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

Mingze Xu(✉), Chenyou Fan, Yuchen Wang, Michael S. Ryoo, and David J. Crandall

School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA
{mx6,fan6,wang617,mryoo,djcran}@indiana.edu

Abstract. In a world of pervasive cameras, public spaces are often captured from multiple perspectives by cameras of different types, both fixed and mobile. An important problem is to organize these heterogeneous collections of videos by finding connections between them, such as identifying correspondences between the people appearing in the videos and the people holding or wearing the cameras. In this paper, we wish to solve two specific problems: (1) given two or more synchronized third-person videos of a scene, produce a pixel-level segmentation of each visible person and identify corresponding people across different views (i.e., determine who in camera A corresponds with whom in camera B), and (2) given one or more synchronized third-person videos as well as a first-person video taken by a mobile or wearable camera, segment and identify the camera wearer in the third-person videos. Unlike previous work which requires ground truth bounding boxes to estimate the correspondences, we perform person segmentation and identification jointly. We find that solving these two problems simultaneously is mutually beneficial, because better fine-grained segmentation allows us to better perform matching across views, and information from multiple views helps us perform more accurate segmentation. We evaluate our approach on two challenging datasets of interacting people captured from multiple wearable cameras, and show that our proposed method performs significantly better than the state-of-the-art on both person segmentation and identification.

Keyword: Synchronized first- and third-person cameras

1 Introduction

There will be an estimated 45 *billion* cameras on Earth by 2022—more than five times the number of people [25]! In a world with so many cameras, it will be commonplace for a scene to be simultaneously recorded by multiple cameras of different types. For example, a busy city street may be recorded by not only fixed surveillance cameras, but also by mobile cameras on smartphones, laptops,

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11205, pp. 656–672, 2018.
https://doi.org/10.1007/978-3-030-01246-5_39

tablets, self-driving cars, and even wearable devices like GoPro [1] and Snap Spectacles [2]. As cameras continue to multiply, new techniques will be needed to organize and make sense of these weakly-structured collections of video. For example, a key problem in many applications is to detect, identify, and track people. Combining data from multiple cameras could significantly improve performance on this and other scene understanding problems, since evidence from multiple viewpoints could help resolve ambiguities caused by occlusion, perspective distortion, etc. However, integrating evidence across heterogeneous cameras in unconstrained dynamic environments is a challenge, especially for wearable and mobile devices where the camera is moving unpredictably.

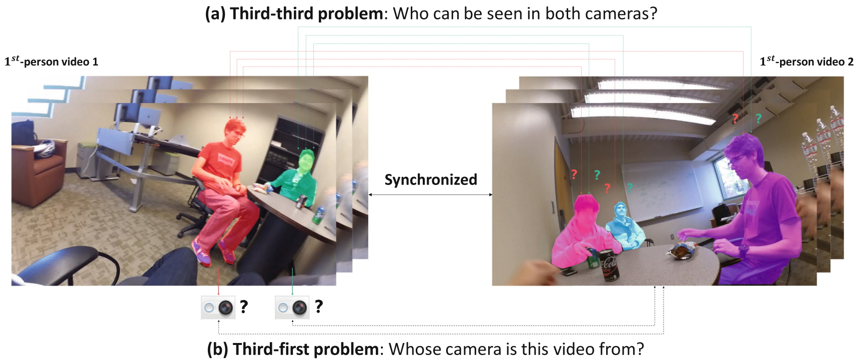

Fig. 1. Two or more people move around an environment while wearing cameras. We are interested in two specific, related problems: (a) given one or more synchronized third-person videos of a scene, segment all the visible people and identify corresponding people across the different videos; and (b) given one or more synchronized third-person videos of a scene as well as a video that was taken by a wearable first-person camera, identify and segment the person who was wearing the camera in the third-person videos.

For example, consider a law enforcement scenario in which multiple police officers chase a suspect through a crowded square. Body-worn police cameras (which nearly 95% of U.S. police departments use or plan to deploy [24]) record events from the officers' perspectives. Investigators later want to reconstruct the incident by combining the first-person wearable camera videos with third-person views from surveillance cameras and civilian smartphone videos uploaded to social media. In any given frame of any given camera, they may want to identify: (1) fine-grained, pixel-level segmentation masks for all people of interest, including both the suspect and the officers (e.g., for activity or action recognition), (2) the instances in which one of the camera wearers (officers) was visible in another camera's view, and (3) instances of the same person appearing in different views at the same time. The scene is complex and crowded, requiring fine-grained segmentation masks to separate individual people (since frequent occlusions would cause bounding boxes to overlap). The wearable camera videos are particularly challenging because the cameras themselves are moving rapidly.

