Dense Invariant Feature Based Support Vector Ranking for Cross-Camera Person Re-identification

Shoubiao Tan, Feng Zheng, Li Liu, Jungong Han, and Ling Shao*, *Senior Member, IEEE*

***Abstract*—Recently, support vector ranking has been adopted to address the challenging person re-identification problem. However, the ranking model based on ordinary global features cannot well represent the significant variation of pose and viewpoint across camera views. To address this issue, a novel ranking method which fuses the dense invariant features is proposed in this paper to model the variation of images across camera views. An optimal space for ranking is learned by simultaneously maximizing the margin and minimizing the error on the fused features. The proposed method significantly outperforms the original support vector ranking algorithm due to the invariance of the dense invariant features, the fusion of the bidirectional features and the adaptive adjustment of parameters. Experimental results demonstrate that the proposed method is competitive with state-of-the-art methods on two challenging datasets, showing its potential for real-world person re-identification.**

***Index Terms*—Person Re-identification, Support Vector Ranking (SVR), Dense Invariant Feature (DIF), Feature Fusion**

I. Introduction

The target of person re-identification (re-id) is to recognize individuals from a distributed multi-camera surveillance system [1]. It is an important problem in people retrieval and tracking applications. It is also a very challenging task because of the variation of the persons and the environment [2]. Images of different persons in the same view may be more similar than those of the same person in different views, because views from multiple cameras vary significantly due to the change of illumination, background, viewpoint, person's pose, occlusion, etc. This makes it difficult to recognize different images belonging to a same person. Moreover, it is still an open problem to obtain useful information from cameras to aid the recognition process because of the uncertainty of capture time and space of images [3]. Although the results obtained from existing solutions are encouraging, more effective ranking methods can be still exploited to improve the performance of person re-id.

Our work on the challenging re-id problem is motivated by two aspects. 1) Support vector machine (SVM) is a well-known technique and achieved remarkable performance on many tasks. Support vector ranking (SVR) is a technique deduced from SVM for solving ranking problems. It was also adopted in [4] to address the person re-id problem and achieved impressive results. Therefore, it is worthwhile to reexamine the role of SVR for person re-id to further exploit its potentiality. 2) In unsupervised ranking methods, score is computed directly from a low-level feature of each image pair to rank the images [5]. Because the low-level features are usually designed to capture the invariant information of different views, some of them can achieve good performance for person re-id [6]. However, a learning process could be helpful to further boost the identification performance for those unsupervised methods.

Inspired by the patch-based similarity metric used in [6], a dense invariant feature (DIF) is designed to capture the invariant information across camera views. Based on the obtained DIF, a feature-fusing SVR method is proposed to learn a discriminative model to enhance person re-identification performance. The contributions of our method are summarized as follows:

1) The DIF, which is a generic descriptor for combining low-level features, is designed to model the variation across the camera views for reliable person similarity ranking. Motivated by the fact that local features usually performs better than global features, images are first divided into densely-sampled patches. Moreover, for a pair of images to be matched, considering that the discriminative parts of persons usually appear in different regions of different views, the corresponding patch of a patch in the first image is searched in the neighboring area of the second image. This operation can avoid misalignment caused by the significant variation of the pose and viewpoint across the cameras, which finally lead to a view-invariant representation. Experimental results show that the DIF brings a significant improvement over the global feature used in [4].

2) Based on the DIF, SVR is used to learn a discriminative model to improve the person re-identification performance. SVR can learn the transformation across the views through the training procedure. We show its effectiveness by comparing the performance of our method with the unsupervised method in [6].

3) Both the forward and backward DIFs are extracted and fused in the ranking process to further improve the recognition rate and the robustness of the method, because they are useful to model the similarity between an image pair in different directions. Since they are built on different views, the fusion procedure can effectively alleviate the influence of noise and large similar regions.

4) It is known that the parameter C in the SVR has significant influence on the ranking performance. A parameter adaptive adjustment process is developed to automatically

S. Tan is with the Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, and the School of Electronics and Information Engineering, Anhui University, Hefei, 230601 P.R. China.

F. Zheng is with the Department of Electronic and Electrical Engineering, The University of Sheffield.

L. Liu, J. Han and L. Shao are with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle upon Tyne, U.K.

*Corresponding author: Ling Shao. E-mail: ling.shao@ieee.org.

select the optimal value of C, which can further improve the ranking performance.

II. Related Work

A. Person re-identification

Generally speaking, solving a person re-id problem includes three steps, i.e., extracting a discriminative feature as the person descriptor, calculating similarities between the probe image and the images in the gallery, and ranking the similarities to yield a matched result. Many unsupervised and supervised pair ranking methods are proposed to address this problem. These methods mainly concentrate on the first two steps: some pay much attention to the construction of discriminative feature representations while the others focus on learning a good similarity measurement model.

Feature representation: There definitively exists some characteristics of a person that keeps invariant across cameras despite the change of viewpoint, illumination, etc. Some methods are dedicated to exploring those invariant characteristics, and measuring their similarity using a standard distance metric. Since the information related to the cameras is often unreliable, most of the representations are built based on visual appearance features [7], [8], [9]. To make the representations more robust, several types of low level features such as color, shape, texture and interest points are often combined together [10], [11], [12]. Also, the features are often extracted from a collection of parts [4], [5], [13] or densely-sampled patches [14]. Salience [6], [15], mid-level semantic attributes [16], [17] and other techniques [18] are also used to enhance the discriminative power of feature representations.

