Head Motion Signatures from Egocentric Videos

Yair Poleg[1], Chetan Arora[2], and Shmuel Peleg[1]

[1] The Hebrew University of Jerusalem, Israel
[2] IIIT Delhi, India

Abstract. The proliferation of surveillance cameras has created new privacy concerns as people are captured daily without explicit consent, and the video is kept in databases for a very long time. With the increasing popularity of wearable cameras like Google Glass the problem is set to increase substantially. An important computer vision task is to enable a person ("subject") to query the video database ("observer") whether he/she has been captured on the video. Following a positive answer, the subject may request a copy of the video, or ask to be "forgotten" by erasing this video from the database. Two properties such queries should have are: (i) The query should not reveal more information about the subject, further breaching his privacy. (ii) The query should certify that the subject is indeed the captured person before sending him the video or erasing it. This paper presents a possible solution when the subject has a head mounted camera, e.g. Google Glass. We propose to create a unique signature, based on pattern of head motion, that could identify that the subject is indeed the person seen in a video. Unlike traditional biometric methods (face, gait recognition etc.), the proposed signature is temporally volatile, and can identify the subject only at a particular time. It is of no use for any other place or time.

1 Introduction

Most people are captured on security cameras many times every day. In addition to high security places like airports, train, and bus stations, cameras are also installed in most shops. Most recorded video is kept in databases for a long time. With the increasing popularity of wearable cameras, the number of times each person is recorded on a video by complete strangers is going to increase substantially. In many countries it is a basic right of people to learn what information about them is kept in databases, and in some cases even to request removal of such information. While this issue has been approached extensively in text based databases, the case of video recordings is yet to be resolved.

Consider the case where a pedestrian is possibly captured by a static security camera, or a moving wearable camera. For the purpose of this paper we refer to the pedestrian as *subject* and the entity holding the video of the subject as *observer*. The subject would like to query if he has been captured in observer's video. The subject can provide an identity signature to the observer for matching in his video. Since the subject and the observer do not trust each other, the signature should not reveal more information about the subject than what is

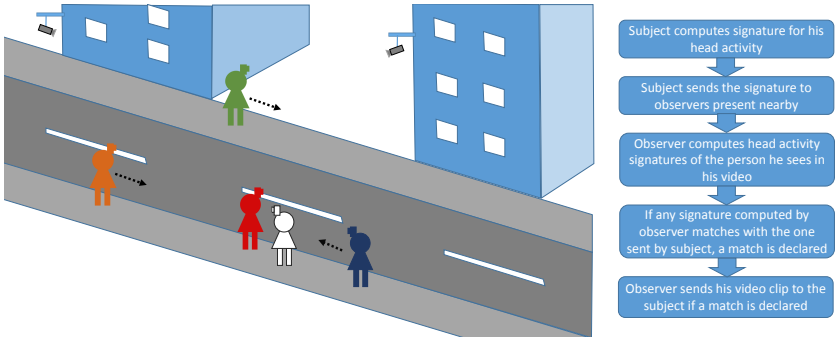


Fig. 1. With the advancement of technology, there is a growing possibility to get captured by surveillance and wearable cameras. In the schematic above, the *subject* (in white) might be captured on video by several *observers*: surveillance cameras, the person he is interacting with him (red) and multiple people around him with wearable cameras. All these observers might have captured the subject on their video. The observers can share their video with the subject or may be asked to erase it from their storage.

being already held by the observer. For example, subject giving observer an image of his face is ruled out, since even if the observer has never captured the subject, he will know following the query the identity of the subject, and he could even use the face he received and search for the subject in other videos. Similar argument hold against most biometric signatures like Gait patterns etc. [1]. Another important consideration is to ensure that the subject is the owner of the claimed identity. For example, the subject may try to impersonate any other person in order to extract videos from innocent observers. Therefore even if the subject provides a face classifier which matches against a person the observer sees, it does not prove that the subject is indeed the person being watched. After all, anybody can create a face classifier of President Obama or Shakira!

Additional potential application of such privacy preserving authentication scheme is for video sharing. Video sharing is particularly necessary since a wearable camera can not capture the wearer. If the wearer would like to see himself in video, the video must be taken by other people's cameras. Consider again the scenario where a pedestrian is being captured in an observer's camera. The subject should be able to prove to the observer that he has been captured in the video, and request to share his video with him. The observer may not be willing to share his entire video, but might agree to share the portion of the video where the subject appears. Such arrangement preserves the privacy of all parties involved: subject, observer and other persons appearing in observer's video.

