IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 12, 2017, accepted December 27, 2017, date of publication January 8, 2018, date of current version March 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2790969

Salient Region-Based Least-Squares Log-Density Gradient Clustering for Image-To-Video Person Re-Identification

TIEZHU LI[1,2], LIJUAN SUN[1,3], CHONG HAN[1,3], AND JIAN GUO[1,3]

[1]College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2]School of Software, Henan University, Kaifeng 475000, China
[3]Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Corresponding author: Lijuan Sun (sunlj@njupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572261, Grant 61373139, and Grant 61702284, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150868, in part by the China Postdoctoral Science Foundation under Grant 2014M551635, in part by the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant 1302085B, in part by NUPTSF under Grant NY214013, and in part by the Major Scientific Research Project of Higher Learning Institution of Henan Province under Grant 18A520022.

ABSTRACT Because of its distinction and reliability, the person-salient region has been applied to pedestrian re-identification (re-id) across disjoint camera views. Despite the great progress achieved, a few works have studied the more challenging image-to-video person re-id problem, in which the gallery consists of videos and the same pedestrian appears in a continuous video sequence. The intrinsic high redundancy of such video sequences makes it more difficult to obtain high performance for pedestrian re-id. To solve this problem, in this paper, we propose a new salient region clustering approach for image-to-video person re-id. Specifically, we use the mean shift to extract the person regions and obtain the salient region by computing the saliency for each region. Then, all salient regions are clustered by least-squares log-density gradient clustering, after which the salient regions that are extracted from the same person are marked as the same group. Finally, the rank of the person re-id can be obtained by computing the distance between the probe-salient region and the gallery clustered salient regions. We evaluate the proposed approach on two public pedestrian sequence data sets (PRID 2011 and MARS), and the experimental results validate the effectiveness of the proposed approach for the image-to-video person re-id.

INDEX TERMS Salient region, image-to-video, clustering, person re-identification, distance learning.

I. INTRODUCTION

With the increasing attention of public security, tens of thousands of surveillance cameras have been deployed to cover dense public venues, such as campuses, transport junctions and shopping malls, which are also becoming the basic public facilities of ordinary residential districts. By leveraging this large number of cameras, we can implement person re-identification (re-id), which has important applications in video surveillance, such as pedestrian retrieval and pedestrian tracking. Since computers can match features among the pedestrians, human effort can be saved from exhaustively retrieving a person from a large number of videos. However, person re-id is a notably challenging task. The mass redundant person frames can significantly reduce the re-id efficiency. In addition, variations in illumination, poses, viewpoints and cluttered background will make the persons appearance dramatically change [1]–[3]. In recent years, some video-based person re-id methods have been presented [3]–[10]. Since videos inherently contain more information, video-based methods can obtain better re-identification results than image-based methods. In [6], the features of a person were first extracted from each video frame by using a convolutional neural network and further forwarded to a Long Short-Term Memory (LSTM) network to encode the temporal information of the video sequence. Finally, the feature vectors of the probe image and the video sequence were further forwarded to the similarity sub-network for distance metric learning. You *et al.* [7] proposed a top-push distance learning model (TDL) to make the matching model more effective by selecting more discriminative features to distinguish different people.

Since the salient features of pedestrians are mainly contained in the salient region, salient regions usually differ from person to person, which makes the salient regions in

pedestrian frames a valuable piece of information for the person re-id [8]. In addition, by extracting the salient region, many non-salient regions will be abandoned, which can reduce the computation in the person re-id process and significantly improve the person re-id speed. Based on these advantages, salient features for human have attracted increasing attention in the task of person re-id [8], [11]–[16]. Although the video-based person re-id method can obtain more information from videos, many challenging problems have not been thoroughly resolved. For example, there is less feature difference between images of the same person and different people, and mass redundant person images will reduce the efficiency of the person re-id. To solve the above problems, we propose a method to cluster the salient regions of the same person and realize the person re-id using the salient region matching based on the clustered salient regions. The proposed framework consists of a mean shift (MS)-based salient region extraction method, a least-squares log-density gradient person salient region clustering (RLSLDG) method, and a color-histogram-based salient-region matching method. In the salient region extraction step, we use the MS to segment the person frame regions and compute every region saliency based on the Global Contrast Based Salient Region Detection [17]. For the salient regions of the same person in video sequences, we use the RLSLDG method to cluster the salient regions abstracted from the video sequence. Then, we can match the probe salient region to the non-redundant gallery salient regions extracted from the output of the RLSLDG. The architecture of the proposed framework is shown in Fig. 1. In the proposed framework, the probe image usually has salient features, and the gallery consists of videos, which maintain mixed frames of different pedestrians. The idea of the proposed approach is to cluster the frames that belong to the same person into one group after extracting all salient regions. By computing the distance between the probe and the gallery salient region sets, we can obtain the rank of gallery region sets that are similar to the probe.

The main contributions of this paper include the following.

- We propose a salient region extraction approach based on the mean shift and global contrast-based salient region detection. Then, with the designed-feature difference index d_{index}, we test the performance of our method on two public person re-id datasets (PRID-2011 and MARS), which validates the superiority of the proposed salient region in increasing the feature difference between the salient regions of the same person and those of different people.
- We develop a new image-to-video person re-id method, which first clusters the salient regions of the same person into the same group. Then, the distance between the probe and the gallery salient regions is computed by point to set the distance metric learning. The experimental results on the PRID-2011 and MARS datasets demonstrate that the proposed method can achieve rank-1 matching rate of 53.2% and 62.1%, respectively.

The rest of the paper is organized as follows. Section II briefly describes the related work. The details of the proposed method are elaborated in Section III. Section IV provides the experimental results on different data sets. Section V summarizes the proposed method and presents future work.

II. RELATED WORKS

With the widespread use of surveillance cameras, person re-id is becoming increasingly important to public security. In this section, we classify prior works related to our approach into three main categories: i) salient-feature representation; ii) features clustering; and iii) distance metric learning.

A. SALIENT-FEATURE REPRESENTATION

The salient features in person images can provide significant information for person re-identification [18]. It can improve the similarity of the intra-class person and enlarge the difference of the inter-class person while decreasing the computation. Therefore, salient features have recently attracted increasing interest [11]–[15], [17]–[20]. Zhu *et al.* [12] implemented a patch-matching-based framework for group re-identification by learning a discriminative salience channel, which can filter out highly unreliable and non-informative patch matches between two group images while retaining true matches with appearance variations. Mingyang *et al.* [13] exploited the multi-feature fusion method include RGB, SIFT and Rotation-invariant LBP (RI-LBP) to improve the salient-feature representation. Xu *et al.* [14] obtained the saliency map by computing and integrating all associated conspicuity maps of three spatial scale features. Cheng *et al.* [17] proposed a regional-contrast-based salient-object detection algorithm by simultaneously evaluating the global contrast differences and spatial-weighted coherence scores. Wang *et al.* [18] improved the matching accuracy by assigning different weights to each patch according to its salience score. Zhou *et al.* [19] proposed an overlapping region-based global context descriptor (OR-GCD) to filter false matches by verifying the initial scale-invariant feature transform (SIFT) matches between images based on the bag-of-visual-words (BOW) quantization. In [20], the gallery data and probe data are projected onto a regularized canonical correlation analysis (RCCA) subspace, and the reference descriptors (RDs) of the gallery and probe data are generated by computing the similarity between them and the reference data. Using a saliency-based matching scheme, the results of comparing the RDs of the probe and the gallery can be improved.

