Groups Re-identification with Temporal Context

Michal Koperski
INRIA
2004 Route Des Lucioles
Sophia Antipolis, France 06902
michal.koperski@inria.fr

Slawomir Bak
Disney Research
4720 Forbes Avenue
Pittsburgh, PA 15213
slawomir.bak@disneyresearch.com

Peter Carr
Disney Research
4720 Forbes Avenue
Pittsburgh, PA 15213
peter.carr@disneyresearch.com

ABSTRACT

Re-identification methods often require well aligned, unoccluded detections of an entire subject. Such assumptions are impractical in real world scenarios, where people tend to form groups. To circumvent poor detection performance caused by occlusions, we use fixed regions of interest and employ codebook-based visual representations. We account for illumination variations between cameras using a coupled clustering method that learns per-camera codebooks with entries that correspond across cameras. Because predictable movement patterns exist in many scenarios, we also incorporate temporal context to improve re-identification performance. This includes learning expected travel times directly from data and using mutual exclusion constraints to encourage solutions that maintain temporal ordering. Our experiments illustrate the merits of the proposed approach in challenging re-identification scenarios including crowded public spaces.

CCS CONCEPTS

• **Computing methodologies** → *Object recognition*; *Object identification*;

KEYWORDS

Computer Vision;People Re-Identification

ACM Reference format:
Michal Koperski, Slawomir Bak, and Peter Carr. 2017. Groups Re-identification with Temporal Context. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 9 pages.
https://doi.org/http://dx.doi.org/10.1145/3078971.3078978

1 INTRODUCTION

Re-identification is the task of finding the same individual across a network of cameras. This problem is very challenging due to significant changes in appearance caused by variations in illumination, viewing angle and a person's pose. In this paper, we focus on re-identification in *crowded environments* with predictable movement patterns. For instance, people tend to maintain their ordering in queues, whether it be checking out at a grocery store or passing through airport security. Such scenarios are popular for re-identification (and extremely challenging), but have not been studied extensively.

Most re-identification approaches assume reliable detections [1, 7, 10, 19, 23, 32], which are not possible in crowd environments. Only a few methods have targeted more challenging scenarios where detections are less reliable and partially occluded [5, 14, 44]. To mitigate the impact of occlusions and poor detection alignment, we focus on re-identifying groups at *pinch points* like queues and doorways.

Instead of running an object detector, we define a fixed region of interest (ROI) on the image plane (see Fig. 1). We propose to represent the visual content within an ROI using a codebook. Codebooks [8, 34, 35] have shown high efficacy in scenarios where specialized object detectors fail. However, standard codebook learning approaches are inefficient in multi-camera environments when there are significant appearance changes between cameras. As a result, we propose a new coupled clustering method that generates per-camera codebooks with entries that correspond across cameras.

A codebook encoding of an ROI is usually not as distinctive as a full body-based descriptor (which can leverage the spatial locations of visual features). Therefore, we enhance our representation by incorporating temporal information. Predictable movement patterns exist in many scenarios, and these can be used to disambiguate people with otherwise similar appearances. In airports, for example, people tend to move from ticketing through security and then to their boarding gates (see Fig. 1). Our main contributions are:

- We propose a new coupled clustering method that learns codebooks for each camera pair with codewords corresponding across cameras. This copes with significant illumination changes of visual appearance between cameras.
- We integrate a temporal model into our matching strategy. The temporal model is jointly optimized using both visual appearance and temporal information. Initial parameters of the temporal model are first estimated using the visual appearance, and then they are used as a feedback to enhance the re-identification. Coherent re-identification matches in return provide better estimation of temporal parameters. By iterating, the method quickly converges to the optimal configuration.
- Because existing re-identification datasets have no temporal information, we collected a new **QUEUE** dataset that simulates a queue scenario in which people move from one location to another with significant variations in illumination. Recorded sequences come from non-overlapping cameras. We also modified existing re-identification datasets to simulate temporal information by adding synthetic timestamps.

ICMR '17, June 06-09, 2017, Bucharest, Romania

ACM ISBN 978-1-4503-4701-3/17/06…$15.00
https://doi.org/http://dx.doi.org/10.1145/3078971.3078978

Figure 1: Predictable movement patterns exist in many scenarios. We model the travel time between cameras to help disambiguate people with similar appearances. Because full body detectors are unreliable in crowded scenarios, we use a fixed region of interest (ROI) on the image plane and represent its visual appearance using a set of learned camera-specific codebooks that have corresponding entries (camera-specific codewords) across cameras. Learning corresponding entries facilitates matching under different imaging (illumination) conditions.

Our experiments illustrate the merits of joint optimization and achieve new state-of-the-art performance on multiple datasets outperforming existing approaches that only considered visual appearance models.

2 RELATED WORK

Re-identification remains an unsolved problem due to large intra-class and inter-class variations caused by changes in lighting, viewing angle and a person's pose. Most effective approaches learn robust metrics for matching [1, 6, 19, 23, 32, 33]. These methods learn a distance function among features from different cameras such that relevant dimensions are emphasized while irrelevant ones are discarded. Explicitly or implicitly, all of these approaches assume matching of full body appearance with negligible amounts of occlusion and cannot operate on images taken in crowded public spaces.

The recent work of [44] introduces *partial person re-identification* to address retrieval in more realistic scenarios where only a partial observation of a person is available for matching. It combines local patch level matching based on sparse coding with global, part-based matching that exploits spatial layout information.

Visual information from surrounding people has also been used to reduce ambiguity in person identification. In [5, 14, 43] we can find that group association between two or more people can give valuable information about the identity of an individual. [16, 21] employ dictionary learning to cope with detection misalignment. The bounding box images are divided into a set of stripes and per-stripe dictionaries are learned assuming well aligned person detections. [17, 18, 28] search for a set of linear transformations, which transform features from different viewpoints to the common embedding space. These methods implicitly assume spatial correspondence between features extracted from both viewpoints (otherwise it would not be clear which transformation should be applied). Thus the mentioned methods are more suitable for recognition of rigid objects. In contrast, our approach does not depend on spatial information. It is mostly invariant to camera viewpoint changes and easily incorporates multiple people into the appearance representation (see Fig. 1).

