Anonymizing Egocentric Videos

Daksh Thapar, Aditya Nigam
Indian Institute of Technology Mandi
dakshthapar.github.io, faculty.iitmandi.ac.in/~aditya

Chetan Arora
Indian Institute of Technology Delhi
www.cse.iitd.ac.in/~chetan

Abstract

In egocentric videos, the face of a wearer capturing the video is never captured. This gives a false sense of security that the wearer's privacy is preserved while sharing such videos. However, egocentric cameras are typically harnessed to wearer's head, and hence, also capture wearer's gait. Recent works have shown that wearer gait signatures can be extracted from egocentric videos, which can be used to determine if two egocentric videos have the same wearer. In a more damaging scenario, one can even recognize a wearer using hand gestures from egocentric videos, or identify a wearer in third person videos such as from a surveillance camera. We believe, this could be a death knell in sharing of egocentric videos, and fatal for egocentric vision research. In this work, we suggest a novel technique to anonymize egocentric videos, which create carefully crafted, but small, and imperceptible optical flow perturbations in an egocentric video's frames. Importantly, these perturbations do not affect object detection or action/activity recognition from egocentric videos but are strong enough to dis-balance the gait recovery process. In our experiments on benchmark EPIC-Kitchens dataset, the proposed perturbation degrades the wearer recognition performance of [42], from 66.3% to 13.4%, while preserving the activity recognition performance of [10] from 89.6% to 87.4%. To test our anonymization with more wearer recognition techniques, we also developed a stronger, and more generalizable wearer recognition method based on camera egomotion cues. The approach achieves state-of-the-art (SOTA) performance of 59.67% on EPIC-Kitchens, compared to 55.06% by [42]. However, the accuracy of our recognition technique also drops to 12% using the proposed anonymizing perturbations.

1. Introduction

Egocentric videos are typically captured through wearable cameras harnessed on a person's head, recording in a hand-free style, without any explicit user intervention. The unique perspective gives a wearer ability to share his/her experience as seen through his/her own eyes.

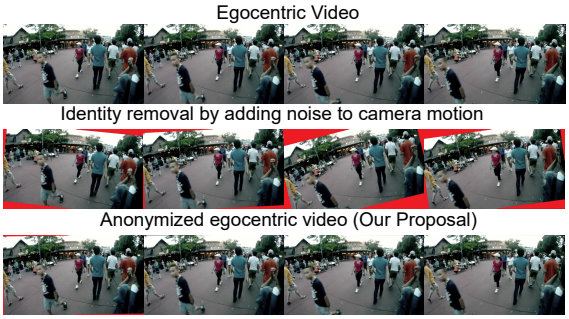

Figure 1: Wearer recognition from egocentric videos is a critical privacy breach. In this paper we propose to add carefully crafted, but subtle 3D rotations in the egocentric videos which can lead to the failure of wearer recognition techniques. Row 1 shows original frames of an egocentric video, whereas 3rd rows shows the frames after proposed transformation. Note that the transformations are imperceptible. We show in our experiments that they also do not affect the performance of other video analysis tasks. On the other hand, naively adding random rotations (row 2) causes visual distortion (note the red portions near the borders due to large rotations), but does not lead to significant deterioration in wearer recognition ability as shown in our experiments.

The potential utility of such cameras in applications like law-enforcement, geriatric care, and life-logging has also spawned serious research in egocentric videos. The computer vision community has responded with a variety of large scale publicly available egocentric video datasets viz. FPSI [8], EGTEA [22], and EPIC-Kitchens [5]. Some notable tasks in egocentric vision taken-up so-far include egocentric video summarization, temporal segmentation, as well as object, action, and activity recognition from first-person viewpoint [16, 45, 20, 33, 28, 19, 9, 46, 15, 44, 24].

While the wearable, hands-free, and always-on nature of egocentric cameras have been one of the reasons for their popularity, they have also brought in unique privacy concerns. Hence, both general users, as well as researchers, have been careful in not collecting the videos in private places such as bedrooms or toilets. However, outdoor egocentric videos, and especially the scenarios when no other person is seen in the videos are generally considered safe and shared freely. This has been due to the popular per-

ception that since wearer's face is not visible in the egocentric videos, hence there is no privacy risk in public sharing of such videos. Hoshen and Peleg [15] have shattered this false belief, and showed that one could train a classifier over optical flow to identify the camera wearer of an egocentric video. Recently, Thapar et al. [41] have further extended the privacy breach to an open-set scenario. Their result is also more alarming, since they show the capability to cross-recognize the wearer gait captured from egocentric videos against a gait captured from a third-person surveillance video. This allows them to even know the face behind an egocentric video! In an extension of their work, the authors have shown a similar capability using hand gestures as seen in egocentric videos [42].

It is important to note that the problem does not exist in videos captured from hand-held, or third-person cameras. Firstly, these cameras do not capture the photographer's gait or any other biometric profile. Secondly, when the face of a person is visible in such a video, it is easy to mask it out by blurring or a similar technique. In the case of egocentric videos, the privacy breach is happening through fine-grained frame-to-frame optical flow, and it is not clear what can be masked out to prevent the leak.