B. FEATURES CLUSTERING

Many images of the same person are extracted from the surveillance video, which often mix with the images of other people. Therefore, how to quickly and accurately classify these mixed images has attracted the attention of many researchers [21]–[25]. Zhang *et al.* [21] proposed an image class method by jointly modeling the object, context and background information (OCB). Ishii *et al.* [22] used the k

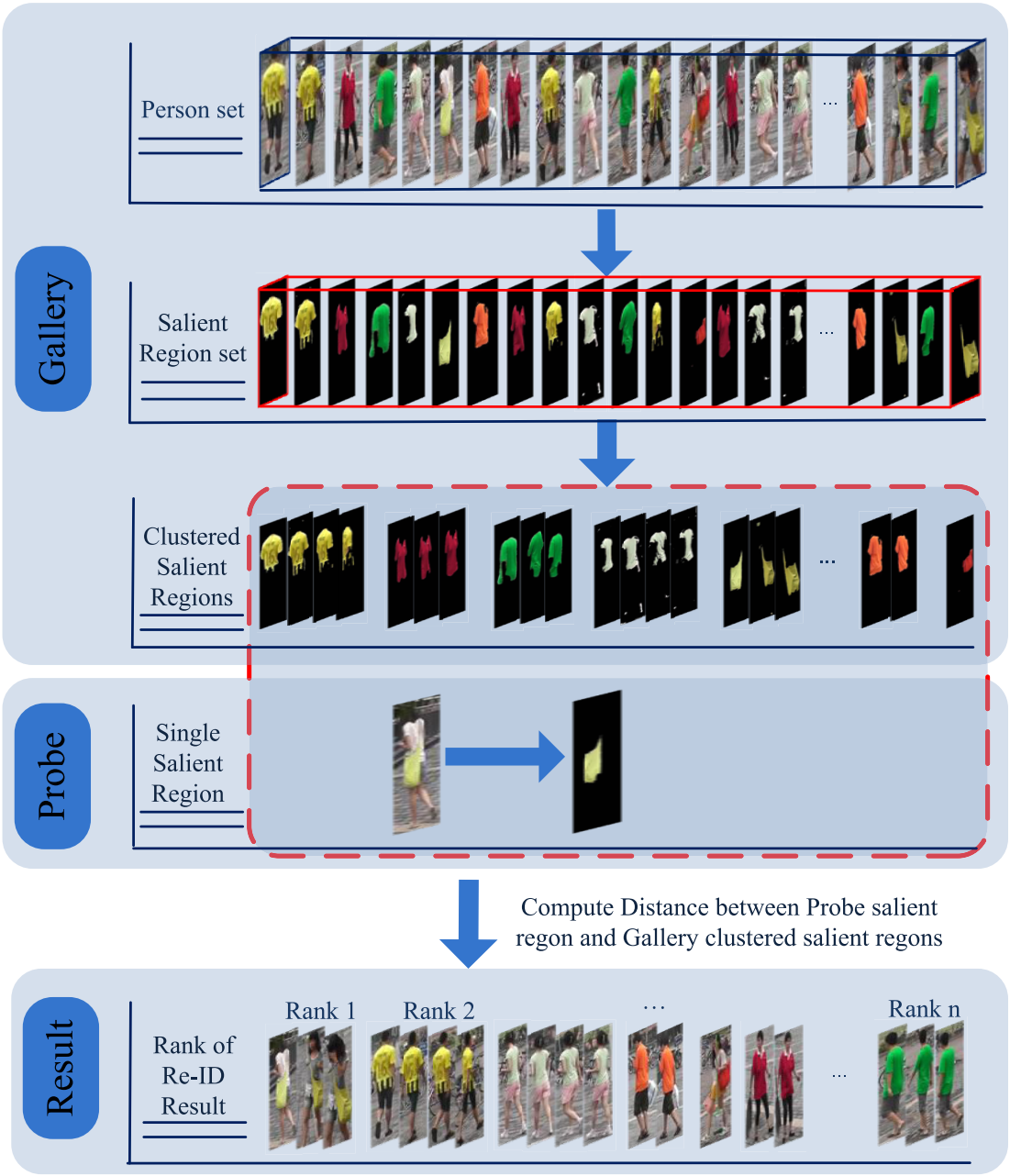

FIGURE 1. Illustration of the proposed approach for image-to-video person re-identification.

nearest neighbors (KNN) to extract k target images that correspond to the top-k similarity scores, which are used to determine the identity of the person that correspond to the probe image. The classical mean shift clustering, which finds the modes of the data probability density, must first estimate the density. However, since good density estimation does not necessarily imply an accurate estimation of the density gradient, such an indirect two-step approach is not reliable. To avoid estimating the data probability density that performs poorly in high-dimensional problems, Sasaki *et al.* [23] proposed the least-squares log-density gradient (LSLDG) method to directly estimate the gradient of the log-density without estimating the density. To further improve the clustering performance of the LSLDG method, Ashizawa *et al.* [24] proposed a Riemannian Direct Log-Density-Gradient Estimation (R-LSLDG) method by introducing Riemannian manifolds into the LSLDG.

C. DISTANCE METRIC LEARNING

The goal of distance metric learning is to learn a proper feature space where the feature vectors of the same person are close, whereas the feature vectors of different people are

far apart. Zhang *et al.* [3] measured the distance between pedestrians by globally aligning the spatio-temporal appearance of a pedestrian. Zhang *et al.* [4] presented an approach to automatically select the most discriminative video fragments from noisy/incomplete image sequences of people, from which reliable space-time and appearance features could be computed while simultaneously learning a video ranking function for the person re-id. Zhu *et al.* [5] proposed a semi-supervised cross-view projection-based dictionary learning approach for video-based person re-id. Zhang *et al.* [6] obtained the feature vector of a video sequence using the LSTM network to first concatenate single image features. Then, a generalized similarity measure, which fuses the affine Mahalanobis distance and Consine similarity, was embedded into the convolutional neural network (CNN) for distance metric learning. You *et al.* [7] integrated a top-push constrain to match video features of people. Li *et al.* [26] segmented images into semantically independent patches to measure the distance between them by estimating and refining an affine transform matrix. In [27], a point-to-set distance learning (PSD) method was proposed by extending the point-to-point distance (PPD), which is based on Mahalanobis distance metric learning.

