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

Abstract

Many existing approaches to person re-identification (Re-ID) are based on supervised learning, 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 two benchmark datasets demonstrate that the proposed approach significantly outperforms the state-of-the-arts.

1 Introduction

Person re-identification (Re-ID) is the problem of matching people across non-overlapping camera views. It has received increasing attention in the past five years due to its huge potentials for security and safety management applications. Despite the best efforts from the computer vision researchers, it remains an unsolved problem. This is because a person's appearance often changes dramatically cross camera views due to changes in pose, occlusion, lighting, and illumination conditions. To overcome these challenges, most existing works [6, 10, 13, 14, 17, 27, 28, 30] employ a large number of labelled matching pairs across each two camera views to learn a matching function that is invariant to the appearance changes. However, these supervised learning-based approaches have poor *scalability*. More specifically, even for a camera network of moderate size, e.g. one installed in an underground station, there can be easily over one hundred cameras. Since hundreds of labelled image pairs are needed for each pair of these cameras, the labelling cost would be prohibitively high. This scalability issue thus severely limits the applicability of the existing methods.

In order to make a person Re-ID model scalable, one solution is to utilise the unlabelled data, which are abundant in the context of Re-ID – in a busy public space, thousands of people pass by in each camera view everyday. There are a few existing efforts on exploiting unlabelled data for unsupervised Re-ID modelling [8, 20, 26]. However, compared to supervised learning approaches, the matching performance of unsupervised models are typically much weaker, rendering them less effective. This is not surprising because none of them benefits from labelled cross-view discriminative information in every camera pairs. Such identity-discriminative information is vital in person re-identification and is the reason why those supervised learning based methods achieve much higher matching accuracy. In general, without cross-view data labels, this information is very difficult to obtain.

In this work, we propose a novel unsupervised learning model for person re-identification that can learn cross-view person identity-discriminative information from unlabelled data. Our model is based on a dictionary learning for sparse coding framework. That is, we attempt to learn a set of dictionary atoms, of which each corresponds to a latent attribute that is invariant across camera views, therefore useful for matching. Conventional dictionary learning approaches are unsupervised [1], designed for learning a set of linear bases to minimise signal reconstruction errors. They are unsuitable for learning any discriminative information across camera views. To overcome this limitation, we introduce a cross-view graph Laplacian regularisation term in our dictionary learning formulation. This term captures a soft cross-view correspondence relationship across camera views, meaning that visually similar people across views are more likely to have the same identity. Our model aims to preserve this relationship in a subspace spanned by the learned dictionary bases. The value of the regularisation term is computed iteratively so that the initially noisy soft-correspondence can be improved when it is computed in the subspace defined by the dictionary atoms rather than the original feature space. Importantly, the same regularisation term can accommodate various amounts of labelled data when available, whilst keeping the ability to exploit the unlabelled data. This makes our model extremely flexible for various application scenarios.

Our contributions are as follows: (1) A novel regularised dictionary learning-based person Re-ID model is proposed for exploiting unlabelled data, which makes the model scalable to large-scale Re-ID problems. (2) The model offers a flexible solution to utilising unlabelled as well as arbitrary amount of labelled cross-view data. Extensive experiments are carried out on two large benchmark datasets. The results show that the proposed model outperforms the state-of-the-art unsupervised approaches. Furthermore, it is clearly superior to the existing semi-unsupervised methods and remains competitive even under the conventional fully supervised setting.

2 Related Work

Supervised Re-ID. Most Re-ID approaches rely on labelled data (cross-view matched image pairs) and are based on supervised distance metric or ranking learning models [6, 10, 13, 14, 17, 27, 28, 30]. Most of the models are non-linear, and many of them are kernelised (e.g. [27]) to cope with the complex appearance variations across camera views. In contrast, the proposed model does not rely on labelled data, thus is not limited by the prohibitively high labelling cost in large scale problems involving hundreds of cameras. Our model also differs in that it is a linear model. Surprisingly it can beat the most competitive non-linear models even under the fully supervised setting which they were designed for.

