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 27, 2023

ONE LOVE. ONE FUTURE

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

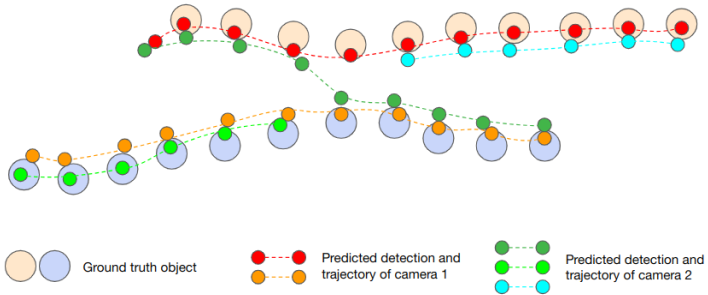


Figure 1: A simple example of confusing track-level matching due to ID switch.

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

- ◇ ignore ID switch which makes mapping at track level counter-intuitive.
- ◇ unreliable evaluation results at frame level.

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

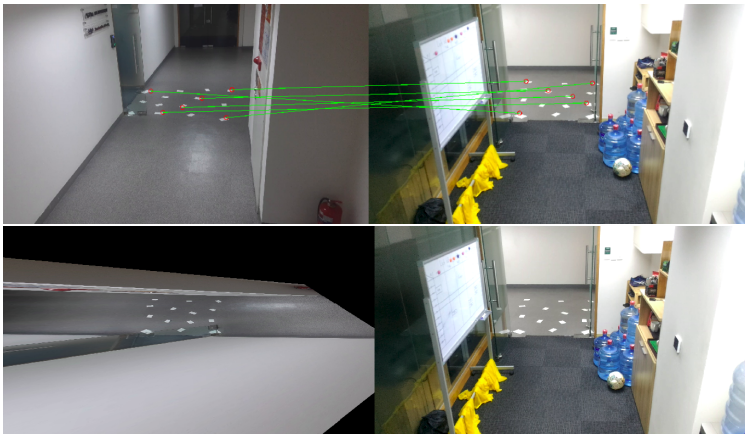
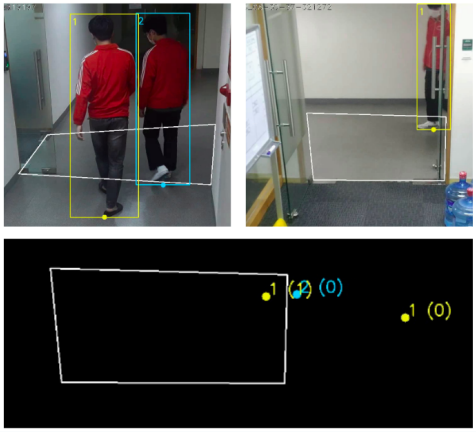
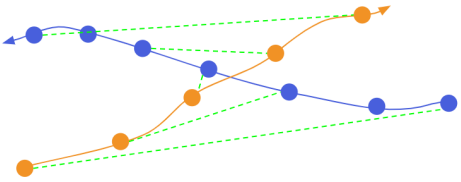


Figure 2: Manually selecting corresponding points in the FOV of a camera pair to build a homography matrix.



(a)



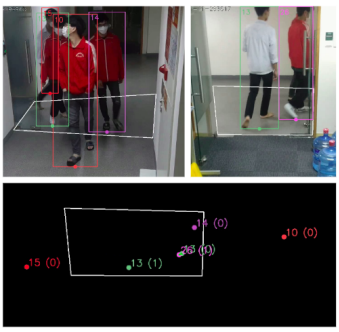
(b)

Figure 3: **a)** foot point interpolation and projection. **b)** Frame-level timestamp matching between 2 tracks.

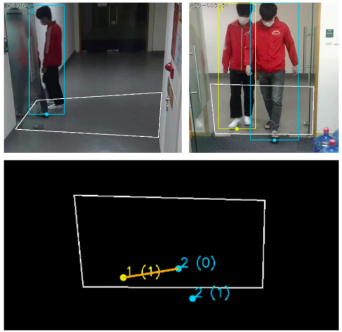
Video set	Baseline	**IQR (20, 80)**
Easy	0.986 (511,7,7)	0.983 (507,7,11)
Medium	0.951 (662,46,22)	0.958 (658,32,26)
Hard	0.913 (966,144,41)	0.927 (959,103,48)
Total	0.941 (2139,197,70)	0.949 (2124,142,85)

Table 1: Baseline method vs. FP filtering. Each cell format is F1(#TP, #FP, #FN).

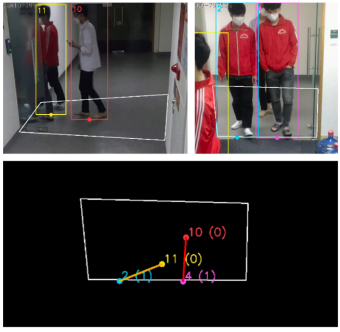
1. Easy set: no significant change.
2. Medium set: #FP $\downarrow$ and #FN $\uparrow$ slightly.
3. Hard set: #FP $\downarrow$ more than Easy and Medium, #FN $\uparrow$ slightly.



(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.

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 2: Baseline method vs. window-based mapping.

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

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.

