A New Patch selection method based on Parsing and Saliency Detection for Person Re-identification

Yixiu Liu[a], Yunzhou Zhang[a,*], Sonya Coleman[b], Bir Bhanu[c], and Shuangwei Liu[a]

[a] *Northeastern University, College of Information Science and Engineering, Shenyang, China, 110819*
[b] *Intelligent Systems Research Centre, University of Ulster, Londonderry, UK*
[c] *University of California, Riverside, Riverside, CA-92521*

Abstract

Person re-identification is an important technique towards automatic recognition of a person across non-overlapping cameras. In this paper, a novel patch selection method based on parsing and saliency detection is proposed. The algorithm is divided into two stages. The first stage, primary selection: Deep Decompositional Network (DNN) is adopted to parse a pedestrian image into semantic regions, then sliding window and color matching techniques are proposed to select pedestrian patches and remove background patches. The second stage, secondary selection: saliency detection is utilized to select reliable patches according to saliency map. Finally, PHOG, HSV and SIFT features are extracted from these patches and fused with the global feature LOMO to compensate for the inherent errors of saliency detection. By applying the proposed method on such datasets as VIPeR, PRID2011, CUHK01, CUHK03, PRID 450S and iLIDS-VID, it is found that the proposed descriptor can produce results superior to many state-of-the-art feature representation methods for person identification.

Keywords: person re-identification, patch selection, pedestrian parsing, saliency detection, feature fusion

*Corresponding author
Email address: zhangyunzhou@mail.neu.edu.cn (Yunzhou Zhang)

1. Introduction

Person re-identification aims to identify pedestrians in non-overlapping cameras. It plays a role in a variety of practical applications, such as pedestrian searching, tracking, and analyzing behavior in different camera scenes. Person re-identification makes a significant contribution in reducing time as it can be used to seek a specific person from large amounts of images or videos rather than a human doing this manually. For the above reasons, person re-identification has gained much attention among researchers [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. However, it remains a challenging problem. A person could undergo significant variations in pose, viewpoint, scale, and illumination when walking through several different cameras. Moreover, background clutter, image blur, and occlusion make the situation even worse. All these problems make intra-person variations even larger than inter-person variations.

In this paper, we focus on constructing robust feature representation to solve these problems. Existing methods for feature representation mostly focus on two different aspects: hand-crafted features and deep features [3].

For hand-crafted features, many of them have been developed to achieve precise matching, such as the covariance descriptor based on bio-inspired features (gBiCov) [5], salient color names based color descriptor (SCNCD) [8], and ensemble color model (ECM) [4]. It can be found that these methods have some common problems. They do not remove the background noise, and the attribute types are simplex. Although ECM fuses different color attribute, it has no gradient and other attributes. However, the features we want to construct should have less noise but more diverse attributes. Therefore, a preprocessing of removing background noise is necessary, and we also consider combining multiple attributes to enhance features. In this paper, Pyramid Histogram of Oriented Gradients (PHOG [12]), HSV and Scale Invariant Feature Transform (SIFT [9]) features representing gradient, color, and extreme points are fused to complement each other.

For deep features, they have continuously updated the highest recognition

rate in recent years. A lot of methods are proposed to extract the deep features based on Convolutional Neural Network (CNN). Some of them try to design new CNN frameworks get better deep features, e.g., JointRe-id [13]. Some works enhance deep features by fusing with multiple hand-crafted features, e.g., FFN [14]. Others obtain more discriminating deep features by modifying the loss function in the training process of CNN, e.g., Quadruplet [15]. Although each of these method has achieved breakthrough results, we still find their weakness in some practical application scenarios. The problem is that data-driven deep learning cannot play a full role if the samples in the training set are insufficient. So we can see that deep learning methods are usually applied to large-scale person re-identification datasets, such as Market1501 [16], DukeMTMC-ReID [17], and MSMT17 [18]. It inspires us to construct a new feature representation to solve the problem of insufficient samples, and the new feature is supposed to improve accuracy more than some deep features.

So we consider about picking out valuable patches from images precise feature matching. Actually, there are already many researches about local feature exist, much like our idea. Whether you design a local feature (e.g., SDALF [19]) or map an existing local feature space to another space (e.g., LFDA [6]), they all have a same problem: feature drift. The location of most similar patches in different images changes cross different camera views, and we call this phenomenon feature drift. Some methods try to solve it by saliency features, and get effective improvement, such as SCNCD [8] and SalMatch [20]. The fly in the ointment is that they ignore the inherent error of saliency. Therefore, we utilize Graph-Based Visual Saliency (GBVS) [21] to change location-guide feature matching into saliency-guide feature matching, so as to effectively solve the problem of feature drift. In addition, we adopt the strategy combined with global feature Local Maximal Occurrence (LOMO [3]) to compensate for the inherent errors caused by saliency detection.

In summary, the proposed method makes the following contributions for person re-identification:

(1) We propose a patch selection method, which can effectively solve the problem of insufficient samples in actual scenarios, and has great significance in engineering applications and has some theoretical value.

(2) In the primary selection, we propose a preprocessing method to remove background noise. We use Deep Decompositional Network (DDN) to divide the picture into semantic regions, and propose sliding window and color matching techniques to remove the background patches.

(3) In the secondary selection, we utilize saliency detection to solve the problem of feature drift and patch unbalance caused by primary selection. It makes us matching features in saliency-guide, rather than location-guide.

(4) We propose a strategy that combines local features with global features to solve the problem of mismatching caused by saliency detection. PHOG, HSV and SIFT features are extracted from the selected patches. LOMO features are extracted from the whole image and fused with them to compensate for the inherent error of saliency detection.

The paper is organized as follows. The review of related work is provided in Section 2. The proposed algorithms are described in detail in Section 3. Experimental results using six public benchmark datasets are presented and analyzed in Section 4. Finally, the conclusions are given in Section 5.

2. Related work

2.1. Deeply-learned methods for person re-identification

Person re-identification is classified into two categories: single-shot case, and multi-shot case. In general, single-shot person re-identification is required to match a single probe image to a single gallery image. As for multi-shot case, a probe image or images can be matched to frames in the gallery and the matching results can be combined to obtain the result for a video sequence.

In recent years, deep learning has been widely used in image recognition tasks and has made great breakthroughs especially in person re-identification. Yi et al. [13] proposed a method which can simultaneously learn features and

a corresponding similarity metric for person re-identification. Chen et al. [22] presented a novel multi-channel parts-based convolutional neural network (CNN) model that utilized a triplet framework. The CNN model was trained by an improved triplet loss function that assigned the same ID for the closer instances in the learned feature space and assigned a different ID for the farther instances. Furthermore, instead of directly training on the sample images, some methods [13, 22, 23, 24] exploited a part or patch-based deep architecture to learn discriminative feature representations, in local regions of people, with CNNs. For example, Yi et al. [24] split the input image into three rectangular overlapping patches from top to bottom firstly, and then extracted the deep features of each patch through CNN architecture.

Through observing the datasets applied by the above methods, we find a rule: the deep learning methods are extremely suitable for large-scale datasets, such as Market1501 [16] and DukeMTMC-ReID [17], and perform well in normal multi-shot datasets, such as CUHK03, but perform relatively poorly in single-shot dataset, such as VIPeR [1]. It inspired us to propose a new method to effectively solve the latter two cases, which is why our approach is only test on datasets such as CUHK03 and VIPeR but not on large-scale datasets.

2.2. Background extraction methods for person re-identification

Background extraction is an important process to improve person re-identification. It separates the target from the background to eliminate the interference of the noisy environment. Based on an improved Random Walks algorithm, Chang et al. [25] proposed an approach that combined the shape prior information and the color seed constraint into the Random Walk formulation, so that each human was divided into several parts where the color features of the HSV histogram and the 1-D RGB signal, along with texture features, were utilized for person re-identification. Le et al., [26] attempted to make a decision on what super-pixels belonged to humans and which others belonged to background through the following two techniques: the combination of super-pixels and local saliency information and the combination of super-pixels and pose estimation.

Background noise is often ignored in many previous feature representations. In this paper, a new background noise removal strategy is proposed. It is a preprocessing technique of patch primary selection. At first, the pedestrian images are parsed into semantic regions with a Deep Decompositional Network (DDN) [27], such as head, body, arms, and legs. Then pedestrian patches are extracted from the environment using sliding windows and color matching.

2.3. Saliency methods for person re-identification

The saliency of an image carries a lot of potential information that is useful for recognition task. The following methods utilizing saliency are mainly related to human perception in person re-identification. Zhao et al., [28] propose a computational model to estimate the probabilistic saliency map and formulate person re-identification as a saliency matching problem. Saliency matching and patch matching were tightly integrated into a unified structural RankSVM framework. Chen et al., [29] establishes a similarity among patches via fusing multi-directional saliency after distribution analysis for the consistency of saliency. Le et al., [26] took full advantage of the saliency for keeping superpixels that display a high saliency score (indicating a human) and removing the others (background). In this paper, saliency detection is used for secondary selection, which changes the matching of local features from location-guide to saliency-guide, so as to obtain more reliable patch sequences.

