Person Re-identification Using Multiple Egocentric Views

Anirban Chakraborty, *Member, IEEE,* Bappaditya Mandal, *Member, IEEE,*
and Junsong Yuan, *Senior Member, IEEE*

***Abstract*—Development of a robust and scalable multi-camera surveillance system is the need of the hour to ensure public safety and security. Being able to re-identify and track one or more targets over multiple non-overlapping camera Field-of-Views in a crowded environment remains an important and challenging problem because of occlusions, large change in the viewpoints and illumination across cameras. However, the rise of wearable imaging devices has led to new avenues in solving the re-identification (re-id) problem. Unlike static cameras, where the views are often restricted or low resolution and occlusions are common scenarios, egocentric/first-person-views (FPVs) mostly get zoomed in, un-occluded face images. In this paper, we present a person re-identification framework designed for a network of multiple wearable devices. The proposed framework builds on commonly used facial feature extraction and similarity computation methods between camera pairs and utilizes a data association method to yield globally optimal and consistent re-id results with much improved accuracy. Moreover, to ensure its utility in practical applications where large amount of observations are available every instant, an online scheme is proposed as a direct extension of the batch method. This can dynamically associate new observations to already observed and labeled targets in an iterative fashion. We tested both the offline and online methods on realistic FPV video databases, collected using multiple wearable cameras in a complex office environment and observed large improvements in performance when compared to the state-of-the-arts.**

***Index Terms*—Person re-identification; Egocentric videos; Wearable devices; Face recognition; Multi-camera surveillance.**

I. Introduction

THE past few years have observed efforts of unprecedented scale to develop robust, reliable and scalable visual surveillance systems, fueled by the advancement of imaging sensor technology. As more sophisticated and cheaper imaging devices become commercially available everyday, a large number of such devices (e.g., networked cameras) are being deployed to continuously monitor very large crowded facilities like shopping malls, public transportation hubs, city streets etc. to ensure public safety and security. It is no longer feasible to manually process and analyze these enormous volumes of data stream every second, due to the amount of human supervision and costs involved. This, in turn, yielded an important and challenging computer vision problem - *person re-identification*.

A. Chakraborty (achak002@ucr.edu) and J. Yuan (jsyuan@ntu.edu.sg) are with the Rapid-Rich Object Search (ROSE) Lab and School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.

B. Mandal (bmandal@i2r.a-star.edu.sg) is with the Institute for Infocomm Research (I^2R), A*STAR, Singapore 138632.

Fig. 1. Illustrative diagram for person re-identification using multiple first-person-view cameras. Three wearable devices (Google Glasses), shown as Cam 1-3, are worn by security personnels at different levels in a multi-stored congested shopping mall. As a target appears in the Field-of-View (FoV) of a camera (shown using trapezoidal regions on the ground plane), uncluttered face shots of the target can be observed and processed for re-identification.

While monitoring large areas under surveillance, the Field-of-Views (FoV) of the cameras in the network are often non-overlapping and targets can disappear in the large *blind gaps* as they move from one camera FoV to another. A typical re-identification (re-id) problem is an inter-camera target association problem, where the task is to automatically keep track of individuals or groups of targets in a non-overlapping camera network. Once a target leaves a camera FoV, the re-id system must be capable of re-associating the same target when it reappears at a different location and time.

Person re-identification has remained a challenging and a partially open problem in the computer vision literature despite continuous research efforts because of a number of challenges associated with the data. Typically, the surveillance cameras are set up to capture wide area videos and hence the individual targets are often few pixels in size in these large FoVs. Naturally, capturing discriminative biometric information for individuals (such as facial features) has been very challenging and unreliable for these surveillance cameras as the faces are often negligible in size and/or heavily occluded because of pose of the target or obstructions in the scene. Hence, visual appearance features (such as color/textures) of the observed targets [1], [2] are still first choices for most person re-id systems. Unfortunately, these features are often non-discriminative because of similar colored clothing of targets and are heavily affected by clutter, occlusion and wide variation of viewpoint and illumination across different camera FoVs. Moreover, because of the aforementioned problems, the

appearance features from the same target may appear very differently from one camera to another.

In recent past, there has been a rapid rise in the development of microelectronic devices, enabling wearable sensors and mobile devices with unprecedented video acquisition and processing capabilities. Google Glasses (GG) [3] and GoPro [4] are just two of many such devices. These wearable devices can capture, record and analyze egocentric (also termed as First-Person-View) video data for human identification [5], [6], [7], which is of paramount interest especially for visual surveillance or monitoring, assistance to elderly, social interactions, etc. These wearable devices (such as GG) can easily be networked so that they can communicate and share information among each other as well as with a remote server.

A network of multiple First-Person-View (FPV) cameras on modern wearable devices such as GGs could, therefore, be a good solution to alleviate the aforementioned challenges in person re-identification, as they can supply zoomed in, uncluttered face shots of targets. Besides, unlike static mounted cameras, the observers wearing these wearable cameras can be placed at any location and hence a more robust and fool-proof surveillance system can be designed. An example scenario of wide area monitoring using three GGs are shown in Fig. 1. Three observers wearing the glasses are monitoring a large multi-storied shopping mall. Whenever a person appears in the FoV of any of the GGs, unconstrained high quality face shots of the target are captured and compared against observations from other cameras for rapid re-identification.

In this paper, we present a framework for person re-identification using multiple wearable cameras supplying egocentric/FPV facial images of each target. For this, we have successfully combined the state-of-the-arts holistic discriminative feature computation methods from the FPV face recognition literature with the robust data association techniques reported in the person re-identification community. The proposed framework starts by extracting facial features from each detected face and then feature similarities are computed between targets across wearable camera pairs. In the network of more than two wearable cameras, multiple paths of association may exist between observations of the same target in different cameras and this often gives rise to the network inconsistency problem. Moreover, unlike classic person re-id problem, not all the persons are observed in all the cameras. All pairwise similarity scores, computed in the first step are, therefore, input to a global data association method known as *Network Consistent Re-identification* (NCR) [8], [9] that yields the final association results and can handle both the aforementioned challenges. We have also collected a wearable device re-id database where FPV videos of 72 targets are captured using 4 GGs in a realistic and complex office environment. Through experiments on this dataset, we show that the NCR not only forces consistency in association results across the network, but also improves the pair-wise re-identification accuracies.

The re-id method, described above is very recently introduced in [10]. However, in this paper, we not only present the method in a substantially more detailed manner, but we also extend it to propose an online person (face) re-identification framework. In a large network of wearable devices, numerous targets are observed every instant and the task is to assign identification labels on each of these observations as and when they become available. Thus, in such a realistic scenario, it is often not feasible to solve the association via an offline optimization problem, as the computational complexity rapidly increases with large number of observations (see Fig. 13). The proposed online person re-id works in an iterative fashion over small successive time windows. At any iteration, the goal is to associate a set of unlabeled observations acquired in the most recent time window to the past observations, given that the associations amongst the past observations are already solved. We utilize the online NCR [9] for efficiently solving this iterative global data association, which limits the size of the problem in each iteration and thereby keeps the large re-id problem tractable. Experiments are done on a new dataset collected using 3 GGs and 14 targets (79 observations as most targets were observed more than once in each GG) and the results indicate robustness of the proposed online re-id method.

A. Related Work

Classic Person Re-identification: Person re-identification using multiple FPVs or egocentric views is a new approach. In the classical person re-identification problem, typically the camera FoVs are wide and whole targets are observed at a distance. Hence, the low resolution of the targets is often the main source of challenge in person re-identification. The existing camera pairwise person re-identification approaches can be roughly divided into 3 categories- (i) discriminative signature based methods [11], [2], [1], [12], [13], [14], [15], (ii) metric learning based methods [16], [17], [18], [19], [20], and (iii) transformation learning based methods [21], [22]. Multiple local features (color, shape and texture) are used to compute person specific discriminative signatures [2], [1], [12], [13], [14]. Metric learning based methods learn optimal non-Euclidean metric defined on pairs of true and wrong matches to improve re-id accuracy [19], [23], [24]. Transformation of features between cameras is learned via a brightness transfer function (BTF) between appearance features [22], a subspace of the computed BTFs [21], linear color variations model [25], or a Cumulative BTF [26] between cameras. In [27], the matching is conducted in a reference subspace after both the gallery and probe data are projected into it.

In a recent work [28], video based modeling is introduced to solve the re-id problem. A deep filter pairing neural network was utilized in [29] to attain better re-id accuracy. Recent approaches based on sparse coding and sparse dictionary learning have reported promising results in person re-identification under occlusion [30] and viewpoint variation [31]. But all of these methods suffer from the inherent challenges in person re-id datasets, viz., weakly discriminative features because of low resolution, occlusion and dependence on color/texture based features because of inability of capturing high-resolution, discriminative facial images.