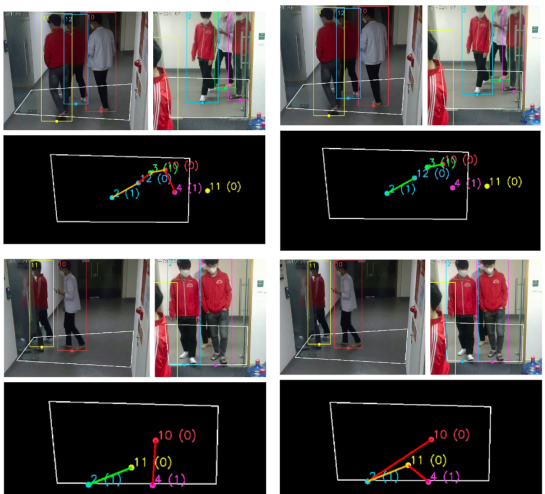


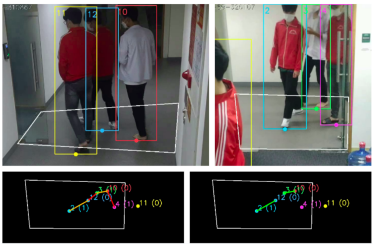
Figure 5: Before (left) and after (right) when window-based mapping works (top) and fails (bottom).

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)

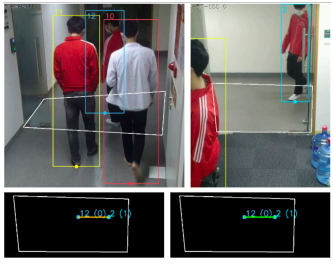
Table 3: Baseline method vs. combination of FP filtering + window-based mapping.

Combining the 2 extensions produces more impressive results than standalone:

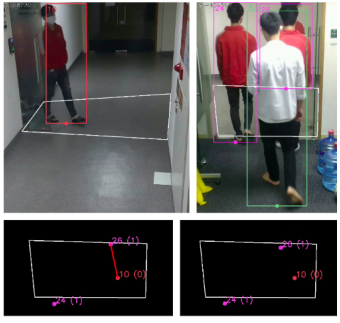
- ◇ Easy set: no significant change.
- ◇ Hard and Medium set: significant ↓ #FP, not much change in #FN.



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

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 4: 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 5: Baseline method using bounding box vs. using pose estimation on individual video.

$\Longrightarrow$ contrary to the expectation

Review expectation 1: more accurate foot points

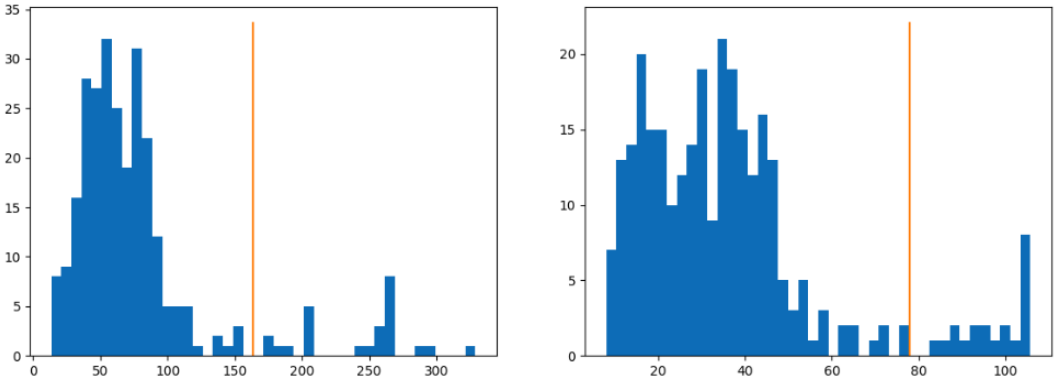
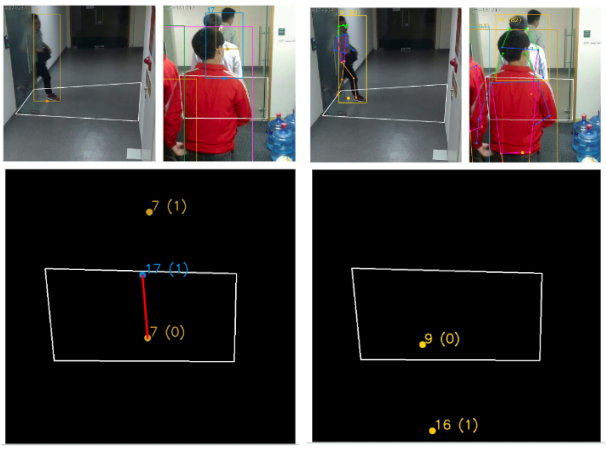


Figure 7: Spatial distance distribution. **a)** using box. **b)** using pose. x-axis is the spatial distance. y-axis is the number of matched pairs. The seam is the upper boundary by IQR(25, 75).

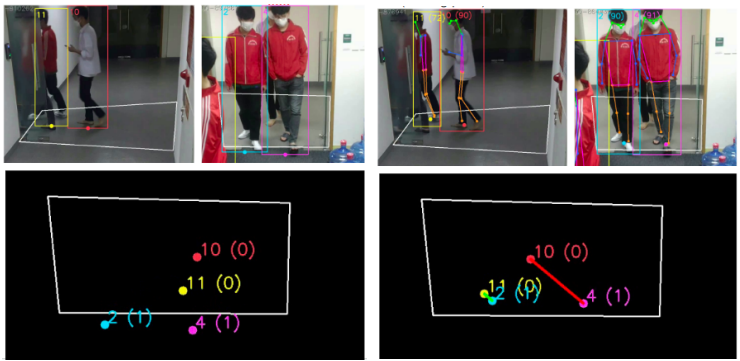
Review expectation 1: more accurate foot points



(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.

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

Demo videos