# Integrating Egocentric Videos in Top-view Surveillance Videos: Joint Identification and Temporal Alignment

Shervin Ardeshir[0000−0001−5760−1665] and Ali Borji[0000−0001−8198−0335]

Center for Research in Computer Vision (CRCV)
University of Central Florida, Orlando, FL. USA

**Abstract.** Videos recorded from first person (egocentric) perspective have little visual appearance in common with those from third person perspective, especially with videos captured by top-view surveillance cameras. In this paper, we aim to relate these two sources of information from a surveillance standpoint, namely in terms of identification and temporal alignment. Given an egocentric video and a top-view video, our goals are to: a) identify the egocentric camera holder in the top-view video (self-identification), b) identify the humans visible in the content of the egocentric video, within the content of the top-view video (re-identification), and c) temporally align the two videos. The main challenge is that each of these tasks is highly dependent on the other two. We propose a unified framework to jointly solve all three problems. We evaluate the efficacy of the proposed approach on a publicly available dataset containing a variety of videos recorded in different scenarios.

## 1 Introduction

The widespread use of wearable devices such as GoPro cameras and smart glasses has created the opportunity to collect first person (egocentric) videos easily and in large scale. People tend to collect large amounts of visual data using their cell phones and wearable devices from the first person perspective. These videos are drastically different from traditional third person videos captured by static surveillance cameras, especially if the third person camera is recording top-down, as there could be very little overlap in the captured frames by the two cameras. Even though a lot of research has been done studying these two domains independently, relating the two views systematically has yet to be fully explored. From a surveillance standpoint, being able to relate these two sources of information and establishing correspondences between them could lead to additional beneficial information for law enforcement. In this work, we take a step towards this goal, by addressing three following problems:

**Self-identification:** The goal here is to identify the camera holder of an egocentric video in another reference video (here a top-view video). The main challenge is that the egocentric camera holder is not visible in his/her egocentric video. Thus, there is often no information about the visual appearance of the camera holder (example in Fig. 1).

**Human re-identification:** The goal here is to identify the humans seen in one video (here an egocentric video) in another reference video (here a top-view video). This problem has been studied extensively in the past. It is considered a challenging problem due to variability in lighting, view-point, and occlusion. Yet, existing approaches

Fig. 1: A pair of top- (left) and egocentric (right) views. Self identification is to identify the egocentric camera holder (shown in red). Human re-identification is to identify people visible in the egocentric video, in the content of the top-view video (orange and purple).

assume a high structural similarity between captured frames by the two cameras, as they usually capture humans from oblique or side views. This allows a rough spatial reasoning regarding parts (e.g., relating locations of head, torso and legs in the bounding boxes). In contrast, when performing human re-identification across egocentric and top-view videos, such reasoning is not possible (examples are shown in Figs. 1 and 2).
**Temporal alignment:** Performing temporal alignment between the two videos directly is non-trivial as the top-view video contains a lot of content that is not visible in the egocentric video. We leverage the other two tasks (self identification and re-identification) to reason about temporal alignment and estimate the time-delay between them.

The interdependency of the three tasks mentioned above encourages designing a unified framework to address all simultaneously. To be able to determine the camera holder's identity within the content of the top-view video (task 1), it is necessary to know the temporal correspondence between the two videos (task 3). Identifying the people visible in the egocentric video in the content of the top-view video (task 2), would be easier if we already knew where the camera holder is in the top-view video at the corresponding time (tasks 1 and 3), since we can reason about who the camera holder is expected to see at any given moment. Further, knowing the correspondence between the people in ego and top views, and temporal alignment between two videos (tasks 2 and 3), could hint towards the identity of the camera holder (task 1). Finally, knowing who the camera holder is (task 1) and who he is seeing at each moment (task 2) can be an important cue to perform temporal alignment (task 3). The chicken-and-egg nature of these problems, encourage us to address them jointly. Thus, we formulate the problem as jointly minimizing the total cost $C_{tot}(l_s, L_r, \tau)$, where $l_s$ is the identity of the camera holder (task 1), $L_r$ is the set of identities of people visible in the egocentric video (task 2), and $\tau$ is the time offset between the two videos (task 3).
**Assumptions:** In this work, we hold assumptions similar to [1]. We assume that bounding boxes and trajectories in top-view are given (provided by the dataset). Therefore, an identity in top-view refers to a set of bounding boxes belonging to one person over time. We further assume that the top-view video contains all the people in the scene (including the ego-camera holder and other people visible in the ego video).

Fig. 2: Sample ego- and top-view bounding boxes. Unlike conventional re-identification instances, rough spatial alignment assumptions do not hold.