III. PROPOSED ALGORITHM

In this section, we introduce the salient region extraction method and analyze the advantage of the salient region by computing the feature difference indices d^s_{index} and d^p_{index}. Then, we describe the clustering method for the mixed salient regions. Finally, we provide a distance model between single salient regions and a salient region set.

A. SALIENT REGION EXTRACTION BASED ON THE MEAN SHIFT ALGORITHM

We extract salient regions by computing and ranking the saliency value of each region for a person frame. Cheng *et al.* [17] introduced the biological vision system to select the pixel whose color was in sharp contrast with all other pixels. We extend the saliency model proposed in [17] and use it to compute the saliency value of each region from the segmentation of the person frame. To segment a person frame, we use the classical mean shift (MS) [28] clustering algorithm based on the HSV (Hue, Saturation, Value) color space.

1) REGION ABSTRACTION BASED ON THE MEAN SHIFT

The Mean Shift clustering algorithm finds the modes of the data probability density by obtaining the zero points of the density gradient. Without requiring the fixing of the number of clusters in advance, the mean shift has been a popular image segmentation algorithm, and the typical implementation of the mean shift is to estimate the density using the kernel density estimation and compute its gradient. Given a set of data $\{x_i\}, i = 1, 2, \ldots, n, x_i \in R^d$, with unknown probability density $f(x)$. In the mean shift algorithm, the probability density $f(x)$ is obtained using the kernel density estimation:

$$\hat{f}(x) = \frac{c_{k,\sigma}}{n} \sum_{i=1}^{n} k(\| \frac{x - x_i}{\sigma} \|^2) \tag{1}$$

where the kernel function $k(\cdot)$ models the correlation between dataset $\{x_i\}$ and density center x; in this paper, we use a typical Gaussian function as $k(\cdot)$. σ is the bandwidth and $c_{k,\sigma}$ is the normalization parameter to ensure that the integration of $\hat{f}(x)$ is equal to 1. In practical use, the Mean Shift analysis over a color image is mainly divided into two domains: the spatial domain and the spectral domain. Therefore, the kernel density estimation $\hat{f}(x)$ in Eq. (1) can be represented as

$$\hat{f}(x) = \frac{c_{k,h_s,h_r}}{n} \sum_{i=1}^{n} k(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2) \tag{2}$$

Since we are concerned about the points where the density gradient is equal to zero, we compute the partial derivative for Eq. (2).

$$\begin{aligned} \frac{\partial \hat{f}(x)}{\partial x_s} &= \frac{c'_{k,h_s,h_r}}{2n} \sum_{i=1}^{n} (x_s - x^i_s) \\ &\quad \cdot k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2) \\ &= \frac{c'_{k,h_s,h_r}}{n} [\sum_{i=1}^{n} k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)] \\ &\quad \cdot [x_s - \frac{\sum_{i=1}^{n} x^i_r \cdot k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)}{\sum_{i=1}^{n} \cdot k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)}] \end{aligned} \tag{3}$$

The gradient of the kernel density estimator $\hat{f}(x)$ can be computed:

$$\nabla \hat{f}(x) = \frac{\partial \hat{f}(x)}{\partial x_s} = \epsilon(x) \cdot m(x)$$

where $\epsilon(x) := \frac{c'_{k,h_s,h_r}}{n} [\sum_{i=1}^{n} k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)] > 0$, and $m(x) := x_s - \frac{\sum_{i=1}^{n} x^i_r \cdot k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)}{\sum_{i=1}^{n} \cdot k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)}$ is the mean shift vector. The mean shift algorithm uses iteration for mode seeking. If we set $\nabla \hat{f}(x) = 0$, the local maximum value can be obtained. More specifically, a necessary condition for $\nabla \hat{f}(x) = 0$ implies that $m(x) = 0$, which indicates that $x = x + m(x)$, and the iteration is complete. The implementation of the mean shift can be summed in two steps: First, compute

$$g_{h_s,h_r}(x) = \frac{\sum_{i=1}^{n} x^i_r \cdot k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)}{\sum_{i=1}^{n} k'(\| \frac{x_s - x^i_s}{h_s} \|^2 + \| \frac{x_r - x^i_r}{h_r} \|^2)},$$

where $g_{h_s,h_r}(x)$ is the centroid of x_i. Second, if $\| g_{h_s,h_r}(x) - x_s \| < \varepsilon$, the iterations are complete. Otherwise, set $x_s = g_{h_s,h_r}(x)$ and enter the next iteration.

2) REGION SALIENCE COMPUTATION

The features in the most salient region have a major effect on the person re-identification, and accordingly, the salient region selection method plays an important role in the re-id process. Our method for salient region selection is based on the Global-Contrast-Based Salient-Region Detection [17], which introduces the biological vision system to select the pixel whose color is in sharp contrast with all other pixels. Thus, the saliency value of the pixel can be represented by its color distance to all other pixels. In addition, to accelerate the calculation process of the saliency value for each region, we reduce the number of colors using a quantized image. We measure the color differences in the HSV color space because it is notably intuitive to express the color in HSV, which is closer to people's feelings. Let $P = \{r_i \mid i = 1, 2, \ldots, m\}$ denote the regions of a person, which are generated by the mean shift clustering algorithm; the saliency value of r_i is represented as follows:

$$S(r_i) = \sum_{r_i \neq r_j} w(r_i) \cdot D(r_i, r_j) \tag{4}$$

where $D(r_i, r_j)$ denotes the distance between regions r_i and r_j, and $w(r_i)$ is the weight of r_i. To make the larger region have a higher chance of being selected as the salient region, $w(r_i)$ is computed in (5).

$$w(r_i) = Sum(r_i) / \sum_{j=1}^{m} Sum(r_j) \tag{5}$$

where the value of $Sum(r_i)$ is represented by the number of pixels in r_i. By counting the frequency of each type of color quantization, we can measure their importance in the region. Let $Sum(r_i, c_l)$ represent the number of pixels whose color is equal to c_l in region r_i; $P(r_i, c_l)$ in Eq. (6) represents the importance of the l-th color in r_i.

$$P(r_i, c_l) = Sum(r_i, c_l) / Sum(r_i) \tag{6}$$

For a quantized image in the HSV color space, we can measure the distance between the l-th color and the k-th color as shown in (7).

$$d(c_l, c_k) = abs(h(c_l) - h(c_k)) + abs(s(c_l) - s(c_k)) + abs(v(c_l) - v(c_k)) \tag{7}$$

After obtaining the distance between the l-th color and the k-th color, we can obtain the distance between regions r_i and r_j,

$$D(r_i, r_j) = \sum_{l=1}^{n(r_i)} \sum_{k=1}^{n(r_j)} P(r_i, c_l) \cdot P(r_j, c_l) \cdot d(c_l, c_k) \tag{8}$$

where $n(r_i)$ is the number of color quantization of region r_i. The procedures of the proposed method to compute the salience values are summarized in Algorithm 1.

