Early Active Learning with Pairwise Constraint for Person Re-identification

Wenhe Liu[1], Xiaojun Chang[2], Ling Chen[1] and Yi Yang[1]

CAI, University of Technology Sydney, Sydney, Australia[1]
LTI, Carnegie Mellon University, Pittsburgh, USA[2]
{allenlwh, cxj273, yeeiyang}@gmail.com, Ling.Chen@uts.edu.au

Abstract. Research on person re-identification (re-id) has attached much attention in the machine learning field in recent years. With sufficient labeled training data, supervised re-id algorithm can obtain promising performance. However, producing labeled data for training supervised re-id models is an extremely challenging and time-consuming task because it requires every pair of images across no-overlapping camera views to be labeled. Moreover, in the early stage of experiments, when labor resources are limited, only a small number of data can be labeled. Thus, it is essential to design an effective algorithm to select the most representative samples. This is referred as *early active learning* or *early stage experimental design* problem. The pairwise relationship plays a vital role in the re-id problem, but most of the existing early active learning algorithms fail to consider this relationship. To overcome this limitation, we propose a novel and efficient early active learning algorithm with a pairwise constraint for person re-identification in this paper. By introducing the pairwise constraint, the closeness of similar representations of instances is enforced in active learning. This benefits the performance of active learning for re-id. Extensive experimental results on four benchmark datasets confirm the superiority of the proposed algorithm.

Keywords: Early active learning, Person re-identification

1 Introduction

The primary target of person re-identification (re-id) is to identify a person from camera shots across pairs of non-overlapping camera views, and research on this topic has attracted considerable attention in recent years [8,9,10,15,29]. In the field of computer vision, re-id can be formed as an image *retrieval* task. Given a *probe* image of a person from one camera view, the difficulty is to identify images of the same person from a *gallery* of images taken by other non-overlapping camera views. Despite the encouraging results reported in previous works, re-id remains a challenge in several respects. The accuracy of identification is often degrades as a result of the uncontrollable and/or unpredictable variation of appearance changes across camera views, such as body pose, view angle, occlusion and illumination conditions [7,20,23].

Supervised re-id methods can achieve promising results if there are sufficient labeled training data. Unfortunately, the human labor necessary for labeling training data is sometimes inadequate. This problem becomes extremely severe in the re-id scenario, since labeling for re-id is difficult to achieve. Unlike other recognition tasks which only requires each image to be labeled, re-id requires all pairs of images across camera views to be labeled. It is a tough task even for humans to identify the same person in different camera views among a potentially huge number of imposters [9,20]. At the same time, pairwise labeled data is required for each pair of camera views in the camera network in re-id, thus the labeling cost will become prohibitively high numbers of cameras in today's world. For example, there might be more than over a hundred in one underground train station [20].

To save labor costs, it is essential to design an effective algorithm that can select a subset of samples that are the most representative and/or informative for training. Active learning is widely studied to solve this kind of sample selection problem. As discussed in [18], active learning methods can be divided into two categories. The first category of algorithms select the most informative samples for labeling when there are already some labeled samples. They include uncertainty sampling methods [11,6,1,22] query by committee methods [21,3]. Most of these active learning methods prefer to select uncertainty data, or data that is difficult to analyze. They thus require a certain number of labeled samples to evaluate the uncertainty of the unlabeled data or sampling bias [18] will result. It is therefore recommended that such methods are only applied in the mid-stage of experiments when there are sufficient labeled data. For the purpose of distinguishing between the two categories, we refer to the first category of active learning methods as *traditional active learning*. The second category of active learning methods is considered for application in the early stage of experiments, when there are limited resources for labeling data. In this case, there are no labeled samples, thus labeling a small number of representative data is desirable for training reliable supervised models. In the category of early active learning, there are clustering-based methods [19,16] and transductive experimental design methods [27]. These kinds of active learning algorithms are referred to as *early active learning* or *early stage experimental design* [18]. We illustrate the procedures of and example of the traditional active learning algorithm, QUIRE [6], and our early active learning algorithm with pairwise constraint (abbreviated as EALPC) in Fig. 1.

In the rest of this paper, we focus on the early active learning methods for person re-identification applications. As mentioned, labeling re-id data is extremely labor-consuming and time-consuming. It is therefore highly desirable to enhance the learning performance in re-id applications by early active learning. Unfortunately, early active learning methods currently merely consider analyzing representative samples with pairwise relationships. Therefore, directly applying them for re-id may be not appropriate.

To overcome the limitations described above, we propose a novel algorithm for person re-identification, Early Active Learning with Pairwise Constraint, abbreviated as EALPC. The main contributions of our work are as follows:

1. We propose a novel Early Active Learning with Pairwise Constraint algorithm for person re-identification. To the best of our knowledge, this is the first method considers to consider both (a) applying early active learning for the re-id application, and (b) extending early active learning schema with pairwise constraint.
2. We introduce the $\ell_{2,1}$-norm to our objective function, which improves the robustness of our methods and suppresses the effects of outliers.
3. We propose an efficient algorithm to optimize the proposed problem. Our optimization algorithm also provides a closed form solution and guarantees to reach the global optimum in the convergence.