## 2 Related Work

**Self-identification** and self-localization of egocentric camera holder have been studied during the last few years. [2] uses the head motion of an egocentric viewer as a biometric signature to determine which videos have been captured by the same person. In [3], egocentric observers are identified in other egocentric videos by correlating their head motion with the egomotion of the query video. Authors in [4] localize the field of view of egocentric videos by matching them against Google street view. Landmarks and map symbols have been used in [5] to perform self localization on a map, and [6, 7] use the geometric structure between different semantic entities (objects and semantic segments) for the problem of self-localization, by relating them to GIS databases. The closest works to ours are [8] and [1, 9]. Please note that the mentioned works [8, 1] do not address the other two problems of re-identification and temporal alignment. Our self-identification problem differs from [8] in three main aspects:
**1.** [8] self identifies the egocentric camera holder in a third person video using a fully supervised method. Please note that even though we perform unsupervised and supervised re-identification and use that as a prior, there is no supervision in the self-identification task. **2.** In [8]'s dataset, each egocentric video contains the majority of other identities, which could hold in settings such as sitting and having a conversation. As a result, cropping one person out from the third person video will have a lot in common with the content of that person's egocentric video. In our dataset, however, this is not the case, as many egocentric viewers do not observe each other at all. **3.** In [8]'s dataset, third person videos have a generic ground level viewpoints, which makes them have similar properties to the egocentric videos in terms of spatial reasoning. The difference between first and third person videos are more severe when the third person video is top-view like ours. Nonetheless, we evaluate [8] on our dataset as a baseline. [1, 9] approached the problem of egocentric self-identification in top-view videos for the first time, leveraging the relationship among different egocentric videos. However, this method is highly dependent on the completeness of the egocentric set and performs poorly when there is only one egocentric video. We use this method as another baseline in our experiments.
**Human Re-identification** This problem has been studied heavily in the past (e.g., [10–14, 14–16]). Deep learning methods have recently been applied to person re-identification [17–19]. Yi [20] uses a Siamese network for learning appearance similarities. Similarly Ahmed [21] uses a two stream deep neural network to determine visual similarity between two bounding boxes. Cheng *et al.* [22] uses a multi-channel CNN in a metric learning based approach. Cho *et al.* [23] proposes using pose priors to perform compar-

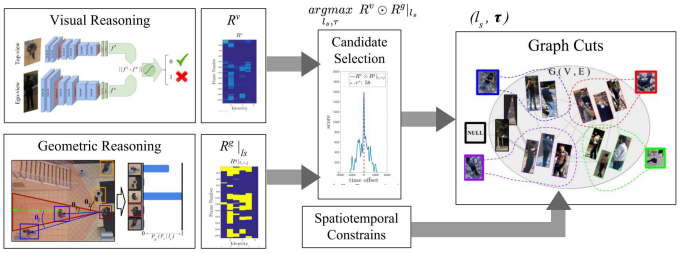



Fig. 3: The block diagram of our proposed method. We use three main cues: visual, geometrical, and spatiotemporal. Visual reasoning is used for initializing re-identification correspondences. Combining geometric and visual reasoning, we generate a set of candidate $(l_s, \tau)$ pairs. Finally, we evaluate the candidates using graph cuts while enforcing spatiotemporal consistency and find the optimum combination of labels and values.

ison between different candidates, and Matsukawa *et al.* [24] uses a region descriptor based on hierarchical Gaussian distribution of pixel features for this task. In the egocentric domain, the study reported in [25] performs person re-identification in a network of wearable devices, and [26] addresses re-identification across time-synchronized wearable cameras. To the best of our knowledge, our work is the first attempt in addressing this problem across egocentric and top-view domains. Visual appearance is often the main cue for human re-identification. This cue can change from one camera to another due to occlusion, viewpoint, and lighting. However, the variation is often relatively low across different static surveillance cameras as the nature of the data is the same (both cameras being ground level or oblique viewpoints). In contrast, in situations where a set of surveillance and egocentric cameras are used, appearance variation is more severe due to egocentric camera motion, more drastic difference in field of views, lighting direction, etc. Thus, we propose a network to unify the representation of human detection bounding boxes across egocentric and top-view videos. In fact our visual re-identification network could be replaced with any other re-identification framework capable of measuring the visual similarity between the egocentric and top-view human detection bounding boxes. We compare and contrast our results with state of the art human identification methods in the experiments section.

**Relating first- and third-person vision:** [27, 28] have explored the relationship between mobile and static cameras for improving object detection. [29] fuses information from the first and third person static cameras and laser range data to improve depth perception and 3D reconstruction. Park *et al.* [30] predict gaze behavior in social scenes using first and third-person cameras. Soran *et al.* [31] have addressed action recognition in presence of one egocentric and multiple static videos and [32] explores transfer learning across egocentric and exocentric actions.

## 3 Framework

We aim to address three different tasks jointly. To find the optimal values for all of the variables in a unified framework, we seek to optimize the following objective:

$$l_s^*, L_r^*, \tau^* = \underset{l_s, L_r, \tau}{argmin}\ C_{tot}(l_s, L_r, \tau) \tag{1}$$

Assuming a set of identities visible in the top-view video as $I^t = \{1, 2, ..., |I^t|\}$, our goal in task 1 is to identify the camera holder (assign a self-identity $l_s$). We assume that the camera holder is visible in the content of the top-view video, thus $l_s \in I^t$. In task 2, we aim to perform human re-identification for the visible humans in the egocentric video. Let $D^e = \{d_1^e, d_2^e, ..., d_{|D^e|}^e\}$ be the set of all human detection bounding boxes across all the frames of the egocentric video. In task 2, we find labeling $L_r = \{l_1^e, l_2^e, ..., l_{|D^e|}^e\} \in |I^t|^{|D^e|}$, which is the set of re-identification labels for human detection bounding boxes. Finally, $\tau$ is the time offset between the egocentric and top-view video, meaning that frame $\tau_0$ in the top-view video corresponds to frame $\tau_0 + \tau$ in the egocentric video. We estimate $\tau$ in task 3. In our notation, we use superscripts to encode the view (*t*: top, *e*: ego).