While person tracking and (re-)identification are well-studied in computer vision [37,44], only recently have they been considered in challenging scenarios of heterogeneous first-person and traditional cameras. Ardeshir and Borji [4] consider the case of several people moving around while wearing cameras, and try to match each of these first-person views to one of the people appearing in a third-person, overhead view of the scene. This is challenging because the camera wearer is never seen in their own wearable video, so he or she must be identified by matching their motion from a third-person perspective with the first-person visual changes that are induced by their movements. That paper's approach is applicable in closed settings with overhead cameras (e.g., a museum), but not in unconstrained environments such as our law enforcement example. Fan *et al.* [14] relax many assumptions, allowing arbitrary third-person camera views and including evidence based on scene appearance. Zheng *et al.* [43] consider the distinct problem of identifying the same person appearing in multiple wearable camera videos (but not trying to identify the camera wearers themselves). But these techniques identify individual people using bounding boxes, which are too coarse in crowded scenes with frequent occlusions. Moreover, these techniques assume that accurate oracle bounding boxes are available (even at test time).

In this paper, we consider the more challenging problem of not only finding correspondences between people in first- and third-person cameras, but also producing pixel-level segmentation masks of the people in each view (see Fig. 1). We define a *first-person* camera to be 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. Our hypothesis is that simultaneous segmentation and matching is mutually beneficial: segmentation helps refine matching by producing finer-grained appearance features (compared to bounding boxes), which are important in crowded scenes with many occlusions, while matching helps locate a person of interest and produce better segmentation masks, which in turn help in tasks like activity and action recognition. We show that previous work [14] is a special case of ours, since we can naturally handle their first- and third-person cases. We evaluate on two publicly available datasets augmented with pixel-level annotations, showing that we achieve significantly better results than numerous baselines.

2 Related Work

We are not aware of work on joint person segmentation and identification in first- and third-person cameras, so we draw inspiration from several related problems.

Object Segmentation in Images and Videos. Deep learning has achieved state-of-the-art performance on semantic image segmentation [5,9,27,28,42], typically using fully convolutional networks (FCNs) that extract low-resolution features and then up-sample. Other approaches [18,26,30,31] are based on region proposals, inspired by R-CNNs [16,32] for object detection. For example, Mask R-CNNs [18] separately predict object masks and their class labels, avoiding competition among classes and improving performance for overlapped instances.

For object segmentation in video [7,20–22,38,39], most methods assume that the object mask in the first frame is known (during both training and testing) and the task is to propagate them to subsequent frames. Khoreva *et al.* [29] propose guided instance segmentation that uses the object mask from the previous frame to predict the next one. The network is pre-trained (off-line) on static images and fine-tuned (on-line) on the first frame's annotations for specific objects of interest. We follow a similar formulation, except that we incorporate both appearance and optical flow in a two-stream network, helping to better update the object mask across time. Our work is also inspired by the pixel-level Siamese matching network of Yoon *et al.* [41] that segments and identifies objects, even those not seen during training. We extend to multiple cameras by using object instances across multiple synchronized videos to learn variations and correspondences in appearance across views. Cheng *et al.* [10] propose a two-stream network which outputs segmentation and optical flow simultaneously, where segmentation focuses on objectness and optical flow exploits motion. Inspired by their observation that segmentation and optical flow benefit each other, we propose a novel architecture that jointly performs person segmentation and identification.

Co-segmentation. Our work is related to co-segmentation of objects appearing in multiple images [33] or videos [8,11,15,17,34]. Several methods use Markov Random Fields with a regularized difference of feature histograms, for example, by assuming a Gaussian prior on the objectness appearance [33] or computing sum squared differences [6]. Chiu *et al.* [11] use distance-dependent Chinese Restaurant Processes as priors on both appearance and motion for unsupervised (not semantic) co-segmentation. Fu *et al.* [15] address video co-segmentation as CRF inference on an object co-selection graph, but segmentation candidates are computed only by a category-independent method [13] and are not refined from information across multiple videos. Guo *et al.* [17] perform iterative constrained clustering using seed superpixels and pairwise constraints, and refine the segmentation with a multi-class MRF. Most of these methods assume that either a target object appears in all videos or that videos contain at least one common target object, and none apply deep learning. To the best of our knowledge, ours is the first paper to propose a deep learning approach to co-segmentation in videos, and is applicable both to single and multiple camera scenarios.

First-Person Cameras. Ardeshir and Borji [4] match a set of first-person videos to a set of people appearing in a top-view video using graph matching, but assume there are multiple first-person cameras sharing the same field of view at any time and only consider third-person cameras that are overhead. Fan *et al.* [14] identify a first-person camera wearer in a third-person video using a two-stream semi-Siamese network that incorporates spatial and temporal information from both views, and learns a joint embedding space from first- and third-person matches. Zheng *et al.* [43] identify people appearing in multiple wearable camera videos (but do not identify the camera wearers themselves).