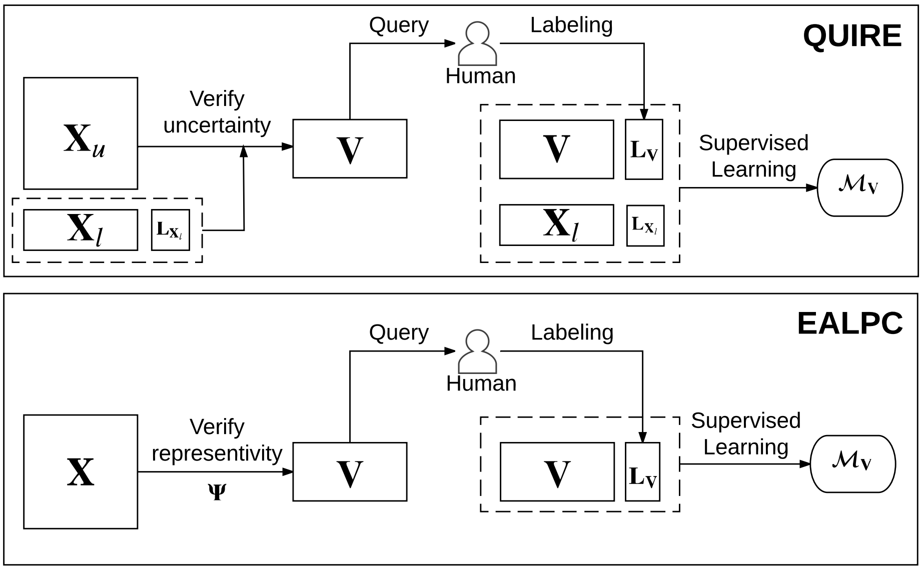


Fig. 1. Procedures of QUIRE [6] (*upper*) and our Early Active Learning with Pairwise Constraint (EALPC) (*lower*). In QUIRE, pre-labeled samples $\mathbf{X}_l$ are used for the uncertainty evaluation on the unlabeled samples $\mathbf{X}_u$. Then, it selects a subset samples $\mathbf{V} \subset \mathbf{X}_u$ for labeling. At last, both $\mathbf{X}_u$ and $\mathbf{V}$ along with their labels are used for supervised learning. In EALPC, unlabeled data $\mathbf{X}$ is analyzed without pre-labeled data. Meanwhile, pairwise constraint $\mathbf{\Psi}$ is introduced to enhance the performance of early active learning for re-id. More details are in Section. 2.

2 The Proposed Framework

In this section, we first revisit the early active learning algorithm and then propose our early active learning with pairwise constraint for re-id.

Notation. Let the superscript $^\top$ denote the transpose of a vector/matrix, $\mathbf{0}$ be a vector/matrix with all zeros, $\mathbf{I}$ be an identity matrix. Let $\mathrm{Tr}(\mathbf{A})$ be the trace of matrix $\mathbf{A}$. Let $\mathbf{a}_i$ and $\mathbf{a}^j$ be the i-th column vector and j-th row vector of matrix $\mathbf{A}$ respectively. Let $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}\mathbf{B}^\top)$ be the inner product of $\mathbf{A}$ and $\mathbf{B}$, and $\|\mathbf{v}\|_p$ be the ℓ_p-norm of a vector $\mathbf{v}$. Then, the Frobenius norm of an arbitrary matrix $\mathbf{A}$ is defined as $\|\mathbf{A}\|_F = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$. The ℓ_2-norm of a vector $\mathbf{a}$ is denoted as $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}}$ and the $\ell_{2,1}$-norm of matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is denoted as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m a_{ij}^2} = \sum_{i=1}^n \|\mathbf{a}^i\|_2$, where a_{ij} is the (i,j)-th element of $\mathbf{A}$ and $\mathbf{a}^i$ is the i-th row vector of $\mathbf{A}$. For analytical consistency, the $\ell_{2,0}$-norm of a matrix $\mathbf{A}$ is denoted as the number of the nonzero rows of $\mathbf{A}$. For any convex function $f(\mathbf{A})$, let $\partial f(\mathbf{A})/\partial \mathbf{A}$ denote its subdifferential at $\mathbf{A}$.We denote $\mathcal{G}$ as a weighted graph with a vertex set $\mathcal{X}$ and an affinity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ constructed on $\mathcal{X}$. The (unnormalized) Laplacian matrix associated with $\mathcal{G}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a degree matrix with $\mathbf{D}(i,i) = \sum_j S(i,j)$.

2.1 Early Active Learning

We first revisit the early active learning algorithm. Given a set of unlabeled samples $\mathbf{X} \in \mathbb{R}^{d\times n}$, the task of active learning is to select a subset of $m < n$ most representative samples $\mathbf{V} \in \mathbb{R}^{d\times m}$. Then, the selected samples are queried labeling for supervised learning. The labeled subset of data is expected to maximize the potential performance of the supervised learning in the early stage of experiment, when the available resource for labeling data is limited, i.e. only a small number of data can be labeled for supervised learning. Generally, we can define the optimization problem of early active learning as follows:

$$\min_{\mathbf{V},\mathbf{A}} \mathbf{R}(\mathbf{X},\mathbf{V},\mathbf{A}) + \alpha \boldsymbol{\Omega}(\mathbf{A}), \ s.t. \ \mathbf{V} \subset \mathbf{X}, \ |\mathbf{V}| = m. \tag{1}$$

where $\mathbf{V}$ is a subset of $\mathbf{X}$, $\mathbf{A}$ is a transformation matrix. In Eq. (1), the first term $\mathbf{R}(\cdot)$ is the reconstruction loss, the second term $\boldsymbol{\Omega}(\cdot)$ is the regularization term and $\alpha > 0$ is a leverage parameter.The major purpose of early active learning is to select a subset $\mathbf{V} \subset \mathbf{X}$ with size $m < n$ that can best represent the whole data $\mathbf{X}$ through the linear transformation matrix $\mathbf{A}$. The selected samples are therefore considered to be the most representative.

In [27], an early active learning via a Transduction Experimental Design algorithm (TED) is proposed with the aim of finding the subset $\mathbf{V} \subset \mathbf{X}$ and a project matrix $\mathbf{A}$ that minimizes the least squared reconstruction error:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{A}} \sum_{i=1}^{n} (\|\mathbf{x}_i - \mathbf{V}\mathbf{a}_i\|_2^2 + \alpha \|\mathbf{a}_i\|_2^2) \\ &\quad s.t. \ \ \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m\times n}, \ \mathbf{V} \subset \mathbf{X}, \ |\mathbf{V}| = m. \end{aligned} \tag{2}$$

where $\mathbf{V}\mathbf{a}_i$ is the representation item of $\mathbf{x}_i$. However, Eq. (2) is an NP-hard problem to solve, thus an approximate solution by a sequential optimization problem is proposed in [27].