The block diagram of our proposed method is shown in Fig. 3. Our method is based on three types of reasonings across the two views. First, we perform visual reasoning across the two videos by comparing the visual appearance of the people visible in the top video to the people visible in the egocentric video. This reasoning will provide us some initial probabilities for assigning the human detection bounding boxes in the ego-view to the identities in the top-view video. It gives an initial re-identification prior based on the likelihood of the human detections matching to top-view identities (Sec 3.1). The second cue is designed to geometrically reason about the presence of different identities in each other's field of view in top-view over time (Sec 3.2), providing us cues for re-identification based on self-identification. We then define two spatiotemporal constraints to enforce consistency among our re-identification labels (Section 3.3) in ego view. In the fusion step (Sec 3.4), we combine visual and geometrical reasoning to narrow down the search space and generate a set of candidate $(l_s, \tau)$ pairs. Finally, we enforce spatiotemporal constraints and evaluate the candidates using graph cuts [33].

### 3.1 Visual Reasoning

The first clue for performing re-identification across the two views is to compare appearances of the bounding-boxes. Since in traditional re-identification works both cameras are static, and they have similar poses (oblique or ground level), there is an assumption of rough spatial correspondence between two human detection bounding boxes (i.e. the rough alignment in location of head-torso-leg between two bounding boxes). Since the viewpoints are drastically different in our problem, the rough spatial alignment assumption does not hold. A few examples are shown in Fig. 2. We perform this task in unsupervised and supervised settings. In the unsupervised setting, we extract some generic features from the two views and directly compare their features. In the supervised setting, we design a two stream network capable of measuring similarity across the two views.

**Unsupervised Baseline** For each bounding box $d_i^e$ in the ego-view, we extract VGG-19 deep neural network features [34] $f_i^e$ (last fully connected layer, 4096 dimensional features). We perform L2 normalization on the features. As mentioned before, top-view bounding boxes are tracked and identities have been assigned to each track (set of bounding boxes belonging to each person). Therefore, for identity $j$ in the top-view video, we extract VGG features from all of its bounding-boxes and represent identity j with the average of its feature vectors $f_j^t$.

To enforce the notion of probability, we measure the probability of ego-view bounding box $d_i^e$ being assigned to label $j$ in top-view ($l_i^e = j$) as:

$$P(l_i^e = j) = \frac{e^{-||f_i^e - f_j^t||}}{\sum_{m=1}^{|I^t|} e^{-||f_i^e - f_m^t||}}. \quad (2)$$

**Supervised Approach: Training:** We train a two stream convolutional neural network to match humans across the two views. As illustrated in Fig. 4, each stream consists of convolution and pooling layers, ending in fully connected layers. The output is defined as the Euclidean distance of the output of the last fully connected layers of each stream passed through a sigmoid activation. If the two bounding boxes belong to the same identity, the output is set to zero (and one, otherwise). This forces the network to find a distance measure across the two views.
**Testing:** We feed bounding box $d_i^e$ to the ego stream of the network and extract $f_i^e$ (We perform L2 normalization). In top-view, for identity $j$ we feed all of its bounding-boxes to the top-view stream and represent identity $j$ with the average of its feature vectors $f_j^t$. Similar to the unsupervised approach, we measure the probability of ego-view bounding box $d_i^e$ being assigned to label $j$ in top-view ($l_i^e = j$) according to Eqn. 2.
**Implementation details of the CNN:** We resize each of the top-view bounding boxes to $40 \times 40$, and each ego-view bounding box to $300 \times 100$ in the RGB format (3 channels). Each stream consists of 3 convolutional blocks, each having two convolutional layers and a pooling layer with $2 \times 2$ pooling. The number of filters for the convolutional layers in order are 16, 16, 32, 32, 64, and 64. Finally, each stream projects to two fully connected layers (top stream: 512, 128; ego-stream: 1024, 128). The euclidean distance of the output of the two streams is then passed through a sigmoid activation in order to enforce the notion of probability. We use Adam optimizer with learning rate of 0.001 and binary cross entropy loss, and train the network end-to-end. The hyper-parameters were fine-tuned on the validation set using grid search in logarithmic scale.

### 3.2 Geometric reasoning

Here, we leverage the geometric arrangement of people with respect to each other in top-view and reason about their presence in each other's field of view. We iterate over different identities in top-view, and perform geometric reasoning assuming the identity is the camera-holder. In Fig.5, we illustrate reasoning about the presence of the identities highlighted with blue and orange bounding boxes, assuming the person highlighted in the red bounding box is the camera holder.

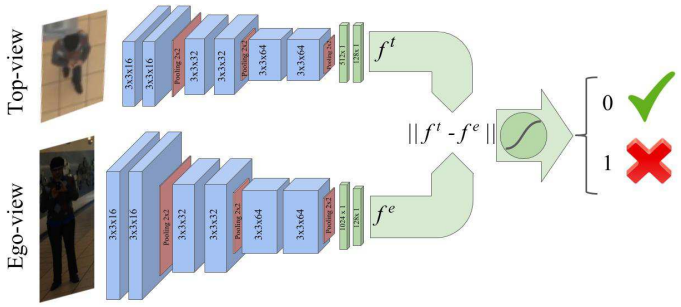



Fig. 4: The architecture of our two stream convolutional neural network trained on pairs of bounding boxes. The Euclidean distance between the output of the last fully connected layers (i.e., top and ego) passed through sigmoid activation is set to 0 when the pair belongs to the same person and 1, otherwise.