Our method is related to [4] and [6]. In [4], [7], a person image is first divided into six horizontal stripes in terms of the characteristics of human body, and then a long feature vector is formed by combining the color and the texture features extracted from the stripes. We name this feature as global six-stripe appearance feature. In [6], low-level features are first extracted from overlapped patches, and then similarities between patches of an image pair are calculated, and two types of local saliency are finally learned to generate the similarity score of an image pair.

Similarity measurement: The distance between two images from different views is a popular similarity measurement of image pairs. Some basic similarity measurements are utilized in those methods which focus on invariant representations. For example, the Euclidean distance and the Bhattacharyya distance are used in [5] to compute the similarity of a pair depending on the characteristics of different features. However, due to the significant diversity across the camera views, there is a huge gap between the feature spaces of two different views. The transformations between the feature spaces can be learned to bridge the gap between the views to enable more accurate distance calculation. That is, to reduce the difference of the feature spaces of the views, a projection can be learned to project one feature space to the other [19], or project the two spaces to a common subspace [10], [20], [21], to obtain a better similarity measurement.

Another type of methods learns complex distance metrics by modeling the variance across the views to attain an accurate similarity measurement, such as LMNN [22], ITML [23], LDML [24], PRDC [7], KISSME [25], LFDA [26], RS-KISS [2] and MCE-KISS [27]. In fact, projection methods can be seen as a type of distance metric, which is proved in [20].

Some other types of techniques are also proposed to improve the performance, such as manifold ranking model [28], locally-adaptive decision function [29] and multi-task distance metric [30].

B. Support vector ranking

Support vector ranking (SVR) is first applied to web page retrieval [31]. A basic SVM algorithm is a classification technique used to divide instances into different groups. A variant of the SVM algorithm, the ranking SVM algorithm is deduced to give high rankings to higher relevance documents and low rankings to lower relevance documents. The order relation of two documents in a query is similar to the label of an instance in the SVM training process. The target of the training process is to minimize the error of binary order relations of the documents. Because the number of the binary order relations is relatively large (there are $N \times (N-1)/2$ binary order relations between N documents in a query), the ranking SVM is computationally expensive when applying it to a large-scale dataset. To overcome this weakness, a primal RankSVM (PRSVM) is proposed in [32] to speed up the existing RankSVM algorithm. The truncated Newton method is adopted to expedite the training process on the primal optimization problem. The cutting-plane algorithm proposed in [33] is another frequently-used efficient algorithm to speed up the SVM training process. Both of them are several orders of magnitude faster than the basic SVR algorithm for large-scale datasets.

Since the person re-id problem can be treated as a ranking problem, SVR is adopted in [4] to deal with it. Although only a global six-stripe appearance feature is used in the method, it still yielded remarkable results. With the rapid development in this area, many other methods have achieved promising results, such as sLDFV[34], RPLM[14], LADF[29], RCCA[10], MLFL[17] and MCE-KISS[27]. In a word, we believe that more efforts can be made to further explore the power of SVR in the research on person re-id.

III. Our Approach

A. Dense invariant feature

As a ranking method, all the images in the gallery are ranked according to their similarity to a probe image and the image with the highest similarity is taken as the matched result. Suppose that there are two camera views A and B. One is treated as the gallery and the other as the probe. Images from different views are grouped into pairs to calculate their similarity. Since global features cannot robustly model the variations of the images, a DIF is designed to represent the similarity of a pair. Fig. 1 illustrates the extraction procedure of the DIF.

Images are first divided into densely-sampled patches with M rows and N columns.

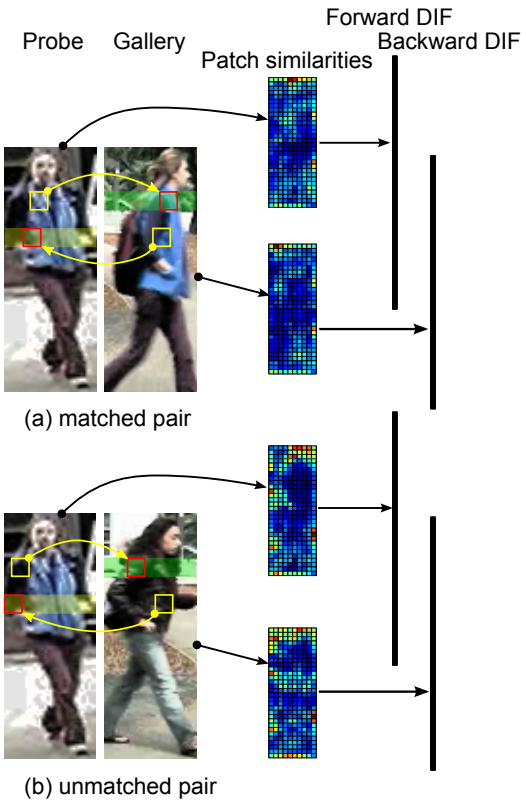


Fig. 1. Extraction of the dense invariant features

Then a local feature is extracted from each patch. $x_{m,n}^{A,u}$ denotes the feature vector of a local patch at the m-th row and n-th column of the u-th image from view A, where $m \in [1..M], n \in [1..N]$.