Face identification in first-person-views: Person identification using faces obtained from static surveillance cameras under unconstrained environment has been a very challenging problem [32]. For humans, identifying individuals at a long distance (low resolution face images and/or with occlusions)

has been easy as compared to machine identification of faces [33]. Using a network of wearable devices, as shown in Fig. 1, we envisage that identifying an individual would be easier as compared to using only static cameras. Unlike static cameras, wearable device cameras (like GGs) can capture faces in non-occluded conditions with good resolutions, especially in cases like social interactions [34], surveillance and monitoring. The good thing about capturing face images and recognizing them is that it does not involve the person to volunteer or the person is not aware and hence it is non-intrusive. In addition to the identity, human face brings many other attributes of the owner such as emotion, trustworthiness, intension, personality, aggressiveness, etc [35]. Hence, FPVs face images captured by the wearable devices are important to analyze.

The main difficulties that face identification (FI) algorithms have to deal with are two types of variations: intrinsic factors (independent of viewing conditions) such as age and facial expressions and extrinsic factors (dependent on viewing conditions) such as pose, occlusion and illumination. The availability of high quality wearable cameras such as GG and GoPro and their networking has helped in capturing face images at multiple instances/places alleviating the problems arising from extrinsic factors. Gan Tian *et al.* in [34] used a network of wearable devices along with other ambient sensors to quantify/evaluate the quality of presenters making presentations in a conference/classroom setting. Many researchers have begun collecting FPV videos for FI or memories for faces on GG as a standalone device and also via bluetooth connection with mobile phones [5], [36]. A large number of local features with many distance measures on a wearable device database is evaluated in [37]. They have shown that when a large number of samples per person are available in the gallery binarized statistical image features (BSIF) outperform many other local features. Face images are of high dimensionality and hence, extracting local features are time consuming. These local features are typically of > 250 dimensions making it unattractive for wearable devices which has limited computational resources [38].

Consistent Data Association: Although the high quality facial features captured using wearable devices are more discriminative in general than the typical color/texture based features used in person re-id, they are still camera pairwise and has to be processed by a global data association method for generating consistent and improved results at the network level. Some recent works aim to find point correspondences in monocular image sequences [39] or links detections in a tracking scenario by solving a constrained flow optimization [40], [41] or using sparse appearance preserving tracklets [42]. Another flow based method for multi target tracking was presented in [43], which allows for one- to-many/many-to-one matching and therefore can keep track of targets even when they merge into groups. The problem of tracking different kinds of interacting objects was formulated and solved as a network flow mixed-integer program in [44]. With known flow direction, a flow formulation of a data-association problem will yield consistent results. But in data-association problems with no temporal or spatial layout information (e.g. person re-identification), the flow directions are not natural and thus the performance may widely vary with different choices of temporal or spatial flow. Recently, in [8], a network-consistent re-identification (NCR) method is presented, which does not require time order information of observations and proposes a scalable optimization framework for yielding globally consistent association results with high accuracy. However, [8] shows experiments on a wide area database and does not utilize face as an important cue for re-identification.

Using the transitivity of correspondence, point correspondence problem was addressed in a distributed as well as computationally efficient manner [45]. However, Consistency and transitivity being complementary to each other, less computation comes at the cost of local conflicts and mismatch cycles in absence of any consistency constraints, requiring a heuristics based approach to correct the conflicts subsequently. The proposed NCR approach, on the other hand, uses maximal information by enforcing consistency and produces a globally optimal solution without needing to correct the correspondences at later stages.

Differences with [10]: As mentioned earlier, a preliminary version of the batch person re-identification method is recently presented in [10]. In the present paper, we extend the batch method to propose a new online person re-id framework (Sec. II-C and Fig. 3). The discussion on the batch NCR method is also expanded substantially (Sec. II-B, Fig. 2). A large number of comparative experiments on old and new (Fig. 10) FPV databases (with and without timestamps) using batch and online methods are performed. Along with PCA, FisherFaces and WSSDA, we have added one more FI method (MSDA) for comparison (parts of Fig. 6, 8, Table I for batch NCR). We have also added ROC curves (Fig. 9) for a better comparison of offline re-id accuracy and shown example test cases (Fig. 7) to highlight improvements attained by NCR in rank-1 performance. Experiments on online re-id are shown in Sec. III-D, Figs. 11, 12. We also provide a comparison between the computation times for the batch and the online methods with increasing number of observations (Fig. 13) and show that the online method is more time and memory efficient.

II. Person Re-identification from Multiple First Person Views

The proposed re-identification pipeline has two distinct parts cascaded to one another -

1. Computation of features from acquired first person view images in each device and subsequent estimation of feature similarity/distance scores between all pairs of observations in each camera pair. Following the general and widely accepted assumption in person re-identification problem set up, we assume that the observations from the same target in the same camera field of vision (FoV) can be clustered a-priori and hence intra-camera similarity score computation is not required in this problem.

2. When observations are acquired using more than two wearable devices/cameras, *network consistency* is enforced using network consistent re-identification framework. The inter-camera similarity scores computed in step 1 are used as inputs to this system and outputs are the final association labels between pairs of observations across any two camera.

The online re-id pipeline is also comprised of the same two components. However, it is an iterative framework that needs to associate newly observed targets in a temporally sliding window to all the past observations, given that the associations between the past observations were already estimated through the previous iterations. Thus, all association labels between the past observations are also utilized as inputs to the second part of the online person re-id framework.

A. Preprocessing and Feature Extraction

In the incoming image captured using wearable device, we apply OpenCV face detector [46] to find faces. If a face is found, we apply OpenCV eye detector [46] and integration of sketch and graph patterns (ISG) [47] based eye detector to locate the pair of eyes in oblique and frontal views. Through the fusion and integration of both eye detectors, high success rate of eye localization in the face images of FPV for both frontal and non-frontal faces at various scales (sizes) are achieved. Its fusion system could achieve over 90% accurate rate for frontal view cases and over 70% accurate rate for non-frontal view cases [5]. Using the detected eye coordinates, faces are aligned, cropped and resized to 67×75 pixels. Same normalization procedure is followed as described in [5]. To overcome the limitations discussed in subsection I-A, we use the whole space subclass discriminant analysis (WSSDA) method for face recognition proposed recently in [48]. This approach extracts holistic discriminant features from diverse face images which are of low dimensions and is attractive among many related approaches and suitable for wearable devices [5].

1) Within-Subclass Subspace Learning for Face Identification: FI performance is constantly challenged by unconstrained pose, lighting, occlusion and expression changes. Classical discriminant analysis methodologies employing between-class and within-class scatter information lose crucial discriminant information [49], [50], [51] and fail to capture the large variances that exist in the appearance of same individual (within-class). For example, mixture subclass discriminant analysis (MSDA), an improvement over subclass discriminant analysis [52] for face recognition, is presented in [53]. In this approach, a subclass partitioning procedure along with a non-Gaussian criterion are used to derive the subclass division that optimizes the MSDA criterion, this has been extended to fractional MSDA and kernel MSDA in [54]. However, these approaches discard the null space of either within-class and within-subclass scatter matrices, which plays a very crucial role in the discriminant analysis of faces.

In WSSDA [48] each class is partitioned into subclasses using spatial partition trees and then eigenfeature regularization methodology [55] is used to alleviate the problems of modeling large variances appearing in within-class face images (images of an individual). This regularization of features has facilitated in computing the total-subclass and between-subclass scatter matrices (depending on the clusters for each person and the number of people in the database) in the original dimensionality of face images. Dimensionality reduction and feature extraction are performed after discriminant evaluation in the entire within-subclass eigenspace.

When training is complete, only the low dimensional gallery features and transformation matrix are stored in the system. For enrollment of a new person, the incoming face images are transformed using the training module (transformation matrix) and only the gallery features are stored. In the recognition phase, any incoming face image vector is converted into a feature vector using the transformation matrix learned by WSSDA method. The feature vector is used to perform recognition by matching it with the gallery features. Using cosine distance measures with 1-nearest neighbor (NN) as the classifier. [48] has evaluated this methodology on the popular YouTube unconstrained face video database [56] and also FPV face videos [37]. For comparison purpose we use the popular holistic features for FI, such as baseline principal component analysis (PCA) [57], FisherFaces using PCA+linear discriminant analysis [58] and mixture subclass discriminant analysis (MSDA) [53] to show that using various methods we can have large improvement in the person re-identification accuracy.

For face recognition, another class of emerging algorithms is the deep learning which uses convolutional neural network and millions of (external) face images for training and obtain very high accuracy rates [59], [60]. However, our chosen method is still attractive because it uses small number of training samples and does not use any external training data but can achieve comparable performances [48].

Global data association: Once the feature similarities are computed between pairs of observations across cameras, the next step is to estimate associations between these observations using a global data association method. As mentioned earlier, the Network Consistent Re-identification (NCR) is used for this purpose.

Re-identification between observations across cameras can be performed via two strategies - (i) batch re-id, where all the observations are available and a globally optimal set of association labels are estimated in one shot, or (ii) online re-id, where more observations are input to the system as time progresses and the objective is to associate the newly observed targets to the past observations as and when they become available. For most practical scenarios, the complete re-id system should be online as, in real life, flow of observations is continuous. Moreover, for a large number of cameras and targets, a batch data association framework is often computationally expensive and hence infeasible.