2.4. Fusion strategy for person re-identification

In this paper, a novel feature representation that combines the global and local features is proposed, which is quite different from other general methods. Most feature-based methods either extract the features from the images directly [3, 8, 30] or use only the local descriptors [1, 9, 19]. Liao [3] designed an efficient feature representation named Local Maximal Occurrence (LOMO), and a subspace and metric learning method known as Cross-view Quadratic Discriminant Analysis (XQDA) [3]. Gray and Tao's work [1] proposed an ensemble of invariant features (EIFs) where the feature representation can effectively handle the

variation of human poses/viewpoints and color difference for matching pedestrians observed under different scenes conditions. Every image was divided into a grid of local patches, and then the color histogram in LAB color space and SIFT features are extracted for metric learning [28].

The difference from the above works is that we not only fuse many types of features, but also consider the relationship between the global and the local. Considering that our patch selection method has a certain mismatching rate caused by saliency detection, we compensate for it by combining the global features with the local features extracted from the selected patches. The fused feature representation is evaluated with several metric methods which are proven to be effective for person re-identification.

3. Technical approach

3.1. Structure of the feature representation

The structure of the technical approach consists of three parts: primary selection, secondary selection, and feature fusion. The overall process of the proposed work is shown in Fig. 1.

As can be seen from Fig. 1, parsing and saliency detection are two important techniques for the two patch selection stages, respectively. Throughout the patch selection, our operation unit is patch. Firstly, as a pre-processing, DNN divides the pedestrian's body and background into semantic regions of different colors (3.2.1). It inspires us to propose a patch based sliding window and color matching method to remove the background patches and preserve the pedestrian patches (3.2.2). Afterwards, saliency detection is utilized to get the saliency map (3.3.1), through with we change location-guide feature matching into saliency-guide feature matching (3.3.2), and obtain the reliable patch sequences with higher saliency scores (3.3.3). Finally, global and local features are extracted and fused (3.4.1) to obtain complete feature representation, and metric learning is performed to evaluate it (3.4.2).

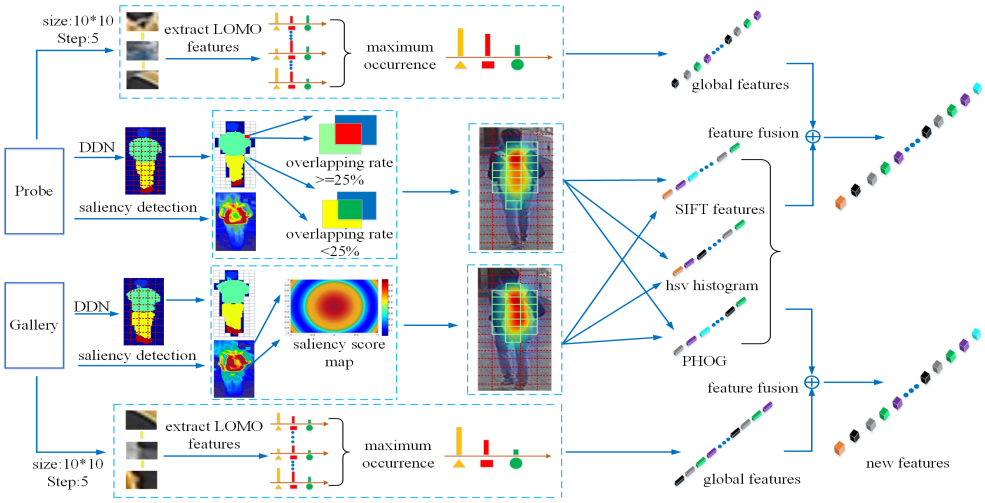

Figure 1: The architecture of the feature representation. It consists mainly of three parts, *i.e.*, primary selection, secondary selection, and feature fusion. (1). For each pair of pictures, we split them into patches and parse them into semantic regions with DDN which will be described in detail in 3.2. The overlap rate is computed to remove the background patches. The threshold of overlap rate is set to 25%. (2). The patches are further selected by saliency detection. The patch sequences with higher saliency scores are obtained. e.g., the most reliable patches that have higher saliency scores are pained in red. (3). We extract the PHOG, HSV histogram and SIFT features from the selected patches and fuse them with global LOMO features.

3.2. Primary selection

3.2.1. Semi-supervised DDN

It is not feasible to fine-tune DDN model directly on person re-identification datasets, because there are no ground truth of label maps. In other words, person re-identification datasets have no label for DDN model. So we modify it into a semi-supervised DDN model, and the training loss function is defined as

$$L = \sum_{x^l} C(y^l, \hat{y}^l) + \lambda \sum_{x^u} E(y^u), \tag{1}$$

The first term in Eq. 1 is the loss function trained on the labelled parsing dataset. The second term is the loss function trained on the unlabelled person re-identification dataset. Before that, let's review the original DDN.

Fig. 2 illustrates the architecture of the DDN which directly maps low-level visual features to the label map of body parts. The input is the feature vector, while the output is a set of label maps of body parts. This architecture is utilized for pedestrian parsing, and mainly consists of one down-sampling layer, two occlusion estimation layers, two completion layers, and two decomposition layers.

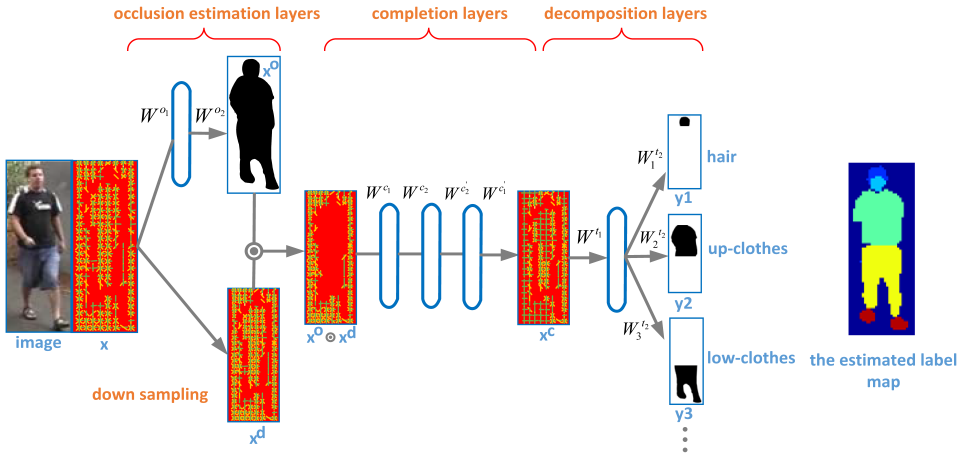

Figure 2: DDN architecture, which achieves parsing and subtraction in a unified deep network.

The input x is down-sampled to x^d. Otherwise, x is mapped into a binary occlusion mask $x^o \in [0,1]^n$ through the weight matrices w^{o_1}, w^{o_2} and the biases b^{o_1}, b^{o_2}. To reduce the number of parameters in the network, x^o and x^d are set to the same size. If the i-th element of the feature is occluded, x_i^o is set to 0, otherwise $x_i^o = 1$. The binary occlusion mask x^o is denoted as

$$x^0 = \tau(W^{o_2}\rho(W^{o_1}x + b^{o_1}) + b^{o_2}), \tag{2}$$

where the function $\tau(x) = 1/(1+\exp(-x))$. For the first layer of occlusion estimation, the rectified linear function [31] $\rho(x) = \max(0, x)$ is utilized as the

activation function and we use a sigmoid function as the activation function in the second layer.

In the architecture of the DDN, the input of the completion layers which are modeled as the denoising autoencoder (DAE)[32] is the element-wise product of x^o and x^d. While the output is the completed feature vector x^c via the weight matrices W^{c_1}, W^{c_2}, $W^{c_1\prime}$, $W^{c_2\prime}$, and the biases b^{c_1}, b^{c_2}, u^{c_1}, u^{c_2}. W' is the transpose of W. Through projecting high dimensional data into a low dimensional space, the encoders W^{c_1} and W^{c_2} find the compact representation of noisy data. The encoders $W^{c_1\prime}$ and $W^{c_2\prime}$ are used to reconstruct the data. We reconstruct x^c with x^o and x^d. The reconstruction process is as follows,

$$z = \rho(W^{c_2}\rho(W^{c_1}(x^o \odot x^d) + b^{c_1}) + b^{c_2}), \tag{3}$$

where $\odot$ represents the element-wise product, and z denotes the compact representation. According to Eq. 3, we can get

$$x^c = \rho(W^{c_1\prime}\rho(W^{c_2\prime}z + u^{c_2}) + u^{c_1}), \tag{4}$$

At the back end of DDN, the completed feature x^c is decomposed into several label maps from y_1 to y_M through the corresponding weight matrices W^{t_1}, $W_1^{t_2}$,..., $W_M^{t_2}$, and biases b^{t_1}, $b_1^{t_2}$,..., $b_M^{t_2}$. We denote the label map $y_i \in [0,1]^n$ as

$$y_i = \tau(W_i^{t_2}\rho(W^{t_1}x^c + b^{t_1}) + b_i^{t_2}) \tag{5}$$

So the loss function for labelled parsing dataset becomes

$$\sum_{x^l} C(y^l, \hat{y}^l) = ||\hat{Y}^l - Y^l||_F^2 \tag{6}$$

Where $Y^l = \{y_i^l\}$ and $\hat{Y}^l = \{\hat{y}^l\}$ are the set of outputs and the set of ground truth labels.