Next, the most similar patch of each patch of an image is searched in the surrounding area of the other image of a pair. $\mathcal{P}(u^A, v^B)$ denotes an image pair of u-th image from view A and v-th image from view B. $s(x_{m,n}^{A,u}, x_{p,q}^{B,v})$ denotes the similarity between the patch at the m-th row and n-th column in the u-th image from view A and one of its surrounding patch in the p-th row and q-th column in the v-th image from view B. In the person re-id problem, due to the change of person's pose and viewpoint, the same part of a person may appear at different positions in different views. The direct method that constructs a feature with the similarity of the patches in the same position of a pair may produce a misaligned feature. Therefore, finding similar patches nearby the corresponding patches in the other image is a more robust way to cope with the changes across different views. Because all the persons in the views have standing posture and the whole body regions are easily extracted, there are only minor changes in the longitudinal direction of the images in different views. Therefore, the most similar patch of a patch shall be searched in the surrounding area in the other image, where the surrounding area can be confined to several adjacent rows, i.e., $p \in max(1, m-l), ..., min(M, m+l)$, where l denotes the range of the adjacent rows. The distance of the features of two patches is used to measure their similarity and a Gaussian function is adopted to increase the discriminative power of the distance:

$$s(x_{m,n}^{A,u}, x_{p,q}^{B,v}) = exp(-\frac{d(x_{m,n}^{A,u}, x_{p,q}^{B,v})^2}{2\sigma_0^2}) \quad (1)$$

where σ_0 is the bandwidth of the Gaussian function and $d(x_{m,n}^{A,u}, x_{p,q}^{B,v})$ is the Euclidean distance between two patches.

Finally, the similarity between each patch of an image and its most similar surrounding patch in the other image of the pair is integrated to build a feature vector, named dense invariant feature of the image in the pair. $\mathfrak{s}(x_{m,n}^{A,u}, x^{B,v})$ denotes the similarity between the patch at the m-th row and n-th column in the u-th image from view A and its most similar surrounding patch in the v-th image from view B. It is defined as:

$$\mathfrak{s}(x_{m,n}^{A,u}, x^{B,v}) = max_{(p,q)} s(x_{m,n}^{A,u}, x_{p,q}^{B,v}) \quad (2)$$

where $p \in [max(1, m-l)..min(M, m+l)]$ and $q \in [1..N]$.

So $F_{(u^A \to v^B)} = \{\mathfrak{s}(x_{m,n}^{A,u}, x^{B,v}) \mid m \in [1..M], n \in [1..N]\}$ is the dense invariant feature of the u-th image from view A to the v-th image from view B.

Similarly, $F_{(v^B \to u^A)} = \{\mathfrak{s}(x_{m,n}^{B,v}, x^{A,u}) \mid m \in [1..M], n \in [1..N]\}$ is the dense invariant feature of the v-th image from view B to the u-th image from view A.

For an image pair $\mathcal{P}(u^A, v^B)$, $F_{(u^A \to v^B)}$ can be called forward DIF of the pair and $F_{(v^B \to u^A)}$ can be called backward DIF.

The similar patch searching process is similar to the cross-correlation method. However, unlike calculating distance directly with the pixel value in cross-correlation, we calculate the distance with the feature extracted from each patch. The extracted feature has more discrimination than the original information of images.

B. Feature fusion for ranking

Due to the improvement on the diversity of representations, combining multiple types of features has achieved great success in many areas [35], [36], [37]. The combination methods can be generally grouped into three categories [36]: descriptor-level fusion, kernel-level fusion [38] and decision-level fusion [39]. It is hard to say which type of methods is the best, because it depends on the correlation of the different type of features. Cai et al. [36] argued that descriptor fusion is probably better when the adopted descriptors have strong dependency. Accordingly, direct descriptor concatenation is adopted to fuse the forward and backward DIFs for ranking because the two features have certain dependency. The different directional DIFs represent different directional space transformation. The fusion of them can alleviate the influence of noise (caused by change of pose, occlusion, etc.) and large similar regions (caused by uniform texture).

For an image pair $\mathcal{P}(u^A, v^B)$, suppose that image u is of person i and image v is of person j, then the image pair can be abbreviated as $\mathcal{P}(i, j)$. Rewrite the forward and backward DIFs of the image pair as F_{ij}^x and F_{ij}^y. $(F_{ij}^x; F_{ij}^y)$ denotes a feature concatenated column vector of the two column vectors F_{ij}^x and F_{ij}^y. The concatenated feature is rewritten as $(F_{ii}^x; F_{ii}^y)$ in the case that it is the representation for a pair

of samples of the same person (i.e., i is equal to j). Then $(F_{ij}^x; F_{ij}^y)$ denotes the pair of samples of different persons. For simplicity, $(F_{ii}^x; F_{ii}^y)$ and $(F_{ij}^x; F_{ij}^y)$ are rewritten as F_{ii}^{xy} and F_{ij}^{xy} respectively. From the perspective of ranking, the learned function f is subject to the following inequality:

$$f(F_{ii}^{xy}) > f(F_{ij}^{xy}), i \neq j. \tag{3}$$

Descriptor concatenation of different features can improve the diversity of a representation, but it may increase the confusion and complexity of the model. We take the classification as an example, because the ranking problem is formulated as a classification problem in the classical framework of RankSVM [31]. As the feature representation is extremely complex, the classification of samples is often non-linear. However, we can convert it to a linear problem using the Kernel trick. In this paper, the space transformation is directly realized by learning a feature projection for one of the two features. Our goal is to make use of the learned projection P to improve the performance of the ranking model. Therefore, the inequality can be rewritten as:

$$f((F_{ii}^x, P^T F_{ii}^y)) > f((F_{ij}^x, P^T F_{ij}^y)), i \neq j. \tag{4}$$

where P is the projection matrix used to transform one feature into the space of the other feature.