In the next subsections we present the NCR method in detail. Please note that the overall re-id strategy is online (see Fig. 3) that iteratively associates new unlabeled observations in a time window to all the labeled observations from the past (online NCR) and a constrained version of the batch NCR can be assumed as the building block of the online method. Therefore, to facilitate a better understanding of the overall re-id scheme, we present the construction of the batch NCR problem first in Sec. II-B. Here, we define the terminologies and the notations associated with NCR, introduce the general objective function and the constraints. Then, once the fundamentals of the NCR (batch) is thoroughly explained, we elaborate on how these objective functions and the constraints can be modified to formulate the online NCR based re-id problem in Sec. II-C.

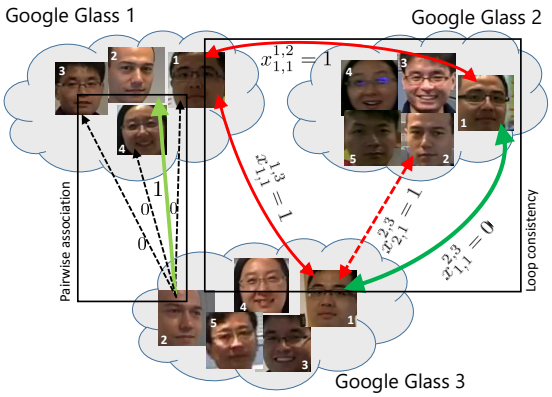


Fig. 2. An illustrative example showing the importance of the pairwise association constraint (left set of arrows) and loop/consistency constraint (right set of arrows) in a data-association problem. It presents a simple person re-identification scenario in a network involving 6 unique targets (data-points/nodes) in 3 wearable devices (groups). A graph is constructed by joining nodes belonging to different groups with edges. The target of re-id (NCR) is to assign labels (0-not associated/1-associated) on these edges under two sets of constraints. According to the pairwise association constraint one target from one camera FoV (e.g. target 2 in GG 3) may have at most one match in another camera (e.g., GG 1). Also, to maintain consistency in associations over the entire network, all paths associating two nodes must conform to one another. For example, target 1 in GG 2 and the same target in GG 3 have no pairwise association, although they are associated via an indirect path through GG 1, thereby violating the loop constraint.

B. Batch Estimation of the Final Associations: Network Consistent Re-identification

The problem of network inconsistency in classic person re-identification tasks was introduced in [8] and later expanded in [9]. A binary integer program to establish consistency in re-identification and thereby improving association accuracy was proposed in these works and termed as Network Consistent Re-identification (NCR) or Network Consistent Data Association (NCDA). We shall use NCR/NCDA interchangeably through this paper to refer to the same optimization method.

Following similar notations used in [8], we denote an observation i in camera/device g as $\mathcal{P}_i^g$. In previous section, we estimate feature similarity/distance between pairs of observations across cameras and let $c_{i,j}^{p,q}$ denote the similarity score estimated between features from observations $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$, observed in camera p and q respectively. The expected output of the NCR framework is a set of association labels between each of these pairs of observations. Thus, if each of the observations are considered as node in a network, clusters of nodes observed in the same camera can be termed as 'groups' and edges can be constructed between pairs of nodes belonging to different groups. The goal is to estimate a label $x_{i,j}^{p,q}$ for each such edge that will denote whether the two nodes associated with this edge are from the same target, i.e., $x_{i,j}^{p,q}=1$, if $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ are the same targets and $=0$, otherwise.

A 'path' between two nodes $(\mathcal{P}_i^p, \mathcal{P}_j^q)$ is a set of edges that connect the nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ without traveling through a node twice. Moreover, each node on a path belongs to a different group. A path between $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ can be represented as the set of edges $e(\mathcal{P}_i^p, \mathcal{P}_j^q) = \{(\mathcal{P}_i^p, \mathcal{P}_a^r), (\mathcal{P}_a^r, \mathcal{P}_b^s), \cdots (\mathcal{P}_c^t, \mathcal{P}_j^q)\}$, where $\{\mathcal{P}_a^r, \mathcal{P}_b^s, \cdots \mathcal{P}_c^t\}$ are the set of intermediate nodes on the path between $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$.

1) Constraints in Data Association: As the first step of NCR, all the observations within each camera FoV (or for the online re-id, all observations in each camera within a time window) are first clustered based on the extracted facial features so that all the image observations in each cluster are from the same target. In consecutive frames in a FPV video, the view point of the observer as well as the illumination in the scene remain more or less constant. Also, the pose of the target face with respect to the camera does not vary substantially in successive frames as, in most situations, motion of the target is straight towards the camera and mostly frontal face shots are observed. The only variable in the observed faces in consecutive time points is the gradually increasing resolution of the face of the target. Even if there are minor misalignments between captured faces of a target, the robust eye localization based preprocessing step aligns the faces. Finally, all the faces are normalized [5] to alleviate the problem of variable size/resolution (see Sec. II-A). This makes the problem of clustering observations from the same target in successive frames an easier task. Given all the detected and aligned faces in a camera FoV, pairwise feature distances are computed using the same methods as in inter-camera (e.g. WSSDA) and a clustering method [61] is employed to group observations from the same targets within each camera FoV.

After clustering, each cluster is treated as one observation and such observations (sets of images belonging to the same target as observed in consecutive frames in one camera) are associated across camera FoVs using the NCR method. Now, because of this a-priori clustering there can be only one observation (image set) from the same target in one camera FoV. As a result, an observation $\mathcal{P}_i^p$ in camera p may have at most one matching observation in any other camera q. If the same set of targets appear in all the camera FoVs, there is an exact one-to-one match between observations across any two camera pairs. However, in a realistic scenario, a target may or may not appear in every camera FoV and hence, $\forall\, x_{i,j}^{p,q} \in \{0,1\}$,

$$\sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \;\forall i = 1 \text{ to } n_p,\; \sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \;\forall j = 1 \text{ to } n_q \quad (1)$$

This is referred to as the 'pairwise association constraint' in NCR/NCDA. An illustrative example of the pairwise constraint is shown in Fig. 2. Of all the edges connecting target 2 in GG 3 to all targets in GG 1, only one has label 1 and the rest must have label 0 to satisfy this constraint.

Now, pairwise associations must also be consistent over the network of camera FoVs. This set of conditions is important when there are three or more cameras/wearable devices to capture FPV images. The consistency condition simply states that if two nodes (observations) are indirectly associated via nodes in other groups, then these two nodes must also be directly associated. Therefore, given two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$, it can be noted that for consistency, a logical 'AND' relationship between the association value $x_{i,j}^{p,q}$ and the set of association values $\{x_{i,a}^{p,r}, x_{a,b}^{r,s}, \cdots x_{c,j}^{t,q}\}$ of any possible path between the nodes has to be maintained. The association value between

the two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ has to be 1 if the association values corresponding to all the edges of any possible path between these two nodes are 1. Keeping the binary nature of the association variables and the pairwise association constraint in mind the relationship can be compactly expressed as,

$$x_{i,j}^{p,q} \geq \left(\sum_{(\mathcal{P}_k^r, \mathcal{P}_l^s) \in e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)} x_{k,l}^{r,s} \right) - |e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)| + 1 \quad (2)$$

$\forall$ paths $e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$, where $|e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)|$ denotes the cardinality of the path $|e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)|$, *i.e.* the number of edges in the path. The relationship holds true for all i and all j. Now, any network containing even a large number of wearable devices/cameras can be exhaustively expressed as a collection of non-overlapping triplet of cameras. For triplets of cameras the constraint in Eqn. (2) simplifies to,

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 2 + 1 = x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \quad (3)$$

The loop/consistency constraint can be easily verified from the example case shown in Fig. 2. Say, the raw similarity scores between pairs of targets across the GGs suggest associations between (target 1 in GG 1, target 1 in GG 2), (2 in GG 2, 1 in GG 3) and (1 in GG 1, 1 in GG 3) independently. However, combining these associations over the entire network leads to an infeasible scenario - targets 1 and 2 in GG2 have the same identity. The constraint in Eqn. (3) also successfully capture this infeasibility, *i.e.*, $x_{1,1}^{2,3} = 0$ but $x_{1,1}^{1,2} + x_{1,1}^{1,3} - 1 = 1$, thus violating the loop constraint.

2) Re-identification as an Optimization Problem: Under the constraints expressed by Eqn. (1) and Eqn. (3), the objective is to maximize the utility $\mathbf{C} = \sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q}$. However, this utility function is only valid for one-to-one re-identification case, as this may reward both true positive and false positive associations (for example, when $c_{i,j}^{p,q} \in [0, 1]$), and hence the optimal solution will try to assign as many positive associations as possible across the network. This will yield many false positive associations. One way of avoiding such a situation in the current framework is to modify the utility function as $\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q}$, where there are m cameras in the network and k is any value within the range of $c_{i,j}^{p,q}$ $\forall i, j, p, q$. The value of k can be learned from the training data (see Sec. III-C1) so that the true-positives are rewarded and false-positives are penalized as much as possible. Therefore, by combining the utility function with the constraints in Eqn. (1) and Eqn. (3), the overall optimization problem for m wearable devices with variable number of observations is written as,

$$\underset{\substack{x_{i,j}^{p,q} \\ i=[1,\cdots,n_p] \\ j=[1,\cdots,n_q] \\ p,q=[1,\cdots,m]}}{\text{argmax}} \left(\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \right)$$

$$\text{subject to } \sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \ \forall i = [1, \cdots, n_p] \, \forall p, q = [1, \cdots, m],$$

$$\sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \ \forall j = [1, \cdots, n_q] \ \forall p, q = [1, \cdots, m], \ p < q$$

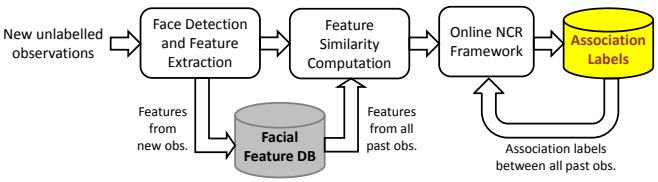


Fig. 3. Online person re-identification system diagram. The online system works in an iterative fashion over small successive time windows. At any iteration, a set of unlabeled observations acquired in the most recent time window is associated to the past observations, given that the associations between the past observations are already solved through the previous iterations.

