.

| Video set | Re-ID | STA |
|-----------|-------------------------|-----------------------|
| Easy | 0.5 (32 - 64 - 0) | 1.0 (32 - 0 - 0) |
| Medium | 0.348 (57 - 211 - 2) | 0.982 (57 - 0 - 2) |
| Hard | 0.380 (54 - 176 - 0) | 0.991 (53 - 0 - 1) |

Table 2: Re-ID vs. the proposed STA. Each cell format is $F1(\#TP, \#FP, \#FN)$ evaluated at track-level.

The proposed baseline method outperformed Re-ID in 3 datasets where people have similar appearance.

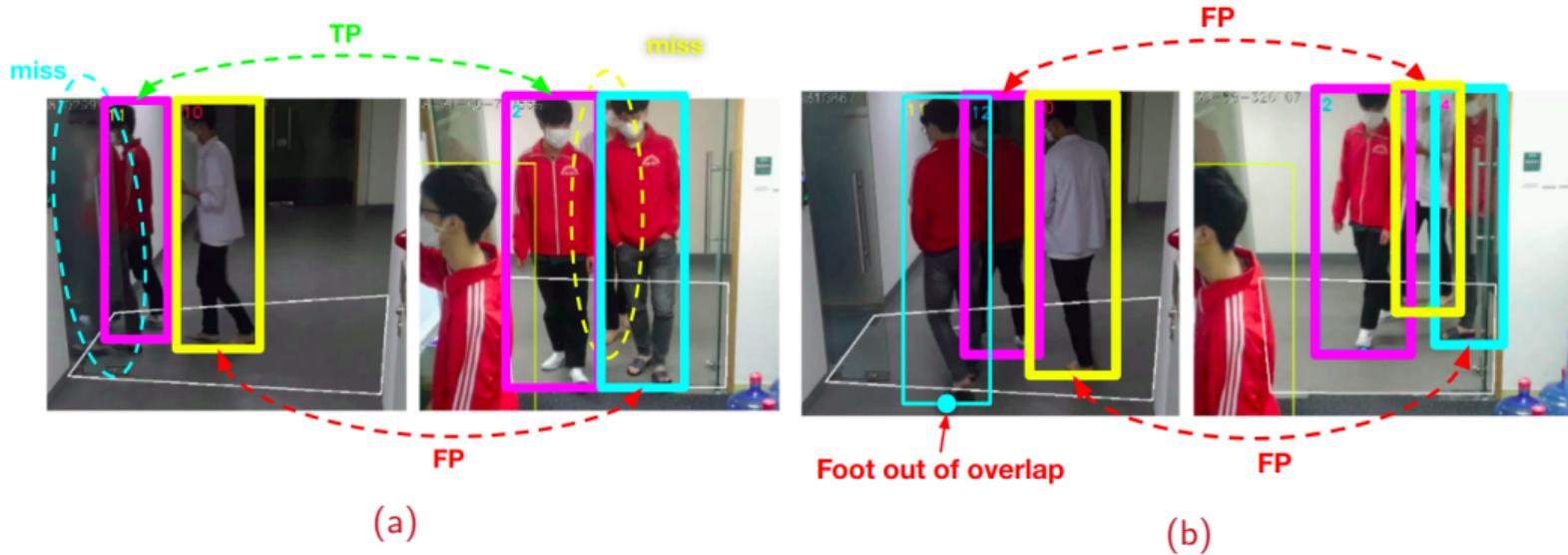


Figure 6: Two main causes of FP and FN:

- a) missing detection.
- b) incorrect foot interpolation.

Assumption: FP due to missing detections have larger spatial distance than TP.
⇒ treated those FP as outliers in the distance distribution.

