Efficient Weighted Histogram Features for Single-Shot Person Re-Identification

Yu-Lun Wei[1] · Chang Hong Lin[1]

Received: 26 November 2015 / Revised: 31 May 2017 / Accepted: 13 June 2017 / Published online: 30 June 2017

Abstract In video surveillance, person re-identification is an important task of recognizing individuals in diverse locations over different non-overlapping camera views under the condition of large illumination variations. In order to deal with these challenges, two different and efficient color-based-methods are proposed for single-shot person re-identification in this article, which uses the Gaussian mixture model to combine with color histograms (low-level feature) and the dense salient patches (mid-level feature) as the color features. Both the two proposed systems are three-stage processes. The first stage is the image enhancement by illumination normalization, and it used to deal the intensity variations. The second stage includes pedestrian segmentation and human region partition, which separates the background (BG) and foreground (FG) and locates the body segments to improve the accuracy of feature extracting and matching. The third stage is to perform feature extracting and matching. A Gaussian mixture model is used in the first system, GMMWCH, to generate the weighted color histogram as the color features, which has a low computation time and a good recognition rate. For the second system, SaliGMMWCH, the dense correspondence is used to link the color histogram weighted by the Gaussian mixture model to find salient regions. Even though that takes more time for computation, the SaliGMMWCH retains a better recognition rate than GMMWCH. In addition, the correct match can be chosen by matching the similarity scores of different feature with an appropriate weight selection. Both the proposed methods have been tested on the benchmark, VIPeR and PRID 2011, for evaluation. The experimental results demonstrate superior recognition rate and execution performance by using the proposed methods compared to other representative methods.

Keywords Single-shot person re-identification · Gaussian mixture model · Weighted color histograms · Video-surveillance

1 Introduction

In recent years, surveillance system plays an important role in public, and this field has attracted more and more research interests. The task for a distributed multi-camera surveillance system to associate people across camera views at different locations and time is known as the person (pedestrian) re-identification problem. It is not only the keypoint of applications such as long-term multi-camera tracking and search missing person and robbers from a crowded, but also a novel and challenging research topic in computer vision due to the large illumination variations, insufficiently robust to viewing condition changes, low resolution images and partial occlusions, as seen in Fig. 1.

For reasons from above, achieving automated person re-identification is a pivotal problem to be solved for public in our lives. Traditionally, re-identification problem has been evaluated as a matching problem. Given a gallery set composing of a number of images of known individuals, for each test image or group of test images of an unknown person, the goal of person re-identification is to return a ranked list of individuals from the gallery set. In view of insufficient information of non-overlapping camera views, we can only use feature of a

✉ Chang Hong Lin
chlin@mail.ntust.edu.tw

Yu-Lun Wei
m10102107@mail.ntust.edu.tw

[1] Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607Taiwan, Republic of China

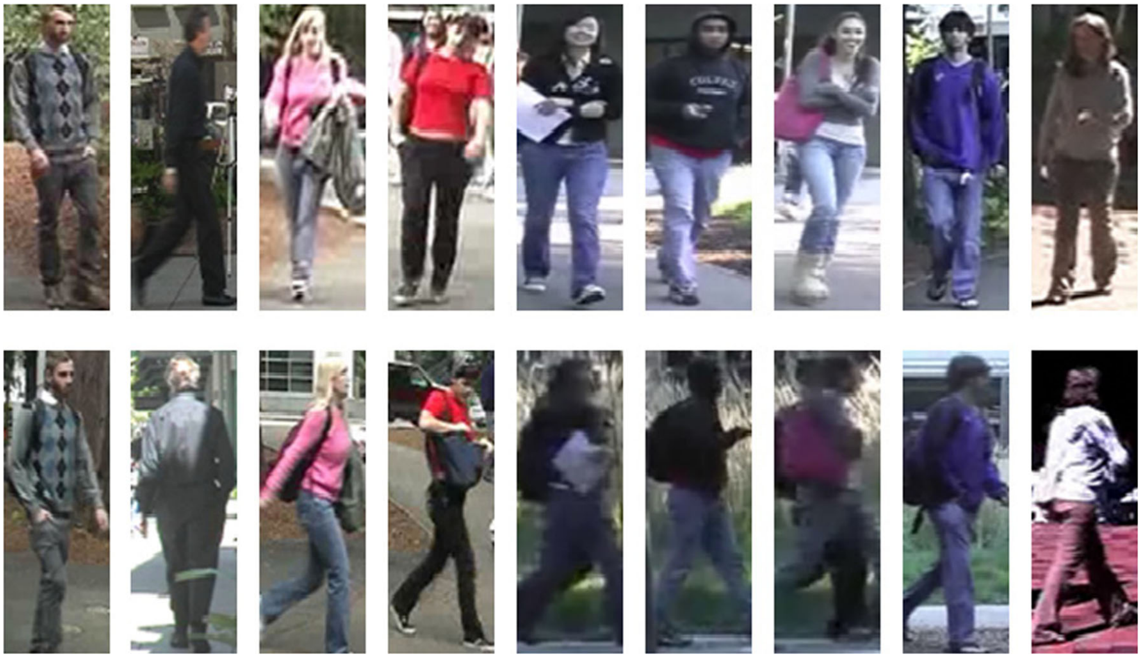


Figure 1 Examples of appearance changes by viewpoint and illumination variations. Each column shows the same person under two different views.

given people to find the best matched person from a lot of candidates. However, each view taken from a different angle and distance leads to degrees of occlusion and other view-specific variables. The descriptor can capture the most distinguishing characteristics of an appearance, while being invariant to camera changes. Although there are many challenges, the person re-identification problem always assumes people wear the same cloths under the multi-camera surveillance network. Moreover, depending on the number of available images per person, the used data can be separated into two types: single-shot cases, if only one individual is available in both the probe and gallery sets; multiple-shot cases, if multiple individuals are available in the probe sets and gallery sets or consisting a short video sequence, as seen in Fig. 2. Because of the inadequate information, the single-shot case is harder than the multiple-shot case.

Person re-identification has a wide range of applications and great commercial value. In order to deal with these challenges, the motivation in this article is to propose two robust algorithms with three-stage processes for appearance matching to improve the performance of single-shot person re-identification. The orientation of the first one: Gaussian Mixture Model of Weighted Color Histograms (GMMWCH) is the computation time; However, the orientation of the second method: Salience Gaussian Mixture Model of Weighted Color Histograms (SaliGMMWCH) is the recognition rate. In contrast to the previous methods, the main contribution of the proposed methods are two-fold: 1) Two unsupervised approaches are provided for the single-shot person re-identification, which implies the proposed methods have more flexibility and can be adopted to a large and variable number of persons. 2) Both the proposed frameworks have a lower computation time with superior accuracy for single-shot person re-identification. A preliminary result of GMMWCH was published in [1], and the SaliGMMWCH has not been previously published.

The rest of the article is organized as follows. In the section II, we briefly review related works for person re-identification. The section III and IV are the cores of this article, detailing the two proposed frameworks. The experimental results are reported in section V. Finally, the conclusions and future perspectives are described in section VI, for extending the proposed algorithm for the person re-identification problem.

2 Realated Wroks

In this section, several previous works relates to person re-identification would be introduced. According to the previous works, we classify them into three categories: supervised methods, unsupervised methods, and other methods. The first one focuses on learning feature representation, and the second focuses on feature extraction.

2.1 Supervised Methods

The first category is supervised methods. Generally, the characteristic of these frameworks, training samples with identity labels are required, which implies these have lower flexibility to be adapted to a large gallery of objects.