Although temporal cues are rarely used in recent re-identification methods[1], they were exploited in early person re-identification and multi-target tracking methods to infer camera topology which enabled tracking across multiple non-overlapping views [15, 26, 29]. These approaches usually exploit space-time cues to learn inter-camera relationships that are then used to constrain the re-identification search space. Camera topology together with temporal information can reduce the number of paths people might take, which constrains potential re-identification matches [31, 36]. Our approach is similar in spirit to these methods but introduces a new formulation that jointly learns models of both visual appearance and movement patterns.

Global motion patterns are learned in [26] for activity modeling. Activities were detected using the density of moving pixels and omitted visual appearance representations. Cross Canonical Correlation Analysis is introduced in [24, 25] to correlate activities across cameras with disjoint FOVs to infer the topology of a camera network. The FOVs of the cameras are segmented into regions within which the activity patterns were similar. Affinity matrices are employed to infer spatio-temporal camera topologies. [9] propose to use cameras topology to improve the global re-identification consistency. Unlike ours, it requires head, torso and leg detectors and it becomes effective in camera networks with at least 3 cameras.

A comprehensive review of camera topology estimation methods is presented in [40]. Similarly, temporal information has been used to constrain the re-identification search space [31, 36]. Re-identification that employs time-ordering constraints is widely applied in Intelligent Transportation Systems (ITS). [13] propose a method for car re-identification based on simple visual appearance model and the expected travel time between cameras. Simple appearance model makes it unsuitable for person re-identification and their temporal model must be initialized using the ground truth. The visual appearance of a car and an inductive loop (extracting speed and induction response amplitude) are used together with time information to predict when the vehicle would re-appear in the second camera [39]. Time ordering constraints are integrated using manually computed temporal windows which controlled feasibility.

[1]This phenomenon may be a result of the available re-identification datasets, which are often only a collection of bounding boxes extracted from disjoint camera views with no temporal information.

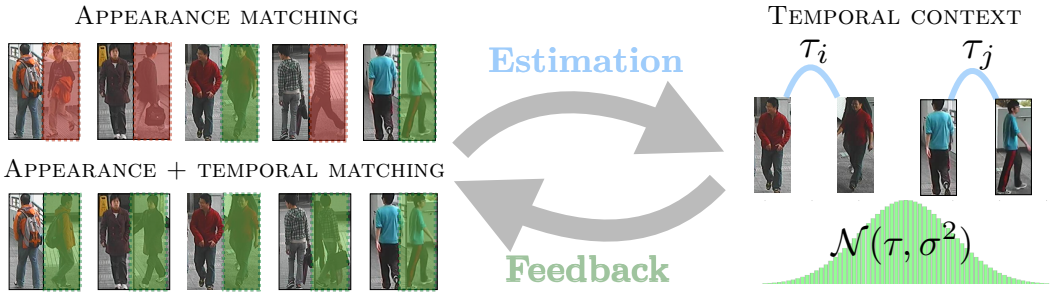

Figure 2: Learning and Leveraging Temporal Context. Actual travel times between cameras are estimated from reliable re-identification matches using only appearance information. Model parameters are then fit to these data samples. Re-identification with temporal context is then used to disambiguate detections with less distinctive visual appearances.

In these approaches, the visual appearance and temporal cues are often modeled independently. Our approach, on the other hand, jointly models visual appearance and temporal context using a feedback loop (see Fig. 2).

Re-identification usually treats every test case (query) as an independent retrieval task. In contrast, multi-target tracking algorithms [3, 41] typically conduct multiple queries simultaneously by posing a linear assignment problem [12, 41]. Prior work [37] extended this idea and proposed to compute m-suboptimal solutions to find consensus assignment. In this paper, we follow a similar idea and incorporate temporal context with multiple queries to simultaneously search for an optimal solution. This enforces mutual exclusion constraints, improving the re-identification accuracy.

3 METHOD

We define a region of interest (ROI) as rectangular bounding box containing sub-image for each camera. In contrast to the common re-identification problem where every ROI bounding box image contains a well detected individual, we consider crowded scenarios where ROI might contain multiple people, occluding each other. As object detectors might be infeasible in such scenario due to large and frequent occlusions, ROI is assumed to be fixed. However, as "pinch points" (doorways, gates, etc.) exist in many scenarios, these are our natural choices for ROIs.

Let A and B denote two cameras in a network. Each camera monitors a local region that people pass through, such as a gate or doorway (see Fig. 1). Suppose there are two ROI bounding boxes $\mathbf{b}_i$ and $\mathbf{b}_j$ captured from cameras A and B respectively at times t_i and t_j, and that appearance descriptors $\mathbf{x}_i$ and $\mathbf{x}_j$ are extracted from the visual content of each ROI bounding box.

We refer to triplets $\mathbf{o}_k = (\mathbf{b}_k, t_k, \mathbf{x}_k)$ as observations. From a Bayesian point of view, re-identification is the likelihood that observations $\mathbf{o}_i$ and $\mathbf{o}_j$ are different images of the same person. Most re-identification methods ignore the spatiotemporal information and only consider the similarity of the extracted appearance descriptors

$$\mathrm{P}_{\mathrm{ReID}}(\mathbf{o}_i, \mathbf{o}_j | i \equiv j) \propto \mathrm{P}_{\mathrm{App}}(\mathbf{x}_i, \mathbf{x}_j | i \equiv j). \tag{1}$$