Given the identity of the camera holder ($l_s$), we compute how likely it is for each person $i$ to be present in $l_s$'s field of view (FOV) at any given time. Following [1], we perform multiple object tracking [35] on the provided top-view bounding boxes (provided by the dataset). Knowing the direction of motion of each trajectory at each moment, we employ the same assumptions used in [1]. We estimate the head direction of each of the top-view camera holders by assuming that people often tend to look straight ahead while walking. Since the intrinsic parameters of the egocentric camera (e.g., focal length and sensor size) are unknown, we consider a lower and upper bound for the angle of the camera holder's FOV ($\theta_1$ and $\theta_2$ in Fig.5) to estimate boundaries on $l_s$'s field of view. As a result, we can determine the probability of each identity being present in the field of view of $l_s$ (i.e., camera holder) at any given time $\zeta$ (Fig. 5, right side). We define the probability of identity $i$ being present in the field of view of the camera holder ($l_s$) at time $\zeta$ as:

$$P_{g_\zeta}(i|l_s) = \begin{cases} 1, & \theta_i < \theta_1 \\ \frac{(\theta_2 - \theta_i)}{(\theta_2 - \theta_1)}, & \theta_1 < \theta_i < \theta_2 \\ 0, & \theta_2 < \theta_i \end{cases} \tag{3}$$

Intuitively, if the bounding box is within the lower bound of the FOV, we assign its presence probability to 1. If its orientation with respect to $l_s$ is outside the upper-bound of the FOV range, we assign its presence probability to 0. For values in between the two bounds (e.g., the person at the bottom-left of Fig. 5), we assign its probability proportional to its orientation with respect to the camera holder. In our experiments, we empirically set $\theta_1$ and $\theta_2$ to $30^\circ$ and $60^\circ$, respectively.

### 3.3 Spatiotemporal Reasoning

The third component of our approach enforces spatiotemporal constraints on the re-identification labels within the egocentric video. We define a cost for assigning the same identity label to a pair of human detection bounding boxes. We later incorporate

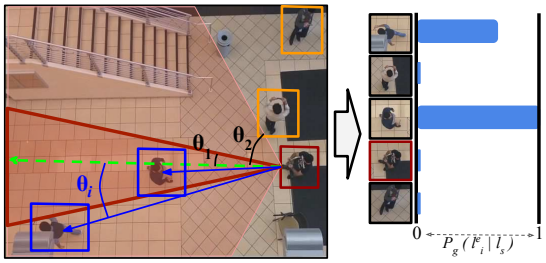



Fig. 5: Geometric reasoning in the top-view video. In this example (left), two identities are present in the field of view of the camera holder (the two red cones showing the lower and upper bound of field of view). Using their orientation (shown by blue arrows) with respect to the camera holder's direction of movement in the top-view (dashed green arrow), we estimate the probability of their presence in the content of the egocentric video. Right bar graph shows the probability of each person being present in the FOV of the camera holder.

this cost in our graph cuts formulation. Two constraints are defined as follows:
**Constraint 1:** Two different bounding boxes present in the same frame cannot belong to the same person. Note that non-maximum suppression is performed in the human detection process. Therefore the binary cost between any pair of co-occurring bounding boxes is set to infinity.
**Constraint 2:** If two bounding boxes have a high overlap in temporally nearby frames, their binary cost should be reduced, as they probably belong to the same identity. We incorporate two constraints in $C_{st}$ cost as follows:

$$C_{st}(d_i^e, d_j^e) = \begin{cases} \infty, \text{ if } \zeta_{d_i^e} = \zeta_{d_j^e} \\ -1, \text{ if } 0 < |\zeta_{d_i^e} - \zeta_{d_j^e}| < \epsilon \text{ and } \frac{A_{d_i^e} \cap A_{d_j^e}}{A_{d_i^e} \cup A_{d_j^e}} > \sigma \\ 0, \ Otherwise \end{cases} \tag{4}$$

where $A_{d_i^e}$ and $A_{d_j^e}$ correspond to the image area covered by human detection bounding boxes $d_i^e$ and $d_j^e$, and $\zeta_{d_i^e}$ and $\zeta_{d_j^e}$ encode the time in which bounding boxes $d_i^e$ and $d_j^e$ are present. If $d_i^e$ and $d_j^e$ have been visible in the same frame, $C_{st}(d_i^e, d_j^e)$ will be set to infinity in order to prevent graph-cuts from assigning them to the same label (constraint 1). The negative cost of $C_{st}(i, j)$ in case of temporal neighborhood ($0 < |\zeta_{d_i} - \zeta_{d_j}| < \epsilon$) and high spatial overlap ($\frac{A_{d_i^e} \cap A_{d_j^e}}{A_{d_i^e} \cup A_{d_j^e}} > \sigma$) will encourage the graph cuts algorithm to assign them to the same label (constraint 2), as they may correspond to the same identity if they have a high overlap. Here, we empirically set $\epsilon$ to 5 frames and $\sigma$ to 0.8.

### 3.4 Fusion

In this section we describe how visual, geometrical, and spatiotemporal reasonings are combined. First, we combine the visual and geometrical reasoning to find a set of candidate $(l_s, \tau)$ pairs. We then examine each candidate pair using graph cuts to measure the cost of its resulting $(L_r, l_s, \tau)$ labeling and select the one with the minimum cost.

**Comparing Visual and Geometrical Priors** In section 3.1, we described how an initial human re-identification prior can be obtained using visual reasoning. In section 3.2, we described how an independent source of information (geometric reasoning) provides yet another set of human re-identification priors given each possible self-identity. In this section, we search over different self-identities and time delays, and choose the one whose patterns of geometric priors is consistent with his/her visual priors.