Du et al. used six kinds of popular color spaces as color features for a random forest to learn the similarity function of pair person images [2]. Discriminative models like SVM and

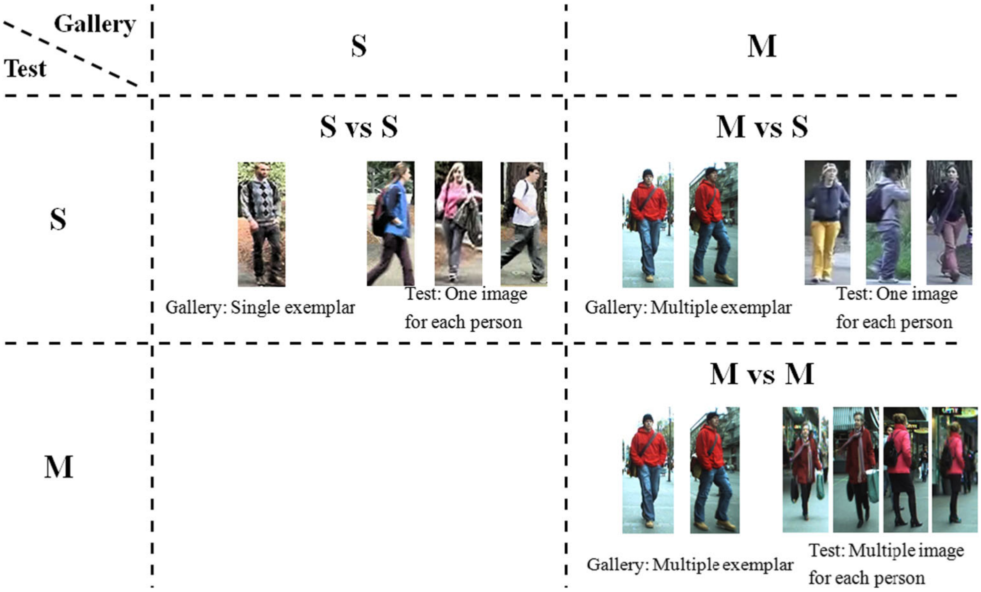

Gallery
Test
S
M
S vs S
M vs S
M vs M
S
M
Gallery: Single exemplar
Test: One image for each person
Gallery: Multiple exemplar
Test: One image for each person
Gallery: Multiple exemplar
Test: Multiple image for each person

Figure 2 Types of the person re-identification.

boosting [3–6] are widely used for feature learning. Gray and Tao used ensemble of localized features (ELF) with the Adaboost learning to recognize viewpoint invariant pedestrian [3]. Both [2, 3] proved the color histograms were effective. Similar to [3], Bak et al. also used AdaBoost to select the most discriminative Haar-like features for each individual [4]. Prosser et al. developed an alternative global selection approach by considering person re-identification as a relative ranking problem [5]. Color and texture-based features are extracted from six equal-sized horizontal strips in order to roughly capture the head, upper and lower torso and upper and lower legs. Schwartz et al. [7] proposed a method for learning discriminative appearance models by partial least square (PLS), which is known as the feature selection approach. This work can not only reduce background influence but also increase the best feature type for each person, but it is not flexible for variable number of persons. Leng et al. [8] not only consider the feature distance between the probe and gallery sets, but also use the contextual similarity between person images. Finally, they combine feature and contextual distance as a single feature distance to get the rank scores. Tao et al. [9] proposed regularized smoothing KISS metric learning (RS-KISS) by seamlessly integrating smoothing and regularization techniques for robustly estimating covariance matrices. Furthermore, they develop incremental RS-KISS (IRS-KISS) to deal the problem that the model needs to be updated to incorporate the information carried by the new labeled training samples. They conduct PCA [10] to obtain the low-dimension representation for each sample. Finally, by training RS-KISS or updating the distance metric by using IRS-KISS, they can find the rank scores. Zhao et al. [11] proposed an salience learning method by exploiting the pairwise salience distribution relationship between ped-estrian images. They integrate salience matching and patch matching in a feature and feed them into the structural RankSVM [12] learning to provide good performance.

Supervised method is mainly feature representation to find similarity functions with known objects. Therefore, these methods require training samples with supervised identity labels, and have no flexibility to be adapted to a large gallery objects. On the other hand, the proposed method does not require know identity labels.

2.2 Unsupervised Methods

The second category is unsupervised methods, which mainly focuses on feature design of given images. The pro-posed methods in this article are classified as this category.

Farenzena et al. [13] use the STEL [14] model to extract foreground, and an asymmetry human partition was proposed to separate head, torso and legs, while symmetry was used to divide the left and right parts. Moreover, the weighted HSV color histogram (wHSV), maximally stable color regions (MSCR) and recurrent high-structured patches (RHSP) were exploited to extract features. The result matching distance was the combination of the distances computed on each features. Chang et al. presented the approach based on an improved Random Walk algorithm, which segmented the human foreground by combining the shape information and the color seed into the Random Walks formulation [15]. The HSV color histogram, 1-D RGB signal and texture feature as the local binary pattern (LBP) and scale invariant local ternary pattern (SILTP) were employed to do feature matching. Zhao et al. [16] proposed an unsupervised salience learning method to exploit discriminative features, including color histogram and SIFT [17] feature. They used adjacency search and K-Nearest Neighbor algorithm (KNN) to select the possible candidates. The result matching score was computed by bidirectional similarity on each features. Malocal et al. [18] proposed a new descriptor by combining Fisher vectors with higher order statistics of local features, and to use the resultant representation (Local Descriptors encoded by Fisher vector, LDFV) to describe person images. Ma et al. [19] developed the BiCov descriptor, which relies on the combination of Biologically Inspired Features (BIF) and covariance descriptor, to compute the similarity of the BIF features at neighboring scales. With the BiCov descriptor, they can handle illumination change and background variations well.

For unsupervised methods, these approaches do not fully utilize the abundant information represented by their feature designs. Moreover, most of them take too much time for feature extracting and matching so that cannot be effectively applied in real-time systems. Therefore, both the proposed methods provide low computation and good recognition rate for single-shot case by exploiting Gaussian mixture model to generate the weighted color histogram. The first proposed method GMMWCM can achieve similar recognition rate 100 times faster than the above mention methods. Furthermore, by incorporating the salience information among pedestrian images in the second system, it can achieve much higher accuracy with similar execution time.

2.3 Other Methods

Different from these two categories, other works developed to handle pose variations, light condition and occlusions [20–24] are classified as the third category.

Wang et al. [20] proposed a multi-layer appearance modeling framework for computing the similarity between image regions to extract discriminative features robust to illumination and misalignment. The computational complexity is the major concern of this work. Gheissari et al. [21] focus on the algorithms that use the overall appearance of an individual. They develop two approaches which use interest operators and model fitting for establishing spatial correspondences between individuals. Bak et al. [22] focus on human signature computation, they first applied Histogram of Oriented Gradient (HOG) as the body detector to establish the correspondence between body parts, then using spatial covariance regions extracted from human body parts to handle pose variation. Cheng et al. [23] adopt Pictorial Structures (PS) to localize the body parts, extract and match their descriptors to deal the pose variation challenge. However, these approaches are not flexible enough and only applicable while the pose estimation work accurately. Zheng et al. [24] consider person re-identification as a distance learning problem and use a novel Probabilistic Relative Distance Comparison (PRDC) model to learn optimal distance to improve the accuracy. Their experiments demonstrate the improvement is more significant while the training sample size is small, which implies they do not need a large of training samples. That performance make this approach noteworthy in the distance learning methods.

3 Gaussian Mixture Model of Weighted Color Histograms

A preliminary GMMWCH system was published in [1]. This article contains more details of the proposed system, as well as analyses in different color spaces. A three-stage process is introduced in this system, as seen in Fig. 3. The first stage is the image enhancement by illumination normalization. The second stage includes pedestrian segmentation and human region partition. Both the STEL model from [14] and the symmetry-based partition method proposed in SDALF [13] are explored. The third stage is used to perform feature extracting and matching.

3.1 Illumination Normalization

The proposed approach is based on the wHSV combined with spatial information of body segments, which is derived from the SDALF [13]. Therefore, intensity variations between probe and gallery images play a key role in correct matching rate. In view of this, illumination normalization is employed as the proposed system's first stage, after loading the probe and gallery images. As depicted in Fig. 4a, there are large light variations in person's appearance, due to uncontrolled changes in illumination, viewing direction, and camera parameters. These make incorrect feature matching even on small quantized bins in the intensity channel of HSV color space. To overcome this problem, the illumination normalization is exploited here. Unlike normal normalization step in the RGB color space, the illumination normalization only normalizes the Y channel of the YCbCr color space, because the insufficient light source and shadows are the main reasons of incorrect feature matching. For that V channel of HSV color space represents intensity of a color, which is decoupled from the color information in the represented image, the Y channel is used instead of directly using the V channel for normalization.

Results are shown in Fig. 4b, the appearance of human can be obtained after the illumination normalization. The performance of the wHSV with/without the illumination normalization process is evaluated by the Cumulative Matching Characteristic (CMC) curve [25], which represents the probability of finding the correct match in a range of top *n* rank, on the VIPeR [25] dataset in Fig. 4c.

