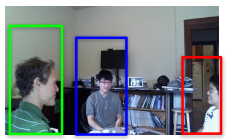
Identifying First-person Camera Wearers in Third-person Videos

Chenyou Fan[1], Jangwon Lee[1], Mingze Xu[1], Krishna Kumar Singh[2], Yong Jae Lee[2],
David J. Crandall[1], Michael S. Ryoo[1]
[1]Indiana University, [2]University of California Davis

1. Introduction

Goal: establish person-level correspondences across egocentric (first-person) and third-person videos. Given a first-person video, decide who is the wearer in the third-person video.

Q: Who took video A, and who took B?



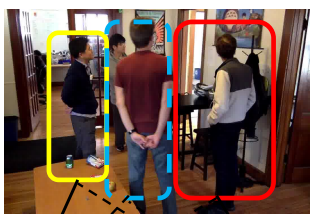
Video A

Video B

2. Data collection

Several (3-5) people appeared in scene, two wearing Xiaoyi cameras.
- 7 sets of synced videos of 5-10 mins each (5 training, 2 test)
- Training: 3,489 correct pairs, 7,399 incorrect pairs
- Test: 1,051 correct pairs, 2,455 incorrect pairs

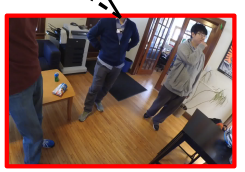
3rd-person camera field-of-view



Correct pair

Incorrect pair

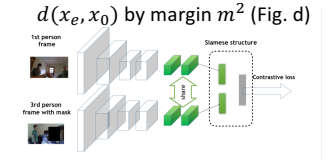
Person A's field-of-view



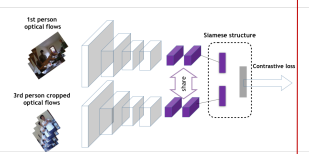
Person B's field-of-view

3. Two-Stream Semi-Siamese Models

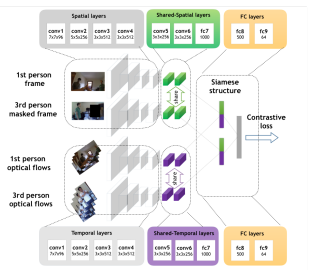
Learning embedding spaces shared by 1st and 3rd-person videos
- spatial overlap between correct pair: spatial CNN (Fig. a)
- temporal correlation: temporal CNN (Fig. b)
- combination: two-stream CNN (Fig. c)
- semi-Siamese: sharing last two convolution layers
- *contrastive loss*: incorrect pair distance $d(x_e, x_p)$ larger than margin
- *triplet loss*: Incorrect pair distance $d(x_e, x_1)$ larger than correct pair $d(x_e, x_0)$ by margin m^2 (Fig. d)



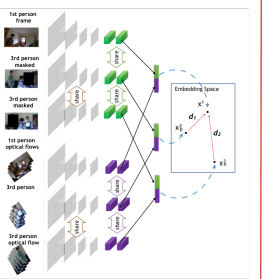
(a) Spatial-domain semi-Siamese network



(b) Motion-domain semi-Siamese network



(c) Two-stream semi-Siamese network



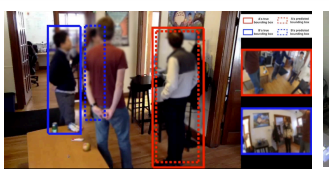
(d) Two-stream semi-triplet network

$$L_{siam}(\theta) = \sum_{i}^{B} y_i ||x_e^i - x_p^i||^2 + (1-y_i)\max(m - ||x_e^i - x_p^i||, 0)^2$$

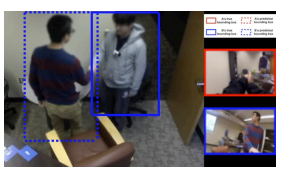
$$L_{trip}(\theta) = \sum_{i}^{B} ||x_e^i - x_1^i||^2 + \max(0, m^2 - (||x_e^i - x_0^i||^2 - ||x_e^i - x_1^i||^2))$$

$y_i = 1$ if x_e and x_p are correct pair, otherwise 0

4. Successful detection and Failure Cases



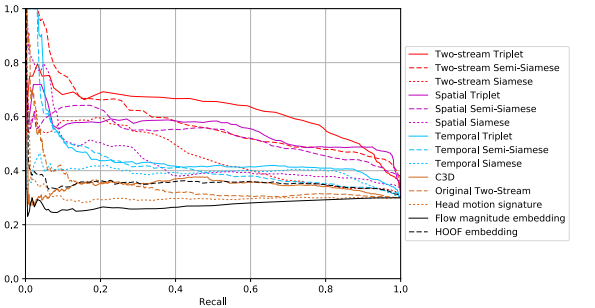
Spatial confusion: person A and B are heavily occluded, then masking fails.



Temporal confusion: person A and B happen to have very similar motion.

5. Results

Binary classification: decide whether a given 1st-person frame was taken by the person in a 3rd-person frame.
Multi-class classification: assign a given 1st-person frame to the correct one of K people appearing in a 3rd-person frame.



Network setting		Evaluation	
Type	Method	Binary AP	Multi Accuracy
Baselines	Flow magnitude to magnitude	0.285	0.250
	HOOF to HOOF	0.316	0.336
	Odometry to HOOF	0.302	0.493
	Velocity to flow magnitude	0.279	0.216
	HOOF embedding	0.354	0.388
	Magnitude embedding	0.276	0.216
	Head Motion Signature [19]	0.300	0.290
	Original Two-stream [25]	0.350	0.460
	C3D [27]	0.334	0.505
Spatial	Siamese	0.481	0.536
	Semi-Siamese	0.528	0.585
	Triplet	0.549	0.588
Temporal	Siamese	0.337	0.372
	Semi-Siamese	0.389	0.445
	Triplet	0.452	0.490
Two-Stream	Siamese	0.453	0.491
	Not-Siamese	0.476	0.554
	Semi-Siamese	0.585	0.639
	Triplet	**0.621**	**0.693**

Table 1. Evaluation in terms of average precision and multi-way classification for baselines and variants of our approach.

Test generality of our approach: Treat one 1st-person video as 3rd-person, use only temporal cue for identification.

Network setting		Evaluation	
Type	Method	Binary AP	Multi Accuracy
Baselines	Flow magnitude to magnitude	0.389	0.442
	HOOF to HOOF	0.382	0.365
	Odometry to HOOF	0.181	0.077
	Velocity to flow magnitude	0.310	0.327
	HOOF embedding	0.405	0.365
	Magnitude embedding	0.406	0.442
	Head Motion Signature [19]	0.359	0.462
	C3D [27]	0.380	0.327
	Two-stream [25] (temporal part)	0.336	0.365
Ours	Temporal Semi-Siamese	**0.412**	**0.500**
	Temporal Triplet	0.386	**0.500**

Table 2. Results for multiple wearable camera experiments.

6.Conclusion

Distance metrics exist between correct pairs of 1st- and 3rd-person video. Three innovations achieved best results in learning this metric: (1) semi-Siamese as opposed to full-Siamese, (2) two-stream CNN to combine spatial and motion cues, and (3) triplet loss instead of Siamese contrastive loss.

[19] Y. Poleg, C. Arora, and S. Peleg. Head motion signatures from egocentric videos. ACCV 2014.
[25] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. NIPS 2014.
[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. ICCV 2015.

This work was supported in part by the NSF (CAREER IIS-1253549), and the IU Office of the Vice Provost for Research.