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

Multi-Camera Tracking for Employee Behavior Monitoring

Trần Quốc Lập – 20194443

July 20, 2023

ONE LOVE. ONE FUTURE

INTRODUCTION & MOTIVATION

Multi-camera tracking (MCT) are widely used today, e.g.:

- ◇ in sales, to track customers' behaviors
- ◇ in sport, to track players' performance

In **employee management**, employees can be tracked to:

- ◇ estimate productivity
- ◇ detect irregular behaviors
- ◇ assign work equally

Example: Amazon's worker surveillance

Figure 1: An example of multi-camera tracking in retails to track customers' movements and interactions with products.

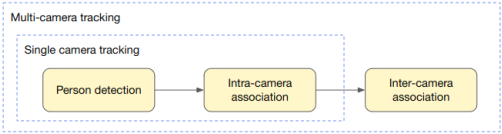


Figure 2: General pipeline for multi-camera tracking.

- ◇ Person detection: detect persons in single frames. E.g: R-CNN, YOLO.
  Issue: missing detection due to occlusion
- ◇ Intra-camera association: associate detections within each camera. E.g: SORT, ByteTrack, DeepSORT.
  Issue: ID switch due to people move close optically or missing detection
- ◇ Inter-camera association: associate tracks across cameras. Popular methods rely on visual similarity to re-identify (Re-ID) people.

Issue of Re-ID:

- ◇ large distance for same person, small distance for different people due to changes in viewpoints, poses, etc.
- ◇ need to perform verification rather than a retrieval, because the query person may not appear in the gallery.

In employee management, employees look very similar due to uniforms.
$\Longrightarrow$ visual-based Re-ID easily matches people incorrectly.

Figure 3: People with similar looks that fool Re-ID. Left to right: decrease in similarity. Blue border: same person. Red border: different people.

Thesis work: develop a solution for MCT, which can work well when people have similar appearance.

Figure 4: Multi-camera tracking with overlapping field of views.

Objectives:

- ◇ **Research**: Propose a **Spatio-Temporal Association** (STA) solution (i.e use space and time info) to replace Re-ID.
- ◇ **Technology**: Master ML, DL, CV frameworks and experimental design.
- ◇ **Application**: Develop a software system to showcase the applicability and demand for the proposed solution.

Scope:

- ◇ Cameras have **small** overlapping FOV and synchronized time.
- ◇ Track employees wearing uniforms in a store.
- ◇ Employee behavior:
  - ○ Stay in designated work position.
  - ○ Move between multiple cameras at a normal pace.
- ◇ Test scenario comprises
  - ○ Maximum 4 people in the overlap simultaneously.
  - ○ 3 cameras covering all areas of the store.

PROPOSED METHOD & EVALUATION

No public dataset satisfies the defined scope.

$\Longrightarrow$ need to collect manually 36 videos (12 for each camera):

- ◇ Easy set: 2 people
- ◇ Medium set: 3 people
- ◇ Hard set: 4 people

Figure 5: Camera setup and FOV

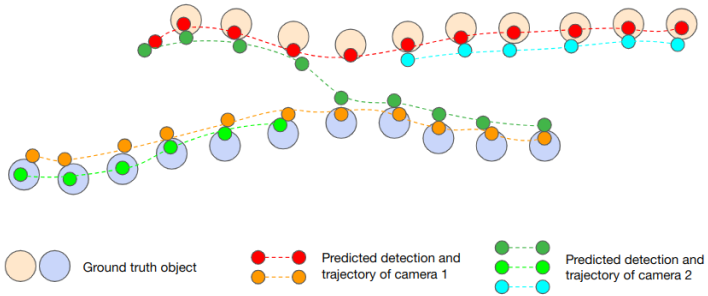


Figure 6: A simple example of confusing track-level matching due to ID switch.

Previous studies on MCT and STA focused on track-level matching:

- ◇ ignore ID switch which makes mapping very complicated.
- ◇ unreliable evaluation results at frame level.