Now we use the current DDN to train the unlabelled person re-identification dataset. The training follows the hypothesis of low-density separation [33]. Specifically, the object of our training is to make the probability that the output

Figure 3: The test results of the VIPeR dataset for person re-identification with DDN.

tends to a class close to 1, and the sum of the probabilities toward other classes tend to be zero. We define the loss as an entropy

$$\sum_{x^u} E(y^u) = -\sum_{i=1}^{N} y_i^u \ln(y_i^u), \tag{7}$$

Where N denotes the number of samples and y_i^u is the output. Finally we get the semi-supervised DDN loss function

$$L = ||\hat{Y}^l - Y^l||_F^2 - \sum_{i=1}^{N} y_i^u \ln(y_i^u) \tag{8}$$

3.2.2. Background noise removal

After segmenting the images of pedestrians into a set of semantic regions, we propose a method based on the use of sliding windows and color matching to remove the cluttered environment around the pedestrians. At first, every image is divided into a grid of local patches, and then the background is masked through computing the overlap rate between the mask and patches. This mask is preset, such as the pedestrian's upper body is green and the background is dark blue. The whole process is shown in Fig. 4.

Every image is divided into patches of size 10×10, with a step size of 5 pixels.

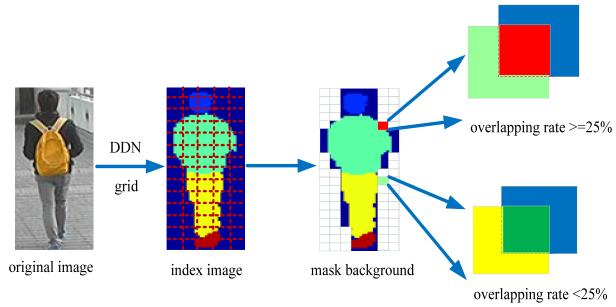

Figure 4: The process of masking the background based on sliding windows and color match.

To determine if a patch will be masked, we apply with following equation:

$$c(P_{ij}) = \frac{u(M - P_{ij})}{x_p * y_p}, \tag{9}$$

where P_{ij} indicates the patch at the i-th row and j-th col of the image, $i, j \in N_+$, $\{i, j | i <= m, j <= n\}$. $c(P_{ij})$ denotes the overlapping rate between the sliding mask M and the P_{ij} and $u(x)$ is indicated as the number of non-zero elements in matrix x. x_p and y_p represent the number of patches in the horizontal and vertical direction, respectively. The patches for which $c(P_{ij}) <= 25\%$ are reserved, whereas others are masked. Because the background of the same pedestrian often changes under different cameras, background noise removal focuses features on pedestrian patches by removing background patches, making feature matching more accurate. We define all the reserved patches of each image as the set $S1$.

3.3. Secondary selection

The primary selection may cause two problems, one is feature drift and the other is patch unbalance. The location of most similar patches in different images changes cross different camera views. The number of pedestrian patches selected from different images may not be same, which may lead to different feature lengths. Person saliency is distinctive and reliable in pedestrian matching across disjoint camera views. If the patches of two images from the same per-

son are matched, the saliency values of these patches should be similar to each other, regardless of their location. In addition, the number of patches is easily controlled by saliency scores, so as to keep the features consistent in length.

3.3.1. Saliency detection

Based on human focus of attention [28], salient regions are defined with the following properties: 1) making the pedestrian more distinctive than other distractors; 2) being reliable to search for the same pedestrian across different

Figure 5: Silent region could be the part of the human body or the decorations the person carries. The salient regions are circled with the yellow dotted lines.

camera views. Compared with the abstract features, it's easier for a human to identify the same person, because if the salient region occurs in one camera view, it usually remains salient in another camera view. For example, in Fig. 5, a human would easily identify that there is a red bag on the shoulder of person $p1$, $p2$ carries a yellow bag, $p3$ has a red umbrella in his hand while $p4$ holds a green parcel in his hand.

A reliable approach to map the salient regions is saliency learning [28]. It divides pedestrians into different parts and manually merges super-pixels that are coherent in appearance. Then the segmented body part is randomly selected and presented to a labeler. The labeler is allowed to select the most likely image from the list based on visual perception. However, this method requires a significant amount of man hours, so it is impractical for large datasets. In this paper, GBVS [21] is employed to automatically detect the salient regions.

Moreover, to reduce the huge cost of matching time, we select only 25 patches whose saliency scores are relatively higher than others. This number is the empirical result of the compromise between computation time and matching accuracy.

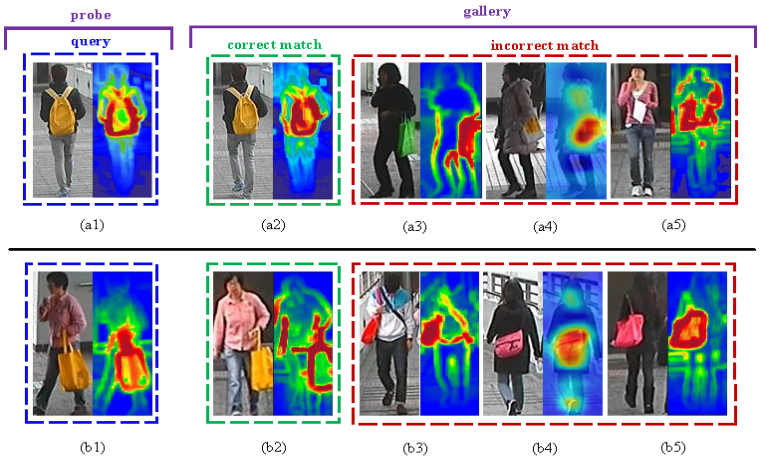

Figure 6: Illustration of saliency detection with the GBVS algorithm and the saliency map of the pedestrian image is shown. Best viewed in color.

As we can see from Fig. 6 that the salient region is detected by the GBVS algorithm that computes bottom-up saliency maps which show a remarkable consistency with the attentional deployment of human subjects. In many cases, different persons from different camera views have different spatial distribution, whereas the salient region of the same pedestrian under different camera views is discriminative from others. For example, the salient region in $(a1)$ is a backpack. The similar salient region also exists in $(a2)$, so $(a2)$ is the correct match of $(a1)$. There is a green bag hanging on the pedestrian's arm in $(a3)$. The yellow bag on the shoulder of the woman in $(a4)$ is very eye-catching. While the woman in $(a5)$ holds a white paper in his hand. They are all the incorrect matches of $(a1)$. For the same reason, $(b2)$ is the correct match of $(b1)$. $(b3)$, $(b4)$, $(b5)$ are the incorrect matches of $(b1)$.

3.3.2. Saliency-guide matching

We hope to match the features of similar patches in different images, but in fact, due to the change of pose and views under different cameras, they are offset in position, and even some patch features may shift from pedestrian to background. Now we change the feature matching from location-guide to saliency-guide, which effectively solves the problem of mismatching caused by feature drift.

At first, the image is constructed as a Gaussian pyramid to extract multi-scale features in the down-sampling process.

$$R(\sigma) = I(x, y) \otimes G(x, y, \sigma), \tag{10}$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{(-\frac{x^2+y^2}{2\sigma^2})}, \tag{11}$$

where $R(\sigma)$ is the initial feature map using the GBVS model, $I(x, y)$ represents the image, $G(x, y, \sigma)$ denotes Gaussian pyramid, σ is the scale factor or bandwidth of Gaussian pyramid and $\otimes$ in Eq. 10 denotes the convolution operator.

Secondly, the activation maps are formed using the feature maps, and the most important thing is to construct the Markov matrix. We assume that the scale of the feature graph is constant. In other words, we ignore the scale σ. We then define the dissimilarity of $R(x, y)$ and $R(p, q)$ as

$$d((i, j)||(p, q)) \triangleq \left| \log \frac{R(i, j)}{R(p, q)} \right|, \tag{12}$$

where $R(x, y)$ and $R(p, q)$ represent the feature value of the pixels at (i, j) and (p, q), respectively. We obtain the fully-connected directed graph G_A through connecting every node of the lattice R, labelled with the indices (i, j) or (p, q). The directed edge from node (i, j) to node (p, q) will be assigned a weight

$$w_1((i, j), (p, q)) \triangleq d((i, j)||(p, q)) \cdot F(i - p, j - q), \tag{13}$$

$$F(a, b) \triangleq \exp\left(-\frac{a^2 + b^2}{2\sigma^2}\right), \tag{14}$$

where σ is a constant which denotes the free parameter. The Markov chain is defined on directed graph G_A. We normalize the weights on the edges of G_A to be 1. Now the stationarity of the Markov chain is utilized to obtain the probability that the state node transforms to another, thereby estimating the saliency of the directed graph and obtaining the saliency map A.

Finally, we normalize the saliency map A, and construct the directed graph G_N. We redefine a Markov chain on G_N, and introduce an edge from (i, j) to (p, q) with weight:

$$w_2((i,j),(p,q)) \triangleq A(p,q) \cdot F(i-p, j-q), \tag{15}$$

where A denotes the final saliency map; every element inside represents the saliency value of the pixel in this position. The size of A is the same as the original image. Every image is divided into patches of size 10×10 with a step size of 5 pixels, and the patches which have the higher saliency value are selected by

$$s\left(\{p_A\}(i,j)\right) = average\left(\{p_A\}(i,j)\right), \tag{16}$$

where $p_A(i, j)$ denotes the patch at the i-th row and j-th column of A, $s(p_A(i, j))$ is the average saliency value of $p_A(i, j)$. We use 0.6 as the empirical value of $s(p_A(i, j))$. The patches are reserved corresponding to the original image where $s(p_A(i, j)) >= 0.6$, while others are removed. We define all the reserved patches of each image as the set $S2$.