Such a serious and well-exposed privacy breach demands an immediate solution. Once the community heeds to this warning seriously, sharing egocentric video data at a large scale would become impossible. This can be a huge potential threat for ongoing egocentric research as most of the recent results in the field have been obtained through data-intensive deep neural network architectures.

The focus of this work is to generate a perturbation to transform any given egocentric video (v^i), corresponding to some camera wearer (i) into an anonymized egocentric video ($\tilde{v}^i$), which blocks wearer recognition techniques from extracting gait signatures of the wearer i from $\tilde{v}^i$. To preserve the utility of these videos, such a transformation must not significantly modify the video such that, (a) alteration should be imperceptible, and (b) the transformed video ($\tilde{v}^i$) should remain usable for other egocentric video analysis tasks viz. object, or activity recognition. Such an anonymizing transformation for egocentric videos is possible because most analysis tasks, e.g., activity or attribute recognition performed over egocentric videos, typically do not need subject-specific features/cues. To the contrary, the techniques for these tasks should be identity agnostic (irrespective of the identity of the camera wearer) to avoid dataset over-fitting. Hence identity features can be safely suppressed or perturbed without affecting the performance of other egocentric video analysis tasks of interest.

A candidate solution for anonymizing transformation is the addition of noise to the egocentric videos. Here, we note that most state-of-the-art (SOTA) egocentric video analysis techniques are based on deep neural networks (DNNs), and are trained on original and augmented data (with noise). Hence, they are already resistant to some amount of noise in the data. The addition of a large amount of noise may block wearer identity recognition. However, it may also cause significant collateral damage by degrading the overall video quality, and adversely affecting the performance of other egocentric video analysis tasks as well.

Contributions: The specific contributions of this paper are: **(1)** We propose a novel video transformation technique to prevent privacy leak in egocentric videos by blocking wearer recognition, but without affecting performance of other video analysis tasks. This plugs an immediate and urgent security hole exposed by recent works and restores the safe sharing of egocentric datasets. **(2)** We show that the proposed transformation degrades the wearer recognition performance of [42] on benchmark EPIC-Kitchens dataset from 66.3% to 13.4%, whereas recognition performance of [41] degrades from 85.5% to 15.8% on FPSI dataset [8]. On the other hand, despite such degradation in recognition performance, the activity recognition accuracy of the SOTA [10] on perturbed videos merely changes from 92% to 91%. **(3)** To further consolidate our work, we propose a novel wearer recognition technique utilizing self-attention and arc-softmax loss. The technique is more generalizable, and achieves SOTA performance of 59.67% in comparison to 55.06% by [42]. However, the proposed perturbation causes our recognition technique to also fail causing drop in accuracy to 12%.

2. Related Work

Egocentric Video Analysis: Some notable works in general egocentric video analysis include camera wearer's activity and action recognition [28, 19, 44, 2, 35, 36, 30, 31], wearer gaze estimation [16], temporal segmentation [24], video summarization [45, 32]. We note that traditional video stabilization techniques could have also introduced small random perturbations in the input videos helping to block wearer recognition. However, [13] has noted that these techniques do not work for egocentric videos, and have instead proposed sub-sampling based joint stabilization and fast-forwarding. We note that the temporal subsampling is not suitable for anonymizing as it will affect the performance of other video analysis tasks as well.

Attribute Extraction from Egocentric Videos: Finocchiaro et al. [9] estimated camera height from egocentric videos by extending [31]. In [7, 1] authors suggest combining first and third-person camera views, to identify the camera wearer in the third person view: sharing fields-of-view [7], and common scene observation [1]. Researchers have also suggested to estimate the wearer's location directly [14, 23] or indirectly (using gaze, social interactions, etc.) [26, 37]. In [29, 47], wearer identity is revealed using

head motion obtained from optical flow. Wearer's pose has been estimated in [17] using 3D joint regression.

Wearer Recognition from Egocentric Videos: Hoshen and Peleg [15] has shown that to identify the camera wearer one can train a classifier from the optical flow in egocentric videos. Thapar et al. [41] extended this privacy breach to an open-set scenario where they can reveal an unseen wearer's identity captured in an egocentric video by cross-domain gait matching in third-person surveillance videos. Building over their earlier work, the same authors [42] have also shown that one could identify a camera wearer from the hand gestures as seen in the egocentric video. Importantly the hand gestures could be identified while performing the same or even an activity unseen at the train time. Taken together, the three works above indicate significant privacy breach by sharing one's egocentric videos.

Adversarial Perturbations in Neural Networks: The geometrical perturbations proposed in our method to scuttle wearer identify recognition can also be seen as introducing adversarial perturbations in a video. We introduce a specific kind of perturbation which only affects one particular task and do not affect others. Szegedy et al. [40] were the first ones to reveal the sensitivity of modern deep neural networks to artificial perturbations using gradient-based algorithms. Goodfellow et al. [12] improved the efficiency of detecting adversarial perturbations using Fast Gradient Sign Method (FGSM), which were then used to train the neural networks improving its robustness. Kurakin et al. [21] and Sharif et al. [34] showed that adversarial examples can be extended to real-world scenarios. Chen et al. [4] show that physical world adversarial examples can be created for object detection networks such as Faster R-CNN. Su et al. [39] and Narodytska and Kasiviswanathan [25] showed that modifying one pixel using Differential Evolution (DE) is sufficient to fool classifiers. Brown et al. [3] and Karmon et al. [18] introduced adversarial patches as a more practical adversarial attack. We are not aware of any technique which insert adversarial perturbation at the video level, or specifically in optical flow.