2.2 Early Active Learning with Pairwise Constraint

In this work, we focus on early active learning in the person re-id problem. As mentioned previously, person re-id is formed as an image *retrieval* task which aims to re-identify the same person across non-overlapping camera views given a probe image of the person. The analysis of pairwise relationships of images in different camera views is therefore required. For this purpose, we introduce a pairwise constraint to early active learning:

$$\boldsymbol{\Psi}_{\mathbf{V}}(\mathbf{A}) = \sum_{i,j=1}^{n} \|\mathbf{V}\mathbf{a}_i - \mathbf{V}\mathbf{a}_j\|_2^2 S_{\mathbf{V}}(i,j), \tag{3}$$

where $\mathbf{V}\mathbf{a}_i$ is the representation item of $\mathbf{x}_i$ and $S_{\mathbf{V}}(i,j)$ is the (i,j)-th element of similarity matrix $\mathbf{S}$. It is the similarity between the i-th and the j-th representations. In this work we define $S_{\mathbf{V}}(i,j)$ as a Gaussian similarity:

$$S_{\mathbf{V}}(i,j) = \begin{cases} \exp(-\frac{\|\mathbf{V}\mathbf{a}_i - \mathbf{V}\mathbf{a}_j\|^2}{\sigma^2}), & \textit{if } \mathbf{V}\mathbf{a}_i \in \mathcal{N}_k(\mathbf{V}\mathbf{a}_j) \textit{ and } \mathbf{V}\mathbf{a}_j \in \mathcal{N}_k(\mathbf{V}\mathbf{a}_i) \\ 0 & , \textit{otherwise}, \end{cases} \tag{4}$$

where $\mathcal{N}_k(\mathbf{x})$ denotes the set of k-nearest neighbors of $\mathbf{x}$. We can then reformulate the pairwise constraint in Eq. (3) by inducing a Laplacian matrix:

$$\boldsymbol{\Psi}_{\mathbf{V}}(\mathbf{A}) = \sum_{i,j=1}^{n} \|\mathbf{V}\mathbf{a}_i - \mathbf{V}\mathbf{a}_j\|_2^2 S_{\mathbf{V}}(i,j) = \mathrm{Tr}((\mathbf{V}\mathbf{A})\mathbf{L}_{\mathbf{V}}(\mathbf{V}\mathbf{A})^T), \tag{5}$$

where $\mathbf{L}_{\mathbf{V}} = \mathbf{D} - \mathbf{S}_{\mathbf{V}}$ is the Laplacian matrix and $\mathbf{D}$ is the degree matrix with each element $\mathbf{D}_{ii} = \sum_j S_{\mathbf{V}}(i,j)$. As discussed in [9], minimizing the pairwise constraint will force the similar representations to be close to each other. Following the assumption that visually similar images of a person have a high probability of sharing the similar representation features in re-id [9], this will make early active learning schema more suitable for re-id applications.

After introducing the pairwise constraint, the early active learning for person re-identification can be formulated as:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{A}} \mathbf{R}(\mathbf{X},\mathbf{V},\mathbf{A}) + \alpha\boldsymbol{\Omega}(\mathbf{A}) + \beta\boldsymbol{\Psi}_{\mathbf{V}}(\mathbf{A}) \\ &\quad s.t. \quad \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m\times n}, \mathbf{V} \subset \mathbf{X}, \ |\mathbf{V}| = m. \end{aligned} \tag{6}$$

where $\alpha > 0$ and $\beta > 0$ are leverage parameters of regularization terms. After substituting Eq. (2) and Eq. (5) into Eq. (6) we obtain:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{A}} \sum_{i=1}^{n} (\|\mathbf{x}_i - \mathbf{V}\mathbf{a}_i\|_2^2 + \alpha\|\mathbf{a}_i\|_2^2) + \beta\mathrm{Tr}((\mathbf{V}\mathbf{A})\mathbf{L}_{\mathbf{V}}(\mathbf{V}\mathbf{A})^T) \\ &\quad s.t. \quad \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m\times n}, \mathbf{V} \subset \mathbf{X}, \ |\mathbf{V}| = m. \end{aligned} \tag{7}$$

Finding the optimal subset $\mathbf{V} \subset \mathbf{X}$ in Eq. (7) is NP-hard. Inspired by [18], we relax the problem to the following problem by introducing the $\ell_{2,0}$-norm for structure sparsity:

$$\begin{gathered}\min_{\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \alpha\|\mathbf{A}\|_{2,0} + \beta \mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T) \\ s.t. \quad \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{n\times n}, \ \ \|\mathbf{A}\|_{2,0} = m. \end{gathered} \tag{8}$$

However, the $\ell_{2,0}$-norm makes Eq. (8) a non-convex problem. At the same time, the least squared loss used in Eq. (8) is sensitive to the outliers [18], which makes the algorithm not robust.

We note that in previous researches [17,18,26], the $\ell_{2,1}$-norm is used instead of the $\ell_{2,0}$-norm. It is shown in [18] that the $\ell_{2,1}$-norm is the minimum convex hull of the $\ell_{2,0}$-norm when row-sparsity is required. In other words, minimization of $\|\mathbf{A}\|_{2,1}$ will achieve the same result as $\|\mathbf{A}\|_{2,0}$ when $\mathbf{A}$ is row-sparse. As analyzed in [18,30], the $\ell_{2,1}$-norm can suppress the effect of outlying samples. We therefore reformulate Eq. (8) as a relaxed convex optimization problem:

$$\min_{\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_{2,1} + \alpha\|\mathbf{A}\|_{2,1} + \beta \mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T). \tag{9}$$

In Eq. (9), we adopt the $\ell_{2,1}$-norm instead of both the least square reconstruction loss term and the $\ell_{2,0}$-norm structure sparsity term for robustness and suppression of outliers. By inducing the matrix formulation, Eq. (9) is rewritten as follows:

$$\min_{\mathbf{A}} \|(\mathbf{X} - \mathbf{X}\mathbf{A})^T\|_{2,1} + \alpha\|\mathbf{A}\|_{2,1} + \beta \mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T). \tag{10}$$

After obtaining the optimal solution of $\mathbf{A}$, the importances of samples can be ranked by sorting the absolute row-sum values of $\mathbf{A}$ in the decreasing order. A subset of the representative samples then can be selected corresponding to the top m largest values and query labeling.