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1$$
$$\forall i = [1, \cdots, n_p], \ j = [1, \cdots, n_q], \ k = [1, \cdots, n_r],$$
$$\forall p, q, r = [1, \cdots m], \text{ and } p < r < q$$
$$x_{i,j}^{p,q} \in \{0, 1\} \ \forall i = [1, \cdots, n_p], \ j = [1, \cdots, n_q],$$
$$\forall p, q = [1, \cdots, m], \ p < q \quad (4)$$

This is a binary integer linear program (ILP) and optimal solution can be efficiently computed using exact algorithms.

C. Online Person Re-identification

The person re-identification framework presented in the last subsection (Sec. II-B) is targeted towards the classical re-id problem, where all observations are assumed available a-priori before the data association is solved in a batch setup. However, in a realistic setup it is hardly the case. In a large network of wearable cameras deployed for the purpose of surveillance, numerous targets are observed every instant and the task is to assign identification labels on each/many of these observations within a short turnaround time. Besides, in such a network of wearable cameras it is often not feasible to solve the association via a batch optimization problem, as the computational complexity rapidly increases with large number of observations. An online method, on the other hand, only works on a small subset of these observations in an iterative fashion and hence solves a much smaller data association problem in each iteration. This can make the real-world re-identification tractable.

The formulation for the online generalized NCR is presented in [9]. It is a direct theoretical extension of the batch problem (Eqn. (4)), as all the constraints (pairwise/loop) from the batch NCR are preserved. Additionally, the online implementation is capable of handling another realistic scenario that the batch NCR is not designed to. If the same target reappears in the same camera FoV after being observed in some other cameras in the network, the online NCR, unlike the batch method, can correctly re-id the target while maintaining global consistency.

The online person re-identification works in an iterative fashion over small successive time windows. At any iteration, the goal is to associate a set of unlabeled observations acquired in the most recent time window to the past observations, given that the associations between the past observations are already solved. A system diagram showing the online re-id workflow is presented in Fig. 3. For a set of unassociated observations obtained in the most recent time window, first the facial features are extracted and the feature similarities are computed

between these new observations as well as between the new and all past (labeled) observations. The extracted features are stored for future usage. Finally, the similarity scores are fed to the online NCR method that optimally associates the new observations with the past as well as amongst each other. The mathematical details of the online NCR is briefly given below.

Let us assume that there are m groups (cameras) of observations upto time point t and the number of unique observations in group k is $n_k^{(t)}, k = 1, 2, \cdots m$. Thus, until time t, the total number of unique observations is $N^{(t)} = \sum_1^m n_k^{(t)}$. Let us also assume that the $N^{(t)}$ observations are already associated and the association is represented using a set of estimated labels $x_{i,j}^{p,q} = {}^{(t)}x_{i,j}^{p,q}, \ \forall i = [1, \cdots n_p^{(t)}], \forall j = [1, \cdots n_q^{(t)}], p, q = [1, \cdots m], \ p < q$.

In the next time window $[t, t+w]$, say, there are $l^{(w)}$ new observations across different cameras and the objective is to associate these new observations to the already observed targets and among each other. Now, some of these $l^{(w)}$ observations may have temporal overlap with some other new observation and therefore may not be associated with each other. The $l^{(w)}$ new observations can therefore be partitioned into s subsets where no two observations within a subset may have come from the same target. This partitioning problem is analogous to the problem of finding the strongly connected components in a graph (where the observations are nodes and two nodes are connected by a link if they have temporal overlap) and can be efficiently solved using a 'depth-first search'. Thus, $n_{m+1} + n_{m+2} + \cdots n_{m+s} = l^{(w)}$, where n_p is the number of unique observations in the p^{th} subset. Now, based on our definition of a'group', each of these s subsets can be called a *virtual/dummy* 'group'. Thus in the aforementioned time window, the data-association problem can be solved using NCR with a total of $N^{(t)} + l^{(w)}$ nodes and $m + s$ groups.

Each node in a dummy group is connected by edges from all the nodes in the other $m + s - 1$ groups. The goal, now is to optimally assign labels (0/1) to each of these unlabeled edges, given that the data-association between all the past observations ($N^{(t)}$ in m groups) is already solved and available. Let, the set containing all unlabeled edges at any iteration be represented as E_u. Each of these edges involves (at least) one node from the new $l^{(w)}$ nodes. Depending on the design of the online problem (such as the width of the time window, number of cameras etc.), the number of unlabeled edges per iteration ($|E_u|$) can be kept substantially small and hence the problem remains tractable.

The objective function is same as that of generalized NCR (Eqn. (4)), though it is defined only on the set of unlabeled edges (E_u) for the online NCR, i.e.,

$$\sum_{\substack{p,q=m+1 \\ p<q}}^{m+s} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \ + \sum_{\substack{p=m+1, \\ q=1}}^{m+s,m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \tag{5}$$

The association constraints between pairs of groups of observations are same as Eqn. (1), except the fact that at least one of the groups must be a dummy group. This reduces the number of constraints by a large margin. The set of groups over which the pairwise association constraints are defined for online NCR are, $\mathcal{E}^{(w)} = \{(p,q) : p, q \in [1, \cdots m+s], p < q\} \backslash \{(p,q) : p \leq m, q \leq m\}$.

The loop constraints remain the identical as in Eqn. (2) or in the simplified Eqn. (3), but defined on a much smaller subset. In online NCR, each of these inequality constraints must involve at least one unlabeled edge, *i.e.*, at least one edge from the set $\{(\mathcal{P}_i^p, \mathcal{P}_j^q) \cup (\mathcal{P}_i^q, \mathcal{P}_k^r) \cup (\mathcal{P}_i^p, \mathcal{P}_k^r)\}$ must belong to the set of unlabeled edges E_u. So, by combining all the constraints together, the online NCR problem for person re-identification in time window $[t, t+w]$ can be written as,

$$\begin{aligned}
&\underset{\substack{x_{i,j}^{p,q} \\ i=[1,\cdots,n_p],\ j=[1,\cdots,n_q] \\ (p,q)\in\mathcal{E}^{(w)}}}{\operatorname{argmax}} \left(\sum_{\substack{p,q=m+1 \\ p<q}}^{m+s} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \ + \sum_{\substack{p=m+1, \\ q=1}}^{m+s,m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \right) \\
&\text{subject to } \sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \ \forall i = [1, \cdots, n_p], \ \forall (p,q) \in \mathcal{E}^{(w)} \\
&\sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \ \forall j = [1, \cdots, n_q], \ \forall (p,q) \in \mathcal{E}^{(w)} \\
&x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\
&\text{and, } \{(\mathcal{P}_i^p, \mathcal{P}_j^q) \cup (\mathcal{P}_i^q, \mathcal{P}_k^r) \cup (\mathcal{P}_i^p, \mathcal{P}_k^r)\} \ \cap E_u \neq \emptyset \\
&\forall\, i = [1, \cdots, n_p],\ j = [1, \cdots, n_q],\ k = [1, \cdots, n_r] \\
&\forall\, p, q, r = [1, \cdots m+s], \text{ and } p < r < q \\
&x_{i,j}^{p,q} = {}^{(t)}x_{i,j}^{p,q}, \ \forall i = [1, \cdots n_p], \forall j = [1, \cdots n_q], \\
&p, q = [1, \cdots m], \ p < q, \quad \text{and,} \\
&x_{i,j}^{p,q} \in \{0, 1\} \ \forall i = [1, \cdots n_p], \ j = [1, \cdots n_q], \ (p,q) \in \mathcal{E}^{(w)}
\end{aligned} \tag{6}$$

Once the association labels are obtained by solving Eqn. (6), the dummy groups are dissolved and the new observations, labeled according to the association results, are put back to the original groups they belong to. If an observation is associated with a past observation from the same group (camera), they are clubbed together into one node using any suitable fusion strategy.

The Eqn. (6) as well as Eqn. (4) are binary integer linear programs. As the constraint matrices are not consistently totally unimodular, a LP relaxation is not guaranteed to give integer solutions. Hence, we choose to employ exact algorithms to solve the NCR optimization problems. In particular, a 'branch and cut' method is used that combines the branch-and-bound and 'cutting plane' methods. We also set an upper limit on the run-time, and our solution is guaranteed to be feasible.