C. Optimization

Following the framework of RankSVM, a linear function is considered for ranking the sample features $f((F_{ij}^x, P^T F_{ij}^y)) =< (w^x; w^y), (F_{ij}^x; P^T F_{ij}^y) > +b$, where $<,>$ denotes the inner product. The ranking function can be organized as $f((F_{ij}^x; P^T F_{ij}^y)) =< (w^x; Pw^y), (F_{ij}^x; F_{ij}^y) > +b$. Thus, our objective function is defined as:

$$\begin{aligned} (w^*, P^*) &= \min_{(w,P)} ||(w^x, Pw^y)||^2 + C\sum \xi_{ij} \\ s.t. \quad & f((F_{ii}^x, P^T F_{ii}^y)) > f((F_{ij}^x, P^T F_{ij}^y)) + 1 - \xi_{ij}, i \neq j. \end{aligned} \tag{5}$$

The objective function is non-convex in general. However, when we fix w, we can see that the objective function with respect to P is convex. On the other hand, when projection P is fixed, the problem becomes a classical RankSVM only when concatenating two types of feature. Thus, following [40], we can iteratively optimize the two parameters in an alternate way. Here we focus on the optimization of P.

Suppose that the optimization is divided into t steps. The optimal ranking function is found in the last step. So far, we can find the optimal projection P to improve the performance of ranking. Removing the unrelated items, the objective function is given by:

$$\begin{aligned} P^* &= \min_P((w^y)^T P^T P w^y + C\sum \xi_{ij}) \\ s.t. \quad & (Pw^y)^T D_{ij}^y > -(w^x)^T D_{ij}^x + 1 - \xi_{ij}, \xi_{ij} > 0, i \neq j. \end{aligned} \tag{6}$$

where $D_{ij}^s = (F_{ii}^s - F_{ij}^s), s \in \{x, y\}$. From the condition, we can learn that we can get a proper (Pw^y) to make the distance between $f(F_{ii}^{xy})$ and $f(F_{ij}^{xy})$ larger and then the samples easier to be ranked.

For simplification, we suppose that the projection is defined as $P = \lambda I$. Therefore, the objective function can be written as:

$$\begin{aligned} \lambda^* &= \min_\lambda(\lambda^2 (w^y)^T w^y + C\sum \xi_{ij}) \\ s.t. \quad & \lambda (w^y)^T D_{ij}^y > -(w^x)^T D_{ij}^x + 1 - \xi_{ij}, \xi_{ij} > 0, i \neq j. \end{aligned} \tag{7}$$

Set $a = (w^y)^T w^y$, $u_{ij} = (w^y)^T D_{ij}^y$ and $v_{ij} = (w^x)^T D_{ij}^x - 1$, then the above objective function can be written as:

$$\begin{aligned} \lambda^* &= \min_\lambda(a\lambda^2 + C\sum \xi_{ij}) \\ s.t. \quad & u_{ij}\lambda > -v_{ij} - \xi_{ij}, \xi_{ij} > 0, i \neq j. \end{aligned} \tag{8}$$

Note that a, u_{ij} and v_{ij} are constants because w^y is a column vector. The objective function can be solved by the Lagrange multiplier method and we can easily deduce the following equations:

$$\begin{aligned} \lambda^* = \min_\lambda \max_{\alpha_{ij}>0, \beta_{ij}>0} (& a\lambda^2 + C\sum \xi_{ij} - \\ & \sum \alpha_{ij}(u_{ij}\lambda + v_{ij} + \xi_{ij}) - \sum \beta_{ij}\xi_{ij}) \end{aligned} \tag{9}$$

Then we can obtain the following equations according to Eq.(9):

$$\lambda = \frac{\sum \alpha_{ij} u_{ij}}{2a} \tag{10}$$

$$\begin{aligned} \alpha^* &= \max_{\alpha_{ij}>0}(-\frac{(\sum \alpha_{ij}u_{ij})^2}{4a} - \sum \alpha_{ij} v_{ij}) \\ &= \max_{\alpha_{ij}>0}(-\alpha^T \frac{uu^T}{4a}\alpha - v^T\alpha) \\ &= \min_{\alpha_{ij}>0}(\alpha^T \frac{uu^T}{4a}\alpha + v^T\alpha) \end{aligned} \tag{11}$$

Algorithm 1 Feature-fusion Support Vector Ranking

Input: $S_{trn}^x, S_{trn}^y, S_{tst}^x, S_{tst}^y, C$ ▷ $S = \{F_{ij}\}$ is the feature set

Output: cmc, λ ▷ cmc is the Cumulative Matching Characteristic performance

```
Calculate D_ij of each feature set.
λ_0 = 0, λ = 1;
while |λ − λ_0| > 10^{-2} do
    λ_0 = λ;
    S_trn = (S^x_trn, λS^y_trn), S_tst = (S^x_tst, λS^y_tst);  ▷ Fusing feature
    [cmc, w] = PRSVM(S_trn, S_tst, C);  ▷ PRSVM is proposed in [32]
    Calculate a, u_ij, v_ij according to their definitions;
    Calculate α according to Eq.(11);
    λ = (Σ α_ij u_ij) / 2a;
end while
return cmc, λ;
```

It can be seen that α^* is a typical quadratic programming (QP) problem and can be solved by convex optimization. After α is obtained from the QP algorithm, λ can be easily calculated according to Eq.(10).

