EgoReID: Person re-identification in Egocentric Videos Acquired by Mobile Devices with First-Person Point-of-View

Emrah Basaran\textsuperscript{1}\textsuperscript{*} \hspace{1cm} Yonatan Tariku Tesfaye\textsuperscript{2}\textsuperscript{*} \hspace{1cm} Mubarak Shah\textsuperscript{2}
basaranemrah@itu.edu.tr \hspace{0.5cm} yonantariku@knights.ucf.edu \hspace{0.5cm} shah@crcv.ucf.edu

\textsuperscript{1}Dept. of Computer Engineering, Istanbul Technical University
\textsuperscript{2}Center for Research in Computer Vision, University of Central Florida

Abstract

Widespread use of wearable cameras and recording devices such as cellphones have opened the door to a lot of interesting research in first-person Point-of-view (POV) videos (egocentric videos). In recent years, we have seen the performance of video-based person Re-identification (ReID) methods improve considerably. However, with the influx of varying video domains, such as egocentric videos, it has become apparent that there are still many open challenges to be faced. These challenges are a result of factors such as poor video quality due to ego-motion, blurriness, severe changes in lighting conditions and perspective distortions. To facilitate the research towards conquering these challenges, this paper contributes a new, first-of-its-kind dataset called EgoReID. The dataset is captured using 3 mobile cellphones with non-overlapping field-of-view. It contains 900 IDs and around 10,200 tracks with a total of 176,000 detections. Moreover, for each video we also provide 12-sensor meta data.

Directly applying current approaches to our dataset results in poor performance. Considering the unique nature of our dataset, we propose a new framework which takes advantage of both visual and sensor meta data to successfully perform Person ReID. In this paper, we propose to adopt human body region parsing to extract local features from different body regions and then employ 3D convolution to better encode temporal information of each sequence of body parts. In addition, we also employ sensor meta data to determine target’s next camera and their estimated time of arrival, such that the search is only performed among tracks present in the predicted next camera around the estimated time. This considerably improves our ReID performance as it significantly reduces our search space.

1. Introduction

Person Re-Identification (ReID) aims at associating the same pedestrian across multiple cameras\cite{11, 54}. This task has drawn increasing attention in recent years due to its importance in applications, such as surveillance\cite{43}, activity analysis\cite{26} and tracking\cite{48}. Person ReID remains a challenging problem because of complex variations in camera viewpoints, human poses, lighting, occlusions, and background clutter.

Currently, in almost all attempts to solve person ReID problem, the source of data is from fixed cameras with wider field of view (FOV). Mostly, in such datasets, full bodies of pedestrians are captured and intra camera variations of targets are very small, as pose of targets often do not change within the same camera. On the contrary, videos from a moving first-person point-of-view (POV) devices pose unique challenges, such as frequent occlusion of targets due to close proximity of camera and targets, blurriness caused by camera motion, severe lighting and background changes, both within and across...
Figure 1: (a) shows three tracklets of the same identity from the same camera, it's evident to see that due to both camera and target motions a target is captured in different poses, backgrounds and illuminations. (b) each row shows tracklets of the same identity captured from camera 1, 2 and 3, respectively. Due to close proximity of target and the cameras, different body regions of the target are missing in Camera 1 and 2.

cameras, appearance difference due to different perspectives. Moreover, due to constant motion of both target and cameras, there is a frequent inter camera appearance changes between sequences of the same target. This is mainly due to the fact that a target could be recorded in several poses in a single camera as shown in Fig. 1 (a). The scene is more dynamic as compared to a traditional fixed camera scenario.

Besides the unique challenges posed by the nature of the video, these devices (such as cellphones) provide additional source of information. Most new generation devices are equipped with additional sensors such as gyros, accelerometers, magnetometers, GPS and are Internet-enabled, it is now possible to obtain large amounts of first-person point-of-view (POV) data un-intrusively.

Aiming to facilitate the research towards such unique video domain, we propose a new dataset called EgoReID. Different from existing datasets EgoReID presents several new features. 1) 3 synchronized mobile cellphones with non-overlapping FOV are used to record the videos. 2) Throughout the recording, all 3 cameras are moving around campus covering larger area, resulting in complex scene transformation and background. 3) 12 sensor meta data along with the videos are also collected. 4) YOLO9000 [30] and FCDSC [39] are used for pedestrian detection and tracklet generation, respectively, which will be provided with the dataset.