3.2 Pedestrian Segmentation and Human Region Partition

Finding correct features is a key point to recognition performance. Therefore, the method to separate foreground

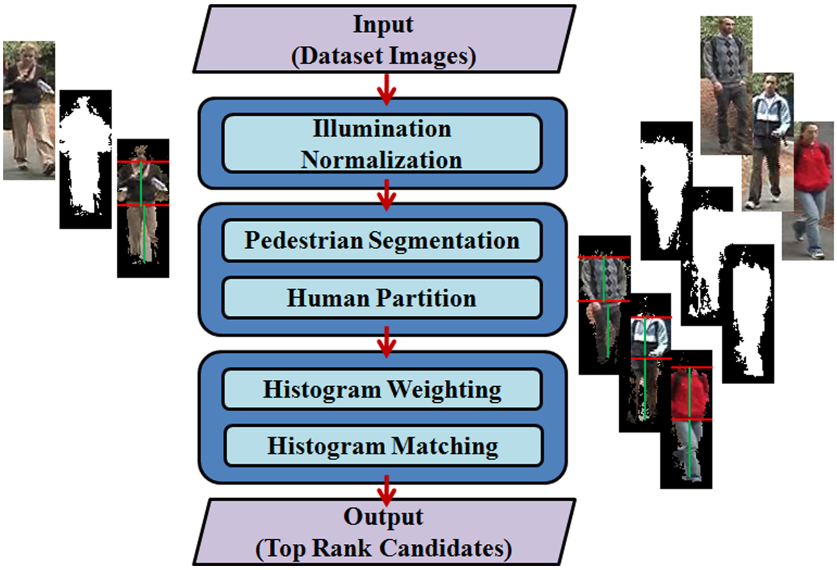


Figure 3 System flowchart.

(FG) and background (BG) is applied here to improve the accuracy of feature extracting and matching, the STEL generative model [14] has been customized here for the FG/BG separation, as seen in Fig. 5. For human region partition, the symmetry based silhouette partition method proposed in SDALF [13] is used to locate the torso and legs. The method is dependent on the visual and positional information of the clothes, and it is robust to viewpoint variations and low resolution. Since the head does not carry enough information due to the low image resolution, it has to be discarded. As depicted in Fig. 6, The values i_{HT} and i_{TL} isolate three regions, $R_k(k = \{0,1,2\})$, approximately corresponding to the head, body and legs, respectively. Similarly, the values j_{Lr1} and j_{Lr2} separate the body and legs into left part and right part.

3.3 Gaussian Mixture Model of Weighted HSV Histograms

In order to achieve the goal of low computation time, a fast color-based feature is proposed for person re-identification. The wHSV, the most contributed color feature in the SDALF [13], uses single one-dimensional Gaussian kernel G(μ, σ) to generate the weighted color histogram. Figure 7a shows the result of human region partition and the corresponding one-dimensional Gaussian kernel, and the darker pixels mean the relevant color pixels are more important. In other words, the pixel values near the j_{Lr1} and j_{Lr2} count more in the final histogram. However, using single one-dimensional Gaussian kernel to weight both the upper and lower body's histograms might not be

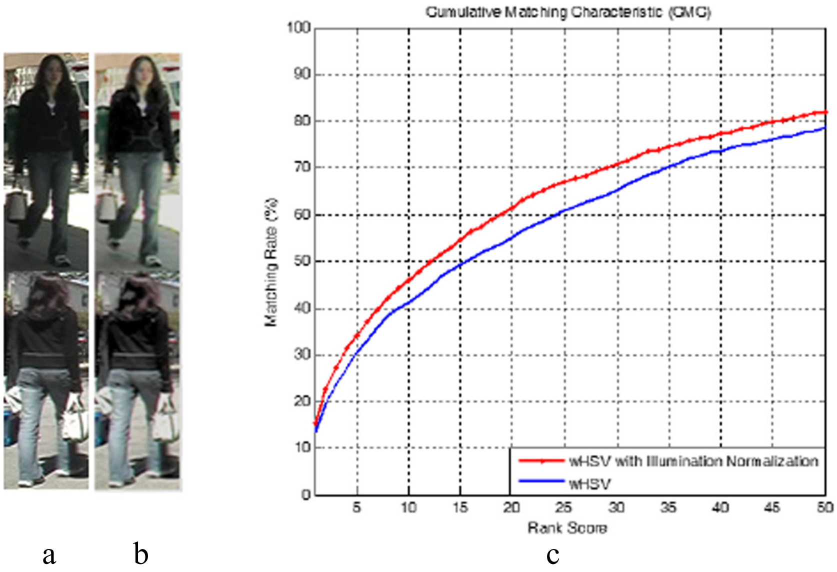


Figure 4 **a** Sample images with large intensity variations. **b** Results after applying illumination normalization. **c** CMC curve of the performance of person re-identification. (on VIPeR [25]).

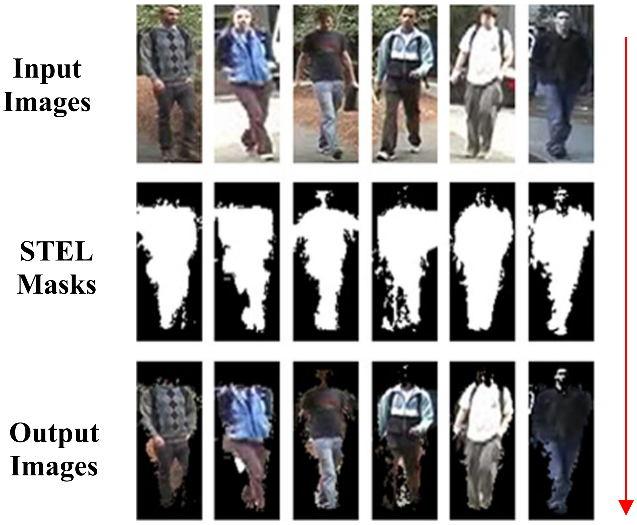


Figure 5 Flowchart of the Pedestrian segmentation.

enough. For the upper body, human's clothes are the most important feature, and it would need more than one single Gaussian kernel to get the adequate color information. For the lower body, if human's legs are not very closed when human walks, then the j_{Lr2} won't stay in the middle of human's legs, and the Gaussian kernel cannot get the best weighted color histograms. In fact, because of human's walking posture, it's unlike for human's legs to stay closed when walking, which implies that using a single Gaussian kernel for the lower body would not be sufficient.

In order to deal with the problem, the Gaussian mixture models $G_{Mix}(\mu, \sigma)$ (GMM) is employed instead of the single Gaussian kernel to weight HSV color histogram obtained as:

$$G_{Mix}(\mu, \sigma) = \sum_{i=1}^{N} w_i \cdot G(\mu - x_i, \sigma_i) \quad (1)$$

The w_1 and w_2 are the weight parameters of each single Gaussian kernel, where μ is the y-coordinate of the j_{Lrk}. In the experiment, the σ_1 and σ_2 is priori sets to 6.5 for upper body and 5.5 for lower body. It is because the upper body is often wider than the lower body, so a larger deviation can cover a wider area, which can better represent the upper body. Figure 7b illustrates the result of human region partition and the corresponding GMM, and darker pixels mean the relevant color pixels are more important. Compare Fig. 7b to Fig. 7a, the upper GMM can get more color histogram information for upper body, and lower GMM can match lower body's histograms more correctly. With both the upper and lower GMMs, we can get sufficient histogram information. The performance of person re-identification with/without using the GMMs (without illumination normalization) is evaluated as well. The results are shown in Fig. 7c. Different color spaces like as the RGB, YCbCr and LAB (CIELAB) model are also evaluated, but the HSV color space has been shown to be superior than others. It also prove that the HSV model is excellent to against different environmental illumination conditions and camera acquisition settings. The person re-identification performance in different color space on VIPeR [25] is shown in Fig. 8.

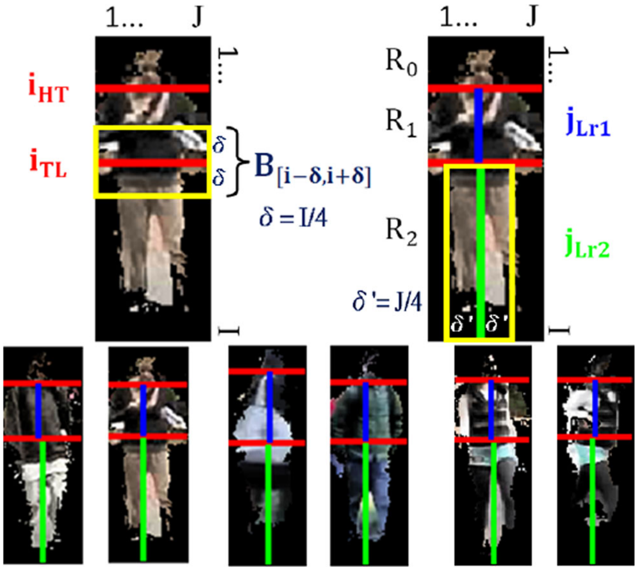


Figure 6 Symmetry-based Silhouette Partition. On the top row, overview of the region partition method in SDALF [13]. On the bottom row, examples of symmetry-based partitions on images from the VIPeR dataset.