**Temporal representation:** In section 3.1, we described how we can compute $P_v(l_i^e = j)$ for any given egocentric human detection bounding box $d_i^e$ and top-view identity $j$. We can form a $T^e \times |I^t|$ matrix $R^v$, where $T^e$ is the number of frames in the egocentric video and $|I^t|$ is the number of identities visible in the top-view video. Intuitively, $R^v(\zeta, j)$ captures the probability of visibility of top-view identity $j$ in the field of view of egocentric camera holder at time $\zeta$. Let $D_\zeta^e = \{d_{\zeta_1}^e, d_{\zeta_2}^e, ..., d_{\zeta_{|D_\zeta^e|}}^e\}$ be the set of human detection bounding boxes visible in frame $\zeta$ of the egocentric video. We define $R^v(\zeta, j) = \sum_{i=1}^{|D_\zeta^e|} P_v(l_i^e = j)$. Since the sum of the probabilities might lead to a value higher than 1, we truncate the value at 1. In other words $R^v(\zeta, j) \leftarrow min(1, R(\zeta, j))$. An example $R^v$ matrix is shown in Fig.6 (center panel).

We can form a similar matrix based on the geometric reasoning for each self-identity. As described in section 3.2, given the self identity of the camera holder ($l_s$), we can compute $P_g(l_i^e = j|l_s)$. Similar to $R^v$, we can form $T^t \times |I^t|$ matrix $R^g$ where $T^t$ is the number of frames in the top-view video, and $R^g(\zeta, j)|_{l_s} = P_{g_\zeta}(i|l_s)$, which is computed according to Eqn. 3. Intuitively $R^g(\zeta, j)|_{l_s}$ is the probability of visibility of identity $j$ in the field of view of self-identity $l_s$ at time $\zeta$ of the top-vie video, geometrically (an example shown in Fig. 6-left). Forming $R^v$ and $R^g|_{l_s}$ for different self-identities ($l_s$), we expect them to have similar patterns for the correct $l_s$. For each top-view identity $l_s$, we compute the cross correlation of its $R^g|_{l_s}$ matrix with $R^v$ across the time dimension in order to evaluate their similarities across different time delays ($\tau$). This cross correlation results in a 1D signal encoding the similarity score of the two matrices given different time offsets. As shown in Fig. 6, we estimate the time offset between the two videos (assuming self-identity $l_s$) by finding the maximum of that score. We search across all self identities and sort them based on their maximum cross correlation score.

$$l_s^*, \tau^* = \underset{l_s, \tau}{argmax} \;\; R^v \odot R^g|_{l_s} \tag{5}$$

where $\odot$ denotes element-wise multiplication. Please note that all the videos in our dataset are captured with the same frame rate. Thus, we can perform all of these computations in a frame-based manner. Otherwise, a pre-processing and quantization on the temporal domain would be necessary to correlate the two matrices.

**Graph Cuts:** Given a set of suggested $(l_s, \tau)$ pairs from the previous section, we evaluate the overall labeling cost as the cost of assigning $l_s$ to the self identity, $\tau$ to the time delay, and $L_r$ to the re-identification labels. Graph cuts allows the re-identification labels to adjust to the spatiotemporal constraint.

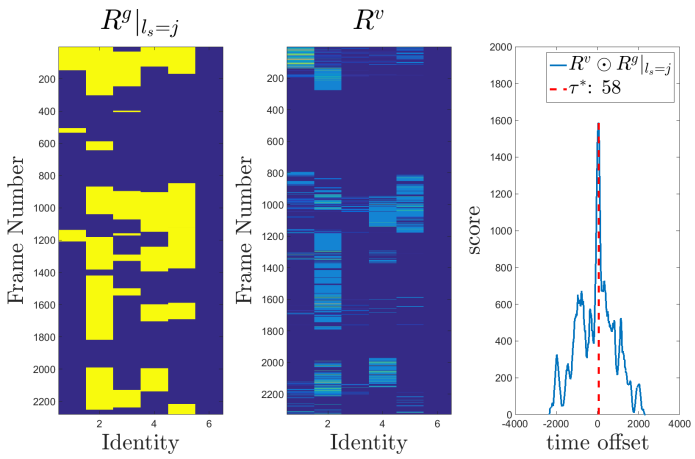



Fig. 6: An example of estimating the self-identity and temporal offset. For a certain self-identity ($l_s$), the geometric reasoning is performed and the suggested re-identification priors are stored in matrix $R^g|_{l_s}$ (values color-coded). The matrix acquired by visual reasoning (in this case the supervised CNN based method) is shown in the middle ($R^v$). The similarity between the patterns in two matrices suggests that the self identity ($l_s$) is a good candidate. By correlating the two matrices across the time domain (the rightmost panel), we can observe a peak at $\tau = 58$. This suggests that if the camera holder has in fact identity $l_s$, the time-offset of his egocentric video with respect to the top-view video is 58 frames. Also, the score of self-identity $l_s$ is the maximum value of the cross correlation which is 1587 in this case. By computing this value for all of the possible self-identities, we can pick the most likely self identity as the one with the highest score.