Learning from unlabelled data. Existing unsupervised Re-ID models rely on hand-crafted

appearance features [8, 20, 22] or localised saliency statistics [26, 29]. Both types of methods have their limitations: for the hand-craft feature- based methods, it is very hard to obtain effective identity-discriminative features by manual design, due to the unknown large cross-view covariates. Saliency-based methods, on the other hand, rely on a representative reference set, and again are not able to explicitly exploit the cross-view identity-discriminative information. These unsupervised methods thus typically achieve much weaker matching accuracy than those supervised methods. As a comprise between scalability and matching accuracy, recently a semi-supervised Re-ID model is proposed which is based on a coupled dictionary learning method [21]. Similar to our model, the use of a dictionary learning for sparse coding model enables the model to utilise unlabelled data. However, those data are only used to minimise the reconstruction error within each camera view independently from other views. Instead, we use the cross-view soft correspondence relationship to learn more discriminative information from the unlabelled data. Our experiments (Section 4) show that our model is superior to existing unsupervised and semi-supervised models for Re-ID.

Dictionary learning and Sparse coding. Beyond person Re-ID, dictionary learning for spare coding has been extensively studied [1, 15]. Graph Laplacian regularisation has also been explored in a sparse coding formulation before, for problems such as unsupervised clustering [9, 31], or supervised face verification/recognition [11]. Unlike these works, we are dealing with an unsupervised verification problem. As a result, the regularisation term is computed differently to capture the soft cross-view correspondence relationship between camera views. Furthermore, our model is learned iteratively with the regularisation term updated in each iteration to improve the cross-view correspondence relationship captured by the regularisation term. In particular, initialised in the noisy visual feature space, the soft-correspondence is computed in a subspace of lower-dimension defined by the dictionary atoms learned from the previous iteration. Such a space is progressively more discriminative for matching people across camera views.

3 Methodology

3.1 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 $X^a = [x^a_1, \ldots, x^a_{m_1}] \in \mathbb{R}^{n\times m_1}$ contains the feature vectors of m_1 images in view A as columns, while $X^b = [x^b_1, \ldots, x^b_{m_2}] \in \mathbb{R}^{n\times m_2}$ 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.

3.2 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 $Y^a = [y^a_1, \ldots, y^a_{m_1}] \in \mathbb{R}^{k \times m_1}$ and $Y^b = [y^b_1, \ldots, y^b_{m_2}] \in \mathbb{R}^{k \times m_2}$, 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 [21]. 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

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

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. Formally, given X^a and X^b we construct a nearest neighbour graph G across cameras with m vertices, where each vertex represents a data point. W is then computed as the weight matrix of G. More precisely, if x^a_i is among the K-nearest neighbours of x^b_j or vice versa, $W_{i,j} = ((x^a_i)^T x^b_j)/(||x^a_i||||x^b_j||)$; otherwise, $W_{i,j} = 0$. Given this regularisation term, we essentially make an assumption that visually similar images are more likely to contain people of the same identity, and their sparse code vectors should also be similar, i.e. having a small distance measured by $||y^a_i - y^b_j||_2^2$.

3.3 Optimisation

To solve the optimisation problem in Eq. (2), we first rewrite it as:

$$(D^*, Y^*) = \underset{D,Y}{\operatorname{argmin}} \|X - DY\|_F^2 + \alpha||Y||_1 + \beta Tr(YLY^T) \;\; s.t. \;\; ||d_i||^2 \leq 1, \;\; i = 1, ..., k \tag{3}$$

where the Laplacian matrix L is defined as $L = Q - W$, where $Q_{ii} = \sum_j W_{ij}$ is the degree of the i^{th} node. It is important to point out that Eq. (3) is not convex for D and Y simultaneously, although it is convex for each of them separately. We thus deploy an alternating optimisation method to solve it. In particular we alternate between the following two subproblems:
(1) *Fix Y, update D*: Given Y, the objective function becomes

$$D^* = \underset{D}{\operatorname{argmin}} \|X - DY\|_F^2 \;\; s.t. \;\; ||d_i||^2 \leq 1, \;\; i = 1, ..., k \tag{4}$$

To solve Eq. (4), we use the Lagrange dual method [18]. The analytical solution of D can be computed as: $D^* = XY^T(YY^T + \Lambda^*)^{-1}$, where Λ^* is diagonal matrix constructed from all the optimal dual variables. In practice, $YY^T + \Lambda^*$ is not guaranteed to be invertible, therefore pseudoinverse is used in place of computing it directly.