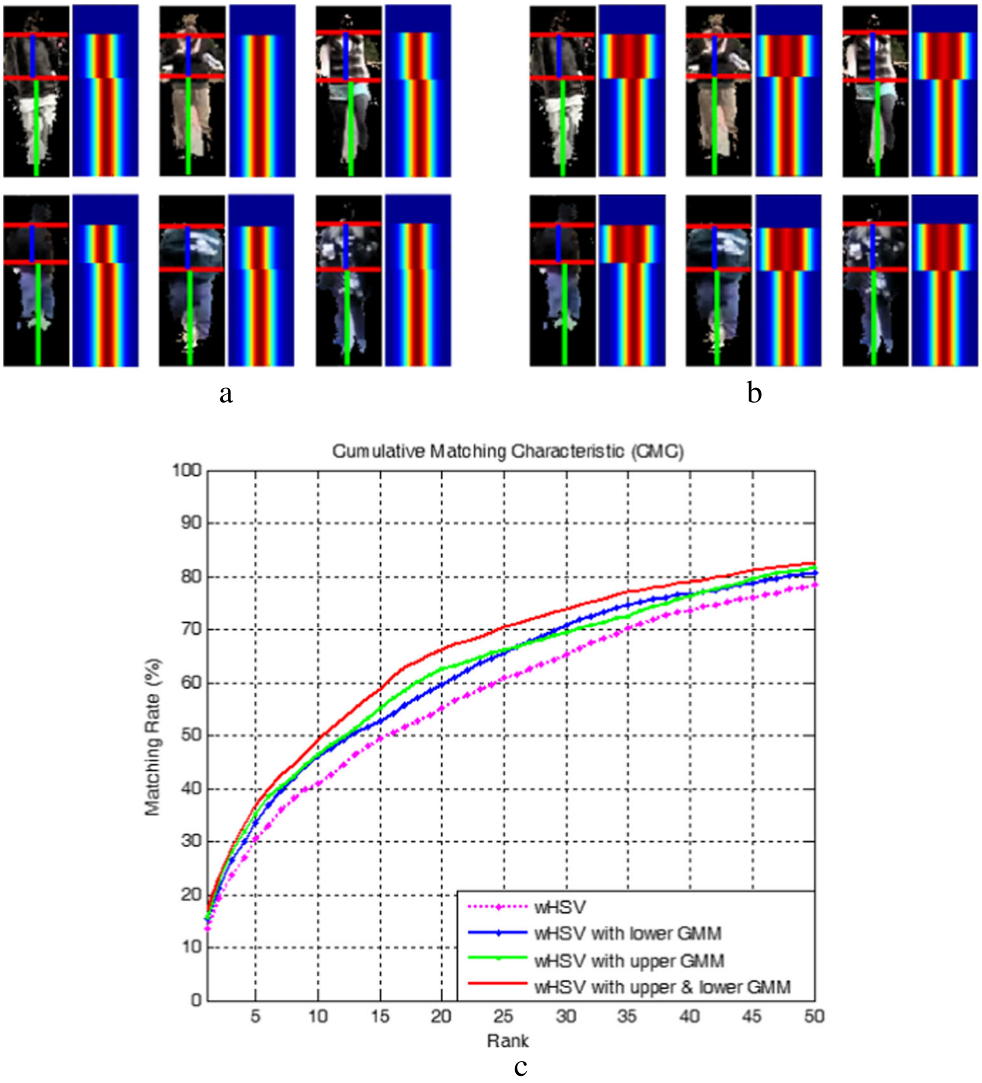
4 Salience Gaussian Mixture Model of Weighted Color Histograms

A three-stage process is introduced in this proposed system, as seen in Fig. 9. The first stage is the image enhancement by illumination normalization. The second stage is the human region partition. The symmetry-based partition method proposed in SDALF [13] is explored here. The third stage is used to perform feature extracting and matching.

4.1 Dense Gaussian Mixture Model of Weighted LAB Histograms

The dense correspondence combined with keypoint feature matching and patch matching has the characteristics of robust alignment [26, 27]. Since many people tend to dress in very similar ways, it is important to capture as fine image details as possible. In order to get low computation time and superior recognition rate simultaneously, the mid-level local patch feature, Dense Gaussian Mixture Model of Weighted LAB Histograms (dGMMwLAB), is adopt here for assoc.-iating persons.

Figure 7 **a** Result of human region partition and corresponding Gaussian kernel. **b** Result of human region partition and corresponding GMM. **c** CMC curve of the performance of person re-identification. (on VIPeR [25]).



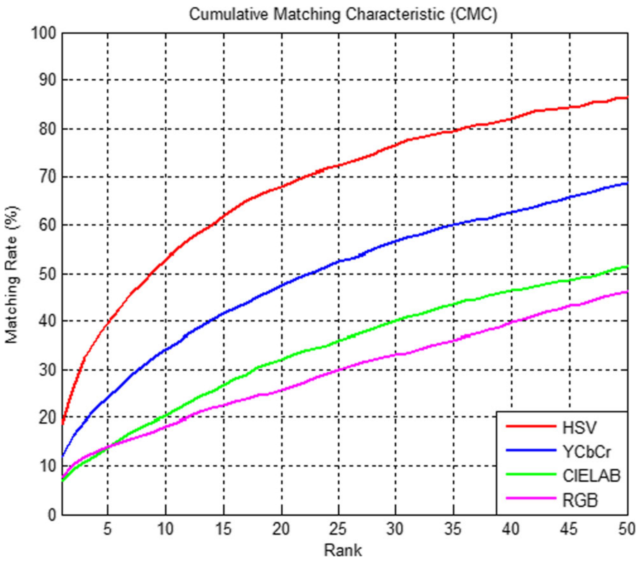


Figure 8 Performances of GMMWCH method in different color spaces.

The proposed dGMMwLAB is a multi-dimensional descriptor vector for each patch. Before building the dense correspondence, the illumination normalization and human region partition are first applied as same as GMMWCH method. Unlike GMMWCH, SaliGMMWCH does not use the pedestrian segmentation step before the human region partition. This is because a more complete image can provide more histogram information for patch matching to improve the patch matching accuracy. Figure 10 illustrates the result of human region partition and the corresponding GMMs, and darker pixels mean the relevant color pixels are more significant. Afterwards, considering appropriate resolution of human images captured by far-field surveillance cameras, the GMMwHSV step is utilized to attain robust color histogram information. As depicted in Fig. 10, different from the GMMWCH method, SaliGMMWCH consider the "head" part for getting the color information as well.

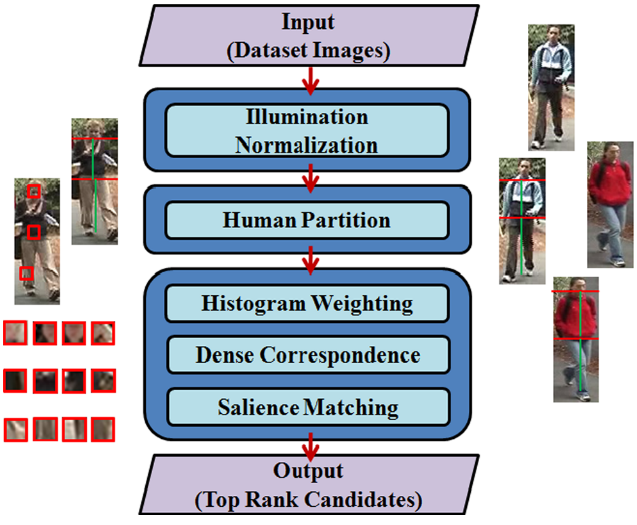


Figure 9 System flowchart.

Finally, local patches on a dense grid is extracted for each person image. The detail parameters of dGMMwLAB feature extraction in the experiment are as follows: We sample the 10 × 10 patch size on a dense grid with a grid step size of 4; According to the image in the VIPeR dataset, the size of these images is 128 × 40, and the average human face size is around 10 × 10, so we choose 10 × 10 as the patch size. Using a large step size can reduce the number of patches in the order of two, and can speed up accordingly. However, the accuracy would decrease as the step size increases. A step size of 4 has almost the same accuracy as the step of 1, but can be 16 times faster. In summary, there are total 30 × 10 salient patches for each 128 × 48 image. Each salient patch computes the three 30-bin histograms in L, A, B channels respectively, and in order to robustly capture the color information, the LAB weighted color histogram is downsampled with scaling factors 0.75 and 1. Each salient patch is finally represented by a discriminative descriptor with length 180 (30 × 3 × 2) feature vector, and there are some examples of patches from VIPeR [25] dataset depicted in Fig. 11.

4.2 Dense Adjacency Constrained Search

The camera views at different locations, misalignment and vertical articulation may lead to the vertical movement of the human body in the image. In order to handle spatial variations and make patch matching more adaptive in person re-identification problem, a simple but effective horizontal adjacency search is employed.