3. Proposed Methodology

Our Strategy: We seek to anonymize a target egocentric video by perturbing it with an adversarial geometric transformation which makes wearer recognition impossible. Since a general 3D transformation may require knowing the depth in the scene, we restrict our attention to Homography, which does not need any depth information, and correspond to causing adversarial 3D rotations in an egocentric camera. The first part of our strategy is to generate a classifier which can recognize a wearer based upon the 3D rotations observed in the egocentric camera. The motivation is that if we can do it successfully, then we can pick an arbitrary egocentric video from the gallery, recognize a wearer from it, and then back-propagate through this network to compute precise subset/component of the input which caused a particular wearer classification. This is akin to finding a wearer's signature in the input space comprising of 3D rotation matrices. We hypothesize that if we take this arbitrary camera wearer's signatures and overlay it on the target video, then the target video will have an arbitrary mix of two identities (original and the one derived from arbitrary gallery video). This will cause enough perturbation leading to failure of the target wearer identification techniques [15, 42, 41]. If we take an analogy from face recognition, this is akin to taking an arbitrary face and blending it with features from another arbitrary face, with the hope that mix of two arbitrary faces will make a face recognition system unable to identify anyone. As we show in our experiments, the strategy indeed causes failure of the target techniques [15, 42, 41].

Visual Distortions: Note that in terms of visual distortions, we end up causing a small Homography based transformation to each frame. Similar to image space adversarial perturbations, the proposed perturbation is too small to be perceived by humans, and, as we show in our experiments, does not cause any deterioration in perforance of other video analysis task, for which this simply means a little more noise in wearer head motion, which they are anyways robust against. Below we describe our implementation of the above mentioned strategy in detail.

Design of Transformation Module: The proposed transformation module has been divided into following key stages: (i) Compute camera rotation matrices from an egocentric video using any third party tool. (ii) Train a novel module named EgoIdNet on a train dataset, and learn to identify a camera wearer based upon the recovered rotation matrices from a video. (iii) Take an arbitrary video from the train set or otherwise, called gallery video, and use EgoIdNet to recover the wearer identity from it. Back-propagate through EgoIdNet to reveal salient features in the input space (15 rotation matrices) for the video. (iv) Compute corresponding Homographies from the 3D rotation matrices computed using back-propagation with a gallery video as input. Take a video which needs to be anonymized. Warp each frame using the homographies computed from previous step. Below we describe each step in detail.

Step 1 – Compute Rotation Matrices: For a given input video, we extract the camera rotation matrices using a SLAM (simultaneous localization and mapping) algorithm. In our implementation, we use the one proposed by Patra et al. [27] due to its demonstrated robustness for egocentric videos[1]. We use 3D Euler angle representation for a rotation matrix: $R = (r_x, r_y, r_z)$. Here, the direction of vector

[1] We thank the authors of [27] for providing their code

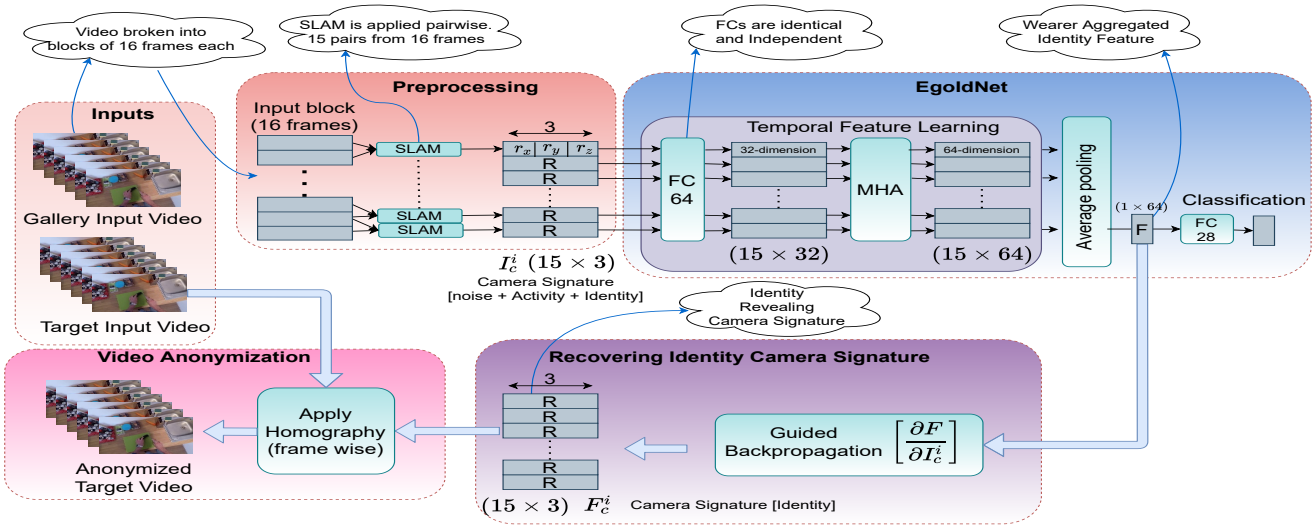

Figure 2: The proposed framework for anonymization first person videos. FC stands for Fully connected network. MHA stands for Multi Head attention netwerk. The first FC is applied to each temporal feature separately.

represents axis of rotation, and its magnitude is the rotation angle (in degrees). This specific input format ensures that signal received by back-propagation through the network at a later stage always corresponds to a valid rotation matrix. We normalize R vectors element-wise to a Gaussian distribution with zero mean, and unit variance.