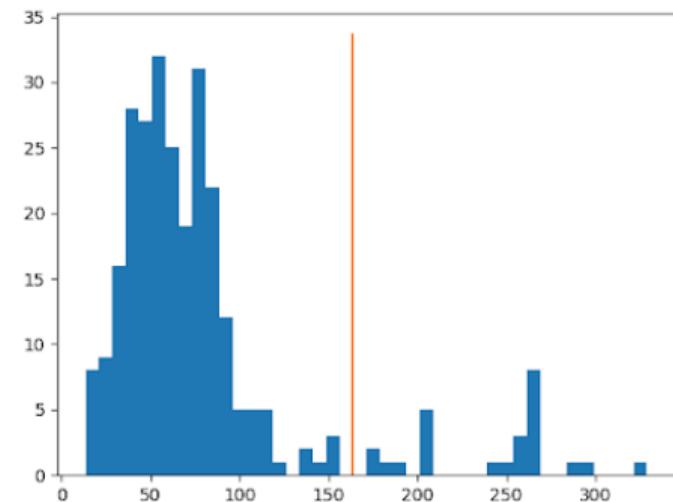
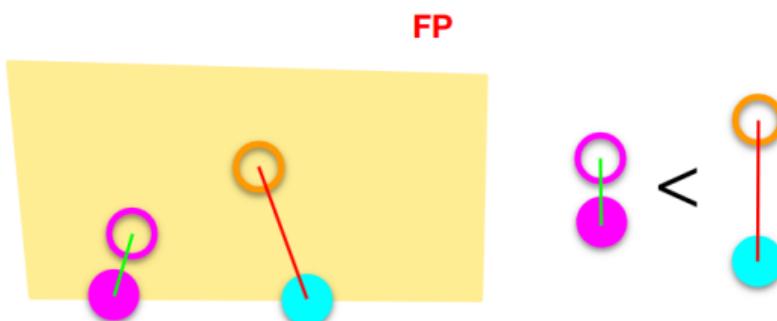
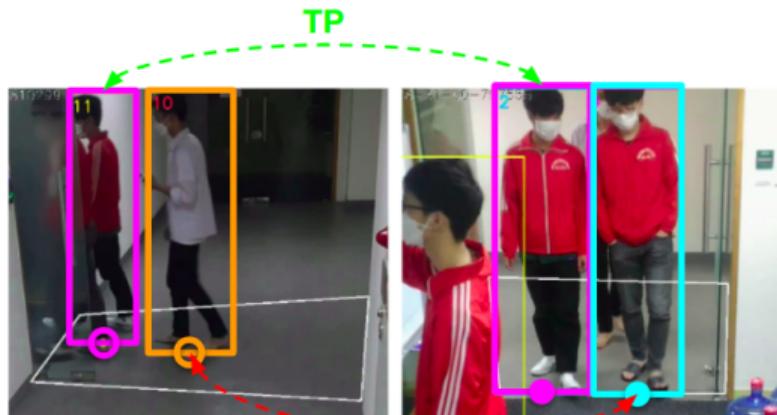


Figure 7: Distance distribution of matched pairs. x-axis: distance. y-axis : #matched pairs. Seam: upper bound by $IQR(25, 75)$.

Assumption: Sometimes a mismatch (FP and FN) can be solved by taking matches in previous and next timestamps into account
⇒ average distance over neighboring frames as an input cost for the Hungarian.