3.3.3. Aligned patch sequences

After primary selection and secondary selection, we obtain corresponding patch sets $S1$ and $S2$ respectively. Now we define their intersection $S = S1 \cap S2$ as a set of reliable patches. Due to the different views under different cameras, the proportions of pedestrian and background are also different. Some images have more pedestrian patches than background patches, while others are the opposite. This problem of patch unbalance results in a different number of reliable patches selected per image, and correspondingly different lengths of extracted features.

In order to ensure that the dimension of the local features extracted from each image is the same, 25 patches are selected from each image from camera A which have a relatively high saliency value within the set S. Using the priori saliency spatial distribution of these patches, we find 25 patches corresponding to the previous 25 patches from each image from camera B with the nearest neighbor classifier for saliency.

Now the similarity of saliency between the patch pairs for different images across disjoint camera views is defined as

$$sim_{saliency}(P^{A,u}, P^{B,v}) = \exp(-\frac{d(p_i^{A,u}, p_j^{B,v})^2}{2\sigma_d^2}) \tag{17}$$

Saliency patches of a pedestrian image are represented as $P^{A,u} = \{p_i^{A,u} | i = 1, 2..., 25\}$, where (A, u) denotes the u-th image under camera A, i denotes the position of the patch in this image, and $p_i^{A,u}$ is the saliency vector of the patch. $d(\cdot)$ is the Euclidean distance, and σ_d is a bandwidth parameter. Finally, we get the corresponding patches of images from camera B.

$$I^{B,u} = find(\min(sim_{saliency}(P^{A,u}, P^{B,v}))) \tag{18}$$

The special form is

$$I_i^{B,u} = find(\min(\exp(-\frac{d(p_i^{A,u}, p_j^{B,v})^2}{2\sigma_d^2}))), \tag{19}$$

where $find(\cdot)$ denotes finding the indexes of patches of an image from camera B according to the saliency matching with the patches of the image from camera A. $I_i^{B,u}$ is an element of $I^{B,u}$ which denotes the indexes set as mentioned above, $i \in \{1, 2...25\}$.

3.4. Feature fusion and metric learning

In order to overcome the shortcomings of either of the methods and take advantage of them, the global features and local descriptors are fused in the process of metric learning, so that we can clearly separate the different pedestrians.

3.4.1. Feature extraction and fusion

The features we fuse consist of one global feature (LOMO) and three local features (PHOG, HSV, SIFT). We adopt the strategy of combining global feature and local features to compensate for inherent errors of saliency detection which may result in mismatching. Specifically, PHOG contains oriented gradient, HSV reflects color distribution, and SIFT captures extreme points in images. As about the global feature (LOMO), although it also contains some color information, it reflects the color distribution of the whole image, using image pair matching rather than saliency-guide patch pair matching. In a word, they complement each other without redundancy.

The LOMO algorithm analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Besides, to handle illumination variations, the Retinex transform [3] and a scale invariant texture operator are applied. To make person re-identification easier than using original images, we apply the HSV color histogram to extract features which has $8 \times 8 \times 8$ bins $= 512$ dimensions. The Scale Invariant Local Ternary Pattern (SILTP) [34] descriptor is also extracted for reducing the impact of illumination invariant. SILTP is an improved operator over the well-known Local Binary Pattern (LBP) [35]. We utilize sliding windows with a size of 10×10 pixels and an overlapping step of 5 pixels to locate local patches in 128×48 pixel images. Two scales of SILTP histograms ($SILTP_{4,3}^{0.3}$ and $SILTP_{4,5}^{0.3}$) are extracted, and the dimension of SILTP is $3^4 \times 2 = 81$. A three-scale pyramid representation is built for utilizing the multi-scale information, which down-samples the original 128×48 image by two 2×2 local average pooling operations and then repeats the above feature extraction procedure. So the final feature has ($8 \times 8 \times 8$ color bins + $3^4 \times 2$ SILTP bins) $\times$ ($24 + 11 + 5$ horizontal groups) $= 26960$ dimensions.

On the other hand, the PHOG, HSV histogram and SIFT features are extracted from every selected patch. PHOG is the Pyramid Histogram of Oriented Gradient, which is an effective descriptor for classification; it is the splicing of

the HOG features at different scales. In this work, the number of layers of pyramids is $L = 3$, and the number of bins of gradient division is $n = 8$. The dimension of PHOG features is $(1 + 4 + 16 + 64) \times 8 = 680$. Color histogram is a significant descriptor that performs outstandingly in the recognition task. To obtain the HSV histogram features, the RGB image is converted to a HSV image at first. The dimension of the HSV histogram feature is $8 \times 8 \times 8 = 512$. Besides, we extract the SIFT features that has 128 dimensions. The schematic representation of feature extraction is shown in Fig. 1.

After finishing the feature extraction, we obtain features with the dimension of 26960+$(680 + 512 + 128) \times 20 = 53360$. Before concatenating them, we fuse them based on metric learning. Due to the diversity of our features and the complexity of the processing process, we count the time consumption of feature extraction and algorithm execution. First, we count the time of feature extraction and fusion on the VIPeR dataset, and the average time for each image was 46.6ms. The experiment is repeated 10 times and averaged (i7-6700 CPU, 2.60GHz, Matlab, Windows). Then we perform our algorithm on CUHK03 dataset including semi-supervised DDN training, which takes a total 2586,463ms $\approx$ 43m 6s, about half the time of FFN [14] (Titan xp, 12GB video memory, GPU, Linux).

3.4.2. Metric learning

We define $dist_{i,j}$ as the distance between the features x_i and x_j cross different camera views.

$$\begin{aligned} &dist^2(x_i, x_j) = ||x_i - x_j||_2^2 \\ &= w_1 \cdot dist_{ij,1}^2 + w_2 \cdot dist_{ij,2}^2 + ... + w_d \cdot dist_{ij,d}^2 \\ &=(x_i - x_j)^{\mathrm{T}} W (x_i - x_j), \end{aligned} \tag{20}$$

where $w_i \geq 0$, $W = diag(w)$ is a diagonal matrix, and $(W)_{ii} = w_i$. W can be determined by learning. d denotes the dimension of the feature which is equal to 53360 in this paper. We replace W with a common semi-definite symmetric matrix M, so we get Mahalanobis distance.

$$dist_{mah}(x_i, x_j) = (x_i, x_j)^{\mathrm{T}} M (x_i, x_j) = ||x_i - x_j||_M^2, \tag{21}$$

M denotes the metric matrix which is obtained through metric learning. Note that M is the semi-definite symmetric matrix. M is directly embedded into the evaluation of the neighbor classifier, and we obtain M through optimizing the performance of the evaluation. Now we discuss the acquisition of M with the Neighbourhood Component Analysis (NCA) as an example.

Neighbour classifiers use the majority voting method when making a decision. Each sample in the neighbourhood casts one vote, and the samples outside the field casts zero votes. For sample x_j, the probability of its effect on x_i classification is

$$p_{i,j} = \frac{\exp(-||x_i - x_j||_M^2)}{\sum_l \exp(-||x_i - x_l||_M^2)}, \tag{22}$$

where l is the number of the samples. As can be seen from Eq. 22, $p_{i,j}$ is the largest when $i = j$. If we recognize the maximum accuracy as an optimal object, the accuracy based on leave-one-out (LOO) is computed as follows

$$p_i = \sum_{j \in \Omega_i} p_{ij}, \tag{23}$$

where Ω_i represents the set of subscripts that belong to the same class as x_i. The accuracy for the entire sample set is

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \sum_{j \in \Omega_i} p_{ij} \tag{24}$$

Then we substitute the Eq. 22 into the 24 and make $M = PP^{\mathrm{T}}$, we get the NCA optimal object

$$\min_P = 1 - \sum_{i=1}^{m} \sum_{j \in \Omega_i} \frac{\exp(-||P^{\mathrm{T}}x_i - P^{\mathrm{T}}x_j||_2^2)}{\sum_l \exp(-||P^{\mathrm{T}}x_i - P^{\mathrm{T}}x_l||_2^2)} \tag{25}$$

Through solving Eq. 25, we obtain the metric matrix M that maximizes the accuracy of the neighbour classifier.

Finally, we get Cumulative Match Characteristic (CMC) curves of person re-identification. Using several different metric methods, we do experiment on different datasets to prove that our proposed method is more effective than many state-of-the-art methods.

4. Experimental results

There are several existing challenging benchmark datasets for person re-identification. In this work, we perform experiments using six datasets, VIPeR [1], PRID2011 [37], CUHK01 [38], CUHK03 [36], PRID 450S [39], iLIDS-VID [40], which are public benchmarks available to conduct experiments. We emphasize that our approach can effectively solve the problem of insufficient samples in actual scenarios. This is why our method has not been tested in large datasets such as Market1501 [16], DukeMTMC-ReID [17], and MSMT17 [18].