$\Longrightarrow$ the proposed method involves frame-level matching.

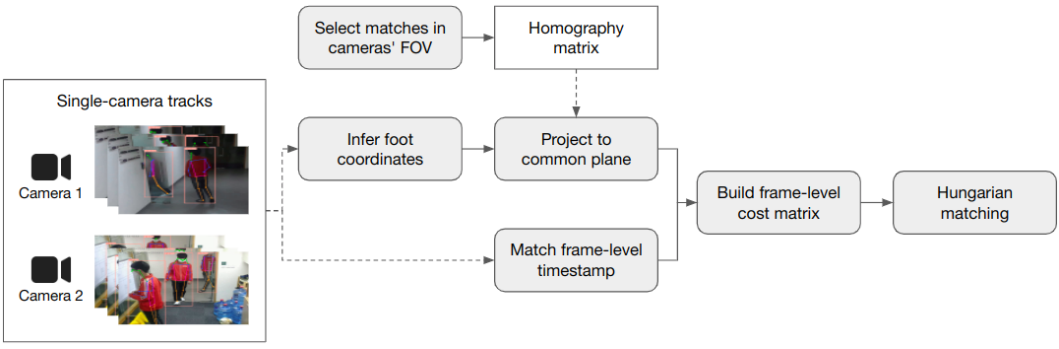


Figure 7: Overview of the proposed method STA.

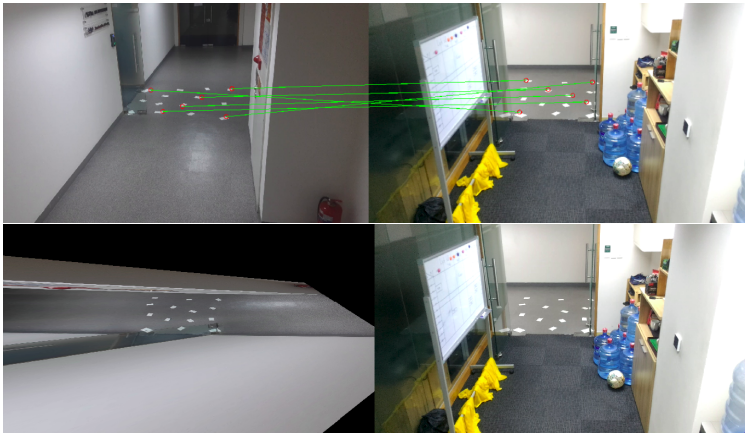
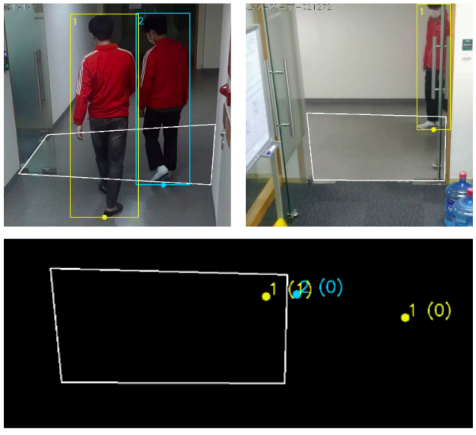
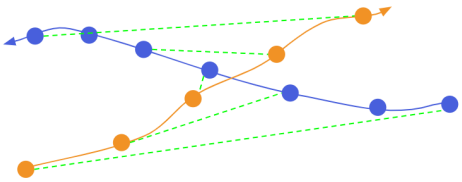


Figure 8: Manually selecting corresponding points in the FOV of a camera pair to build a homography matrix.



(a)



(b)

Figure 9: **a)** foot point interpolation and projection. **b)** Frame-level timestamp matching between 2 tracks.

