Fast Person Re-identification via Cross-camera Semantic Binary Transformation

Jiaxin Chen[†‡], Yunhong Wang[†‡*], Jie Qin[†‡], Li Liu[§♮] and Ling Shao[♭]
[†]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, China
[‡]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[§]Malong Technologies Co., Ltd, [♭]School of Computing Sciences, University of East Anglia, U.K.
chenjiaxinX@gmail.com, yhwang@buaa.edu.cn, qinjiebuaa@gmail.com
li.liu@malongtech.cn, ling.shao@ieee.org

Abstract

Numerous methods have been proposed for person re-identification, most of which however neglect the matching efficiency. Recently, several hashing based approaches have been developed to make re-identification more scalable for large-scale gallery sets. Despite their efficiency, these works ignore cross-camera variations, which severely deteriorate the final matching accuracy. To address the above issues, we propose a novel hashing based method for fast person re-identification, namely Cross-camera Semantic Binary Transformation (CSBT). CSBT aims to transform original high-dimensional feature vectors into compact identity-preserving binary codes. To this end, CSBT first employs a subspace projection to mitigate cross-camera variations, by maximizing intra-person similarities and inter-person discrepancies. Subsequently, a binary coding scheme is proposed via seamlessly incorporating both the semantic pairwise relationships and local affinity information. Finally, a joint learning framework is proposed for simultaneous subspace projection learning and binary coding based on discrete alternating optimization. Experimental results on four benchmarks clearly demonstrate the superiority of CSBT over the state-of-the-art methods.

1. Introduction

In the last few years, person re-identification (ReID) has attracted more and more research interest, due to its wide range of applications such as long-term tracking [44], searching people of interest (e.g. criminals or terrorists) and activity analysis [48]. This task aims to match a certain person across multiple non-overlapped cameras, which is very challenging due to the cluttered backgrounds, severe occlusions, illumination changes and pose variations.

A variety of approaches have been proposed to address the above problem [10, 20, 24, 46, 51, 53, 59, 62], by representation learning or building robust signature matching. However, most of them focused on improving the matching accuracy, but neglected to consider the re-identification efficiency. As a consequence, high computation costs and memory load are required by conventional methods, making them unable to provide timely responses, especially when dealing with large-scale gallery sets. Meanwhile, recently, there has been an explosive growth of wearable and mobile devices with limited computation capability. It is therefore highly desirable to develop a re-identification system that can quickly retrieve the target person from numerous gallery images with low memory load and fast speed.

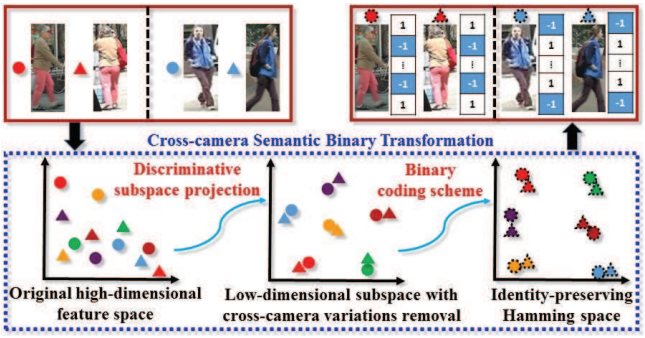


Figure 1. Illustration of the proposed framework. Different colors (shapes) indicate different person identities (cameras).

Recently, hashing has emerged as a promising way for large-scale data processing [9, 38], and has a wide range of applications such as action recognition [27, 36] and image retrieval [26, 27, 28]. Inspired by this, several supervised hashing based approaches have been developed for efficient person re-identification [4, 54, 60]. These methods attempt to build discriminative binary vectors, and subsequently construct identity-preserving hash functions. By virtue of the learned hash functions, original high-dimensional feature vectors are transformed into short binary codes, which can be stored efficiently. More importantly, very fast matching could be accomplished by calculating the Hamming dis-

*Yunhong Wang is the corresponding author.

tance. However, these methods concentrate on learning discriminative binary codes. The intrinsic *cross-camera variations*, i.e., large intra-person disturbances and small inter-person discrepancies, in the raw data are not considered, which may severely deteriorate the matching accuracy.

To deal with the aforementioned drawbacks of existing works, we propose a novel hashing based framework, namely Cross-camera Semantic Binary Transformation (CSBT), for fast person re-identification. As illustrated in Fig. 1, CSBT aims to transform original high-dimensional feature vectors into compact identity-preserving binary codes. To that end, we first employ a discriminative projection to alleviate the cross-camera variations, by minimizing intra-person distances and maximizing inter-person discrepancies in the projected subspace. Subsequently, a binary coding scheme is proposed to learn a set of binary vectors and a hash function. Specifically, the binary vectors are constructed by simultaneously preserving the semantic *pairwise relationships* (a pair of persons have the 'same' or 'different' identities) and local affinity information embedded in the subspace. Inspired by [39], a discrete learning procedure is developed to further guarantee the quality of generated binary codes. A hash function is finally built by fitting the projected feature vectors and corresponding binary codes. In order to avoid correlated code bits, orthogonal constraints are further introduced into the hash mapping, which can be efficiently solved. By adopting the joint subspace projection learning and binary coding, we can obtain a binary transformation that is more robust to cross-camera variations.

The main contributions of this paper are three-fold:

(1) We propose a novel binary transformation framework (CSBT) for fast person re-identification. This framework jointly learns a subspace projection and a binary coding scheme, which can seamlessly alleviate cross-camera variations in raw data and generate high quality binary codes. As a consequence, we can construct a more robust binary transformation to improve the matching accuracy with guaranteed re-identification efficiency.

(2) A new binary coding scheme is proposed by simultaneously incorporating the semantic pairwise relationships and local affinity information embedded in the subspace. A discrete alternating optimization algorithm is further introduced, by virtue of which we can obtain representative binary codes to preserve person identities across cameras.

(3) We extensively conduct experiments to evaluate the performance of the proposed method. The results demonstrate the efficiency of the proposed method, especially on large-scale gallery sets.

2. Related Work

In the literature, most of existing works on person re-identification can be divided into three categories: building robust appearance [10, 20, 59] and spatial-temporal representations [5, 25, 31, 47], developing discriminative similarity metrics [6, 16, 20, 35, 62], or designing more reliable matching strategies [5, 32]. Most of these works concentrate on promoting matching rates, and ignore their scalability to large volumes of gallery images during test. As a result, both the time and memory costs of existing re-identification methods will grow sharply and become unaffordable, as the number of gallery images increases.

Recently, several hashing based approaches have been proposed to address fast person re-identification. In [54], a deep regularized similarity comparison hashing method (DRSCH) was developed by incorporating regularized triplet-based formulation and bit-scalable hashing generation into a deep convolutional neural network. DRSCH together with the Deep Semantic Ranking Hashing (DSRH) [56], which preserves multi-level semantic similarity between multi-label images, were then evaluated in the context of person re-identification. In [60], the cross-view binary identities (CBI) were learned by constructing two sets of hash functions, through minimising the intra-person Hamming distance and maximising the cross-covariances.