In this work, we incorporate the expected movement patterns as well. Because we have restricted our camera views to each focus on a single localized region, the extracted bounding boxes locations are effectively constant (see Fig. 1). As a result, we formulate the likelihood as a product of appearance similarity and expected travel time between the two locations[2]

$$\mathrm{P}_{\mathrm{ReID}}(\mathbf{o}_i, \mathbf{o}_j) \propto \mathrm{P}_{\mathrm{App}}(\mathbf{x}_i, \mathbf{x}_j)\mathrm{P}_{\mathrm{Time}}(t_i, t_j), \tag{2}$$

where the conditional dependency $i \equiv j$ has been omitted for notational convenience. Re-identification is often used to match observations between cameras: given an observation $\mathbf{o}_j$ from camera B, the most likely observation of the same individual from camera A is

$$\mathbf{o}_i^* = \arg\max_i \mathrm{P}_{ReID}(\mathbf{o}_i, \mathbf{o}_j) \tag{3}$$

$$= \arg\min_i d^2_{ReID}(\mathbf{o}_i, \mathbf{o}_j) \tag{4}$$

$$= \arg\min_i \left(\lambda d^2_{App}(\mathbf{x}_i, \mathbf{x}_j) + (1-\lambda) d^2_{Time}(t_i, t_j)\right), \tag{5}$$

where $\lambda \in [0, 1]$ is a parameter that controls the relative influence of similar visual appearance and expected travel time. We will now describe the distance functions that reflect our appearance (d^2_{App}) and temporal models (d^2_{Time}).

3.1 Appearance Model

Crowded scenarios produce detections that are poorly aligned and/or partially occluded. To cope with these difficulties, we propose a visual appearance descriptor $\mathbf{x}_i$ based on a codebook approach.

Codebooks [8, 34, 35] have been used previously in re-identification [27, 42], but only in situations with well-aligned bounding boxes. Different codebooks were learned for different regions of the bounding box by dividing it into a fixed layout of rectangular patches (horizontal stripes or a dense grid). Although spatial structure is beneficial for recognition, it requires unoccluded, reliably aligned detections, which is impractical in crowded scenarios. Instead we learn a single codebook for the entire *ROI*. To achieve that we

[2] Our methodology easily extends to a model which does not require pre-defined fixed locations.

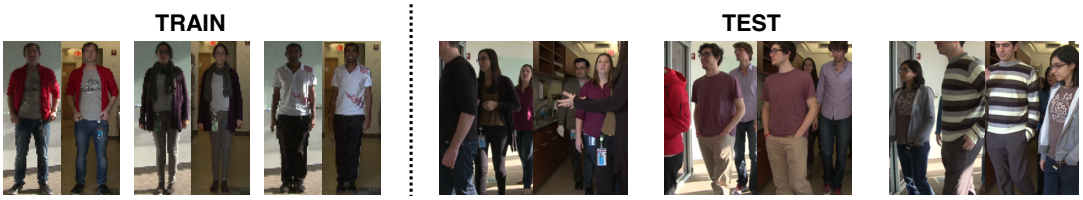

Figure 3: Sample images from the QUEUE dataset. The left column illustrates image pairs from QUEUE01 that were used for training codebooks; the right column illustrates image pairs from QUEUE02 simulating the queue scenario.

propose to densely sample random patches from *ROI*. And then compute codebook based on descriptors extracted from patches.

Two popular choices for converting visual features into codewords are *Bag-of-words* (BoW) [8] and *Fisher Vector* (FV) [34, 35]. Let $\mathcal{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_F\}$ be a set of image features extracted from training data. Image features are usually high-dimensional vectors (*e.g.* SIFT descriptors concatenated with color histograms) and are typically clustered into codewords. For example, $BoW = \{\mu_k : k = 1, \ldots, K\}$ uses k-means clustering to define codewords as the centers of the learned clusters. The number of the clusters K is the codebook size. BoW maps each image feature $\mathbf{f}_f$ to a codeword k using nearest neighbour lookup. The descriptor is a K-dimensional histogram of codeword frequencies.

FV is an extension of BoW. The distribution of features is modeled by a Gaussian Mixture Model (GMM). Let $FV = \{\mu_k, \Sigma_k, w_k : k = 1, \ldots, K\}$ be the parameters of a GMM fitting the distribution of image features, where w_k is the weight of the k^{th} Gaussian with mean μ_k and covariance Σ_k. An image feature $\mathbf{f}_f$ is represented by the gradients of the likelihood of this feature being generated by a certain Gaussian. The gradients are computed over both μ and Σ variables. The resulting descriptor is a $2K$-dimensional vector (see [34, 35] for details).

Both BoW and FV can be learned using features from different cameras (*e.g.* from camera A and camera B) by combining them into a single set $\mathcal{F}^{A,B} = \{\mathbf{f}_1^A, \ldots, \mathbf{f}_{F^A}^A, \mathbf{f}_1^B, \ldots, \mathbf{f}_{F^B}^B\}$ and then performing clustering on $\mathcal{F}^{A,B}$ [27]. If there are strong appearance changes across the two cameras, a common mapping from image features to codewords might not be effective. Instead, we modify the underlying clustering method (k-means for BoW and GMM for FV) such that corresponding features will map to the same cluster even though they have different appearances due to illumination changes.

Coupled clustering. We propose to train codebooks on features extracted from corresponding images which are well aligned between cameras (see Train images from Fig. 3). Note that we only require correspondences when learning a codebook. At test time, our method does not require well-aligned detections.

Given corresponding features $\mathbf{f}^A$ and $\mathbf{f}^B$, we build $\mathcal{F}^{A|B} = \{(\mathbf{f}^A|\mathbf{f}^B)_1, \ldots, (\mathbf{f}^A|\mathbf{f}^B)_{F^{A|B}}\}$, which is a set of concatenated corresponding image features from camera A and camera B. On such feature set we perform either k-means (in case of BoW) or GMM (in case of FV) clustering to obtain model parameters. For example, in case of k-means, we divide the coupled codebook $BoW^{A,B} = \{\mu_{A|B} : k = 1, \ldots, K\}$ into two codebooks $BoW^A = \{\mu_k^A : k = 1, \ldots, K\}$ and $BoW^B = \{\mu_k^B : k = 1, \ldots, K\}$ by extracting the first and second halves of the cluster center dimensions respectively. FV can be split analogously obtaining two GMM models: $FV^A = \{\mu_k^A, \Sigma_k^A, w_k : k = 1, \ldots, K\}$ and $FV^B = \{\mu_k^B, \Sigma_k^B, w_k : k = 1, \ldots, K\}$, where w_k are shared across the models. The appearance transfer function between cameras is learned implicitly due to the proposed coupled clustering method.

