Person Re-identification Using Multiple First-Person-Views on Wearable Devices

Anirban Chakraborty (NTU, Singapore), Bappaditya Mandal (I²R, A*STAR, Singapore), Hamed Kiani Galoogahi (IIT, Italy)

NANYANG TECHNOLOGICAL UNIVERSITY

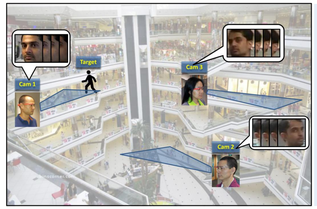
Institute for Infocomm Research A*STAR

WACV Lake Placid 2016

IEEE 2016 Winter conference on Applications of Computer Vision

Introduction

Problem: To re-identify and track one or more targets over multiple non-overlapping camera field-of-views in a crowded environment.



Challenges:

- ❖ Low resolution images of targets.
- ❖ Partially/fully occluded targets.
- ❖ Wide variation in viewpoints and hence, pose of targets.
- ❖ Variation in illumination across multiple cameras.

Unavailability of robust, discriminative features.

- ✓ With wearable cameras (Google Glasses™, GoPro™), it is possible to capture unoccluded, high resolution face shots of targets.
- ✓ In this work, we present a person re-identification framework designed for egocentric/ first-person-view videos obtained from a network of wearable devices.

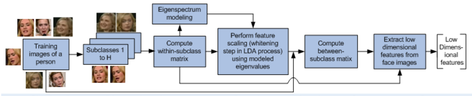
Contributions

❖ We propose a re-identification pipeline that 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 scores between all pairs of observations in each device pair.

2. The computed scores are used as inputs to a global data-association method to estimate 'network consistent' final association labels between pairs of observations across any two devices.

Pairwise feature similarity computation



❖ Images containing target faces are normalized, each class (target) is divided into subclasses using spatial partition trees.

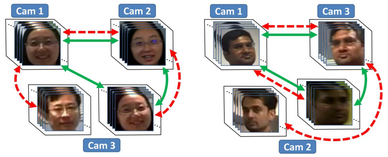
❖ Subclass discriminant analysis and eigenfeatures feature regularization methodology is used to handle the problem of modelling large variances in within-class face images.

❖ Using the regularized features, total-subclass and between subclass scatter matrices are computed.

❖ Incoming face image vector is converted into a feature vector using the learned transformation matrix. Cosine distance measure with 1-NN is used as the classifier to generate pairwise similarity scores.

[1]A.Das and A.Chakraborty et. al., Consistent Re-identification in a Camera Network, ECCV, 2014.
Download NCR/NCDA code from: https://github.com/dasabir/NCR_Code

Network consistent re-identification (NCR)[1]



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

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

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

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

$$\forall\, i = [1,\cdots n_p],\ j = [1,\cdots n_q]\, \forall\, p,q = [1,\cdots m],\ \text{and } p<q$$

$$\forall \text{ paths } e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$$

$$x_{i,j}^{p,q} \in \{0,1\} \ \forall i,j,p,q$$

❖ On a network of cameras, both direct and indirect paths of association exist.

❖ Associations via different paths must match → *Network Consistency*.

❖ Enforced using a global data association (optimization) method.

❖ Input: Feature similarity score $c_{i,j}^{p,q}$.

❖ Output: Association labels $x_{i,j}^{p,q} \in \{0,1\}$.

❖ Utility function rewards true-positives and penalizes false-positives (via k).

❖ Pairwise association constraints: A target from one camera FoV may have **at most** one match in another camera.

❖ Works even when → 'Not all targets are observed in all the cameras'.

❖ Consistency constraints: Direct and all indirect paths of associations point to the same association between two obs.