Algorithm 2 C-adaptive FFSVR

Input: G, P ▷ G denotes the gallery images and P indicates the probe images

Output: cmc, λ, C ▷ cmc is the Cumulative Matching Characteristic performance

Divide G and P randomly into two parts G_{trn}, P_{trn} and G_{tst}, P_{tst}, one for training and the other for testing;
Extract forward DIF and backward DIF from the two parts into S^x_{trn}, S^y_{trn} and S^x_{tst}, S^y_{tst} according to Section III-A;
Divide training samples S^x_{trn}, S^y_{trn} randomly into two sub-parts S^x_{trn1}, S^y_{trn1} and S^x_{trn2}, S^y_{trn2} to learn the best C;
while the best cmc is not reached **do**
    Predict a new C according to the cmc with previous C
    $cmc = FFSVR(S^x_{trn1}, S^y_{trn1}, S^x_{trn2}, S^y_{trn2}, C)$;
end while
▷ test on the probe images with the learned best C
$[cmc, \lambda] = FFSVR(S^x_{trn}, S^y_{trn}, S^x_{tst}, S^y_{tst}, C)$;
return cmc, λ, C;

D. Algorithm and Implementation

In an SVR algorithm, parameter C makes a great influence on the training process and the ranking performance. A successive approximation procedure is adopted in this paper to find the best C by the training samples to obtain the best performance. The major procedure "Feature-fusion Support Vector Ranking (FFSVR)" and the total algorithm of our method "C-adaptive FFSVR" are summarized in Alg. 1 and Alg. 2. Note that the triangle in the algorithms represents the start of a comment.

In the algorithm, some processes are performed using existing codes or tools. The code provided by [6] is adopted to extract low-level features and calculate the similarity between patches. A program is designed to compute the patch similarity into forward and backward DIF. Positive and negative samples are also constructed in the program. Then RankSVM [31] is adopted and extended to implement our feature fusion SVR process. The QP solver of Matlab is adopted to address the QP problem in the SVR training process (The quadprog function of Matlab is compiled to dll and called from C code). Like many other person re-id methods, feature construction is the most time-consuming step of the whole algorithm because the DIFs are calculated for all the image pairs.

IV. Experiments and Results

A. Datasets

Two public datasets, the VIPeR dataset [41] and the CUHK Campus dataset [35], are adopted to evaluate our approach. The VIPeR dataset is a commonly used dataset for person re-id evaluation, and the CUHK Campus dataset is a recently released large-scale dataset. Both of them are very challenging for person re-id because their images were captured from different views with low resolutions and vary significantly on background, illumination, person's pose and viewpoint. The VIPeR dataset (VIPeR) contains 632 image pairs captured from two cameras, and each pair consists of two images of the same person from the two cameras, i.e., there are totally 1264 images in the dataset. The CHUK Campus dataset (CAMPUS) contains 3884 images of 971 persons captured from two cameras. Each person has two images in each camera view.

B. Experimental setup

The dColorSIFT proposed by Zhao et al. [6] is used to represent each patch for all images. Although DIF is a generic descriptor for combination of low-level features, it is also sensitive to the choice of the feature. dColorSIFT is a good feature to make the DIF achieve a good ranking performance. Following [15], all images in VIPeR are first normalized to the size of 128×48 and all images in CAMPUS are normalized to 160×60. Then all the images are divided into patches with the size of 10×10 and an overlap of 6×6. Then the dColorSIFT feature [6], i.e., 32-bin color histograms computed in 3 channels (L, A, B) and 3 scaling levels (0.5, 0.75, 1), and 8-bin dense SIFT extracted in 4×4 cells and 3 color channels, are extracted from each local patch. The patch similarity is calculated according to Eq.(1) with $\sigma_0 = 2.8$ for VIPeR and $\sigma_0 = 3.2$ for CAMPUS [6]. The number of adjacent rows is confined to 1 to search the most similar patches for each patch of an image in the other image [6].

C. Experimental results

Our ranking methods (denoted by DSVR_XXX) with a single forward DIF, a single backward DIF, and the fused feature are tested on two datasets. In the experiments, both the fixed C and adaptive C are evaluated. Some state-of-the-art methods, including RankSVM [4], eSDC [6], SalMatch [15], eLDFV [34], sLDFV [34], RPLM [14], LF [26], RS-KISS [2], MCE-KISS [27], RCCA [10], LADF [29], LMNN [35], ITML [35], GenericMetric [35] and MLFL [17], are compared with our methods. All the experimental results are obtained using the same evaluation criterion: 50% of all images are randomly selected for training and the rest are used for testing. All the reported results are averaged over 10 independent trials, and shown in standard Cumulated Matching Characteristics (CMC) curves [42].