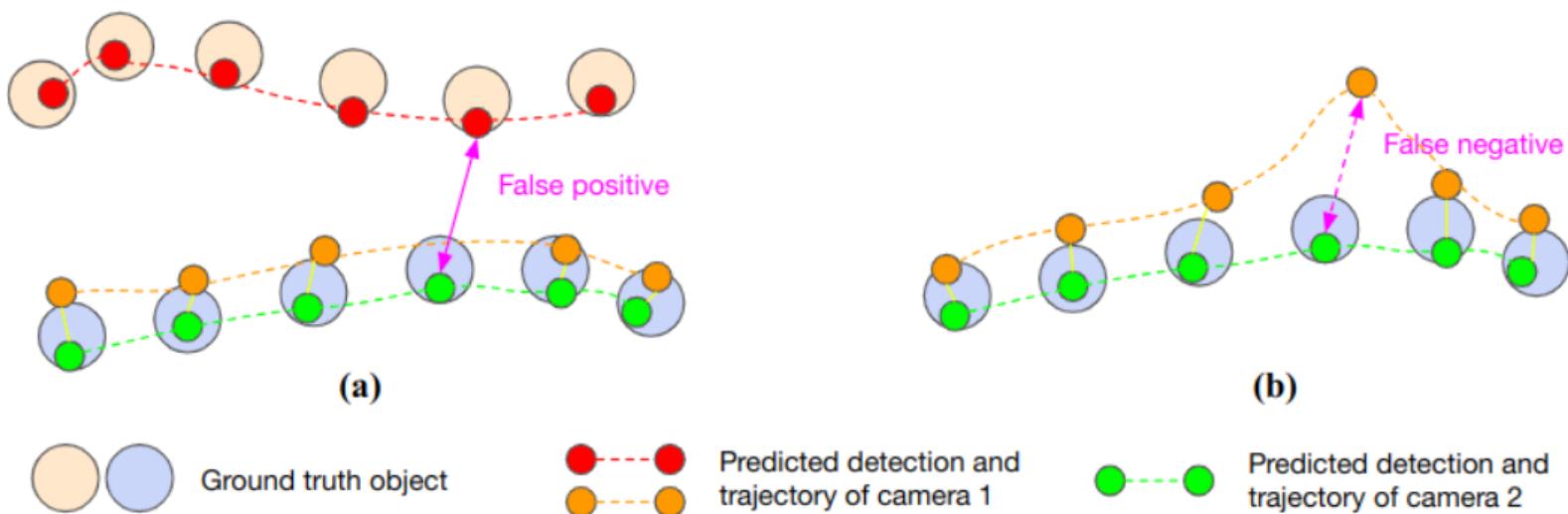


Figure 8: window-based mapping may reduce a) FP and b) FN.

Issue with rectangular bounding box: Even if it fits the body well, the foot point (midpoint of bottom edge) may not be accurate. E.g: when legs apart.

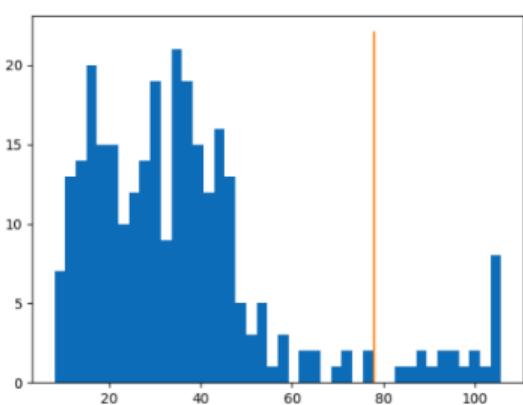
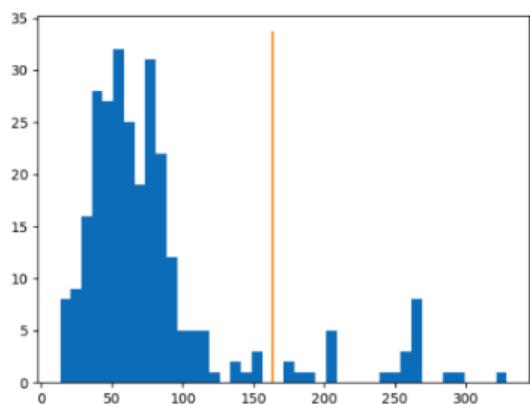


Figure 9: Distance distribution. **a)** using box. **b)** using pose.

Experiments prove that using pose:

- ◊ gives more accurate foot points.
- ◊ must combine with FP filtering, or else it will be worse than using box.

| Set | Baseline | FP filtering (1) | Window-based (2) | (1) + (2) [*] | (1) [*] + (2) ⁺ + Pose |
|--------|------------------------|---|---|---|---|
| Easy | 0.986 (511,7,7) | 0.983 (507,7,11) | 0.994 (515,3,3) | 0.986 (509,5,9) | 0.985 (657,3,17) |
| Medium | 0.951 (662,46,22) | 0.958 (658,32,26) | 0.955 (665,43,19) | 0.969 (662,20,22) | 0.991 (886,8,8) |
| Hard | 0.913 (966,144,41) | 0.927 (959,103,48) | 0.921 (975,135,32) | 0.938 (963,83,44) | 0.960 (1179,36,63) |
| Total | 0.941 (2139,197,70) | 0.949 (\uparrow 0.8%) (2124,142,85) | 0.948 (\uparrow 0.7%) (2155,181,54) | 0.959 (\uparrow 1.9%) (2134,108,75) | 0.976 (\uparrow 3.7%) (2722,47,88) |