In a typical person ReID pipeline, first, each pedestrian video sequence is represented by a single feature vector and then is matched in a task-specific metric space, where the feature vectors of same target are expected to have smaller distance than that of different target [46, 23, 24]. These approaches show impressive results, when target sequences do not undergo drastic view changes or occlusion. However, in egocentric videos pedestrians undergo drastic change in both appearance and pose due to frequent motion of both camera and targets. Furthermore, frequent occlusion of different body parts of targets is common due to close proximity of a target to the camera as shown in Fig. 1 (b).

To address aforementioned challenges posed by the first-person POV videos, we propose a framework which uses human semantic parsing to extract local features from human body parts. It is natural to make use of semantic body regions in this particular setup, since mostly pedestrians are in close proximity from cameras, thus, different body regions are more visible. However, most of Person Re-ID datasets deal with image-based re-ID problem and do not deal with videos, therefore very few video-based Person re-ID datasets are available for training video deep networks for re-ID. In this paper, we propose a novel approach to leverage network trained on 10 image-based re-ID datasets for video-based Person Re-ID problem. Here, we employ a model, trained on these image-based datasets, to extract frame level local features for each semantic region, then we employ 3D convolutions to encode the temporal information in each sequence of regions. Instead of directly encoding the whole sequence, we use sequence of each regions to learn discriminative local features and compare respective regions separately to find a match. In addition to traditional video features, we successfully demonstrate the use of additional sensor meta data to complement appearance cues. In particular, we use meta data information to estimate target’s next camera and its time of arrival, that way we significantly reduce the search space by pruning several unreliable pairs which violate the estimated
time of arrival constraint.

The contributions of this paper can be summarized as follows: 1) A unique and challenging EgoReID dataset is collected and will be released with detections, tracks and sensor meta data. Compared to existing datasets, EgoReID has a unique feature which can help advance research in solving realistic and challenging task of person re-identification in a first-person POV videos. 2) We propose a new model to solve person ReID problem in egocentric videos, where human semantic parsing is employed to extract local visual cues and compare corresponding semantic regions to find a match. 3) We propose a method to employ sensor meta data information to significantly reduce our search space by first predicting the next camera of a target then estimating its time of arrival, thus, we only search for our match among tracks which satisfy the time constraint.

2. Related Works

In this section, we review related works in the areas of video-based person re-identification and egocentric vision.

**Video-based person re-identification:** The success of deep learning in wide range of computer vision areas has been inspiring a lot of studies in person re-identification. In recent years, researchers are considering more realistic scenarios such as larger dataset [53, 52], complex scenarios [55, 31] and combining different modalities such as text descriptions in their approaches [56, 19].

The success of deep learning in image classification [16], has inspired several studies in person re-identification. The effectiveness of Convolutional Neural Network (CNN) in learning discriminative image representations from large scale datasets have been exploited effectively in [40, 41, 25, 17, 6, 51, 58, 49, 38]. Utilization of video data is further facilitated by the powerful feature learning ability of Convolutional Neural Networks. In video-based ReID [29, 58, 46, 42, 59, 28], the learning algorithm is given pair of video sequences instead of images. Authors in [29], introduce an RNN model to encode temporal information. The features from all timesteps are then combined using temporal pooling to give an overall appearance feature for the complete sequence and then compute the feature similarity of two videos. In [42] the most discriminative video fragments are selected from noisy/incomplete image sequences of people from which reliable space-time and appearance features can be computed, whilst simultaneously learning a video ranking function for person ReID. Authors in [28], introduce a new space-time person representation by encoding multiple granularities of spatio-temporal dynamics in form of time series. In [58] attention weights are combined per-frame with visual features and the forward propagated RNN hidden variables, in order to weaken the influence of the noisy samples. Authors in [45] employ spatial pooling layer to select regions from each frame, while temporal attention pooling is performed to select informative frames over the sequence, both pooling guided by the information from their distance matching. In [36], the dual attention mechanism is used, in which both intra-sequence and inter-sequence attention strategies are employed for feature refinement and feature-pair alignment, respectively.