Step 2(a) – Preparing Input: The previous step gives the camera pose corresponding to each frame (relative to a chosen reference frame, usually the first one) as a 3-dimensional R vector. For training EgoIdNet, we divide each input video (V_c) into clips of 16 frames each $(C_1, \ldots, C_n)$. For each clip C_i, we first compute the pairwise rotation between a frame t and $(t+1)$. This gives us 15 vectors per clip. We concatenate these vectors into a 15×3 matrix, and denote it by I_c^i (indicating features corresponding to i$^{\text{th}}$ clip of unknown arbitrary camera wearer c).

Step 2(b) – Feature Scaling: EgoIdNet extracts temporal features from I_c^i, to obtain wearer-specific identity features (see Fig. 2). The first stage of the model consists of a fully connected layer comprising of 64 neurons. The layer has its weight tied for each 3D rotation vector, but is applied to each vector separately. Note that this is equivalent to performing non-linear, learnable feature scaling which converts a 3 dimensional vector to a 64 dimensional one, and facilitates highly accurate wearer recognition.

Step 2(c) – Multi-head attention module: To learn the temporal relationship between the frames of a clip, in the second stage of EgoIdNet, we use a multi-head attention layer having eight narrow heads of self-attention [43]. The output features are averaged out to obtain an aggregated wearer identity feature F. Finally, the classification layer is applied over F to perform wearer recognition. The basic building block of multi-head attention [43] is the scaled dot product mechanism. The mechanism is a sequence to sequence operation, which given a sequence of n input vectors $v_1, \ldots, v_n$ (called *value* vectors), learns to output sequence of vectors $y_1, \ldots, y_n$, based on a sequence of *query*: $q_1, \ldots, q_n$, and *key* vectors: $k_1, \ldots, k_n$. Here, $n = 15$, and each vector in the sequence is of dimension 64 in our case. The module performs following steps:

1. Learns weight matrices, each of size 64×64, to transform each of the three (value, query and key) sequences: $\tilde{q}_i = W_q q_i$, $\tilde{k}_i = W_k k_i$, $\tilde{v}_i = W_v v_i$.
2. Compute y_i as a weighted average of the transformed value vector $\tilde{v}_j$: $y_i = \sum_j w_{ij} \tilde{v}_j$. Here j, iterates over the whole sequence, and w_{ij} is computed as the dot product between query and key sequences: $w_{ij} = \text{softmax}(w'_{ij})$, where $w'_{ij} = \frac{\tilde{q}_i^T \tilde{k}_j}{\sqrt{d}}$. Note that, in order to obtain the weights (w_{ij}) to compute a particular output y_i, the corresponding query vector $\tilde{q}_i$ is compared (via. scaled dot product) with all key vectors, $\tilde{k}_1, \ldots \tilde{k}_n$. This dot product is interpreted as attention for each value vector $\tilde{v}_1, \ldots, \tilde{v}_n$, and the scaled dot product attention mechanism is utilized multiple times in a multi-head attention.
3. We stack n, 64 dimensional vectors: value, query, and key, into a matrix of size $n \times 64$. Each matrix is subdivided into 8 sub-matrices of size $n \times \frac{64}{8}$. Each sub-matrix is processed through a separate attention head, having independent weight matrices of $\frac{64}{8} \times \frac{64}{8}$. The attention heads jointly produce 8 output matrices of size $n \times 8$. We concatenate all sub-matrices to get the final output of size $n \times 64$.

Step 2(d) – Loss Function: We train EgoIDNet with arc-softmax loss [6], due to its better class compactness property under open-set scenarios. The loss is defined as:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{i=n}\log\frac{e^{\cos(\theta_{y_i}+m)}}{e^{\cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{n}e^{\cos\theta_j}}.$$

Here n is the number of classes, m is the angular margin enforced between features of different classes. Further, θ_{y_i} is the angle formed between the wearer's aggregated identity feature and the weight vector of i^{th} neuron in the final fully connected layer (see [6] for details).