The concatenation step is crucial because it guides a clustering algorithm to find clusters that represent similar visual features from both cameras. In addition, it ensures codeword correspondence across cameras which is crucial for computing similarity between images (see Eq. (6)). Without the concatenation step, a clustering algorithm would find clusters based only on appearance features and would ignore which camera they were extracted from. Object appearance is usually more similar in multiple images from the same camera compared to images taken from two disjoint cameras. In standard clustering where the source of the image is ignored, the clusters may contain patches from only one camera.

In test stage, to compare two *ROIs* from different cameras, we densely sample random patches from *ROI* and based on them we compute codebook representation. We follow the same steps for a second *ROI*. Thanks to the proposed coupled clustering, the codebook representation of the same group of people across cameras, is much more consistent than representation based on standard codebook. Moreover, because we pick random patches from *ROI*, our method is much more robust in terms of misalignment and occlusions.

Appearance Dissimilarity Measure. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be extracted appearance representations (in our case, $\mathbf{x}_i$ and $\mathbf{x}_j$ are histograms (BoW) or *Fisher Vectors*). We compute the visual appearance dissimilarity using ℓ_2 norm:

$$d^2_{\text{App}}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2. \tag{6}$$

3.2 Temporal Model and Joint Optimization

For simplicity, we assume that the amount of time a subject takes to travel between cameras A and B can be modeled as a normal distribution $\mathcal{N}(\tau, \sigma^2)$. Given the parameters ($\tau$,$\sigma^2$) of the distribution and the times t_i and t_j of the two bounding boxes, the **Time Context Dissimilarity Measure** between the actual and expected elapsed time is

$$d^2_{\text{Time}}(t_i, t_j) = \frac{(t_j - t_i - \tau)^2}{2\sigma^2}. \tag{7}$$

Parameter Estimation. In practice, τ and σ are estimates of an unknown distribution $\mathcal{N}(\tau^*, \sigma^{*2})$. The parameters of the distribution could be estimated heuristically using the distribution of typical walking speeds if the path distance between cameras A and B is known. When that information is not available, we propose the following solution. Given an appearance dissimilarity function $d_{\text{App}}(\mathbf{x}_i, \mathbf{x}_j)$, we construct a binary threshold classifier $h(\mathbf{x}_i, \mathbf{x}_j) \mapsto \{0, 1\}$ to predict whether observations $\mathbf{o}_i$ and $\mathbf{o}_j$ contain indeed the same subject. We select the threshold ξ that achieves a 95% confidence level on the training data and use only these observation pairs $(\mathbf{o}_i, \mathbf{o}_j)$ — *e.g.* people with distinctive appearances that can be re-identified reliably — to estimate the parameters (τ, σ^2) of the normal distribution (see Algorithm 1 steps 1 - 3).

Threshold ξ is learned on the training set, but parameters τ and σ are estimated at test time. Image pairs with distances below threshold ξ can be false positives. We exploit the fact that the variance of timestamp differences $t_{ij} = t_i - t_j$ of the true positives is low, while t_{ij} values of false matches form outliers. Based on this fact, we use Minimum Covariance Determinant [4, 38] (MCD) to estimate τ and σ parameters. The MCD method was designed to estimate unimodal Gaussian parameters from data containing outliers which is exactly our case.

Algorithm 1 Parameter Estimation algorithm

1: For each test query o_j compute:
   $o_i^* \leftarrow \arg\min_i d^2_{App}(x_i, x_j)$
2: Construct $\Delta T = \{t_{ij}\}$: $d_{App}(x_i, x_j) < \xi$
3: Compute initial estimates $\tau, \sigma \leftarrow MCD(\Delta T)$
4: **loop**
5: For each test query o_j compute:
   $o_i^* \leftarrow \arg\min_i d^2(o_i, o_j)$ // see Eqs. 4-5
6: Reconstruct ΔT using all (o_i^*, o_j)
7: Update estimates $\tau, \sigma \leftarrow MCD(\Delta T)$
8: **end loop**

To improve the estimations of τ^* and σ^*, we propose to **jointly optimize** the parameters τ^* and σ^* based on information from re-identification stage and the current temporal model parameters (see Algorithm 1 steps 4 - 8). After the initial estimation of τ^* and σ^* based on the appearance-only model, we propose to re-estimate τ and σ. We do it based on current re-identification matches, which are obtained from our model using jointly appearance and temporal context (see Eq. (5)). Because the re-identification accuracy of the joint appearance and temporal context model is higher than the appearance only model, we can re-estimate (jointly optimize) τ and σ by taking into account timestamps of current matches. Better temporal context enables matching individuals with less discriminative visual appearance (e.g. common clothing colors such as black winter coats). These new matches are then used to refine the current temporal context model.

The Univariate Gaussian could be replaced with any density estimation technique. With sufficient data, we can fit more complex models (*e.g.* GMM). However as time information is not available in re-id benchmarks and our QUEUE dataset is relatively small, a single Gaussian distribution is effective.