(2) *Fix D, update Y*: When D is fixed, we optimise each column of Y, y_i alternatively rather than optimise them simultaneously. Specifically, to optimise each y_i, we fix the sparse codes $y_j (j \neq i)$ for other local features. So the optimisation of the objective cost of Eq. (3) is equivalent to optimising the following objective function:

$$y_i^* = \underset{y_i}{\operatorname{argmin}} L(y_i) + \alpha ||y_i||_1, \tag{5}$$

where $L(y_i) = ||x_i - Dy_i||^2 + \beta(y_i^T(YL_i) + (YL_i)^T y_i - y_i^T L_{ii} y_i)$, and L_i is the i^{th} column of L and L_{ii} is the entry located in the i^{th} column, i^{th} row of L. We follow the widely used feature sign search algorithm [9, 18] to estimate y_i.

3.4 Iterative Updating the Regularisation Term

Note that when we compute the soft correspondence matrix W in Eq. (2), we used the cosine distance of the low-level features to measure the visual similarity. However, the low-level features are inevitably noisy and sensitive to the pose and lighting changes cross camera views. This is precisely why we wanted to do the matching in a new lower dimensional subspace defined by D rather than the n-dimensional low-level feature space. Now starting with the noisy W, and after obtaining D and Y using the alternative optimisation algorithm described above, we assume that the soft correspondence matrix W can now be better computed using Y rather than X. Given updated W, we repeat the alternating optimisation process to estimate D and Y. This iterative procedure stops when the cost function value converges. Note that this is very different from existing methods with a graph Laplacian regularisation term [9, 31], which stick to the same W or L matrix throughout the optimisation procedure. We observe from our experiments that (1) This iterative updating procedure converges rapidly (<5); and (2) it produces a marked improvement on the Re-ID performance compared to the model learned without updating the regularisation term. We summarise all steps of our method in Algorithm 1.

Algorithm 1: Dictionary learning with iterative graph Laplacian regularisation

Input: Training samples X, weights α and β, the initial Laplacian matrix L_0, number of iterations T given the current L, a threshold Th
Output: The learned dictionary D

Initialisation: iteration index $i = 0$, objective function value $O_0 = 100$;
while $O_i - O_{i-1} > Th$ **do**
  for $t = 1, 2, \ldots, T$ **do**
    Update sparse code Y using Eq. (5);
    Update dictionary D using Eq. (4);
  end
  Compute objective function O_i using Eq. (2);
  Compute the Laplacian matrix L_i;
  $i = i + 1$;
end

3.5 Matching

After learning the dictionary D using unlabelled training data X, given a pair of test samples x_i^a and x_i^b, we first compute their sparse codes y_i^a and y_i^b by solving the following functions:

$$y_i^{a*} = \underset{y_i^a}{\operatorname{argmin}} \|x_i^a - D y_i^a\|_F^2 + \alpha ||y_i^a||_1. \tag{6}$$

$$y_i^{b*} = \underset{y_i^b}{\operatorname{argmin}} \|x_i^b - D y_i^b\|_F^2 + \alpha ||y_i^b||_1. \tag{7}$$

These are standard LASSO problems [25] and can be solved very efficiently using the SPAMS toolbox [23]. After obtaining y_i^{a*} and y_i^{b*}, their matching is done by computing the cosine distance between their respective sparse code vectors. Alternatively, we can use L_2 constraint instead of L_1- constraint on coefficients. In this case, it will be simply regularised least squares problem, which has a closed-form solution: $y_i^{a*} = (D^T D + \alpha I)^{-1} D^T x_i^a$ and $y_i^{b*} = (D^T D + \alpha I)^{-1} D^T x_i^b$.

3.6 Extension to Semi-Supervised and Supervised Re-ID

Our model is designed primarily for unsupervised learning without any data labels. However, it can be readily extended to other settings with minimal modification. Specifically, when there are partially labelled cross-view image pairs, we simply set the corresponding $W_{i,j}$ to 1 to turn that part of W to be hard correspondence, whilst keep the rest of the $W_{i,j}$ computed as KNN graph weights as before. The matrix W thus becomes a hybrid of hard and soft correspondence matrix with only the soft part updated iteratively. When all the data are labelled, i.e. the fully supervised setting, all values of $W_{i,j}$ will be either 1 or 0 depending whether the corresponding cross-view image pair contains the same person. Matrix W thus becomes a hard correspondence matrix. In summary, when variable amount of data labels are available, ranging from zero to full, the model remain unchanged apart from the W or the Laplacian matrix L being computed. Our model is thus extremely flexible for deployment under different settings.