III. Experimental Results and Analysis

A. Database 1 for Offline Re-identification

4 GGs are used to collect FPV videos of 72 people, out of which 37 are male and 35 are female resulting in about 7077 images. These videos are captured using egocentric

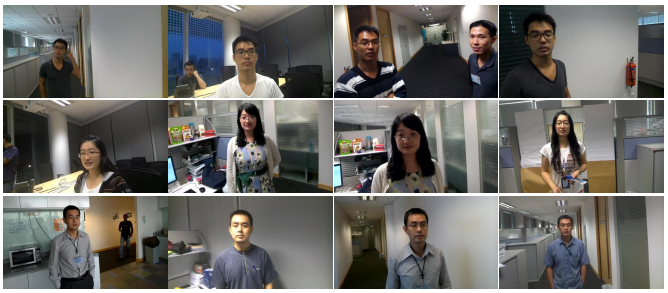
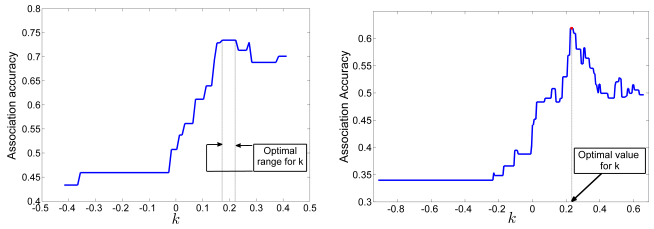


Fig. 4. Original images in Database 1, as captured and seen on the GGs. Each row contains four images of the same persons, observed in 4 GG FoVs (different locations), The targets in the images show diverse appearances and are captured at different scales, poses and illumination conditions.



(a) (b)

Fig. 5. Estimation of optimal k (Eqn. (4)) from an annotated training set. k is varied over the range of similarity scores in the training set and overall association accuracy is computed for each k. (a) shows the variation of accuracy with k for a training set of WSSDA similarity scores and (b) shows the same for PCA based pairwise measures.

views at different levels in corridors, lifts, escalators, pantries, downstairs eateries, passage ways, etc, of a large multi-storied office environment. Cam 1, 2, 3 and 4 (corresponding to the 4 GGs) observe 52, 40, 43 and 50 persons in their respective FoVs. Since both capturing and target people are moving the images are often blurry in nature and they are sometimes out of camera focus. The face and eye detectors as described in section II-A serve as filters to remove images with large motion blur or poor image quality.

Fig. 4 shows some good sample images as captured by the GGs. Each of the three rows shows four images (on 4 GGs) containing the same person at different locations and times. The dataset poses a tough challenge for person re-id as the targets are captured at widely varying scales, poses and illumination conditions. It can be also be observed from Fig. 4 that the same targets often appear in different clothings in different cameras and hence a typical appearance feature based person re-id system may not be applicable in such situations. More details on the database 1 is given in the suppl. materials.

B. Pairwise Similarity Score Generation

Using the normalized images as described in section II-A, we extract features applying various FI algorithms as described in section II-A1. We perform training using various FI algorithms: PCA, FisherFaces, MSDA and WSSDA on the FPV face image database and use the same training strategy as described in [5], [52]. Each class (person) is partitioned by the same number of subclasses (equally balanced). k-means tree is built based on nearest neighbor (NN) clustering of face appearances [5].

During training, we obtain the transformation matrices for each of the methods using the same 42 people for training comprising of 305 images. During the testing phase, novel images are transformed using the transformation matrices obtained from each of the methods into low-dimensional feature vectors. We limit the dimensionality of the final transformation matrix to 80 features ($\times$ the dimensionality of face image vector [5]), so that the final features obtained are of 80 dimensions for each of the normalized face images. We use cosine distance measures with 1-nearest neighbor (NN) as the best match for each of the faces in a frame to generate pairwise scores between the persons observed in each of cameras FoVs.

C. Network Consistent Re-identification (Offline)

1) Test-Train Partitions - Learning k From Training Data: With the pairwise similarity scores generated (as explained in the previous section), the next step is to optimally combine them using the aforementioned Network Consistent Data Association (NCR) method, which yields the final association results. As shown in Eqn. (4), the value of k in the objective function of the integer program is specific to the distribution of the pairwise similarity scores and hence has to be learned from a training set before solving for the association labels.

As we have used four different methods, viz., PCA, FisherFaces, MSDA and WSSDA for pairwise similarity score generation, we generate four separate sets of consistent association results - one for each of these baseline methods. We refer to them as PCA+NCR, FisherFaces+NCR, MSDA+NCR and WSSDA+NCR respectively throughout the rest of the paper. For each of these four methods, we generate 10 sets of exhaustive training-testing partitions (non-overlapping) from the collected dataset. Each set contains 24 randomly selected targets (a third of the dataset) in the training set and the remaining 48 (two thirds of the dataset) are used for testing. The final test results including re-identification accuracies for each method are averaged over these 10 test sets.

To learn k for each of the training sets, first the range of the pairwise similarity scores are identified. As the optimum value of k must lie within this interval, we vary k and compare the accuracy of data association against the ground truth on the annotated training data. The accuracy is computed as $\frac{(\#\text{ true positive}+\#\text{ true negative})}{\#\text{ of unique people in the trainset}}$ and the value of k corresponding to the maximum association accuracy is estimated as the optimal of k and fixed during testing. We show examples of variation of training accuracy with k in Fig. 5. If the maximum accuracy is observed over a range of k (as seen in Fig. 5(a) for WSSDA + NCR case), the mean k over that range is taken as the optimum value. Fig. 5(b) shows another similar plot for learning optimum k for the PCA+NCR experiments.

2) Re-identification Performance Comparisons: Before and After NCR: The re-identification performances of the individual pairwise methods (PCA, FisherFaces, MSDA and WSSDA) are presented and compared - both before and after enforcing the network consistency. First, comparative evaluations are shown in terms of recognition rate as Cumulative Matching Characteristic (CMC) curves and normalized

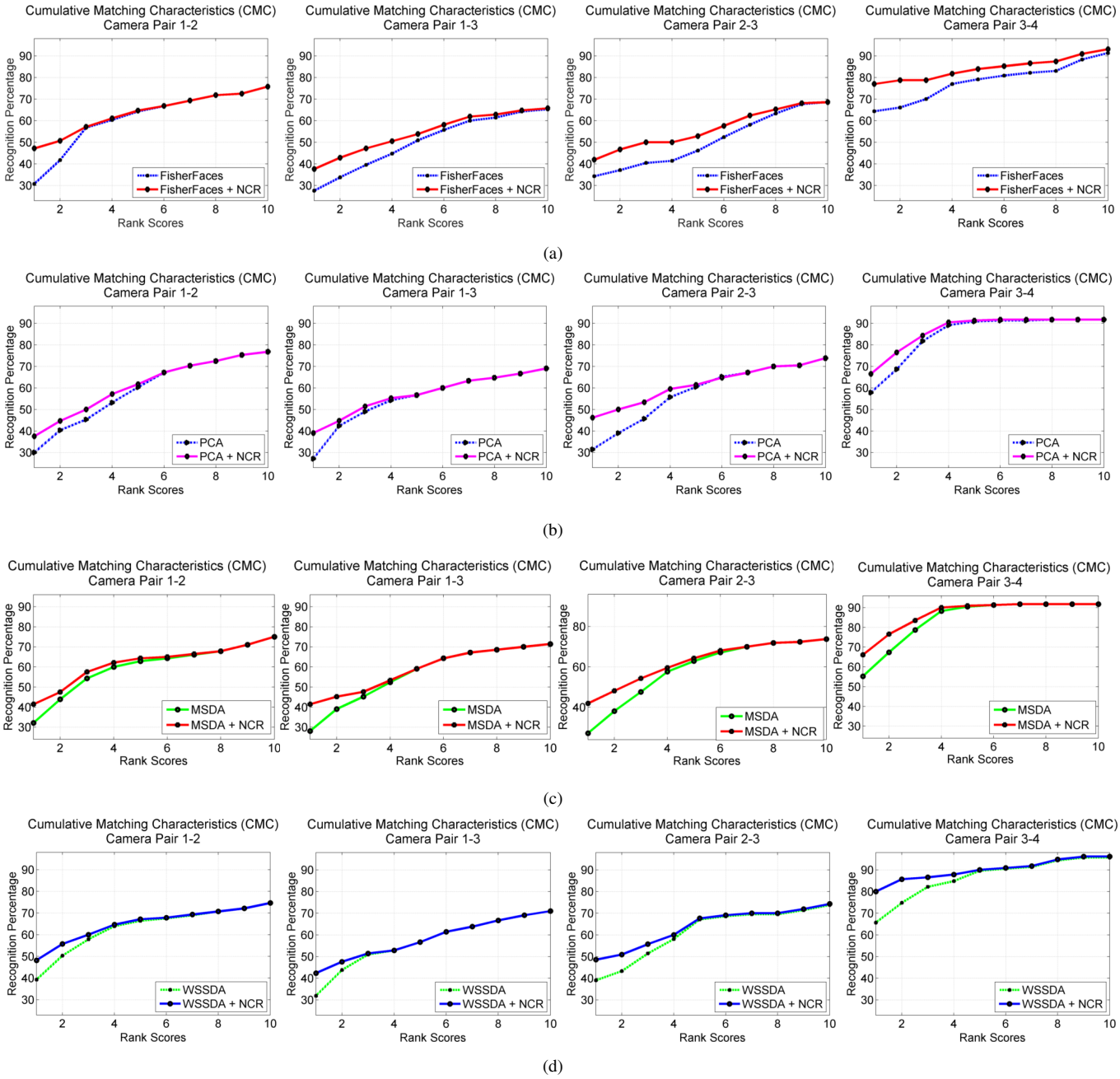


Fig. 6. CMC curves comparing methods (a) FisherFaces, (b) PCA, (c) MSDA and (d) WSSDA respectively, both before and after NCR.