Kernelization The proposed algorithm can be extended to the kernel version for non-linear high dimensional space. We define $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathcal{H}$ as a mapping from the Euclidian space to a Reproducing Kernel Hilbert Space (RKHS) as $\mathcal{H}$. It can be induced by a kernel function $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \boldsymbol{\Phi}(\mathbf{x})^T\boldsymbol{\Phi}(\mathbf{y})$. Then we can project $\mathbf{X}$ to RKHS space as $\boldsymbol{\Phi}(\mathbf{X}) = [\boldsymbol{\Phi}(\mathbf{x}_1), \cdots, \boldsymbol{\Phi}(\mathbf{x}_n)]$. The proposed problem thus becomes:

$$\min_{\mathbf{A}} \|(\boldsymbol{\Phi}(\mathbf{X}) - \boldsymbol{\Phi}(\mathbf{X})\mathbf{A})^T\|_{2,1} + \alpha\|\mathbf{A}\|_{2,1} + \beta \mathrm{Tr}((\boldsymbol{\Phi}(\mathbf{X})\mathbf{A})\mathbf{L}_{\mathbf{X}}(\boldsymbol{\Phi}(\mathbf{X})\mathbf{A})^T). \tag{11}$$

We denote our Early Active Learning with Pairwise Constraint algorithm in Eq. (10) as EALPC and the kenerlized version of our algorithm in Eq. (11) as EALPC_K.

3 Optimization

We provide an efficient algorithm for optimizing the proposed objective function. Taking the derivative w.r.t. $\mathbf{A}$ in Eq. (10) and setting it to zero, we obtain [1]:

$$\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{P} - \mathbf{X}^T\mathbf{X}\mathbf{P} + \alpha\mathbf{Q}\mathbf{A} + \beta\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{L}_{\mathbf{X}} = \mathbf{0}, \tag{12}$$

where $\mathbf{P}$ is a diagonal matrix and its i-th diagonal element is $p_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$. $\mathbf{Q}$ is a diagonal matrix and its i-th diagonal element is $q_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. Then by setting the derivative of Eq. (12) w.r.t. $\mathbf{a}_i$ to zero for each i, we obtain:

$$p_{ii}\mathbf{X}^T\mathbf{X}\mathbf{a}_i - p_{ii}\mathbf{X}^T\mathbf{x}_i + \alpha\mathbf{Q}\mathbf{a}_i + \beta\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{L}_i = \mathbf{0}, \tag{13}$$

where $\mathbf{L}_i$ is the i-th column vector of $\mathbf{L}_{\mathbf{X}}$. It is sample to verify that $\mathbf{A}\mathbf{L}_i = l_{ii}\mathbf{a}_i + \sum_{k\neq i} l_{ki}\mathbf{a}_k$, where l_{ii} and l_{ki} are the (i,i)-th and (k,i)-th element of $\mathbf{L}_{\mathbf{X}}$ respectively and $\mathbf{a}_k$ is the k-th column vector of $\mathbf{A}$. Therefore, the optimal solution $\mathbf{a}_i^*$ can be calculated by the closed form solution:

$$\mathbf{a}_i^* = (p_{ii}\mathbf{X}^T\mathbf{X} + \alpha\mathbf{Q} + \beta\mathbf{X}^T\mathbf{X}l_{ii})^{-1}(p_{ii}\mathbf{X}^T\mathbf{x}_i - \beta\mathbf{X}^T\mathbf{X}\sum_{k\neq i}\mathbf{a}_k l_{ki}). \tag{14}$$

In Eq. (12), $\mathbf{P}$ and $\mathbf{Q}$ are dependent on $\mathbf{A}$, thus they also need to be determined in each iteration. We propose an iterative algorithm to solve this problem. The detailed algorithm is described in Algorithm 1. In the next section, we will prove that Algorithm 1 converges to the global optimal solution of Eq. (10).

4 Convergence Analysis

We first introduce a lemma proposed in [17]:

Lemma 1. *For any arbitrary vector* $\mathbf{m}$ *and* $\mathbf{n}$ *there is*

$$\|\mathbf{m}\|_2 - \frac{\|\mathbf{m}\|_2^2}{2\|\mathbf{n}\|_2} \leq \|\mathbf{n}\|_2 - \frac{\|\mathbf{n}\|_2^2}{2\|\mathbf{n}\|_2}. \tag{15}$$

Next, in the following theorem we prove the convergence of our algorithm:

Theorem 1. *Algorithm 1 monotonically decreases the objective function value of Eq. (10) in each iteration.*

[1] In practice, when $\mathbf{x}_i - \mathbf{X}\mathbf{a}_i = 0$, p_{ii} can be regularized as $p_{ii} = \frac{1}{2\sqrt{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \eta}}$. Similarly when $\mathbf{a}_i = \mathbf{0}$, we set $q_{ii} = \frac{1}{2\sqrt{\|\mathbf{a}^i\|_2^2 + \eta}}$. η is a very small constant. It can be verified that when $\eta \to 0$ the problem with η reduces to the original problem in Eq. (12).