4 Experiments

4.1 Datasets and settings

Datasets. Two widely used benchmark datasets were used for experiments. **VIPeR** [7] contains 632 image pairs of people captured outdoor from two non-overlapping camera views. Following the standard setting, the dataset was randomly split into two sets of 316 image pairs, one for training and the other for testing. For the test set, all images from one view is used as the gallery set and the others probe set. The results for all evaluations were obtained by averaging over the 10 splits publicly available from [8]. **PRID** [12] is different from other available datasets in that the gallery and probe sets do not have exactly the same set of people, which is a more realistic setting in practice. Specifically, it has two camera views. View *A* captures 385 people, whilst View *B* contains 749 people. Only 200 people appear in both views. In our experiments we used the single shot version of the dataset as in [10, 14], i.e. one image per person per view. In each data split, 100 people with one image from each view were randomly chosen from the 200 present in both camera views for the training set, while the remaining 100 of View *A* were used as the probe set, and remaining 649 of View *B* were used as gallery (containing the 100 people in the probe). Experiments were carried out on the 10 splits as in [10, 14] with the averaged results reported.

Features. For person appearance representation, we used the histogram-based image descriptor introduced in [10]. Specifically, three types of features, (1) Colour histogram, (2) HOG [5] and (3) LBP [2], were concatenated resulting in a 5138-D feature vector [10].

Evaluation metric. We obtain conventional Cumulative Matching Characteristics (CMC) curves for our models and other models with codes available. However, to compare with a wider ranges of baselines, for which no code is available, we report cumulative matching accuracies at different ranks which correspond to key points on the CMC curves.

Parameter settings. The parameters of our model were set to the following: the number of nearest neighbours for computing W in Eq. (2), $K = 3$ (and we obtained similar results when $K < 6$); the weight for the sparsity penalty term in Eq. (2), $\alpha = 0.0001$ (we found empirically that α should be in the range of 0.0001 and 0.3); the weight on the Laplacian regularisation, $\beta = 1$ for VIPeR, $\beta = 0.5$ for PRID; the number of iteration, $T = 50$ (see Algorithm 1); and the dictionary size, $k = 256$.

4.2 Unsupervised Re-ID

Competitors. Under this setting, we compared our approach with (1) the hand-crafted feature-based methods including SDALF [8] and CPS [4]. These features are designed to be view invariant. (2) The saliency learning-based eSDC [29] and GTS [26]. (3) The sparse representation classification-based ISR [20]. Note that ISR is transductive in that it uses the gallery set of the test data whilst none of other methods does.

Comparative results. From Table 1, the following observations can be made: (1) Our regularised sparse coding-based Re-ID model is clearly superior to all existing unsupervised methods on both datasets. (2) The margin is bigger on the more challenging PRID dataset which has a smaller training set and much bigger test gallery set than VIPeR. (3) Our unsupervised ReID model is competitive even compared to the state-of-the-arts supervised methods. In particular, compared to the supervised Re-ID results in Table 3, our Rank 1 matching accuracy for VIPeR (29.6%) is slightly better than a number of recently proposed models such as LDFA [24] (29.3%) and MLF [30] (29.1%), and significantly better than earlier

models such as RankSVM [8] (25.2%) and KISSME [16] (25.4%). On the more challenging PRID datasets, our unsupervised method (21.1% at Rank 1) outperforms all existing supervised learning methods (see Table 3, the best results are 15% by RPLM [12] and EIML [13]). This demonstrates clearly the effectiveness of the proposed new unsupervised Re-ID model. (4) All learning-based models clearly outperform the handcrafted feature-based methods (SDALF and CPS). (5) Both of the best two models (ours and ISR) are based on sparse learning. But there are vital differences: ISR uses the test gallery set directly as dictionary, whilst our model learns a dictionary from an unlabelled training set. Our model is less expensive to compute and more flexible as once learned the sparse code is unchanged for any test gallery and probe images. In contrast, using ISR, all codes need to be recomputed when new people are added to the test gallery.