Area Under Curve (nAUC) values, which are the common practice in the literature. The CMC curve is a plot of the recognition percentage versus the ranking score and represents the expectation of finding the correct match inside top t matches. nAUC gives an overall score of how well a re-identification method performs irrespective of the dataset size. Please note that, we are presenting our results in the most generalized test setup where targets may not be visible in all the camera FoVs. Hence while estimating the CMC and nAUC values between any pair of cameras i and j, only those targets in camera i are considered that are also observed in camera j's FoV.

As explained before, the output of NCR based re-id is a set of binary association labels (matched/not matched) between pairs of observations and the similarity scores cannot be re-computed based on these labels. However, to compare improvements obtained by a re-id method before and after NCR, we employ the following strategy to compute the CMC

TABLE I
COMPARISON OF PCA, FISHERFACES (FF), MSDA AND WSSDA WITH THEIR NCR COUNTERPARTS BASED ON NAUC VALUES (UPTO RANK 10).

Cam pair	PCA	FF	MSDA	WSSDA	PCA + NCR	FF + NCR	MSDA + NCR	WSSDA + NCR
1-2	0.5978	0.6187	0.5439	0.6387	0.6179	0.6393	0.5600	**0.6544**
1-3	0.5614	0.5077	0.5155	0.5741	0.5743	0.5484	0.5317	**0.5847**
1-4	0.5349	0.5183	0.4890	0.6508	0.5521	0.5410	0.4957	**0.6717**
2-3	0.5849	0.5090	0.5381	0.6172	0.6185	0.5646	0.5664	**0.6407**
2-4	0.6455	0.5571	0.5900	0.6717	0.6513	0.5817	0.5893	**0.6950**
3-4	0.8570	0.7826	0.7648	0.8708	0.8763	0.8423	0.7863	**0.9017**

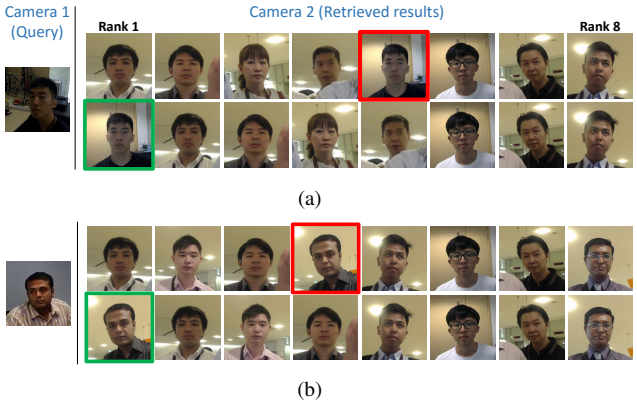


Fig. 7. Improvements in rank-1 re-identification performance are attained when pairwise feature distance computation methods are combined with NCR. (a) (Top row) For the target 13 in GG 1 as query, top 8 retrieved results using PCA from GG 2. The correct match is ranked 5^{th}. (Bottom row) Re-ranking using PCA+NCR which puts the correct match at rank-1. (b) (Top row) For the target 5 in GG 1 as query, top 8 retrieved results using FisherFaces from GG 2. Correct match was ranked as 4^{th} before application of NCR. (Bottom row) FisherFaces+NCR puts target 5 in camera 2 at the rank-1 position, thereby improving the re-id accuracy.

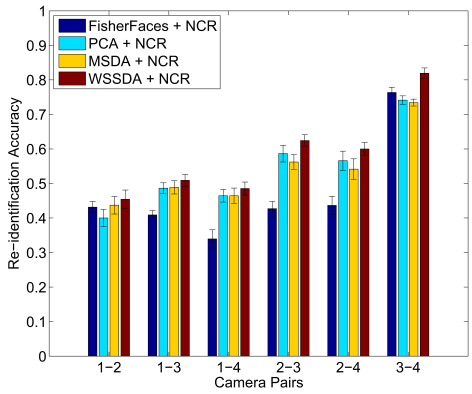


Fig. 8. Comparison of overall re-identification accuracies (combining both true-positives and false-positives).

curves. For the baseline methods (before NCR), the CMC curves are drawn as usual using the similarity scores (real valued, normalized between 0 and 1). Once NCR assigns 0/1 labels to the pairs of observations across cameras, we place the observation corresponding to label 1 at rank 1 position and regenerate the modified rankings. This is analogous to changing the similarity score of the label 1 associations to 1.0 (or to the maximum possible similarity value) and then re-compute the CMC curves.

Figs. 6(a), 6(b), 6(c), 6(d) present the CMC curves for FisherFaces, PCA, MSDA and WSSDA respectively and in each plot, comparisons of the recognition performances are shown before and after application of NCR (e.g., PCA and PCA+NCR in Fig. 6(b)). Plots are shown for camera pairs 1-2, 1-3, 2-3 and 2-4 for every feature computation method. Each CMC is plotted upto rank 10. As observed, amongst the four pairwise re-identification methods, WSSDA is superior to all the other three methods. Moreover, for each of the features and every camera pair, individual pairwise methods are substantially outperformed by their respective NCR counterparts. In particular, WSSDA+NCR achieves the highest rank-1 performances across all camera pairs, such as 49% in camera pairs 1-2 and 2-3 and 80% in camera pair 3-4.

These observations are further established by the nAUC values (computed from CMC until rank 10), as shown in Table I. PCA+NCR, FisherFaces+NCR, MSDA+NCR and WSSDA+NCR individually perform better than the pairwise methods PCA, FisherFaces, MSDA (in 5 of 6 pairs) and WSSDA respectively with WSSDA+NCR showing the best nAUC scores across all 6 camera pairs.

After estimating the binary association labels using NCR, the associations with label 1 between cameras p and q are processed to generate rank-1 matches for a target in camera p (analogous to query) from all targets in camera q (treated as the gallery set). Thus, if the correct match is not returned as rank-1 by a pairwise association method (any face verification method), a further processing through NCR can rectify the error by re-ranking the correct match as the top ranked result. Two example test cases from database 1 are shown in Fig. 7 for camera pair 1-2. For each of these examples, one target from camera 1 is selected as the query and all targets in camera 2 is treated as the gallery set. Fig. 7(a) and (b) top rows show the top 8 retrieved results from camera 2 for PCA and FisherFaces respectively whereas the bottom rows show the same when NCR is combined with the pairwise methods. The correct matches (ranked as 5 and 4 respectively for the pairwise methods) are retrieved as rank-1 when NCR is applied.

3) Overall Re-identification Accuracy by Combining both True Positive and False Positive: A correct re-identification result in a realistic dataset such as ours not only contains correct matches (true positives) but also constitutes of the true negatives, when a target is only observed in a subset of cameras. Hence, the overall accuracy of person re-identification across any pair in the network of wearable devices should be estimated as $\frac{(\#\text{ true positive}+\#\text{ true negative})}{\#\text{ of unique targets in the testset}}$. We compare these accuracy values obtained by NCR when applied on each of PCA, FisherFaces, MSDA and WSSDA similarity measures. From Fig. 8, it can be observed that NCR on WSSDA is more accurate than PCA+NCR, FisherFaces+NCR and MSDA+NCR across all 6 camera pairs, with the best accuracy of more than 80% observed in camera pair 3-4.

We further plot ROC curves to show variation of true positive rate (TPR/recall) with the false positive rate (FPR) for all the baseline and baseline+NCR methods (Fig. 9). For camera pairs 1-4, 2-4 and 3-4, WSSDA+NCR shows best recall values for both low FPR (< 0.1) and for FPR > 0.3. For camera pair 1-2, however, FisherFaces+NCR shows better TPR at low FPR values. For all the camera pairs and all baseline methods, baseline+NCR methods outperform baseline (pairwise) only methods at low FPR values.