Algorithm 1: Algorithm for solving problem in Eq. (10)

```
  Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, parameters $\alpha$ and $\beta$.
1 Initialize $\mathbf{A} \in \mathbb{R}^{n\times n}$.
2 while not converge do
3     Compute the diagonal matrix $\mathbf{P}$, where the $i$-th diagonal element is
      $p_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$.
4     Compute the diagonal matrix $\mathbf{Q}$, where the $i$-th diagonal element is
      $q_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$.
5     Update $\mathbf{A}$ by each column $\mathbf{a}_i$ as in Eq. (14):
      $\mathbf{a}_i^* = (p_{ii}\mathbf{X}^T\mathbf{X} + \alpha\mathbf{Q} + \beta\mathbf{X}^T\mathbf{X}l_{ii})^{-1}(p_{ii}\mathbf{X}^T\mathbf{x}_i - \beta\mathbf{X}^T\mathbf{X}\sum_{k\neq i}\mathbf{a}_k l_{ki}).$
6 end
  Output: The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.
```

Proof. Suppose in an iteration the updated $\mathbf{A}$ is $\mathbf{A}^+$. According to Step 5 in Algorithm 1 we know that:

$$\mathbf{A}^+ = \arg\min_{\mathbf{F}} f(\mathbf{F}), \tag{16}$$

where we denote the function

$$f(\mathbf{F}) = \mathrm{Tr}((\mathbf{X} - \mathbf{X}\mathbf{F})\mathbf{P}(\mathbf{X} - \mathbf{X}\mathbf{F})^T) + \alpha\mathrm{Tr}(\mathbf{F}\mathbf{Q}\mathbf{F}^T) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{F})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{F})^T).$$

Thus, in each iteration when updating $\mathbf{A}$ to $\mathbf{A}^+$ we have

$$\begin{aligned}&\mathrm{Tr}((\mathbf{X} - \mathbf{X}\mathbf{A}^+)\mathbf{P}(\mathbf{X} - \mathbf{X}\mathbf{A}^+)^T) + \alpha\mathrm{Tr}((\mathbf{A}^+)\mathbf{Q}(\mathbf{A}^+)^T) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A}^+)\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A}^+)^T)\\ &\leq \mathrm{Tr}((\mathbf{X} - \mathbf{X}\mathbf{A})\mathbf{P}(\mathbf{X} - \mathbf{X}\mathbf{A})^T) + \alpha\mathrm{Tr}(\mathbf{A}\mathbf{Q}\mathbf{A}^T) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T).\end{aligned} \tag{17}$$

According to the definition of $\mathbf{P}$ and $\mathbf{Q}$, we thus obtain:

$$\begin{aligned}&\sum_{i=1}^{n}\left(\frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i^+\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2} + \alpha\frac{\|\mathbf{a}^{i+}\|_2^2}{2\|\mathbf{a}^i\|_2}\right) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A}^+)\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A}^+)^T)\\ &\leq \sum_{i=1}^{n}\left(\frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2} + \alpha\frac{\|\mathbf{a}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T).\end{aligned} \tag{18}$$

Meanwhile, according to Lemma 1, we can induce the following inequalities:

$$\sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i^+\|_2 - \frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i^+\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}\right) \leq \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2 - \frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}\right), \tag{19}$$

and

$$\sum_{i=1}^{n}\left(\|\mathbf{a}^{i+}\|_2 - \frac{\|\mathbf{a}_i^+\|_2^2}{2\|\mathbf{a}_i\|_2}\right) \leq \sum_{i=1}^{n}\left(\|\mathbf{a}^i\|_2 - \frac{\|\mathbf{a}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right). \tag{20}$$

After summing Eq. (19)and Eq. (20) in the both sides of Eq. (18), we conclude that:

$$\begin{aligned}&\sum_{i=1}^{n}(\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i^{+}\|_2+\alpha\|\mathbf{a}^{i^{+}}\|_2)+\beta\mathrm{Tr}((\mathbf{X}\mathbf{A}^{+})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A}^{+})^{T})\\ &\leq\sum_{i=1}^{n}\left(\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i\|_2+\alpha\|\mathbf{a}^{i}\|_2\right)+\beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^{T}).\end{aligned} \tag{21}$$

The above inequality indicates that the objective function value of Eq. (10) monotonically decreases in Algorithm 1. □

Meanwhile, let $\partial f(\mathbf{A})/\partial\mathbf{A}=0$ is equal to solving Eq. (12), thus in convergence, $\mathbf{A}$ will satisfy Eq. (10). As Eq. (10) is a convex problem, $\mathbf{A}$ is the global optimum solution to our problem. Overall, Algorithm 1 will converge to the global optimum solution of Eq. (10).

5 Experimental Study

In the experiments, we compare our proposed EALPC algorithm with five state-of-the-art and classic active learning algorithms. After determining and labeling the most representative samples, we train the re-id models with these samples using five popular re-id algorithms. All experiments are operated on four widely referenced re-id benchmark datasets. We report the average performance of 10 trials of independent experiments on each dataset.

5.1 Datasets and Settings

Datasets We analyze performance of active learning for re-id on four widely referred benchmark datasets for person re-identification.

1. **VIPeR** [4] The VIPeR dataset contains 1,264 images of 632 persons from two non-overlapping camera views. Two images are taken for each person, each from a different camera. Variations in viewpoint and illumination conditions occur frequently in VIPeR.
2. **PRID** [5] The PRID dataset contains images of 385 individuals from two distinct cameras. Camera B records 749 persons and Camera A records 385 persons, 200 of whom are same persons.
3. **i-LID** [30] The i-LID dataset records 119 individuals captured by three different cameras in an airport terminal. It contains 476 images with large occlusions caused by luggage and viewpoint changes.
4. **CAVIAR** [2] The CAVIAR dataset contains 72 individuals captured by two cameras in a shopping mall. The number of the images is 1,220, with 10 to 20 images for each individual. The size of the images in the CAVIAR dataset varies significantly from 39 × 17 to 141 × 72.

In the experiments, we use the recently proposed Local Maximal Occurrence (LOMO) features for person image representation [12]. As in [14,20], all person images are scaled to 128 × 48 pixels. We then use the default setting in [12] to produce a 29,960 dimension feature for each image.