4.1. Parameters and implementation details

The parameter settings in this paper are shown in the Table 1. In addition, we perform fine tuning on the basis of original DDN, so the initialization of parameters and bias are the result of previous training. Two scales of center/surround Retinex is used for image preprocessing when LOMO features are extracted. For all the experiments, we repeat the procedure 10 times to calculate an average performance.

Table 1:

Parameter settings

Paraneters	Values	Descriptions
s_p	10×10	the size of patches
s_w	10×10	sliding windows
c_{thr}	0.25	the threshold of overlapping rate $c(P_{i,j})$
n_P	25	the number of patches we selected based on saliency detection
s_{thr}	0.6	the threshold of saliency value
σ	0.5	the scale factor of Eq. 11
σ_d	0.5	bandwidth parameter of Eq. 17

4.2. Comparison with state-of-the-art methods

We perform a number of experiments and the results show that the proposed algorithm achieves better performance than many of the existing methods. In

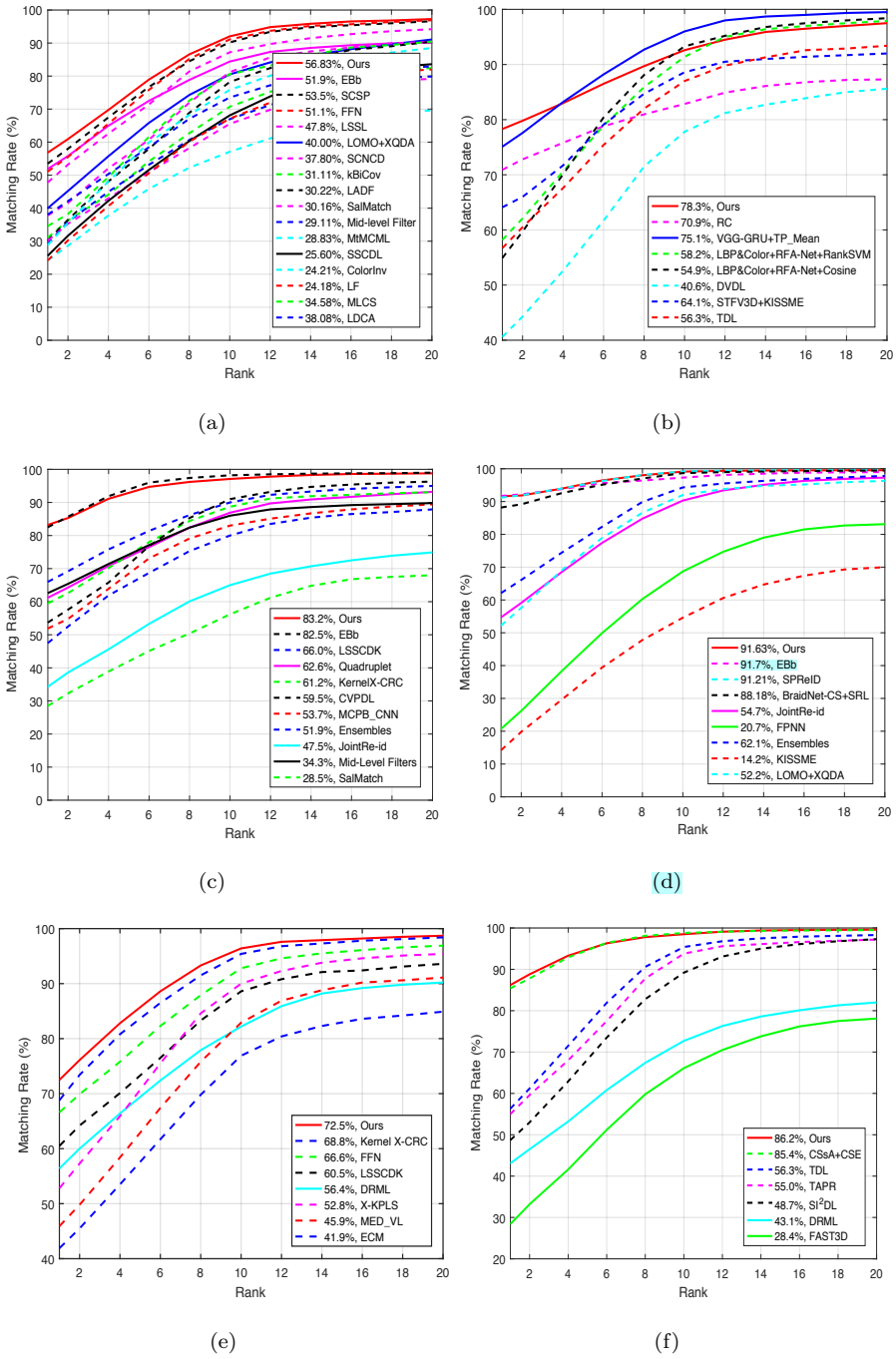

Figure 7: CMC curves of VIPeR, PRID2011, CUHK01, CUHK03, PRID 450S, iLIDS-VID datasets.

order to demonstrate the advantages of our method in the case of insufficient samples, we also compared the with many deep learning methods, which are listed separately in the tables. Fig. 7 shows the CMC curves for different methods on every dataset. The red solid lines represent the results of our algorithm. It can be seen from Fig. 7 that our method has the highest matching rate.

Table 2:

TOP r RANK MATCHING ACCURACY (%) ON VIPeR DATASET.

Method	**rank=1**	**rank=10**	**rank=20**	**Reference**
Ours	**56.83**	**92.03**	**97.27**	Proposed
FFN [14]	51.1	91.4	96.9	2016 WACV
EBb [41]	51.9	84.8	90.2	2018 CVPR
MLCS [42]	34.58	80.59	90.43	2017 TCSVT
LDCA [11]	38.08	73.52	82.91	2017 CVPR
SCSP [43]	53.5	90.2	96.6	2016 CVPR
LSSL [44]	47.8	87.6	94.2	2016 AAAI
LOMO+XQDA [3]	40.00	80.51	91.08	2015 CVPR
SCNCD [8]	37.80	81.20	90.40	2014 ECCV
kBiCov [5]	31.11	70.71	82.45	2014 IVC
SalMatch [20]	30.16	65.54	79.15	2013 ICCV
Mid-level Filter [9]	39.11	65.95	79.87	2014 CVPR
SSCDL [45]	25.60	68.10	83.60	2014 CVPR
MtMCML [46]	28.83	75.82	88.51	2014 TIP
ColorInv [35]	24.21	57.09	69.65	2013 TPAMI
LF [6]	24.18	67.12	82.00	2013 CVPR

4.2.1. Experiments on VIPeR

The VIPeR dataset contains two cameras, each of which captures one image per person. It also provides the viewpoint angle for each image. It has been used by many researchers and is still one of the most challenging datasets. The VIPeR dataset contains 632 pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions. It is randomly split into two subsets containing the same number of pictures for training and test respectively.

We evaluated the proposed algorithm and several state-of-the-art algorithms, Fig. 7(a) shows the results of the comparisons through CMC curves on the VIPeR dataset. The cumulative matching scores (%) at rank 1, 10, and 20 are listed in Table 2. From Table 2 it can be seen that our method is superior to all compared state-of-the-arts, surpassing the 2nd best method by 3.33% (56.83 - 53.5) in Rank-1, 0.63% (92.03 - 91.4) in Rank-10, and 0.37% (97.27 - 96.9) in Rank-20. Compared to eliminating background-bias (EBb) method, our method improves the rank-1 by 4.93%, rank-10 by 7.23%, and rank-20 by 7.07%. It indicates the superiority of primary patch selection by background noise removal. Compared to the deep learning method FFN, our method improves the Rank-1 by 5.73%. This indicates that in the single-shot case, our feature fusion strategy is more effectively than FFN that fuses deeply learning features with multiple hand-crafted features.

4.2.2. Experiments on PRID2011

The PRID2011 dataset consists of images extracted from multiple person trajectories recorded from two different, static surveillance cameras. Images from these cameras contain a viewpoint change and a stark difference in illumination, background and camera characteristics. The PRID dataset has 385 trajectories from camera A and 749 trajectories from camera B. Among them, only 200 people appear in both cameras.

Fig. 7(b) shows the results of the comparisons through CMC curves on the PRID2011 dataset. The cumulative matching scores (%) at rank 1, 10, and 20 are listed in Table 3. From Table 3 it can be seen that our method is

Table 3:

TOP r RANK MATCHING ACCURACY (%) ON PRID2011 DATASET.

Method	**rank=1**	**rank=10**	**rank=20**	**Reference**
Ours	**78.3**	92.6	97.5	Proposed
VGG-GRU+TP_Mean [47]	75.1	**97.5**	**99.5**	2017 ICIC
LBP&Color+RFA-Net+RankSVM [48]	58.2	93.4	97.9	2017 ECCV
LBP&Color+RFA-Net+Cosine [48]	54.9	93.7	98.4	2017 ECCV
RC [49]	70.9	82.7	87.3	2018 CVPR
DVDL [50]	40.6	77.8	85.6	2015 ICCV
STFV3D+KISSME [2]	64.1	89.9	92.0	2012 CVPR
TDL [51]	56.7	87.6	93.4	2016 CVPR

superior to all compared state-of-the-arts in Rank-1. It surpasses the 2nd best VGG-GRU+TP_Mean by 3.2% (78.3 - 75.1) in Rank-1. Although it is 4.9% (97.5-92.6) and 2.0% (99.5-97.5) lower than VGG-GRU+TP_Mean in rank-10 and rank-20 respectively, it is not inferior to the suboptimal LBP&Color+RFA-Net+Cosine. As can be seen that compared with multiple hits, our method has a obvious advantage in accuracy of one hit.