D. Online Person Re-identification

In this section, we show experimental results of NCR when applied to the problem of online person re-identification from multiple wearable cameras. The experimental setup for the online re-id is fundamentally different from a classical re-identification problem, where all observations across all

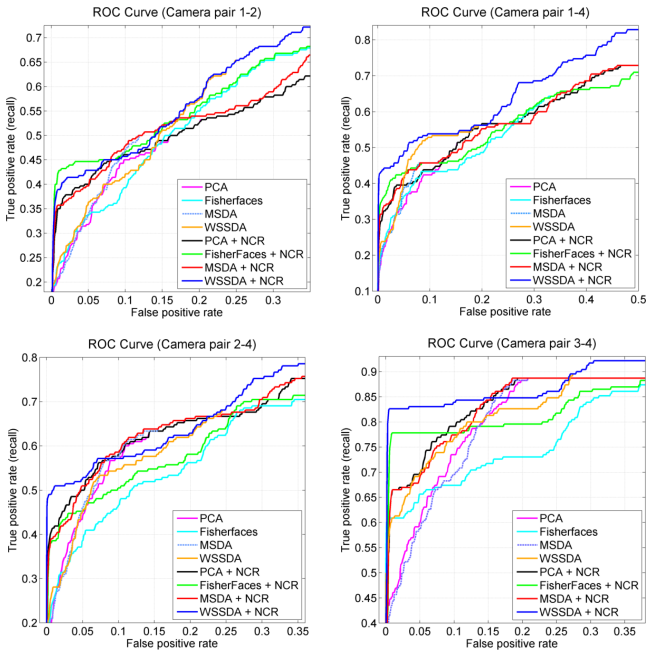


Fig. 9. Portions of ROC curves (FPR$<$ 0.4) for four camera pairs shown for all four baseline methods (viz., PCA, FisherFaces, MSDA and WSSDA) along with the four (baseline+NCR) methods (best viewed in color).

cameras are available a-priori and the data associations are solved in batch. In the online case, however, more observations become available as time progresses and NCR is sequentially applied to associate clusters of observations in the current time window to all past observations. Thus, to generate a temporal stream of data, time information associated to each observation is needed to be known. We assume that the input to the NCR method is a set of tracklets (a temporally consecutive series of observations/images from the same target obtained from within the same camera FoV), which were made available to the NCR at their respective times of appearance. Moreover, as explained in Sec. II-C, the tracklets which are temporally overlapping (even at different camera FoVs) may not be associated with one another. Hence, during runtime of the online NCR, tracklets having temporal overlaps were clustered into the same dummy group, with the pairwise similarity scores computed as described earlier in Sec. II-A1. This necessitated collection of a second similar FPV dataset, where the time information for each observation is available.

Database 2 for Online Re-identification: The FPV dataset for the online person re-id was collected in a large multi-storied office environment with 3 GGs. Videos of 14 persons were collected as they walked along the office corridors in unconstrained environment and were observed by the persons wearing the GGs, at different locations of the office complex. This resulted in around 4900 detected faces across the dataset. Time stamp for every frame is stored as metadata. Out of the 14 persons, 11 are males and 3 are females and 10 of them were wearing glasses (a challenging scenario for finding eyes). Some good sample images from the database are shown in Fig. 10. 9 out of these 14 targets were observed in all three cameras twice, whereas the rest were not observed in camera 3. This

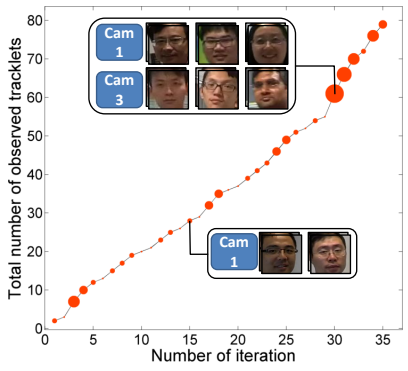
Fig. 10. Database 2 for online re-identification. Left, original image captured by Google Glass. Right four columns show 4 persons with 3 images each, captured using 3 different Google Glasses at different locations/views.

yielded a test set containing 84 tracklets, one for each tracklet in each camera FoV, which were to be associated with one another using online NCR based on their identities.

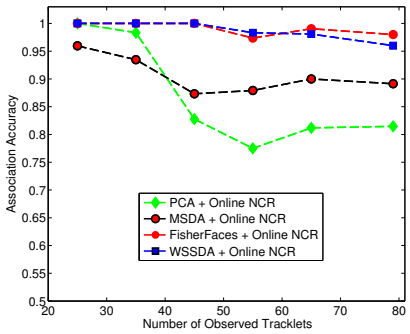
Like the batch problem, the training phase of the online re-id also has two main steps - 1. training the pairwise FI systems (such as MSDA, WSSDA etc.) and 2. estimating optimal k. As the database 2 is collected in the same office environment as database 1 and the same set of features from the batch re-id are used in online re-id experiments, we re-use the same training data in online re-id as well. Thus, like the batch version, the FI systems for the online re-id are identically trained on the FPV face image database in [5], and we use the same set of estimated k values for each of the four methods (as obtained in the batch method and described in Sec. III-C1).

As the time information for each tracklet is available, the tracklets are first clustered based on their temporal overlap into 35 groups. Tracklets in each group have temporal overlap (co-occuring) with one another and no two tracklets from two different groups have time overlap. These clusters of observations are further time ordered and at each iteration of the online NCR, one such cluster is introduced as input. It can be noted that each observation cluster may contain one or more tracklet(s). Fig. 11(a) shows how new observations are available at each iteration of the online NCR, and how the total number of observed/labeled tracklets evolve. Each bubble represents one cluster of tracklets, fed to the online NCR at each iteration and the radius of the bubble is proportional to the number of tracklets in the observed cluster. Note that the tracklets in each cluster belong to unique targets as they have temporal overlap with one another. As an example, the 30^{th} cluster has 6 unique temporally overlapping observed tracklets (3 in camera 1 and 3 in camera 3), as shown by target's face images associated to each tracklet.

At each iteration, the new tracklets are associated with the previously observed ones and labeled accordingly. The association accuracy at each iteration is estimated as $\frac{(\#\text{ true positive}+\#\text{ true negative})}{\#\text{ of unique faces in the testset}}$. The change in estimated accuracy with the increasing number of observations is plotted in Fig. 11(b) for all four methods, viz., PCA+NCR, FisherFaces+NCR, MSDA+NCR and WSSDA+NCR. As observed, both FisherFaces and WSSDA, when combined with NCR maintain very high association accuracy (more than 95%) even when majority of the tracklets are observed, with FisherFaces marginally outperforming WSSDA. PCA+NCR, on the other hand stabilizes to around 82% average accuracy after initial deterioration. MSDA, FisherFaces and WSSDA show very



(a)



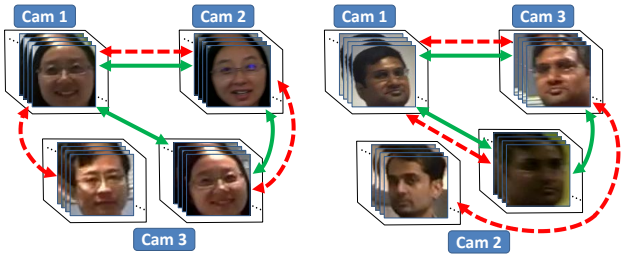
(b)

Fig. 11. Experiments and results on online person re-identification. (a) shows the increase in total number of observed tracklets (sets of faces for unique individuals) with iterations (analogous to time) in the online re-id setup. The bubbles represent clusters of temporally overlapping tracklets in each iteration (observational time window) and their radii are proportional to the number of tracklets in each of them. Face images associated to individual tracklets are shown for the 15^{th} and 30^{th} clusters. (b) shows the time evolution of re-identification accuracy for 4 methods - PCA+NCR, FisherFaces+NCR, MSDA+NCR and WSSDA+NCR. Consistently high accuracies are observed for both FisherFaces and WSSDA, with very slow rate of decrement for FisherFaces, MSDA and WSSDA as more observations are available.

slow decrement in accuracy as time goes by and more and more observations become available.

Two example cases are chosen from the experimental results to show how NCR can yield consistent re-id where the pairwise baseline methods fail. They are shown in Fig. 12. At first, re-identification is performed independently over each of the three pairs of cameras (GGs) using the WSSDA feature similarity computation. In Fig. 12(a), independent pairwise associations (red dashed lines) were correct between camera pairs 1-2 and 2-3. However, the incorrect associations between cameras 1-3 (red dashed line) make the association across the 3 cameras inconsistent. Similarly, in Fig. 12(b), incorrect pairwise re-identifications between targets across camera pair 2-3 make the overall results inconsistent. However, in both the cases, NCR enforces network consistency and makes the resultant data association results across the cameras correct (as shown using solid green arrows).

Comparison of average run-times: We have also compared the average run-times of the two global data association methods, viz., the batch NCR and the online NCR, by gradually increasing the number of unlabeled observations (and hence the number of variables to solve for) from 4 to 64. With increasing number of observations, the batch NCR needs to solve much



(a) (b)

Fig. 12. Example cases from the experimental results showing how the inconsistent associations are rectified. The red dashed lines denote re-identification performed independently over each of the three pairs of cameras (GGs) using the WSSDA feature similarity computation. Note that, incorrect associations between targets in camera pair 1-3 in (a) and between camera pair 2-3 in (b) render the overall association incorrect. The NCR algorithm, when applied over the same similarity scores generated by WSSDA, enforces the consistency requirements and makes the resultant associations across the cameras correct (as shown using green solid lines).

larger sized problems. The online NCR, on the other hand, has to run for more number of iterations, proportional to the number of observations, but solves a substantially small and about fixed size problem per iteration. As seen in Fig. 13, the batch NCR takes substantially longer time than the online NCR to solve the same association problem, especially as the number of observations increases. Moreover, unlike batch NCR, the online method is more memory efficient thereby making it an automatic choice for large scale re-id problems.