The above work assumes that the people have been detected with accurate bounding boxes in both training *and test* datasets. We build on these methods, proposing a novel architecture that simultaneously segments and identifies camera wearers and others. We find that segmenting and identifying are mutually beneficial; in the law scenario described above with crowded scenes and occluded people, for example, fine-grained segmentation masks are needed to accurately extract visual features specific to any given person, while identity information from multiple views helps accurately segment the person in any individual view.

3 Our Approach

Given two or more videos taken from a set of cameras (potentially both static and wearable cameras), we wish to segment each person appearing in these videos, identify matches between segments that correspond to the same person across different views, and identify the segments that correspond to the wearer of each first-person camera. The main idea is that despite having very different perspectives, synchronized cameras recording the same environment should be capturing some of the same people and background objects. This overlap permits finding similarities and correspondences among these videos in both visual and motion domains, as long as differences caused by differing viewpoints are ignored. Unlike prior work [14] which assumes a ground truth bounding box is available for each person in each frame, we perform segmentation and matching simultaneously. We hypothesize that these two tasks are mutually beneficial: person segmentations provide more accurate information than coarse bounding boxes for people matching, while people's appearance and motion from different perspectives produce better segmentation masks.

More concretely, we formulate our problem as two separate tasks. The *third-third problem* is to segment each person and find person correspondences across different views captured from a pair of third-person cameras. The *third-first problem* is to segment and identify the camera wearer of a given first-person video in third-person videos. We first introduce a basic network architecture for both problems: a two-stream fully convolutional network (FCN) that estimates a segmentation mask for each person using the current RGB frame, stacked optical flow fields, and segmentation result of the previous frame (which we call the *pre-mask*) (Sect. 3.1). We then introduce a Siamese network for each of our two problems, that incorporates the FCN and allows person segmentation and identification to benefit each other (Sect. 3.2). Finally we describe our loss used for segmentation and distance metric learning (Sect. 3.3).

3.1 Two-Stream FCN Network

We use FCN8s [28] as the basis of our framework but with several important modifications. We chose FCN8s due to their effectiveness and compactness, although other architectures such as DeepLabv3+ [9] and Mask R-CNN [18] could be easily used. Figure 2 presents our novel architecture. To take advantage of video and incorporate evidence from both appearance and motion, we expand FCN8s to a two-stream architecture, where a *visual stream* receives RGB frames (top of Fig. 2) and a *motion stream* receives stacked optical flow fields (bottom).

This design is inspired by Simonyan and Zisserman [35], although their network was proposed for a completely different problem (action recognition from a single static camera). To jointly consider both spatial and temporal information, we use "early" fusion to concatenate features at levels `pool3`, `pool4`, and `pool5` (middle of Fig. 2). Following FCN8s to incorporate "coarse, high level information with fine, low level information" [28] for more accurate segmentation, we combine the fused features from these different levels.

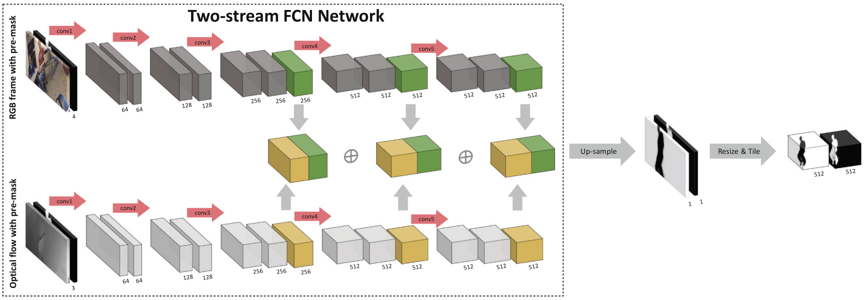

Fig. 2. Visualization of our two-stream FCN network. We feed RGB frames with pre-masks to the visual stream (top, dark grey) and stacked optical flow fields with pre-mask to the motion stream (bottom, light grey). The spatial and temporal features at `pool3`, `pool4`, and `pool5` are fused to predict the segmentation of the target person. We downsample the extracted features of the `softmax` layer by 16, then tile the background and foreground channels by 512, separately.

However, in contrast to Long *et al.*'s FCN8s, our two-stream FCN targets instance segmentation: we want to segment specific people, not just all instances of the "person" class. We address this with an instance-by-instance strategy in both training and test, in which we only consider a single person at a time. In order to guide the network to segment a specific person among the many that may appear in a frame, we append that person's binary pre-mask (without any semantic information) to the input of each stream as an additional channel. This pre-mask provides a rough estimate of the person's location and his or her approximate shape in the current frame. In training, our network is pre-trained by taking ground truth pre-masks as inputs, and then fine-tuned with estimated masks from the previous frame. In testing, we assume that we have a (possibly quite coarse) segmentation of each person in the first frame and propagate this mask forward by evaluating each subsequent unlabeled frame in sequence. A pixel-level classification loss function is used to guide learning (Sect. 3.3).

