Is Sharing of Egocentric Video Giving Away Your Biometric Signature?

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

[1] Indian Institute of Technology Mandi, Mandi, India
d18033@students.iitmandi.ac.in, aditya@iitmandi.ac.in
[2] Indian Institute of Technology Delhi, New Delhi, India
chetan@cse.iitd.ac.in

1 Datasets Used

In this section we have described all the datasets used in the proposed work. In order to perform first person recognition, two publically available datasets FPSI and EVPR have been used along with our indigenous dataset.

1.1 Datasets for first-person recognition

First Person Social Interactions dataset (FPSI) [1]: 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) [2]: 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 ((IITMD-WFP): To validate our technique in multiple scenarios, we have created a new dataset of our own as well. The dataset consists of 3.1 hours of videos captured by 31 different subjects. The videos have been taken on two different days for each subject, introducing variability in the dataset. *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. The dataset is broken down into three parts: (1) (DB-01) The dataset collected indoor. (2) (DB-02) The dataset collected outdoor. (3) (DB-03) The combined indoor and outdoor dataset. 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 experimenting in a much larger scenario, we also test our system with a combined dataset, merging all the above datasets. We refer to this as DB-04, which consists of

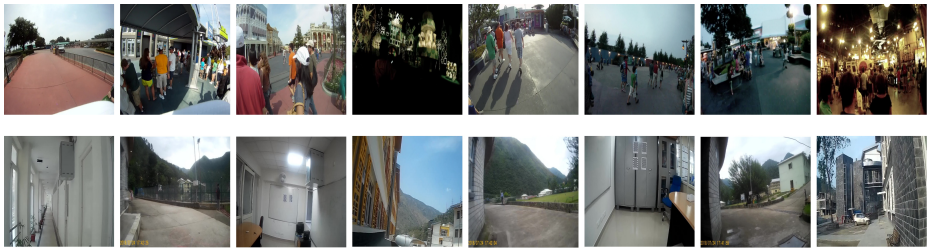


Fig. 1: Representative images for the 2 datasets, FPSI (First person social interaction) dataset and our inhouse collected dataset (IITMD-FPR). The first row shows the different scenerios of FPSI dataset and the second row shows the different scenerios of IITMD-FPR dataset.

Table 1: Dataset Specifications. Every clip of 4 seconds duration is one data input.

Dataset	Subjects	Training data	Testing data	Genuine Matching	Imposter Matching
FPSI	6	5693	7268	5255000	26275000
EVPR	32	1652	1708	7680	53760
DB-01	31	2012	2384	30526	5625489
DB-02	31	2276	2345	35621	7526489
DB-03	31	4288	4729	50498	10256412
DB-04	69	9981	11997	262775510	131376005

sequences from 69 distinct subjects. Figure 1 shows representative images from each dataset. We do not include images from EVPR since the authors have only made available optical flow data and not the RGB images.

1.2 Dataset for the first-person to third-person matching

Our Dataset ((IITMD-WTP): To the proposed first person to third person matching framework, we have collected a dataset that contains both third-person as well as 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. We have captured ten rounds of videos for each subject, and we use the first five rounds for training and the last five for testing.

2 Experiments and Results

To validate our proposed approach, we have performed three types of experimental analysis. First is the analysis of the classifier. The classifier is validated for classification and verification tasks according to the procedure used by Hoshen and Peleg [2]. The second is the analysis of the learned distance metric for both recognition as well as verification problems. Here, we test the proposed system in an open set recognition

Table 2: Model nomenclature and details (FE refers to Feature Extractor).

Proposed Model	Gait cycle analysis				Gait cycle merging	Training Loss
	Spatio-temporal FE	Temporal FE	Model Complexity	Multi Scale		
Classification Models						
C1	C3D	LSTM	Low	No	LSTM	Cross entropy
C2	3DResNet	LSTM	High	No	LSTM	Cross entropy
C3	I3D	LSTM	Moderate	Yes	LSTM	Cross entropy
Verification Models						
Baseline Models						
V1	C3D		Low	No	Voting	Triplet Loss
V2	3DResNet		High	No	Voting	Triplet Loss
V3	I3D		Moderate	Yes	Voting	Triplet Loss
Inter Gait Cycle Models						
IGCV1	C3D		Low	No	LSTM	Triplet Loss
IGCV2	3DResNet		High	No	LSTM	Triplet Loss
IGCV3	I3D		Moderate	Yes	LSTM	Triplet Loss
Spatio-temporal Cascaded LSTM Models						
SCLV1	C3D	LSTM	Low	No	LSTM	Triplet Loss
SCLV2	3DResNet	LSTM	High	No	LSTM	Triplet Loss
SCLV3	I3D	LSTM	Moderate	Yes	LSTM	Triplet Loss