Active Learning Algorithms We choose five active learning algorithms and compare them with our proposed algorithm.

1. **Random** As a baseline algorithm, we randomly select samples and query labeling.
2. **K-means** We use the K-means algorithm as another baseline algorithm as in [18]. In each experiment, samples are ranked by their distances from the K cluster centers in ascending order.
3. **QUIRE** [6] Active learning by Querying Informative and Representative Examples is an algorithm which queries the most informative and representative examples for labeling using the min-max margin-based approach.
4. **TED** [27] Active learning via Transduction Experimental Design is an algorithm that selects a subset of informative samples from a candidate dataset. It formulates a regularized linear regression problem which minimizes reconstruction error.
5. **RRSS** [18] Early active learning via Robust Representation and Structured Sparsity is a early active learning algorithm. It uses the $\ell_{2,1}$-norm to introduce structured sparsity for sample selection and robustness. However, RRSS does not consider the pairwise relations in re-id. We also introduce the kernelized RRSS denoted as **RRSS_K**.
6. **EALPC** Our proposed early active learning with pairwise constraint algorithm is denoted as EALPC. We also use a kernelized version of our algorithm denoted as **EALPC_K**. For kernelization, we construct a Gaussian kernel for the candidate dataset, i.e. $\mathcal{K}(x_i, x_j) = \exp(-\alpha\|x_i - x_j\|^2)$.

To seek the optimal parameters (if any), we apply a grid search in a region of $\{10^{-4}, 10^{-3}, \cdots, 1, \cdots, 10^3, 10^4\}$ with a five-fold cross validation strategy to determine the best parameters.

Re-identification Algorithms Five state-of-the-art supervised re-id algorithms are chosen for the performance analysis of the proposed early active learning algorithms on person re-id.

1. **NFST** [28] Null Foley-Sammon Transform space learning is a re-id algorithm for learning a discriminative subspace where the training data points of each of the classes are collapsed to a single point.
2. **KCCA** [14] Kernel Canonical Correlation Analysis algorithm seeks a common subspace between the proposed images extracted from disjoint cameras and projects them into a new space.
3. **XQDA** [12] Cross-view Quadratic Discriminant Analysis learns a discriminant low dimensional subspace by cross-view quadratic discriminant analysis for metric learning.
4. **kLFDA** [24] Kernelized Local Fisher Discriminant Classifier is a closed form method that uses a kernelized method to handle large dimensional feature vectors while maximizing a Fischer optimization criterion.
5. **MFA** [25] Marginal Fisher Analysis method is introduced for dimensionality reduction by designing two graphs that characterize the intra-class compactness and interclass separability.

Settings We report the average performance of 10 independent trials. In each trial, we divide each dataset into two equal-sized subsets as training and test sets, with no overlapping of person identities. Following the setting in [20], we divide the probe and gallery sets for re-id as follows: for datasets recording two camera views, e.g. VIPeR and PRID, images of one view are randomly selected for the probe sets, and images from the other view are chosen for the gallery sets. For a multi-view dataset, e.g. i-LID, images of one view are randomly selected as gallery sets and others are chosen as probe images. For the training set, we apply active learning methods to select a subset of training samples and query human labeling. The supervised re-id algorithms are then trained with the labeled samples. For evaluation measurement, we evaluate the performance of re-id by Cumulative Matching Characteristic (CMC) curve, which is the most commonly used performance measure for person re-id algorithms [7,13,12]. CMC calculates the probability that there exists a candidate image in the rank k gallery set that appears to match the prob image. In the experimental study, we also report the Rank One Matching Accuracy from CMC for simplicity.

Dataset	CAVIAR					VIPeR				
Algorithm	NFST	KCCA	XQDA	kLFDA	MFA	NFST	KCCA	XQDA	kLFDA	MFA
Random	23.65	23.47	21.38	27.55	25.87	26.65	23.01	27.23	22.78	23.64
K-means	26.90	25.99	22.05	27.74	27.40	27.59	26.16	27.59	23.15	24.39
TED	29.78	28.70	29,42	27.94	28.08	27.45	28.53	28.43	25.75	26.09
QUIRE	30.66	30.87	31.56	28.18	26.16	28.39	27.43	28.54	26.25	25.13
RRSS	31.87	30.69	33.57	30.95	29.01	31.56	28.54	30.71	27.34	28.04
RRSS_K	31.69	33.03	35.56	31.41	31.13	31.61	28.73	31.46	28.51	29.40
EALPC	34.12	33.57	37.45	33.09	31.16	32.61	29.45	31.82	28.54	29.56
EALPC_K	**35.00**	**35.20**	**38.75***	**33.29**	**31.91**	**33.66**	**30.44**	**34.29***	**29.18**	**30.03**

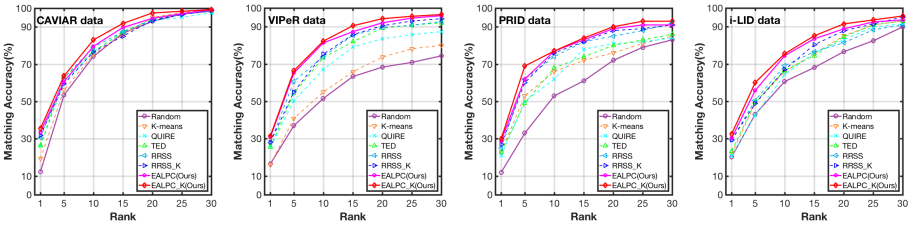
Dataset	PRID					iLIDS				
Algorithm	NFST	KCCA	XQDA	kLFDA	MFA	NFST	KCCA	XQDA	kLFDA	MFA
Random	24.49	25.47	24.00	23.50	20.00	25.96	23.40	25.00	23.35	25.00
K-means	26.16	27.54	27.01	24.70	21.20	27.02	23.94	27.00	25.57	25.20
TED	27.72	27.71	29.32	24.33	22.11	29.15	25.33	28.13	27.33	29.20
QUIRE	27.24	26.90	29.33	24.40	22.50	28.72	25.74	28.03	29.48	30.20
RRSS	29.21	28.44	30.00	25.09	23.97	28.11	27.66	30.82	30.08	30.55
RRSS_K	30.33	29.03	31.05	25.30	24.10	29.17	27.37	32.00	30.30	31.10
EALPC	32.22	30.63	31.03	25.90	25.60	29.26	27.66	32.34	30.43	31.60
EALPC_K	**32.70**	**31.50**	**33.40***	**26.06**	**25.70**	**31.19**	**28.72**	**34.00***	**31.60**	**32.47**