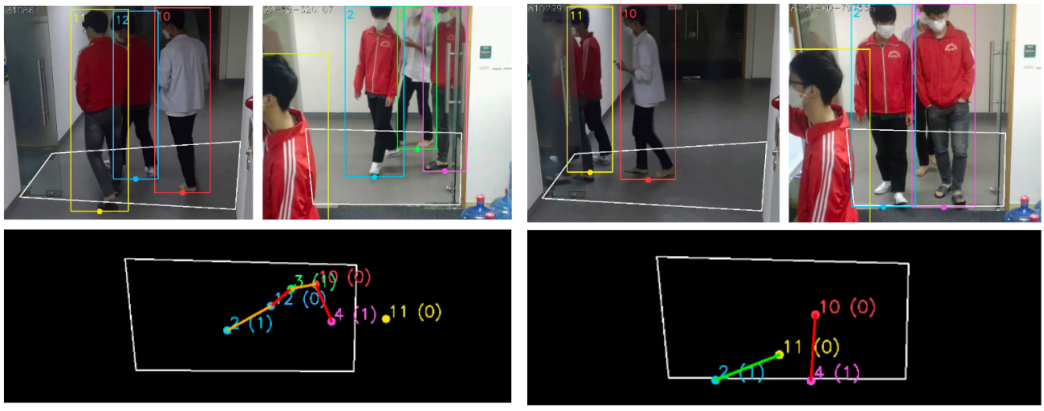
Evaluation metric:

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

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: Performance of the proposed baseline method on the recorded videos.

- ◇ Promising but decreases with complexity.
- ◇ More FP than FN, especially with hard and medium sets.



(a) (b)

Figure 10: Cause of FP and FN due to **a)** incorrect foot point interpolation. **b)** missing detection. Green line indicates TP. Red line indicates FP. Yellow line indicates FN.

Assumption: False positives due to missing detections have larger spatial distance than true positives.

$\Longrightarrow$ treated those FP as outliers in the distance distribution.

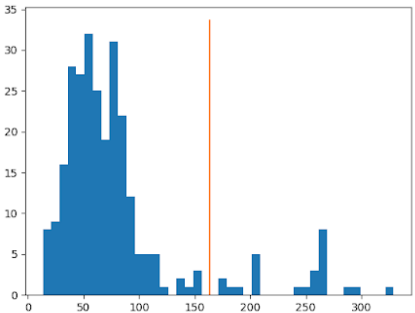


Figure 11: Distance distribution of matched pairs. x-axis is the spatial distance. y-axis is the number of matched pairs. The seam is the upper bound by IQR(25, 75).

Video set	Baseline	**IQR (20, 80)**
Easy	0.986 (511,7,7)	0.983 (507,7,11)
Medium	0.951 (662,46,22)	0.958 (658,32,26)
Hard	0.913 (966,144,41)	0.927 (959,103,48)
Total	0.941 (2139,197,70)	0.949 (2124,142,85)

Table 2: Baseline method vs. FP filtering. Each cell format is F1(#TP, #FP, #FN).

1. Easy set: no significant change.
2. Medium set: #FP $\downarrow$ and #FN $\uparrow$ slightly.
3. Hard set: #FP $\downarrow$ more than Easy and Medium, #FN $\uparrow$ slightly.

Assumption: Sometimes a mismatch (FP and FN) can be solved by taking matches in previous and next timestamps into account

$\Longrightarrow$ average distance over neighboring frames as an input cost for the Hungarian.