**Egocentric vision:** Recently, visual analysis of egocentric videos has been a hot topic in computer vision, and the work ranging from object detection [9] to recognizing daily activities [8, 7], predicting gaze behavior [23, 3], and video summarization [27] have been proposed.

3. EgoReID Dataset

In this section, we briefly introduce our new EgoReID dataset.

3.1. Dataset Description

During collection, we use 3 synchronized Samsung mobile cellphones with non-overlapping field of views. Throughout the period of recording, all 3 cameras were constantly moving around and covered larger area. To generate tracklets, we first use YOLO9000 [30] to detect pedestrians and then employ FCDSC tracker [39] to generate tracklets. Sample tracklets are shown in Fig. 1 and more are given in the supplementary material. EgoReID consists of around 900 different identities with 10,200 tracklets with a total of 176,000 human detection bounding boxes. Camera 1 and 2 have 190 IDs in common while camera 2 and 3 share 256 IDs in common. Around 103 IDs are present in all three cameras. 3164, 2748 and 4353 tracklets are present in
camera 1, 2 and 3, respectively.

Moreover, we have collected 12 sensor metadata, namely, Longitude, Latitude, Speed, Distance, Time, Accelerometer, Heading, Gyro, Magnetic, Gravity, Orientation and Rotation vectors (each is 3-D vector). Please refer to Fig. 2.

**Evaluation protocol:** Similar to most of previous datasets, we utilize the Cumulative Matching Characteristics (CMC) curve to evaluate the ReID performance. For each query, multiple true positives could be returned. Therefore, we also consider person ReID as a retrieval task, and also employ mean Average Precision (mAP) as the evaluation metric.

### 4. Proposed Method

This section briefly presents the proposed model which consists of 3 main components: frame-level feature extraction, semantic segmentation and video-level feature extraction. The overview of the proposed network is shown in Figure 3.

Our model is inspired by the model [14], where human parsing is successfully applied in image-based ReID. Important feature of their model is that it is trained on ten different dataset providing robust system. In the unique nature of our dataset, since targets are close to the cameras, different body regions of a target are clearly visible, therefore above method is pretty suitable for our problem. However, their method is image-based, in this paper, we extend person ReID as a retrieval task, and also employ mean Average Precision (mAP) as the evaluation metric.

#### Feature sequence extraction module:
In our model, we adopt Inception-V3 [37] as a backbone of our feature extraction module. Given a video sequence $X \in \mathbb{R}^{H \times W \times 3 \times S}$ of length $S$, where each frame (RGB) is of size $H \times W$, each frame in video at a time-step is passed to Inception-V3 to produce the frame-level feature maps. At the top of the Inception-V3, using bilinear interpolation, the feature maps are scaled up to the size of the segmentation maps.

#### Semantic segmentation module:
For the segmentation module, as in the feature extraction module, we use the Inception-V3 architecture. However, we make two important modification to this architecture compared to the one used for feature extraction. The first modification is that the output stride of the model is reduced from 32 to 16 to get the feature maps with adequate resolution for semantic segmentation task. The extra computation cost resulting from this modification is eliminated by the replacing the convolutions in the last Inception block with the dilated convolutions [47]. The second modification is the use of atrous spatial pyramid pooling [5], instead of global average pooling to exploit multi-scale information. Refer to [14] for more details about the feature extraction and semantic segmentation modules.

#### Video-based person ReID model:
Given a video sequence, each frame in a video at a time-step is passed to an Inception-V3 to produce the frame-level feature maps. Simultaneously, our semantic segmentation module is fed the same sequence, which generates segmentation (probability) maps of 3 semantic regions for each frame in the sequence. We pool the output of our feature extractor model using the 3 segmentation maps. This results in a set of 3 maps, each focusing on a different body region of the sequence. Next, to encode the temporal information, we pass the activation sequence of each region to 3 consecutive 3D convolutions that do not share the weights. As a result of this encoding process, a tensor is obtained for each body region and feature vectors are constructed using the Global Average Pooling.
Figure 2: In (a) Camera orientation-pitch and FOV of camera (in terms of images) at each step are shown. (b) Camera rotations around x axis and FOV (in terms of images) at each step are show. (c) Top row shows sample detection results and corresponding tracking results are shown at the bottom row (each track is shown by a different color). (d) shows sample GPS locations of three cameras as red dots.

Figure 3: Proposed Network: First, the input video clip consisting of S frames are input to the Inception-V3 models, the feature maps and 3 segmentation maps for each frame (f-1 to f-S) are generated in the upper and lower branches of the model, respectively. We pool the feature maps using the segmentation maps and obtain sequences of 3 feature maps, each focusing on a different body regions: lower, upper and full. Next, we aggregate the temporal information within these sequences utilizing 3D convolutions, which operate with the stride of 2 and with the kernel size of 3 for the temporal 1 for the spatial dimensions. With this encoding process, we obtain a tensor of size (30x30x2048) for each body region and apply global average pooling (GAP) to construct the feature vectors. In addition to these vectors, we construct another feature vector taking the average of the output feature maps of Inception-V3 (top branch) and applying GAP on the resulting tensor.
ing (GAP) on these tensors. We also use one more feature vector, constructed by taking the average of the feature maps produced in the feature extractor module and applying GAP on it. During the training of the model, we pass the 3 feature vectors of the body regions to different soft-max layers that do not share the weights and compute the classification losses separately. We do not use the global features while training the model. At the test time, we concatenate all region and global features to get the final descriptors for the input videos.

5. Improving ReID Employing Sensor Meta-data

In this section, we discuss the proposed method for employing sensor meta data information to further help refine our re-identification results. In particular, we use the heading, speed and GPS (longitude and latitude) of the camera, which are captured by our recording devices.

Let $T_i^a$ represents an observation of a person $i$ in camera $a$. Its trajectory is given by a set of points $T_i^a = [d_{i,1}^a, ..., d_{i,t^r}^a]$, where $t^s$ and $t^r$ represent the time it entered and exited camera $a$, respectively. $\rho(T_i^s)$ denote appearance feature representation of tracklet $T_i^a$ generated by the proposed model. GPS and heading of camera $a$ at time $t^r$ are respectively represented by $C_{i,t^r}^{\text{gps}}$ and $C_{i,t^r}^{\text{h}}$. The heading of a target $i$ in camera $a$, $T_i^a$, is inferred from the camera $a$’s heading at time $t^r$, that is, if target is moving in the same direction as camera (target is walking away from the camera), then $T_i^a = C_{i,t^r}^{\text{h}}$. Where as, if a target is moving towards the camera, like the target in Fig. 4, then $T_i^a = C_{i,t^r}^{\text{h}} + 180^\circ$, that is, we assign the opposite heading w.r.t the camera’s heading. The direction of target’s movement is determined from the tracklet direction.

**Estimating the next camera:** Naively, one can select the closest camera to the target at time $t^r$, as the next camera without considering the target’s direction of motion. As can be seen from the example in Figure 4, such approach will end up selecting the wrong camera, which is located on the other side of the target. So, first, we need to select candidate cameras which are spatially located in the direction of target’s heading. Then, we select the closest camera among those based on their GPS distance from the target at time $t^r$. In this set up, we not only select the closest camera but also we ensure the next camera is located along the way of target’s motion. Let define a function $H(GPS_1, GPS_2)$, which returns heading angle between two GPS points. In particular, it determines which direction GPS2 is located w.r.t GPS1 and is computed as follows:

$$H(GPS_1, GPS_2) = \text{atan2}(X, Y),$$

where, $X = \sin(\lambda_2 - \lambda_1) * \cos(\varphi_2)$ and $Y = \cos(\varphi_1) * \sin(\varphi_2) - \sin(\varphi_1) * \cos(\varphi_2) * \cos(\lambda_2 - \lambda_1)$. $(\lambda_1, \varphi_1)$ and $(\lambda_2, \varphi_2)$ represent tuple of longitude and latitude of $GPS_1$ and $GPS_2$, respectively.

For a camera, $C^m$, to be in a candidate set, $\mathcal{K}$, of target $T_i^a$, the following should hold: $T_i^b \equiv H(C_{i,t^r}^{\text{h}}, C_{i,t^r}^{\text{h}})$. That is, for a camera $C^m$ to be selected as a candidate, its heading w.r.t the location of current camera of the target, $C_{i,t^r}^{\text{h}}$, is the same as
target’s heading $T_i^{a,h}$. Finally, the next camera for target $i$ in camera $a$ is estimated by selecting the closest camera at time $t^\ast$ among the candidate cameras, and is given by: $\arg \min_{C_j} ||T_i^{a_s} - C_j^s||$, $\forall C_j \in K$, where, $T_i^{a_s}$ is GPS of target $i$ of camera $a$ at time $t^\ast$ and due to very close proximity of the cameras to the targets, we can approximate target’s GPS by the camera $d$’s GPS at time $t^\ast$, $C_i^{d_s}$.

**Estimating time of arrival:** Next, we estimate the time required to travel between camera $a$ (current camera) and $j$ (the next camera). Since our cameras topology is very dynamic (due to camera motion), we implicitly compute the time as a division of distance and speed. The speed of a target, $T_i^{a_s}$, is estimated to be 1.3 m/s, which is the average walking speed of individual [32]. While the speed of camera $j$, $C_j^s$, can be inferred from the sensor metadata data. The distance between the next camera, $C_j$, and the target is computed between their GPS. Formally, the time required by target, $T_i^{a}$, to reach camera $j$ from camera $a$ can be computed as follows:

$$E_{T_i^{a}, C_j} = \begin{cases} \frac{||T_i^{a_s} - C_j^s||}{|T_i^{a_s} - C_j^s|}, & \text{if } T_i^{a_s} \equiv C_j^s, \\ \frac{||T_i^{a_s} - C_j^s||}{|T_i^{a_s} + C_j^s|}, & \text{Otherwise} \end{cases}$$

where $T_i^{a_s}$ is the heading of target $i$ at the time of exit, $t^\ast$, from camera $a$ and is inferred from the heading of camera $a$ at time $t^\ast$.

Note that in eq. 1, since both target and camera are moving, time computation depends on the direction of their motion. If both are moving in the same direction (i.e. same heading), first case in eq. 1, then we divide the distance by the difference of their speeds, otherwise, we estimate the time by dividing the distance with the sum of their speed.

Finally, we impose our time constraint on our appearance similarity. The final affinity between query target, $T_i^{a}$, and the rest of tracks in camera $j$ is updated as follows:

$$M(T_i^{a_s}, T_j^l) = \begin{cases} \psi(\rho(T_i^{a_s}), \rho(T_j^l)), & \text{if } T_{l,t^\ast} \geq E_{T_i^{a_s}, C_j} \\ 0, & \text{Otherwise} \end{cases}$$

where $\psi(\rho(T_i^{a_s}), \rho(T_j^l))$ is appearance similarity between two tracks features, $T_{l,t^\ast}$ is an entrance time of track $T_i^{a_s}$ and $|T_j|$ is total number of tracks in camera $j$. As can be inferred from eq. 2, during ReID, we only compare track $T_i^{a_s}$ with tracks in camera $j$ which appear after the estimated time of arrival of track $T_i^{a_s}$ at camera $j$. $E_{T_i^{a_s}, C_j}$. Thus, we can significantly reduce our search space by pruning several wrong matching which violate our time constraint.

6. Experiments

6.1. Datasets

In addition to the EgoReID, we present evaluation on widely used, fixed cameras dataset, MARS [52]. It consists of 6 cameras and 1261 different pedestrians. There are 625 identities for training and 636 identities for testing.

**Evaluation settings:** Training/Testing split of EgoReID dataset contains 567 identities for training and 309 identities for testing. For MARS dataset, we follow the same setting as the authors of the dataset. To evaluate performance for each algorithm, we report the Cumulative Matching Characteristic (CMC) metric and mean average precision (mAP).

**Training the Network:** We first train the semantic segmentation module of our proposed model using Look Into Person (LIP) [10] dataset. We then train our feature extractor module using the probability maps produced with segmentation module. Since only 3 semantic regions are used in the model, we group the probability maps of different regions to create the probability maps for foreground, upper-body and lower-body. During the training of the feature extractor module, following the same settings mentioned in [14], we use a training set consists of 10 image-based re-identification datasets. We use a training set, which is constructed by the aggregation of 10 image-based re-identification datasets. Next, we fine-tune the feature extractor by image-based training using the images from the video datasets (MARS or EgoReID). In the last step of the training, we freeze feature extractor and train 3D convolution layers using video clips of 15 frames. As mentioned in Section 4, there are three consecutive 3D convolution layers for each region. Each of these layers produce a tensor of size $(30 \times 30 \times 2048)$, by encoding the temporal information with the kernel size of
and with the stride of 1 and 2 in spatial and temporal dimension, respectively. We construct the feature vectors from these tensors applying GAP and compute the loss for each region by performing multi-class classification separately. Finally, the loss of the model is obtained by summing the three losses. More implementation details are provided in the supplementary material.

**Ablation study of our approach:** We investigate the effect of each component of our model by conducting several experiments. In Table. 1, we show the results of each component in the proposed network. We evaluate the effects of each body regions and sensor meta data. As can be noted from Table. 1, lower body features perform worst on EgoReID, while they achieve reasonable result on MARS. This is mainly due to the fact that in EgoReID dataset lower body is frequently missing as the camera is very close to the target. We can also observe from Table. 1 that, jointly using features from different body regions, leads to improvements in the performance than using only one body region.

We can also see that the proposed approach for using sensor information to prune several unreliable matching pairs, significantly improves the ReID performance of our approach on EgoReID dataset. As shown in Table. 1, we are able to improve rank-1 and mAP by around 15% and 12%, respectively.

To further show the effectiveness of the proposed model, we compare our results with [14], where we apply their method on frame by frame bases and then employ average pooling over temporal dimension to generate features from each region. As can be seen from Table. 2, our approach gives significantly better results in both rank-1 and mAP (i.e. 20% and 34% respectively). This shows the effectiveness of the proposed 3Dconv module in encoding the temporal information.

| Methods                          | R-1 | R-5 | mAP |
|----------------------------------|-----|-----|-----|
| SPReID [14] (Avg pooling)        | 58.23 | 72.07 | 31.21 |
| Ours                             | **79.19** | **89.70** | **65.91** |

Table 2: Comparison of SPReID [14] and the proposed approach on MARS dataset.

**Comparison to state-of-the-art Methods:** In Table. 3 and Table. 4, we compare our approach against the state-of-the-art methods on EgoReID and MARS, respectively. As can be observed from Table. 3, our approach significantly outperforms state-of-the-art approach on EgoReID dataset. This shows the effectiveness of our human semantic region based local feature extraction approach on our dataset. This is mainly due to different body parts of pedestrians being clearly visible in EgoReID. In Table. 4, we observe that our method has competitive r-1 results, while it has better results in r-5 and mAP. This further demonstrate the robustness of our approach in handling different domains of video.

| Methods                          | R-1 | R-5 | mAP |
|----------------------------------|-----|-----|-----|
| PSE [35]                         | 15.17 | 25.79 | 8.58 |
| Ours                             | **51.31** | **62.21** | **41.72** |

Table 3: Comparison to state-of-the-art on EgoReID dataset.

| Methods                          | R-1 | R-5 | mAP |
|----------------------------------|-----|-----|-----|
| K-Res [57]                       | 70.51 | - | 55.12 |
| MSCAN [17]                       | 71.77 | - | 56.05 |
| SpaAttn [18]                     | **82.30** | - | 65.8 |
| PSE+ ECN [35]                    | 76.70 | - | 71.8 |
| MGCAM [34]                       | 77.17 | - | 71.17 |
| DuATN [36] [50]                  | 81.16 | 92.47 | 67.73 |
| Ours                             | 71.20 | 85.70 | 71.8 |
| Ours+RR                          | 79.19 | 89.70 | 65.91 |

Table 4: Comparison to state-of-the-art on MARS dataset. RR is re-ranking using [35].

| Methods                          | R-1 | R-5 | mAP |
|----------------------------------|-----|-----|-----|
| PSE+ ECN [35]                    | 82.07 | **93.84** | **77.96** |

Table 4: Comparison to state-of-the-art on MARS dataset. RR is re-ranking using [35].
7. Conclusion

In this paper, we presented a new EgoReID dataset which is captured using 3 mobile cellphones with non-overlapping FOV. EgoReID presents substantial variations in lighting, scene, background, human pose, etc. Compared with existing video-based ReID datasets, EgoReID poses unique and realistic challenges to person ReID task. Moreover, we have recorded 12 sensor meta data for each video. We have also proposed a novel approach which employs human semantic parsing to extract local discriminative features and compare corresponding semantic regions to find a match. Experiments conducted on MARS and EgoReID showed the effectiveness of our approach in different video domains. In addition, we have also successfully employed sensor meta data information to determine target’s next camera and its estimated time of arrival, thus, we only search for a target in the predicted camera around the estimated time of arrival. This significantly improved our ReID performance by reducing our search space.

Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, pages 59–64. ACM, 2011. 12
[2] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 8624–8628. IEEE, 2013. 12
[3] A. Borji, D. N. Sihite, and L. Itti. What/where to look next? modeling top-down visual attention in complex interactive environments. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(5):523–538, 2014. 3
[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 4724–4733. IEEE, 2017. 4
[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017. 4
[6] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, 2017. 3
[7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 407–414. IEEE, 2011. 3
[8] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In European Conference on Computer Vision, pages 314–327. Springer, 2012. 3
[9] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On, pages 3281–3288. IEEE, 2011. 3
[10] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In CVPR, volume 2, page 6, 2017. 7
[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy. Person re-identification. Springer, 2014. 1
[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), volume 3, pages 1–7. Citeseer, 2007. 12
[13] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In Scandinavian conference on Image analysis, pages 91–102. Springer, 2011. 12
[14] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1062–1071, 2018. 4, 7, 8, 12
[15] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset
for tracking people across multiple cameras. In 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision, volume 5. Citeseer, 2014. 12

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012. 3

[17] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 384–393, 2017. 3, 8

[18] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 369–378, 2018. 8

[19] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. arXiv preprint arXiv:1702.05729, 2017. 3

[20] W. Li and X. Wang. Locally aligned feature transforms across views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3594–3601, 2013. 12

[21] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In Asian Conference on Computer Vision, pages 31–44. Springer, 2012. 12

[22] W. Li, R. Zhao, T. Xiao, and X. Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 152–159, 2014. 12

[23] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In 2013 IEEE International Conference on Computer Vision, pages 3216–3223. IEEE, 2013. 2, 3

[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2197–2206, 2015. 2

[25] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In European Conference on Computer Vision, pages 816–833. Springer, 2016. 3

[26] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. 2009. 1

[27] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 2714–2721. IEEE, 2013. 3

[28] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. Pattern Recognition, 65:197–210, 2017. 3

[29] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1325–1334, 2016. 3

[30] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. arXiv preprint, 2017. 2, 3

[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision, pages 17–35. Springer, 2016. 3

[32] T. Robin, G. Antonini, M. Bierlaire, and J. Cruz. Specification, estimation and validation of a pedestrian walking behavior model. Transportation Research Part B: Methodological, 43(1):36–56, 2009. 7

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015. 12

[34] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 8

[35] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. arXiv preprint arXiv:1711.10378, 2017. 4, 8

[36] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. arXiv preprint arXiv:1803.09937, 2018. 3, 8

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016. 4

[38] Y. T. Tesfaye, E. Zemene, M. Pelillo, and A. Prati. Multi-object tracking using dominant sets. IET computer vision, 10(4):289–298, 2016. 3

[39] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple
non-overlapping cameras using constrained dominant sets.  
[40] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In European Conference on Computer Vision, pages 791–808. Springer, 2016.  
[41] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1288–1296, 2016.  
[42] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. IEEE Trans. Pattern Anal. Mach. Intell., 38(12):2501–2514, 2016.  
[43] X. Wang. Intelligent multi-camera video surveillance: A review. Pattern recognition letters, 34(1):3–19, 2013.  
[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850, 2016.  
[45] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. arXiv preprint arXiv:1708.02286, 2017.  
[46] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1345–1353, 2016.  
[47] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.  
[48] S.-I. Yu, Y. Yang, and A. Hauptmann. Harry potter’s marauder’s map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 3714–3720. IEEE, 2013.  
[49] E. Zemene, Y. Tariku, H. Idrees, A. Prati, M. Pelillo, and M. Shah. Large-scale image geo-localization using dominant sets. arXiv preprint arXiv:1702.01238, 2017.  
[50] J. Zhang, N. Wang, and L. Zhang. Multi-shot pedestrian re-identification via sequential decision making. arXiv preprint arXiv:1712.07257, 2017.  
[51] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1077–1085, 2017.  
[52] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In European Conference on Computer Vision, pages 868–884. Springer, 2016.  
[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, pages 1116–1124, 2015.  
[54] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984, 2016.  
[55] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, et al. Person re-identification in the wild. In CVPR, volume 1, page 2, 2017.  
[56] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. arXiv preprint arXiv:1701.07717, 2017.  
[57] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 3652–3661. IEEE, 2017.  
[58] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 6776–6785. IEEE, 2017.  
[59] Z. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. IEEE Transactions on Image Processing, 27(11):5683–5695, 2018.  