3.3 Ordering Preference via Mutual Exclusion

In many re-identification experiments, each test query $\mathbf{o}_i$ is evaluated independently

$$\mathbf{o}_i^* = \arg\min_i d^2(\mathbf{o}_i, \mathbf{o}_j). \tag{4}$$

If multiple queries are conducted, it is entirely possible that different queries $\mathbf{o}_j$ and $\mathbf{o}_k$ from camera B will have the same best match $\mathbf{o}_i$ in camera A. By conducting multiple queries simultaneously, we can enforce mutual exclusion constraints and ensure that each query has a unique match. In scenarios where people are observed while passing through confined regions, such as a gates or doorways, and there is an expected path between cameras, the order of people observed in A should be similar to the order of people observed in B. By incorporating temporal context into the cost function and enforcing mutual exclusion constraints, we can implicitly encourage a preference for preserved ordering to improve the recognition of observations that do not have distinctive appearance descriptors. Given N observations from camera A and M observations from camera B, we formulate simultaneous re-identification queries as a linear assignment problem

$$\{\mathbf{o}_i^*\} = \arg\min_\pi \left(\sum_{j=1}^{M} d^2(\mathbf{o}_{\pi(j)}, \mathbf{o}_j) \right), \tag{8}$$

where π is M-vector mapping observations from camera B to observations in camera A. We encode the linear assignment problem into an $M \times N$ cost matrix $\mathbf{C} = \{c_{ij}\}_{M \times N}$ where $c_{ij} = d^2(\mathbf{o}_j, \mathbf{o}_i)$ and determine the optimal assignment using the Kuhn-Munkres (Hungarian) algorithm [20]. Our experiments evaluate the merits of conducting queries independently versus simultaneously with a preference for preserving ordering.

For clarity, we have described the simple case where every observation in B has a corresponding observation in A (another typical bias in most re-identification datasets). For more practical scenarios, our linear assignment formulation can be encoded in an augmented matrix to include 'no match' conditions as well [12].

4 EXPERIMENTS

We carried out experiments on four challenging datasets: **VIPeR** [11], **CUHK01** [22], **iLIDS-groups** [43], and our **QUEUE** dataset. The results are analyzed in terms of recognition rate using *rank-1 accuracy* as well as the *cumulative matching characteristic* (CMC) [11] curve. The CMC curve represents the expectation of finding the correct match in the top r matches. The curve can be characterized by a scalar value computed by normalizing the area under the curve referred to as *nAUC* value.

4.1 Datasets

VIPeR [11] is one of the most popular person re-identification datasets. It contains 632 image pairs of pedestrians captured by two outdoor cameras. **VIPeR** images contain large variations in lighting conditions, background, viewpoint, and image quality. Each bounding box is cropped and scaled to be 128×48 pixels.

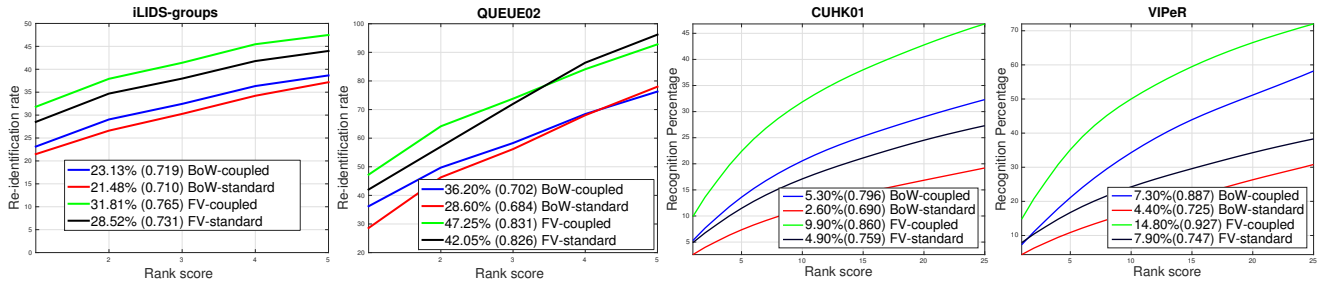

Figure 4: Performance comparison of different codebook models on iLIDS-groups, QUEUE02, CUHK01 and VIPeR. *Rank-1* identification rates as well as *nAUC* values (provided in brackets) are shown in the legend next to the method name. Coupled clustering significantly outperforms standard approaches for both BoW and FV model.

CUHK01 [22] contains 971 people captured with two cameras. For each person, 2 images from each camera are available. The first camera captures the side view of a pedestrian and the second captures a frontal or rear view. Each bounding box is scaled to be 160×60 pixels.

iLIDS-groups [43] contains 64 groups captured with two cameras in the airport. In most cases 4 images of each group are available. Thus the dataset contains 274 images of cropped groups. This dataset along with standard challenges like illumination changes across cameras, introduces some other challenges including occlusions because people in group tend to occlude each other. Additionally dataset was recorded at an airport where people are often occluded by luggage. Moreover people in groups very often change their relative positions across the cameras. We mentioned in Section 3.1 that we need patch correspondence between cameras to train the proposed coupled clustering codebook. Because iLIDS-groups is very noisy and contains multiple people, it is difficult to extract corresponding patches automatically. In the result we train a codebook using iLIDS-MA [2] dataset, which contains 3680 manually cropped images of 40 individuals, well aligned and acquired by the same camera pair.

QUEUE is our new dataset that contains two scenarios. The first **QUEUE01** consists of 23 individuals with 3379 images of people registered by two cameras in significantly different lighting conditions (see Fig. 3). This data was used for training the codebook. In the second scenario (**QUEUE02**), 23 individuals from **QUEUE01** were asked to move from the the one location to another, simulating a queue scenario (see Fig. 3). We manually annotated 15 groups; each group is described by unique group id and frame time-stamp from the video stream. As can be seen in Fig. 3 the queue is dense so some individual belongs to more than one group. To our best knowledge the **QUEUE02** is the only group re-identification dataset which provides timestamp information. **QUEUE02** is the only dataset which contains both: people groups and timestamps. We chose **iLIDS-groups**, because it contains groups of people. **VIPeR** and **CUHK01**, though they do not contain neither timestamp information nor groups of people were selected to show relative improvement of the proposed coupled clustering method compared with standard Bag of Words approaches.

Dataset	Codebook size				
	16	**20**	**64**	**128**	**256**
iLIDS-groups	29.3	31.8	29.2	29.0	-
QUEUE02	47.2	44.2	46.2	41.9	-
CUHK01	8.7	8.6	9.6	9.9	9.0
VIPeR	11.4	11.6	12.7	14.8	12.4

Table 1: Rank-1 accuracy w.r.t. codebook size for coupled clustering based on Fisher Vector.