Additionally, a few existing methods for cross-view hashing were also applied to person re-identification, including: Cross-Modality Similarity Sensitive Hashing (CMSSH) [2], Cross-View Hashing (CVH) [17], Predictable Dual-view Hashing (PDH) [37], Collective Matrix Factorisation Hashing (CMFH) [8] and Canonical Correlation Analysis based hashing (CCA) [60]. These approaches were originally designed to address binary coding for multi-modal data, through exploiting the correlations between distinct sources of representations. By taking each camera view as one modality, they can be straightforwardly applied to person re-identification.

Despite the promising efficiency achieved by existing hashing based methods, all of them neglect to deal with intrinsic cross-view variations in raw data. Meanwhile, DRSCH requires a huge amount of labeled training data, which is not easy to acquire in practice. And CBI can only train a model between a pair of camera views, which is not flexible to scenarios with multiple camera views.

3. The Proposed Framework

Suppose that N D-dimensional training samples $\{\mathbf{x}_i\}_{i=1}^N$ together with corresponding labels $\{\mathbf{y}_i\}_{i=1}^N$ are available, where $\mathbf{y}_i$ indicates the person identity (ID) of $\mathbf{x}_i$. We treat $(\mathbf{x}_i, \mathbf{x}_j)$ as a positive sample pair if $\mathbf{y}_i = \mathbf{y}_j$, and a negative pair otherwise. Our target is to transform high dimensional feature vectors $\{\mathbf{x}_i\}_{i=1}^N$ into a set of binary codes $\{\mathbf{b}_i\}_{i=1}^N$ with L bits, based on which a hash function $H : \mathbb{R}^D \rightarrow \{-1, 1\}^L$ can be trained via regression.

Traditional works learn $\{\mathbf{b}_i\}_{i=1}^N$ by exploiting either the intrinsic local affinity of raw data $\{\mathbf{x}_i\}_{i=1}^N$, or the semantic

similarities from labels $\{\mathbf{y}_i\}_{i=1}^N$. In our work, we attempt to combine these two kinds of information. However, in person re-identification, the local affinity information in the original feature space is too noisy due to cross-camera variations. To address this problem, we introduce a discriminative subspace, where intra-class distances are forced to be smaller than inter-class distances. Through this way, we can obtain two advantages: 1) The binary transformation learned from the embedded subspace can be more robust to cross-camera variations; 2) The local affinity of embedded data will contain more useful information to train discriminative binary codes.

Based on the aforementioned motivations, we formulate the general framework of CSBT as follows:

$$\begin{array}{ll}\min\limits_{\mathbf{B},\mathbf{P},F} & \ell_{\mathrm{ML}}(\mathbf{P}) + \beta\ell_{\mathrm{H}}(\mathbf{B},\mathbf{P}) \\ \text{s.t.} & \mathbf{b}_i = \mathrm{sgn}(F(\mathbf{P}^T\mathbf{x}_i)),\ i = 1,\cdots,N,\end{array} \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1^T;\cdots;\mathbf{x}_N^T] \in \mathbb{R}^{N\times D}$, $\mathbf{Y} = [\mathbf{y}_1^T;\cdots;\mathbf{y}_N^T] \in \mathbb{R}^N$, $\mathbf{B} = [\mathbf{b}_1^T;\cdots;\mathbf{b}_N^T] \in \{-1,1\}^{N\times L}$, $\mathbf{P} \in \mathbb{R}^{D\times d}$ is the subspace projection, ℓ_{ML} is the loss function for subspace projection learning, ℓ_{H} is the loss function for binary coding, $H = \mathrm{sgn}(F(\cdot))$ is the hash function, and β is a positive trade-off parameter. Here, $F(\cdot)$ is the linear mapping function, and $\mathrm{sgn}(\cdot)$ is the sign function.

In terms of the loss function ℓ_{ML}, we harness the principle of metric learning to train the subspace projection matrix $\mathbf{P}$, or equivalently, a Mahalanobis distance function

$$\begin{aligned} D_{\mathbf{P}}^2(\mathbf{x}_i,\mathbf{x}_j) &= \|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|^2 \\ &= (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i-\mathbf{x}_j)\end{aligned} \tag{2}$$

to measure the distance between samples, where $\mathbf{M} = \mathbf{P}\mathbf{P}^T$.

We adopt the log-logistic loss function as in [20, 62], which can provide a soft margin to separate different classes, and is particularly useful for classification problems. Specifically, we utilize the following loss function

$$\ell_{\mathrm{ML}}(\mathbf{P}) = \sum w_{i,j}\log(1 + e^{y_{i,j}(D_{\mathbf{P}}^2(\mathbf{x}_i,\mathbf{x}_j)-\mu)}), \tag{3}$$

where

$$y_{i,j} = \begin{cases} 1, \text{ if } \mathbf{y}_i = \mathbf{y}_j, \\ -1, \text{ if } \mathbf{y}_i \neq \mathbf{y}_j, \end{cases} \quad w_{i,j} = \begin{cases} 1/N_p, \text{ if } \mathbf{y}_i = \mathbf{y}_j, \\ 1/N_n, \text{ if } \mathbf{y}_i \neq \mathbf{y}_j. \end{cases} \tag{4}$$

N_p and N_n are the numbers of positive and negative sample pairs, respectively. μ is a constant bias, which is applied considering that $D_{\mathbf{P}}^2$ has a lower bound of zero.

Since $\log(1+e^z)$ is monotonically increasing and $\log(1+e^{-z})$ is monotonically decreasing, we can observe that: positive sample pairs (with the same person IDs) from different cameras are pulled close, and negative sample pairs (with different person IDs) are pushed apart in the projected subspace by minimizing ℓ_{ML} shown in Eq. (3). Through this way, we expect to learn a discriminative projection $\mathbf{P}$ that can mitigate cross-camera variations.

As for the loss function ℓ_{H}, our target is to learn binary codes $\{\mathbf{b}_i\}_{i=1}^N$ of the embedded feature vectors $\{\mathbf{P}^T\mathbf{x}_i\}_{i=1}^N$, by exploiting both the semantic and local data affinity information. Concretely, we utilize the following loss function

$$\ell_{\mathrm{H}}(\mathbf{B},\mathbf{P}) = \sum w_{i,j}y_{i,j}a_{i,j}d_h(\mathbf{b}_i,\mathbf{b}_j), \tag{5}$$

where $d_h(\mathbf{b}_i,\mathbf{b}_j) = |\{k|b_{i,k} \neq b_{j,k}, 1 \leq k \leq L\}|$ indicates the Hamming distance [30], and $a_{i,j}$ encodes the semantic and local data affinity between embedded samples $\mathbf{P}^T\mathbf{x}_i$ and $\mathbf{P}^T\mathbf{x}_j$. In this paper, we define that

$$a_{i,j} = \begin{cases} 1 & \text{, if } \mathbf{y}_i = \mathbf{y}_j; \\ 1 - e^{-\frac{\|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|_2^2}{2\sigma^2}} & \text{, if } \mathbf{y}_i \neq \mathbf{y}_j. \end{cases} \tag{6}$$