Algorithm 1 Salience-Value-Computing Algorithm

Input: $P = \{r_i \mid i = 1, 2, \ldots, m\}$.
Output: S, the saliency value of region set P.
$S = \{S(r_i) \mid i = 1, 2, \ldots, m\}$.

```
for i = 1 : step : m do
    Compute the weight of r_i based on Eq. (5).
    for j = 1 : step : m do
        if i == j then
            continue;
        endif
        for l = 1 : step : n(r_i) do
            Compute the proportion of the l-th color
            in r_i based on Eq. (6);
            for k = 1 : step : n(r_j) do
                Compute the proportion of the k-th
                color in r_j based on Eq. (6);
                Compute the distance between the
                l-th color and the k-th color based on
                Eq. (7);
            endfor
        endfor
        Compute the distance between regions r_i and
        r_j based on Eq. (8);
    endfor
    Compute the salience value S(r_i) of region r_i
    based on Eq. (4);
endfor
return S.
```

3) ANALYSIS OF SALIENT REGION

Compared with the global person frame, the salient region only contains the most significant feature information of the person. Thus, the salient region can reduce the feature differences between the salient regions of the same person and improve the feature differences between the salient regions of different people. The feature differences between intra-person set and inter-person set are illustrated in Fig. 2.

To analyze the advantage of the salient region of people, we provide the feature difference index d^s_{index} between salient regions and feature difference index d^p_{index} between global person frames. d^s_{index} and d^p_{index} are calculated as follows:

$$d^s_{index} = \frac{\bar{d}^s_{differentperson} - \bar{d}^s_{sameperson}}{\bar{d}^s_{differentperson}} \tag{9}$$

$$d^p_{index} = \frac{\bar{d}^p_{differentperson} - \bar{d}^p_{sameperson}}{\bar{d}^p_{differentperson}} \tag{10}$$

where $\bar{d}^p_{differentperson}$ is the feature differences in an inter-person set, $\bar{d}^p_{sameperson}$ is the feature differences in an

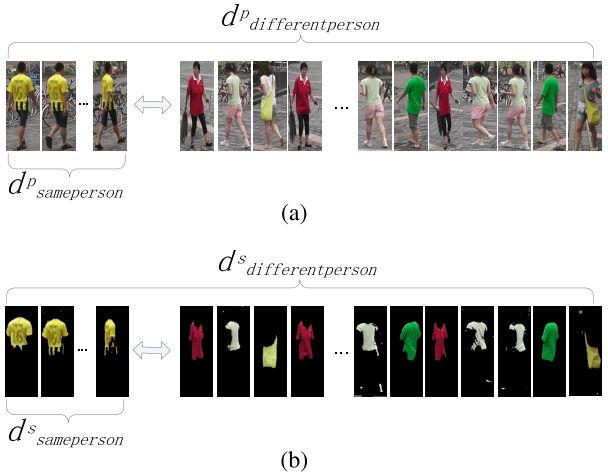

FIGURE 2. The illustration of the difference between intra-person set and inter-person set. (a) The difference based on global person. (b) The difference based on salient region.

intra-person set, and they are calculated as follows:

$$\bar{d}^{s}_{sameperson}(S_{same}) = \frac{2}{n*(n-1)} \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D(r_i, r_j) \quad (11)$$

$$\bar{d}^{p}_{differentperson}(S_{mix}) = \frac{2}{n*(n-1)} \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D(r_i, r_j) \quad (12)$$

where S_{same} is the set of salient regions that belong to the same person, and S_{mix} is the set of salient regions that belong to different people. $\bar{d}^{s}_{sameperson}$ and $\bar{d}^{p}_{differentperson}$ can be similarly obtained from the corresponding global person frames.

In experiments, we randomly select 50 pedestrians from the PRID-2011 and MARS datasets and 10 frames for every pedestrian. Then, we compute d^{p}_{index} and d^{s}_{index} on the selected person frames and their corresponding salient regions. The Features Difference Index in PRID-2011 and MARS is shown in Fig. 3. Fig. 3 shows that the feature difference index d^{s}_{index} is higher than d^{p}_{index} in the PRID-2011 and MARS datasets, which verifies the advantage of the salient region of people.

B. CLUSTERING FOR SALIENT REGIONS

In this section, we implement the salient region clustering following the approach proposed in [24] and analyze the performance of the clustering method by conducting a series of experiments on the PRID-2011 dataset. Specifically, to find the effect of the number of pedestrians on the clustering accuracy, we change the number of pedestrians from 3 to 10 in the experiment process.

1) CLUSTERING BASED ON THE LEAST-SQUARES LOG-DENSITY GRADIENT

$f(x)$ is the unknown probability density function of dataset $\{x_i\}$, $i = 1, 2, \ldots, n, x_i \in R^d$. As the gradient of the log-density, $g(x)$ is represented as:

$$\begin{aligned} g(x) &:= (g^{(1)}(x), \ldots, g^{(d)}(x))^T \\ &= \nabla log f(x) = \frac{\nabla f(x)}{f(x)} \end{aligned} \quad (13)$$

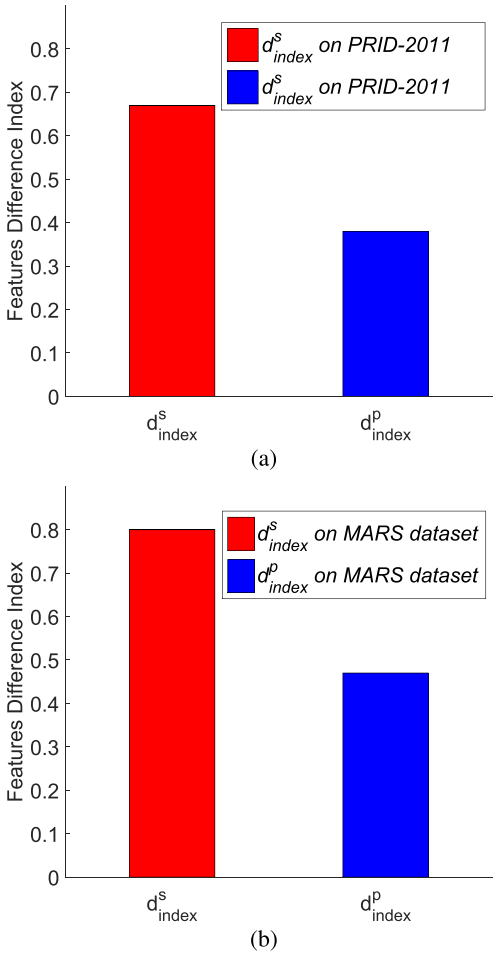

FIGURE 3. The features difference index on the two person re-identification datasets. (a) features difference index in PRID-2011, (b) features difference index in MARS.