Dataset	VIPeR				PRID			
Ranks	Rank 1	Ranks 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
eSDC [29]	26.7	50.7	62.4	76.4	-	-	-	-
SDALF [8]	19.9	38.9	49.4	65.7	16.3	29.6	38.0	48.7
ISR [20]	27.0	49.8	61.2	73.0	17.0	34.4	42.0	54.3
CPS [4]	22.0	44.7	57.0	71.0	-	-	-	-
GTS [28]	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 1: Unsupervised Re-ID results on VIPeR and PRID

4.3 Semi-supervised Re-ID

Setting and competitors. In this experiment, one third of the labels of the training data are provided as in [21]. Only one semi-supervised Re-ID method exists: SSCDL[21] which is also based on dictionary learning but does not exploit cross-view identity-discriminative information using the unlabelled data as in our model. In addition to SSCDL, we also compared with a number of fully-supervised models including the classic RankSVM [8] and KISSME [16], and the state-of-the-arts MFA [27], kLFDA [27] and KCCA [10]. These fully supervised model can only use the one-third labelled training data. All of their codes are publically available therefore we used the same features. In contrast, SSCDL is a patch-based matching approach, thus their reported results are used for comparison.

Comparative results. Table 2 reveals the the following findings: (1) Again, our model as a semi-supervised method is clearly superior to the existing approach namely SSCDL. This highlights the importance of learning cross-view discriminative information from unlabelled data. (2) Compared to our results under the unsupervised setting (Table 1), the improvements indicate that our model benefits from both labelled and unlabelled data. (3) The supervised models, learned using the one third labelled training data only, are clearly inferior. This is particularly the case on the PRID dataset. For PRID, one third of the training set only gives 33 labelled pairs, which is evidently not enough for the existing supervised learning models to learn a useful matching function, as indicated by the significantly worse results obtained by KLFDA and KCCA (14.1% and 5.3% respectively, compared to 22.1% by our model).

4.4 Supervised Re-ID

Competitors. With all training data labelled, we compared with 9 methods: RankSVM [8], KISSME [16], kLFDA [27], MFA [27], KCCA [10], MLF [30], LFDA [24], EIML [13],

Dataset	VIPeR				PRID			
Ranks	Rank 1	Ranks 5	Ranks 10	Ranks 20	Rank 1	Rank 5	Rank 10	Rank 20
RankSVM [8]	20.7	41.8	54.6	68.1	-	-	-	-
KISSME [16]	18.5	43.7	57.9	74.5	5.1	15.2	24.1	40.1
kLFDA [27]	27.5	56.0	69.6	82.6	14.1	33.7	44.0	56.2
KCCA [10]	24.6	56.2	71.7	**85.6**	5.3	15.7	25.8	37.0
MFA [27]	25.3	53.6	66.7	78.8	13.3	32.5	43.3	56.4
SSCDL [21]	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**

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

RPLM [14], most of which are distance metric learning-based ones.

Comparative results. The results in Table 3 show that our model, when deployed under the conventional setting, is still very competitive. In particular, on VIPeR, our Rank 1 matching rate of 38.9% is only slightly worse than that of KLFDA (40.7%), whilst being better than all other compared methods[1]. On the more challenging PRID, our result is significantly better than all competitors. Note that our model is designed for unsupervised learning and under the fully supervised setting, both the soft-correspondence and iterative Laplacian regularisation features are inapplicable. The significant advantage of our model over the others can only be explained by the ability to jointly learn a set of view-invariant and identity-discriminative latent attributes by dictionary learning.

Dataset	VIPeR				PRID			
Ranks	Rank 1	Ranks 5	Ranks 10	Ranks 20	Rank 1	Rank 5	Rank 10	Rank 20
RankSVM [8]	25.2	48.1	60.3	74.8	-	-	-	-
KISSME [16]	25.4	53.3	67.7	82.1	10.2	26.1	37.4	53.2
kLFDA [27]	**40.7**	70.0	81.2	90.8	19.7	44.9	56.4	65.9
MLF [30]	29.1	52.3	66.0	79.9	12.3	20.9	27.1	35.1
MFA [27]	33.5	65.2	77.2	87.3	17.4	39.1	52.6	64.5
LDFA [24]	29.3	61.0	76.0	88.1	18.9	42.9	54.4	66.3
KCCA [10]	37.2	**71.8**	**84.6**	**92.7**	14.5	34.3	46.6	59.1
RPLM [14]	27.0	-	69.0	83.0	15.0	-	42.0	54.0
EIML [13]	22.0	-	63.0	78.0	15.0	-	38.0	50.0
Ours	38.9	70.8	78.5	86.1	**25.2**	**51.9**	**62.9**	**71.6**

Table 3: Supervised Re-ID results on VIPeR and PRID.