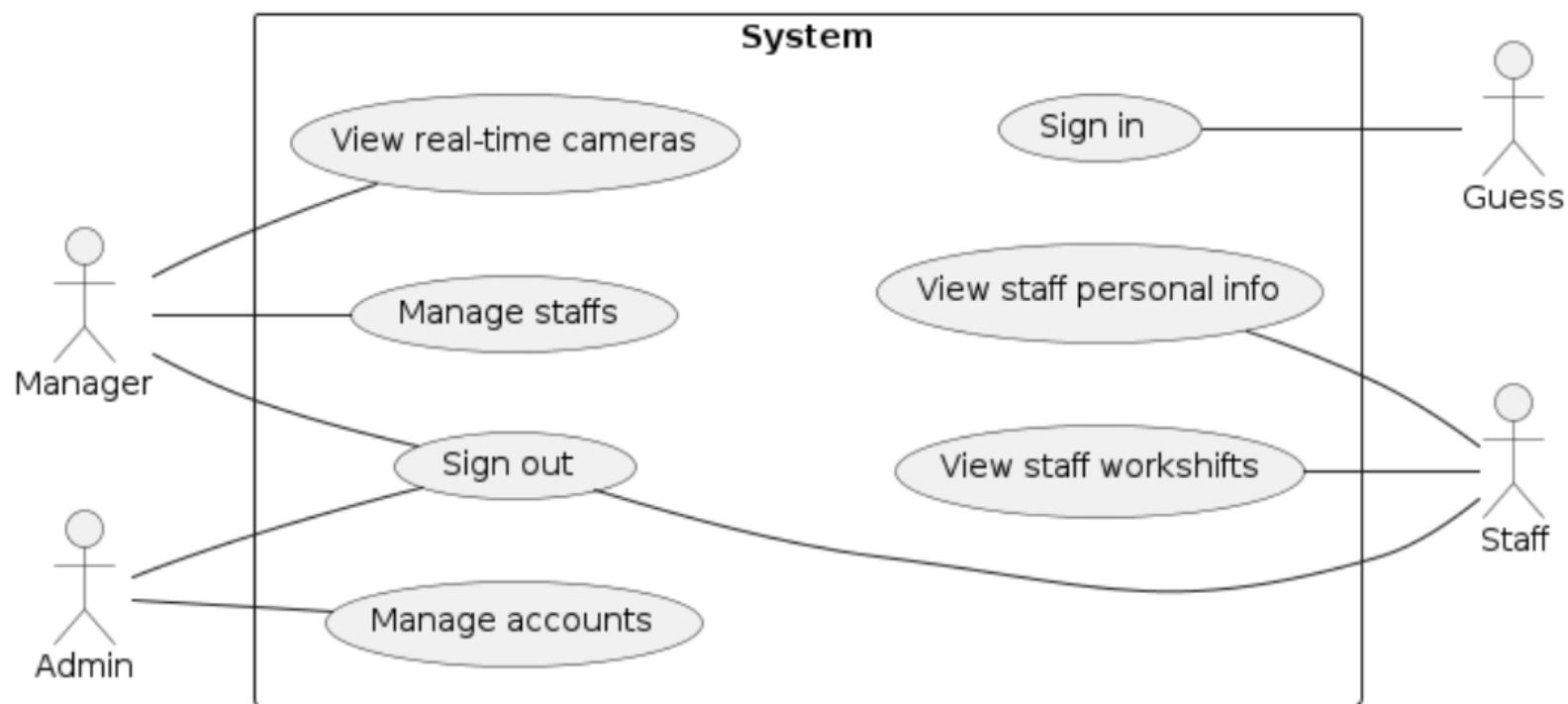
Table 3: Each cell format is $F1(\#TP, \#FP, \#FN)$ evaluated at frame-level. (1): $IQR(20, 80)$. (2): window-size = 15. (1)^{*}: $IQR(25, 75)$. (2)^{*}: window-size = 11. (2)⁺: window-size = 7.