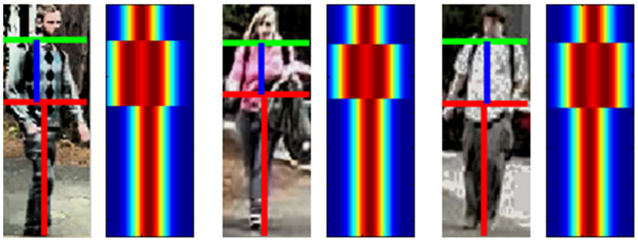


Figure 10 Silhouette partition and the corresponding Gaussian kernel.

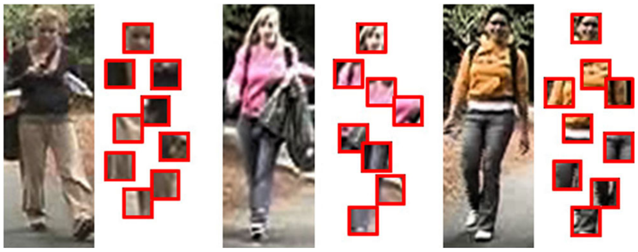


Figure 11 Examples of salient patches.

The collection of all dGMMwLAB features in a pedestrian image are represented as $x^{A,p} = \left\{x_{m,n}^{A,p}\right\}$, where *(A, p)* denotes the *p*-th image in the camera A, and (m, n) denotes the patch centered at the *m*-th row and the *n*-th column of the image *p*. The collection of dGMMwLAB features in *m*-th row of image p from the camera A is represented as $Pat^{A,p}(m) = \left\{x_{m,n}^{A,p} \middle| n = 1, 5, 9, \ldots, N\right\}$, where N is the number of the columns. The patch matching between all patches in $Pat^{A\,,p}(m)$ and corresponding patch set *S* in image *q* from the camera B is represented as:

$$S\left(x_{m,n}^{A,p}, x^{B,q}\right) = Pat^{B,q}(m), \forall x_{m,n}^{A,p} \in Pat^{A,p}(m) \tag{2}$$

The patch set *S* restricts the search set in image *q* within the *m*-th row. However, the human pose variations caused by uncontrolled camera view changes lead to the pedestrian is not always well aligned and have some vertical movements of the body. In order to deal with the spatial variations, the strict horizontal constraint search is relaxed to have a larger search range like as [16]:

$$\hat{S}\left(x_{m,n}^{A,p}, x^{B,q}\right) = \left\{Pat^{B,q}(b) \mid b \in \mathcal{N}(m)\right\}, \forall x_{m,n}^{A,p} \in Pat^{A,p}(m) (3)$$

where $N(m) = \{m\text{-}l, \ldots, m, \ldots, m + l\}$, $m\text{-}l \geqq 0$, $m + l \leqq M$, and *M* is the number of the rows. The value *l* defines the half height of the relaxed adjacent vertical space. If *l* is very small, the small search space cannot tolerate the large spatial variation, and the target patch cannot find the correct corresponding patch. While *l* is set to be very large, the large search space will increases the chance of mismatch. In order to strike the balance between the vertical toleration and the chance of mismatch, $l = 2$ is chosen in the experiment setting. Figure 12 shows some visually similar patches returned by the adjacency constrained search.

4.3 K-Nearest Neighbor Salience Matching

For problems such as symmetry detection and object detection, we would like to compute more than one single nearest neighbor

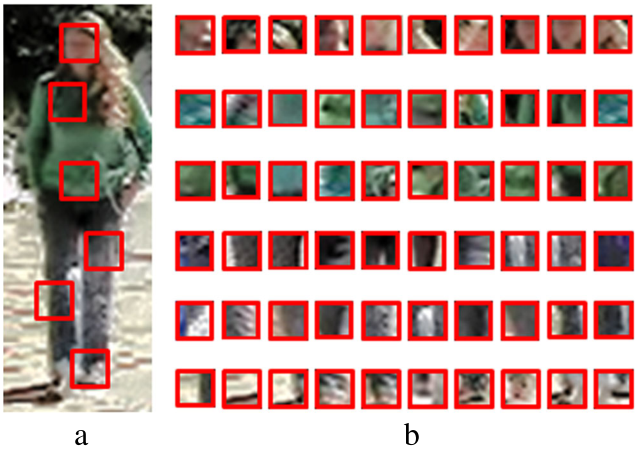


Figure 12 Examples of adjacency search. **a** A test image from the VIPeR [25] dataset. *Red* boxes show six patches of different body parts. **b** The top ten nearest neighbor patches by adjacency search are shown. Note that the ten nearest neighbor patches are from ten different pedestrians.

at different positions. This can be done by collecting the k nearest neighbors for each patch. The salient patches which computed based on previously-built dense correspondence posses the uniqueness property. In order to select those features, the K-Nearest Neighbor (KNN) algorithm [28] is utilized here to find patch samples in the minority of the corresponding set in the same spirit of [16]. With this strategy, human salience is better adapted to the pedestrian re-identification problem. Afterwards, we find salient patches that retain the property of uniqueness among the reference set *R*. Then we fix size N$_r$ for the number of images in the reference set. For a person's image $x^{A,p}$= $\{x^{A,p}_{m,n}\}$, a nearest neighbor (NN) set of size N_r is built for every test patch $x^{A,p}_{m,n}$ to find similar patches obtained as:

$$X_{NN}\left(x^{A,p}_{m,n}\right) = \left\{x \mid argmin_{x^{B,q}_{i,j}}\, d\left(x^{A,p}_{m,n}, x^{B,q}_{i,j}\right), x^{B,q}_{i,j} \in \hat{S}\left(x^{A,p}_{m,n}, x^{B,q}\right), \quad q = 1, \dots, N_r\right\} \tag{4}$$

where $\hat{S}\left(x^{A,p}_{m,n}, x^{B,q}\right)$ is the adjacency search set of the salient patch $x^{A,p}_{m,n}$ in Eq. (3). In order to get similarity of two salient regions, the function *d(·)* is used here to compute the Euclidean distance between two salience distributions.

The purpose of computing human salience is to identify patches with unique characteristic. So that the similar scheme in [28] is applied to $X_{NN}\left(x^{A,p}_{m,n}\right)$ of each test patch here, and the KNN distance is exploited to define the salience score obtained as:

$$score_{KNN}\left(x^{A,p}_{m,n}\right) = W_k\left(X_{NN}\left(x^{A,p}_{m,n}\right)\right) \tag{5}$$

where W_k represents the weight of the *k*-th nearest neighbor. The value W_k is defined by *1/D$_k$*, and D_k is the Euclidean distance to other neighbors. The goal of salience detection is to identify persons with special appearance, thus we assume the reference set can reflect the test scenario well, so that the test patches can only find the limited number $(k = \alpha N_r,\ 0 < \alpha \leq 1)$ of visually similar neighbors. Furthermore, we believe that more than half of the pedestrian in the reference set R are dissimilar with him/her, so α = *1/2* is set in the experiment. The best choice of k should depend upon the data. Generally, the larger value of k can reduce the effect of noise on the classification, but make boundaries between classes less distinct. While the *k* is small, it will increases the chance of mismatch, but has low computation costs. In order to strike the balance and consider the robustness of the proposed patch feature, N_r= *80* and k = N_r*/2* = *40* is a proportion parameter reflecting our expectation in the experiment. Since k depends on the size of the reference set *R*, the defined salience score can work well even if the reference set's size is very large.

In order to make salience score in Eq. (5) more adapted to the experiments, $score_{KNN}\left(x^{A,p}_{m,n}\right)$ is normalized as:

$$NorScore_{KNN}\left(x^{A,p}_{m,n}\right) = \frac{score_{KNN}\left(x^{A,p}_{m,n}\right) - lwDist}{upDist - lwDist} \tag{6}$$

where *upDist* is the maximum of $score_{KNN}\left(x^{A,p}_{m,n}\right)$ and *lwDist* is the minimum of $score_{KNN}\left(x^{A,p}_{m,n}\right)$. By Eq. (6), the value of $NorScore_{KNN}\left(x^{A,p}_{m,n}\right)$ is located between 0 and 1 to adapted to the similarity score computation.

In order to incorporate salience information into dense correspondence matching, a weighting mechanism is built to return a ranked list of individuals from a lot of candidates after we get the normalized salience scores of those patches. In fact, the same person's images taken by different non-overlapping cameras would still be likely to have more similar salient patches than those of the different pedestrians. Consequently, the difference of two persons' salience scores is used as a penalty to the similarity score, and the product of salience scores are used to enhance the similarity score of matched patches. Finally, the weighting mechanism is formulated as follows:

$$SimScore_{sali}\left(x^{A,p}, x^{B,q}\right) = \sum_{m,n} \frac{NorScore_{KNN}\left(x^{A,p}_{m,n}\right) \cdot NorScore_{KNN}\left(x^{B,q}_{i,j}\right) \cdot d\left(x^{A,p}_{m,n},\ x^{B,q}_{i,j}\right)}{\varepsilon + \left|NorScore_{KNN}\left(x^{A,p}_{m,n}\right) - NorScore_{KNN}\left(x^{B,q}_{i,j}\right)\right|} \tag{7}$$

where ε is the very small number in order to prevent the denominator being zero, and the function *d(·)* is the Euclidean distance used to compute the similarity between two salient patch features.