Table 1. Rank One Matching Accuracy(%) on four benchmarks. Percentage of selected instances for labeling is 20% of all samples. Each column is an active learning algorithm and each row is a re-id algorithm. The best result of each re-id algorithm is marked in bold numbers. The best result of the algorithms overall is marked with an asterisk($*$).

5.2 Experimental Result Analysis

Performance of Re-id We illustrate the performance of the active learning algorithms for re-id application in Table 1. In Table 1, each row corresponds to an active learning algorithm, and each column corresponds to a supervised re-id method. On each benchmark dataset, we select 20% of training samples via active learning algorithms and query labeling. The labeled subsets of samples are then adopted by supervised re-id algorithms for training models. We report the rank one matching accuracy in Table 1.

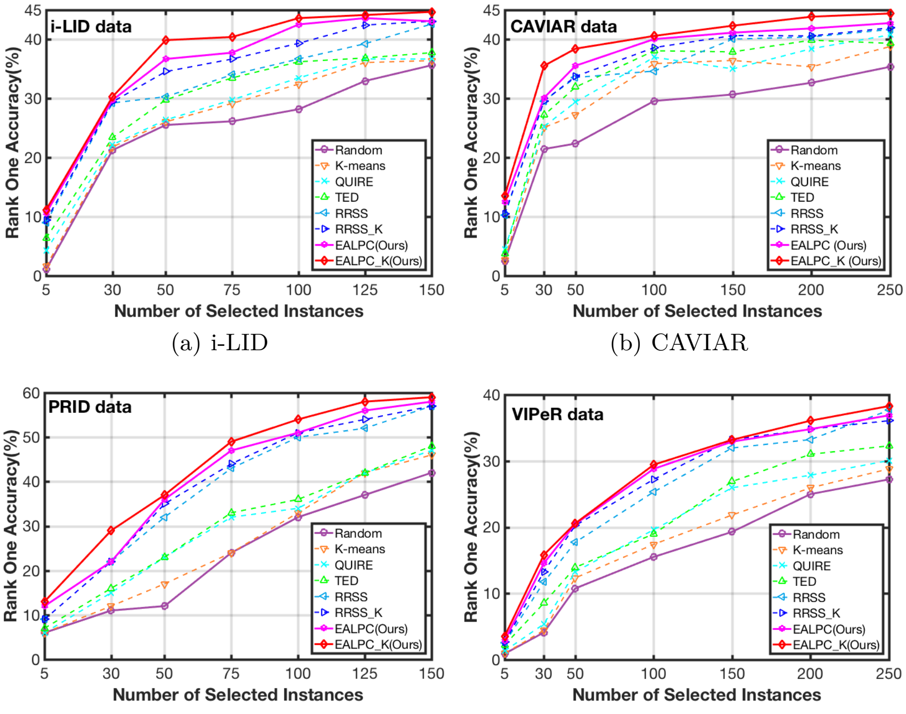
As shown in Table 1, we observe that: 1) All active learning algorithms perform better than Random selection. This indicates that active learning algorithms can select useful samples to improve the performance of re-id. 2) Our algorithms consistently outperform the other active learning algorithms. The table also confirms that our algorithms are better than the RRSS and TED method by around 5% on rank one matching accuracy. RRSS and TED have a similar optimization target to our algorithm but without pairwise constraint. This implies that our method is much suitable for re-id applications as a result of introducing the pairwise constraint. 3) The performance of the kernelized methods is better than the performance of the linear methods with our algorithm. This is consistent with the mathematical analysis in [18] that kernelization produces more discriminative representation by mapping data into high-dimensional feature space. 4) The active learning algorithms with XQDA method for report better rank one matching accuracy than those with LOMO features.



(a) CAVIAR (b) VIPeR (c) PRID (d) i-LID

Fig. 2. CMC Performance Comparison of Active Learning algorithms. XQDA is chosen as the re-id algorithm. The percentage of selected samples is set to 10% of all samples.

In Figure 2, we illustrate the performance via CMC curves of active learning methods with XQDA as the re-id algorithm. The percentage of the labeled training sample is set to only 10% to present a more challenging task. We choose XQDA as it returned the best re-id results in the previous experiments. As shown in Fig. 2, we can observe that: 1) Our algorithms outperforms other algorithms consistently on all four benchmark datasets. 2) Compared to the results in Table 1, all algorithms suffer a decrease in the rank one matching accuracy when the percentage of labeled samples is halved from 20% to 10%. However, our algorithm only decreases around by 5% on rank one matching accuracy whereas the accuracy of others, e.g. Random and K-means, reduces approximately 10%. This indicates that our algorithm is more robust. 3) The matching accuracy of our algorithm is the only one to reach 90% with rank 15 on CAVIAR and VIPeR, and the only one to reach 90% on rank 20 on PRID and i-LID. This implies that our algorithm is more effective on re-id.