The results of the methods on the VIPeR dataset are shown in Table I and Fig. 2, whereas the results on the CUHK Campus dataset are shown in Table III and Fig. 3. In the following tables and figures, our ranking methods are named as:

- DSVR_B: ranking with backward DIF and fixed C,
- DSVR_BA: ranking with backward DIF and adaptive C,
- DSVR_F: ranking with forward DIF and fixed C,
- DSVR_FA: ranking with forward DIF and adaptive C,
- DSVR_FS: ranking with fused DIFs and fixed C,
- DSVR_FSA: ranking with fused DIFs and adaptive C.

From the experimental results , we have the following five observations:

1) Our method with the fused feature achieves an average accuracy of 29.35% on VIPeR and 33.46% on CAMPUS. Both of them are better than most of the compared methods. Also

TABLE II
PARAMETERS C AND FUSION WEIGHT LEARNED IN OUR METHOD FOR THE VIPER DATASET

Trial No	1	2	3	4	5	6	7	8	9	10
C	3.16	3.16	10	0.17	0.32	3.65	1.19	1.77	3.65	42.1
Weight	0.1	0.1	0.5	0.3	0.7	0.7	0.5	0.5	0.3	0.5

TABLE IV
PARAMETERS C AND FUSION WEIGHT LEARNED IN OUR METHOD FOR THE CUHK CAMPUS DATASET

Trial No	1	2	3	4	5	6	7	8	9	10
C	0.01	0.1	0.037	0.032	0.032	0.1	0.21	0.32	0.075	0.042
Weight	1	0.5	0.5	0.5	0.7	0.5	0.5	0.3	0.7	0.5

TABLE I
TOP RANKED RATES WITH 316 PERSONS ON THE VIPER DATASET

Rank	1	5	10	20
eLDFV[34]	22.34	47	60.04	71
sLDFV[34]	26.53	56.4	70.88	84.63
RPLM[14]	27	50	69	83
LF[26]	24.18	53.5	67.12	81.38
RS-KISS[2]	24.4	51.3	66.3	81.6
MCE-KISS[27]	28.2	58.5	72.1	85.9
RCCA[10]	30	56	75	87
LADF[29]	29.34	61.7	75.7	87.7
MLFL[17]	29.11	52.7	66	79.9
RankSVM[4]	16.27	38.23	53.73	69.87
eSDC[6]	26.31	46.61	58.86	72.77
SalMatch[15]	**30.16**	52.9	65.8	79.4
DSVR_B[1]	23.73	46.55	58.42	71.52
DSVR_BA	24.40	46.42	58.32	70.95
DSVR_F[1]	27.56	50.15	61.74	75.09
DSVR_FA	28.23	49.81	61.55	75
DSVR_FS[2]	28.35	50.69	61.99	74.74
DSVR_FSA	**29.35**	50.66	61.93	74.94

[1] C is fixed at 1.
[2] C is fixed at 1.

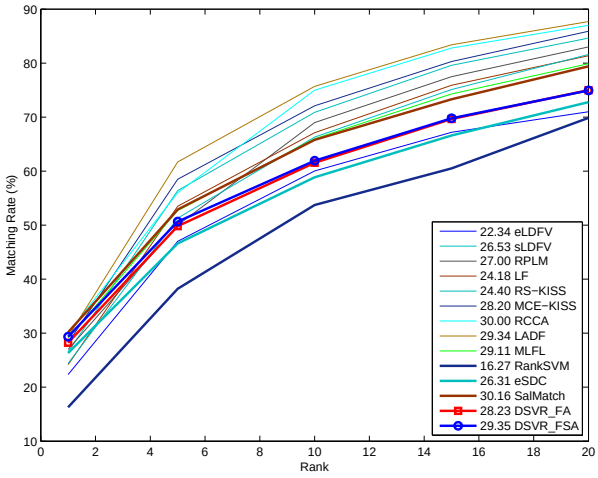


Fig. 2. CMC curves with 316 persons on the VIPeR dataset

TABLE III
TOP RANKED RATES WITH 486 PERSONS ON THE CUHK CAMPUS DATASET

Rank	1	5	10	20
LMNN[35]	13.45	31.6	42.5	54.2
ITML[35]	15.98	35.6	45.8	59.6
GenericMetric[35]	20	44.02	56.07	69.47
MLFL[17][1]	**34.3**	54.8	64.9	75.3
SalMatch[15]	28.45	45.85	55.67	67.95
DSVR_B[2]	25.43	44.53	55.29	66.54
DSVR_BA	26.21	44.05	54.20	65.45
DSVR_F[2]	30.04	48.97	58.88	69.69
DSVR_FA	30.6	48.44	57.9	69.09
DSVR_FS[3]	32.82	51.5	61.31	71.33
DSVR_FSA	**33.46**	50.88	60.97	70.97

[1] 43.39% is reported by combining with the best performing LADF.
[2] C is fixed at 0.1.
[3] C is fixed at 0.1.