4.2.3. Experiments on CUHK01

The CUHK01 dataset contains two images for every identity from each camera. This dataset has one pair of disjoint cameras and the image quality of this dataset is relatively good. It contains 971 persons captured from two camera views. Camera A captures the frontal or back views of pedestrians while camera B captures them in a side view.

The CMC curves of comparison with other algorithms are described in Fig. 7(c). All the corresponding data are recorded in Table 4. The proposed method achieved 83.2% at rank-1, which slightly outperforms the second one EBb with an improvement of 0.7% (83.2 - 82.5). The proposed method is comparable to the most advanced algorithms EBb at rank-15 and rank-20, which achieved accuracy at 98.4% and 98.8%, respectively. Compared to SalMatch, we far surpassed it in all the results. It shows that using background noise removal,

Table 4:

TOP r RANK MATCHING ACCURACY (%) ON CUHK01 DATASET.

Method	**rank=1**	**rank=10**	**rank=15**	**rank=20**	**Reference**
Ours	**83.2**	97.1	98.4	98.8	Proposed
Quadruplet [15]	62.6	86.0	88.9	89.8	2017 CVPR
MCPB_CNN [22]	53.7	91.0	95.4	96.3	2016 CVPR
JointRe-id [13]	47.5	80.0	86.8	87.9	2015 CVPR
EBb [41]	82.5	**98.2**	**98.7**	**99.0**	2018 CVPR
LSSCDK [52]	66.0	90.0	93.3	95.0	2016 CVPR
Kernel X-CRC [53]	61.2	87.3	91.2	93.2	2019 JVCIR
CVPDL [54]	59.5	89.7	91.7	93.1	2015 ICOAI
Ensembles [55]	51.9	83.0	88.5	89.4	2015 CVPR
Mid-Level Filters [9]	34.3	65.0	71.2	74.9	2014 CVPR
SalMatch [20]	28.5	55.7	66.1	68.0	2014 ICCV

patch selection, feature fusion techniques is far more effective than just using saliency matching.

4.2.4. Experiments on CUHK03

The CUHK03 is one of the highest cited person re-identification dataset which consists of five different pairs of camera views, and the number of pictures in this dataset exceeds 14,000. There are 13,164 bounding boxes detected by a Deformable Part Model (DPM) of 1,467 different identities in CUHK03 dataset.

Fig. 7(d) and Table 5 provide the matching results of all the compared algorithms. It can be seen that the proposed method is superior to the 2nd best SPReID by 0.6% (91.8 - 91.2) in Rank-1, and ties with SPReID in Rank-20. SPReID extracts local features from human body parts obtained by human semantic parsing. Both SPReID and our method use parsing for person re-identification. SPReID focuses on parsing to make pedestrian body segmentation more accurate, while we focus on selecting reliable patch sequences for precise matching. So if we learn from SPReID, the background noise will be smaller. Then the selected patch sequence will be theoretically more reliable,

Table 5:

TOP r RANK MATCHING ACCURACY (%) ON CUHK03 DATASET.

Method	**rank=1**	**rank=10**	**rank=15**	**rank=20**	**Reference**
Ours	**91.8**	99.1	99.4	**99.6**	Proposed
BraidNet-CS+SRL [56]	88.2	98.7	99.2	99.5	2018 CVPR
JointRe-id [13]	54.7	91.5	96.8	97.3	2015 CVPR
FPNN [36]	20.7	68.7	80.1	83.1	2014 CVPR
SPReID [57]	91.2	**99.2**	**99.5**	99.6	2018 CVPR
EBb [41]	91.7	-	98.7	99.0	2018 CVPR
Ensembles [55]	62.1	94.3	97.2	97.8	2015 CVPR
KISSME [2]	14.2	52.6	66.4	70.0	2012 CVPR
LOMO+XQDA [3]	52.2	92.1	95.6	96.3	2015 CVPR

and finally the accuracy will be improved.

4.2.5. Experiments on PRID 450S

The PRID 450S dataset contains 450 pairs of single-shot pedestrian images, which are captured from two adjacent cameras. It is another challenging dataset, similar to the VIPeR dataset, for background interference, partial occlusion and viewpoint changes.

Table 6:

TOP r RANK MATCHING ACCURACY (%) ON PRID 450S DATASET.

Method	**rank=1**	**rank=10**	**rank=15**	**rank=20**	**Reference**
Ours	**72.5**	**96.4**	**97.8**	**98.7**	Proposed
FFN [14]	66.6	92.8	96.6	96.9	2016 WACV
Kernel X-CRC [53]	68.8	95.9	97.3	98.4	2019 JVCIR
LSSCDK [52]	60.5	88.6	92.2	93.6	2016 CVPR
DRML [58]	56.4	82.2	88.9	90.2	2016 ICIP
X-KPLS [59]	52.8	90.0	94.8	95.4	2017 ICPR
MED_VL [60]	45.9	82.9	89.8	91.1	2016 AAAI
ECM [4]	41.9	76.9	82.6	84.9	2015 WACV

We evaluated the proposed approach by comparing the state-of-the-art approaches on the PRID 450S dataset. This evaluation was conducted using the images of detected persons. It can be seen from Table 6 that our method is superior to all compared state-of-the-arts, surpassing the 2nd best Kernel X-CRC by 3.7% (72.5 - 68.8) in Rank-1, 0.5% (96.4 - 95.9) in Rank-10, 0.5% (97.8 - 97.3) in Rank-15 and 0.3% (98.7 - 98.4) in Rank-20. Compared to Kernel X-CRC, our local features contain gradient, color, and extreme points, not just the color model as Kernel X-CRC does. It indicates the superiority of diverse features. Fig. 7(e) describes the matching results of all the compared algorithms on the PRID 450S dataset.

4.2.6. Experiments on iLIDS-VID

The iLIDS-VID dataset involves 300 different pedestrians observed across two disjoint camera views in a public open space. It comprises 600 image sequences of 300 distinct individuals, with one pair of image sequences from two camera views for each person. Each image sequence has variable length ranging from 23 to 192 image frames, with an average of 73 frames. The iLIDS-VID dataset is very challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and random occlusions. Fig. 7(f) and Table 7 show the matching results of all the

Table 7:
TOP r RANK MATCHING ACCURACY (%) ON ILIDS-VID DATASET.

Method	**rank=1**	**rank=10**	**rank=15**	**rank=20**	**Reference**
Ours	**86.2**	98.5	**99.4**	**99.6**	Proposed
CSsA+CSE [61]	85.4	**98.8**	99.2	99.5	2018 CVPR
TDL [51]	56.3	95.6	97.9	98.3	2016 CVPR
TAPR [62]	55.0	93.8	96.9	97.2	2016 ICIP
SI^{2}DL [63]	48.7	89.2	96.6	97.3	2016 IJCAI
DRML [58]	43.1	72.7	80.0	82.0	2016 ICIP
FAST3D [64]	28.4	66.7	75.2	78.1	2016 ICIP

compared methods. We can see that our method is superior to all the state-of-the-art methods, surpassing the 2nd best method by 0.8% (86.2 - 85.4) in Rank-1, 0.2% (99.4 - 99.2) in Rank-15, and 0.1% (99.6 - 99.5) in Rank-20. And it's only 0.3% (98.8 - 98.5) lower than CSsA+CSE. Moreover, the proposed method far surpasses recent methods (TDL, TAPR, SI^2DL and DRML) in all results. These validate that a combination of techniques may be more effective than just using a single technique for person re-identification.

4.3. Ablation analysis

To further illustrate the rationality of each step of our process, we conduct ablation experiments for our method on VIPeR dataset. We verify the roles of four key parts of our algorithm, including background noise removal, saliency detection, local features, and global features through experiments. We take turns to remove key part (C1-C4) and compare with the complete method (C5), as shown in Table 8. It can be seen that Rank-1 and Rank-10 results of all incomplete methods are inferior to the complete method, which implies the importance of the default part.

Table 8:

Experimental results for different configurations on VIPeR datasets.

Config.	Background noise removal	Saliency detection	Local features	Global features	Rank-1	Rank-10
C1	×	√	√	√	50.26	90.54
C2	√	×	√	√	43.31	86.52
C3	×	×	×	√	40.00	80.51
C4	√	√	√	×	51.74	89.62
C5	√	√	√	√	56.83	92.03

Firstly, we remove the process of background noise removal (C1), in other words, there is no primary selection, $S1 = S$. As can be seen from Table 8 that due to background noise, the performance degrades 6.57% and 1.49% for the Rank-1 and Rank-10 accuracy, respectively. Then we investigate the role of saliency detection (C2). We eliminate secondary selection based on saliency detection, and select 25 patches from $S1$ according to the principle of proximity.