When ℓ_{H} is minimized, from Eqs. (5) and (6) we have the following observations: 1) The Hamming distance between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ will be diminished/increased, if they consist a positive/negative pair; 2) For two negative sample pairs ($\mathbf{x}_i$,$\mathbf{x}_j$) and ($\mathbf{x}_i$,$\mathbf{x}_k$), if $\|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|_2 < \|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_k\|_2$, then $a_{ij} < a_{ik}$, implying that a larger weight will be imposed on maximizing the Hamming distance between binary codes of $\mathbf{x}_i$ and $\mathbf{x}_k$. As a result, $d_h(\mathbf{b}_i,\mathbf{b}_k)$ is preferred to be larger than $d_h(\mathbf{b}_i,\mathbf{b}_j)$. This indicates that the learned binary codes are forced to preserve both the semantic information and local affinity in the embedded subspace, by reducing the loss $\ell_{\mathrm{H}}(\mathbf{B},\mathbf{P})$ in (5).

Finally, with respect to $F(\cdot)$, we adopt the widely used linear mapping [39], i.e., $F(\mathbf{z}) = \mathbf{W}^T\mathbf{z}$, where $\mathbf{W} \in \mathbb{R}^{d\times L}$ is the mapping matrix. To further avoid severely correlated binary code bits, we introduce orthogonal constraints on $\mathbf{W}$, i.e., $\mathbf{W}^T\mathbf{W} = \mathbf{I}_L$ [13], where $\mathbf{I}_L$ is the identity matrix with order L. Moreover, inspired by [39], we replace $\mathbf{b}_i = \mathrm{sgn}(F(\mathbf{P}^T\mathbf{x}_i))$ by a regularization loss $\|\mathbf{b}_i - \mathrm{sgn}(F(\mathbf{P}^T\mathbf{x}_i))\|^2$, for optimization convenience. Meanwhile, by employing a regularization term on $\mathbf{P}$ and adopting matrix notations, we obtain the final formulation of the proposed framework:

$$\begin{array}{ll}\min\limits_{\mathbf{B},\mathbf{P},\mathbf{W}} & \mathcal{L}(\mathbf{B},\mathbf{P},\mathbf{W}) \\ \text{s.t.} & \mathbf{W}^T\mathbf{W} = \mathbf{I}_L, \mathbf{B} \in \{-1,1\}^{N\times L}\end{array} \tag{7}$$

where $\mathcal{L}(\mathbf{B},\mathbf{P},\mathbf{W}) = \ell_{\mathrm{ML}}(\mathbf{P}) + \beta\ell_{\mathrm{H}}(\mathbf{B},\mathbf{P}) + \frac{\gamma}{2}\|\mathbf{B} - \mathbf{X}\mathbf{P}\mathbf{W}\|_F^2 + \frac{\nu}{2}\|\mathbf{P}\|_F^2$. γ and ν are trade-off parameters.

Based on the learned $\mathbf{P}$ and $\mathbf{W}$ by solving the optimization problem (7), an unseen test data sample $\mathbf{x}$ can then be transformed into binary codes by using $\mathrm{sgn}\left(\mathbf{W}^T\mathbf{P}^T\mathbf{x}\right)$.

4. Optimization

Since (7) is a non-convex optimization problem, it is difficult to find the global optimum. In this paper, we develop

an alternating iteration algorithm to achieve a locally optimal solution. Specifically, we alternate updates of $\mathbf{B}$, $\mathbf{P}$ and $\mathbf{W}$, i.e., optimize one variable whilst fixing the rest.

B-Step. When fixing $\mathbf{P}$ and $\mathbf{W}$, based on the fact that $d_h(\mathbf{b}_i, \mathbf{b}_j) = \frac{L - \mathbf{b}_i^T \mathbf{b}_j}{2}$, problem (7) can be reformulated into

$$\min_{\mathbf{B} \in \{-1,1\}^{N \times L}} -\frac{\beta}{2} \sum s_{i,j} \mathbf{b}_i^T \mathbf{b}_j + \frac{\gamma}{2} \|\mathbf{B} - \mathbf{XPW}\|_F^2, \quad (8)$$

where $s_{i,j} = w_{i,j} y_{i,j} a_{i,j}$ encodes the semantic and data affinity correlation of the i-th and j-th samples.

Problem (8) is fundamentally NP-hard. Inspired by [39], we propose to discretely learn $\mathbf{B}$ by adopting an alternating optimization procedure. Concretely, we learn $\mathbf{B}$ sample-by-sample, i.e., optimize $\mathbf{b}_i$ whilst fixing the remaining $N-1$ samples $\{\mathbf{b}_1, \cdots, \mathbf{b}_{i-1}, \mathbf{b}_{i+1}, \cdots, \mathbf{b}_N\}$. By setting $\mathbf{z} = \mathbf{b}_i$ and using the fact $\mathbf{z}^T \mathbf{z} = L$, we attain the following results

$$-\frac{\beta}{2} \sum s_{i,j} \mathbf{b}_i^T \mathbf{b}_j = \mathbf{z}^T \left(-\frac{\beta}{2} \sum_{j \neq i} s_{i,j} \mathbf{b}_j \right) + const, \quad (9)$$

$$\|\mathbf{B} - \mathbf{XPW}\|_F^2 = -2\mathbf{z}^T \mathbf{W}^T \mathbf{P}^T \mathbf{x}_i + const. \quad (10)$$

By taking Eqs. (9) and (10) back into problem (8), we finally derive the optimization problem below

$$\min_{\mathbf{z} \in \{-1,1\}^L} \mathbf{z}^T \left(-\frac{\beta}{2} \sum_{j \neq i} s_{i,j} \mathbf{b}_j - \gamma \mathbf{W}^T \mathbf{P}^T \mathbf{x}_i \right), \quad (11)$$

which has the following closed-form solution

$$\mathbf{z} = \operatorname{sgn}\left(\tfrac{\beta}{2} \textstyle\sum_{j \neq i} s_{i,j} \mathbf{b}_j + \gamma \mathbf{W}^T \mathbf{P}^T \mathbf{x}_i \right). \quad (12)$$

From Eq. (12), we can observe the update of each sample $\mathbf{b}_i$ relies on the remaining $N-1$ binary vectors.

P-Step. By fixing $\mathbf{B}$ and $\mathbf{W}$, problem (7) turns into

$$\min_{\mathbf{P} \in \mathbb{R}^{D \times d}} \mathcal{F}(\mathbf{P}), \quad (13)$$

where $\mathcal{F}(\mathbf{P}) = \sum w_{i,j} \log(1 + e^{y_{i,j}(D_{\mathbf{P}}^2(\mathbf{x}_i, \mathbf{x}_j) - \mu)}) + \beta \sum w_{i,j} y_{i,j} a_{i,j} d_h(\mathbf{b}_i, \mathbf{b}_j) + \frac{\gamma}{2} \|\mathbf{B} - \mathbf{XPW}\|_F^2 + \frac{\nu}{2} \|\mathbf{P}\|_F^2$.