Step 3 – Computing Perturbation: We compute the desired perturbation which can scuttle a wearer recognition technique by back-propagating in EgoIdNet. For this, we take arbitrary video from a train set or otherwise, compute rotations, and crop a 16 frame clip from it, as described earlier. We then perform inference for wearer recognition from the input features. However, in this case we are not interested in the recognition but in finding the salient parts of the input feature which led to the decision. For this, we de-convolve through the model using guided back-prop [38] (see Fig. 2). We compute the partial derivative of the output of wearer aggregated identity features (F) with respect to the input (I_c^i), (denoted by $F_c^i = \partial F/\partial I_c^i$). The guided back-prop computes the derivative of n^{th} layer output X_{n+1} with respect to the input X_n as follows:

- For a convolutional layer represented by $X_{n+1} = X_n * K_n$, where K_n is the convolutional kernel, the gradient of feature map F with respect to X_n is: $\frac{\partial F}{\partial X_n} = \frac{\partial F}{\partial X_{n+1}} * K_n^t$ where K_n^t is the flipped version of kernel K_n.
- For a fully connected layer represented by $X_{n+1} = X_n \times W_n$, where W_n is the weight matrix, the gradient of feature map F with respect to X_n is: $\frac{\partial F}{\partial X_n} = \frac{\partial F}{\partial X_{n+1}} \times W_n^t$ where W_n^t is the flipped version of kernel W_n.
- For RELU activation defined as $X_{n+1} = \max(X_n, 0)$, sub-gradient takes the form: $\frac{\partial F}{\partial X_n} = \frac{\partial F}{\partial X_{n+1}} \times \mathbb{I}(X_n > 0)$ where $\mathbb{I}$ is element-wise indicator function.
- For sigmoid and tanh activations defined as $X_{n+1} = \sigma(X_n)$ and $X_{n+1} = \tanh(X_n)$, sub-gradient takes the form: $\frac{\partial F}{\partial X_n} = \frac{\partial F}{\partial X_{n+1}} \times \sigma(X_n) \times (1-\sigma(X_n))$, and $\frac{\partial F}{\partial X_n} = \frac{\partial F}{\partial X_{n+1}} \times (1 - \tanh^2(X_n))$

Post guided back-prop operation, F_c^i contains the identity signatures corresponding to the wearer of the particular video. We use F_c^i as the perturbation that will be added for anonymizing a new video. As mentioned earlier, we note that which video to use for computing perturbation is not crucial in our technique. Addition of said perturbation in a new video results in overlap of the features corresponding to two different identities in that video. This overlap is the primary reason for failure of a wearer recognition technique, and not which identity has been used for computing the perturbation.

Step 4 – Creating Anonymized Videos: Note that the perturbation computed in the previous step is in the input space. i.e. 15 rotation vectors. For adding the perturbation to a new video, we convert each recovered vector $R = (r_x, r_y, r_z)$, to a rotation matrix R_{mat}. We then convert the rotation matrix R_{mat} to a homography $H = KR_{\text{mat}}K^{-1}$, where K is camera intrinsic matrix of the target video. We then divide the target video to anonymise into clips of 16 frames, and warp frame 2–16 of the clip using the computed Homography. To further minimize the the impact of perturbation we reduce the magnitude of recovered rotation vectors by a factor α before computing Homography.

4. Dataset and Evaluation Methodology

Datasets: We validate the performance of our anonymizing strategy on the same benchmark egocentric datasets as used by the attack techniques [15, 41, 42]: **EPIC-Kitchens [5]:** dataset consists of 55 hours of egocentric videos from 32 subjects, and contains 125 labeled activities performed by the subjects. As Thapar et al. [42] has validated the person recognition system on the five activities, we have also chosen the same five activities viz *cut, mix, put, take*, and *wash*. **FPSI [8]:** dataset consists of videos captured by 6 people wearing cameras mounted on their hat, and spending their day at Disney World Resort in Orlando, Florida. **IITMD-WFP:** dataset [41] consists of 3.1 hours of videos captured from 31 different subjects. The dataset has been captured under two different scenarios: indoor and outdoor, which are referred to as DB-01 (indoor), and DB-02 (outdoor). The combined dataset is referred as DB-03. This is inline with the original nomenclature [41].

Compared Techniques: The primary objective of this work is to anonymize egocentric videos without degrading the performance of other analysis tasks. To validate the performance degradation of person recognition on FPSI dataset, we have used 2 publicly available SOTA models: Hoshen and Peleg [15] and Thapar et al. [41] and refer to them as "Hoshen", and "Th_fpsi" respectively. For EPIC-Kitchens dataset, there is only one attack technique, available from Thapar et al. [42] (hereinafter "Th_epic"). To validate the performance of activity recognition on FPSI dataset we have used Poleg et al. [31] (referred to as "Poleg"), and for EPIC-Kitchens, we have used [44], [11], and [10] and refer to them as "Verma", "Ghadiyaram", and "Furnari" respectively. We got the code of [44] by requesting the authors. The code for others is publicly available.

Naive Noise Model: A naive solution to the anonymization objective could be the addition of noise to the egocentric videos. To implement this model, we introduce a varying

amount of noise in the original videos by randomly generating a rotation vector with magnitude sampled from a uniform distribution (ranging from 0–β radians). For the axis direction, we sampled angle of rotation axis with x, y, and z axis, uniformly between $[-\pi, +\pi]$. To add noise, we compute the Homography matrix using this random rotation vector and warp each frame with it.

Comparison Metrics: For evaluating the performance of person verification, we have used the following performance parameters: Equal Error Rate (EER), Correct Recognition Rate (CRR), and Decidability Index (DI) as used in [41]. For good performance, EER should be low, whereas CRR and DI should be high. For evaluating the performance of activity recognition, we have used Accuracy (%) and F-score as the performance parameters.

Experiment Setup: We perform our experiments under both open set and closed set scenarios. For gait based recognition, we follow the protocol of [41] and test on the FPSI dataset. For the open set scenario, we take the videos of the first three subjects for training our EgoIdNet, other wearer recognition models, as well as the activity recognition model. Whereas, the videos from the other three subjects have been taken for testing the attack (wearer recognition), and anonymization model and check for any degradation in the activity recognition. For the hand gesture based recognition, we use EPIC-Kitchens dataset and follow the protocol of [42]. Here, for the open set testing, videos for the first half of the subjects (16) are taken for training, and second half of the subjects have been taken for testing.

5. Experimental and Results

Recognition performance of EgoIdNet: We have analyzed the recognition capability of EgoIdNet under both closed-set (wearers are known and trained for during training) and open-set (wearers are unseen during training) scenarios. Table 1 compares the performance of EgoIdNet with the current SOTA [41, 42]. We observe a significant performance boost-up for open-set scenarios while marginally better results for the closed-set. The better open-set performance indicates improved generalization due to our reliance on camera rotation directly rather than optical flow. Note that optical flow is easily corrupted in dynamic scenes compared to camera ego-motion.

Performance Comparison with Naive Noise Model: Fig. 3 shows the result of addition of varying amount of random noise, and our perturbation to the videos of "cut" activity from EPIC-Kitchens dataset. We compare the anonymization performance to block open set wearer recognition, where the bar corresponding to "Th_epic" show the wearer recognition by [42]. The other three bars show activity recognition performance of SOTA. We see that with increasing level of random noise, the activity recognition performance drops at a similar rate as the wearer recognition, indicating significant interference of random noise in other video analysis task as well. On the other hand, with increasing amount of the proposed perturbation, wearer recognition falls while activity recognition falls marginally.

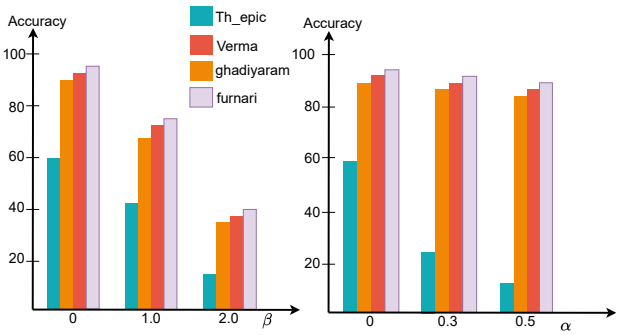

Figure 3: Comparison of anonymization performance and degradation in activity recognition on videos of "cut" activity in EPIC-Kitchens dataset after fine-tuning each model with the respective noise. The left plot corresponds to adding random noise with various level of β. The right plot is performance of proposed perturbation for various levels of α. See text for details of α and β. "Th_epic" show the wearer recognition by [42] (lower is better), whereas the other three bars show activity recognition performance of SOTA on the perturbed videos (higher values indicate no degradation and is better). Noise level 0 indicates no noise is added (i.e. original dataset).

Anonymization Performance for Open Set Wearer Recognition: The results of performance analysis on FPSI dataset for each of the target subject in an open set scenario is shown in Table 2. The table also shows similar results on the EPIC kitchens dataset for the five activities. In the table, original refers to the original performance of each model, anonymized refers to the performance of the model on the dataset transformed/anonymized using our technique, and fine-tuned refers to the performance of the models after fine-tuning on the anonymized dataset. Similar to the previous results, the two tables show that as we add stronger perturbation using our technique, the performance of wearer recognition significantly decreases, whereas performance on the activity recognition is maintained on each of the subject, as well as activity.

Open Set Anonymization Performance: As noted by other authors as well [41], open set wearer recognition is a difficult but more practical setting for wearer recognition and in turn easier for a technique trying to anonymize such videos by blocking wearer recognition. We have already successfully shown the ability of our proposed perturbations in causing wearer recognition techniques to fail in this scenario. However, to estimate the capability of our anonymization in a more difficult setting, we also test in closed set scenario. Note that, in a closed-set situation, the wearers are pre-known to a wearer recognition techniques

Table 1: Comparative analysis of our system with [41] and [42] for wearer recognition in egocentric videos. CA, and EER denote the classification accuracy, and Equal Error Rate respectively in percentage. Higher CA and lower EER is better.

Dataset	Closed Set Analysis				Open Set Analysis			
	EgoIDNet		[41]		EgoIDNet		[41]	
	CA↑	EER↓	CA↑	EER↓	EER↓	CRR↑	EER↓	CRR↑
FPSI	82.4	18.76	82.0	19.71	–	–	–	–
DB-01	99.1	2.47	99.2	2.79	5.87	86.33	6.43	83.67
DB-02	97.6	3.54	97.3	3.81	7.67	84.28	8.23	82.77
DB-03	99.0	4.12	98.7	4.35	6.52	83.46	9.39	80.56
	EgoIDNet		[42]		EgoIDNet		[42]	
EPIC	71.21	12.46	71.04	12.32	14.32	59.67	15.28	55.06

Table 2: Demonstration of proposed anonymization technique on FPSI, and EPIC-Kitchens dataset. "Original" indicates the performance of a model on original dataset. "Anonymized" indicates the performance of a model after adding proposed perturbation. "Finetuned" indicates that the model has been fine-tuned over the anonymized dataset. We observe that the performance of SOTA wearer recognition techniques ([41], and [42]) degrades significantly, whereas the performance of the SOTA activity recognition techniques ([31], and [10]) is maintained after addition of proposed perturbation in the videos.

Results on FPSI Dataset

Subject	Person recognition: Th_fpsi[41]				Activity Recognition: Poleg[31]			
	Parameter	Original	Anonymized	Finetuned	Parameter	Original	Anonymized	Fine-tuned
Subject-1	EER ↑	16.5	55.2	51.6	ACC↑	92.6	84.7	91.2
	CRR ↓	85.5	13.6	15.8	F-score↑	0.9	0.8	0.9
Subject-2	EER ↑	17.8	59.4	54.6	ACC↑	90.5	81.6	89.2
	CRR ↓	82.1	12.9	14.6	F-score↑	0.9	0.8	0.8
Subject-3	EER ↑	25.4	67.6	63.8	ACC↑	85.4	79.9	84.6
	CRR ↓	76.4	8.2	10.3	F-score↑	0.8	0.8	0.8

Results on EPIC-Kitchens Dataset

Activity	Wearer recogntition: Th_epic[42]				Activity Recognition: Furnari[10]			
	Parameter	Original	Anonymized	Finetuned	Parameter	Original	Anonymized	Fine-tuned
Cut	EER↑	19.2	48.7	44.3	ACC↑	92.4	85.2	91.8
	CRR↓	59.9	15.4	17.8	F-score↑	0.9	0.8	0.9
Mix	EER↑	16.6	52.9	49.4	ACC↑	89.6	81.8	87.4
	CRR↓	66.3	19.2	21.4	F-score↑	0.8	0.8	0.8
Wash	EER↑	20.8	64.5	61.2	ACC↑	91.3	84.2	90.5
	CRR↓	57.2	8.4	11.6	F-score↑	0.9	0.8	0.9
Put	EER↑	21.2	62.8	59.3	ACC↑	80.8	72.6	79.3
	CRR↓	58.1	10.2	14.9	F-score↑	0.8	0.7	0.8
Take	EER↑	15.2	50.8	49.4	ACC↑	82.6	77.7	82.1
	CRR↓	52.5	11.6	13.4	F-score↑	0.8	0.7	0.8

which makes it easier to train a wearer recognition model for such subjects. For this experiment we add proposed perturbation on unseen videos of the same subject and show that wearer recognition technique [15] fails to recognize a wearer now. Even after we fine-tune the model by training with videos perturbed using our technique, it still fails to recognize wearer in such videos. We conduct the experiment on FPSI dataset, and follow the protocol of Hoshen and Poleg [15]. The videos of each subject captured in the morning have been taken for training and the videos captured in the afternoon have been taken for testing. The results of the close-set analysis on FPSI dataset are shown in Table 3. From the table, it can be observed that even if we train the wearer recognition model over particular subjects that have to be anonymized, still our framework can fool the Hoshen classifier. This strengthens the belief that our proposed perturbation framework will be able to anonymize any given egocentric video successfully.

5.1. Qualitative Analysis

Fig. 4 shows the visual output after adding proposed perturbation, or random noise in two videos to block wearer

Figure 4: Visual comparison of the distortion caused by adding random rotation, and proposed transformation. 1st and 4th rows show the frames from original videos from FPSI and EPIC kitchens dataset respectively. 2nd and 5th rows show the output after noise addition in 1st and 4th, respectively. 3rd and 6th rows show the corresponding outputs from our technique. Note the large distortion by adding rotation as evident from black regions near the border.

recognition. The 1st and 4th rows show the frames from original videos from FPSI and EPIC kitchens datasets respectively. 2nd and 5th rows show the output after noise addition in 1st and 4th, respectively. 3rd and 6th rows show the corresponding outputs from our technique. There are almost no distortions visible in the output produced from our technique highlighting the subtle nature of the distortion introduced.

6. Conclusion

Recent works have shown a significant privacy breach by demonstrating the wearer recognition capabilities through one's egocentric videos. In this paper, we have studied techniques to plug the breach by fooling state of the art wearer recognition models. We add a subtle but systematic 3D rotations in the egocentric videos which are imperceptible to humans but are significant enough to fool all known wearer recognition techniques. We note that most state of the art wearer recognition techniques work on optical flow. Hence, as part of our efforts to test our anonymization with other style of wearer recognition techniques, we propose a novel wearer recognition method using camera egomotion, which is more robust to compute in dynamic scenes. While our proposed wearer recognition achieves state of the art performance, it still fails on the videos with proposed anonymizing perturbations. Importantly, the proposed perturbation does not affect other egocentric video analysis tasks such as wearer's activity recognition from egocentric videos. We hope that our work will prove to be an important milestone for the safe sharing of egocentric videos both for the consumer as well as research applications. The code and pretrained models for the work will be released publicly post acceptance.

Table 3: We demonstrate the performance of proposed perturbation technique on FPSI dataset in a closed set scenario. "Original" indicates the performance of a model on original dataset. "Anonymized" indicates the performance of the model after adding proposed perturbation. "Finetuned" indicates that the model has been finetuned over the anonymized dataset.

Wearer Recognition

Model	EER↑	CRR↓	DI↓
Hoshen[15] (Original)	20.34	76.0	0.25
Hoshen[15] (Anonymized)	70.22	08.7	0.02
Hoshen[15] (Fine-tuned)	68.75	10.4	0.02
Th_fpsi[41] (Original)	19.71	82.0	0.27
Th_fpsi[41] (Anonymized)	65.33	11.2	0.03
Th_fpsi[41] (Fine-tuned)	62.10	15.5	0.03

Activity Recognition

Model	ACC↑	F-Score↑
Poleg[31] (Original)	89	0.81
Poleg[31] (Anonymized)	82	0.76
Poleg[31] (Fine-tuned)	89	0.80

References

[1] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV (5)*, volume 9909 of *Lecture Notes in Computer Science*, pages 253–268. Springer, 2016. 2

[2] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, CV Jawahar, and KCIS CVIT. Unsupervised learning of deep feature representation for clustering egocentric actions. In *IJCAI*, pages 1447–1453, 2017. 2

[3] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 3

[4] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Robust physical adversarial attack on faster r-cnn object detector. *arXiv preprint arXiv:1804.05810*, 2(3):4, 2018. 3

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 5

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5

[7] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J. Crandall, and Michael S. Ryoo. Identifying first-person camera wearers in third-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[8] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 1, 2, 5

[9] Jessica Finocchiaro, Aisha Urooj Khan, and Ali Borji. Egocentric height estimation. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1142–1150. IEEE, 2017. 1, 2

[10] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019. 1, 2, 5, 7

[11] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019. 5

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[13] Tavi Halperin, Yair Poleg, Chetan Arora, and Shmuel Peleg. Egosampling: Wide view hyperlapse from single and multiple egocentric videos. 2

[14] Joel A Hesch and Stergios I Roumeliotis. Consistency analysis and improvement for single-camera localization. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22. IEEE, 2012. 2

[15] Yedid Hoshen and Shmuel Peleg. An egocentric look at video photographer identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4284–4292, 2016. 1, 2, 3, 5, 7, 8

[16] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and actions. *arXiv preprint arXiv:1901.01874*, 2019. 1, 2

[17] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 3

[18] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. *arXiv preprint arXiv:1801.02608*, 2018. 3

[19] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR 2011*, pages 3241–3248. IEEE, 2011. 1, 2

[20] Johannes Kopf, Michael F Cohen, and Richard Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):78, 2014. 1

[21] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 3

[22] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018. 1

[23] Ana Cristina Murillo, Daniel Gutiérrez-Gómez, Alejandro Rituerto, Luis Puig, and Josechu J Guerrero. Wearable omnidirectional vision system for personal localization and guidance. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–14. IEEE, 2012. 2

[24] Pravin Nagar, Mansi Khemka, and Chetan Arora. Concept drift detection for multivariate data streams and temporal segmentation of daylong egocentric videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1065–1074, 2020. 1, 2

[25] Nina Narodytska and Shiva Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318. IEEE, 2017. 3

[26] Hyun S Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. In *Advances in Neural Information Processing Systems*, pages 422–430, 2012. 2

[27] Suvam Patra, Kartikeya Gupta, Faran Ahmad, Chetan Arora, and Subhashis Banerjee. Ego-slam: A robust monocular slam for egocentric videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 31–40. IEEE, 2019. 3

[28] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE*

Conference on Computer Vision and Pattern Recognition, pages 2847–2854. IEEE, 2012. 1, 2

[29] Yair Poleg, Chetan Arora, and Shmuel Peleg. Head motion signatures from egocentric videos. In *Asian Conference on Computer Vision*, pages 315–329. Springer, 2014. 2

[30] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2544, 2014. 2

[31] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. Compact cnn for indexing egocentric videos. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 2, 5, 7, 8

[32] Anuj Rathore, Pravin Nagar, Chetan Arora, and CV Jawahar. Generating 1 minute summaries of day long egocentric videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2305–2313, 2019. 2

[33] Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3137–3144. IEEE, 2010. 1

[34] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 3

[35] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016. 2

[36] Suriya Singh, Chetan Arora, and CV Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognition*, 62:45–55, 2017. 2

[37] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3503–3510, 2013. 2

[38] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 5

[39] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 3

[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[41] Daksh Thapar, Aditya Nigam, and Chetan Arora. Is sharing of egocentric video giving away your biometric signature? In *The European Conference on Computer Vision (ECCV)*, August 2020. 2, 3, 5, 6, 7, 8

[42] Daksh Thapar, Aditya Nigam, and Chetan Arora. Recognizing camera wearer from hand gestures in egocentric videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2095–2103, New York, NY, USA, 2020. Association for Computing Machinery. 1, 2, 3, 5, 6, 7

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[44] Sagar Verma, Pravin Nagar, Divam Gupta, and Chetan Arora. Making third person techniques recognize first-person actions in egocentric videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2301–2305. IEEE, 2018. 1, 2, 5

[45] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2244, 2015. 1, 2

[46] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7593–7602, 2018. 1

[47] Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. Ego-surfing first-person videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2