A. Appendices

Here, we provide several details on our dataset. We have included plots, sample tracklets and videos to help better explain our EgoReID dataset. To further elaborate the effectiveness of the proposed person ReID model, we have included qualitative results of some randomly selected queries.

B. EgoReID Dataset

In fig. 8 - 15, we have included plots for several sensor meta data from camera 2, namely, heading, accelerometer, gravity, gyroscope, magnetic, orientation, rotation and speed.
Fig. 5 shows sample tracklets from each camera. Each row contain tracklets of the same person, while each column represent different camera. As can be noted from fig. 5, our dataset contains several illumination, pose and background changes both within and across cameras.

Fig. 6 and 7 provide more detailed statistics on our dataset. For instance, camera 3 captures the most IDs and tracklets. Camera 2 captures the second most IDs but ranked third in total number of tracklets it contain.

C. Qualitative Results

**Person re-identification:** In Fig. 16 - 17, given different example queries from EgoReID dataset, top 10 retrieved tracklets from the gallery are shown for different methods. In all examples, the first row is a result from Inception-v3 where we extract frame level global features and perform average pooling for each tracklet. The second row represent results where we apply [14] on every frame in a tracklet and apply average pooling along the temporal dimension. While the third and fourth rows show results of our method before and after employing sensor meta data information. Correctly matched tracklets are shown with green border while the incorrect ones are show in red. To avoid clutter, we only show three frames per tracklet.

D. Implementation Details

We train the feature extractor module for 200K iterations using the training set consists of 10 image-based ReID datasets (Market-1501 [53], CUHK01 [21], CUHK02 [20], CUHK03 [22], DukeMTMC-reID [56], 3DPeS [1], PRID [13], PSDB [44], Shinpuhkan [15] and VIPeR [12]), then fine-tune it on the video datasets for 20K iterations. The initial learning rates for these two processes are 0.01 and 0.001, respectively, and we decay them 10 times. The minibatch size is set to 8 and the input images with the size of $512 \times 512$ are used. The other settings are similar to the mentioned above. We perform the training of the re-identification and the segmentation models using Nesterov Accelerated Gradient [2] and used the pre-trained InceptionV3 models on ImageNet [33].
Figure 5: Sample tracklets from EgoReID. Each row corresponds to different identities, while each column represents tracklets from cameras 1, 2 and 3. (from left to right). As can be noted, in addition to changes in background, pose and illumination of the same tracklet across camera (each row), due to camera motion, different tracklets from the same camera (each column) are also captured with different background, illumination and poses. This makes our dataset challenging but more close the reality.

Figure 6: The number of IDs captured by each camera.

Figure 7: The number of tracklets per each camera.
Figure 8: Heading plot.

Figure 9: Accelerometer (x,y,z) plot.
Figure 10: Gravity (x,y,z) Plot.

Figure 11: Gyroscope (x,y,z) plot.
Figure 12: Magnetic (x,y,z) Plot.

Figure 13: Orientation (x,y,z) plot.
Figure 14: Rotation Vector (x,y,z) plot.

Figure 15: Speed plot.
Figure 16: As can be seen from the results, Inception-v3 fail to correctly match the query in its top ten results while our approach without sensor meta data is able to find a match in rank-2 and after applying sensor meta data information our result further refined to rank-1.

Figure 17: Inception-v3 and our approach without meta data are able to find the correct match in their top-5. However, after refining our result using sensor meta data, we are able to get the correct match in rank-1.