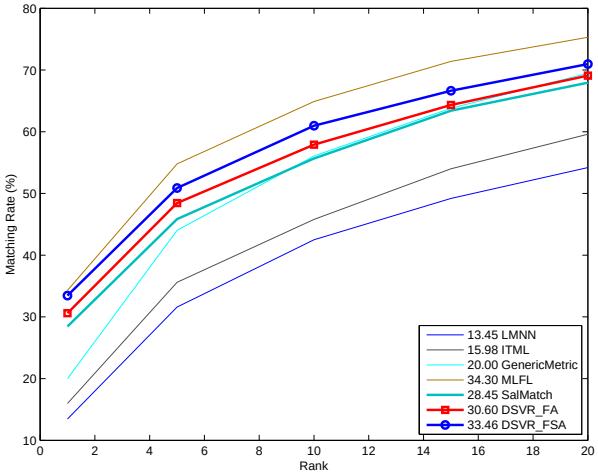


Fig. 3. CMC curves with 486 persons on the CUHK campus dataset

based on dColorSIFT feature and the SVM learning method, SalMatch slightly outperforms our method on VIPeR because a large number of complex saliency models are constructed for every image patch before its learning process. However, even with the complex saliency models, SalMatch is worse than our method by 5% on CAMPUS, showing that our method is more robust. Similarly, MLFL slightly outperforms our method on CAMPUS, because it combines a high level descriptor with a low-level feature to achieve high performance. Altogether, we can conclude that the proposed method is competent for

person re-id applications.

2) RankSVM is the first SVR based method for person re-id, where a global appearance feature is used. Our method with only forward DIF also outperforms RankSVM by a useful margin of about 12% on VIPeR. It shows that the proposed DIF can significantly enhance the performance for person re-id.

3) Similar to our method, eSDC also uses the dColorSIFT feature and patch-based similarity metric. Our method outperforms eSDC on VIPeR even when it is combined with wHSV and MSCR [5]. We conjecture that the advantage of our method is attributed to the SVR learning method. Through the training process of SVR, the transformation across the views is learned. In other words, the information can be learned about which patches are dominant (more likely to appear in both views).

4) On both datasets, the fused feature makes our method more powerful than the single forward or backward DIF. It shows that the feature fusion is indeed beneficial to boost the performance for the SVR-based method. It can be also seen that the performance improvement is not significant on VIPeR. This is because the two directional features have certain dependency. The fusion procedure mainly takes effect on alleviating the influence of noise. In addition, the forward DIF performs better than the backward DIF on both datasets, which indicates that constructing DIF based on the probe image exploits more discriminative power than gallery images.

5) Our methods with adaptive C beat those with fixed C in all the experiments, which shows that the C-adaptive procedure in our method can automatically find an optimal value of C and hence improve the ranking performance.

The parameter C and the feature-fusion weights learned in 10-trial experiments on VIPeR and CAMPUS are shown in Tables II and IV.

Finally, the results in Table I show that the performances of our method and some other methods are not consistent across different ranks on VIPeR. The possible reason is that the methods also reduce the distance of other images with the minimization of the distance of some image pairs. This may cause more false matches in larger ranks. The samples in the training process are only labeled as positive or negative, such that the learned model cannot guarantee a good performance in larger ranks. A potential solution is to define a more discriminative objective function in the training process, which will be considered in our future work.

D. Computational Time

All the experiments are run on a workstation with 16 Intel(R) Xeon(R) CPUs (E5-2660: 8 cores, 2.20GHz) and 64GB RAM. The dColorSIFT extraction and DIF construction are implemented with Matlab and multi-threading technique is used. The learning and test of SVR is implemented with single-threading C. The detailed runtime on VIPeR is: 16 minutes for dColorSIFT extraction, 25 minutes for DIF construction, 10 minutes for model training, and 1 second for the test process.

V. Conclusion and Future Work

A novel ranking method which fuses the dense invariant features has been presented in this paper to model the relationship between an image pair across different camera views to solve the challenging person re-id problem effectively. Experimental results show that the designed DIF is a useful descriptor for an image pair which leads to a significant improvement on ranking performance. The fusion of bidirectional DIFs in the ranking process further improves the performance due to the reduction of the noise. Results on two challenging datasets verify the robustness of the proposed method. We will pay much attention to learning aligned patches from images of the probe view in our future work.

Acknowledgments

The authors are grateful to the support of the Anhui Provincial Natural Science Foundation (Grant No. 1508085MF120) and the S&T project of China State Grid Corporation (Grant No. 5212D01502DB). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Anhui University.

References

[1] S. F. Tahir and A. Cavallaro, "Cost-effective features for reidentification in camera networks," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 24, no. 8, pp. 1362–1374, Aug. 2014.

[2] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.

[3] Y. Wang, R. Hu, C. Liang, C. Zhang, and Q. Leng, "Camera compensation using a feature projection matrix for person reidentification," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 24, no. 8, pp. 1350–1361, Aug. 2014.

[4] B. Prosser, W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *BMVC*, no. 3, 2010, p. 5.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[6] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.

[7] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.

[8] S. Zhang, H. Yao, X. Sun, and X. Lu, "Sparse coding based visual tracking: Review and experimental comparison," *Pattern Recogn.*, vol. 46, no. 7, pp. 1772–1788, Jul. 2013.

[9] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 25, no. 11, pp. 1749–1760, Nov. 2015.

[10] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *AVSS*, 2013, pp. 244–249.

[11] S. Zhang, H. Yao, H. Zhou, X. Sun, and S. Liu, "Robust visual tracking based on online learning sparse representation," *Neurocomputing*, vol. 100, no. 1, pp. 31–40, Jan. 2013.