We form a graph $G(V, E)$ in which nodes are the human detection bounding boxes in ego-view $V = \{d_1^e, d_2^e, d_3^e, ..., d_{|D^e|}^e\}$ (See Fig. 7 for an illustration.). The goal is to assign each node to one of the top-view labels. Edges of the graph encode the spatiotemporal constraints between the nodes (as described in Section 3.3). Given the self identification label and time delay, we can perform graph cuts with its cost defined as:

$$C_{tot}(l_s, \tau) = \sum_{i=1}^{|D^e|} \Big[ C_u(l_i^e|\tau, l_s) + \sum_{j=1, j\neq i}^{|D^e|} C_{st}(l_i^e, l_j^e) \Big] \tag{6}$$

The first term in rhs of Eqn. 6 encodes the unary cost for assigning $d_i^e$ to its label $l_i^e$, given self-identity $l_s$ and relative temporal offset ($\tau$) between the two videos. We set $C_u$ as:

$$C_u(l_i^e = j|\tau, l_s) = 1 - P^v(l_i^e = j) R^g(\zeta_i^e - \tau, j)|l_s \tag{7}$$

where $\zeta_i^e$ is the time in which human detection bounding box $d_i^e$ appears in the ego-view. Intuitively Eqn. 7 means that the probability of bounding box $d_i^e$ (appearing at time $\zeta_i^e$ in the ego-view) being identity $j$ in top-view, is the probability of the visibility of identity $j$ at the field of view of $l_s$ at time $\zeta_i^e - \tau$ in the top-view, multiplied by

its likelihood of being identity $j$ visually. The binary terms determine the costs of the edges and encode the spatiotemporal cost described in section 3.3. The output of this method provides us with a cost for each $(l_s, \tau)$ pair, alongside with a set of labellings for the human detection bounding boxes $L_r$. The pair with the minimum cost and its corresponding $L_r$ is the final solution of our method (i.e., $l_s^*, L_r^*, \tau^*$).

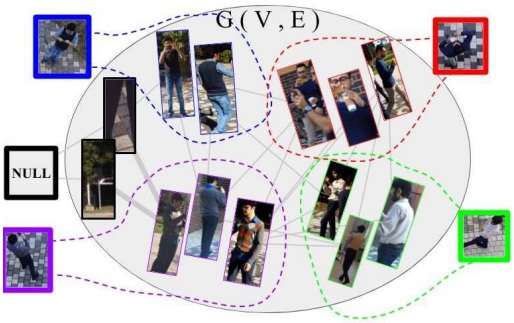



Fig. 7: An illustration of the graph formation. The silver oval contains the graph G(V,E) in which each node is one of the ego-view human detection bounding boxes. The squared bounding boxes highlight different top-view labels in different colors. The graph cuts are visualized using the dashed colored curves. We always consider an extra NULL class for all of the human detection bounding boxes that do not match any of the classes.

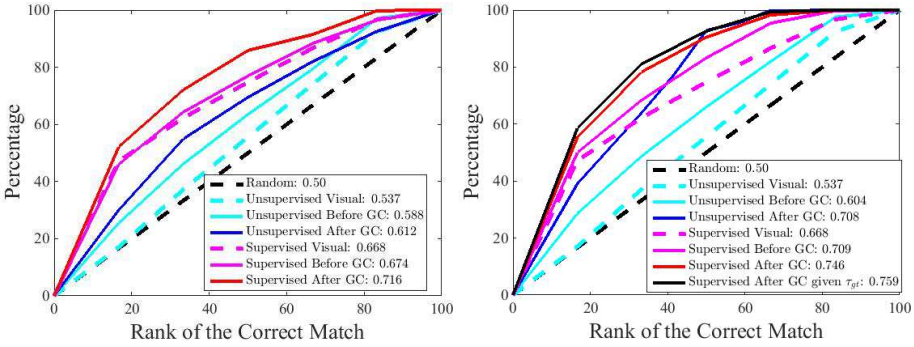
## 4 Experimental Results

### 4.1 Dataset

We use the publicly available dataset [1]. It contains sets of videos shot in different indoor and outdoor environments. Each set contains one top-view and several egocentric videos captured by the people visible in top-view. Each ego-top pair is used as an input to our method. We used three sets for training our two stream neural network and the rest for testing. There are 47 ego-top test pairs and therefore 47 cases of self-identification and temporal alignment. The total number of human detection bounding boxes, and therefore human re-identification instances is 28,250. We annotated the labels for all the 28,250 human detection bounding boxes and evaluated the accuracy for re-identification and self-identification. The number of people visible in top-view videos vary from 3 to 6, and lengths of the videos vary from 1,019 frames (33.9 seconds) up to 3,132 frames (104.4 seconds).

### 4.2 Evaluation

We evaluate our proposed method in terms of addressing each objective and compare its performance in different settings. Moreover, we analyze the contribution of each component of our approach in the final results.




(a) (b)

Fig. 8: (a) shows the re-identification performance of different components of our method. (b) shows the same evaluation given the ground truth self identification labels.