Specifically, $P(\cdot)$ in Eq (17) is redefined as the coordinate of the patch rather than saliency value. It can be seen that the accuracy drops sharply, degrading 12.15% and 6.86% for the Rank-1 and Rank-10, respectively. Next, we remove the local features (C3),which means that only global feature LOMO is used and there is no patch selection. The feature representation is the same as [3]. Rank-1 and Rank-10 become 40.00% and 80.51%. It also proves the necessity of the patch selection method we proposed. Finally, global feature (C4) is removed to demonstrate its role in compensating for inherent errors in saliency detection. As can be seen from Table 8 that without the assistance of LOMO, there is s slight decrease in accuracy, with rank-1 and rank-10 dropping by 5.09% and 2.42% respectively.

4.4. Comparison with the most relevant methods

In this paper, the three key points of the proposed approach are utilizing local descriptors with the global features, background noise removal, and saliency detection. There are three corresponding algorithms, including LOMO+XQDA, saliency learning, and super-pixel segmentation for person re-identification. The following will introduce their differences with our proposed method and the experimental results.

Table 9:

PERSON RE-ID MATCHING RATES(%) AT DIFFERENT RANKS ON VIPER, CUHK01, AND CUHK03 DATASETS

	VIPeR				CUHK01				CUHK03			
Method	rank@1	10	15	20	1	10	15	20	1	10	15	20
Ours	**56.8**	**92.0**	**96.2**	**97.2**	**69.2**	**92.8**	**96.1**	**97.8**	**68.2**	**95.2**	**97.8**	**98.4**
LOMO+XQDA [3]	40.0	80.5	88.3	91.0	61.8	86.5	91.5	93.7	52.2	92.1	95.4	96.3
Saliency learning [28]	44.1	81.8	88.4	91.2	28.5	55.7	66.4	68.0	56.8	93.8	96.2	97.5
BackSub-reid [26]	27.2	64.2	75.2	77.8	19.2	44.8	65.8	68.7	40.2	72.1	84.5	86.4

We compared the matching rate with LOMO features, saliency learning, and the method based on super-pixel segmentation for person re-identification on VIPeR, CUHK01, and CUHK03 datasets. Table 9 records the results of the

experiments which indicates that the proposed method is always better than others at rank-1.

4.5. Parameter analysis of the proposed method

4.5.1. The threshold of overlapping rate in background noise removal

The proposed system achieves accurate salient person re-identification through background removal based on super-pixel segmentation. However, in this paper, we apply the pedestrian parsing via a DDN network to achieve the background removal. The experiments show that the proposed method has obvious advantages over the other methods.

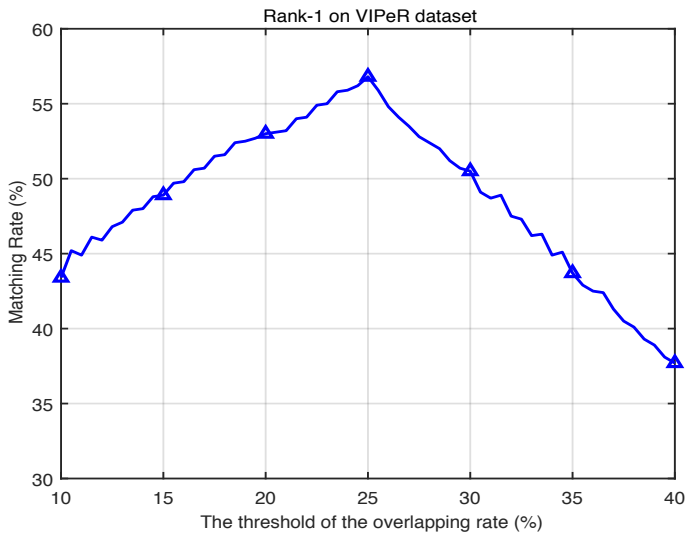

Figure 8: The relationship between matching rate and the threshold of overlapping rate (%) on VIPeR dataset.

In this paper, we take the pedestrian parsing as an important method for removing the background. We parse the pedestrians with the DDN network which allows the background to be removed from the edges of a human. It is an important preprocessing for picking up the pedestrian patches.

In the process of removing the background noise, we set the threshold of overlapping rate $c(P_{i,j})$ to 25%, which was empirically determined after many

experiments. It directly determines whether the patch belongs to the pedestrian or the background. The experimental result for choosing the overlap rate threshold are shown in the Fig. 8. We compared the matching rate when selecting different thresholds ranging from 0.1 to 0.4 at rank-1 on the VIPeR dataset, which shows that 25% as the threshold is appropriate.

4.5.2. The number of selected patches

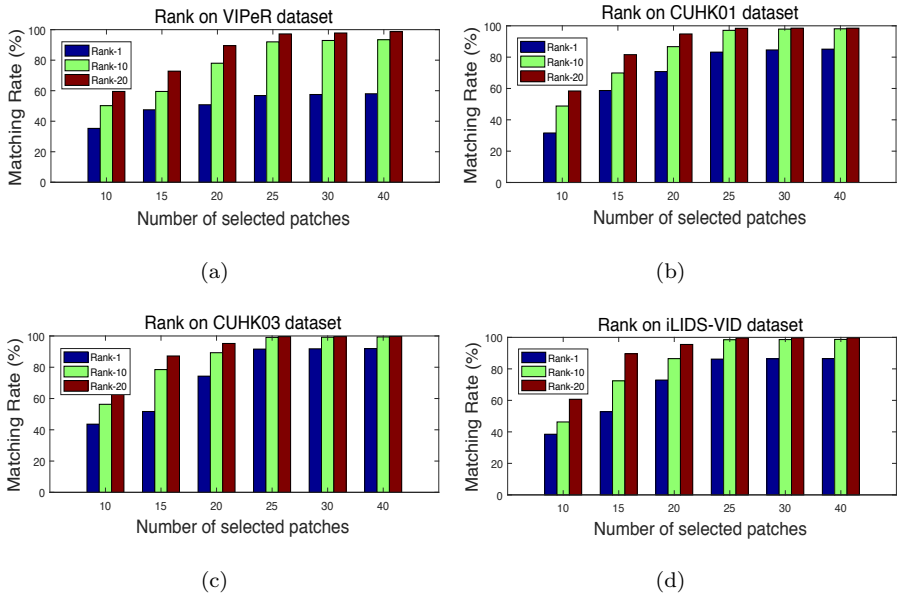

Figure 9: The relationships between matching rate and the number of selected patches on VIPeR, CUHK01, CUHK03 datasets.

The number of patches in S has a great impact on the matching rate and execution efficiency. If the number is too small, effective information will be missed, resulting in lower accuracy. Too many will increase the computation time and reduce the execution efficiency.

Fig. 9 describes the relationship between the number and the matching rate at rank-1 on the VIPeR, CUHK01, CUHK03, and iLIDS-VID datasets. As we can see from Fig. 9, the matching rate increases as the number of selected

patches increases. However, after the number exceeds 25, the rate of growth becomes very slow, while the cost of time is multiplied. Finally, we selected 25 patches, which provides a compromise between computation time and matching accuracy.

5. Conclusions

In this paper, we proposed a new patch selection method based on parsing and saliency detection for person Re-identification. We solve the problem of feature drift and patch imbalance of local features, and effectively compensate for the inherent errors caused by saliency detection by combining local features with global features. It provides more ideas for solving related problems. In addition, our method can effectively deal with the real scenario of insufficient samples, which has a strong engineering application value. It's another highlight of our work.

Acknowledgments

The research is supported by National Natural Science Foundation of China(No.61471110, 61733003), National Key R&D Program of China(No.2017YFC0805005, 2017YFB1301103), Advance Research Project(41412050202), Fundamental Research Funds for the Central Universities(N172608005) and Natural Science Foundation of Liaoning(No.20180520040).

References

[1] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision - ECCV 2008, European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings*, 2008, pp. 262–275.

[2] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[4] X. Liu, H. Wang, Y. Wu, J. Yang, and M. H. Yang, "An ensemble color model for human re-identification," in *Applications of Computer Vision*, 2015, pp. 868–875.

[5] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification ," *Image & Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.

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

[7] X. Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, and J. Y. Yang, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1363–1378, 2017.

[8] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," , vol. 8689, no. 9, pp. 536–551, 2014.

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

[10] Y. Wang, R. Hu, C. Liang, C. Zhang, and Q. Leng, "Camera compensation using a feature projection matrix for person reidentification," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 24, no. 8, pp. 1350–1361, 2014.

[11] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," pp. 7398–7407, 2017.

[12] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *ACM International Conference on Image and Video Retrieval*, 2007, pp. 401–408.

[13] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[14] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in *Applications of Computer Vision*, 2016, pp. 1–8.

[15] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1320–1329.

[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015.

[17] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 3774–3782.

[18] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 79–88.

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

[20] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision*, 2014, pp. 2528–2535.

[21] B. Schlkopf, J. Platt, and T. Hofmann, "Graph-based visual saliency," in *International Conference on Neural Information Processing Systems*, 2006, pp. 545–552.

[22] 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 *Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[23] 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," 2016.

[24] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *International Conference on Pattern Recognition*, 2014, pp. 34–39.

[25] Y. C. Chang, C. K. Chiang, and S. H. Lai, "Single-shot person re-identification based on improved random-walk pedestrian segmentation," in *International Symposium on Intelligent Signal Processing and Communications Systems*, 2013, pp. 1–6.