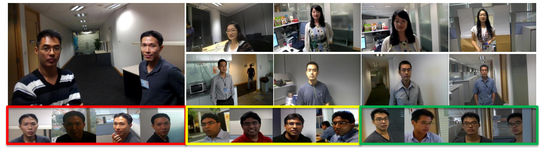
❖ For a triplet of cameras/devices, each constraint simplifies to,

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1$$

$$\forall\, i,j,k = [1,\cdots n], \forall\, p,q,r = [1,\cdots m],\ \text{and } p<r<q$$

Solved as **binary integer linear program**.

Realistic first person view dataset



❖ A large multi-storied office environment.

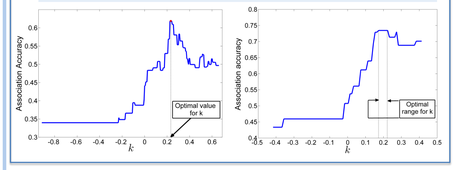
❖ 4 Google Glasses.

❖ 72 total targets (37 males, 35 females).

- ✓ Variable number of observed targets (52, 40, 43, 50) in each Google Glass field-of-view.
- ✓ Motion blur in zoomed-in face images because of rapid head movement of the observers.
- ✓ Face and eye detectors are employed to filter blurry, occluded and low resolution images.

Experiments and results

Estimation of k



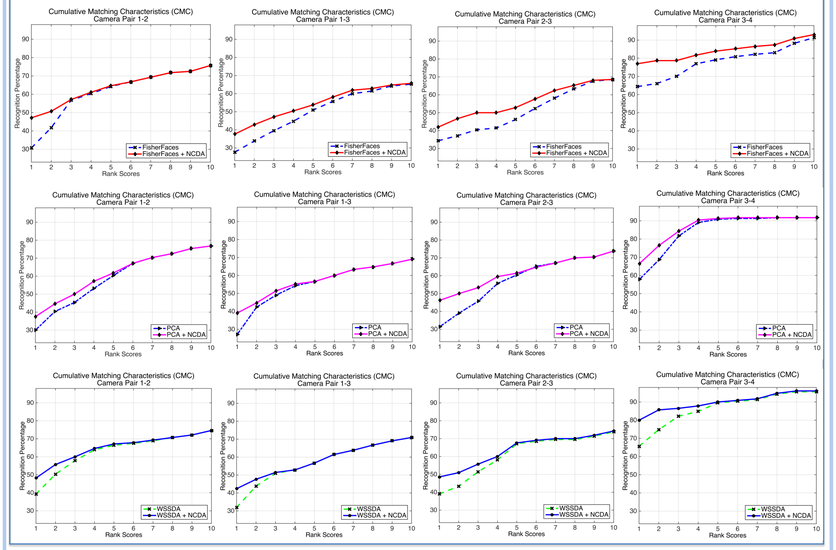
❖ Baseline methods PCA, Fisherfaces, WSSDA.

❖ Compared against [Baseline + NCR].

❖ One third targets used to train NCR.

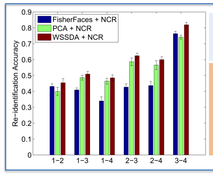
❖ Results averaged over 10 random partitions.

Cumulative Matching Characteristics - Baseline *vs.* [Baseline + NCR]



nAUC Values (Upto rank 10)

Cam pair	PCA	FF	WSSDA	PCA + NCR	FF + NCR	WSSDA + NCR
1-2	0.5978	0.6187	0.6387	0.6179	0.6393	**0.6544**
1-3	0.5614	0.5077	0.5741	0.5743	0.5484	**0.5847**
1-4	0.5349	0.5183	0.6508	0.5521	0.5410	**0.6717**
2-3	0.5849	0.5090	0.6172	0.6185	0.5646	**0.6407**
2-4	0.6455	0.5571	0.6717	0.6513	0.5817	**0.6950**
3-4	0.8570	0.7826	0.8708	0.8763	0.8423	**0.9017**



Overall accuracy (TP+TN)

[WSSDA+NCR] is the best performing face re-id method overall.

The authors acknowledge the help of I²R, Singapore staff researchers in collection of the FPV video datastet used here.