**4.2.1. Self-identification:** We evaluate our proposed method in terms of identifying the camera holder in the content of the top-view video. Since we perform self-identification based on initial re-identification probabilities (visual reasoning), we evaluate self-identification based on supervised and unsupervised re-identification results, alongside with state-of-the-art baselines. We also evaluate the performance in each setting before and after the final graph cuts step to assess the contribution of the spatiotemporal reasoning. Upper-bounds of the proposed method are also evaluated by providing the ground-truth re-identification and temporal alignment. The cumulative matching curves are shown in Fig.**??** left. The solid yellow curve is the performance of [1]. As explained before, [1] highly relies on the relationship among multiple egocentric videos and does not perform well when it is provided with only one egocentric video. The dashed yellow curve shows the performance of [8]. The network provided by the authors was used. As explained in the related work section, this framework is not designed for scenarios such as ours. The cyan and blue curves show our self-identification accuracy in the unsupervised setting before and after the graph cuts step, respectively. The magenta and red curves show the performance in supervised setting, before and after the graph cuts step, respectively. The dashed black curve shows random ranking (performance of chance). The advantage of graph cuts and the spatiotemporal constraints can be observed by comparing before and after graph cuts curves. The contribution of our two stream visual reasoning is evident by comparing the unsupervised curves with their corresponding supervised settings. The effect of the geometrical reasoning could be seen by comparing visual reasoning results, and the before GC curves. The numbers in the figure legend show the area under each curve for quantitative comparison. The margin between the supervised and unsupervised approaches shows the effect of re-identification quality on self-identification performance, confirming the interconnectedness of the two tasks. The solid green and solid black curves show the upper-bounds of the proposed method. We evaluate self-identification, when providing ground-truth re-identification labels and the time-delay to the proposed approach.

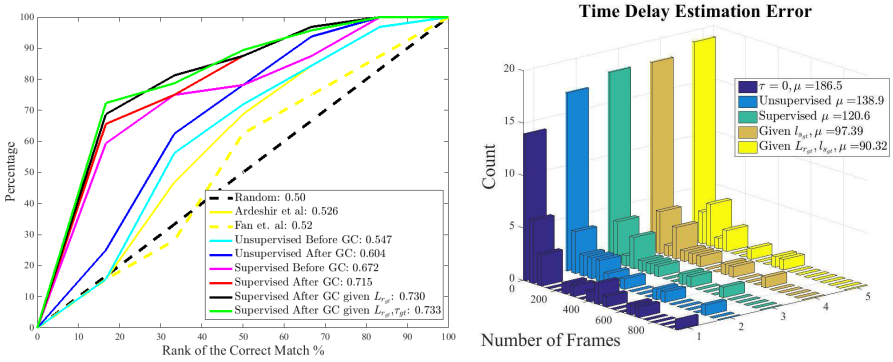



Fig. 9: Left shows the cumulative matching curves illustrating the performance in the self-identification task. Right shows the distribution of time-delay estimation errors using our supervised and unsupervised methods, compared to the baselines and upper-bounds.

**4.2.2. Cross-view human re-identification:** We compute the human re-identification performance in supervised and unsupervised settings, before and after graph cuts (shown in Fig. 8a). In order to better assess the performance, we compute the performance of our proposed method given the ground truth self identification label ($l_{s_{gt}}$), and ground truth time delay $\tau_{gt}$ (Fig. 8b), which results in upper-bounds for re-identification performance. In both figures (a and b), the dashed black line shows the chance level performance. The dashed cyan and magenta curves show the performance of direct visual matching across the two views using our unsupervised and supervised visual reasonings, respectively. Solid cyan and magenta curves show the performance of our unsupervised and supervised visual cues combined with geometric reasoning. Which is re-identification solely based on unary confidences in Eqn. 7 and before applying graph cuts. Finally, blue and red curves show performance of the unsupervised and supervised methods (in order) after the graph cuts step, which enforces the spatio-temporal constraints. Black solid curve in Fig 8b shows the performance of the proposed method, given the ground truth time delay between the two videos in addition to the ground truth self-identity. Comparing the red curves of Fig. 8a and 8b shows the effect of knowing the correct self identity on re-identification performance and thus confirming the inter dependency of the two tasks. Comparing the red and black solid curves in Fig. 8b shows that once the self-identity is known, correct time-delay does not lead to a high boost in re-identification performance which is consistent with our results on self-identification and time delay estimation. Comparing Fig. 8 a and b shows that knowing the correct self identity improves re-identification. As explained before, any re-identification method capable of producing a visual similarity measure could be plugged into our visual reasoning component. We evaluate the performance of two state of the art re-identification methods in Table 1. Before Fusion is the performance of each method in terms of Area under curve of cumulative matching curve (similar to Fig. 8a). After fusion is the over-

all performance after combining the re-identification method with our geometrical and spatiotemporal reasoning.

| Method | Before Fusion | After Fusion |
|---|---|---|
| Ours (Unsupervised) | 0.537 | 0.612 |
| Ahmed [21] | 0.563 | 0.621 |
| Cheng[22] | 0.581 | 0.634 |
| Ours (supervised) | **0.668** | **0.716** |

Table 1: Performance of different re-identification methods. Before Fusion is the performance of the re-identification method directly applied to the bounding boxes (only visual reasoning). After fusion shows the performance of our method if we replace our two stream network with the methods mentioned above.

**4.2.3. Time-delay estimation:** Defining $\tau_{gt}$ as the ground truth time offset between the egocentric and top-view videos, we compute the time-offset estimation error ($|\tau^* - \tau_{gt}|$) and compare its distribution with that of baselines and upper bounds. Fig. **??** shows the distribution of time-offset estimation error. In order to measure the effectiveness of our time-delay estimation process, we measure the absolute value of the original time-offset. In other words, assuming $\tau^* = 0$ as a baseline, we compute the offset estimation error (shown in the dark blue histogram). The mean error is also added to the figure legend for quantitative comparisons. Please note that the time delay error is measured in terms of the number of frames (all the videos have been recorded at 30fps). The baseline $\tau = 0$ leads to 186.5 frames error (6.21s). Our estimated $\tau^*$ in the unsupervised setting, reduces this figure to 138.9 frames (4.63s). Adding visual supervision reduces this number to an average of 120.6 frames (4.02s). To have upper bounds and evaluate the performance of this task alone, we isolate it from the other two by providing the ground-truth self identification ($l_{s_{gt}}$) and human re-identification labels ($L_{r_{gt}}$). Providing $l_{s_{gt}}$ will lead to 97.39 frames error (3.24), and providing both $l_{s_{gt}}$ and $L_{r_{gt}}$ reduces the mean error to 90.32 (3.01s). Similar to our re-identification upper-bounds, knowing the self-identity improves performance significantly. Once self-identity is known, the ground truth re-identification labels will improve the results by a small margin.