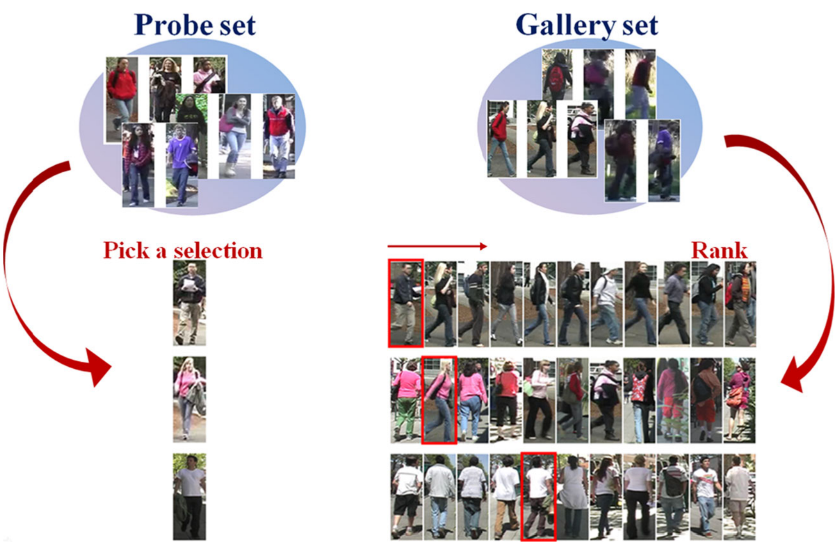


Figure 13 Evaluation process.

By finding the maximal similarity score between a pair of person images, the best matched image can be determined obtained as:

$$target_{person} = \underset{q}{argmax}\, SimScore_{sali}\left(x^{A,p}, x^{B,q}\right) \tag{8}$$

where $x^{A,p} = \left\{x_{m,n}^{A,p}\right\}_{m\in M, n\in N}$ and $x^{B,q} = \left\{x_{i,j}^{B,q}\right\}_{i\in M, j\in N}$ are collection of salient patch features in two person images. With the Eq. (8), we can get the best matched person from candidates, and this is what person re-identification does.

4.4 Feature Combination for Ranking Score

In this section, we illustrate how the different features are jointly used as a single similarity score for matching to improve the recognition rate. The purpose of person re-identification is to associate each person of set A to the corresponding person of set B captured in distributed locations at different times, which I_A is an image belongs to the gallery set and I_B is an image from the probe set. After we obtain the dGMMwLAB feature, we combine it with the similarity score of GMMwHSV feature into a single sim-ilarity score for matching images obtained as:

$$\begin{aligned} &SimScore(I_A, I_B) \\ &= \beta_{sali} \cdot SimScore_{sali}(I_A, I_B) + \beta_{wHist} \cdot SimScore_{wHist}(I_A, I_B) \end{aligned} \tag{9}$$

where $SimScore_{sali}(I_A, I_B)$ is the similarity scores of the dGMMwLAB feature between I_A and I_B, and the $SimScore_{wHist}(I_A, I_B)$ is similarity scores of GMMwHSV feature. *The* β_{sali} and β_{wHist} are the weight parameters for the dGMMwLAB and GMMwHSV feature respectively.

Proposed approaches have the complementary characteristic of existing approaches to further improve the recognition rate. Therefore, the similarity score is extended by combining the similarity scores of existing approaches with the similarity score in Eq. (9), the final similarity score between a pair of images is defined as follows:

$$\begin{aligned} &eSimScore(I_A, I_B) \\ &= \sum_i \beta_i \cdot d_i(f_i(I_A), f_i(I_B)) + SimScore(I_A, I_B) \end{aligned} \tag{10}$$

where $f_i(I_A)$ is the feature of the I_A, and the distance d_i evaluates feature's similarity between two persons. The β_i is the

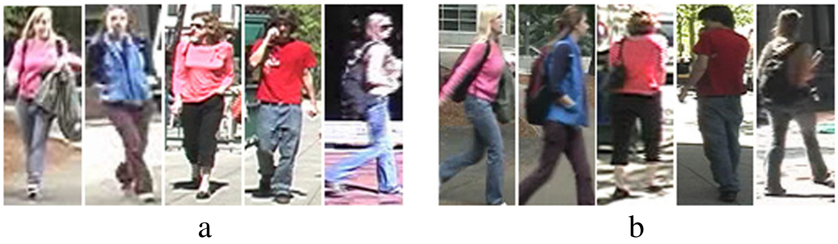


Figure 14 Examples of viewpoint change between a pair of images on the VIPeR [25]. **a** Images captured by Camera A are mainly from 0 degree to 90 degree. **b** Images captured by Camera B are mainly from 90 degree to 180 degree.

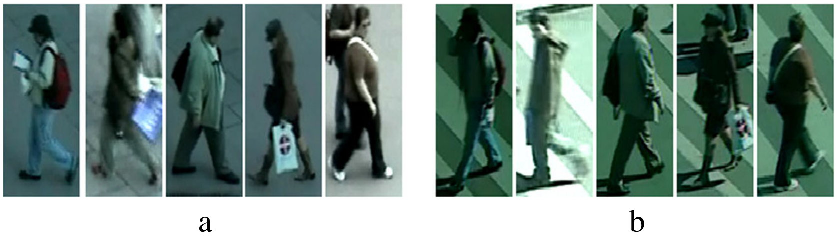


Figure 15 Examples of viewpoint change between a pair of images on the PRID 2011 [6]. Upper and lower row correspond to different camera views.

parameter to control the weight for the *i*-th distance measure. In the experiment, we combine the MSCR feature in [13] with the proposed framework, SaliGMMWCH. The experiment result of the combination with other existing app-roaches, eSaliGMMWCH, is shown in experimental results.

5 Experimental Results

In this section, we would show several experiments and analyses. We first describe our developing platform and introduce the used benchmark dataset, VIPeR [25]. Then we show performance of the proposed methods in different color spaces for comparison. Finally we show extensive experi-ments to evaluate the approaches. The evaluation process is shown in Fig. 13. For performance evaluation, we use the Cumulative Matching Characteristic (CMC) curve [25] to show the recognition rate on each top rank. Rank-*k* reco-gnition rate indicates the probability to find each probe image matched correctly to the right gallery image at rank *k*, and the CMC curve [25] is the cumulated values of the recognition rate at all ranks. Comparisons with other methods in the state-of-the-art benchmark datasets are also provided.

Developing Platform Both the proposed frameworks are implemented partly in C++ and partly in MATLAB without any particular optimization or parallelization. The operating system is Microsoft Windows 7. All the experimental results are executed on a general computer with an Intel i7–3770 CPU and 8GB memory.

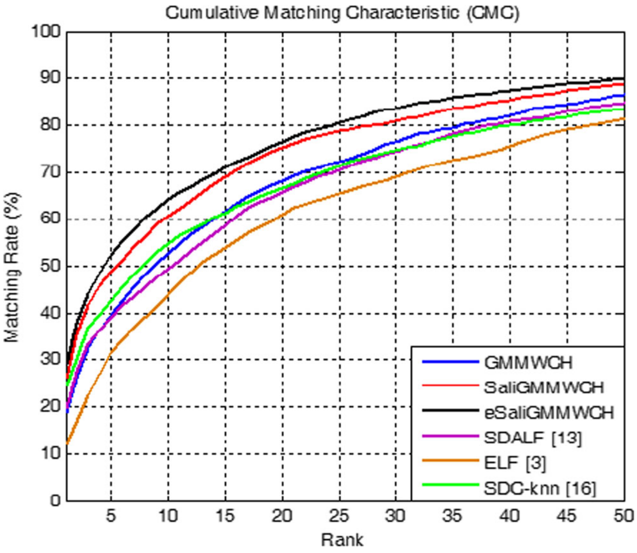


Figure 16 CMC curve of the performance comparison on VIPeR [25].