The problem of video based authentication scheme which does not violate privacy lies in the general framework of privacy preserving secure multiparty communication (SMC) [2]. In a general SMC problem, the two parties hold a portion of data and want to evaluate a function on the union of the data without revealing their data to each other. In our problem, the function to be evaluated

is whether the subject appears in a video clip and the data held by the two parties is the video held by the observer and the identification information held by the subject. Secure multiparty communication techniques are known to be computationally intensive with large communication overhead. This makes their application hard for large data sets such as images or videos. On the other hand, domain specific constraints applicable to these large data sets allow to devise new strategies which are not generic but are efficient for targeted problems. Avidan and Butman [3] address several techniques and applications of privacy preserving computation protocols within computer vision context. Privacy preserving content based image retrieval is addressed by [4]. While both methods provide means against leaking private information, they do not address our requirement of verification of ownership of data, which leaves them vulnerable to impersonation attacks in the scenarios we present.

The focus of the paper is to suggest a novel protocol and an algorithm for privacy preserving signatures for querying and certifying the identity of a person captured in a video. We propose a novel use of videos from head mounted camera (a.k.a. egocentric videos) to generate temporally volatile personal signature of the wearer based upon his head motion. Most importantly, this signature does not contain any private information of the wearer. Intuitively the instantaneous optical flow in egocentric video is dominated by the motion of wearer's head. E.g., if a wearer moves his head to the left the optical flow for most parts of the frame is to the right and vice-versa. The optical flow over a set of frames therefore provide a compact representation of wearer's *head activity* (a sequence of instantaneous head motions) over a short period of time. We show that the head activity at a resolution of $1/30^{\text{th}}$ of a second along with coarse location and time information is discriminative enough to differentiate one person from another. The signature consists of relative motions of the face with respect to the torso. When the observer's face have enough pixels to be recognizable, there is enough resolution for computing the signature as well. Older low resolution cameras, where people appear as blobs, may not have enough resolution to work with the proposed techniques. However, such low resolution videos have no use in the context of privacy preservation and video sharing application anyway.

The organization of the paper is as follows. We begin with describing the first step in our algorithm, which is to compute the signature of the subject's head activity. Section 2 describes how to compute these signatures from subject's egocentric video. Section 3 describes how to compute head activity signatures from observer's point of view. In Section 4 we propose a method for matching the two signatures and present theoretical bounds for the uniqueness of the signatures. Section 5 details results on various experiments conducted by us to ascertain the accuracy of the proposed matching scheme. We conclude with our thoughts on future work in Section 6.

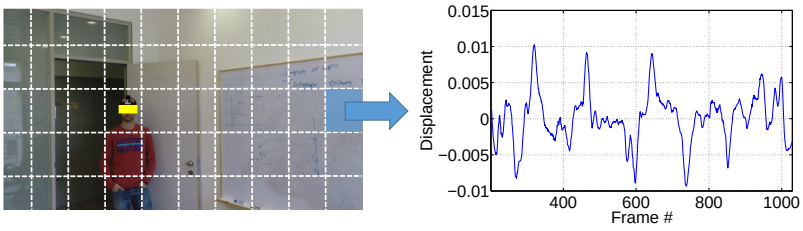


Fig. 2. Optical flow is computed at fixed image locations in the egocentric video. The optical flow is mostly proportional to the angular velocity of the wearer's head. The left image shows the fixed image locations (blocks) in which the optical flow is computed. The graph on the right shows an example of optical flow for one specific block. We estimate the optical flow in 50 such blocks and average them to compute a single global motion vector. Concatenation of the x and y components of this motion vector over a period of time is used as a signature of wearer's head activity.

2 Head Activity Signature from Subject's Camera

An ideal way of computing subject's head activity from the subject's egocentric video is to estimate the camera's egomotion and pose at each frame. Computing egomotion is a well studied area in computer vision [5, 6] with various commercial and research software available [7–9]. However, egomotion computation becomes difficult in the case of egocentric videos due to large and rapid changes in the viewing direction caused by the wearer's natural head motion. Unstructured environment coupled with lack of constraints on lighting and moving objects in the scene make the problem further challenging. Our experiments with various egomotion computation software [7–9] did not yield much success.

Computing instantaneous optical flow between two consecutive frames is a more robust estimation procedure [10]. While it doesn't yield exact camera location and pose, it provides us with enough information for our needs, as we explain next. We note that optical flow observed in an egocentric video comprises of two main components. The first component is in radially outwards direction due to forward motion of the camera wearer. The second and more dominant component is due to the head motion (rotation) of the wearer. Neglecting the first component, the optical flow observed in the egocentric video is proportional to the angular velocity of the wearer's head and can be considered as a signature of wearer's head activity. The change from egomotion to optical flow allows the head activity signatures to be computed in a robust and efficient manner, making it attractive for mobile devices with relatively low compute power.[3]

[3] We note that inertial devices (as used by [11]) could have been used for computing head activity signatures as well. However, our experiments with such inertial devices have yielded a very noisy signal which is not useful for our case. In any case, the requirement of additional hardware restricts the potential application areas, while using image based solution widens the scope of application.