Generally, $\mathcal{F}(\mathbf{P})$ is non-convex with respect to $\mathbf{P}$. It is therefore difficult to find a global optimal solution. In this paper, we aim to derive a local optimal solution by using the *gradient descent* method. Concretely, given the point $\mathbf{P}^{(k-1)}$ at iteration $k-1$, $\mathbf{P}$ is updated by

$$\mathbf{P}^{(k)} = \mathbf{P}^{(k-1)} - \eta^{(k)} \nabla \mathcal{F}(\mathbf{P}^{(k-1)}), \quad (14)$$

where $\mathbf{P}^{(k)}$, $\eta^{(k)}$, and $\nabla \mathcal{F}(\mathbf{P}^{(k)})$ are the value of $\mathbf{P}$, the step length, and the gradient of $\mathcal{F}(\mathbf{P})$ at the k-th iteration, respectively. Here, $\nabla \mathcal{F}(\mathbf{P})$ is formulated as follows:

$$\nabla \mathcal{F}(\mathbf{P}) = \sum_{i,j} \left(g_{i,j}(\mathbf{P}) + \beta h_{i,j}(\mathbf{P}) \right) (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{P} + \gamma (\mathbf{X}^T \mathbf{XPWW}^T - \mathbf{X}^T \mathbf{BW}^T) + \nu \mathbf{P},$$

Algorithm 1 Cross-camera Binary Transformation

Input: Data matrix $\mathbf{X}$, labels $\mathbf{Y}$, the maximal iteration number $T_{\max}$, bit length L and trade-off parameters β, γ, ν.
Output: Binary codes $\mathbf{B}$, subspace projection matrix $\mathbf{P}$, and linear mapping $\mathbf{W}$.

1: **repeat**
2: **B-Step:** Update $\mathbf{B}$ by Eq. (12).
3: **P-Step:** Update $\mathbf{P}$ by Eq. (14).
4: **W-Step:** Update $\mathbf{W}$ by Eq. (19).
5: **until** *converged or reach the maximal iteration* $T_{\max}$

where $g_{i,j}(\mathbf{P}) = \frac{2 w_{i,j} y_{i,j}}{1 + e^{-y_{i,j}(D_{\mathbf{P}}^2(\mathbf{x}_i, \mathbf{x}_j) - \mu)}}$, and

$$h_{i,j}(\mathbf{P}) = \begin{cases} 0 & , \text{if } \mathbf{y}_i = \mathbf{y}_j; \\ \dfrac{w_{i,j} y_{i,j} d_h(\mathbf{b}_i, \mathbf{b}_j)}{\sigma^2 e^{\frac{D_{\mathbf{P}}^2(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}}} & , \text{if } \mathbf{y}_i \neq \mathbf{y}_j. \end{cases} \quad (15)$$

In order to guarantee the convergence of the gradient descent method depicted in Eq. (14), we choose the step length $\eta^{(k)}$ that satisfies the Wolfe conditions by using backtracking line search, according to [33].

W-Step. By fixing $\mathbf{B}$ and $\mathbf{P}$, we can rewrite (7) into

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times L}, \mathbf{W}^T \mathbf{W} = \mathbf{I}_L} \mathcal{G}(\mathbf{W}) := \tfrac{1}{2} \|\mathbf{B} - \mathbf{XPW}\|_F^2. \quad (16)$$

The above problem is a nonlinear optimization problem with orthogonal constraints. Inspired by [49], we adopt the *Crank-Nicolson-like update scheme* to find a feasible solution, due to its simplicity and computational efficiency.

Specifically, given a feasible point $\mathbf{W}^{(k-1)}$ at iteration $k-1$ and the corresponding gradient

$$\nabla \mathcal{G}(\mathbf{W}^{(k-1)}) = \mathbf{P}^T \mathbf{X}^T \mathbf{XPW}^{(k-1)} - \mathbf{P}^T \mathbf{X}^T \mathbf{B}, \quad (17)$$

a skew-symmetric matrix $\mathbf{A}^{(k)} = \nabla \mathcal{G}(\mathbf{W}^{(k-1)}) {\mathbf{W}^{(k-1)}}^T - \mathbf{W}^{(k-1)} \nabla \mathcal{G}(\mathbf{W}^{(k-1)})^T$ is firstly calculated. The new trial point $\mathbf{W}^{(k)}$is then obtained by doing curvilinear search along the path

$$\mathbf{Y}^{(k)}(\tau) = (\mathbf{I}_d + \frac{\tau}{2} \mathbf{A}^{(k)})^{-1} (\mathbf{I}_d - \frac{\tau}{2} \mathbf{A}^{(k)}) \mathbf{W}^{(k-1)}. \quad (18)$$

Similar to the **P-step**, we utilize the backtracking line search [33] to find a proper step length $\tau^{(k)}$, based on which $\mathbf{W}$ is updated by

$$\mathbf{W}^{(k)} = \mathbf{Y}^{(k)}(\tau^{(k)}). \quad (19)$$

By repeating the aforementioned procedure, we can finally obtain a feasible $\mathbf{W}$, which achieves a local optimum.

The overall solution is summarized in Algorithm 1. In the **B-step**, we can alternatively infer $\mathbf{B}$ by directly adopting $\mathbf{B} = \operatorname{sgn}(\mathbf{W}^T \mathbf{P}^T \mathbf{X})$ in the **B-step**. However, this strategy may introduce large cumulative quantization errors. In

contrast, we propose a new discrete learning method to ensure the high quality of learned **B** in our work. **W** and **P** can be obtained based on the optimal **B** iteratively. This training/testing strategy is widely adopted by recent hashing methods [29, 39].

4.1. Convergence Analysis

From Eqs. (4) and (6), we can observe that $0 < w_{i,j} \leq 1$ and $0 < a_{i,j} \leq 1\,(\forall i, j = 1, \cdots, N)$. Since $\log(1+e^x) > 0$ for any $x \in \mathbb{R}$, we can then derive that

$$\ell_{\mathrm{ML}}(\mathbf{P}) = \sum w_{i,j} \log\left(1 + e^{y_{i,j}\left(D^2_{\mathbf{P}}(\mathbf{x}_i, \mathbf{x}_j) - \mu\right)}\right) > 0.$$

Moreover, we can easily deduce the following results

$$\ell_{\mathrm{H}}(\mathbf{B}, \mathbf{P}) \geq \sum -d_h\left(\mathbf{b}_i, \mathbf{b}_j\right) \geq -N^2 L.$$

It is then straightforward to see that the objective function $\mathcal{L}(\mathbf{B}, \mathbf{P}, \mathbf{W})$ in (7) has a lower bound. On the other hand, $\mathcal{L}(\mathbf{B}, \mathbf{P}, \mathbf{W})$ consistently decreases, when iteratively conducting **B-Step**, **P-Step** and **W-Step**. We can therefore conclude that Algorithm 1 converges to a local minimum (see empirical studies in the supplementary material).