VIPeR Dataset [25] The VIPeR dataset is the most widely used for evaluation and reflect most of the challenges in real-world person re-identification applications. That is captured by two non-overlapping cameras from arbitrary viewpoints in outdoor environment with two images for each person, under significant viewpoint change, pose variation, varying illumination conditions, low resolution, occlusions and so on. It is composed by 632 pedestrian pairs, each pair is made up of images of the same pedestrian shoot from two different camera views, camera A and camera B. Each image is scaled to 128 × 48 pixels for experiments. We also show some examples of viewpoint variation on the VIPeR [25] dataset in Fig. 14. As can be seen from the figure, camera A captured images mainly from 0 degree to 90 degree while camera B mostly from 90 degree to 180 degree, and most of the examples contain viewpoint change over 90 degree. Due to its complexity, few researchers only have published their quantitative results on the VIPeR [25] dataset. It is the most challenging datasets currently available for single-shot person re-identification problem. In the experiment, 316 image pairs are first evenly chosen from dataset to build the gallery and probe sets. Then each image of the gallery set is matched with the images of the probe set. This provides a ranking for every

Table 1 Top ranked matching rates in [%] (VIPeR [25]).

Method	Rank 1	Rank 10	Rank 20
GMMWCH	18.76	52.87	68.26
SaliGMMWCH	25.85	60.73	75.09
eSaliGMMWCH	29.08	64.11	76.52
ELF [3]	12.24	42.58	59.85
SDALF [13]	19.52	49.48	65.64
SDC_knn [16]	24.28	53.73	66.61
eLDFV [18]	22.34	60.04	71.00
eBiCov [19]	20.66	56.18	68.00
PRDC [24]	15.66	53.86	70.09
OAR [35]	21.4	41.5	71.5
Mahalanobis [36]	16	54	72

Table 2 Top ranked matching rates in [%] (PRID 2011 [6]).

Method	Rank 1	Rank 10	Rank 20
SaliGMMWCH	14.7	41.6	52.4
eSaliGMMWCH	16.6	43.1	54.6
Mahalanobis [29]	16	41	51
KISSME [30] [29]	15	39	52
EIML [31] [29]	16	39	51
LMNN [32] [29]	10	30	42
LMNN-R [33] [29]	9	32	43
ITML [34] [29]	12	36	47
OAR [35]	41.5	82.5	86.7
Mahalanobis [36]	16	41	51

image in the gallery with respect to the probe. This whole evaluation procedure is repeated 10 trials in order to provide a robust statistics. For a fair comparison, the same way is used to choose these 316 image pairs as the public data from the SDALF [13] framework.

PRID 2011 Dataset [6] The PRID 2011 dataset created in 2011 by the Austrian Institute of Technology consists of person images recorded from two different cameras. Both multi-shot and single-shot scenarios are provided in this dataset. Since the proposed method are focusing on the single-shot case, we use only the latter one. Typical challenges on this dataset are significant illumination change, pose variation and occlusions due to the differences in environment and camera characteristics. 385 persons will filmed by camera A, and 749 persons were filmed by camera view B, with 200 of them appearing in both views. Each image is scaled to 128 × 64 pixels for experiments. We also show some persons on the PRID 2011 [6] dataset in Fig. 15. For the evaluation, these image pairs are randomly split into the gallery and probe sets of equal size. Thus, searching the 100 first persons of one camera view in other persons of the other view. This whole evaluation procedure is repeated 10 trials to provide a robust statistics.

We compare proposed methods with state-of-the-art techniques on the public available benchmark dataset, VIPeR [25] and PRID 2011 [6], for the evaluation. Since ELF [3], SDALF [13] and SDC_knn [16] have published their results on the VIPeR [23] dataset, they are used for the comparison. The same splitting assignments in these approaches are used in the experiments. As can be seen from the Fig. 16, we compare the performance of the proposed frameworks with the ELF, SDALF and SDC_knn by the CMC curves [25]. The experimental results show that the proposed implementations of the GMMWCH and SaliGMMWCH attains the performance which are better than most of the three benchmarking approaches. In particular, rank 1 matching rate is around 18.8% for GMMWCH and 25.85% for SaliGMMWCH, versus 12.2% for ELF, 19.5% for SDALF, 24.3% for SDC_knn, 22.34% for eLDFV [18], 20.66% for eBiCov [19] and 15.66% for PRDC [24]. The matching rate at rank 10 is around 52.9% for GMMWCH, and 60.7% for SaliGMMWCH, versus 42.6% for ELF, 49.5% for SDALF, and 53.7% for SDC_knn, 60.04% for eLDFV [18], 56.18% for eBiCov [19] and 53.86% for PRDC [24]. Furthermore, by combining with other existing feature descriptors, MSCR [13] feature, the rank 1 matching rate of eSaliGMMWCH goes to 29.1%, and the matching rate at rank 10 goes to 64.1%. This result shows the well complementarity of proposed approaches to other features. More comparison results on VIPeR [25] and PRID 2011 [6] datasets are show in Tables 1 and 2. Since the proposed GMMWCH need masks to separate the foreground and background for improving accuracy, the GMMWCH didn't be evaluated on PRID 2011 [6] datasets.

The proposed GMMWCH and SaliGMMWCH not only have better rthan those methods, but also have outstanding execution performance. The experimental results in Table 3 report that the proposed frameworks have a superior average computation time. The GMMWCH took only 0.006 s to extract features and performed a match in 0.036 s, and SaliGMMWCH took 0.58 s to extract features and performed a match in 0.77 s. Both the two proposed methods were implemented partly in C++ and partly in MATLAB without any particular optimization or parallelization. The SDALF and SDC_knn evaluated by using the publicly available source codes provided by the authors [13] [16], and were implemented partly in C++ and MATLAB as well. As a qualitative comparison, the SDALF requires over 2.3 s to extract features, and performs a match in 2.5 s. The SDC_knn requires over 1 s to extract features, and performs a match in 3 s.

The improvement of proposed methods can be explained in three aspects: First, most of the false positives are due to severe lighting changes, which the illumination normalization step can handle it effectively. Second, since many people tend to dress in very similar ways, it is important to capture as fine image details as possible. This is what the color histogram weighted by Gaussian mixture model does, and it further provide a not only robust but also efficient feature to deal those situations.

Table 3 Average computation time per person (VIPeR [25]).

Method	Feature extracting	Matching	Total
GMMWCH	0.006 s	0.036 s	0.04 s
SaliGMMWCH	0.58 s	0.77 s	1.35 s
eSaliGMMWCH	0.63 s	3.55 s	4.18 s
SDALF [13]	2.36 s	2.57 s	4.93 s
SDC_knn [16]	1.07 s	3.01 s	4.08 s

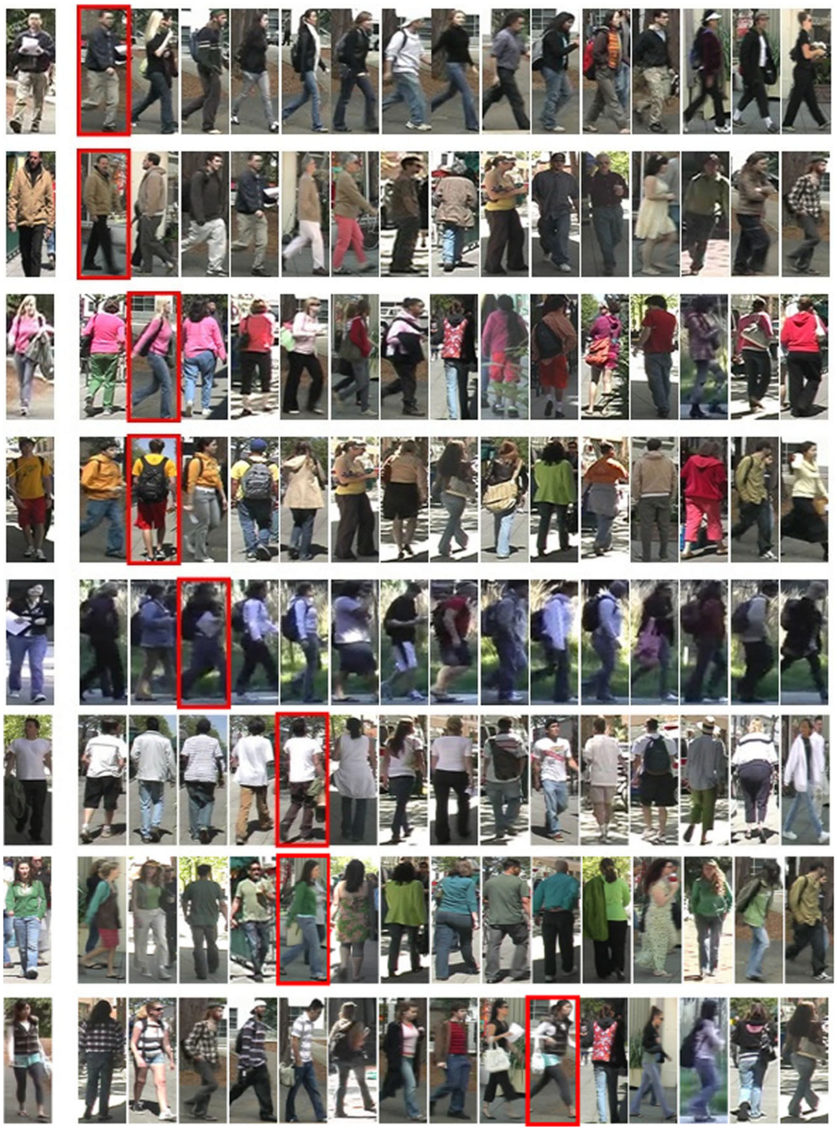


Figure 17 Examples of person re-identification on VIPeR [25]. The first column indicates the probe image, and the remaining columns shows the ranked results with the correct match in red. And the rank of the correct match are 1,1,2,2,3,5,5,10.