3.2 Siamese Networks

The network in the last section learns to estimate the segmentation mask of a specific person across frames of video. We now use this network in a Siamese structure with a contrastive loss to match person instances across different third- and first-person views. The main idea behind our Siamese networks is to learn an embedding space such that features captured by different cameras from different perspectives are close together only if they actually belong to the same person—i.e., so that a person's appearance features are invariant to camera viewpoint. The Siamese formulation allows us to simultaneously learn the viewpoint-invariant embedding space for matching identities and the pixel-wise segmentation network described above in an end-to-end fashion. Moreover, our Siamese (or semi-Siamese) FCN architecture improves the invariance of object segmentation across different perspectives and transformations. In contrast to co-segmentation methods that require pairs of images or videos in both training and testing, our approach only need pairs in the training phase. In testing, our two-stream FCN network can be applied to any single stream input, and uses the embedding space to match with others. To allow the segmentation network to receive arbitrary sizes of inputs, our contrastive loss function is generalized to a 3D representation space, with a Euclidean distance for positive exemplars and a hinge loss for negative ones.

In particular, we explore two Siamese network structures, customized for our two tasks: the third-third problem of segmenting and matching people across a pair of cameras, and the third-first problem of segmenting a person of interest and identifying if he or she is the wearer of a first-person camera. The third-third problem considers a more general case in which the cameras may be static or may be wearable, but they are all viewing a person of interest from a third-person viewpoint; we thus use a full-Siamese network that shares all convolution layers in the FCN branch and the embedding layers. In contrast, the third-first problem must match feature representations from different perspectives (identifying how a camera wearer's ego-motion visible in a first-person view correlates with that same motion's appearance from a third-person view). As in [14], our third-first network is formulated in a semi-Siamese structure, where separate shallow layers capture different low-level features while deeper ones are shared.

Third-Third Network. Figure 3 shows the architecture of our third-third network, which segments and matches people in common from a pair of third-person camera views. We use a Siamese structure with two branches of the FCN network from Fig. 2 (and discussed in Sect. 3.1), where all corresponding convolution layers are shared. The Siamese branch is thus encouraged to learn relationships between people's appearance in different views by optimizing a generalized embedding space. The key idea is that despite being captured from very different perspectives, the same person in synchronized videos should have some correspondences in both visual and motion domains.

In more detail, given an RGB frame and optical flow fields (appended with the pre-mask of the person of interest) as inputs, each of size $W \times H$, the FCN branch estimates a binary-valued person segmentation mask of the same size. The Siamese branch is then appended to the `pool5` layer of both visual and motion streams with an input size of $512 \times W' \times H'$, where $W' = \frac{W}{16}$ and $H' = \frac{H}{16}$, for matching. To obtain more accurate representations for each

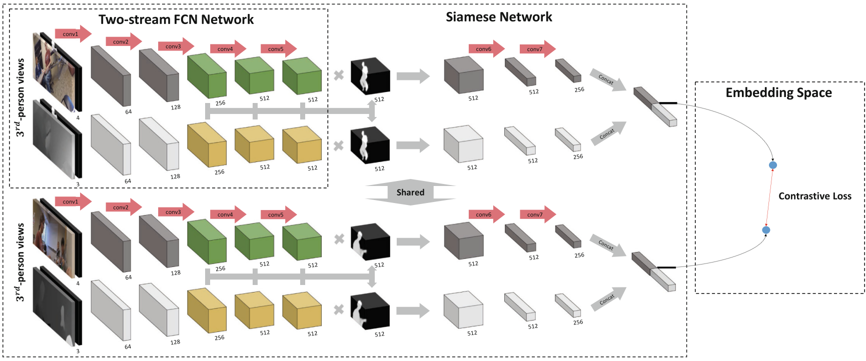

Fig. 3. Our third-third network segments and identifies the people in common across different videos. The network is composed of two FCN branches with a Siamese structure, where all convolution layers (shown in the same color) are shared. (Color figure online)

"target" person, we re-weight the spatial and temporal features by multiplying them with the confidence outputs of the FCN branch. To emphasize the pixel positions belonging to the person while retaining some contextual information, we use soft attention maps after the softmax layer rather than the estimated segmentation mask. We first resize the soft attention of the foreground from $1 \times W \times H$ to $1 \times W' \times H'$ and tile it to $512 \times W' \times H'$ to fit the size of `pool5` outputs. For both visual and motion streams, we multiply this resized confidence map with the features, which gives a higher score to the person's pixels and a low score to the background. By "cropping out" the region corresponding to a person from the feature maps, the match across two views should receive a higher correspondence. This correspondence will also back-propagate its confidence to improve segmentation. Finally, the re-weighted spatial and temporal features are concatenated together for matching each person instance.