where $g^{(j)}(x) := \partial_j log f(x)$ and $\partial_j = \frac{\partial}{\partial_{x^{(j)}}}$. The key idea of the Least-Squares Log-Density Gradient is to fit a model $\widetilde{g}^{(j)}(x)$ to the true log-density gradient $g^{(j)}(x)$ under the squared loss:

$$\begin{aligned} J^{(j)}(\widetilde{g}^{(j)}(x)) &:= \int (\widetilde{g}^{(j)}(x) - g^{(j)}(x))^2 f(x) dx - C \\ &= \int \widetilde{g}^{(j)}(x)^2 f(x) dx - 2 \int \widetilde{g}^{(j)}(x) \partial_j f(x) dx \\ &= \int \widetilde{g}^{(j)}(x)^2 f(x) dx + 2 \int \partial_j \widetilde{g}^{(j)}(x) f(x) dx \end{aligned} \quad (14)$$

where $C = g^{(j)}(x)^2 f(x) dx$; the last equality follows from integration by parts under $lim_{x^{(j)} \to \pm\infty} \widetilde{g}^{(j)}(x) f(x) = 0$. Then, the empirical approximation of $J^{(j)}$ is:

$$\hat{J}^{(j)}(\widetilde{g}^{(j)}(x)) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{g}^{(j)}(x)^2 + \frac{2}{n} \sum_{i=1}^{n} \partial_j \widetilde{g}^{(j)}(x_i) \quad (15)$$

$\widetilde{g}^{(j)}(x)$ in Eq. (15) can be represented by a linear-in-parameter model.

$$\hat{J}^{(j)}(\widetilde{g}^{(j)}(x)) = \theta^{(j)^T} \psi^{(j)}(x) = \sum_{k=1}^{m} \theta_k^{(j)} \psi_k^{(j)}(x) \quad (16)$$

where m is the number of parameters; $\theta^{(j)} \in \mathbb{R}^m$ is the parameter vector, and $\psi^{(j)}(x) \in \mathbb{R}^m$ is the basis functions. By adding an $\ell_2 - regularizer$ to Eq. (15), the problem of Eq. (15) can be represented as the following optimization problem:

$$\hat{\theta}^{(j)} = argmin_{\theta^{(j)} \in R^m}[\theta^{(j)^T} \hat{G}^{(j)} \theta^{(j)} + 2\theta^{(j)^T} \hat{h}^{(j)} + \lambda \theta^{(j)^T} \theta^{(j)}] \tag{17}$$

where $\lambda \geq 0$ is the regularization parameter, and $\hat{G}^{(j)}$ and $\hat{h}^{(j)}$ are defined as follows:

$$\hat{G}^{(j)} := \frac{1}{n} \sum_{i=1}^{n} \psi^{(j)}(x_i) \psi^{(j)}(x_i)^T,$$
$$\hat{h}^{(j)} := \frac{1}{n} \sum_{i=1}^{n} \partial_j \psi^{(j)}(x_i) \tag{18}$$

The optimal solution $\theta^{(j)}$ can be analytically obtained as

$$\theta^{(j)} = -(\hat{G}^{(j)} + \lambda^{(j)} I_m)^{-1} \hat{h}^{(j)} \tag{19}$$

where I_m is an $m \times m$ identity matrix.

2) ANALYSIS OF THE CLUSTERING ACCURACY

To analyze the performance of the proposed clustering method in this paper using salient regions, we randomly select 5-20 frames from every persons folder and mix them together to simulate the original pedestrian frame set. We also use different numbers of pedestrians (3-10) to analyze its effect on the clustering accuracy. The clustering accuracy based on salient regions and global person with different number of people in the PRID-2011 dataset is shown in Fig. 4. Fig. 4 shows that the highest and average clustering accuracies based on salient regions are 91% and 70%, whereas the corresponding values of the global person are 75% and 47%. The clustering accuracy based on salient regions is 69% when the number of people is 10. The highest and average clustering accuracies based on the salient region are significantly higher than those based on the global person frame mainly because the salient region can eliminate a lot of useless non-salient information, which can adversely affect the salient region set distinguishability.

C. DISTANCE BETWEEN THE PERSON SALIENT REGION AND SALIENT REGION SETS

A set of person salient regions usually can be represented by a hull. To rule out the meaningless points that are too far from the sample mean, the hull of a set of samples $S = [s_1, \ldots, s_i, \ldots, s_n]$ in [29] is defined as follows:

$$H(S) = \{\sum_{i=1}^{n} s_i a_i \mid \sum a_i = 1, \| a \|_{l_p} \leq \sigma\} \tag{20}$$

Based on the defined person salient region set, the task of person re-identification in the proposed method can be modeled as the problem of computing the distance between a sample salient region x and a set of samples S. In [27], the point-to-set distance metric learning (PSDML) method was proposed,

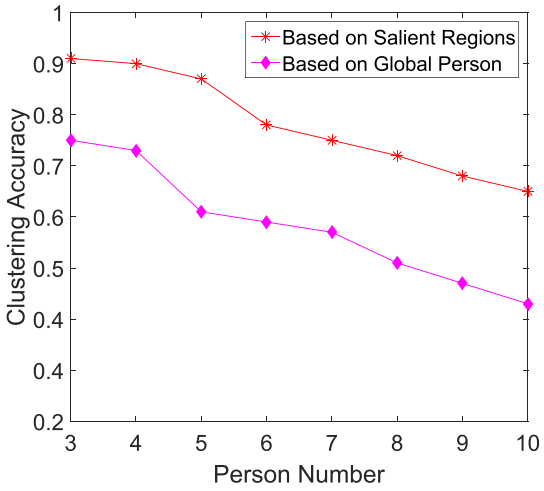

FIGURE 4. Clustering results in PRID-2011 dataset with different number of people.

and we use PSDML to match the sample person salient region and clustered person salient region set $H(S)$. With $d(x, S)$ as the distance between a sample person salient region x and salient region set S, the distance can be represented as:

$$d(x, S) = \| x - S\hat{a} \|_2 \tag{21}$$

where $\hat{a} = argmin_a \| x - H(S) \|_2^2$. To obtain a more accurate re-identification result, a projection matrix P is usually introduced. Then, Eq. (21) is rewritten as:

$$\begin{aligned} d_M(x, S) &= \| P(x - S\hat{a}) \|_2^2 \\ &= (x - S\hat{a})^T P^T P (x - S\hat{a}) \end{aligned} \tag{22}$$

where $\hat{a} = argmin_a \| x - S(a) \|_2^2$, and we can obtain $\hat{a}$ by least-square regression as $(S^T S + \lambda I)^{-1} S^T x$. With $M = P^T P$ and m person salient region sets $\{S_1, S_2, \ldots, S_m\}$, for the query sample x and S_i, $\hat{a}_i$ is:

$$\hat{a}_i = (S_i^T M S_i + \lambda I)^{-1} S_i^T M x \tag{23}$$

Then, the distance between x and S_i is:

$$d_M(x, S_i) = (x - s_i \hat{a}_i)^T M (x - S_i \hat{a}_i) \tag{24}$$

By computing all distances between the query salient region and all salient region sets, we can obtain the re-identification rank after sorting the computed distances. According to Eq. (24), matrix M plays a notably important role in the computation of $d_M(x, S_i)$, and we obtain an appropriate M by PSDML [27] through learning from the training salient region sets $\{S_1, S_2, \ldots, S_n\}$.

IV. PERFORMANCE EVALUATION

To evaluate the effectiveness of our approach, we conduct our experiments on two public datasets and compare the results to other state-of-the-art approaches.

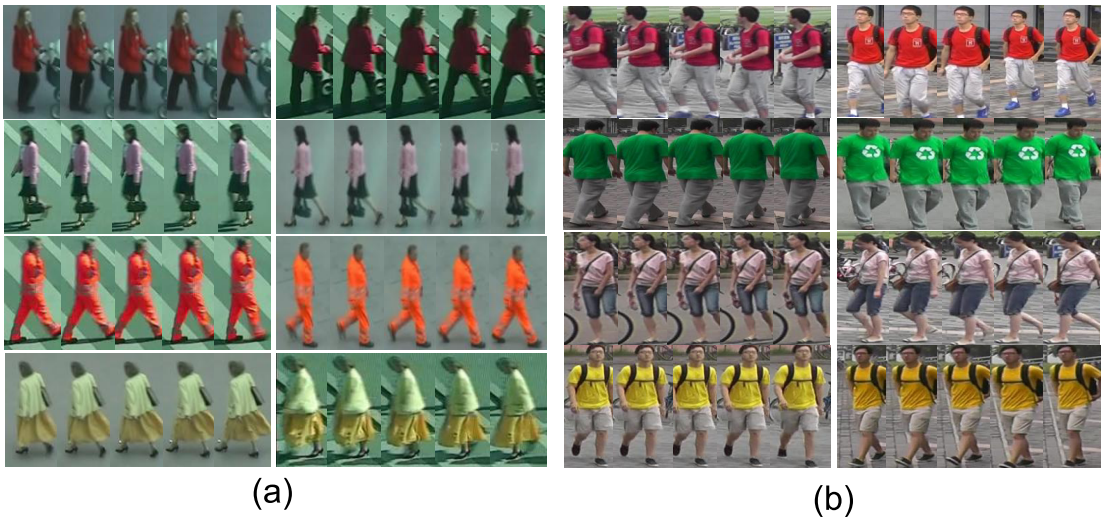

FIGURE 5. Example pairs of the image sequence of the same pedestrian in different camera views from (a) PRID-2011 and (b) MARS, Only five frames are choose from each sequence and shown in a row.

TABLE 1. The summary of datasets of PRID-2011 and MARS.

Dataset	pedestrians	cameras	image size	Label
PRID-2011	200	2	128×64	Hand
MARS	1261	6	256×128	DPM+GMMCP

A. DATASETS

In this paper, we evaluate the performance of the proposed method on the two public person re-id datasets: PRID-2011 [30] and MARS [31]. The details of the two datasets are shown in Table 1.

1) PRID-2011 DATASET

The PRID-2011 dataset includes 385 and 749 person sequences in two disjoint cameras (Cam-A and Cam-B). Each sequence has 5-675 frames with an average of 84. Among them, the first 200 people appear in both views, and we only use them for evaluation in our experiments.

2) MARS DATASET

The MARS (Motion Analysis and Re-identification Set) dataset is an extension of the Market1501 dataset [32], which capture 1,261 different pedestrians using six near-synchronized cameras on campus. Five 1080p HD cameras and one SD camera were used to record the video information.

B. EXPERIMENTAL SETTING AND EVALUATION PROTOCOL

In the experiment, we compare our approach with several state-of-the-art video-based person re-id methods, including discriminative video fragment selection and ranking (DVR) [1] and its three enhancement versions (Salience + DVR, MS-Colour&LBP + DVR and SDALF [33] + DVR), Local Maximal Occurrence (LOMO) [3], Cross-view Quadratic Discriminant Analysis (XQDA) [33], HistLBP + XQDA [33], hist3D + KISSMR [34], recurrent convolutional (RCN) [36], and Bow + KISSME [35]. Following the setting of [6], the data extracted from the PRID-2011 and MARS datasets are randomly divided into two groups of equal size. We use the people in the first group for training and another group for the test.

1) SALIENT REGION EXTRACTION

In the experiments, we use the mean shift [28] algorithm to segment the person images into regions and control the number of regions at 3-6. The HSV color space is notably intuitive to express the color in Hue, Saturation and Value, and it is closer to people's feelings. Moreover, it is insensitive to illumination changes. Based on these advantages, we compute the saliency of every region in the HSV color space using Algorithm 1 and select the most salient region.

2) SALIENT REGION CLUSTERING

The pedestrian images in the PRID-2011 and MARS datasets were artificially marked and classified, where the images of each person were selected and saved in one folder. However, the original pedestrian images that were extracted from the surveillance video may contain several people. Thus, we randomly select 5-20 images from every persons folder and mix them together to simulate the original pedestrian image set. Since pedestrians usually appear in successive frames of the video, we limit the number of people to no more than 10 in our experiments with the clustering results in Fig. 4. For the PRID-2011 dataset, the salient regions of pedestrians are normalized as 50×50 color images, and we extract four features (Hue, Saturation, Value, texture) for every pixel, especially, the texture feature are extracted based on Gabor filter. Furthermore, we reduce the dimension of the salient region of $50 \times 50 \times 4$ to 576 using the PCA. The salient regions that are extracted from the MARS dataset are

TABLE 2. TOP Rank-r MATCHING RATES (%) IN PRID-2011 dataset.

Methods	PRID-2011			
Rank	1	5	10	20
Salience+DVR	41.7%	64.5%	77.5%	77.1%
MS-colour & LBP+DVR	37.6%	63.9%	75.3%	77.5%
Bow+XQDA	31.8%	58.5%	81.3%	85.3%
SDALF+DVR	31.6%	58%	81.3%	85.3%
Hog3D+DVR	21.7%	51.7%	83.5%	87%
Ours	52.3%	72.1%	82.6%	85.2%

normalized as 100×100 color images and processed through the PCA. Then, these features are represented by the Low-Rank Representation (LRR) model on Grassmann Manifold, and the result of Z are used for the clustering algorithm (LSLDG) with λ set to 0.1.

3) DISTANCE LEARNING

We randomly select 80 people in the test group and 10 salient regions for every person to construct the gallery set; meanwhile, the probe salient region is also selected. We compute the distance between the probe salient region and the gallery clustered salient region sets based on PSDML [27]. For the evaluation on PRID-2011 and MARS, the value of λ in Eq. (23) is set to 0.03 and 0.01, respectively.

In the experiments, we use the standard cumulated matching characteristic (CMC) curve as our evaluation metric and report the Rank-k average matching rates of 10 trials. All algorithms are coded using Matlab R2015a and implemented in a quad-core Intel®Core^{TM} I5-5200U 2.20GHzCPU machine with 8G RAM.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) EXPERIMENTAL RESULTS ON THE PRID-2011 DATASET

With the above experimental setting and evaluation protocol, we obtain the image-to-video person re-id result on the PRID-2011 dataset using our method. Some example person sequences of the PRID-2011 dataset with saliency features are shown in Fig. 5(a). Table 2 and Fig. 6 show the top ranked matching rates of our method and the state-of-the-art methods on the PRID-2011 dataset. As shown in Table 2, our method achieve a rank-1 accuracy of 52.3%, which has comparable matching rates with other methods. In addition, the increase in matching rate decreases when the rank is over 10 because the proposed salient region in this paper can increase the feature difference between the salient regions of the same person and those of different people, which can effectively reduce the distance between the right matching salient regions and the probe salient region. Thus, the matching rate rapidly increases when the rank is in the top 10.

2) EXPERIMENTAL RESULTS ON MARS

The frames in the MARS dataset have better image quality than those in PRID-2011, and the salient features of pedestrians are better, which is favorable for improving the person re-id matching rate. Some example person sequences of the PRID-2011 dataset with saliency features are shown in Fig. 5(b). Fig. 7 shows the CMC curves of the compared methods, and Table 3 reports the detailed rank-1, rank-5, rank-20 matching rates of all compared methods. The proposed method achieves a rank-1 accuracy of 62.1%, which is comparable to other state-of-the-art methods, and the matching rate of our proposed method rapidly increases in the 5 top ranks. There are two main reasons that the proposed method can achieve such performance results: 1) the salient region extracted from the person frame can efficiently increase the feature difference in the inter-group of salient region sets

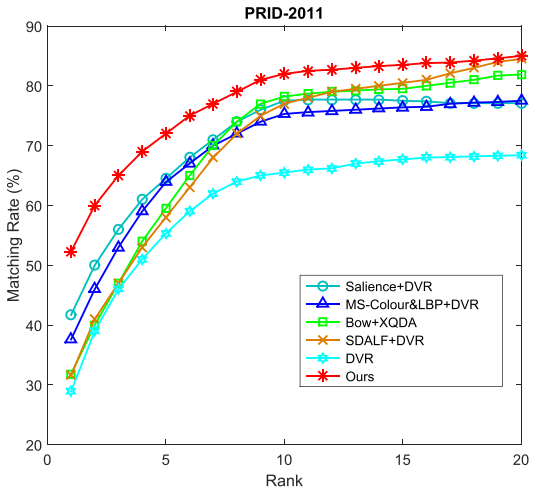

FIGURE 6. CMC curves of average matching rates in the PRID-2011 dataset.

TABLE 3. TOP Rank-r MATCHING RATES (%) IN MARS dataset.

Methods	MARS		
Rank	1	5	20
LOMO+XQDA	30.7%	46.6%	60.0%
Bow+KISSME	30.6%	46.2%	59.2%
HistLBP+XQDA	18.6%	33%	45.9%
RCN	44.2%	58.6%	77.6%
TMSL	56.5%	70.6%	83.5%
Ours	62.1%	75.3%	83.7%

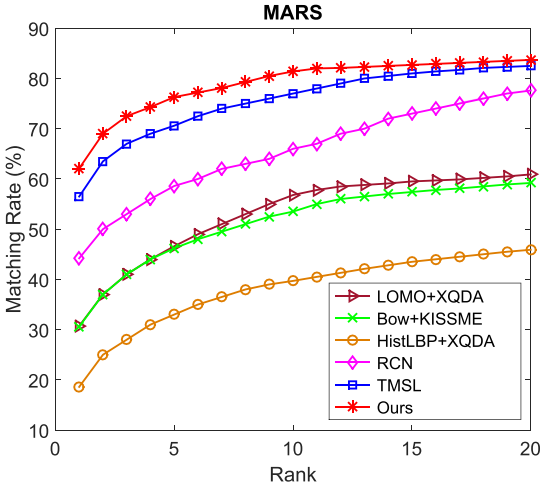

FIGURE 7. CMC curves of average matching rates in the MARS dataset.

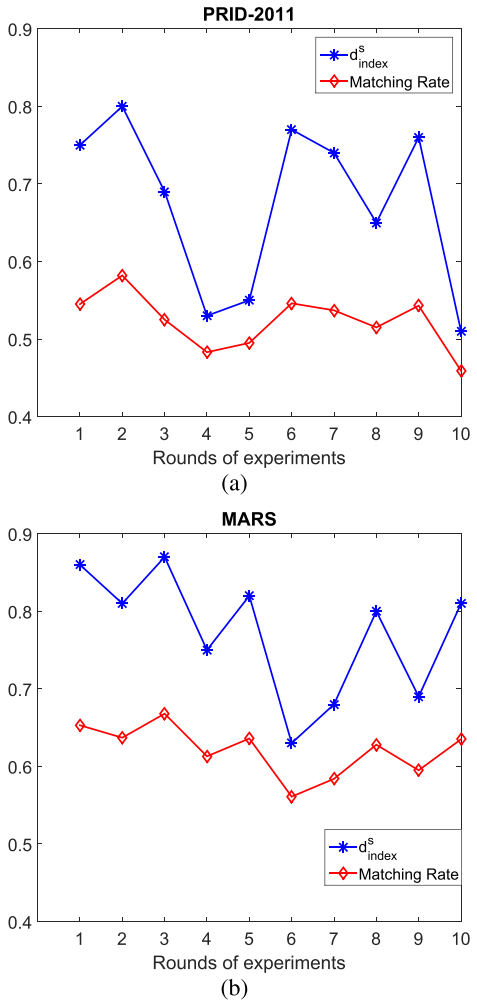

FIGURE 8. The features difference index d^s_{index} VS Matching Rate (%) in PRID-2011 (a) and MARS (b).

while decreasing the feature difference of the intra-group salient region set; 2) through the low-rank representation for the salient region and by clustering them on Grassmann manifold, the clustering accuracy of the gallery salient region sets can be further improved, which can indirectly affect the final matching rate.

3) EFFECT OF FEATURE DIFFERENCE INDEX d^s_{index}

To analyze the relation between feature difference index d^s_{index} of the gallery salient region sets and the matching rate of person re-id, we compute and save the d^s_{index} of gallery salient region sets in each round for 10 trials. The value of d^s_{index} and the matching rate of each round of experiment is shown in Fig. 8. The matching rate of person re-id has similar fluctuations with feature difference index d^s_{index}. When d^s_{index} of the gallery salient region sets increases, the matching rate is more accurate. This phenomenon also appears in the PRID-2011 and MARS dataset. Since the gallery salient region sets in the MARS dataset has a larger d^s_{index} than the PRID-2011 dataset, the MARS dataset has a 9.8% higher rank-1 matching rate than the PRID-2011 dataset (62.1% vs. 52.3%). The relation between the feature difference index and the matching rate shows that improving the feature difference of inter-group people in the gallery is effective to increase the person re-id accuracy.

V. CONCLUSION

Image-to-video person re-id attempts to solve the problem where the probe is a single image and the gallery consists of videos of multiple people from non-overlapping cameras. In this paper, we propose a new framework for the image-to-video person re-id problem. The proposed method extracts the salient region from a person frame, which can efficiently increase the feature difference between inter-groups of salient region sets. Then, we cluster the mixed person salient regions into groups using the Least-Squares Log-Density Gradient clustering method. Finally, we obtain the ranks of the person frame sets by learning the distance model between the probe salient region and the gallery salient region sets. Experimental results on the PRID-2011 and MARS datasets validate the efficiency of the proposed method and reveal the relation between the feature difference index and the matching rate of person re-id. However, since the salient region of the person frame is critical to the final matching rate, the performance will suffer from certain degradation when the person frame has low color contrast because of worse salient region extraction. We will further study a more robust salient region extraction method from person frames in our future work.

REFERENCES

[1] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, Dec. 2016.
[2] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang, "Person re-identification by dual-regularized KISS metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2726–2738, Jun. 2016.
[3] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2042–2054, Apr. 2017.
[4] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[5] X. Zhu *et al.*, "Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[6] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai, "Image-to-video person re-identification with temporally memorized similarity learning," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[7] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proc. IEEE Conf. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 1345–1353.
[8] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *Proc. IEEE Conf. ICCV*, Dec. 2015, pp. 3810–3818.
[9] 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 *Proc. Int. Conf. IJCAI*, Jul. 2016, pp. 3552–3559.
[10] A. Bedagkar-Gala and S. K. Shah, "Multiple person re-identification using part based spatio-temporal color appearance model," in *Proc. IEEE Conf. ICCV*, Barcelona, Spain, Nov. 2011, pp. 1721–1728.
[11] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.
[12] F. Zhu, Q. Chu, and N. Yu, "Consistent matching based on boosted salience channels for group re-identification," in *Proc. IEEE Conf. ICIP*, Phoenix, AZ, USA, Sep. 2016, pp. 4279–4283.

[13] Y. Mingyang, W. Wanggen, H. Li, and Z. Yifan, "Person re-identification using human salience based on multi-feature fusion," in *Proc. Int. Conf. ICSSC*, Shanghai, China, Jul. 2015, pp. 195–199.
[14] Y. Xu, J. Li, J. Chen, G. Shen, and Y. Gao, "A novel approach for visual saliency detection and segmentation based on objectness and top-down attention," in *Proc. Int. Conf. ICIVC*, Chengdu, China, Jun. 2017, pp. 361–365.
[15] H. Chenini, "An embedded FPGA architecture for efficient visual saliency based object recognition implementation," in *Proc. Int. Conf. ICSC*, Batna, Algeria, May 2017, pp. 187–192.
[16] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1265–1274.
[17] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
[18] C. Wang, S. Tang, S. Zhu, and X. Jing, "Person re-identification based on saliency," in *Proc. Chin. Control Conf. CCC*, Chengdu, China, Jul. 2016, pp. 3887–3890.
[19] Z. Zhou, Y. Wang, Q. M. J. Wu, C.-N. Yang, and X. Sun, "Effective and efficient global context verification for image copy detection," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 48–63, Jan. 2017.
[20] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person reidentification with reference descriptor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 776–787, Apr. 2016.
[21] C. Zhang, G. Zhu, C. Liang, Y. Zhang, Q. Huang, and Q. Tian, "Image class prediction by joint object, context and background modeling," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[22] M. Ishii, H. Imaoka, and A. Sato, "Fast k-nearest neighbor search for face identification using bounds of residual score," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May/Jun. 2017, pp. 194–199.
[23] H. Sasaki, A. Hyvärinen, and M. Sugiyama, "Clustering via mode seeking by direct estimation of the gradient of a log-density," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2014, pp. 19–34.
[24] M. Ashizawa, H. Sasaki, T. Sakai, and M. Sugiyama, "Least-squares log-density gradient clustering for Riemannian manifolds," in *Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Fort Lauderdale, FL, USA, 2017 vol. 54.
[25] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Multi-objectives-based text clustering technique using K-mean algorithm," in *Proc. Int. Conf. CSIT*, Amman, Jordan, Jul. 2016, pp. 1–6.
[26] J. Li, X. Li, B. Yang, and X. Sun, "Segmentation-based image copy-move forgery detection scheme," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 507–518, Mar. 2015.
[27] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *Proc. IEEE Conf. ICCV*, Sydney, NSW, Australia, Dec. 2013, pp. 2664–2671.
[28] S. Susan and A. Kumar, "Auto-segmentation using mean-shift and entropy analysis," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, New Delhi, India, Mar. 2016, pp. 292–296.
[29] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. IEEE Conf. CCPR*, Beijing, China, Jun. 2015, pp. 1565–1573.
[30] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Image Anal.*, 2011, pp. 91–102.
[31] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.
[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Conf. ICCV*, Dec. 2015, pp. 1116–1124.
[33] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. CVPR*, San Francisco, CA, USA, Jun. 2010, pp. 2360–2367.
[34] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. CVPR*, Boston, MA, USA, Jun. 2015, pp. 2197–2206.
[35] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. CVPR*, Providence, RI, USA, Jun. 2012, pp. 2288–2295.
[36] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 1325–1334.

TIEZHU LI received the B.S. and M.S. degrees in computer application from Henan University in 2005 and 2008, respectively. He is currently pursuing the Ph.D. degree with the School of Computer, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research focuses on computer vision problems, including issues like pedestrian re-identification, person tracking, and person behavior analysis.

LIJUAN SUN received the B.S. degree in radio engineering from Southeast University in 1985, the M.S. degree in signal, circuit and system, and the Ph.D. degree in communication and information system from the Nanjing University of Posts and Telecommunications in 1988 and 2007, respectively. She is currently a Professor with the School of Computer Science and Technology and a Ph.D. Supervisor with the School of Software, Nanjing University Posts and Telecommunications, Nanjing, China. Her main research interests include multimedia wireless sensor networks and evolutionary computation.

CHONG HAN received the M.S. degree in computer application from Henan University in 2010, and the Ph.D. degree in information network from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2013. He is currently a Lecturer with the Nanjing University of Posts and Telecommunications. His research interests include wireless visual/video sensor networks and multimedia information processing.

JIAN GUO received the B.S. and M.S. degrees in computer science and the Ph.D. degree in information network from the Nanjing University of Posts and Telecommunications, China, in 2000, 2007, and 2013, respectively. He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications. His research interests cover evolutionary computation, swarm intelligence, and multimedia wireless sensor networks.