Third, with incorporating human salience information to dense correspondence matching, it can tolerate larger extent of pose and appearance variations.

Even though the main purpose of the person re-identification is to find the rank of interest person as top as possible, the first concern in both the two proposed approaches is the balance between the recognition rate and execution performance in order to make proposed systems more adapted to real-time applications. Finally, Fig. 17 shows some examples of ranked results of person re-identification of 8 probe images.

6 Conclusions

In this article, we proposed two different unsupervised frameworks for solving the single-shot person re-identification problem. Both the proposed methods are based on the color histogram feature. The first proposed method, GMMWCH, includes the illumination normalization step to make it robust to changing illumination conditions. It retains the original concept of wHSV [13] and employs the Gaussian mixture model to get sufficient color histogram information. It can accordingly improve its recognition rate and execution performance. In addition, the GMMWCH consider person re-identification as a matching problem, which implies it has more flexibility and can be adopted to a large and variable number of persons. For the second framework, SaliGMMWCH, a similar scheme like as the GMMWCH method is provide. The illumination normalization and human partition step are applied for reducing the chance of patch mismatch. Then, the salient patch matching combined with our color histogram weighted by the Gaussian mixture model is utilized with adjacency constraint search for handling the viewpoint and pose variation. With this strategy, it shows great flexibility in matching across large viewpoint change and excellent performance.

Experiments show that both the proposed approaches not only greatly improve the recognition rate of single-shot person re-identification but also have a lower computational cost than state-of-the-art techniques on benchmark dataset. The GMMWCH takes only 0.04 s for feature extracting and matching, and SaliGMMWCH takes 0.58 s to extract features and performed a match in 0.77 s. Therefore, the proposed frameworks have more possibility to be applied in the real-time applications.

Acknowledgments This work was supported by the National Science Council (NSC), Taiwan under No. 102-2221-E-011-139-.

References

1. Wei, Y.-L., & Lin, C.-H. (2013). Single-shot person re-identification by Gaussian mixture model of weighted color histograms. In intelligent signal processing and communication systems (ISPACS).
2. Du, Y., Ai, H., Lao, S. (2012). Evaluation of color spaces for person re-identification. In International Conference on pattern recognition (ICPR), pp. 1371–1374.
3. Gray, D., & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensamble of localized features. In European Conference on computer vision (ECCV), pp. 262-275.
4. Bak, S., Corvee, E., Bremond, F., Thonnat, M. (2010). Person re- identification using Haar-based and DCD-based signature. In Workshop on Activity Monitoring by Multi-Camera Surveillance Systems.
5. Prosser, B., Zheng, W.-S., Gong, S., Xiang, T. (2010) Person re-identification by support vector ranking. In British Machine Vision Conference (BMVC).
6. Hirzer, M., Beleznai, C., Roth, P., Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. Image Analysis.
7. Schwartz, W., & Davis, L. (2009). Learning discriminative appearance-based models using partial least squares. In XXII Brazilian symposium on computer graphics and image processing (SIBGRAPI).
8. Leng, Q., Hu, R., Liang, C., & Wang, Y. (2013). Person re-identification based on contextual characteristic. *Electronics Letters, 49*, 1074–1076.
9. Tao, D., Jin, L., Wang, Y., Yuan, Y., & Li, X. (2013). Person re-identification by regularized smoothing KISS metric learning. *IEEE Circuits and Systems for Video Technology, 23*, 1675–1685.
10. Hotelling, H. (1993). Analysis of a complex of statistical variables into principal components. *Journal of Education & Psychology, 24*(7), 417–441.
11. R. Zhao, W. Ouyang, and X. Wang (2013) Person re-identification by salience matching. In International Conference on Computer Vision (ICCV).
12. T. Joachims, T. Finley, and C.-N. J. Yu (2009). Cutting-plane training of structural svms. *Machine Learning, 77*(1), 27–59.
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In computer vision and pattern recognition (CVPR), pp. 2360-2367.
14. Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B. (2009). Stel component analysis: Modeling spatial correlations in image class structure. In computer vision and pattern recognition (CVPR), pp. 2044-2051.
15. Chang, Y.-C., Chiang, C.-K., Lai, S.-H. (2012). Single-shot person re-identification based on improved random-walk pedestrian segmentation. In intelligent signal processing and communication systems (ISPACS).
16. Zhao, R., Ouyang, W., Wang, X. (2013) Unsupervised salience learning for person re-identification. In Computer Vision and Pattern Recognition (CVPR).
17. Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision, 60*, 91–110 Springer.
18. Ma, B., Su, Y., Jurie, F. (2012). Local descriptors encoded by fisher vectors for person re-identification. In European Conference on computer vision (ECCV).
19. Ma, B., Su, Y., Jurie, F. (2012). Bicov: A novel image representation for person re-identification and face verification. British Machive Vision Conference, Sep 2012, Guildford, United Kingdom, p. 11.
20. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P. (2007). Shape and appearance context modeling. In International Conference on computer vision (ICCV).
21. Gheissari, N., Sebastian, T., Hartley, R. (2006). Person re-identification using spatiotemporal appearance. In Computer Vision and Pattern Recognition (CVPR).
22. Bak, S., Corvee, E., Bremond, F., Thonnat, M. (2010). Person re-identification using spatial covariance regions of human body parts. In Advanced Video and Signal-Based Surveillance.
23. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V. (2011). Custom pictorial structures for re-identification. In British Machine Vision Conference (BMVC).
24. Zheng, W-S., Gong, S., Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In Computer Vision and Pattern Recognition (CVPR).
25. Gray, D., Brennan, S., Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In 10th IEEE International workshop on performance evaluation of tracking and surveillance (PETS), 09/2007.
26. Ma, K., & Ben-Arie, J. (2012). Vector array based multi-view face detection with compound exemplars. In computer vision and pattern recognition (CVPR).
27. Liu, C., Yuen, J., Torralba, A. (2011). Sift flow: Dense correspondence across scenes and its applications. In Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
28. Byers, S., & Raftery, A. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association, 93*(442), 577–584.
29. Roth, P. M., Hirzer, M., Köstinger, M., Beleznai, C., Bischof, H. (2014). Mahalanobis distance learning for person re-identification. In Person Re-Identification, Advances in Computer Vision and Pattern Recognition, pp 247–267.
30. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., Bischof, H. (2012). Large scale metric learning from equivalence constraints. In Computer Vision and Pattern Recognition (CVPR).
31. Hirzer, M., Roth, P. M., Bischof, H. (2012). Person re-identification by efficient metric learning. In Advanced Video and Signal-Based Surveillance.
32. Weinberger, K. Q., & Saul, L. K. (2008). Fast solvers and efficient implementations for distance metric learning. In Machine Learning.
33. Dikmen, M., Akbas, E., Huang, T. S., Ahuja, N. (2010). Pedestrian recognition with a learned metric. In Asian Conference on computer vision (ACCV).

34. Davis, J. V., Kulis, B., Jain, P., Sra, S., Dhillon, I. S. (2007) Information-theoretic metric learning. In International Conference on Machine learning (ICML), pp 209-216.

35. R. Layne, T.M. Hospedales and S. Gong (2014). Attributes-based Re-Identification. In S. Gong, M. Cristani, S. Yan, C. C. Loy (Ed.), Book: Person Re-Identification. Chapter: 4. (pp. 93–117). Berlin, Springer.

36. P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, (2014) Mahalanobis distance learning for person re-identification. In S. Gong, M. Cristani, S. Yan, C. C. Loy (Ed.), Book: Person Re-Identification. Chapter: 12. (pp. 247–267). Berlin, Springer.

Yu-Lun Wei received the M.S. degree in electronic and computer engineering from National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include ubiquitous camera framework and image processing.

Chang Hong Lin received the B.S. and M.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, and the M.A. and Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 1997, 1999, 2003, and 2007, respectively. He is an Associate Professor in the department of electronic and computer engineering of National Taiwan University of Science and Technology. His research interests include ubiquitous camera framework, code compression for embedded system, and hardware/software co-synthesis. He is the corresponding author of this article.