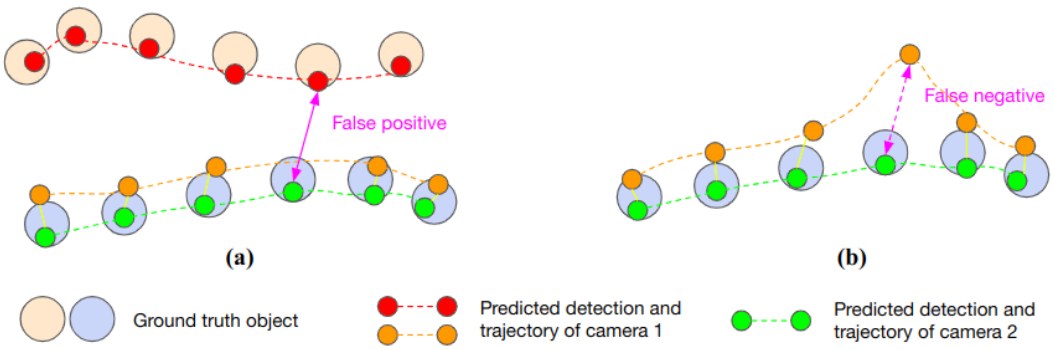


Figure 12: window-based mapping may reduce **a)** FP and **b)** FN.

Video set	Baseline	**Size = 15**
Easy	0.986 (511,7,7)	0.994 (515,3,3)
Medium	0.951 (662,46,22)	0.955 (665,43,19)
Hard	0.913 (966,144,41)	0.921 (975,135,32)
Total	0.941 (2139,197,70)	0.948 (2155,181,54)

Table 3: Baseline method vs. window-based mapping.

- ◇ slight ↓ in both #FP and #FN in all 3 sets.
- ◇ more improvement on hard set.

Video set	Baseline	**IQR(20,80) + size = 11**
Easy	0.986 (511,7,7)	0.986 (509,5,9)
Medium	0.951 (662,46,22)	0.969 (662,20,22)
Hard	0.913 (966,144,41)	0.938 (963,83,44)
Total	0.941 (2139,197,70)	0.959 (2134,108,75)

Table 4: Baseline method vs. combination of FP filtering + window-based mapping.

Combining the 2 extensions produces more impressive results than standalone:

- ◇ Easy set: no significant change.
- ◇ Hard and Medium set: significant ↓ #FP, not much change in #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.

Expectations on using pose:

1. more accurate foot points than bounding box.
2. greater and positive impact on complex cases than on easy ones.

Video set	Baseline with box	**Baseline with pose**
Easy	0.986 (511,7,7)	0.985 (665,11,9)
Medium	0.951 (662,46,22)	0.950 (883,81,11)
Hard	0.913 (966,144,41)	0.928 (1200,145,42)
Total	0.941 (2139,197,70)	0.948 (2748,237,62)

Table 5: Baseline method using bounding box vs using pose estimation.

However, when examining the evaluation results on each individual video:

- ◇ 2/4 hard videos improve remarkably.
- ◇ 2/4 hard videos and 1/4 medium video drop with sharp increase in #FP.
- ◇ other videos show no significant change.

Set	ID	**Baseline with box**	Baseline with pose
Medium	10	0.914 (154,20,9)	0.889 (196,40,9)
Hard	11	0.875 (186,36,11)	0.859 (211,48,21)
	12	0.895 (218,42,9)	0.833 (252,84,17)

Table 6: Baseline method using bounding box vs. using pose estimation on individual video.

$\Longrightarrow$ contrary to the expectation

Video set	ID	Baseline with box	Baseline with pose	IQR(20, 80) size = 11 with box	**IQR(25, 75) size = 7 with pose**
Easy		0.986 (511,7,7)	0.985 (665,11,9)	0.986 (509,5,9)	0.985 (657,3,17)
Medium		0.951 (662,46,22)	0.950 (883,81,11)	0.969 (662,20,22)	0.991 (886,8,8)
Hard	11	0.875 (186,36,11)	0.859 (211,48,21)	0.898 (158,18,18)	0.894 (200,15,32)
	12	0.895 (218,42,9)	0.833 (252,84,17)	0.948 (219,16,8)	0.938 (253,17,16)
	all	0.913 (966,144,41)	0.928 (1200,145,42)	0.938 (963,83,44)	0.960 (1179,36,63)
Total		0.941 (2139,197,70)	0.948 (2748,237,62)	0.959 (2134,108,75)	0.976 (2722,47,88)

For the sake of comparison with Re-ID, track level matching is still needed, and was obtained by manually correcting ID switch.

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)

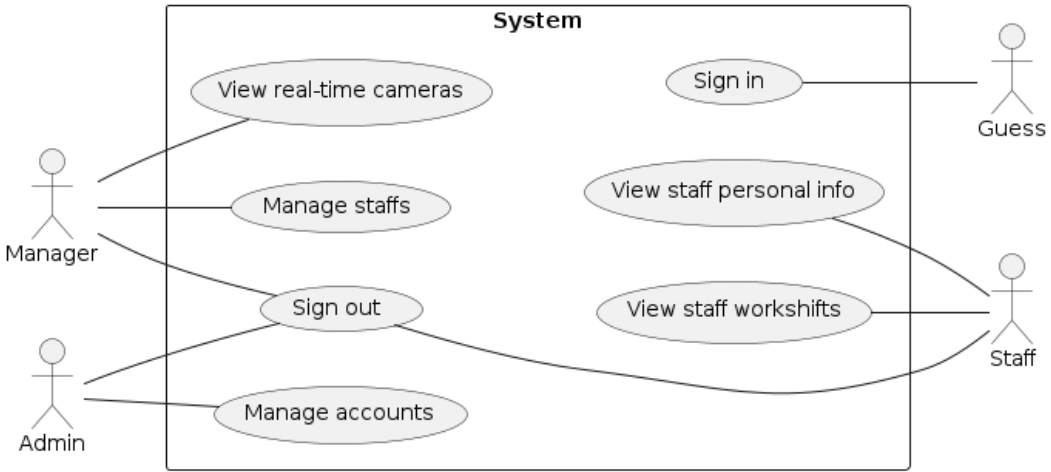
Table 7: Re-ID vs. the proposed STA method. Note that this evaluation was done at track-level after fixing ID switch cases.

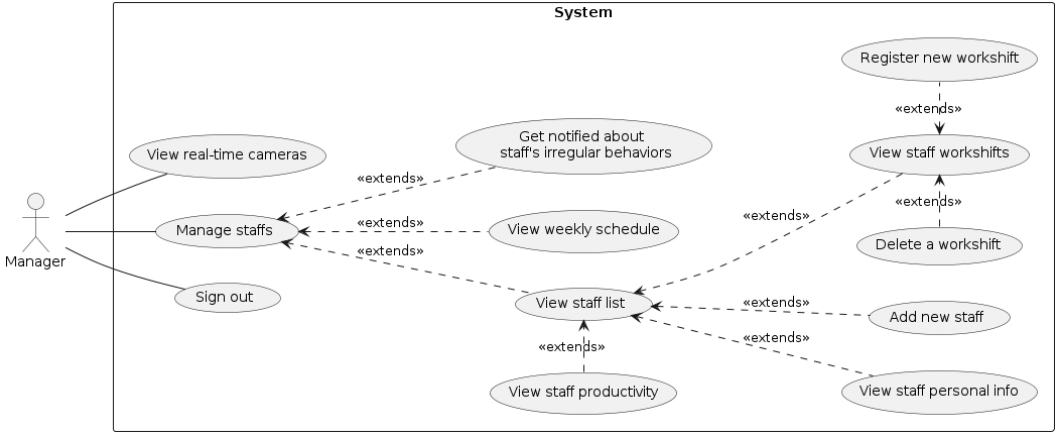
Figure 13: The matching results using Re-ID.

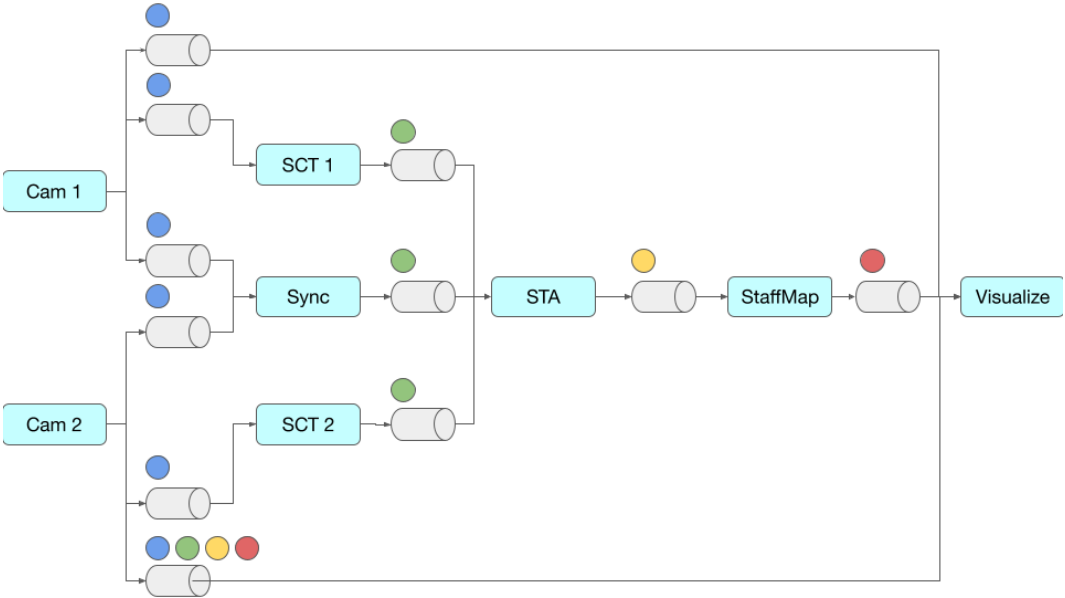
With Re-ID, all individuals wearing the same uniform are considered the same person.

- ◇ Mapping at the frame level is more reliable and can still produce track-level matches.
- ◇ Issues:
  - missing detection: improve with FP filtering and window-based mapping.
  - accurate foot point interpolation: improve with Pose, but need FP filtering to be applied.
- ◇ Those extensions have more impact on complex cases and can be combined to produce a more impressive result.

APPLICATION SYSTEM DEVELOPMENT


System
View real-time cameras
Sign in
Guess
Manager
Manage staffs
View staff personal info
Sign out
View staff workshifts
Staff
Admin
Manage accounts

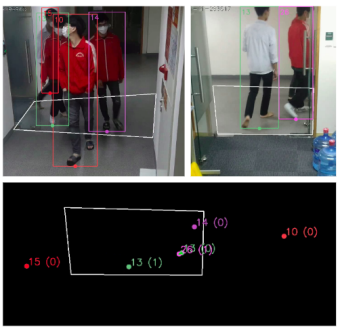

System
Manager
View real-time cameras
Manage staffs
Sign out
Get notified about
staff's irregular behaviors
View weekly schedule
View staff list
View staff productivity
Register new workshift
View staff workshifts
Delete a workshift
Add new staff
View staff personal info
«extends»
«extends»
«extends»
«extends»
«extends»
«extends»
«extends»
«extends»
«extends»


Cam 1
Cam 2
SCT 1
SCT 2
Sync
STA
StaffMap
Visualize

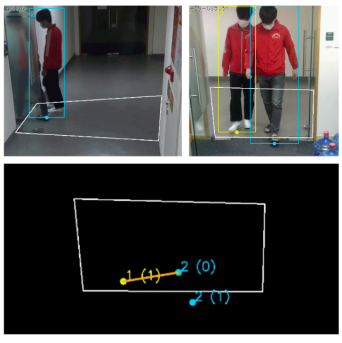
APPENDIX

#identities (GTs), #tracks (IDs) by YOLOv7 + ByteTrack, and #ID switches.

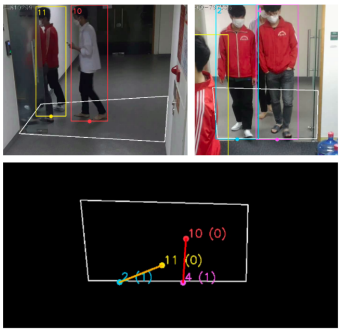
Video set	ID	Camera 1			Camera 2			Camera 3		
		GTs	IDs	SWs	GTs	IDs	SWs	GTs	IDs	SWs
Easy	1	2	4	2	2	6	4	2	2	0
	2	2	4	2	2	5	3	2	5	3
	3	2	4	2	2	6	4	2	3	1
	4	2	5	5	2	8	6	2	3	1
Medium	5	3	6	3	3	10	7	3	5	7
	6	3	9	6	3	10	7	3	7	5
	9	4	8	7	4	9	7	4	5	1
	10	3	8	5	3	9	7	3	4	1
Hard	7	5	13	10	5	15	13	6	9	5
	8	5	14	9	5	15	11	5	18	16
	11	4	11	7	4	12	11	4	6	3
	12	4	13	9	4	14	14	4	21	23



(a) FP filtering works.



(b) FP filtering removes a TP.



(c) FP filtering fails to remove a FP.

Overall, FP filtering works as expected:

- ◇ more impact on complex cases: significant ↓ #FP, slight ↑ #FN.
- ◇ less impact on simple cases.

Overall, window-based mapping works as expected:

- ◇ ↓ #FP and ↓ #FN in most cases, especially complex ones.
- ◇ Compared to FP filtering: FP filtering ↓ #FP more significantly, but ↑ #FN slightly.

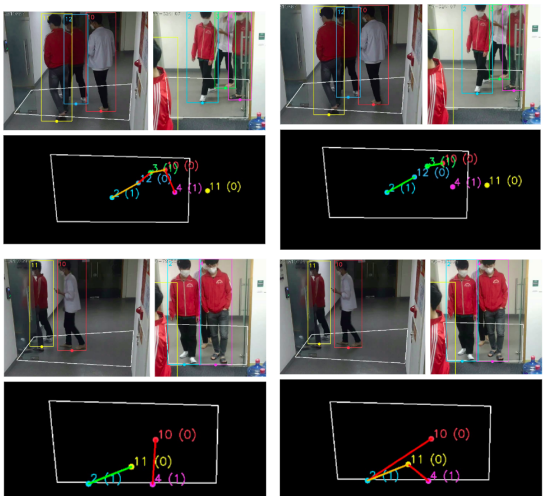
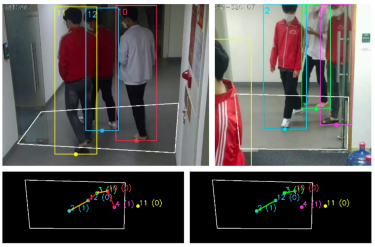
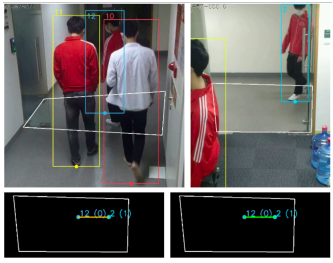


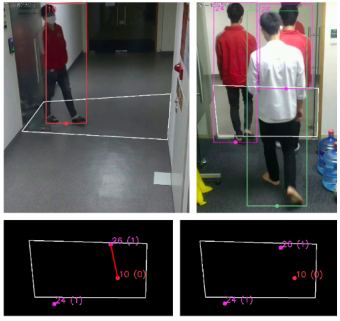
Figure 15: Before (left) and after (right) when window-based mapping works (top) and fails (bottom).



(a) FP filtering when FP's distance < FN's distance



(b) FP filtering when FN's distance > filtering threshold.



(c) window-based on continuous frames of incorrect matching

Figure 16: Examples where FP filtering + window-based mapping is better than...
The bottom left/right of each subfigure is the result by standalone/combined extension.

Review expectation 1: more accurate foot points

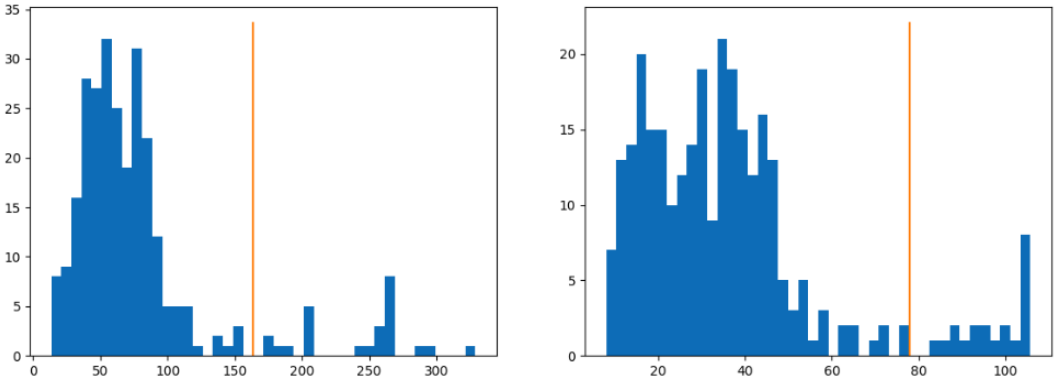
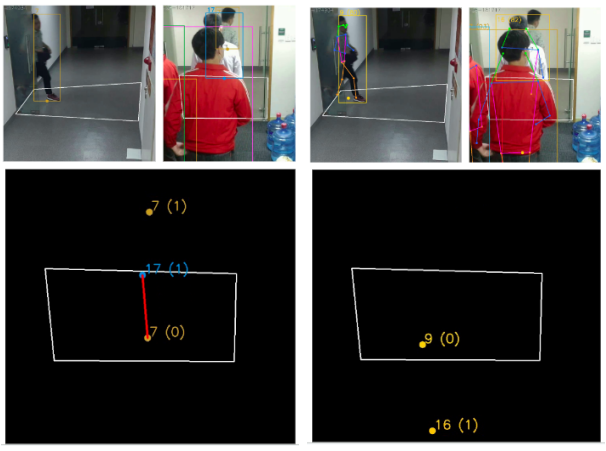


Figure 17: Spatial distance distribution. **a)** using box. **b)** using pose. x-axis is the spatial distance. y-axis is the number of matched pairs. The seam is the upper boundary by IQR(25, 75).

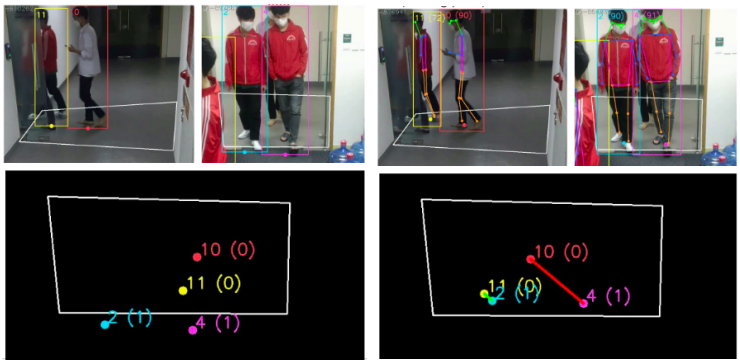
Review expectation 1: more accurate foot points



(a) using box (b) using pose

Pose estimation interpolates foot points more accurately, even with partial occlusion.

Review expectation 2: greater and positive impact on complex cases.



(a) using box (b) using pose

Sometimes, if a person:

- ◇ enters camera, pose has the foot inside the overlap earlier than box.
- ◇ exits camera, pose has the foot outside the overlap later than box.

$\Longrightarrow$ a longer interval inside the overlap $\Longrightarrow \uparrow$ FP if missing detection happens.
$\Longrightarrow$ might be addressed with FP filtering.

Demo videos