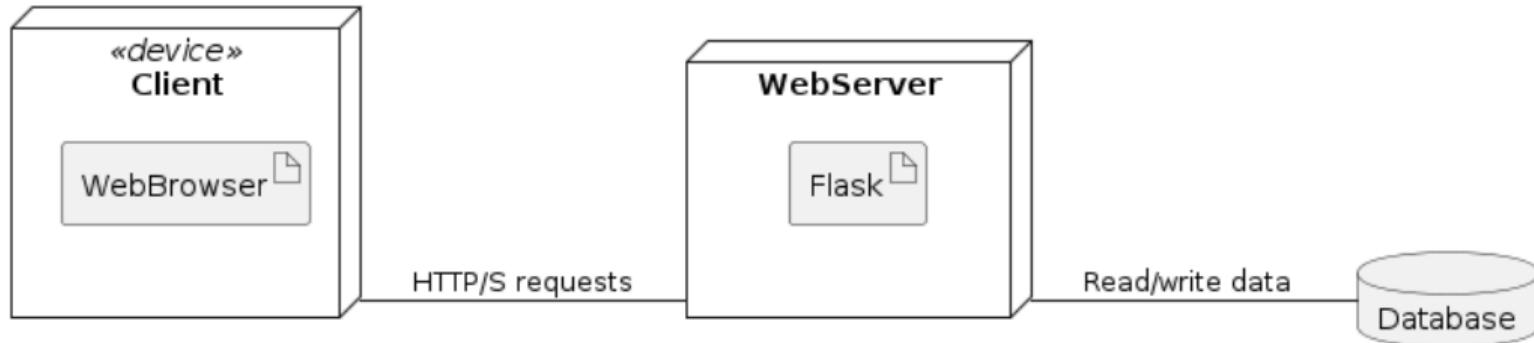
In general, all the extensions have **greater and positive impact** on **complex** cases than on **easy** cases.

- ◊ The proposed Spatio-Temporal Association outperformed visual-based Re-ID in 3 datasets where people have similar appearance.
- ◊ Enhancements:
 - address missing detection: FP filtering and window-based mapping ($\uparrow 1.9\%$).
 - address inaccurate footpoint: Pose estimation but need FP filtering to be applied ($\uparrow 3.7\%$)
- ◊ Those extensions have more impact on complex cases.

APPLICATION SYSTEM DEVELOPMENT

Application: Develop a software system to showcase the applicability and demand for the proposed solution.





Frameworks used in research and developments:

- ◊ Machine learning and Computer Vision:
 - NumPy, Scikit-Learn, SciPy.
 - PyTorch, OpenCV.
- ◊ Web development: Flask, SQLite.

Employee Management

Camera view

user <5> signed in
Nguyen Van A

Employee Management

Productivity report: Nguyen Van A

| Day | Date | Day shift | Arrival | Staying time |
|-----------|------------|-----------|-------------------------------|--------------|
| Wednesday | 12/04/2023 | morning | 08:30:20 (a few seconds late) | 10% |
| Tuesday | 11/04/2023 | morning | 08:30:11 (a few seconds late) | 88.9% |
| Monday | 10/04/2023 | morning | 08:30:19 (a few seconds late) | 56.4% |
| Friday | 07/04/2023 | morning | 08:30:02 | 61.4% |

Employee Management

Messages

9 minutes ago
Tran Van B was absent from work area since 08:31:04

9 minutes ago
Tran Van B is back to work area at 08:30:58

9 minutes ago
Tran Van B was absent from work area since 08:30:34

9 minutes ago
Pham Van C is back to work area at 08:30:38

9 minutes ago
Pham Van C was absent from work area since 08:30:29

9 minutes ago
Tran Van B is back to work area at 08:30:33

9 minutes ago
Pham Van C arrived late at 08:30:29

Contributions:

- ◊ A Spatio-Temporal Association for MCT that outperformed Re-ID when people have similar appearance.
- ◊ Extensions that incrementally improved the proposed method.
- ◊ A software system that showcased the applicability of the proposed method.

Future works:

- ◊ Investigate the precision of the proposed method on different datasets.
- ◊ Determine the parameters of the extensions automatically that are not specific to any dataset.
- ◊ Develop a solution to combine Re-ID and STA in general MCT problems.