4.2 Coupled Clustering

In this section we evaluate the proposed coupled clustering method from Section 3.1. We extract image features by densely sampling 24× 12 pixel patches from ROI, ignoring their spatial locations. For each patch, we extract LAB and HSV histograms, each with 30 bins per channel. To keep only informative patches, we applied background subtraction. Please note that for **QUEUE02** we use fixed ROI, but for the other datasets (since we do not have access to original images containing whole scene) we use cropped images provided by the authors of datasets. Please note that in case **QUEUE02** our appearance model is based on Fisher Vector of randomly sampled patches within ROI and ignores patch locations, thus is invariant to small shifts of ROI. The resulting 180-dimensional feature vectors were used to generate codebooks using both BoW and FV. The appropriate size of each codebook was determined using cross validation: **CUHK01** FV(128) BoW (500), **VIPeR** FV(128) BoW (500), **iLIDS-groups** FV(20) BoW (300), **QUEUE02** FV(16) BoW(200).

In case of **CUHK01** and **VIPeR**, we followed the common evaluation protocol in re-identification. We split subjects from each dataset into training/testing: **VIPeR** - 316/316, **CUHK01** - 486/485.

In evaluation of the **iLIDS-groups** dataset we followed the same evaluation protocol as in [43], where the dataset was introduced. We randomly select one image from each group to build the gallery set and the rest of the images form the probe set. The selection procedure was repeated 10 times. To train the codebooks we use the iLIDS-MA dataset.

For **QUEUE02**, we split the data set into 11/12 groups, and similarly, we split QUEUE01 dataset to 10/13 individuals. In this way we assure that individuals from QUEUE01 who were used to train the codebook do not appear in the group from **QUEUE02** in the testing phase.

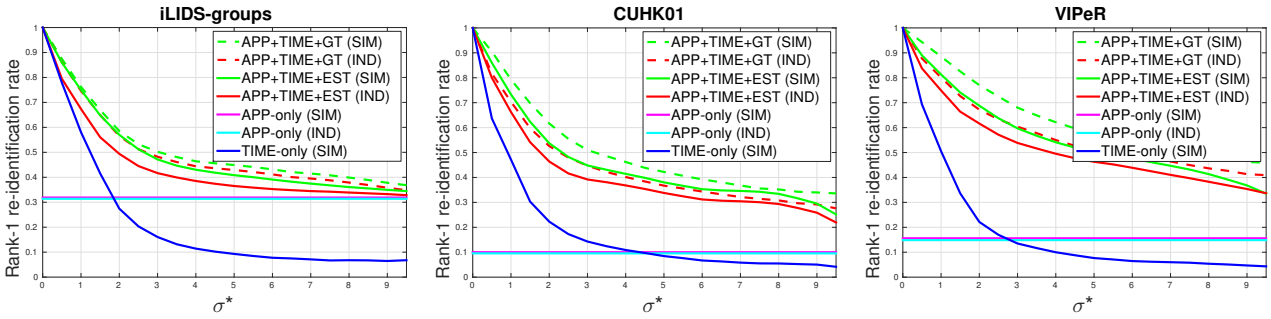

Figure 5: Rank-1 recognition rate based on the reliability of an expected travel time (simulated by synthesizing timestamps from different normal distributions with different standard deviations σ^*). Appearance-only models (magenta and cyan) are not affected by temporal context. Temporal context on its own (dark-blue) is powerful but degrades quickly as the variance in expected travel time increases (making temporal context less informative). Combining appearance and temporal context gives a significant boost in re-identification performance compared to using either model independently. Incorporating mutual exclusion constraints (green) generally leads to improved performance over independent queries (red).

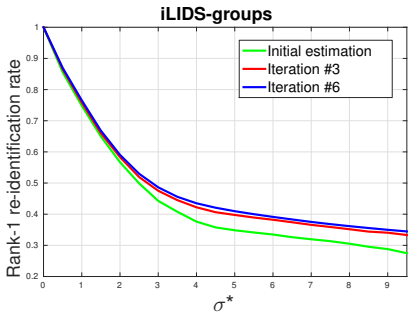

Figure 6: Rank-1 recognition rate with varying travel time consistency (using different standard deviations σ^* for synthetic timestamps). Iterative re-estimation of the unknown σ^* parameter improves overall Rank-1 accuracy.

When testing, we evaluate a single-shot scenario and repeat the codebook generation procedure 10 times for computing averaged CMC curves. Figure 4 illustrates CMC curves on the four datasets. It is apparent that our coupled clustering approach has a large margin of improvement over standard clustering on all datasets for both BoW and FV models. The significant performance gain is especially evident using FV codebooks. In table 1 we show Rank-1 accuracy w.r.t. FV codebook size. This experiment illustrates the merits of coupled clustering and confirms our claims that learning a generic codebook that maps features to codewords is not as effective as learning camera specific codebooks that map corresponding features across cameras to the same codewords. Note that the relatively small gain in performance on the **iLIDS-groups** dataset can be explained by the characteristics of the selected evaluation scheme. In the evaluation scheme proposed by the authors of the dataset, it is possible that both images from the same camera might be assigned to gallery and probe sets. Because re-identification on images from same camera is a much easier task, the advantages of the proposed coupled clustering are not evident on **iLIDS-groups** dataset.

4.3 Temporal Context

In this section, we evaluate the performance of the joint appearance and temporal context model (see Section 3.2 and Eq. (5)). We employ the previously learned coupled clustering FV codebooks (see Section 4.2) for an appearance model. We evaluate the benefit of temporal context by comparing the performance of the appearance-only model (see Eq. (1)) versus the appearance+temporal context model (see Eq. (2)). In both cases, the performance is measured when queries are conducted independently (IND) (see Eq. (5)) and by enforcing the proposed mutual exclusion constraints (Section 3.3) when performing all queries simultaneously (SIM) (see Eq. (8)).

VIPeR, **CUHK01** and **iLIDS-groups** do not contain timing information, so we simulate a queue scenario by assigning random timestamps to images from the first camera. Timestamps for images from the second camera are generated by sampling a normal distribution $\mathcal{N}(\tau^*, \sigma^{*2})$. In real-life τ^* and σ^{*2} parameters would be fixed and depend on crowd behavior. In practice, σ^{*2} controls how much the temporal order is preserved *w.r.t* the second camera. We set τ^* to 0 because it represents the expected time difference between images from two cameras and can be always normalized to 0. The above experimental setup gives us the opportunity to study the limits of the proposed method when temporal context is no longer informative.

Figure 5 illustrates the rank-1 recognition accuracy *w.r.t* σ^*. Performance of standard appearance-only re-identification (magenta and cyan) is invariant to σ^*. Temporal context on its own (dark blue) is quite powerful, but degrades quickly as flow from one camera to another is less ordered and the validity of the underlying assumption breaks down. Hybrid cost functions that consider both appearance and temporal context give significant improvement in performance.

The effect of estimating parameters τ and σ are evaluated by plotting the upper performance limit (dashed red & green) using

		CUHK01		VIPeR		iLIDS-groups		QUEUE02	
Method		*rank-1*	*rank-5*	*rank-1*	*rank-5*	*rank-1*	*rank-5*	*rank-1*	*rank-5*
KISSME [19]		16.41	38.02	19.60	49.53	9.57	31.64	32.14	79.46
XQDA [23]		63.21	83.00	40.00	68.13	26.15	46.76	25.89	92.86
Assoc. Groups of People [43]		-	-	-	-	22.50	45.0	-	-
Our (App - IND)		9.54	26.81	14.85	37.32	**31.38**	**47.37**	**47.25**	**96.20**
Our (App - SIM)		10.00	29.83	15.59	38.81	**31.88**	**48.65**	**55.46**	**100.00**
Our (App+Time+Est - IND)	s=0.5	79.98	92.00	83.77	98.52	71.55	100.00	100.00	100.00
	s=9.0	21.94	46.11	33.50	59.02	33.42	55.69		
Our (App+Time+Est - SIM)	s=0.5	**86.00**	**94.50**	**88.97**	**100.00**	**86.48**	**100.00**	**100.00**	**100.00**
	s=9.0	25.16	48.49	36.66	59.50	36.09	60.15		

Table 2: Performance comparison on CUHK01, VIPeR, iLIDS-groups and QUEUE02 datasets. We report *rank-1* and *rank-5* accuracy. Our appearance model outperforms standard metric learning approaches on group datasets. Integrating the temporal model together with our joint optimization consistently improves the re-identification performance in all datasets.

the true values of τ^* and σ^*. We then compare the performance of our temporal context model (solid red & green) estimated based on Algorithm 1. The results illustrate that there is a small performance drop using our estimation method. Incorporating mutual exclusion constraints (green) generally leads to improved performance over independent queries (red).

In Table 2 we report the performance on **QUEUE02**. This dataset contains real timestamps, so τ^* and σ^* are fixed. We use our proposed parameter estimation method to find τ and σ. Temporal context provides a significant boost in performance compared to appearance-only methods. Also, mutual exclusion leads to performance improvements.

Time Context Parameter Estimation In Section 3.2, we introduced an iterative algorithm to estimate the unknown parameters τ^* and σ^*. The experiments on **iLIDS-groups** shown in Fig. 6 confirm that iterative parameter re-estimation improves Rank-1 re-identification accuracy. After three iterations the algorithm converges and further parameter re-estimation does not translate into improvement in accuracy. Finally, the advantages of iterative parameter re-estimation is especially visible for cases with high $\sigma*$ (the ordering of people between cameras is rarely maintained). In such cases, the number of examples used for initial parameter estimation is small and insufficient to accurately estimate τ and σ.

4.4 State-of-the-art comparison

In this section we compare with two most common metric learning approaches: KISSME [19] and its extension - XQDA [23] that showed to be very effective while applied to re-identification problem [30]. We also report the results of [43] that proposed a dedicated descriptor for group re-identification. In Table 2 we report *rank-1* and *rank-5* accuracies on several re-identification datasets: **CUHK01**, **VIPeR** which contain well detected individuals, and group datasets: **iLIDS-groups** and **QUEUE02** which contain multiple people and a large amount of occlusion. The order of the datasets (from left column to right one) reflects the difficulty (*e.g.* VIPeR contains more variations in pose and image quality than CUHK01, and QUEUE02 contains significantly larger amount of occlusion than iLIDS-groups). Notice, that metric learning approaches perform relatively well on datasets that contain well detected individuals (CUHK01, VIPeR) but their performance degrades on group datasets (**iLIDS-groups**, **QUEUE02**). Our appearance model (App) ignores the spatial location of features, thus it is outperformed by metric learning approaches. However, with increasing difficulty of datasets, our codebook-based model outperforms metric learning approaches as the spatial information of features become less reliable. Notice that for iLIDS-groups and QUEUE02, our appearance-only model outperforms all other approaches, illustrating benefits of the proposed coupled clustering method. From Table 2 it is apparent that integrating the temporal model and using our joint optimization significantly increases the re-identification accuracy. In addition, we show that by applying mutual exclusion, we further improve the *Rank-1* and *Rank-5* accuracies (compare IND to SIM results). For datasets with simulated temporal information we also provide performance for two extreme values of σ^*. The results indicate that even if the ordering is rarely maintained ($\sigma^* = 9$) the proposed joint optimization still outperforms appearance-only model. The performance gain is even stronger when ordering is usually maintained between cameras ($\sigma^* = 0.5$).

5 SUMMARY

Many real-world scenarios have large amounts of occlusion and background clutter. Re-identification methods that require reliable detection will struggle in these circumstances. We propose a coupled clustering method that learns codebooks with entries that correspond across cameras and that is robust to both occlusion and alignment errors. We make use of common movement patterns by incorporating temporal context into the re-identification process. We show how the parameters of our model can be estimated from data, and illustrate the benefit of jointly optimizing appearance and temporal models. Our experiments demonstrate that methods based on coupled clustering and temporal context provided significant performance gains on challenging group re-identification datasets. Furthermore, we show how mutual exclusion constraints between multiple simultaneous queries (to preserve an ordering preference) also helps improve re-identification performance. Finally, our method has high practical impact in real-world scenarios, not only in terms of robustness, but also in usability (high Rank-1 accuracy).

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K. Marks. 2015. An Improved Deep Learning Architecture for Person Re-Identification. In *CVPR*.

[2] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. 2012. Boosted human re-identification using Riemannian manifolds. *Image and Vision Computing* 30, 6-7 (2012), 443 – 452.

[3] Ben Benfold and Ian Reid. 2011. Stable Multi-Target Tracking in Real-Time Surveillance Video. In *CVPR*.

[4] R. W. Butler, P. L. Davies, and M. Jhun. 1993. Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics* 21, 3 (09 1993), 1385–1400.

[5] Y. Cai, V. Takala, and M. Pietikainen. 2010. Matching Groups of People by Covariance Descriptor. In *2010 20th International Conference on Pattern Recognition*. 2744–2747. https://doi.org/10.1109/ICPR.2010.672

[6] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang. 2015. Similarity Learning on an Explicit Polynomial Kernel Feature Map for Person Re-Identification. In *CVPR*.

[7] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. 2011. Custom Pictorial Structures for Re-identification. In *BMVC*.

[8] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. 2004. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*.

[9] Abir Das, Anirban Chakraborty, and Amit K. Roy-Chowdhury. 2014. Consistent Re-identification in a Camera Network. In *ECCV*.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

[11] D. Gray, S. Brennan, and H. Tao. 2007. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *PETS* (2007).

[12] Chang Huang, Bo Wu, and Ramakant Nevatia. 2008. Robust Object Tracking by Hierarchical Association of Detection Responses. In *ECCV*.

[13] Timothy Huang and Stuart Russell. 1997. Object Identification in a Bayesian Context. In *IJCAI*.

[14] Haruyuki Iwama, Yasushi Makihara, and Yasushi Yagi. 2012. Group Context-aware Person Identification in Video Sequences. *IPSJ Transactions on Computer Vision and Applications* 4 (2012), 87–99. https://doi.org/10.2197/ipsjtcva.4.87

[15] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. 2008. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding* 109, 2 (2008), 146 – 162.

[16] Xiao-Yuan Jing, Xiaoke Zhu, F. Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. 2015. Super-resolution Person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*.

[17] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2012. Multi-view Discriminant Analysis. In *Proc. of the 12th European Conf. on Computer Vision*. 808–821.

[18] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. 2006. Learning Discriminative Canonical Correlations for Object Recognition with Image Sets. In *Proc. of the 9th European Conference on Computer Vision*. 251–262.

[19] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2012. Large Scale Metric Learning from Equivalence Constraints. In *CVPR*.

[20] H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955).

[21] Sheng Li, Ming Shao, and Yun Fu. 2015. Cross-view Projective Dictionary Learning for Person Re-identification. In *IJCAI*.

[22] Wei Li, Rui Zhao, and Xiaogang Wang. 2012. Human Reidentification with Transferred Metric Learning. In *ACCV*.

[23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In *CVPR*.

[24] C.C. Loy, T. Xiang, and S. Gong. 2009. Multi-camera activity correlation analysis. In *CVPR*.

[25] Chen Change Loy, Tao Xiang, and Shaogang Gong. 2010. Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding. *International Journal of Computer Vision* 90, 1 (2010), 106–129.

[26] C. C. Loy, T. Xiang, and S. Gong. 2012. Incremental Activity Modeling in Multiple Disjoint Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 9 (Sept 2012), 1799–1813.

[27] Bingpeng Ma, Yu Su, and Frederic Jurie. 2012. Local Descriptors Encoded by Fisher Vectors for Person Re-identification. In *ECCVW*.

[28] Y. Makihara, A. Mansur, D. Muramatsu, Z. Uddin, and Y. Yagi. 2015. Multi-view Discriminant Analysis with Tensor Representation and Its Application to Cross-view Gait Recognition. In *Proc. of the 11th IEEE Conf. on Automatic Face and Gesture Recognition (FG 2015)*. Ljubljana, Slovenia, 1–8.

[29] D. Makris, T. Ellis, and J. Black. 2004. Bridging the gaps between cameras. In *CVPR*.

[30] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. 2016. Hierarchical Gaussian Descriptor for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Riccardo Mazzon, Syed Fahad Tahir, and Andrea Cavallaro. 2012. Person re-identification in crowd. *Pattern Recognition Letters* 33, 14 (2012), 1828 – 1837. Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context.

[32] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. 2015. Learning to rank in person re-identification with metric ensembles. In *CVPR*.

[33] Sateesh Pedagadi, James Orwell, Sergio A. Velastin, and Boghos A. Boghossian. 2013. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *CVPR*.

[34] Florent Perronnin, Yan Liu, Jorge Sanchez, and Herve Poirier. 2010. Large-Scale Image Retrieval with Compressed Fisher Vectors. In *CVPR*.

[35] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. 2010. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*.

[36] A. Rahimi, B. Dunagan, and T. Darrell. 2004. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *CVPR*.

[37] S. H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. 2016. Joint Probabilistic Matching Using m-Best Solutions. In *CVPR*.

[38] Peter J. Rousseeuw and Katrien Van Driessen. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41, 3 (Aug. 1999), 212–223.

[39] C. C. Sun, G. S. Arr, R. P. Ramachandran, and S. G. Ritchie. 2004. Vehicle Reidentification using multidetector fusion. *Transactions on Intelligent Transportation Systems* 5, 3 (2004).

[40] Xiaogang Wang. 2013. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters* 34, 1 (2013), 3 – 19. Extracting Semantics from Multi-Spectrum Video.

[41] Li Zhang, Yuan Li, and R. Nevatia. 2008. Global data association for multi-object tracking using network flows. In *CVPR*.

[42] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2014. Learning Mid-level Filters for Person Re-identification. In *CVPR*.

[43] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2009. Associating Groups of People. In *BMVC*.

[44] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. 2015. Partial Person Re-Identification. In *ICCV*.