Ignoring perspective effects in the instantaneous optical flow, we seek to estimate frame to frame homography using a translation only model. It would have sufficed to estimate a single homography in the ideal case. However, given the likelihood of moving objects in the scene, we divide the frame into non-overlapping tiles and compute the translation independently for each tile (see Fig. 2). In our experiments we divide the frame to 10×5 non-overlapping tiles and compute the optical flow using our own implementation of LK [12], similar to [10]. We chose LK due to its efficiency, simplicity and robustness. Other methods for optical flow estimation can be used as well [13].

Each tile in the frame gives an independent optical flow estimation. We average the x and y components independently to arrive at single two dimensional motion vector for every two consecutive frames. There are more robust methods than averaging, but our experiments show that simple averaging is sufficient. Concatenating these motion vectors over a period of time gives a signature of wearer's activity over the duration. Formally, let $(u_t^{i,j}, v_t^{i,j})$ be the (x, y) optical flow computed for tile $(i, j) \in (M \times N)$ at frame t. The average motion vector for frame t is defined as $(\bar{u}_t, \bar{v}_t) = (\frac{1}{MN}\sum_{i,j} u_t^{i,j}, \frac{1}{MN}\sum_{i,j} v_t^{i,j})$. The *subject's signature* for time period $[a, b]$ is then:

$$\mathcal{S}_{[a,b]} = ((\bar{u}_a, \bar{v}_a), (\bar{u}_{a+1}, \bar{v}_{a+1}), ..., (\bar{u}_b, \bar{v}_b))^T. \quad (1)$$

In the context of our problem, the subject computes the signature $\mathcal{S}_{[a,b]}$ whenever he'd like to query observers whether he appears on their video or not. The signature is then sent to the observer who then computes another signature, the *observer's signature* using a procedure we describe in next section. The observer then matches the two signatures to verify that the subject is visible in his video. It may be noted that although we call the signatures as subject's and observer's signature, they both describe the activity of subject's head. The notation is only to disambiguate who computes the signature.

It may be noted that in the proposed protocol, the subject need not have seen the observer to obtain his own head activity signature. However, the observer must see the subject to be able to obtain a signature of the subject's head activity. Therefore, the requirement of having seen each other in the protocol is asymmetric and is reflective of prevailing situation in surveillance as well as video sharing applications where the subject might not have seen the observer.

3 Head Activity Signature from Observer's Camera

The observer is willing to share (or erase) parts of his video with subjects who can provide evidence that they appear in the observer's video. In the previous section we presented the subject's signature, which can serve as an evidence for the subject's appearance in the observer's video. In order for the observer to verify the signature, he first needs to find candidate subjects within his video. Therefore, the observer first checks if there are any candidate subjects visible in his video clip. We assume the availability of off-the-shelf person/pedestrian detector for this purpose [14]. We then detect and track feature points on the head

(a) (b) (c)

Fig. 3. Signature from observer's camera. (a) The original feature point locations on the subject (in blue) and in other frames (in red). The observed displacement is due to the combination of movement of candidate subject's head, torso as well as motion of the observer. (b) Tracked feature point after warping by the homography (shown in red). All the points on the torso are mapped to their original location. However, the feature point on the head is mapped to a different location due to change in the subject's head pose. (c) Curve showing the subject's head activity computed as concatenation of displacement of warped point on the head, over a period of time.

and torso of the candidate subjects separately. Any human parts based model can be used for detecting head and torso [15]. We note that the instantaneous displacement of a feature point on the head of a candidate subject (as seen from the observer's camera) has three major components. The first component is due to the walking of the candidate subject. As the candidate subject walk towards or away from the observer, there is a displacement in the feature point. This is true even when the head of the candidate subject was perfectly stationary with respect to his torso. The second component is due to the observer's camera motion. The third component is due to the relative motion of candidate subject's head with respect to torso. For computing observer's signatures we are interested in finding a function measuring the third component and invariant to first and second components (see Fig. 3).

Let us assume that the observer is sufficiently far from the candidate subjects he captures on the video, such that the body of a candidate subject (face as well as torso) could be treated as a plane. This makes it possible to express the displacement of feature points on the candidate subject's body by finding a planar homography. We detect and track multiple points on the torso and head of each of the candidate subjects appearing in the observer's video. We fit a homography (with respect to some reference frame r) using only the points on torso. Ideally, transforming the points using the computed homography should cancel the displacement in the feature points due to the motion of the torso. Any remaining displacement (x as well as y) observed in the feature points on the head of the candidate subject is entirely due to the change in the candidate subject's head pose with respect to his torso. The concatenation of the remaining displacement computed as described above over a time duration gives the signature of head activity of the candidate subject from observer's point of view (see Fig. 3).