4.5 Further Analysis

Effects of iterative Laplacian Regularisation. One of the key features that distinguishes our model from conventional dictionary learning for sparse coding models is that we introduced a cross-view graph Laplacian regularisation term in our dictionary learning formulation whose value is updated iteratively. Table 4 compares the performance of Re-ID under both unsupervised and semi-supervised settings with and without the iterative updating of the regularisation term. It shows that there is marked improvement when the regularisation term is updated. The improvement is bigger under the unsupervised setting – without any labelled data, the soft-correspondence matrix W is obtained entirely using low-level features, thus noisier than that under semi-supervised setting; iterative updating to improve the matrix is therefore more beneficial.

[1]Note that MLF [30] achieved a higher Rank 1 result (43.39%) when combined with LADF [19].

	VIPeR		PRID	
	Unsupervised	Semi-supervised	Unsupervised	Semi-supervised
without iteration	25.4	28.8	16.2	19.9
with iteration	**29.6**	**32.5**	**21.1**	**22.1**

Table 4: Rank 1 matching accuracy of our model with and without iterative updating the graph Laplacian regularisation term.

Running costs. Our model can run very efficiently. For example, on VIPeR, for each image, once the features are extracted, the sparse coding part took 0.023s for each image based on a desktop machine with Intel CPU at 3.30GHz and memory of 8.0 GB with MATLAB implementation. Following that, matching one pair of images only involves computing a cosine distance between their sparse codes.

5 Conclusion

We have proposed a novel unsupervised person Re-ID method based on a regularised dictionary learning approach. Compared to existing models, our method is unique in that it can exploit unlabelled data to learn cross-view identity-discriminative information due to a new graph Laplacian formulation updated iteratively. Experiments on two benchmark datasets show that the proposed model significantly outperforms existing methods under various settings including unsupervised, semi-supervised and fully supervised.

Acknowledgements

The authors were funded in part by the European Research Council under the FP7 Project SUNNY (grant agreement no. 313243).

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 2006.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[3] P. Bryan, Z. Wei-Shi, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. BMVC*, 2010.

[4] D. S. Cheng, M. Cristani, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Proc. BMVC*, 2011.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

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

[7] G. Douglas, B. Shane, and T. Hai. Evaluating appearance models for recognition, reacquisition and tracking. In *PETS*, 2007.

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

[9] Sh. Gao, I. Tsang, L. Chia, and P. Zhao. Local features are not lonely laplacian sparse coding for image classification. In *Proc. CVPR*, 2010.

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

[11] H. Guo, Z. Jiang, and L. S. Davis. Discriminative dictionary learning with pairwise constraints. In *Proc. ACCV*, 2014.

[12] M. Hirzer, C. Beleznai, M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. SCIA*, 2011.

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

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

[15] K. Kenneth, M.Joseph, R. Bhaskar, E. Kjersti, L. Te-Won, and S. Terrence. Dictionary learning algorithms for sparse representation. *Neural Computing*, 15(2), February 2003.

[16] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, 2012.

[17] R. Layne, T. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *Proc. BMVC*, 2014.

[18] H. Lee, A. Battle, R. Raina, , and Y. Ng. Andrew. Efficient sparse coding algorithms. In *Proc. NIPS*, 2007.

[19] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[20] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[21] 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.

[22] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *Proc. BMVC*, 2012.

[23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. In *Journal of Machine Learning Research*, volume 11, pages 19–60. 2010.

[24] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proc. CVPR*, 2013.

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996.

[26] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *Proc. BMVC*, 2014.

[27] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Proc. ECCV*, 2014.

[28] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Li. S. Salient color names for person re-identification. In *Proc. ECCV*, 2014.

[29] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proc. CVPR*, 2013.

[30] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Proc. CVPR*, 2014.

[31] M. Zheng, J. Bu, Ch. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph regularized sparse coding for image representation. In *IEEE Transactions on Image Processing*, 2011.