setting. The third is the analysis of matching of first-person videos with the third person videos. This analysis is also done in an open set recognition setting. It may be noted that given the limited amount of related work, comparison with other techniques was not possible. Therefore, to validate our contributions, we have performed a rigorous ablation study using different network backbones: C3D, I3D [3], and 3D-ResNet [4]. The details of the frameworks used in the ablation study are shown in Table 2. Here the C1, C2, and C3 models are the proposed systems trained using cross-entropy loss function for the classification task. V1, V2, and V3 are simple verification frameworks (baselines), in which a single 3-D CNN is used to extract the features from 1 gait cycle, and the final decision (for 4 seconds video) is computed by voting. IGCV1, IGCV2, and IGCV3 also have a single 3-D CNN that is used to extract the features from 1 gait cycle, but LSTM computes the final feature vector. Finally, SCLV1, SCLV2, and SCLV3 are the proposed frameworks, in which a 3-D CNN is used as a spatio-temporal feature extractor, and an LSTM finally computes the feature of a gait cycle. These features are then again combined using LSTM.

2.1 Analysis of classifier

We have analyzed our classifier for both identification and the verification task, as suggested in [2]. For identification, the classification accuracy is computed over all the datasets. For verification, one vs. rest strategy is used. We report *Equal Error Rate* (EER) for the verification, which is the rate at which the false accept rate is equal to the false

Table 3: Comparative analysis of classification accuracies (%) of the proposed system with [2] for first person recognition task.

Dataset	[2]	C1	C2	C3
FPSI	76.0	82.0	83.3	82.8
EVPR	90.0	92.5	93.3	93.2
DB-01	95.1	99.2	99.2	99.3
DB-02	93.7	97.3	97.7	98.0
DB-03	94.0	98.7	98.5	98.8
DB-04	85.6	89.9	90.6	90.1

Table 4: Comparative analysis of Equal Error Rate (%) of the proposed system with [2] for first person recogniAtion task.

Dataset	[2]	C1	C2	C3
EVPR	11.3	9.8	9.1	9.4

reject rate. Lower EER is better. The comparative analysis of all the datasets is shown in Table 3. For classification, the values for EVPR and FPSI dataset have been taken from their paper, whereas for DB-01, DB-02, DB-03, and DB-04, we computed the results using the code provided by [2]. For verification, Hoshen and Peleg have given results for the EVPR dataset, which is 11.3% EER. Accordingly, the comparison is made on EVPR only. The comparative analysis on the EVPR dataset is shown in Table 4. From the Table 3 and 4, it is clear that for each dataset our proposed system performs better than the state-of-the-art [2].

2.2 Analysis of Learned Distance Metric and Ablation Study

For analyzing the distance metric, we validated it for verification and recognition problems. We computed the equal error rate (EER) and decidability index for all the datasets in the verification framework. Decidability index [5] is a commonly used score in biometrics to evaluate the discrimination between genuine and impostor matching scores in a verification task. A large decidability index indicates strong distinguishability characteristics, i.e., high recognition accuracy and robustness. We report rank 1 correct recognition rate (CRR) for all the datasets.

To show the robustness of the proposed system, we first computed the performance of only 3-D CNN's without LSTM's. The C3D was pretrained only on RGB data, whereas I3D and 3D-ResNet were pretrained on both RGB as well as optical flow data. Results of the proposed technique on various datasets and on various 3D CNN backbones (C3D, I3D [3], and 3D-ResNet [4]) pretrained on RGB data are shown in Table 5. Results of the proposed technique on various datasets and various 3D CNN backbones (C3D, I3D, and 3D-ResNet) pretrained on optical flow data are shown in Table 6. Each 3-D CNN takes 1 second of video as input, whereas the proposed approach takes 4 seconds of inputs. Hence to compare the results here, we have used the average voting principle and have shown the results of 4 seconds of video duration. Finally, we computed the performance

Table 5: Performance analysis of the 3-D CNN's pretrained on RGB data for person verification task.

Dataset	**EER**(%)			**CRR**(%)			**DI**		
	V1	V2	V3	V1	V2	V3	V1	V2	V3
FPSI	22.80	24.26	23.52	62.50	62.17	63.05	0.26	0.27	0.27
EVPR	14.79	15.45	14.97	68.12	67.4	68.25	1.95	1.98	1.96
DB-01	3.99	4.54	4.86	87.54	88.2	87.85	2.19	2.45	2.5
DB-02	5.12	5.72	5.64	85.71	84.67	84.2	2.06	2.07	2.42
DB-03	7.82	8.4	9.94	85.99	84.77	84.56	2.15	2.15	2.02
DB-04	19.26	20.02	19.8	70.51	69.46	69.9	1.78	1.63	1.57

Table 6: Performance analysis of the 3-D CNN's pretrained on optical flow data for person verification task.

Dataset	**EER**(%)		**CRR**(%)		**DI**	
	V2	V3	V2	V3	V2	V3
FPSI	22.84	21.63	63.33	62.97	0.28	0.27
EVPR	15.23	14.67	70.07	69.32	2.01	1.99
DB-01	4.21	4.38	89.41	88.26	2.47	2.53
DB-02	5.29	5.03	87.17	87.58	2.24	2.29
DB-03	6.2	5.72	86.28	87.04	2.12	2.26
DB-04	18.36	19.64	71.75	71.09	1.66	1.69

of the proposed system, including LSTM. The results for each 3-D CNN backbone for analyzing the gait cycle with LSTM over every dataset and results of the proposed system are shown in Table 7. It is evident from the tables that the performance over DB-01 is best as it is collected in a controlled indoor environment. The performances on DB-02 and DB-03 are very similar to those on DB-01.

Moreover, the performance of C3D as the backbone is superior to that of I3D and 3D-ResNet architectures. This is because the C3D backbone has fewer parameters than their counterparts hence avoiding overfitting. However, the I3D backbone performs better than 3DResNet. This could be because of the multi-scale information caught by I3D, which is missing in 3DResNet.

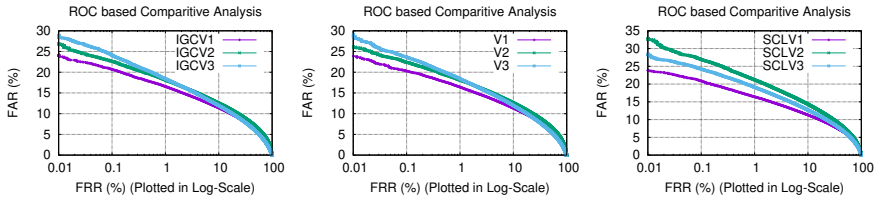
The result over DB-04 is not that promising and is closer to that of FPSI. This is because the number of samples in FPSI is enormous, and the combined dataset is dominated by it. However, it is interesting to observe that the decidability index over DB-04 is sufficiently better than that of FPSI, showing that the network generalizability has increased.

The ROC curves for the EVPR dataset are shown in Figure 2. The X-axis is plotted on a logarithmic scale for better visibility of variations.

To further strengthening the generalizability claim, we have also performed cross-dataset testing in two different scenarios. In the first scenario, we trained the proposed system on DB-01 (indoor) dataset and tested on the DB-02 (outdoor) dataset and vice versa. In the second, we trained on DB-03 (our dataset) and tested on the EVPR dataset

Table 7: Performance analysis of the various proposed models for person verification task.

Method	**EER** (%)					
	FPSI	EVPR	DB-01	DB-02	DB-03	DB-04
V1	22.80	14.79	3.99	5.12	7.82	19.26
V2	22.84	15.23	4.21	5.29	6.2	18.36
V3	21.63	14.67	4.38	5.03	5.72	19.64
IGCV1	21.47	13.76	3.26	5.39	6.95	17.96
IGCV2	20.92	13.30	3.32	4.76	5.48	16.89
IGCV3	20.67	12.95	3.87	4.86	5.02	16.71
SCLV1	**19.71**	12.00	**2.79**	**3.81**	**4.35**	15.49
SCLV2	19.73	12.32	2.82	4.84	4.84	15.88
SCLV3	20.34	**11.88**	3.30	4.28	4.54	**15.44**



(a) (b) (c)

Fig. 2: ROC curves of the proposed systems on EVPR dataset (a) IGCV1, IGCV2 and IGCV3 (b) V1, V2, and V3, and (c) SCLV1, SCLV2, and SCLV3.

and vice versa. The results for cross-dataset testing is shown in table 8. From the table, it can be clearly seen that despite not have seen any data at all of a given dataset, the system is still able to recognize the camera wearer.

2.3 Open Set Verification

We also performed open set verification on the indoor dataset (DB-01), outdoor dataset (DB-02), indoor and outdoor combined dataset (DB-03), combined dataset (DB-04), and EVPR. Openset analysis is not performed over the FPSI dataset as the number of subjects is very small. Half of the subjects from each of the individual datasets were taken for training and rest half for testing. *We believe that this mimics the anonymous and uncooperative wearers, which have not been seen at the train time, but we would still like to verify them at the test time.* The results of open set verification are shown in Table 9. Comparing these results with closed set results of the Table 7, it is evident that there is only a minor decrease in the performance of the network, which still has a low error rate. Hence, we can conclude that the proposed model can verify unseen camera wearers also.

Table 8: Performance analysis for cross-dataset testing.

Dataset		**EER** (%)		
Trained on	Test on	SCLV1	SCLV2	SCLV3
DB-01	DB-02	10.1	10.04	9.87
DB-02	DB-01	7.33	7.26	7.23
DB-03	EVPR	15.89	16.07	15.56
EVPR	DB-03	11.42	11.68	11.02

Table 9: Performance analysis for openset verification on EVPR and DB-04.

Dataset	**EER** (%)		
	SCLV1	SCLV2	SCLV3
EVPR	14.35	13.96	14.32
DB-01	6.43	6.31	5.92
DB-02	8.23	7.49	7.26
DB-03	9.39	8.71	8.94
DB-04	20.61	19.21	19.28

References

1. 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
2. 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
3. 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
4. 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
5. Ravikanth, C., Kumar, A.: Biometric authentication using finger-back surface. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2007) 1–6