Is Sharing of Egocentric Video Giving Away Your Biometric Signature?

Daksh Thapar[1], Chetan Arora[2], and Aditya Nigam[1]

[1] Indian Institute of Technology Mandi, Mandi, India
[2] Indian Institute of Technology Delhi, New Delhi, India
https://egocentricbiometric.github.io/

Abstract. Easy availability of wearable egocentric cameras, and the sense of privacy propagated by the fact that the wearer is never seen in the captured videos, has led to a tremendous rise in public sharing of such videos. Unlike hand-held cameras, egocentric cameras are harnessed on the wearer's head, which makes it possible to track the wearer's head motion by observing optical flow in the egocentric videos. In this work, we create a novel kind of privacy attack by extracting the wearer's gait profile, a well known biometric signature, from such optical flow in the egocentric videos. We demonstrate strong wearer recognition capabilities based on extracted gait features, an unprecedented and critical weakness completely absent in hand-held videos. We demonstrate the following attack scenarios: (1) In a closed-set scenario, we show that it is possible to recognize the wearer of an egocentric video with an accuracy of more than 92.5% on the benchmark video dataset. (2) In an open-set setting, when the system has not seen the camera wearer even once during the training, we show that it is still possible to identify that the two egocentric videos have been captured by the same wearer with an Equal Error Rate (EER) of less than 14.35%. (3) We show that it is possible to extract gait signature even if only sparse optical flow and no other scene information from egocentric video is available. We demonstrate the accuracy of more than 84% for wearer recognition with only global optical flow. (4) While the first person to first person matching does not give us access to the wearer's face, we show that it is possible to match the extracted gait features against the one obtained from a third person view such as a surveillance camera looking at the wearer in a completely different background at a different time. In essence, our work indicates that sharing one's egocentric video should be treated as giving away one's biometric identity and recommend much more oversight before sharing of egocentric videos. The code, trained models, and the datasets and their annotations are available at https://egocentricbiometric.github.io/

1 Introduction

With the reducing cost and increasing comfort level, the use of wearable egocentric cameras is on the rise. Unlike typical point and shoot versions, egocentric cameras are usually harnessed on a wearer's head and allow to capture one's perspective. While the hands-free mode and the first-person perspective make these cameras attractive for adventure sports, and law enforcement, the always-on mode has led to its popularity for life-logging, and geriatric care applications. The broader availability of first-person

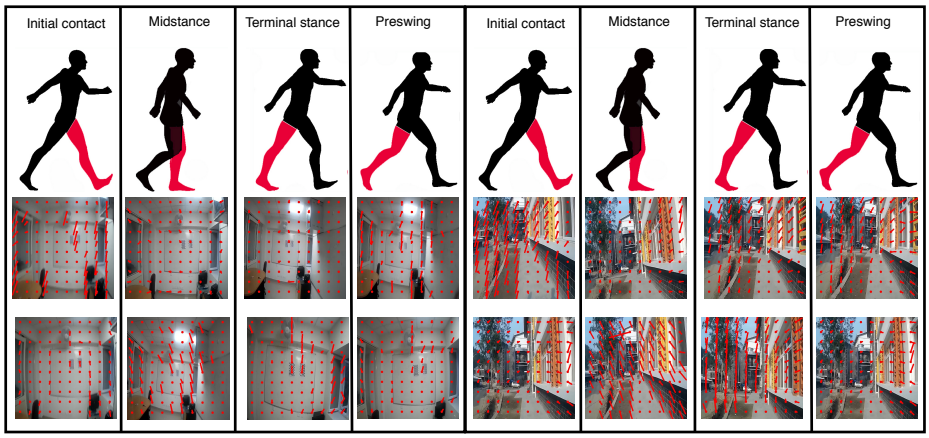

Fig. 1: The figure motivates the presence of the signal to identify a wearer from his/her first person video, even when a wearer is never seen in such videos. Here, we show the relation of optical flow vectors computed from egocentric videos with respect to the gait stance of the camera wearer for two different subjects. The first row shows an indicative third-person stance corresponding to the first person frame. Whereas, the second and third rows show the actual frames captured using the first person camera at the above-specified instance. We synchronized the first-person and third-person videos for purposes of this illustration. We overlay the optical flow vectors for the two different subjects on the respective RGB frames to illustrate the significant difference between the two subjects' optical flow. We draw the reader's attention to the large optical flow observed in the initial contact and pre-swing phases for the first subject (2nd row), whereas for the second subject (3rd row), large optical flow is observed in mid and terminal stance. In this work, we show that it is possible to extract and match the camera wearer's gait features from such optical flow in an open set recognition setting.

videos has attracted interest from computer vision community, with specialized techniques proposed for egocentric video summarization, temporal segmentation, and object, action, and activity recognition from first-person viewpoint [1–9].

One exciting feature of egocentric videos is that the camera wearer is never visible in them. This has led to many novel applications of egocentric videos, exploiting the unavailability of user identity in such videos. For example, Poleg et al. [10] has observed that since an egocentric camera is mounted on the wearer's head, the head motion cues are embedded in the observed motion of the captured scene. They have suggested to freely share the observed optical flow in the first-person video to be used as a temporally volatile, authentication signature of the wearer. Their premise is that the optical flow from egocentric videos does not reveal any private identifying information about the wearer. We speculate that the same belief may also be one reason for the wider public sharing of egocentric videos.