5. Experiments

In this section, we evaluate the proposed method on four datasets: VIPeR [14], CUHK01 [18], CUHK03 [19] and Market-1501 [61]. Several samples are shown in Fig. 2.

VIPeR is one of the most widely used datasets, which contains 632 pedestrians from two non-overlapping cameras. This dataset is very challenging due to the low image quality, together with large variations in illumination, poses and viewpoints. For evaluation, the single-shot setting is used in our experiments as in [62]. We follow the standard settings to randomly select $p = 316$ persons for test, and the rest 316 persons for training. This is repeated for 10 times and the averaged performance is reported.

CUHK01 includes 3,884 images of 971 pedestrians captured by two disjoint cameras, with each person having two images under each camera. Different from VIPeR, images in CUHK01 are of higher resolutions. On this dataset, both the 485/486 and 871/100 training/test settings (multi-shot) are widely used. We therefore report results for these two different partitions over 10 trials.

CUHK03 contains 13,164 images of 1,360 pedestrians under six surveillance cameras, with each person observed by two disjoint cameras and having an average of 4.8 images in each view. We follow [1, 45, 54], and use the 20 training/test splits provided in [19] with manually cropped images under the single-shot setting.

Market1501 contains 32,688 bounding boxes of 1501 identities, most of which are cropped by the Deformable Parts Model (DPM) [11]. Each person is captured by 2∼6 cameras. This dataset is the largest publicly available person re-identification dataset to date. Similar to [61], we use 12,936 images for training. During test, we utilize 3,368 images for query and 19,732 images for gallery under the single-query evaluation settings.

Figure 2. Sample images: VIPeR (left) and CUHK01 (right). Images in the same column/row belong to the same person/camera.

5.1. Experimental Setup

Image Representation. We adopt the Local Maximal Occurrence (LOMO) feature [20] for person representation. Specifically, all images are normalized to 128×64 pixels. A set of sliding windows are then generated, where both color and texture histograms are extracted. Maximal occurrences of patterns encoded by histogram bins are calculated and concatenated into a 26,960-dimensional feature vector.

Parameter Settings. In our evaluations, the dimension of subspace d and balancing parameters β, γ, ν in (7) are selected by cross-validation. The maximal iteration number $T_{\max}$ is set to 16. For computational efficiency, we employ PCA to reduce the dimension of LOMO features to 3000. Since the bit length L significantly affects the performance of hashing based approaches, we fine-tune L in the range [64,1024] with step-size 64, and choose the bit length that achieves the highest rank 1 accuracies.

Evaluation Metrics. Similar to most publications, we use the Cumulated Matching Characteristics (CMC) curve to evaluate the performance of various person re-identification methods. Since the mean average precision (mAP) is a widely used evaluation metric for hashing methods, we also report mAP when comparing CSBT with the state-of-the-art hashing approaches.

5.2. Comparison with Hashing Methods

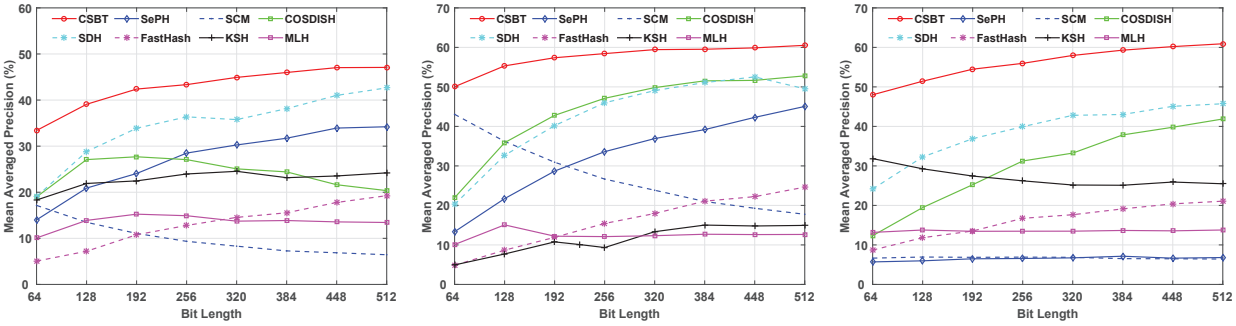
In this section, we evaluate CSBT on VIPeR, CUHK01 and CUHK03. We choose the following state-of-the-art hashing methods for comparisons: single-view hashing including MLH [34], KSH [30], FastHash [22], SDH [39], COSDISH [15], and cross-view hashing including SCM [52], SePH [23], CBI [60]. Note that [54] reported results by using two deep learning based hashing, i.e., DRSCH [54] and DSRH [56] on CUHK03. We also compare with these two methods by directly adopting results from [54].

Since the bit length significantly affects the performance of hashing methods, we demonstrate the rank 1 matching

Table 1. Rank 1 matching rate (%) with different bit length L using different hashing approaches.

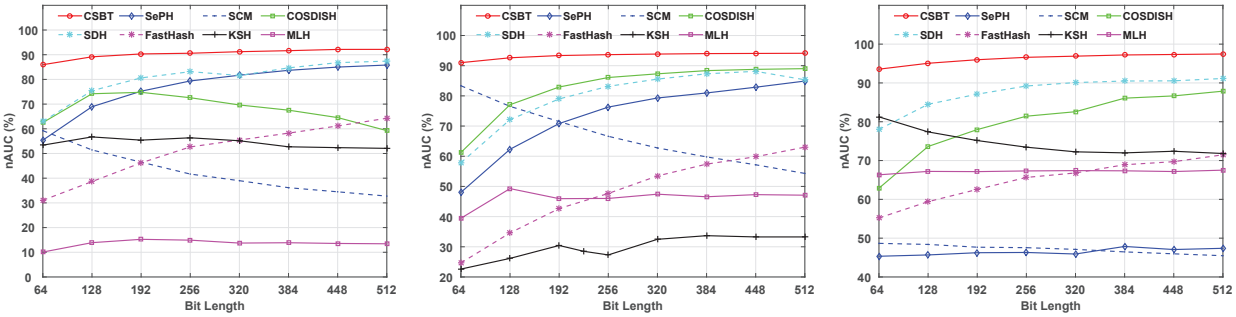
Method		Reference	VIPeR				CUHK01 (P=486)				CUHK03			
			64 bits	128 bits	256 bits	512 bits	64 bits	128 bits	256 bits	512 bits	64 bits	128 bits	256 bits	512 bits
Cross-View	CBI*	IJCAI2016 [60]	13.9	18.0	23.1	26.3	-	-	-	-	-	-	-	-
	SePH	CVPR2015 [23]	6.2	10.4	15.9	20.1	6.7	12.5	22.1	32.7	1.3	1.6	1.9	2.2
	SCM	AAAI2014 [52]	8.9	6.5	3.9	2.3	30.7	25.1	17.1	10.3	2.0	2.0	2.2	2.1
Single-View	COSDISH	AAAI2016 [15]	9.8	16.5	16.8	12.4	13.1	24.4	34.9	41.0	4.4	9.3	19.1	29.0
	SDH	CVPR2015 [39]	9.6	17.6	23.6	29.5	11.9	22.3	34.5	38.2	12.9	19.3	25.0	31.4
	DRSCH	TIP2015 [54]	-	-	-	-	-	-	-	-	22.0	18.7	-	-
	DSRH	CVPR2015 [56]	-	-	-	-	-	-	-	-	14.4	8.1	-	-
	FastHash	CVPR2014 [22]	1.5	2.4	5.7	10.3	1.7	4.1	8.7	15.4	2.6	4.9	8.6	12.1
	KSH	CVPR2012 [30]	10.6	13.6	15.9	16.5	10.6	13.6	15.9	10.2	19.1	18.0	15.3	15.0
	MLH	ICML2011 [34]	4.6	7.6	8.8	7.7	4.7	8.5	6.1	6.5	5.2	5.5	5.0	5.5
	CSBT	**Ours**	**20.3**	**24.7**	**29.5**	**33.1**	**36.2**	**42.3**	**45.5**	**48.0**	**33.1**	**36.2**	**40.3**	**46.2**

('*': The best results are adopted from [60] for each bit length. '-': The source codes or experimental results are not available.)



(a) VIPeR (b) CUHK01 (c) CUHK03

Figure 3. Comparison of mAP for state-of-the-art hashing methods by using different bit lengths.



(a) VIPeR (b) CUHK01 (c) CUHK03

Figure 4. Comparison of nAUC for state-of-the-art hashing methods by using different bit lengths.

rate and mAP across different bits in Table 1 and Fig. 3. Particularly, on the VIPeR dataset, [60] provided rank 1 matching accuracies of CBI with various bit lengths, by using various kinds of features (including the LOMO feature). To make a compact description, we only demonstrate the best results of CBI for each bit length. On the CUHK01 dataset, only the results with optimal bit lengths were reported for CBI in [60]. Since *the source code of CBI is not publicly available*, and the deep hashing methods (e.g., DRSCH and DSRH) perform poorly on small datasets, we just provide comparison results on CUHK01 by re-implementing MLH, KSH, FastHash, SDH, COSDISH, SePH and SCM using the LOMO feature. As shown in Table 1 and Table 2, even using very short bit lengths (e.g., 64 bits), the performance of CSBT is obviously superior to CBI with the optimal bit lengths. On the CUHK03 dataset, [54] reported the rank 1 matching rates of DRSCH and DSRH with 64 and 128 bits. We directly copy the results and show them in Table 1. Considering that top ranked results are usually desirable in practice besides rank 1, we additionally report the normalized Area Under the CMC Curve (nAUC) as in [60] to make a more comprehensive study. In our work, we follow [60] and summarize nAUC at top 85 ranks by using different bit lengths in Fig. 4.

From Table 1, Fig. 3 and Fig. 4, we have the following observations: 1) CSBT consistently outperforms compared hashing methods over different bit lengths. 2) As the bit length decreases, the performance of compared methods (e.g., SePH, SDH, COSDISH) drops sharply. In contrast, CSBT achieves more robust performance, since it still yields promising results even using short codes (e.g., less than 128 bits). This is due to that CSBT performs binary

Table 2. Matching rates (%), average query time (in seconds) and memory usage (in kilobytes) for storing gallery data, by comparing with state-of-the-art approaches on various datasets.

Method		Reference	VIPeR r=1	r=5	r=20	Time (s)	Mem. (KB)	CUHK03 (Labeled) r=1	r=5	r=20	Time (s)	Mem. (KB)
Hashing	CBI*	IJCAI2016 [60]	31.3	57.3	81.6	**1.4e-06**	**2.72e+01**	-	-	-	-	-
	DRSCH*	TIP2015 [54]	-	-	-	-	-	22.0	48.4	81.0	-	-
	DSRH*	CVPR2015 [56]	-	-	-	-	-	14.4	43.4	79.2	-	-
Non Hashing	DCSL	IJCAI2016 [55]	44.6	73.4	92.1	-	5.68e+03	**80.2**	**97.7**	**99.7**	-	5.18e+02
	Gated CNN	ECCV2016 [43]	37.8	-	-	-	5.68e+03	-	-	-	-	-
	EDM	ECCV2016 [40]	40.9	-	-	-	5.68e+03	61.3	-	-	-	5.18e+02
	SIR+CIR	CVPR2016 [45]	35.8	-	-	-	5.68e+03	52.2	-	-	-	5.18e+02
	JSTL	CVPR2016 [50]	38.6	-	-	-	5.68e+03	75.3	-	-	-	5.18e+02
	SCSP	CVPR2016 [3]	53.5	-	-	3.78e-02	2.96e+02	61.3	-	-	-	-
	NSL	CVPR2016 [53]	42.3	71.5	92.1	-	6.66e+04	58.9	85.6	96.3	-	2.11e+04
	KCCA+DCIA	ICCV2015 [12]	**63.9**	**78.5**	-	-	1.20e+04	-	-	-	-	-
	Improved Deep	CVPR2015 [1]	34.8	63.6	84.5	-	5.68e+03	54.8	86.2	98.5	-	5.18e+02
	Semantic	CVPR2015 [41]	31.1	68.6	**94.9**	-	-	-	-	-	-	-
	MLF	CVPR2014 [59]	29.1	52.3	79.9	-	-	-	-	-	-	-
	DeepReID	CVPR2014 [19]	-	-	-	-	5.68e+03	20.7	50.1	80.0	-	5.18e+02
	SalMatch	ICCV2013 [57]	30.2	52.3	79.2	-	-	-	-	-	-	-
	eSDC	CVPR2013 [58]	26.3	50.8	76.5	1.14+01	4.98e+05	8.8	27.0	55.1	3.45e+00	1.58e+05
	KISSME	CVPR2012 [16]	19.6	47.9	77.2	9.2e-03	8.39e+01	14.2	41.1	70.1	-	-
	SDALF	CVPR2010 [10]	20.0	38.7	65.7	3.6e+00	1.17e+03	5.6	23.5	52.0	1.22e+00	3.72e+02
	CSBT*	**Ours**	36.6	66.2	88.3	1.68e-06	3.45e+01	55.5	84.3	98.0	**4.83e-07**	**1.01e+01**