(a) i-LID

(b) CAVIAR

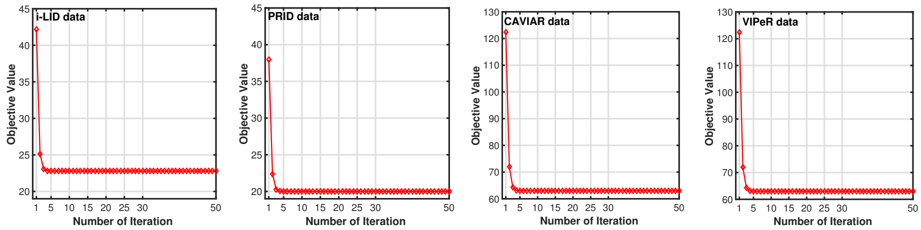
(c) PRID

(d) VIPeR

Fig. 3. Rank One Matching Accuracy(%) w.r.t. Number of Selected Instances. We use XQDA as the re-id algorithm and train it with varying numbers of samples selected by the active learning methods.

Effects on the Number of Selected Instances Figure 3 illustrates the performance of re-id when the number of instances that selected by active learning methods varies. As displayed in Fig. 3, we observe that: 1) Generally, rank one matching accuracy of all re-id algorithms increases gradually when the number of selected instances increases. 2) All active learning methods report better performances than Random selection. This indicates that active learning algorithms can improve the performance of re-id applications. 3) Our algorithm consistently performs better than the other active learning algorithms when the number of selected instance increases. More specifically, for our algorithm, kernelized methods is better than the linear methods.

Convergence In Figure 4, we draw the objective value of the first 50 iterations of our algorithm on benchmark datasets. In the experiments, we fix the leverage parameters as $\alpha = 0.1$ and $\beta = 1$ and set the percentage of selected samples to 20%. As shown in Fig. 4, the object values of our algorithm decrease dramatically and barely change after the first five iterations on all the benchmark datasets. This indicates that our algorithm converges very rapidly on all the datasets, which is consistent with our theoretical analysis of convergence.



(a) i-LID (b) PRID (c) CAVIAR (d) VIPeR

Fig. 4. Convergence Analysis of EALPC on Benchmark Datasets. The parameters are set as $\alpha = 0.1$ and $\beta = 1$. The percentage of selected samples is 20%.

6 Conclusion

In this work, we have proposed a novel early active learning algorithm with a pairwise constraint for person re-identification. The proposed method is designed for the early stage of supervised re-id experiments when there are limited labor resources for labeling data. Our algorithm introduces a pairwise constant for analyzing graph structures specifically for re-identification. A closed form solution is provided to efficiently weight and select the candidate samples. Extensive experimental studies on four benchmark datasets validate the effectiveness of the proposed algorithm. The experimental results demonstrate that our methods achieve encouraging performance against the state-of-the art algorithms in the filed of early active learning for person re-identification. In future work, our algorithm can be applied to other applications that consider the pairwise relatedness, such as in social network analysis, etc.

Acknowledgements This work was partially supported by the Data to Decisions Cooperative Research Centre www.d2dcrc.com.au and partially supported by the National Science Foundation under Grant No. IIS-1638429.

References

1. Balcan, M.F., Broder, A., Zhang, T.: Margin based active learning. In: International Conference on Computational Learning Theory. pp. 35–50. Springer (2007)
2. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference (BMVC) (2011)
3. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning 28(2), 133–168 (1997)
4. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS). vol. 3 (2007)
5. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis. pp. 91–102. Springer (2011)
6. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. In: Advances in Neural Information Processing Systems (NIPS). pp. 892–900 (2010)
7. Karanam, S., Li, Y., Radke, R.J.: Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4516–4524 (2015)
8. Karanam, S., Li, Y., Radke, R.J.: Sparse re-id: Block sparsity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 33–40 (2015)
9. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Person re-identification by unsupervised l 1 graph learning. In: European Conference on Computer Vision. pp. 178–195. Springer (2016)
10. Kodirov, E., Xiang, T., Gong, S.: Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In: Proceedings of the British Machine Vision Conference (BMVC). vol. 3, p. 8 (2015)
11. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Proceedings of the Eleventh International Conference on Machine Learning. pp. 148–156 (1994)
12. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2197–2206 (2015)
13. Lisanti, G., Masi, I., Bagdanov, A.D., Del Bimbo, A.: Person re-identification by iterative re-weighted sparse ranking. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(8), 1629–1642 (2015)
14. Lisanti, G., Masi, I., Del Bimbo, A.: Matching people across camera views using kernel canonical correlation analysis. In: Proceedings of the International Conference on Distributed Smart Cameras. p. 10. ACM (2014)
15. Ma, A.J., Li, P.: Semi-supervised ranking for re-identification with few labeled image pairs. In: Asian Conference on Computer Vision. pp. 598–613. Springer (2014)
16. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proceedings of the Twenty-first International Conference on Machine Learning. ACM (2004)
17. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint l2, 1-norms minimization. In: Advances in Neural Information Processing Systems (NIPS). pp. 1813–1821 (2010)

18. Nie, F., Wang, H., Huang, H., Ding, C.H.: Early active learning via robust representation and structured sparsity. In: International Joint Conference on Artificial Intelligence (IJCAI) (2013)
19. Nie, F., Xu, D., Li, X.: Initialization independent clustering with actively self-training method. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(1), 17–27 (2012)
20. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1306–1315 (2016)
21. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 287–294. ACM (1992)
22. Twomey, N., Diethe, T., Flach, P.: Bayesian active learning with evidence-based instance selection. In: Workshop on Learning over Multiple Contexts, European Conference on Machine Learning (ECMLâĂŹ15) (2015)
23. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1249–1258 (2016)
24. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: European Conference on Computer Vision. pp. 1–16. Springer (2014)
25. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(1) (2007)
26. Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X.: l2, 1-norm regularized discriminative feature selection for unsupervised learning. In: International Joint Conference on Artificial Intelligence (IJCAI) (2011)
27. Yu, K., Bi, J., Tresp, V.: Active learning via transductive experimental design. In: Proceedings of the 23rd international conference on Machine learning. pp. 1081–1088. ACM (2006)
28. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1239–1248 (2016)
29. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
30. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. IEEE Transactions on Image Processing 20(5), 1327–1336 (2011)