Formally, let $\mathcal{C}_t^i$ be the i^{th} candidate subject detected in the observer's video at time $t \in [a, b]$ and let $\mathcal{P}_{i,t}^{\text{torso}}$ be the sets of feature points tracked on the torso

of $\mathcal{C}^i$. For simplicity, let us have a single feature point $p^{\text{head}}_{i,t} = (x, y)$ tracked on the head of $\mathcal{C}^i$ at time t and let $r \in [a, b]$ be a reference frame of our choice, in which $\mathcal{C}^i$ appears. For each frame t, we find an homography $\mathcal{H}_{t\to r}$ using $\mathcal{P}^{\text{torso}}_{i,r}$ and $\mathcal{P}^{\text{torso}}_{i,t}$. We then calculate the warped point $w^{\text{head}}_{i,r} = \mathcal{H}_{t\to r} \cdot p^{\text{head}}_{i,t}$. The displacement $\delta_{i,t} = (\delta^x_{i,t}, \delta^y_{i,t}) = w^{\text{head}}_{i,r} - p^{\text{head}}_{i,t}$ is proportional to the head pose of $\mathcal{C}^i$ relative to his torso. We concatenate the displacements for all $t \in [a, b]$ to get: $\widetilde{\mathcal{O}}_{i,[a,b]} = \Big((\delta^x_{i,a}, \delta^y_{i,a}), (\delta^x_{i,a+1}, \delta^y_{i,a+1}), \ldots, (\delta^x_{i,b}, \delta^y_{i,b})\Big)^T$. Note that while the subject's signature $\mathcal{S}$ is proportional to the subject's head angular velocity, $\widetilde{\mathcal{O}}$ is proportional to the change in head displacement with respect to frame r. We therefore temporally derive $\widetilde{\mathcal{O}}$ to get the *observer's signature*: $\mathcal{O} = \frac{d\widetilde{\mathcal{O}}}{dt}$

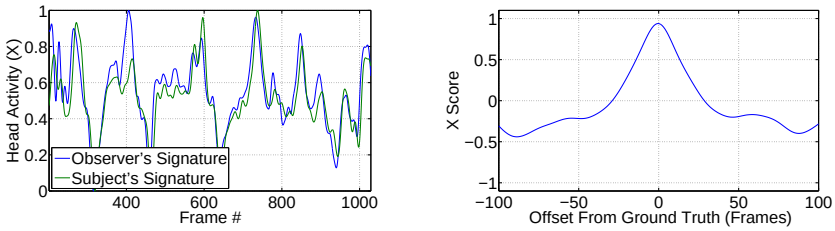
In our implementation we manually select one feature point on the head and at least 10 feature points on the torso of each candidate subject in the first frame of the sequence. In case of multiple feature points on a candidate subject's head, the average of the $\delta_{i,t}$ displacements per frame can be used as the elements of $\widetilde{\mathcal{O}}$.

It may be noted that, while the subject's signature indeed describe the angular velocity of the head, observer's signature describes it as observed after projecting it on the observer's image plane. The two signals are therefore not identical. However, the projection can only change the magnitude of instantaneous displacement but not the sign of the displacement. The matching strategy as we describe in the next section should therefore ignore the magnitude and focus on the sign of instantaneous displacement.

4 Matching Head Activity Signatures

Once a subject has presented his signature $\mathcal{S}$, the observer would like to verify if the subject's signature matches with the observer's signatures corresponding to any of the candidate subjects that appears in the observer's video. It may be noted that the subject's and the observer's signatures have been produced from two videos which may have very little in common (looking to opposite sides, different cameras/resulotion/FPS etc.). Even the derivation process of the signatures is not the same and therefore one might consider the signatures as originating from different modalities. While there are various methods available for matching signals obtained from different modalities [16], we believe that problem in our case is much simpler. Most importantly, both signatures have the same dimension and are measured in pixels. We observe that for our case the scale of the two signatures can be very different but they should agree in their phase for a correct match. We propose using Pearson Correlation Coefficient as a score for signature match. The Pearson's correlation coefficient ρ between two variables X and Y is defined as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]}{\sigma_X \sigma_Y}, \tag{2}$$



(a) (b)

Fig. 4. (a) The curves showing subject's (in green) and observer's (in blue) signatures. (b) Correlation between observer's and subject's signatures at various temporal alignments. The zero offset implies the ground truth alignment. We show the correlation scores with various alignments in the range of ± 100. The score is significantly higher at the time instance of correct match (offset = 0).

where Cov is the covariance between X and Y, σ_X is the standard deviation of X, μ_X is the mean of X and $\mathbb{E}$ is the expectation. We denote the x and y components of the subject's signature as $\mathcal{S}_x, \mathcal{S}_y$. Similarly, the components of the observer's signature is denoted by $\mathcal{O}_x, \mathcal{O}_y$. We compute independently $\rho_x = \rho(\mathcal{S}_x, \mathcal{O}_x)$ and $\rho_y = \rho(\mathcal{S}_y, \mathcal{O}_y)$. Recall that $\rho \in [-1, +1]$. We add $|\rho_x|$ and $|\rho_y|$ to obtain a total score for the match.[4] In an abuse of notation, we call the total score as *correlation*, ρ_x as x-correlation, and ρ_y as y-correlation.

Fig. 4 shows the matching between signatures for an indicative experiment. The subject provides his signature with an indication of time where he claims to be present in observer's video. Observer computes observer's signature but tries matching it with various alignments around the time claimed by subject. This is to allow synchronization errors between subject and observer. In our implementation, we have empirically chosen a threshold of $T = 1.1$ (which is just above the half way mark) and declare that the signatures match if the total score is greater than this threshold. Note that we claim and show in the experiments later that the signatures are unique and even if observer would have chosen to match over the entire range for which subject's signature was available, he would have found a match only in case of valid subject at the correct time. The decision to search only in the search window around the claimed time is due to efficiency.

4.1 Signature Uniqueness

An important consideration at this stage is to quantitatively evaluate uniqueness of the signatures. The question we seek to answer is: How hard is it for a malicious attacker to fool the observer into sharing his video clip by 'guessing' another

[4] The observed displacement in observer's signature could be opposite or in phase with subjects's signature depending upon the case that the subject is seen from front or back by the observer.

subject's signature? For the discussion in this section we consider each of the $\mathcal{S}_x, \mathcal{S}_y, \mathcal{O}_x$ and $\mathcal{O}_y$ as vectors in d-dimensional space, where d is the duration of the signatures in frames. Note that correlation score is invariant to scale and shift. Therefore, we assume each vector to be normalized to unit norm and zero mean. The question we are asking now is: How easy it is for a malicious attacker to 'guess' signature vectors $\mathcal{S}_x$,$\mathcal{S}_y$ such that the correlation with the observer's vectors $\mathcal{O}_x$, $\mathcal{O}_y$ comes out to be more than $\frac{T}{2} = 0.55$ per vector (x and y).[5]

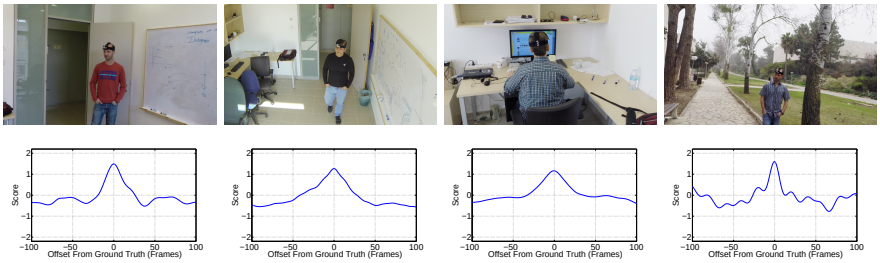
Geometric interpretation of the correlation coefficient views it as the cosine of the angle between the two vectors [17, 18]. A correlation score of $\frac{T}{2} = 0.55$ corresponds to an angle of about 60° between the vectors. With this interpretation it is easy to observe that the probability of getting another vector within 60° degrees of the first one is practically zero for a large enough d. In other words, for large enough dimensions, two random vectors sampled uniformly are almost orthogonal with probability 1.[6] In our context it implies that if the values of the signature at different time instance are $i.i.d$ (identically independent distributed), then the chances of hitting a correlation of $\frac{T}{2} = 0.55$ (or for that matter any non-zero correlation) by random guessing the vector is practically zero.

The above argument is naïve in the sense that we assume the values are i.i.d. In our case, the value of the signature at a particular time instance represent the velocity of the head at that time instance. Therefore, the velocities at two consecutive time instances are strongly dependent. The head of a human being never comes to still immediately after moving at some other velocity or the other way around. A more reasonable approach would be to bound the acceleration of the head motion. Let us assume that the difference of values of the signature (head velocity) at two consecutive time instance can not be more than ϵ. Assume the signature value is ϵ at a time instance t. With the bound on acceleration, the possibilities for the next value are 2ϵ, ϵ or 0. We restrict the possibilities further and say that the next value can only be ϵ or 0. Note that this is equivalent to providing additional information to malicious subject. We show that even with this additional information it is not possible for the malicious observer to guess the signature vector. Observe that with the additional restriction, the vector can be treated as a binary vector. It is easy to see that for sufficiently high dimension, the chance of hitting non-zero correlation, even for binary vector, is practically zero.

The above theoretical analysis is still naïve in the sense that we do not account for patterns that may arise from unique individual behaviour observed over long time (i.e. collecting millions of signatures of the same subject). We are not aware of any work that attempts to deal with such long term patterns from egocentric videos. However, the possibility of such pattern-based attacks can not

[5] Any unequal division of the total score requirement would be more difficult to meet.

[6] In our implementation, the dimension of the vectors (length of the signature) is usually more than 200 frames (corresponding to 3-4 seconds of video at 60 frames per second). This is a sufficiently large dimension for the proposed probabilistic analysis.



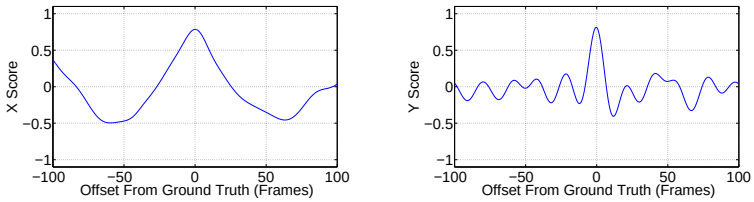
(a) Standing (b) Surveillance (c) Workstation (d) Crossover

Fig. 5. Matching scores between subject's and observer's signature at different temporal offsets as described in Fig. 4. First row shows a sample frame for each sequence. (a) In this experiment subject and observer are standing in place and talking. Note that the two are not exactly stationary and there is a bit of walking by observer towards the subject. (b) Matching scores for the case when the subject is captured walking towards a surveillance camera hanged from a room's ceiling. (c) The proposed scheme also works in the scenario when observer haven't seen the 'face' of the subject. In this case, subject is working on the computer and observer is watching his back. The matching score can still accurately find the match at correct time instance. (d) The proposed scheme can also handle cases where the subject and/or observer are moving. In this scenario, the subject and observer are walking towards each other. The original video clips corresponding to results shown here can be found at the project page.

be ruled out. We validate the theoretical analysis in this section with empirical evidence in Section 5.

5 Experiments

We have conducted our experiments using both self shot videos and a publicly available dataset [19, 10]. We have used GoPro Hero3+ camera in narrow view mode for shooting our videos. The videos are in full HD resolution at 60 FPS and are available at the project page: `http://www.vision.huji.ac.il/egosig/`. We use our own implementation of LK for computing the subject's signature. We divide each frame to 10×5 grid and compute optical flow independently in each grid region. The choice of grid is to be robust against moving objects in the scene and other sources of optical flow failure. Other choices of grid size did not yield much improvement in the results. For computing the observer's signature we detect GFTT features and track them using the LK. Both implementations are available in OpenCV [20]. The points are detected and tracked separately in head and torso regions. For the head region, we choose one point visible for longest duration, whereas for torso region we use all the points visible during entire length of the video. The homography between the torso points is computed using Matlab's geometric transformation estimator. The homography estimation is done by using RANSAC with outlier removal options set. We observe that homography estimation is not stable between frames separated by long time



(a) (b)

Fig. 6. (a) In case of walking there is a periodicity in subject's head activity due to natural head motion associated with walking steps. This leads to weak periodicity in the x-correlation scores. However, the head activity is not exactly same and there is still a significant peak in the correlation score at correct time instance. (b) The y-correlation scores are not effected.

duration. We therefore choose a new *hop-over* frame after every 60 frames (1 second in our videos). The homography is computed between the current and the hop-over frame and then multiplied with the homography between the hop-over and the reference frame to find the overall transformation. The time duration for the observer's signature is chosen at-least 200 frames to ensure low false-positive rates. The signatures are declared as a match if there is a correlation score (sum of x and y correlations) of more than $T = 1.1$ observed at any point. The time of match is declared at the point of maximum correlation score.

We conducted experiments under various interaction scenarios between the observer and the subject. Fig. 5 shows the correlation scores for indicative experiments. As described in previous section, we compute the observer's signature and then find the correlation score with subject's signature in a window of ± 100 frames around the time claimed by subject. Note that in our self-shot experiments the videos are approximately synchronized and therefore we should see a peak at origin. The proposed scheme correctly recognizes the subject while standing, talking or walking. The signatures can be successfully matched, even when the observer sees the back of the subject. This exposes the strength and novelty of the proposed method with respect to face pose estimation approaches. Not only is the face pose estimation much harder inference compared to proposed tracking based approach, but the face pose estimation by definition can be only applied if observer sees the 'face' of the subject.

We notice that there is some periodicity in the matching score in the walking sequence. The reason for this could be understood from the fact that there is a natural head motion associated with the stepping during walking. Since the stepping speed doesn't change too much over a period of few seconds there is a gross similarity in the head activity in x direction (see Fig. 6). The periodicity is limited to x direction only and is not visible in y-correlation score. Even for x correlation, although the head motion is periodic at gross level, the activity don't match precisely between two (walking) steps. Therefore, the peak in the score is still observed at correct time instance.

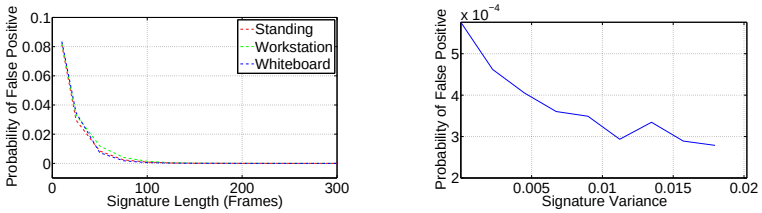


Fig. 7. (a) Average probability of false positive for various signature length corresponding to observer's signature obtained from different sequences. The false positive probability is non-zero for short signatures but quickly goes to 0 for signatures longer than 100 frames. In our implementation we use signatures of length more than 200. (b) Probability of false positive with respect to signatures of different variance. The length of the signature chosen in this case is 100. The probability is practically zero even for signatures with low variance.

In Section 4.1, we presented theoretical analysis on the uniqueness of the proposed matching technique. We showed that the probability of a subject to randomly 'guess' a signature which can match observer's signature with a correlation score of more than $T = 1.1$ ($\frac{T}{2} = 0.55$ for each x and y-correlation) is practically 0. The results holds for sufficiently large dimension and we claimed that the dimension of 200 in our implementation is sufficiently large. We validate our claim with experimental evidence here. We created a repository of more than a million subject's signatures based on the videos from GeogiaTech's First-Person Social Interactions Dataset [19]. In all, the repository is based on more than 30 hours of egocentric videos. For the observer's signatures, we used the signatures from the workstation, standing and whiteboard sequences that were shot by us. We evaluated the matching score between each observer signature and the entire and subject's signatures repository. The probability of a false match is the number of instances where the correlation of more than 1.1 is observed against the number of evaluations. We repeat this experiment for various signature length. Fig. 7(a) shows the results. Expectedly, the probability of false positive is non-zero for shorter signatures, but goes to 0 quickly for signatures of length more than 100 frames. We also confirmed by choosing observer's signature having different variance. Fig. 7(b) shows the results. The probability of a false match is practically 0 even for signatures with low variance. To further verify the above findings, we repeated this experiment with a slight change. Instead of matching the observer's signatures (which are based on our self shot videos) with the subject's signature repository, we randomly picked signatures from the same repository and matched them against the rest of the repository. This process yielded similar results, which proves the uniqueness of the signatures.

One can think of special cases where even long signatures of more than 100 frames can be guessed. One such example is a subject's signature that represents no head activity at all (signature is a 'flat line'). E.g, the subject's camera is

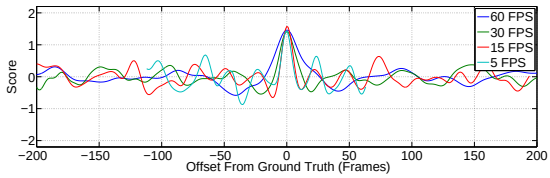


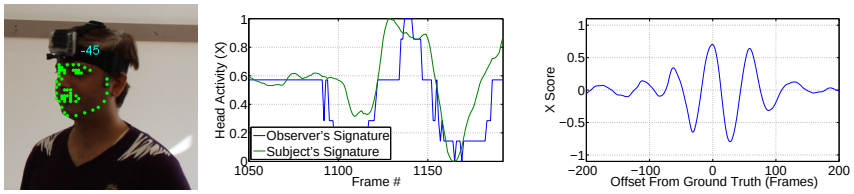
Fig. 8. Comparison of correlation scores for videos at various frames per second (FPS) corresponding to whiteboard sequence with signature length of 200 frames. We observe that although the best correlation score is similar at all FPS, the signal to noise ratio, measured as the ratio of highest peak to second highest is best at 60 FPS.

placed on a statue's head. It is clear that if the observer is presented with multiple such signatures from multiple statue-like subjects, there is no way of telling who produced which signature. A simple way to overcome this case is to require that the signature contain a minimal amount of information, measured as variance or entropy.

The experiments conducted as above merely serve the purpose to show that it is practically impossible to randomly guess the signature. The question of whether it is possible or not to make an 'educated' guess of a subject's signature (based on long-term observation on the specific subject) remains open.

It may be noted that the correlation scores are not strongly dependent upon the frame rate of input videos. Although we use videos at 60 frames per second (FPS), the result do not change much at lower FPS. Fig. 8 shows the score comparison at various FPS corresponding to the whiteboard sequence. The length of the signature is chosen to be same for all videos. Note that the correlation peak is at the correct place and of similar strength at all FPS. However, the signal to noise ratio measured as ratio of highest peak to second highest peak is best for 60 FPS. We therefore, recommend videos at 60 FPS for the problem. The 60 FPS videos have additional advantage that the observer has to keep the subject in view for a shorter amount of time (for same signature length) thereby implying more flexibility.