[12] S. Zhang, H. Zhou, H. Yao, Y. Zhang, K. Wang, and J. Zhang, "Adaptive normalhedge for robust visual tracking," *Signal Process.*, vol. 110, pp. 132–142, May 2015.

[13] T. Zhou, M. Qi, J. Jiang, X. Wang, S. Hao, and Y. Jin, "Person re-identification based on nonlinear ranking with difference vectors," *Inform. Sciences*, vol. 279, pp. 604–614, Sep. 2014.

[14] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012, pp. 780–793.

[15] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *ICCV*, 2013, pp. 2528–2535.

[16] R. Layne, T. M. Hospedales, S. Gong *et al.*, "Person re-identification by attributes," in *BMVC*, 2012, p. 3.

[17] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014, pp. 144–151.

[18] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *ECCV*, 2012, pp. 806–820.

[19] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: what features are important?" in *ECCV*, 2012, pp. 391–401.

[20] L. Liu, X. Lu, Y. Yuan, and X. Li, "Person re-identification by bidirectional projection," in *ICIMCS*, 2014, p. 1.

[21] S. Zhang, X. Lan, Y. Qi, and P. C. Yuen, "Robust visual tracking via basis matching," *IEEE Trans. Circ. Syst. Video Tech.*, DOI: 10.1109/TCSVT.2016.2539860, 2016.

[22] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *NIPS*, 2005, pp. 1473–1480.

[23] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007, pp. 209–216.

[24] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009, pp. 498–505.

[25] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[26] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013, pp. 3318–3325.

[27] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based kiss metric learning," *IEEE Trans. Cyb.*, vol. 45, no. 2, pp. 242–252, Feb. 2015.

[28] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *ICIP*, no. 4, 2013, p. 5.

[29] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013, pp. 3610–3617.

[30] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.

[31] T. Joachims, "Optimizing search engines using clickthrough data," in *KDD*, 2002, pp. 133–142.

[32] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Inform. Retr.*, vol. 13, no. 3, pp. 201–215, Mar. 2010.

[33] T. Joachims, "Training linear svms in linear time," in *KDD*, 2006, pp. 217–226.

[34] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV*, 2012, pp. 413–422.

[35] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012, pp. 31–44.

[36] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *CVPR*, 2014, pp. 596–603.

[37] J. Yu, D. Tao, J. Li, and J. Cheng, "Semantic preserving distance metric learning and applications," *Inform. Sciences*, vol. 281, pp. 674–686, Oct. 2014.

[38] D. Tao, L. Jin, W. Liu, and X. Li, "Hessian regularized support vector machines for mobile image annotation on the cloud," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 833–844, Apr. 2013.

[39] D. Tao, J. Cheng, M. Song, and X. Lin, "Manifold ranking-based matrix factorization for saliency detection," *accepted by IEEE Trans. Neural Networks Learn. Syst.*, 2015.

[40] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2006, pp. 801–808.

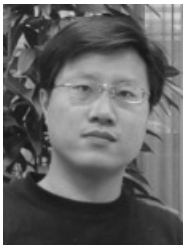
[41] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS*, vol. 14, no. 2, 2007, pp. 1524–1531.

[42] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *ICCV*, 2007, pp. 1–8.

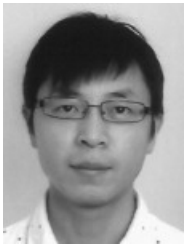
Shoubiao Tan is an associate professor in Anhui University, P.R. China. He received the Ph.D. degree in 2004 from University of Science and Technology of China. He has ever been an academic visitor of the University of Sheffield, U.K. between Feb. 2014 and Feb. 2015. His current research interests include computer vision and pattern recognition.

Feng Zheng is currently working toward the Ph.D. degree in the Department of Electronic and Electrical Engineering, the University of Sheffield, UK. His research interests include visual tracking and ensemble learning.

Li Liu received the B.Eng. degree in electronic information engineering from Xian Jiaotong University, Xian, China, in 2011, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., in 2014. He is currently a Research Fellow with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K. His current research interests include computer vision, machine learning, and data mining.



Jungong Han received his Ph.D. degree in Telecommunication and Information System from Xidian University, China, in 2004. During his Ph.D. study, he spent one year at Internet Media group of Microsoft Research Asia, China. From 2005 to 2010, he was with Signal Processing Systems group at the Technical University of Eindhoven (TU/e), The Netherlands. In December of 2010, he joined the Multi-Agent and Adaptive Computation group at the Centre for Mathematics and Computer Science (CWI) in Amsterdam. From 2012-2015, he has been a senior scientist with Civolution technology in Eindhoven (a combining synergy of Philips Content Identification and Thomson STS). In September of 2015, he started a senior lecturer position at Northumbria University, UK. Dr. Hans research interests include multimedia content identification, multi-sensor data fusion, and computer vision. He has written and co-authored over 80 papers. He is an associate editor of Elsevier Neurocomputing, and has organized several special issues on international journals.



Ling Shao is a Professor with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle upon Tyne, U.K. Previously, he was a Senior Lecturer (2009-2014) with the Department of Electronic and Electrical Engineering at the University of Sheffield and a Senior Scientist (2005-2009) with Philips Research, The Netherlands. His research interests include Computer Vision, Image/Video Processing and Machine Learning. He is an associate editor of the IEEE Transaction on Image Processing, the IEEE Transaction on Cybernetics, and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology.