Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification

Elyor Kodirov
e.kodirov@qmul.ac.uk

Tao Xiang
t.xiang@qmul.ac.uk

Shaogang Gong
s.gong@qmul.ac.uk

School of Electronic Engineering and
Computer Science,
Queen Mary University of London
London E1 4NS, UK

1 Introduction

Person re-identification (Re-ID) is the problem of matching people across non-overlapping camera views. Despite the best efforts from the computer vision researchers, it remains an unsolved problem. This is because a person's appearance often changes dramatically across camera views due to changes in pose, occlusion, lighting, and illumination conditions.

Many existing approaches to person re-identification (Re-ID) are based on supervised learning [1, 2, 3, 4] which requires hundreds of matching pairs to be labelled for each pair of cameras. This severely limits their scalability for real-world applications. This work aims to overcome this limitation by developing a novel unsupervised Re-ID approach. The approach is based on a new dictionary learning for sparse coding formulation with a graph Laplacian regularisation term whose value is set iteratively. As an unsupervised model, the dictionary learning model is well-suited to the unsupervised task, whilst the regularisation term enables the exploitation of cross-view identity-discriminative information ignored by existing unsupervised Re-ID methods. Importantly this model is also flexible in utilising any labelled data if available. Experiments on VIPeR and PRID benchmark datasets demonstrate that the proposed approach significantly outperforms the state-of-the-art.

2 Methodology

Problem Definition. Suppose a set of training person images are collected from a pair of camera views denoted as A and B respectively. An n-dimensional feature vector is computed from each person's image to represent ones appearance. Let's denote the training data matrix as $X = [X^a,\ X^b] \in \mathbb{R}^{n\times m}$ where $X^a = [x_1^a,\ \dots\ ,x_{m_1}^a] \in \mathbb{R}^{n\times m_1}$ contains the feature vectors of m_1 images in view A as columns, while $X^b = [x_1^b,\ \dots\ ,x_{m_2}^b] \in \mathbb{R}^{n\times m_2}$ does the same for the m_2 images in view B. We thus have $m = m_1 + m_2$. Note, the training data are *unlabelled* therefore it is unknown which person observed in view A corresponds to a given person in view B and vice versa. The objective of unsupervised person Re-ID is to learn a matching function f from X, so that given x^a and x^b representing two test person images from A and B respectively, $f(x^a, x^b)$ can be used for matching their identities.

Dictionary Learning with Graph Laplacian Regularisation. Our solution to the problem defined above is to learn a shared dictionary $D \in \mathbb{R}^{k\times m}$ for the two camera views using X. With this dictionary, each n-dimensional feature vector, regardless which view it comes from, is projected into a lower k-dimensional subspace spanned by the k dictionary atoms (columns of D) so that they can be matched by the cosine distance in this subspace. The underpinning idea is that each atom or the dimension of the subspace corresponds to a latent appearance attribute which is invariant to the camera view changes, thus useful for cross-view matching. Formally, we aim to learn the optimal dictionary D, such that the sparse code of X, denoted as $Y = [Y^a,\ Y^b] \in \mathbb{R}^{k\times m}$, where $Y^a = [y_1^a,\ \dots\ ,y_{m_1}^a] \in \mathbb{R}^{k\times m_1}$ and $Y^b = [y_1^b,\ \dots\ ,y_{m_2}^b] \in \mathbb{R}^{k\times m_2}$, can be used for matching the training data; and we wish the same D can be generalised to match unseen test image pairs from the two views.

Using a conventional dictionary learning formulation, D and Y can be estimated as:

$$(D^*, Y^*) = \underset{D,Y}{\operatorname{argmin}} \|X - DY\|_F^2 + \alpha ||Y||_1 \tag{1}$$

where $\|X - DY\|_F^2$ is the reconstruction error term evaluating how well a linear combination of the learned atoms can approximate the input data, and $||.||_F$ denotes the matrix Frobenious norm; $||Y||_1$ is the sparsity term favouring small number of atoms to be used for reconstruction; this term is weighted by α. It is clear from this formulation that the conventional dictionary learning model only cares about how to best reconstruct X using D and Y, without giving any consideration to whether the sparse code is meaningful for matching people cross camera views. In order to make the learned dictionary discriminative for cross-view matching, one must exploit cross-view identity discriminative information. With cross-view labels, this can be achieved by forcing the two matched images to have identical sparse codes [5]. However, without any labels available under our unsupervised setting, it is not possible to use this conventional formulation for person Re-ID.

To overcome this problem, we introduce a graph Laplacian regularisation term in the dictionary learning formulation, and rewrite Eq. (1) as

$$(D^*, Y^*) = \underset{D,Y}{\operatorname{argmin}} \|X - DY\|_F^2 + \alpha ||Y||_1 + \beta \sum_{i,j}^{m} ||y_i^a - y_j^b||_2^2 W_{ij} \tag{2}$$

where β is the weight of the new regularisation term, and $W \in \mathbb{R}^{m\times m}$ is a cross-view correspondence matrix capturing the identity relationship between the people in X^a and X^b which needs to be preserved after they are projected and become Y^a and Y^b. Note, since the training data are unlabelled, the true cross-view correspondence relationship is unknown. We therefore use W to represent a soft cross-view correspondence relationship. That is, each person in A can correspond to multiple people in B depending on their visual similarity.

3 Experimental Results

Experimental results on VIPeR and PRID benchmark datasets show that our model significantly outperforms the state of the art.

Table 1: Unsupervised Re-ID results on VIPeR and PRID

Dataset	VIPeR				PRID			
Ranks	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
eSDC	26.7	50.7	62.4	76.4	-	-	-	-
SDALF	19.9	38.9	49.4	65.7	16.3	29.6	38.0	48.7
ISR	27.0	49.8	61.2	73.0	17.0	34.4	42.0	54.3
CPS	22.0	44.7	57.0	71.0	-	-	-	-
GTS	25.2	50.0	62.5	75.8	-	-	-	-
Ours	**29.6**	**54.8**	**64.8**	**77.3**	**21.1**	**43.7**	**55.8**	**64.8**

Table 2: Semi-supervised Re-ID results on VIPeR and PRID

Dataset	VIPeR				PRID			
Ranks	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
RankSVM	20.7	41.8	54.6	68.1	-	-	-	-
KISSME	18.5	43.7	57.9	74.5	5.1	15.2	24.1	40.1
kLFDA	27.5	56.0	69.6	82.6	14.1	33.7	44.0	56.2
KCCA	24.6	56.2	71.7	**85.6**	5.3	15.7	25.8	37.0
MFA	25.3	53.6	66.7	78.8	13.3	32.5	43.3	56.4
SSCDL	25.6	53.7	68.2	83.6	-	-	-	-
Ours	**32.5**	**61.8**	**74.3**	84.1	**22.1**	**45.3**	**56.5**	**66.3**

[1] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Proc. ACCV*, 2011.

[2] L. Giuseppe, M. Iacopo, and D. B. Alberto. Matching people across camera views using kernel canonical correlation analysis. In *Proc. ICDSC*, 2014.

[3] M. Hirzer, M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *Proc. AVSS*, 2012.

[4] M. Hirzer, M. Roth, M. Koestinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proc. ECCV*, 2012.

[5] X. Liu, M. Song, D. Tao, X. Zhou, Ch. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proc. CVPR*, 2014.