Third-First Network. Figure 4 shows the architecture of our third-first network, the goal of which is to segment a first-person camera wearer in third-person videos and to recognize the correspondence between the first-person view and its representation in third-person videos. To be specific, given a first-person video, our network must decide which, if anyone, of the people appearing in a third-person video is the wearer of this first-person camera, and to estimate the wearer's segmentation. In contrast to the third-third network which has two FCN branches focusing on the same task (person segmentation), the second branch of the third-first network receives the first-person videos as inputs and is designed to extract the wearer's ego-motion and the visual information of the background, which hopefully also provides constraints for the segmentation. We thus propose a semi-Siamese network to learn the first- and third-person distance metric, where the first-person branch has a similar structure to the FCN but without the up-sampling layers or the segmentation loss. The structure

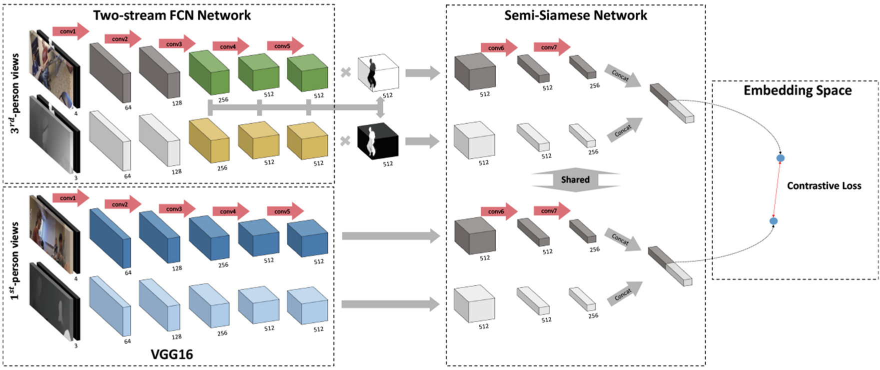

Fig. 4. Our third-first network segments and identifies the first-person camera wearer in third-person videos. The network is formulated in a semi-Siamese structure where only convolution layers of the embedding space (shown in the same color) are shared. (Color figure online)

of the Siamese branch is similar to that of the third-third network, but with a different re-weighting method: we multiply the spatial features with the soft attention of the background but the temporal features with the soft attention of the foreground. We do this because camera wearers do not appear in their own first-person videos (with occasional exceptions of arms or hands), but the backgrounds reflect some similarities between different perspectives; meanwhile, motion features of camera wearers in third-person videos is related to the ego-motion in first-person videos. The re-weighted appearance and motion features are then concatenated after several convolution operations, as discussed above.

3.3 Loss Functions

We propose two loss functions for joint segmentation and distance metric optimization for a batch of N training exemplars. First, *sigmoid cross entropy loss* compares a predicted segmentation mask to ground truth,

$$L_{seg} = -\sum_{i}^{N}\sum_{w}^{W}\sum_{h}^{H}\left(S_{i,w,h}\cdot\log\hat{S}_{i,w,h} + (1-S_{i,w,h})\cdot\log(1-\hat{S}_{i,w,h})\right), \quad (1)$$

where $\hat{S}_i \in \{0,1\}^{W\times H}$ is the predicted segmentation mask of exemplar i and $S_i \in \{0,1\}^{W\times H}$ is the corresponding ground truth mask. Second, *generalized contrastive loss* encourages low distances between positive exemplars (pairs of corresponding people) and high distances between negative ones,

$$\begin{aligned} L_{siam} = \sum_{i}^{N}\sum_{c}^{C}\sum_{w}^{W''}\sum_{h}^{H''} & y_i\,||a_{i,c,w,h}-b_{i,c,w,h}||^2 \\ & + (1-y_i)\max(m-||a_{i,c,w,h}-b_{i,c,w,h}||,\,0)^2, \end{aligned} \quad (2)$$

where m is a constant, a_i and b_i are two features corresponding to exemplar i, and y_i is 1 if i is a correct correspondence and 0 otherwise. This loss enables our model to learn an embedding space for arbitrary input sizes.

4 Experiments

We test our third-third and third-first networks on joint person segmentation and identification in two datasets of synchronized first- and third-person videos, collected by two different authors. We primarily evaluate on the publicly available IU ShareView dataset [14], consisting of 9 sets of two 5–10 min first-person videos. Each set contains 3–4 participants performing a variety of everyday activities (shaking hands, chatting, eating, etc.) in one of six indoor environments. Each person in each frame is annotated with a ground truth bounding box and a unique person ID. To evaluate our methods on person segmentation, we manually augmented a subset of the dataset with pixel-level person annotations, for a total of 1,277 labeled frames containing 2,654 annotated person instances. We computed optical flow fields for all videos using FlowNet2.0 [19].

Since adjacent frames are typically highly correlated, we split the training and test datasets at the video level, with 6 video sets used for training (875 annotated frames) and 3 sets for testing (402 annotated frames). In each set of videos, there are 3–4 participants, two of which wear first-person cameras. Note that a first-person camera never sees its own wearer, so the people not wearing cameras are the ones who are in common across the first-person videos. Since our approach uses sequences of contiguous frames and pairs of instances (either a pair of two people or a pair of one person and one camera view), we divide each video set into several short sequences, each with 10–15 consecutive frames. More specifically, in training we create 484 positive and 1,452 negative pairs for the third-third problem, and 865 positive and 1,241 negative pairs for the third-first problem (about a 1:3 ratio). In testing, each problem has 10 sequences of pairs of videos, and each video has 20 consecutive frames (about 4 s). Thus we have about 400 annotated test frames for evaluating matching, and about 1,000 person instances for evaluating segmentation (since every frame has 2–3 people).

We also evaluate our models on a subset of UTokyo Ego-Surf [40], which contains 8 diverse groups of first-person videos recorded synchronously during face-to-face conversations in both indoor and outdoor environments. Limited by the size of the dataset (only 3 available pairs of short videos including 3–4 participants), we use it only for testing, and still train on IU ShareView. As before, we manually created pixel-level person annotations for 10 sequences of pairs of videos, each with 20 consecutive frames.

4.1 Evaluation Protocol

We implemented our networks in PyTorch [3], and performed all experiments on a single Nvidia Titan X Pascal GPU.

Training. Our training process consisted of two stages: (a) optimizing only the FCN branch supervised by the pixel-level classifier for providing imperfect but reasonable soft attentions, and (b) optimizing the joint model (either the third-third or third-first network) based on the person segmentation and identification tasks, simultaneously. Our two-stream FCN network is built on VGG16 [36], and we initialized both visual and motion streams using weights pre-trained on ImageNet [12]. The FCN branch was then optimized with an instance-by-instance strategy, which only considers one particular person of interest at a time, and uses the ground truth pre-mask as an additional channel to indicate which person the network should focus on. We used stochastic gradient descent (SGD) with fixed learning rate 10^{-4}, momentum 0.9, weight decay 0.0005, and batch size 25. Learning was terminated after 30 epochs. Our joint model was then initialized with the weights of the pre-trained FCN and fine-tuned by considering pairs of instances as inputs for person segmentation and identification. We again used SGD optimization but with learning rate 10^{-5}. For the first 20 epochs, we froze the weights of the FCN branch, and optimized the Siamese branch to make the contrastive loss converge to a "reasonable" range (not too large to destroy the soft attention). We then started the joint learning process, and terminated after *another* 40 epochs.

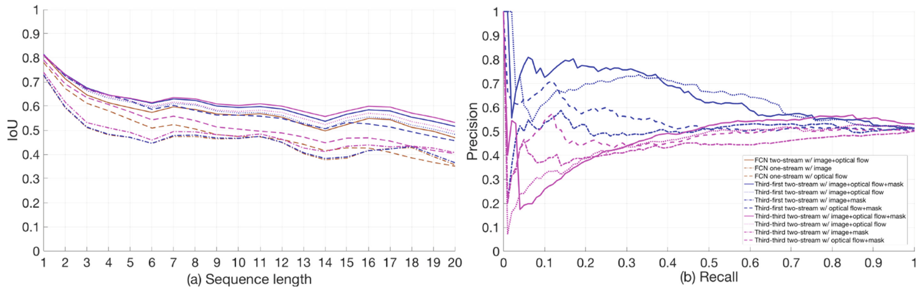

Fig. 5. IoU and precision-recall curves of our models on IU ShareView dataset [14]

Testing. In contrast to training, which requires pairs of videos as inputs, our joint model can be applied to an individual stream, where each video frame is processed to simultaneously estimate each person's segmentation and extract corresponding features for matching between different streams. In testing, all possible pairs of instances are considered as candidate matches: each pair contains either two people from different videos in the third-third problem, or a first-person camera view and a person appearing in a third-person video in the third-first problem. Unlike methods that require a pair of instances as input, our approach only needs to process each person and camera view once.

4.2 Evaluation

For both third-first and third-third problems, we evaluate our method with two tasks: person (1) segmentation and (2) identification across multiple cameras.

Person Segmentation is evaluated in terms of *intersection over union (IoU)* between the estimated segmentation maps and the ground truth. This is measured over each video in the test dataset by applying our models to each frame. Our model sequentially takes the segmentation results from the previous frame (the pre-mask) as input to guide the segmentation of the next frame. In the evaluation, the ground segmentation mask of the first (and only the first) video frame is assumed to be available.

Person Identification is evaluated with *Mean average precision (mAP)* and *Accuracy (ACC)*, each of which takes a different view of the problem. mAP treats people matching as a retrieval problem: given all possible pairs of person instances from two different cameras (i.e., two person instances from different third-person videos in the third-third problem *or* one person instance from third-person video and one first-person video in the third-first problem), we wish to retrieve all pairs corresponding to the same person. ACC evaluates whether the single best match for a given candidate is correct: for every person instance in each view, the classifier is forced to choose a single matching instance in all other views, and we calculate the percentage of matches that are correct. This setting is the same to the one used in Fan *et al.* [14], except that their task is significantly easier because they assume person ground-truth bounding boxes are available during both training and testing, whereas our approach must infer the person's position (as well as segmentation mask) automatically.

4.3 Experimental Results

Baselines. To characterize the difficulty of segmentation in this dataset, we first test several baselines, shown in Table 1 for IU ShareView. Copy First simply propagates the ground truth segmentation mask from the first frame to all following frames in the sequence. In a completely static scenes with no motion, the IoU of Copy First should be 100.0, but our dataset includes frequent motion of both the wearable cameras and people, and thus shows a relatively low IoU of 41.9. A second baseline consisting of a single-stream FCN using only image information achieves somewhat better IoU of 47.1, while a third baseline consisting of a single-stream FCN using only optical flow achieves 50.9. A two-stream baseline FCN that combines both visual and motion performs significantly better than either one-stream network, achieving IoU of 57.3.

Our Models. We next test our approach that jointly performs segmentation with person instance matching. On segmentation, our full model produces an IoU of 62.7 for the third-third scenario and 61.9 for third-first, compared to 57.3 for the two-stream baseline that performs only segmentation. Figure 5(a) reports more detailed analysis of the segmentation performance (Y-axis) based on the length of video sequences (X-axis), and shows that our approach is still able to predict reasonable results on long videos. To permit a fair comparison across models, both the one- and two-stream FCNs were optimized with the same hyper-parameters (discussed in Sect. 4.1). Table 1 also presents results on person instance matching on IU ShareView. We achieved mAP scores of 49.0 and 65.2 on the third-third and third-first problems, respectively, and ACCs of 55.5 and 73.1. We compare these results with the state-of-the-art method of Fan *et al.* [14]. Their task is to match first-person camera views to camera wearers in *static* third-person video, so we extend it to our third-third and third-first problems by re-implementing their best model using VGG16 [36] (instead of AlexNet [23]) and training on our new, augmented dataset. The results show that our joint model outperforms in both third-third (mAP of 49.0 vs. 44.2) and third-first (mAP of 65.2 vs. 64.1) problems. This is likely due to learning a more accurate embedding space, with the help of jointly learning to perform segmentation. More importantly, our approach is able to obtain more accurate feature representations from people's pixel-level locations rather than simply relying on rough bounding boxes. Figure 5(b) compares the precision-recall curves of the different techniques for person matching.

Table 1. Experimental results of our models on IU ShareView dataset [14]

Network architecture					Evaluation		
	Backbone	*Streams*		Re-weighting	*Segmentation*	*Identification*	
		Image	Optical flow		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**

UTokyo Ego-Surf Dataset. We also test our models on our subset of UTokyo Ego-Surf (without retraining), and Table 2 summarizes the results. Though performing worse than on the IU ShareView dataset on which they were trained, the models still give reasonable results, indicating robustness even though the datasets are recorded by different cameras (Xiaoyi Yi vs. Panasonic HX-A500) and scenarios (indoor vs. outdoor).

Table 2. Experimental results of our models on UTokyo Ego-Surf dataset [40]

Network architecture					Evaluation		
	Backbone	*Streams*		Re-weighting	*Segmentation*	*Identification*	
		Image	Optical flow		IoU	mAP	ACC
Third-third	FCN	X	X	W/o	42.1	43.8	36.7
	FCN	X	X	Soft attention	**43.0**	**45.5**	**42.0**
Third-first	FCN	X	X	W/o	41.4	45.2	44.0
	FCN	X	X	Soft attention	**43.6**	**52.0**	**55.2**

Ablation Studies. We also test simpler variants of our technique. To evaluate our re-weighting method that incorporates estimated soft attention maps, we tried *not* re-weighting the spatial and temporal features and simply using `pool5` layer outputs. We also compare with the results of [14], which uses ground truth bounding boxes to "crop out" regions of interest. As shown in Tables 1 and 2, using re-weighting with soft attention not only outperforms for the matching task but also generates better segmentation maps. Our ablation study also tested the relative contribution of each of our motion and visual feature streams. As shown in Table 1, our dual-stream approach performs significantly better than either single-stream optical flow or visual information on both the third-third and third-first problems, evaluated on both segmentation and matching (Fig. 6).