Method		Reference	CUHK01 (p=486) r=1	r=5	r=20	Time (s)	Mem. (KB)	CUHK01 (p=100) r=1	r=5	r=20	Time (s)	Mem. (KB)
Hashing	CBI*	IJCAI2016 [60]	30.6	52.9	69.1	-	-	34.0	63.7	90.5	-	-
Non Hashing	DCSL	IJCAI2016 [55]	**76.5**	**94.2**	**98.7**	-	1.70e+04	**89.6**	**97.8**	99.2	-	3.48e+03
	EDM	ECCV2015 [40]	-	-	-	-	-	86.6	-	-	-	3.48e+03
	SIR+CIR	CVPR2016 [45]	-	-	-	-	1.70e+04	72.5	-	-	-	3.48e+03
	JSTL	CVPR2016 [50]	66.6	-	-	-	1.70e+04	-	-	-	-	-
	NSL	Zhang2016 [53]	65.0	85.0	94.4	-	-	-	-	-	-	-
	Improved Deep	CVPR2015 [1]	47.5	71.6	87.5	-	1.70e+04	65.0	89.0	97.5	-	3.48e+03
	Semantic	CVPR2015 [41]	32.7	51.2	76.3	-	-	-	-	-	-	-
	MLF	CVPR2014 [59]	34.3	55.1	75.0	-	-	-	-	-	-	-
	DeepReID	CVPR2014 [19]	-	-	-	-	1.70e+04	27.9	61.0	88.2	-	3.48e+03
	SalMatch	ICCV2013 [57]	28.5	46.0	67.3	-	-	-	-	-	-	-
	eSDC	CVPR2013 [58]	19.7	32.4	50.2	1.09e+02	2.52e+06	22.8	46.1	70.5	2.61e+01	5.19e+05
	KISSME	CVPR2012 [16]	-	-	-	-	-	29.4	60.2	86.6	-	-
	SDALF	CVPR2010 [10]	9.9	22.7	39.7	2.51e+01	4.02e+03	9.9	45.7	67.8	6.23e+00	9.01e+02
	CSBT*	**Ours**	51.2	76.3	91.8	**9.57e-06**	**9.85e+01**	74.3	93.8	**99.3**	**2.26e-07**	**4.69e+00**

('*': Experimental results with the optimal bit length are adopted. '-': The source codes or implemental details are not available.)

coding scheme in a subspace with less cross-camera variations, making the learned binary transformation more robust. 3) For longer bit lengths (e.g., 256 and 512 bits), SDH can obtain better results than cross-view hashing methods. This is probably because SDH adopts the training strategy similar to **B-step** of CSBT, i.e., discretely learning binary codes without relaxation, while cross-view hashing approaches only generate approximate results.

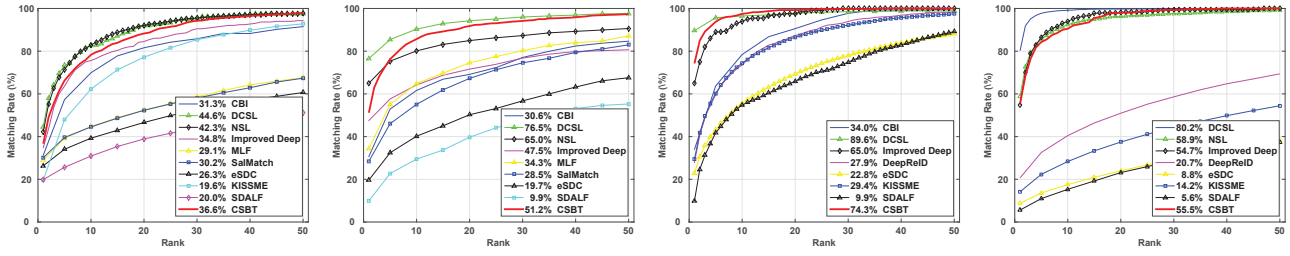
5.3. Comparison with the State-of-the-art Person Re-identification Methods

In this section, we compare CSBT to the state-of-the-art person re-identification approaches on the VIPeR, CUHK01, CUHK03 and Market-1501 datasets. The compared methods include existing hashing based approaches for fast person re-identification: DSRH [56], DRSCH [54] and CBI [60]. Same as [60], we also compare to several representative non-hashing based methods, which adopt different matching strategies during testing: 1) **Metric Learning -** KISSME [16] and NSL [53]; 2) **Local Patches based Matching -** SDC [58], SalMatch [57] and MLF [59]; 3) **Deep Learning -** DeepReID [19], Improved Deep [1], DCSL [55], SIR+CIR [45], EDM [40], SSDL [42] and Gated CNN [43]; 4) **Semantic Attribute based Matching -** Semantic [41]; 5) SDALF [10].

Table 3. Rank 1 matching rate (%), mAP, average query time (in seconds) and memory usage (in kilobytes) for storing gallery data, by comparing with state-of-the-art approaches on Market-1501.

Method	Reference	r=1	mAP	Time (s)	Mem. (KB)
Gated CNN	ECCV2016 [43]	**65.9**	**39.6**	-	4.32e+04
SSDL	ECCV2016 [42]	39.4	19.6	-	-
SCSP	CVPR2016 [3]	51.9	-	-	1.21e+01
NSL	CVPR2016 [53]	55.4	-	-	4.13e+06
BoW+KISSME	ICCV2015 [61]	39.6	17.7	2.08e+00	1.54e+04
eSDC	CVPR 2013 [58]	33.5	13.5	7.47e+02	1.45e+06
KISSME (LOMO)	CVPR 2012 [16]	40.5	19.0	-	-
SDALF	CVPR2010 [10]	20.5	8.2	2.53e+02	9.31e+04
CSBT	**Ours**	42.9	20.3	**4.7e-04**	**1.52e+03**

Table 2 summarizes the comparison results on the VIPeR, CUHK01 and CUHK03 datasets, and Fig. 5 demonstrates the corresponding CMC curves at top 50 ranks. Table 3 shows the rank 1 matching rate together with mAP on Market-1501. In [60], the results of CBI by using different features (including the LOMO feature) and bit lengths are



(a) (b) (c) (d)

Figure 5. CMC curves on various datasets: (a) VIPeR; (b) CUHK01 (486 test persons); (c) CUHK01 (100 test persons); (d) CUHK03.

provided on VIPeR and CUHK01. We simply adopt the best performance of each method with the optimal bit length.

It should be noted that some non-hashing based works proposed recently have reported higher results, by employing deep learning [7, 55], model ensembling [20, 21, 53, 59] or rank optimization [12]. Generally, the matching accuracies by using hashing based methods are relatively lower than non-hashing based ones, due to the binary quantization loss [60]. However, this paper mainly focuses on fast person re-identification. Hashing methods are significantly more efficient compared to non-hashing based approaches, which will be shown in the later experiments. It is therefore a trade-off between accuracy and efficiency. As we pointed out, efficiency was paid little attention in the literature, but is important for large-scale sets. We thus aim to improve the efficiency without sacrificing too much accuracy.

From Tables 2–3 and Fig. 5, it can be observed that CSBT significantly outperforms existing hashing based person re-identification approaches. For instance, CSBT boosts the rank 1 matching rate of CBI by 5.3%, 20.6% and 40.3%, on VIPeR and CUHK01, respectively. Similar observations can be obtained on CUHK03.