[26] C. V. Le, Q. N. Hong, T. T. Quang, and N. D. Trung, "Superpixel-based background removal for accuracy salience person re-identification," in *IEEE International Conference on Consumer Electronics-Asia*, 2017, pp. 1–4.

[27] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *IEEE International Conference on Computer Vision*, 2014, pp. 2648–2655.

[28] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 2, pp. 356–370, 2017.

[29] Y. Chen, Z. Huo, and C. Hua, "Multi-directional saliency metric learning for person re-identification," *Iet Computer Vision*, vol. 10, no. 7, pp. 623–633, 2017.

[30] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[31] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on International Conference on Machine Learning*, 2010, pp. 807–814.

[32] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of Machine Learning Research*, vol. 11, no. 12, pp. 3371–3408, 2010.

[33] A. Z. A. O Chapelle, "Semi-supervised classification by low density separation," *AISTATS*, 2005.

[34] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Computer Vision and Pattern Recognition*, 2010, pp. 1301–1306.

[35] T. Ojala and I. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[36] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[37] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person reidentification by descriptive and discriminative classification," in *Scandinavian Conference on Image Analysis*, 2011.

[38] L. Wei, Z. Rui, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*, 2012.

[39] P. M. Roth, M. Hirzer, M. Kstinger, C. Beleznai, and H. Bischof, *Mahalanobis Distance Learning for Person Re-identification*, 2014.

[40] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, no. 12, pp. 2501–2514, 2016.

[41] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," 2018.

[42] L. An, Z. Qin, X. Chen, and S. Yang, "Multi-level common space learning for person re-identification," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[43] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.

[44] Y. Yang, S. Liao, Z. Lei, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3655–3661.

[45] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3550–3557.

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

[47] Q. N. Hong, N. N. Tuan, T. T. Quang, D. N. Tien, and C. V. Le, "Deep spatio-temporal network for accurate person re-identification," in *International Conference on Information and Communications*, 2017, pp. 208–213.

[48] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," pp. 701–716, 2017.

[49] J. Zhou, B. Su, and Y. Wu, "Easy identification from better constraints: Multi-shot person re-identification from reference constraints," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *IEEE International Conference on Computer Vision*, 2015, pp. 4516–4524.

[51] J. You, A. Wu, X. Li, and W. S. Zheng, "Top-push video-based person re-identification," in *Computer Vision and Pattern Recognition*, 2016, pp. 1345–1353.

[52] Y. Zhang, B. Li, H. Lu, A. Irie, and R. Xiang, "Sample-specific svm learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1278–1287.

[53] R. Prates and W. R. Schwartz, "Kernel cross-view collaborative representation based classification for person re-identification," 2016.

[54] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *International Conference on Artificial Intelligence*, 2015, pp. 2155–2161.

[55] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, "Learning to rank in person re-identification with metric ensembles," in *Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[56] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 1470–1478.

[57] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," 2018.

[58] W. Yao, Z. Weng, and Y. Zhu, "Diversity regularized metric learning for person re-identification," in *IEEE International Conference on Image Processing*, 2016, pp. 4264–4268.

[59] R. F. Prates and W. R. Schwartz, "Kernel hierarchical pca for person re-identification," in *International Conference on Pattern Recognition*, 2017.

[60] A. F. O. A. Intelligence, "Association for the advancement of artificial intelligence," *Hyperfine Interactions*, vol. 6, no. 1, 2011.

[61] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," 2018.

[62] C. Gao, J. Wang, L. Liu, J. G. Yu, and N. Sang, "Temporally aligned pooling representation for video-based person re-identification," in *IEEE International Conference on Image Processing*, 2016, pp. 4284–4288.

[63] X. Zhu, X. Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 3552–3558.

[64] Z. Liu, J. Chen, and Y. Wang, "A fast adaptive spatio-temporal 3d feature for video-based person re-identification," in *IEEE International Conference on Image Processing*, 2016, pp. 4294–4298.

Authors

Yixiu Liu received the B.E. degrees from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2016, and is currently working toward the Ph.D. degree at School of School of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests are in the

area of computer vision, including person re-identification, pedestrian tracking.

Yunzhou Zhang received B.S. degree and M.S. degree in Mechanical and Electronic engineering from National University of Defense Technology, Changsha, China in 1997 and 2000, respectively. He received Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009, where he is currently a professor with at the School of Information Science and Engineering, Northeastern University, China. His research interests include intelligent robot, computer vision, and sensor networks.

Sonya Coleman (M11) received the B.Sc. (Hons) degree in mathematics, statistics, and computing and the Ph.D. degree in mathematics from the University of Ulster, Londonberry, U.K., in 1999 and 2003, respectively. She is currently a Lecturer in the School of Computing and Intelligent System, Magee College, University of Ulster. She has more than 50 publications primarily in the field of mathematical image processing, and much of the recent research undertaken by her has been supported by funding from EPSRC award EP/C006283/11, the Leverhulme Trust, and the Nuffield Foundation. Additionally, she is co-investigator on the EU FP7 funded project RUBICON. She is the author or coauthor of over 70 research papers on image processing, robotics, and computational neuroscience. Dr. Coleman was awarded the Distinguished Research Fellowship by the University of Ulster in recognition of her contribution to research in 2009.

Bir Bhanu (M82F95LF17) received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, and the M.B.A. degree from the University of California at Irvine,

Irvine, CA. He was the Founding Professor of electrical engineering with the University of California at Riverside (UCR), Riverside, CA, and served as its first Chair from 1991 to 1994. He has been the Cooperative Professor of computer science and engineering (since 1991), bioengineering (since 2006), and mechanical engineering (since 2008). He served as the Interim Chair of the Department of Bioengineering from 2014 to 2016. He also served as the Director of the National Science Foundation Graduate Research and Training Program in video bioinformatics with UCR. He is currently the Bourns Presidential Chair in engineering, the Distinguished Professor of electrical and computer engineering, and the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory, UCR. He has published extensively and has 18 patents. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human-computer interactions, and biological, medical, military, and intelligence applications. In 1991, he was a Senior Honeywell Fellow with Honeywell Inc. He is a Fellow of AAAS, IAPR, SPIE, and AIMBE.

Shuangwei Liu received B.S. degree in automation from Northeastern University, Shenyang, China, in 2017. He is a graduate student in pattern recognition and intelligent systems in Northeastern University. He majors in computer vision and image processing, especially deep learning methods and person re-ID.

*Biography of the author(s)
Click here to download Biography of the author(s): Biography.doc

	Yixiu Liu received the B.E. degrees from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2016, and is currently working toward the Ph.D. degree at School of School of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests are in the area of computer vision, including person re-identification, pedestrian tracking.
	Yunzhou Zhang received B.S. degree and M.S. degree in Mechanical and Electronic engineering from National University of Defense Technology, Changsha, China in 1997 and 2000, respectively. He received Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009, where he is currently a professor with at the School of Information Science and Engineering, Northeastern University, China. His research interests include intelligent robot, computer vision, and sensor networks.
	Sonya Coleman received the B.Sc. (Hons) degree in athematics, statistics, and computing and the Ph.D. degree in mathematics from the University of Ulster, Londonberry, U.K., in 1999 and 2003, respectively. She is currently a Lecturer in the School of Computing and Intelligent System, Magee College, University of Ulster. She has more than 50 publications primarily in the field of mathematical image processing, and much of the recent research undertaken by her has been supported by funding from EPSRC award EP/C006283/11, the Leverhulme Trust, and the Nuffield Foundation. Additionally, she is co-investigator on the EU FP7 funded project RUBICON. She is the author or coauthor of over 70 research papers on image processing, robotics, and computational neuroscience. Dr. Coleman was awarded the Distinguished Research Fellowship by the University of Ulster in recognition of her contribution to research in 2009.
	Bir Bhanu received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, and the M.B.A. degree from the University of California at Irvine, Irvine, CA. He was the Founding Professor of electrical engineering with the University of California at Riverside (UCR), Riverside, CA, and served as its first Chair from 1991 to 1994. He has been the Cooperative Professor of computer science and engineering (since 1991), bioengineering (since 2006), and mechanical engineering (since 2008). He served as the Interim Chair of the Department of Bioengineering from 2014 to 2016. He also served as the Director of the National Science Foundation Graduate Research and Training Program in video bioinformatics with UCR. He is currently the Bourns Presidential Chair in engineering, the Distinguished Professor of electrical and computer engineering, and the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory, UCR. He has published extensively and has 18 patents. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics,

	human-computer interactions, and biological, medical, military, and intelligence applications. In 1991, he was a Senior Honeywell Fellow with Honeywell Inc. He is a Fellow of AAAS, IAPR, SPIE, and AIMBE.
	Shuangwei Liu received B.S. degree in automation from Northeastern University, Shenyang, China, in 2017. He is a graduate student in pattern recognition and intelligent systems in Northeastern University. He majors in computer vision and image processing, especially deep learning methods and person re-ID.

*Photo of the author(s)

Click here to download Photo of the author(s): Photos of authors.doc

	Yixiu Liu
	Yunzhou Zhang
	Sonya Coleman
	Bir Bhanu

	Shuangwei Liu

***Source Files - Latex or Word**

Click here to download Source Files - Latex or Word: Source files.zip

***Conflict of Interest Form**

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work. There is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.