In this work, we take position exactly opposite Poleg et al. and posit that the head motion cues contain private information, but they are also highly correlated with the wearer's gait. Human gait is a well known biometric signature [11] and have been

traditionally extracted from the third-person view. Hence, through our exploration, we wish to draw the community's attention to a hitherto unknown privacy risk associated with the sharing of egocentric videos, which has never been seen in the videos captured from hand-held cameras. We focus on following specific questions: (1) Given a set of egocentric videos, can we classify a video to its camera wearer? (2) Given two anonymous videos picked from the public video-sharing website, can we say if the same camera wearer captured the two videos without seeing any other video from the wearer earlier? (3) What is the minimum resolution of the optical flow, which may be sufficient to recognize a camera wearer. Specifically, Poleg et al. has suggested the use of global optical flow as privacy safe, temporally volatile signatures. Is it possible to create a wearer's gait profile based on global optical flow? (4) How strong is the gait profile recovered from an egocentric video. Specifically, if there is a corresponding gait profile from a third-person point of view, say from a surveillance camera, is it possible to match the two gait profiles and verify if they belong to the same person? Our findings and specific contributions are as follows:

1. We analyze the biomechanics of a human gait and design a deep neural network, called EgoGaitNet (EGN) to extract the wearer gait from the optical flow in a given egocentric video. In a closed-set setting, when the set of camera wearers are known a-priori, we report an accuracy of 92.5% on the benchmark dataset.
2. We also explore the open-set scenario in which the camera wearers are not known a-priori. For this we train the EGN with ranking loss, and report an Equal Error Ratio (EER) of 14.85% on the benchmark egocentric dataset containing 32% subjects.[3]
3. We tweak the proposed EGN architecture to work with sparse optical flow and show that even with global optical (2 scalars per frame corresponding to the flow in x and y directions), one can identify the camera wearer with a classification accuracy of 77%.
4. While, the three contributions above give a strong capability to recognize a wearer in the closed set setting or identify other egocentric videos from the wearer in a closed set scenario, and they do not reveal the identity/face of the wearer. We propose a novel Hybrid Symmetrical Siamese Network (HSSN), which can extract the gait from third person videos and match it with the gait recovered from EGN. It may be noted that the first-person and third-person videos for this task may have been captured at a completely different time and context/background. Since there is no benchmark dataset available with the corresponding first person and third person videos of the same person, we experiment with dataset generated by us and report an EER of 10.52% for recognizing a wearer across the views.
5. We contribute two new video datasets. The first dataset contains 3.1 hours of first-person videos captured from 31 subjects with a variety of physical build in multiple scenarios. The second contains videos captured from 12 subjects for both first-person and third-person setting. We also use the datasets to test the proposed models on the tasks as described above.

[3] To put the numbers in perspective, for the gait based recognition from third-person views, state of the art EER (on a different third-person dataset) is 4%

2 Related Work

Gait Recognition from Third Person Viewpoint: We note that there has been a significant amount of work on gait recognition from third-person videos that use the trajectory of the limbs [11], joints [12], or silhouette [13–16]. The focus of our work is on extracting gait from egocentric videos. Hence, these works are not directly relevant to the proposed work. However, they serve to support our hypothesis that the motion of the limbs (or the gait in general) also affects the motion of the head, which ultimately gets reflected in the observed optical flow in an egocentric video. Below, we describe only the works related to wearer recognition from first person videos.

Wearer Recognition from Egocentric Videos: Tao et al. [17] have shown that gait features could also be captured from wearable sensors like accelerometer and gyroscope. Finocchiaro et al. [7] estimated the height of the camera from the ground using only the egocentric video. They have extended the original network model proposed in [18] to estimate the height of the wearer, with an Average Mean Error of 14.04 cm over a range of 103 cm of data. They have reported the classification accuracy for relative height (tall, medium, or short) at 93.75%. Jian and Graumann [19], have infered the wearer's pose from the egocentric camera. They have given a learning-based approach that gives the full body 3D joint positions in each frame. The technique uses both the optical flow as well as static scene structures to reveal the viewpoint (e.g., sitting vs. standing).

Hoshen and Poleg [9] have shown that one could identify a camera wearer in a closed set scenario, based on shared optical flow from his/her egocentric camera. They have trained a convolutional neural network using the block-wise optical flow computed from the consecutive egocentric video frames and showed a classification accuracy of 90%. However, their work assumes critical restrictive assumptions relevant to privacy preservation. First, their framework requires many more samples from the same camera wearer to train the classifier for the identification task. The requirement is unrealistic for anonymous videos typically posted on public video sharing websites, with non-cooperating camera wearers. Secondly, original head motion signatures suggested by Poleg et al. [10] were computed by averaging the optical flows (resulting in 2 scalars per frame), whereas Hoshen and Poleg have used full-frame optical flows. Thirdly, since the work only matches the first-person to first-person videos, the true identity (or face) of the wearer is never revealed.

Wearer Recognition using Egocentric and Third-Person Videos: There have been techniques that assume the presence of another third-person camera (wearable or static) present simultaneously to the egocentric camera and aim to identify the camera wearer in the third-person view. In [20], authors exploit multiple wearable cameras sharing fields-of-view to measure visual similarity and identify the target subject. Whereas, in [21], the common scene observed by the wearer and a surveillance camera has been used to identify the wearer. Other works compute the location of the wearer directly [22, 23] or indirectly (using gaze, social interactions, etc.) [24, 25] which is then used to identify the wearer. Unlike our approach, all these techniques assume the presence of the third-person camera view within the same context and time, which though exciting, does not lead to mounting privacy attacks, which is the focus of our work.

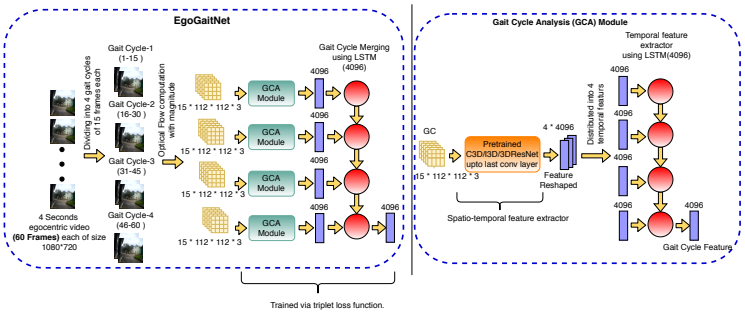

Fig. 2: The network architecture for the proposed first person verification network EgoGaitNet.

3 Proposed Approach

In traditional gait recognition systems, where the subject is visible in the video, the salient features are the limbs' movement. However, in the case of egocentric videos, the subject is not visible, thus ruling out traditional gait recognition methods. Hence for doing so, we look into the biomechanics of gait. A gait cycle consists of multiple gait cycle segments/phases (GCS). Transitioning from these segments causes the overall motion of the body, and hence the correlated motion of the camera harnessed on the head of the camera wearer. Thus, assuming a stationary background, optical flow provides us with information about the GCS transitions.

3.1 Extracting Gait Signatures from Egocentric Videos

In order to extract the gait features from egocentric videos, we propose EgoGaitNet (EGN) model. The architecture of EGN is shown in Figure 2. We have extracted frames from the videos at 15FPS. We resize each frame to the size of $112 \times 112 \times 3$ and divide each video into clips of 4 seconds (*i.e.* 60 frames). We compute dense optical flows between each consecutive frame using Gunner Farneback's algorithm [26]. Hence, for each frame, we get $112 \times 112 \times 2$ optical flow matrix, where the channels depict the flow at each point in x and y directions. We compute the magnitude of flow at each point and append the magnitudes with the flow matrix to make it $112 \times 112 \times 3$ optical flow matrix. We hypothesize that each 4-second clip of size $60 \times 112 \times 112 \times 3$ contains the camera wearer's gait information embedded in the optical flow transitions. We further assume that one gait cycle (half step while walking) is 15 frames (1 second) and divides each clip into four parts of 15 frames each. Our choice of gait cycle time (1 sec) and the number of gait cycles sufficient to extract gait information (4) is inspired by similar work in third person gait recognition [16].

To extract the gait cycle feature from each of the segmented clips, we propose a Gait Cycle Analysis (GCA) module (as shown in Figure 2). It consists of a pre-trained spatio-temporal (3D CNN) feature extractor for extracting the intra-gait cycle segment information. We use the features from the last convolutional layer from the 3D CNN and reshape the spatial channels to 1D and obtain a 4×4096 feature vector representation for inputting to the GCA module. Note that the feature vector is obtained from each

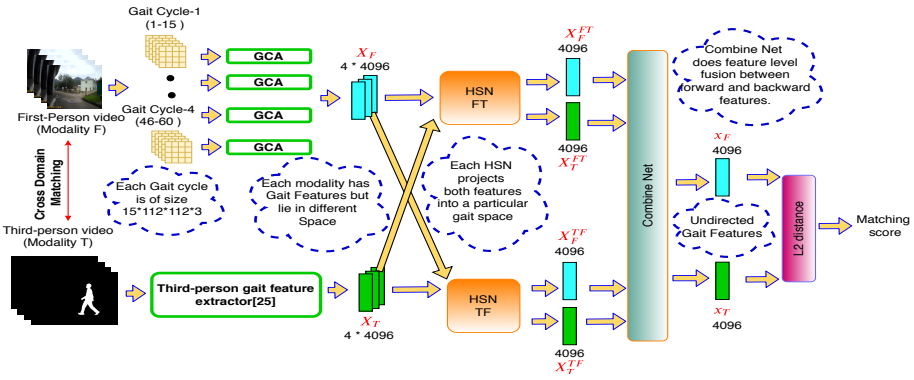

Fig. 3: The network architecture used for the proposed first person to third person Matching Network. The first-person gait feature extractor is taken from the proposed EgoGaitNet. The third-person gait feature extractor is taken from [16].

gait cycle of 15 frames. To further learn features specific to first-person videos, we split the temporal features to make it four vectors of 4096 dimension each. These features are inputted to a temporal feature extractor (LSTM) having 4096 recurrent dimensions (Figure 2(right side)), and giving us a single 4096 dimensional feature vector representation of a gait cycle. We use four gait cycles to extract the gait signature of a wearer. To learn inter gait cycle relationships, we pass the 4096 dimensional features corresponding to a gait cycle to a gait cycle merging process, which is an LSTM based architecture with 4096 recurrent dimensions(Figure 2 (left side)). The output of the LSTM gives us a feature representation of 4096 containing the gait signature of a wearer.

In our experiments, we have done an ablation study to understand the effect of 3D CNN architecture on the performance of EGN. We give the details in the experiment section later as well as in the Supplementary Section.

3.2 Recognizing Wearer from First Person Video

To recognize a wearer from her/his first-person video, we train the EGN network for two scenarios. The first one is closed set recognition, in which the network has already seen the data of every subject during training (classification mode). The second one is the open set scenario in which the testing is done on subjects that have not been seen by the network (metric learning mode).

Closed Set Recognition For closed set recognition, we train the EGN as a classifier for the camera wearer task. This task is not the prime focus of the work but has been done to compare the performance of the architecture with the current state of the art [9]. A classification layer is added at the end of EGN, and the network is trained using a categorical cross-entropy loss function. To perform the verification task, we have trained our network in a one vs. rest fashion as done by [9] for the fair comparison. We have used ADAM optimizer with a learning rate of 0.0005. We apply dropouts with the dropping

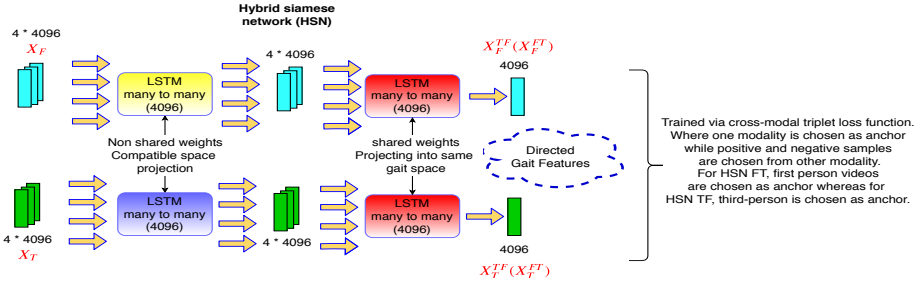

Fig. 4: The network architecture used for the proposed Hybrid Siamese Network (HSN).

probability of 0.5 over the fully connected layer and LSTM except for the classification layer for better regularization. ReLU activation has been used in all the layers except LSTM, where Tanh activation is used. The output of the classification layer is normalized using the softmax activation to convert the output to a pseudo probability vector.

Open Set Recognition To perform open set recognition of camera wearer from egocentric videos, we train the EGN network to learn a distance metric between two head motion signatures using triplet loss function. This enables the network to learn a suitable mapping between a sequence of optical flow vectors to a final feature vector (a point in the embedding space defined by the output layer of the network), such that the L_2 distance between the embeddings of the same camera wearer is small and distance between embeddings of different wearers is large. For efficient training of EGN, we apply semi-hard negative mining and dynamic adaptive margin in triplet loss as described by [27]. We use a step-wise modular training procedure to streamline the training of EGN, as described further. First, we train only the 3D CNN, then freeze the 3D CNN and only train the LSTM of GCA module, followed by freezing the GCA and training the gait cycle merging module. Finally, we fine-tune the complete EGN for the first-person recognition task via. triplet loss. Given two video segments i and j, the network must produce an embedding Θ, such that if i and j belong to the same subject, then $L_2(\Theta^i, \Theta^j)$ should tend to 0, otherwise, $L_2(\Theta^i, \Theta^j) \geq \beta$, where β is the margin. The loss has been defined over 3 embeddings: (1) Θ^i: embedding of an anchor video, (2) Θ^{i^+}: embedding of another video from the same wearer, and (3) Θ^{i^-}: embedding of a video from another arbitrary wearer. Formally: $\mathcal{L}(i, i^+, i^-) = \max(0, (\Theta^i - \Theta^{i^+})^2 - (\Theta^i - \Theta^{i^-})^2 + \beta)$ We sum the loss for all possible triples (i, i^+, i^-) to form the cost function J, which is minimized during the training of the proposed architecture: $J = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(i, i^+, i^-)$

3.3 Extracting Gait from Sparse Optical Flow

One of the questions that we seek to answer from this work is whether the original head motion signatures proposed by [10], which contain only two scalar values per frame, can reveal wearer's identity? A naive way to do this would be to compute the flow at appropriate spatial resolution and follow the same train and test procedure as done for the dense optical flow. However, given the limited information offered by the global optical

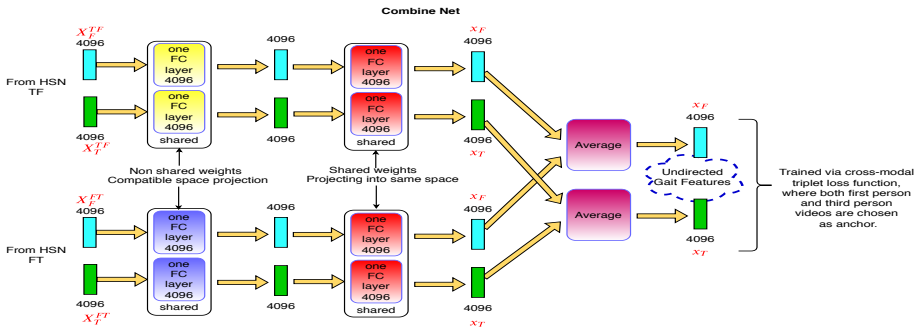

Fig. 5: The network architecture used for the proposed Combine Net.

flow, we observe severe over-fitting using the naive approach. One possible solution is to use a pre-trained network. However, here we propose a simple but extremely effective workaround, as described below.

Given a desired optical flow resolution of $x \times y$, we divide each frame into a same-sized grid. We compute the optical flow per cell independently, which is then given as an input to the EGN. However, instead of giving the optical flow of size $x \times y$, we copy the optical flow, coming from each cell to every pixel underlying the cell. This is equivalent to up-sampling the optical flow image using the nearest neighbor technique. We give this up-sampled optical flow as input to the EGN network. Matching the size of the optical flow vector allows us to use a pre-trained network at a much higher resolution and then only fine-tune it on the lower resolution flow as required. As shown in the experimental section, the simple workaround gives us a reasonably good accuracy and allows us to claim the wearer recognition capability even with the frame-level global optical flow. We understand that more sophisticated methods for optical flow up-sampling, including learnable up-sampling, could have used but have not been explored in our experiments.

3.4 Recognizing Wearer from Third Person Video

The main goal of this paper is to match the gait profile extracted from an egocentric video to the gait profile extracted from a third-person video, which allows us to track a camera wearer based on his/her egocentric video alone. To achieve this, we propose two deep neural network architectures called Hybrid Siamese Network (HSN) and Combine Net. The overall pipeline is shown in Figure 3.

We first extract the third-person gait features using a state-of-the-art third-person gait recognition technique. We have used [16] in our experiments; however, any other similar technique could have been used as well. The input to [16] is 60 RGB frames, divided into four gait cycles as in the case of EGN (which only took optical flow and not RGB). The output of [16] is the gait feature vector of 4×4096 dimension denoted as X_T.

For extracting the gait features from egocentric videos, we use the GCA module described in the EGN and extract a 4096 dimensional feature vector corresponding to each gait cycle segment/phase. Hence for four segments, we get a feature vector of size 4×4096, denoted as X_F in our model. Both X_T and X_F vectors contain the gait

information of the camera wearer, but they lie in entirely different spaces as they are coming from very different viewing modalities. To make them compatible, we pass them through the proposed Hybrid Siamese Network (HSN), which is trained to learn a mapping that can project the two vectors into the same gait space.

The HSN is trained using cross-modal triplet loss function (described later below) in which anchors are coming from one modality, and positives and negatives are coming from other modality. This adds a directional attribute to the HSN, causing the metric function learned by HSN to be asymmetric. Hence, we train two HSN networks, one HSN-TF, where the anchor videos are chosen from third-person, and the second HSN-FT, where the anchors are chosen from egocentric videos. The output embeddings from the two HSN are denotes as X_F^{TF} (X_T^{TF}) and X_F^{FT} (X_T^{FT}) respectively, where subscript T indicates the third person, and F denotes the first-person features.

One way to create an undirectional metric is to merge the matching scores obtained from both HSN TF and HSN FT. Another way is to perform a feature level fusion between the gait features extracted from both the HSN's. The four features (namely $X_F^{\text{TF}}, X_T^{\text{TF}}, X_F^{\text{FT}}, X_T^{\text{FT}}$) transformed by HSN-TF and HSN-FT are not compatible for direct fusion. Hence we propose another neural network *CombineNet* to fuse the features. The details of HSN and Combine Net are given below.

HSN Architecture: As shown in Figure 4, we first pass both 4×4096 dimensional X_T and X_F vectors through a many-to-many LSTM. The weights of the LSTM for the two vectors are not tied. We follow this up with another many-to-one LSTM network, which transforms the two vectors to a common 4096 dimensional feature space. Both the LSTM layers have a recurrent dimension of 4096. As described earlier, we train two HSNs: HSN-TF, and HSN-FT, with different anchor modalities.

CombineNet Architecture: We combine the asymmetrical features received from the two HSN's using the CombineNet. The Combine Net receives four distinct features of size 4096 (X_F^{TF}, X_T^{TF}, X_F^{FT}, and X_T^{FT}). As shown in Figure 5, first, a non-shared fully connected layer (FC) is applied over the features. However, since X_F^{TF} and X_T^{TF} are already in the same feature space and so are X_F^{FT} and X_T^{FT}, this FC layer is shared among them. Finally, to transform all the features into the same space, a shared FC layer is applied over the four feature vectors. As the features are now in the same space, both the first-person features and third-person features are averaged for fusion to provide the undirected gait features (X_F, X_T). The training of CombineNet is explained below.

Training procedure using Cross-modal triplet loss: Both the HSN and CombineNet are trained using cross-modal triplet loss function as described for EGN. However, the selection of triplets is done differently, to learn the desired metrics in both cases. Since the loss function here deals with two modalities, the anchor video is selected from the first modality, whereas the positive and negative videos are selected from the second modality. Despite the different modalities, anchor and positive must belong to the same subject, whereas anchor and negative should belong to different subjects. We train the HSN-FT with cross-modality triplet loss function by selecting the anchors from the first-person videos, whereas for HSN-TF, we select the anchors from third-person videos. We finally freeze the two HSNs and train the CombineNet by selecting both first-person and third-person videos as anchors. For the triplets having first-person videos as the anchor, the positive and negative videos are selected from third-person. Whereas, for

triplets having third-person videos as the anchor, the positive and negative videos are selected from the first-person videos.

4 Datasets Used

First Person Social Interactions dataset (FPSI) [28]: FPSI is a publicly available dataset consisting of video captured by 6 people wearing cameras mounted on their hat, and spending their day at Disney World Resort in Orlando, Florida. We have used only walking sequences from this dataset, where the gait profile of the wearer is reflected in the observed optical flow in the video. *Further, we have tested in the unseen sequence mode where morning videos have been used for training and evening ones for testing.*

Egocentric Video Photographer Recognition dataset (EVPR) [9]: It consists of videos of 32 subjects taken for egocentric first-person recognition. The data is made using two different cameras. *In our experiments, we use videos captured from one of the cameras for training while the remaining videos have been used for testing.*

Our Dataset for Wearer Recognition in Egocentric Video (IITMD-WFP): We also contribute a new egocentric dataset consisting of 3.1 hours of videos captured by 31 different subjects. We introduced variability by taking videos on two different days for each subject. *To maintain testing in unseen sequence settings, we have used the videos from one of the days for training and other for testing.* To introduce further variability in the scene, we have captured in two scenarios: indoor and outdoor, and refer to the respective datasets as DB-01 (indoor), and DB-02 (outdoor). To make sure that the network does not rely on the scene-specific optical flow, we have captured video for each subject in a similar scenario. For both the indoor and outdoor datasets, the path taken by each of the subjects was predefined and fixed, and the videos were captured using the SJCAM 4000 camera. For the biometric applications, it is especially important to show the verification performance over many subjects, since the performance metrics typically degrade quickly with dataset size, due to an exponential increase in the imposter matchings. Hence, we create a combined dataset by merging DB-01, and DB-02, and refer to it as DB-03. We combine EVPR, FPSI, and DB-03 and refer to it as DB-04. After merging, the combined DB-04 dataset contains 69 subjects.

Our Dataset for Wearer Recognition in Third Person Video(IITMD-WTP): To validate our first-person to third-person matching approach, we have collected a dataset containing both third-person and first-person videos of 12 subjects. The third-person videos are captured using Logitech C930 HD camera, whereas the first-person videos from SJCAM 4000 camera. The axis of the third person camera is perpendicular to the walking line of each subject. The total video time of IITMD-WTP dataset is 1 hour 3 minutes having 56,700 frames. For the open-set verification, we use six subjects for training, and remaining unseen subjects for testing. For closed-set analysis, the first five rounds have been used for training and the last five for testing. The representative images and detailed statistics for each dataset have been given in the supplementary material.

5 Experiments and Results

5.1 Hyper-parameters and Ablation Study

Our gait feature extractor module (c.f. Section 3.1), uses 3D CNNs for finding spatio-temporal optical flow patterns correlated with wearer's gait. We have performed a rigorous ablation study using different network backbones: C3D [29], I3D [30], and 3D-ResNet [31], which C3D performs the best and has been used for further analysis. We have also compared our architecture with various combination style for merging features from individual gait cycles, and have finally chosen uni-directional LSTM with four gait cycle input. The detailed ablation study is given in the supplementary material.

5.2 Wearer Recognition in Egocentric Videos

We first analyze our system for recognition capability in egocentric videos. We test in both closed-set (wearers are known and trained for during training) and open-set (wearers are unseen during training) scenarios. Table 1, columns 2–5, compare the performance with [9] for closed-set scenario, in terms of classification accuracy (CA) and Equal Error Rate (EER). The values for EVPR and FPSI datasets have been taken from their paper, whereas for others, we computed the results using the authors' code. It is easy to see that for each dataset, our system improves [9].

Table 1: Comparative analysis of our system with [9] for wearer recognition in egocentric videos. While [9] works only for closed-set scenarios, our system can work both in closed-set as well as open-set scenarios. CA, EER, and DI denote the classification accuracy, Equal Error Rate, and Decidability Index respectively in percentage. Higher CA and lower EER is better.

Dataset	Closed Set Analysis				Open Set Analysis		
	[9]		EgoGaitNet		EgoGaitNet		
	CA	EER	CA	EER	EER	CRR	DI
FPSI	76.0	20.34	82.0	19.71	–	–	–
EVPR	90.0	11.3	92.5	9.8	14.35	68.12	1.95
DB-01	95.1	4.38	99.2	2.79	6.43	83.67	2.35
DB-02	93.7	5.03	97.3	3.81	8.23	82.77	2.15
DB-03	94.0	5.72	98.7	4.35	9.39	80.56	2.02
DB-04	85.6	19.64	89.9	15.44	20.61	62.17	0.27

For the open-set scenario, we establish the validity of the learned distance by our approach using the decidability index (DI) and rank one correct recognition rate (CRR). Decidability index [32] is a commonly used score in biometrics to evaluate the discrimination between genuine and impostor matching scores in a verification task. The score is defined as: $DI = \frac{|\mu_g - \mu_i|}{\sqrt{(\sigma_g^2 + \sigma_i^2)/2}}$, where $\mu_g(\mu_i)$ is the mean of the genuine (impostor) matching scores, and $\sigma_g(\sigma_i)$ is the standard deviation of the genuine (impostor) matching scores. A large decidability index indicates strong distinguishability characteristics, i.e., high recognition accuracy and robustness. The open-set analysis is not performed over the FPSI dataset as the number of subjects is very small. For the rest of the datasets, half of the subjects from each of the individual datasets were taken for training and rest half for testing. *We believe that open set analysis mimics much more practical attack*

scenarios with uncooperative wearers, which have not been seen at the train time, but we would still like to find other videos captured by them. From Table 1, columns 6–8, it is apparent there is only a minor decrease in the performance of the network compared to the closed-set scenario, which still has a very low error rate. Hence, we can conclude that the proposed model can verify unseen camera wearers also.

The ROC curves for our approach on various datasets are shown in Figure 6. It can be seen that performance over the EVPR dataset is better than FPSI. This may be due to the fact that the activities performed by the subjects in FPSI are varied, whereas EVPR contains only walking sequences. We also show the curves for a much larger DB-04 dataset to establish robust recognition performance even with a large number of subjects, indicating significant privacy risk associated with sharing egocentric videos.

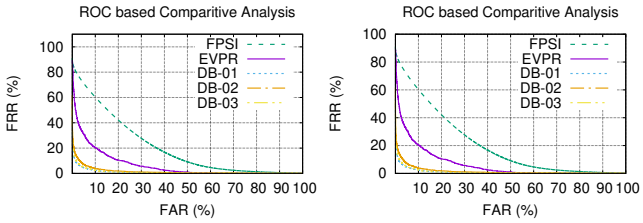

Fig. 6: Left: ROC curves of proposed system on various datasets. Right: The ROC curves for individual datasets when trained and tested on a combined DB-04 dataset. Note that in the combined dataset, an imposter matching increase exponentially and the stable performance of our approach show the technique's strength.

Wearer Height Analysis: A doubt regarding our system's good performance can be that it is differentiating based on the wearer's height. We did a limited analysis to verify that there is no such over-fitting in the system. We segregated three subjects of similar height and tested our model on just those 3. For those three subjects of similar height, we got an equal error rate of only 2.03%, showing us that the proposed model can differentiate successfully between the subjects despite having the same height.

Effect of Spatial Resolution of Optical Flow: The experiments so far have been done on dense optical flow. However, one of the questions that we seek to answer is whether the original head motion signatures proposed by [10], which contain only two scalar values per frame, can reveal wearer's identity. As explained in Section 3.3, we have created a simple workaround by simply up-sampling the optical flow given at a lower resolution to the original resolution using the nearest neighbor approach. This allows us to use a pre-trained network trained with dense optical flow, and fine-tune it with the up-sampled flow. As done in the earlier experiments, we have been careful in separating the unseen wearers at an early stage, which are never shown to the network, either in pre-training or fine-tuning stage. The performance over different sizes of optical flow input is shown in Figure 7. In the figure, x-axis maps to the number of optical flow values in rows and columns. 112 refers to dense optical flow, and 1 refers to the case where the whole optical flow was globally averaged to a single vector as

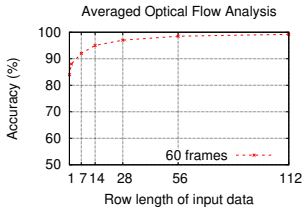

Fig. 7: Performance of our classifier on averaged optical flow, as used in [10].

in [10]. We get a high identification accuracy of 92% when dealing with only a 7×7 optical flow matrix. Even with a single global flow vector, we achieve an accuracy of 84%, indicating that even averaged head motion signatures are enough to recover the gait profile and recognize a camera wearer.

5.3 Wearer Recognition in Third Person Videos

Taking the privacy attack one step further, in this section we show that using HSN proposed in this paper; it is possible to match the gait profile extracted from egocentric videos, even with the one extracted from regular third person videos. For this, we perform experiments on the IITMD-WTP dataset under both closed-set and open-set protocols. For the former, the first five walks of every subject were used for training and the last 5 for testing the system. Whereas in the open-set scenario, only the first six subjects were used for training, and the system has been evaluated on the last six unseen subjects. Table 2 shows the results. We report the scores for both HSN-FT and HSN-TF and the CombineNet, which fuses the features from HSN-FT and HSN-TF.

Table 2: Performance analysis for recognizing a wearer in a third person video. The score fusion approach refers to classifying/verifying a sample by average of HSN-FT and HSN-TF scores.

Model	**Closed-set Analysis**			**Open-set Analysis**		
	EER	**CRR**	**DI**	**EER**	**CRR**	**DI**
HSN-FT	11.45	72.46	1.68	15.84	69.27	1.02
HSN-TF	11.02	75.78	1.70	15.36	69.75	1.02
Score Fusion	8.76	76.24	1.72	13.68	71.65	1.05
CombineNet	9.21	79.86	1.71	14.02	73.36	1.06

5.4 Model Interpretability

We have tried to analyze our model to understand if it can learn the wearer's gait cues. We have visualized the activations of 3-D convolutional filters of the first layer of our model. We extract activations from the optical flow input of 2 different subjects and compare the filters having maximum activation corresponding to the two subjects. Figure 8 shows two such filters for subjects 1 and 2. Recall that the first layer in our model is a 3-D CNN layer with the kernel of size $3 \times 3 \times 3$, and input to the network is of the size $15 \times 112 \times 112 \times 3$, where 3 channels correspond to optical flow in x, and y directions, and its magnitude. Recall that we take a gait cycle of 15 frames. The output of the first layer filter is of size $15 \times 112 \times 112$. Figure 8 shows the activations for 10 frames. In the first and last columns, we have shown the 3rd person gait respective to each of the subjects. The second column from the left and right shows the corresponding 1st person video frame. The 3rd and 4th columns show the activations corresponding to each subject's optical flow input from filter 1. The 5th and 6th columns show the activations corresponding to each subject's optical flow input from filter 2. These activations have been overlaid with the input optical flow vectors. Note that the RGB frames are only for illustration purposes, whereas the proposed model only uses the optical flow.

We observe that filter activations are mostly synchronized with the gait phases. For example, filter 1 activations are high when subject 1 moves his/her one leg while the

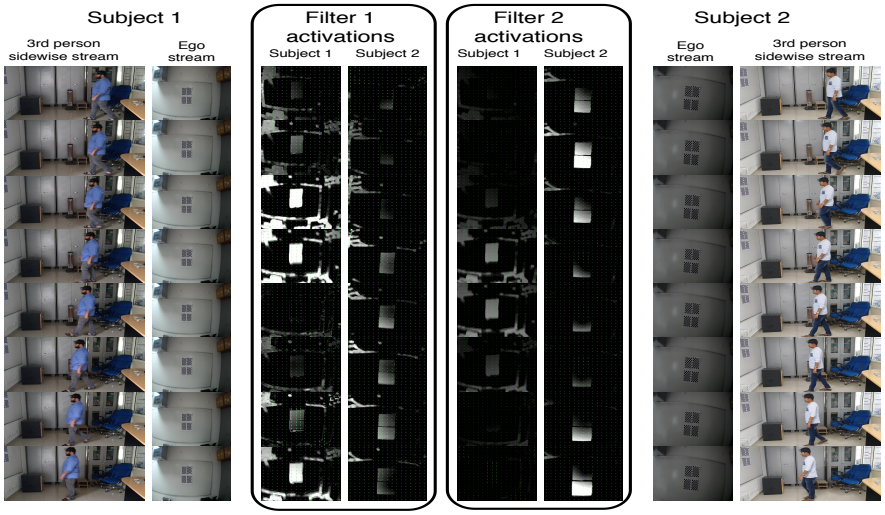

Fig. 8: Filter activations of filter 1 and 2 of first layer for two different subjects with same background and external surroundings. Please refer to the paper text for the details. We speculate on the basis of the visualization that initial layers of the proposed network are temporally segmenting a gait phase. This effectively allows the following layers to learn gait specific features.

other leg is stationary. We observe that similar movement of subject 2 is captured by filter 2. We speculate that the initial layers of our network are trying to segment the gait and trigger on a specific gait phase, which then is combined into distinguishing features by the later layers. Moreover, it can also be seen that the activations are high in the spatially salient parts of the image. In these parts, one can capture useful features for computing optical flows. Since gait features are present in the transition of optical flows from one frame to another, we believe that the network captures the gait features only and not overfitting over the structure of the input scene.

6 Conclusion and Future Work

In this paper, we have tried to create a new kind of privacy attack by using the head-mounting property of wearable egocentric cameras. Our experiments validate a startling revelation that it is possible to extract gait signatures of the wearer from the observed optical flow in the egocentric videos. Once the gait features are extracted, it is possible to train a deep neural network to match it with the gait features extracted from another egocentric video, or more surprisingly, even with the gait extracted from another third person video. While the former allows us to search other first person videos captured by the wearer, the latter completely exposes the camera wearer's identity. We hope that through our work, we will be able to convince the community that sharing egocentric videos should be treated as sharing one's biometric signatures, and strong oversight may be required before public sharing of such videos. To extend this work in the future, we would like to investigate other body-worn devices' ability to extract gait of the wearer.

References

1. Huang, Y., Cai, M., Li, Z., Sato, Y.: Mutual context network for jointly estimating egocentric gaze and actions. arXiv preprint arXiv:1901.01874 (2019)
2. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled egocentric video summarization via constrained submodular maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2235–2244
3. Kopf, J., Cohen, M.F., Szeliski, R.: First-person hyper-lapse videos. ACM Transactions on Graphics (TOG) **33**(4) (2014) 78
4. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2010) 3137–3144
5. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 2847–2854
6. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR 2011, IEEE (2011) 3241–3248
7. Finocchiaro, J., Khan, A.U., Borji, A.: Egocentric height estimation. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2017) 1142–1150
8. Yagi, T., Mangalam, K., Yonetani, R., Sato, Y.: Future person localization in first-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7593–7602
9. Hoshen, Y., Peleg, S.: An egocentric look at video photographer identity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4284–4292
10. Poleg, Y., Arora, C., Peleg, S.: Head motion signatures from egocentric videos. In: Asian Conference on Computer Vision, Springer (2014) 315–329
11. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception & psychophysics **14**(2) (1973) 201–211
12. Carter, J.N., Nixon, M.S.: Measuring gait signatures which are invariant to their trajectory. Measurement and Control **32**(9) (1999) 265–269
13. Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N.P., Roy-Chowdhury, A.K., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Transactions on image processing **13**(9) (2004) 1163–1173
14. Man, J., Bhanu, B.: Individual recognition using gait energy image. IEEE transactions on pattern analysis and machine intelligence **28**(2) (2006) 316–322
15. Hofmann, M., Rigoll, G.: Exploiting gradient histograms for gait-based person identification. In: Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE (2013) 4171–4175
16. Thapar, D., Jaswal, G., Nigam, A., Arora, C.: Gait metric learning siamese network exploiting dual of spatio-temporal 3d-cnn intra and lstm based inter gait-cycle-segment features. Pattern Recognition Letters **125** (2019) 646–653
17. Tao, W., Liu, T., Zheng, R., Feng, H.: Gait analysis using wearable sensors. Sensors **12**(2) (2012) 2255–2283
18. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact cnn for indexing egocentric videos. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2016) 1–9
19. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 3501–3509

20. Fan, C., Lee, J., Xu, M., Singh, K.K., Lee, Y.J., Crandall, D.J., Ryoo, M.S.: Identifying first-person camera wearers in third-person videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
21. Ardeshir, S., Borji, A.: Ego2top: Matching viewers in egocentric and top-view videos. In: ECCV (5). Volume 9909 of Lecture Notes in Computer Science., Springer (2016) 253–268
22. Hesch, J.A., Roumeliotis, S.I.: Consistency analysis and improvement for single-camera localization. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE (2012) 15–22
23. Murillo, A.C., Gutiérrez-Gómez, D., Rituerto, A., Puig, L., Guerrero, J.J.: Wearable omnidirectional vision system for personal localization and guidance. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE (2012) 8–14
24. Park, H.S., Jain, E., Sheikh, Y.: 3d social saliency from head-mounted cameras. In: Advances in Neural Information Processing Systems. (2012) 422–430
25. Soo Park, H., Jain, E., Sheikh, Y.: Predicting primary gaze behavior using social saliency fields. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 3503–3510
26. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis, Springer (2003) 363–370
27. Thapar, D., Jaswal, G., Nigam, A., Kanhangad, V.: PVSNet: Palm vein authentication siamese network trained using triplet loss and adaptive hard mining by learning enforced domain specific features. arXiv preprint arXiv:1812.06271 (2018)
28. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1226–1233
29. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2015) 4489–4497
30. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6299–6308
31. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2018) 6546–6555
32. Ravikanth, C., Kumar, A.: Biometric authentication using finger-back surface. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2007) 1–6