IV. Conclusions and Future Work

In this paper, the problem of re-identification from FPV videos collected using multiple wearable devices such as GGs is presented and discussed in detail. We present a framework for solving this re-identification problem by combining robust feature extraction methods for FPV face recognition with global data association techniques for network-consistent person re-identification (NCR). For real-life large scale person re-id scenarios where the objective is to identify targets shortly after they are observed, an online person re-id pipeline is also proposed that builds on the online implementation of the NCR algorithm.

For testing effectiveness of the proposed frameworks, we have collected two separate FPV databases - one each for the batch and online methods. The database 1 consists of FPV images of 72 targets collected using 4 GGs in a complex office environment. The database 2 (collected for online re-id) consists of continuous videos (including timestamps) for 14 targets captured using 3 GGs as they navigate through office corridors and are observed in the same camera FoVs twice. Analysis of the results indicates robustness of the method (both batch and online) in establishing consistency in association results as well as significant improvements in accuracy over the state-of-the-arts baseline methods across all camera pairs. Moreover, the online re-id method is also shown to be much faster and memory efficient, especially with large number of observations. The future work would include improvement of the method by incorporating other spaio-temporal motion constraints, development and utilization of novel facial features in

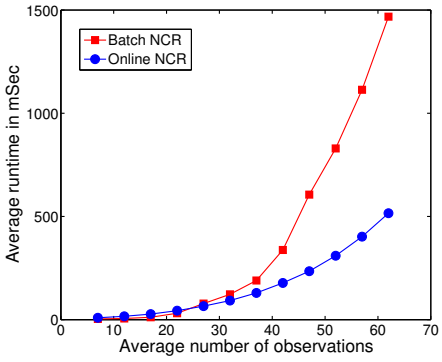


Fig. 13. A comparison between the change in average runtimes for batch and online NCR with increasing number of observations. The experiments were done on a desktop computer with dual core Intel i5 CPU (3.2GHz), 8 GB RAM and 64 bit Windows 7 operating system.

the present re-id framework, combining upper body features with facial features and real-time implementation and testing of the online NCR.

REFERENCES

[1] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification : What features are important ?" in *European Conference on Computer Vision, Workshops and Demonstrations*. Florence, Italy: Springer Berlin Heidelberg, 2012, pp. 391–401.

[2] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, Nov. 2013.

[3] Google, "Google glass," *http://www.google.com/glass/start/*, 2015.

[4] GoPro, *http://gopro.com/*, 2015.

[5] B. Mandal, S. Ching, L. Li, V. Chandrasekhar, C. Tan, and J.-H. Lim, "A wearable face recognition system on google glass for assisting social interactions," in 3^{rd} *International Workshop on Intelligent Mobile and Egocentric Vision, ACCV*, Singapore, Nov 2014, pp. 419–433.

[6] X. Wang, X. Zhao, V. Prakash, W. Shi, and O. Gnawali, "Computerized-eyewear based face recognition system for improving social lives of prosopagnosics," *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, pp. 77–80, 2013.

[7] Y. Utsumi, Y. Kato, K. Kunze, M. Iwamura, and K. Kise, "Who are you?: A wearable face recognition system to support human memory," in *ACM Proceedings of the 4th Augmented Human International Conference*, 2013, pp. 150–153.

[8] A. Das, A. Chakraborty, and A. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European Conference on Computer vision*, 2014.

[9] A. Chakraborty, A. Das, and A. Roy-Chowdhury, "Network consistent data association," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2015.

[10] A. Chakraborty, B. Mandal, and H. K. Galoogahi, "Person re-identification using multiple first-person-views on wearable devices," *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[11] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *European Conference on Computer Vision, Workshops and Demonstrations*, 2012, pp. 381–390.

[12] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *International Conference on Computer Vision and Pattern Recognition Workshops*. Providence, RI: IEEE, Jun. 2012, pp. 31–36.

[13] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[14] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *International Conference on Computer Vision and Pattern Recognition*, 2013.

[15] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869–878, 2015.

[16] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *ArXiv e-prints*, 2013.

[17] A. Alavi, Y. Yang, M. Harandi, and C. Sanderson, "Multi-shot person re-identification via relational stein divergence," in *Image Processing, IEEE International Conference on*, 2013.

[18] L. Yang and R. Jin, "Distance metric learning : A comprehensive survey," Michigan State University, Tech. Rep., 2006.

[19] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Asian conference on Computer vision*, 2010, pp. 501–512.

[20] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.

[21] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, Feb. 2008.

[22] F. Porikli and M. Hill, "Inter-camera color calibration using cross-correlation model function," in *IEEE International Conference on Image Processing (ICIP)*, 2003, pp. 133–136.

[23] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*, 2012, pp. 31–44.

[24] S. Pedagadi, J. Orwell, and S. Velastin, "Local fisher discriminant analysis for pedestrian re-identification," in *International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[25] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *European Conference Computer Vision*, 2006.

[26] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *British Machine Vision Conference*, Sep. 2008.

[27] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person reidentification with reference descriptor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 776–787, 2016.

[28] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, 2014, pp. 688–703.

[29] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[30] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *International Conference on Computer Vision*, 2015, pp. 4678–4686.

[31] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *International Conference on Computer Vision*, 2015, pp. 4516–4524.

[32] M. Grgic, K. Delac, and S. Grgic, "Scface - surveillance cameras face database,," *Multimedia Tools and Applications Journal*, vol. 51, no. 3, pp. 863–879, Feb 2011.

[33] P. Sinha, "Qualitative representations for recognition." Springer-Verlag, 2002, pp. 249–262.

[34] G. Tian, Y. Wong, B. Mandal, V. Chandrasekhar, and M. Kankanhalli, "Multi-sensor self-quantification of presentations," in *ACM Multimedia (ACMMM)*, Brisbane, Australia, Oct 2015, pp. 601–610.

[35] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs," *Journal of Experimental Psychology: General*, vol. 4, no. 142, pp. 1323–1334, 2013.

[36] A. D. Molino, B. Mandal, L. Li, and J.-H. Lim, "Organizing and retrieving episodic memories from first person view," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Torino, Italy, Jul 2015, pp. 1–6.

[37] B. Mandal, W. Zhikai, L. Li, and A. Kassim, "Performance evaluation of local descriptors and distance measures on benchmarks and first-person-view videos for face identification," *Journal of Neurocomputing*, vol. 184, pp. 107–116, 2016.

[38] S. Ching, B. Mandal, Q. Xu, L. Liyuan, and J.-H. Lim, "Enhancing social interaction with seamless face recognition on google glass: Leveraging opportunistic multi-tasking on smart phones," in 17^{th} *International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileCHI)*, Copenhagen, Denmark, Aug 2015, pp. 750–757.

[39] K. Shafique and M. Shah, "A noniterative greedy algorithm for multiframe point correspondence," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 1, pp. 51–65, 2005.

[40] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1806–1819, 2011.

[41] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Multi-commodity network flow for tracking multiple people," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1614–1627, 2014.

[42] H. Ben-Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *International Conference on Computer Vision*, 2011, pp. 137–144.

[43] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2470–2477.

[44] X. Wang, E. Turetken, F. Fleuret, and P. Fua, "Tracking interacting objects using intertwined flows," *Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[45] S. Avidan, Y. Moses, and Y. Moses, "Centralized and distributed multi-view correspondence," *International Journal of Computer Vision*, vol. 71, no. 1, pp. 49–69, 2007.

[46] (2015) Open source computer vision,(http://opencv.org/). [Online]. Available: http://opencv.org/

[47] X. Yu, W. Han, L. Li, J. Shi, and G. Wang, "An eye detection and localization system for natural human and robot interaction without face detection," *TAROS*, pp. 54–65, 2011.

[48] B. Mandal, L. Li, V. Chandrasekhar, and J. H. Lim, "Whole space subclass discriminant analysis for face recognition," in *International Conference on Image Processing (ICIP)*, 2015, pp. 329–333.

[49] W. Liu, Y. Wang, S. Z. Li, and T. N. Tan, "Null space approach of fisher discriminant analysis for face recognition," in *ECCV*, 2004, pp. 32–44.

[50] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, January 2005.

[51] B. Mandal, X. D. Jiang, and A. Kot, "Dimensionality reduction in subspace face recognition," in *IEEE* 6^{th} *International Conference on Information, Communications and Signal Processing (ICICS)*, Singapore, Dec 2007, pp. 1–5.

[52] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, 2006.

[53] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Mixture subclass discriminant analysis," *Signal Processing Letters*, vol. 18, no. 5, pp. 319–322, 2011.

[54] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Mixture subclass discriminant analysis link to restricted gaussian model and other generalizations," *Transactions on Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 8–21, 2013.

[55] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, Mar 2008.

[56] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained video with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2011, pp. 529–534.

[57] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, January 1991.

[58] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, August 1996.

[59] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, Columbus, OH, Jun 2014, pp. 1701–1708.

[60] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[61] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD*, vol. 96, no. 34, 1996, pp. 226–231.