Fig. 6. Sample results of the third-third and third-first problems, where two videos of each sample are from two synchronized wearable cameras. The color of person segmentation masks and camera views indicates the correspondences across different cameras. (Color figure online)

5 Conclusion

We presented a novel fully convolutional network (FCN) with Siamese and semi-Siamese structures for joint person instance segmentation and identification. We also prepared a new, challenging dataset with person pixel-level annotations and correspondences in multiple first- and third-person cameras. Our results demonstrated the effectiveness and robustness of our approach on joint person segmentation and identification. The results suggested that jointly inferring pixel-level segmentation maps and correspondences of people helps perform each individual task more accurately, and that incorporating both visual and motion information works better than either individually.

Although our results are encouraging, our techniques have limitations and raise opportunities for future work. First, the joint models assume people appear in every frame of the video, so that our approach will treat someone who disappears from the scene and then re-enters as a new person instance. While this assumption is reasonable for the relatively short video sequences we consider here, future work could easily add a re-identification module to recognize people who have appeared in previous frames. Second, the joint models perform a FCN forward pass for every individual person in each frame; future work could explore sharing computation costs to improve the efficiency of our method, especially for real-time applications. Lastly, we plan to further evaluate our approach on larger datasets including more diverse scenarios.

Acknowledgments. 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." We would like to thank Sven Bambach for assisting with dataset collection, and Katherine Spoon and Anthony Tai for suggestions on our paper draft.

References

1. http://www.gopro.com/
2. http://www.spectacles.com/
3. http://pytorch.org/
4. Ardeshir, S., Borji, A.: Ego2Top: matching viewers in egocentric and top-view videos. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 253–268. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_16
5. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv:1511.00561 (2015)
6. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: iCoseg: interactive co-segmentation with intelligent scribble guidance. In: CVPR (2010)
7. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017)
8. Chen, D.J., Chen, H.T., Chang, L.W.: Video object cosegmentation. In: ACMMM (2012)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611 (2018)
10. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: SegFlow: joint learning for video object segmentation and optical flow. In: ICCV (2017)

11. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: CVPR (2013)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
13. Endres, I., Hoiem, D.: Category independent object proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_42
14. Fan, C., et al.: Identifying first-person camera wearers in third-person videos. In: CVPR (2017)
15. Fu, H., Xu, D., Zhang, B., Lin, S.: Object-based multiple foreground video co-segmentation. In: CVPR (2014)
16. Girshick, R.: Fast R-CNN. In: CVPR (2015)
17. Guo, J., Cheong, L.F., Tan, R.T., Zhou, S.Z.: Consistent foreground co-segmentation. In: ACCV (2014)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv:1703.06870 (2017)
19. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: CVPR (2017)
20. Jang, W.D., Kim, C.S.: Online video object segmentation via convolutional trident network. In: CVPR (2017)
21. Jun Koh, Y., Kim, C.S.: Primary object segmentation in videos based on region augmentation and reduction. In: CVPR (2017)
22. Khoreva, A., Perazzi, F., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. arXiv:1612.02646 (2016)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
24. Lafayette Group: Survey of technology needs - body worn cameras. Technical report (2015)
25. LDV Capital: 45 billion cameras by 2022 fuel business opportunities. Technical report (2017)
26. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. arXiv:1611.07709 (2016)
27. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (2017)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
29. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: CVPR (2017)
30. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: NIPS (2015)
31. Pinheiro, P.O., Lin, T.-Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 75–91. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_5
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
33. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In: CVPR (2006)

34. Rubio, J.C., Serrat, J., López, A.: Video co-segmentation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7725, pp. 13–24. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37444-9_2
35. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
37. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. PAMI **36**, 1442–1468 (2014)
38. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. arXiv:1704.05737 (2017)
39. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. arXiv:1706.09364 (2017)
40. Yonetani, R., Kitani, K.M., Sato, Y.: Ego-surfing first-person videos. In: CVPR (2015)
41. Yoon, J.S., Rameau, F., Kim, J., Lee, S., Shin, S., Kweon, I.S.: Pixel-level matching for video object segmentation using convolutional neural networks. arXiv:1708.05137 (2017)
42. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. arXiv:1612.01105 (2016)
43. Zheng, K., et al.: Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In: ICCV (2017)
44. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv:1610.02984 (2016)