Using head activity signature from egocentric video as proposed in the paper is simple, efficient and robust enough for the problem we are considering. However, more sophisticated approaches for the signatures could have been used. For example, face pose estimation has been a well studied problem in computer vision community and various algorithms have been proposed recently for the same [21, 22]. We have experimented with [21] for which the source code was publicly available. Our experiments showed ambiguity in the matching due to quantized nature of face pose output. Figure 9 shows the result. It is important to note that using face pose estimation from the said method to infer head activity will restrict our method to forward moving sequences with x head motion only. Furthermore, it would restrict us to cases when the face is seen clearly and at enough resolution. Therefore, we are not advocating its use in the context of our method.



(a) (b) (c)

Fig. 9. (a) Face features and pose estimated by [21]. (b) Observer's signature generated using pose estimation. Also shown is the subject's signature for visual comparison. Note that [21] gives coarse pose estimation in step size of 15 degrees only. The same is visible in observer's signature which can only take few discrete values now. (c) Matching score at various offset. While the peak is at the correct location, there is another high correlation at an offset of 60 frames from ground truth. This is corresponding to matching with next head turn as visible in (b).

6 Conclusion

A novel method for privacy aware sharing of egocentric videos has been presented. The proposed method paves the way for exciting applications by enabling a camera wearer to access the video clips who might have captured him. The focus of the paper is not new technology for tracking or pose detection but to use simple existing techniques to offer a practical solution for the privacy concerns associated with video capture and sharing. There is no personal information disclosure by either parties other than sharing of the requested video clip in which the subject appears. The head activity signatures are temporally volatile and can not be easily used to recognize the subject at any other time or place. Yet, the computed signatures are unique enough to distinguish the correct signature amongst the various signatures presented. The technique relies on simple steps of detecting and tracking features for which many efficient algorithms are available. This broadens the scope of application by making it attractive for mobile devices. The robustness of the tracking algorithms enables the algorithm to be applied from a distant or low resolution cameras as well. The signatures do not depend upon the visibility of a 'face' and can be computed even when the observer see the subject from the back.

A possible weakness of the current algorithm is in requirement to see the head for the duration of the signature. The duration required is not large (typically a few seconds) but even this small duration can be a restriction in an egocentric setting, where the observer's view point may be changing quickly due to natural head motion. We note that there is a possibility of reacquiring the feature points and allowing for 'holes' in the signature during matching. This is an interesting possibility which we would like to explore in the future research.

Acknowledgement: This research was supported by Intel ICRC, by Israel Ministry of Science, and by Isreal Science Foundation.

References

1. Shiraga, K., Trung, N.T., Mitsugami, I., Mukaigawa, Y., Yagi, Y.: Gait-based person authentication by wearable cameras. In: International Conference on Networked Sensing Systems. (2012) 1–7
2. Yao, A.C.C.: How to generate and exchange secrets. In: FOCS. (1986) 162–167
3. Avidan, S., Butman, M.: Blind vision. In: ECCV. Volume 3953. (2006) 1–13
4. Upmanyu, M., Namboodiri, A., Srinathan, K., Jawahar, C.: Efficient privacy preserving video surveillance. In: ICCV. (2009) 1639–1646
5. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. 2 edn. Cambridge University Press, New York, NY, USA (2003)
6. Raudies, F., Neumann, H.: A review and evaluation of methods estimating egomotion. CVIU **116** (2012) 606–633
7. Castle, R.O., Klein, G., Murray, D.W.: Video-rate localization in multiple maps for wearable augmented reality. In: IEEE ISWC. (2008)
8. Wu, C.: VisualSFM : A visual structure from motion system. (`http://ccwu.me/vsfm/`)
9. VISCODA: Voodoo camera tracker. (`http://www.digilab.uni-hannover.de/`)
10. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: CVPR. (2014)
11. Spriggs, E., Torre, F.D.L., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: CVPRW. (2009)
12. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI. (1981) 674–679
13. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. IJCV **92** (2011) 1–31
14. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: Survey and experiments. TPAMI **31** (2009) 2179–2195
15. Ramanan, D.: Part-based models for finding people and estimating their pose. In: Visual Analysis of Humans. Springer (2011) 199–223
16. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: CVPR. (2005) 88–95
17. Jr., J.S.: The relationship between the coefficient of correlation and the angle included between regression lines. The Journal of Educational Research **41** (1947) 311–313
18. Wikipedia: Pearson product-moment correlation coefficient. (`http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`)
19. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR. (2012)
20. Bradski, G.: Opencv ver 2.4.3. (2013)
21. Ramanan, D., Zhu, X.: Face detection, pose estimation, and landmark localization in the wild. IEEE Conference on Computer Vision and Pattern Recognition (2012) 2879–2886
22. Ho, H.T., Chellappa, R.: Automatic head pose estimation using randomly projected dense sift descriptors. In: ICIP. (2012) 153–156