Joint Person Segmentation and Identification in Synchronized First- and Third-person Videos

Mingze Xu Chenyou Fan Yuchen Wang Michael S. Ryoo David J. Crandall
School of Informatics, Computing, and Engineering
Indiana University, Bloomington, IN 47408

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

COMPUTER VISION LAB

1. Introduction

- **Motivation:** Scenes are often captured by cameras of different types, including fixed, hand-held, and wearable.
- **Goal:** Segment, and identify correspondences between, people in the videos and people holding or wearing the cameras.

Fig. 1. In a scene captured by cameras of different types, both **static** and **wearable**, we want to identify corresponding people and camera wearers.

2. Problems

- **Third-third problem:** Given one or more synchronized third-person videos of a scene, **segment all visible people** and **identify corresponding people** across different videos.
- **Third-first problem:** Given one or more synchronized third-person videos of a scene as well as a video from a wearable camera, **identify and segment the person who was wearing the camera** in the third-person videos.

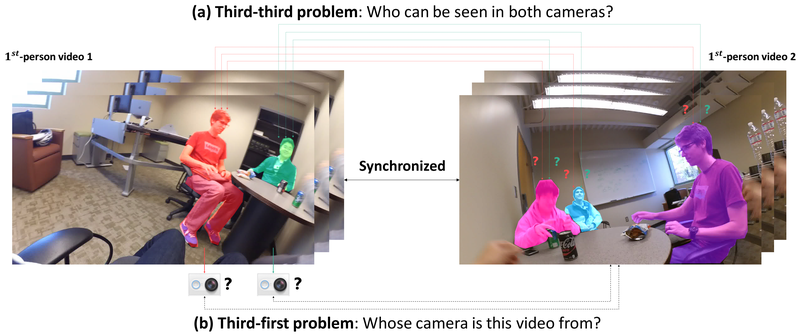

Fig. 2. A *first-person camera* is a wearable camera for which we care about the identity of the camera wearer, while a *third-person camera* is either a static or wearable camera for which we are not interested in determining the wearer.

3. Network Architectures

- **Two-stream Fully Convolutional Network (FCN)**
  - Produces a segmentation mask for the person of interest.
  - Downsamples the extracted features of the softmax layer by 16, and tiles the background and foreground channels by 512.

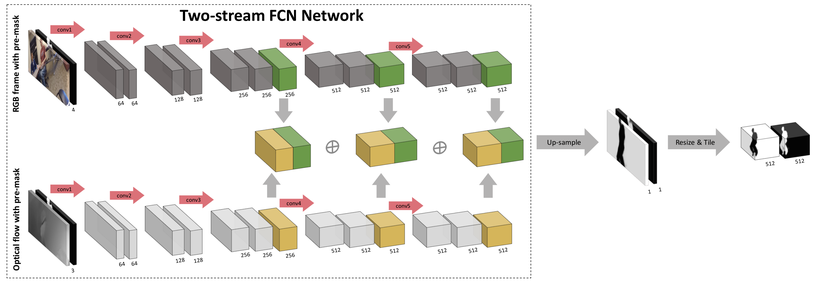

Fig. 3. Our two-stream FCN network.

- **Third-third Network** segments and identifies the people in common across different videos.

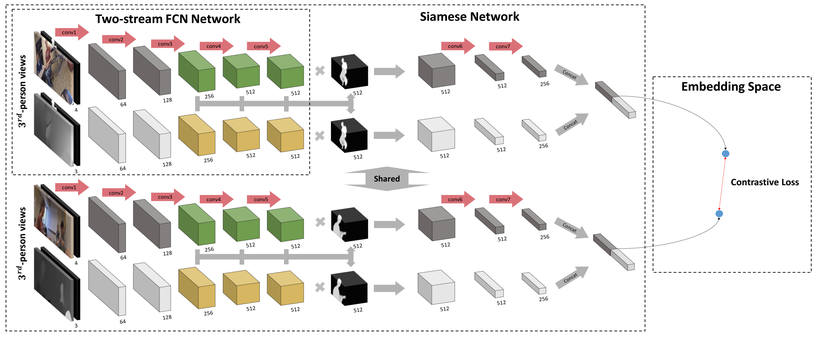

Fig. 4. Our third-third network.

- **Third-first Network** segments and identifies the first-person camera wearer in third-person videos.

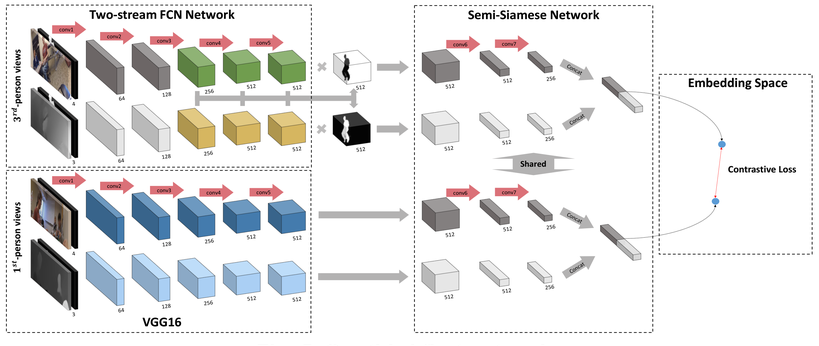

Fig. 5. Our third-first network.

4. Experimental Results

- **IU ShareView dataset** consists of 9 sets of pairs of 5-10 minute **synced first-person videos** in six indoor environments, with a total of **1,277 pixel-level ground truth segmentation maps** of **2,654 annotated person instances**.

Network Architecture					Evaluation		
		Streams			*Segmentation*	*Identification*	
	Backbone	Image	Optical flow	Re-weighting	IoU	mAP	ACC
Baselines	Copy First			-	41.9	-	-
	FCN	X		-	47.1	-	-
	FCN		X	-	50.9	-	-
	FCN	X	X	-	57.3	-	-
Third-Third	VGG	X	X	bounding box [14]	-	44.2	40.1
	FCN	X		soft attention	49.3	44.3	44.5
	FCN		X	soft attention	54.1	48.4	46.2
	FCN	X	X	w/o	60.6	45.6	48.9
	FCN	X	X	soft attention	**62.7**	**49.0**	**55.5**
Third-First	VGG	X	X	bounding box [14]	-	64.1	50.6
	FCN	X		soft attention	47.4	51.4	52.7
	FCN		X	soft attention	58.9	55.1	53.1
	FCN	X	X	w/o	59.8	64.0	61.7
	FCN	X	X	soft attention	**61.9**	**65.2**	**73.1**

Table 1. Experimental results of our models on IU ShareView dataset.

Fig. 6. Sample results. Colors of segmentation and camera views indicate estimated correspondences across different cameras.

5. Conclusion

- Proposed two novel (semi-)Siamese FCNs for joint person segmentation and identification, and evaluated on a new, challenging dataset with pixel-level ground truth and correspondences across first- and third-person cameras.
- Results show that jointly inferring segmentation and people correspondences helps perform each task more accurately.

4. Shervin Ardeshir, Ali Borji. Ego2Top: Matching viewers in egocentric and top-view videos. *ECCV* 2016.
14. Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David Crandall, and Michael Ryoo. Identifying First-person Camera Wearers in Third-person Videos. *CVPR* 2017.

Acknowledgements: *This work was supported by the National Science Foundation (CAREER IIS-1253549), and the IU Office of the Vice Provost for Research, the College of Arts and Sciences, and the School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project "Learning: Brains, Machines, and Children."*