Compared with state-of-the-art non-hashing based methods, CSBT can still achieve competitive performance. As shown in Tables 2–3, the matching rates at rank 1 of CSBT are higher than many existing non-hashing approaches, including deep learning based methods such as SIR+CIR and Improved Deep. However, there remains a gap between the matching accuracy of CSBT and that of non-hashing models such as DCSL. Nevertheless, CSBT has its advantages in the matching efficiency, which we will show shortly.

The efficiency of CSBT. As previously claimed, the proposed hashing based method is significantly more efficient than non-hashing person re-identification approaches. To make it clear, we compare the average time cost for one query during re-identification, i.e., the time for computing the similarities between one query and all gallery data (316, 972/200, 100 and 19732 samples for VIPeR, CUHK01, CUHK03 and Market-1501, respectively). Additionally, we provide the memory load for storing gallery data. All experiments are conducted on a PC with Intel Core CPU (3.4GHz) and 16GB RAM. Since the source codes of SIR+CIR, SSDL, Improved Deep, Semantic and DeepReID are not available, we can not directly evaluate their time and storage efficiency. However, as described in [1, 19, 45], these methods require all raw gallery images during testing. As a consequence, we can compare their memory usage. Besides, the implementation details of KISSME are unknown on CUHK01. We therefore only evaluate its time and memory cost on VIPeR, CUHK03 and Market-1501.

As shown in Table 2 and Table 3, the hashing based methods, i.e., CSBT and CBI, are significantly more efficient than compared ones. Particulary, on the largest Market-1501 dataset, CSBT is at least 1,000 times faster, while requiring much less memory usage, compared to non-hashing based person re-identification approaches.

6. Conclusion

In this paper, we have presented a novel hashing based approach, namely cross-camera semantic binary transformation (CSBT), for fast person re-identification. A joint framework has been proposed to learn subspace projection and discriminative binary codes, which simultaneously preserves identities and mitigates cross-camera variations. Moreover, a new discrete learning optimization method was adopted, which can avoid quantization errors and generate better binary codes. Extensive experimental results demonstrated that CSBT has significantly enhanced the performance of existing hashing based methods. Meanwhile, CSBT can achieve competitive matching accuracy than the state-of-the-art person re-identification approaches, whilst significantly reducing the time and memory costs.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001002, in part by the Hong Kong, Macao, and Taiwan Science and Technology Cooperation Program of China under Grant L2015TGA9004, in part by the National Natural Science Foundation of China under Grant 61573045, and in part by the Foundation for Innovative Research Groups through the National Natural Science Foundation of China under Grant 61421003.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2010.

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016.

[4] J. Chen, Y. Wang, and R. Wu. Person re-identification by distance metric learning to discrete hashing. In *ICIP*, 2016.

[5] J. X. Chen, Y. H. Wang, and Y. Y. Tang. Person re-identification by exploiting spatio-temporal cues and multi-view metric learning. *IEEE Signal Processing Letters*, 23(7):998–1002, 2016.

[6] J. X. Chen, Z. X. Zhang, and Y. H. Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *IEEE Trans. on Image Processing*, 24(12):4741–4755, 2015.

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[8] G. Ding, Y. Guo, and J. Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, 2014.

[9] Q. S. A. H. Z. T. F. Shen, C. Shen. Inductive hashing on manifolds. In *CVPR*, 2013.

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

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[12] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimization by discriminant context information analysis. In *ICCV*, 2015.

[13] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[14] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[15] W. C. Kang, W. J. Li, and Z. H. Zhou. Column sampling based discrete supervised hashing. In *AAAI*, 2016.

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

[17] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, 2011.

[18] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[21] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015.

[22] G. Lin, C. Shen, Q. Shi, A. van den Hengel, and D. Suter. Fast supervised hashing with decision trees for high-dimensional data. In *CVPR*, 2014.

[23] Z. Lin, G. Ding, M. Hu, and J. Wang. Semantics-preserving hashing for cross-view retrievalg. In *CVPR*, 2015.

[24] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, 2013.

[25] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, 2015.

[26] L. Liu, Z. Lin, L. Shao, F. Shen, G. Ding, and J. Han. Sequential discrete hashing for scalable cross-modality similarity retrieval. *IEEE Transactions on Image Processing*, 26(1):107–118, 2017.

[27] L. Liu, M. Yu, and L. Shao. Multiview alignment hashing for efficient image search. *IEEE Transactions on Image Processing*, 24(3):956–966, 2015.

[28] L. Liu, M. Yu, and L. Shao. Unsupervised local feature hashing for image similarity search. *IEEE Transactions on Cybernetics*, 46(11):2548–2558, 2016.

[29] W. Liu, C. Mu, S. Kumar, and S. F. Chang. Discrete graph hashing. In *NIPS*, 2014.

[30] W. Liu, J. Wang, R. Ji, Y. G. Jiang, and S. F. Chang. Supervised hashing with kernels. In *CVPR*, 2012.

[31] Z. Liu, J. Chen, and Y. Wang. A fast adaptive spatio-temporal 3d feature for video-based person re-identification. In *ICIP*, 2016.

[32] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *ICIP*, 2013.

[33] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science and Business Media, 2008.

[34] M. Norouzi and D. M. Blei. Minimal loss hashing for compact binary codes. In *ICML*, 2011.

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

[36] J. Qin, L. Liu, M. Yu, Y. Wang, and L. Shao. Fast action retrieval from videos via feature disaggregation. In *BMVC*, 2015.

[37] M. Rastegari, J. Choi, S. Fakhraei, H. D. III, and L. S. Davis. Predictable dual-view hashing. In *ICML*, 2013.

[38] F. Shen, W. Liu, S. Zhang, Y. Yang, and H. T. Shen. Learning binary codes for maximum inner product search. In *ICCV*, 2015.

[39] F. Shen, C. Shen, W. Liu, and H. T. Tao. Supervised discrete hashing. In *CVPR*, 2015.

[40] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016.

[41] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015.

[42] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.

[43] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[44] B. Wang, G. Wang, K. L. Chan, and L. Wang. Tracklet association with online target-specific metric learning. In *CVPR*, 2014.

[45] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[46] G. Wang, L. Lin, S. Ding, Y. Li, and Q. Wang. Dari: Distance metric and representation integration for person verification. In *AAAI*, 2016.

[47] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[48] X. Wang, K. Tieu, and E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(1):56–71, 2010.

[49] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.

[50] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[51] Y. Yang, Z. Lei, S. Zhang, H. Shi, and S. Z. Li. Metric embedded discriminative vocabulary learning for high-level person representation. In *AAAI*, 2016.

[52] D. Zhang and W. J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, 2014.

[53] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[54] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Trans. on Image Processing*, 24(12):4766–4779, 2015.

[55] Y. Zhang, X. Li, L. Zhao, and Z. Zhang. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*, 2016.

[56] F. Zhao, Y. Huang, L. Wang, and T. Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*, 2015.

[57] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013.

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

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

[60] F. Zheng and L. Shao. Learning cross-view binary identities for fast person re-identification. In *IJCAI*, 2016.

[61] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[62] W. S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.