## 5 Conclusion

We explored three interconnected problems in relating egocentric and top-view videos namely human re-identification, camera holder self-identification, and temporal alignment. We perform visual reasoning across the two domains, geometric reasoning in top-view domain and spatiotemporal reasoning in egocentric domain. Our experiments show that solving these problems jointly improves the performance in each individual task, as the knowledge about each task can assist solving the other two.

## References


1. Ardeshir, S., Borji, A.: Ego2top: Matching viewers in egocentric and top-view videos. In: European Conference on Computer Vision, Springer (2016) 253–268
2. Cheng DS, Cristani M, S.M.B.L.M.V.: Head motion signatures from egocentric videos. InComputer Vision–ACCV. Springer International Publishing. (2014)
3. Yonetani, Ryo, K.M.K., Sato., Y.: Ego-surfing first person videos. Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, (2015)
4. Bettadapura, Vinay, I.E., Pantofaru., C.: Egocentric field-of-view localization using first-person point-of-view devices. Applications of Computer Vision (WACV), IEEE Winter Conference on. (2015)
5. Kiefer, Peter, I.G., Raubal., M.: Where am i? investigating map matching during selflocalization with mobile eye tracking in an urban environment. Transactions in GIS 18.5 (2014)
6. Ardeshir, S., Malcolm Collins-Sibley, K., Shah, M.: Geo-semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2792–2799
7. Shervin Ardeshir, Amir Roshan Zamir, A.T., Shah., M.: Gis-assisted object detection and geospatial localization. In European Conference on Computer VisionECCV (2014) 602–617
8. Fan, C., Lee, J., Xu, M., Kumar Singh, K., Jae Lee, Y., Crandall, D.J., Ryoo, M.S.: Identifying first-person camera wearers in third-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 5125–5133
9. Ardeshir, S., Borji, A.: Egocentric meets top-view. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
10. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
11. Cheng DS, Cristani M, S.M.B.L.M.V.: Custom pictorial structures for re-identification. BMVC (2011)
12. Bak S, Corvee E, B.F.T.M.: Multiple-shot human re-identification by mean riemannian covariance grid. InAdvanced Video and Signal-Based Surveillance (AVSS), 8th IEEE International Conference on (2011)
13. Bazzani L, Cristani M, M.V.: Symmetry-driven accumulation of local features for human characterization and re-identification. omputer Vision and Image Understanding. (2013)
14. Zhao, R., Oyang, W., Wang, X.: Person re-identification by saliency learning. IEEE transactions on pattern analysis and machine intelligence **39**(2) (2017) 356–370
15. Martinel, N., Foresti, G.L., Micheloni, C.: Person reidentification in a distributed camera network framework. IEEE transactions on cybernetics (2016)
16. García, J., Martinel, N., Gardel, A., Bravo, I., Foresti, G.L., Micheloni, C.: Discriminant context information analysis for post-ranking person re-identification. IEEE Transactions on Image Processing **26**(4) (2017) 1650–1665
17. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (June 2014) 152–159
18. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. CoRR **abs/1607.08378** (2016)
19. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. CoRR **abs/1607.08381** (2016)
20. Yi, D., Lei, Z., Li, S.Z.: Deep metric learning for practical person re-identification. CoRR **abs/1407.4979** (2014)

21. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
22. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
23. Cho, Y.J., Yoon, K.J.: Improving person re-identification via pose-aware multi-shot matching. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
24. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
25. Chakraborty, A., Mandal, B., Galoogahi, H.K.: Person re-identification using multiple first-person-views on wearable devices. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2016) 1–8
26. Zheng, K., Guo, H., Fan, X., Yu, H., Wang, S.: Identifying same persons from temporally synchronized videos taken by multiple wearable cameras
27. Alahi, Alexandre, M.B., Kunt., M.: Object detection and matching with mobile cameras collaborating with fixed cameras. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 (2008)
28. Alahi A, Marimon D, B.M.K.M.: A master-slave approach for object detection and matching with fixed and mobile cameras. InImage Processing, 2008. ICIP 2008. 15th IEEE International Conference (2008)
29. Ferland F, Pomerleau F, L.D.C.M.F.: Egocentric and exocentric teleoperation interface using real-time, 3d video projection. InHuman-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on (2009)
30. Park, Hyun, E.J., Sheikh., Y.: Predicting primary gaze behavior using social saliency fields. Proceedings of the IEEE International Conference on Computer Vision. (2013)
31. Soran, B., Farhadi, A., Shapiro, L.: Action recognition in the presence of one egocentric and multiple static cameras. In: Asian Conference on Computer Vision, Springer (2014) 178–193
32. Ardeshir, S., Borji, A.: An exocentric look at egocentric actions and vice versa. Computer Vision and Image Understanding (2018)
33. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 670–677
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
35. Dicle, Caglayan, O.C., Sznaier., M.: The way they move: Tracking multiple targets with similar appearance. Proceedings of the IEEE International Conference on Computer Vision (2013)