HUMAN RE-IDENTIFICATION THROUGH DISTANCE METRIC LEARNING BASED ON JENSEN-SHANNON KERNEL

Yoshihisa Ijiri[1], Shihong Lao[2], Tony X. Han[3] and Hiroshi Murase[4]
[1]*Corporate R&D, OMRON Corp., Kizugawa, Kyoto, Japan*
[2]*OMRON Social Solutions Co. Ltd., Kizugawa, Kyoto, Japan*
[3]*Electrical & Computer Engineering Dept., Univ. of Missouri, Columbia, MO, U.S.A.*
[4]*Graduate School of Information Science, Nagoya Univ., Chigusaku, Nagoya, Japan*
joyport@ari.ncl.omron.co.jp

Keywords: Human Re-identification, Distance Metric Learning, Jensen-Shannon Kernel.

Abstract: Human re-identification, i. e., human identification across cameras without an overlapping view, has important applications in video surveillance. The problem is very challenging due to color and illumination variations among cameras as well as the pose variations of people. Assuming that the color of human clothing does not change quickly, previous work relied on color histogram matching of clothing. However, naive color histogram matching across camera network is not robust enough for human re-identification. Therefore, we learned an optimal distance metric between color histograms using a training dataset. The Jensen-Shannon kernel is proposed to learn nonlinear distance metrics. The effectiveness of the proposed method is validated by experimental results.

1 INTRODUCTION

With the continuous drop of hardware costs, the number of surveillance cameras deployed have been growing drastically, leaving the available human analysts far behind. To fill this gap, many automatic video surveillance techniques and systems have been proposed. Among them, human identification, which is one major problem in video surveillance, enables us to match human tracks in surveillance areas such as stores and shopping malls for the purposes of security or marketing. For law enforcement, with one or more images of suspects, automatic human identification speeds up the process of finding them from a large amount of surveillance camera records. Considering the cost effectiveness and the psychological effects on the citizenry, it is impractical to fully cover the entire surveillance area without any blind spot using the camera network. Therefore, in general, human identification needs to be done across cameras with non-overlapping views. In this paper, this human re-identification problem is studied.

It is natural to tackle the human re-identification through the face recognition approach, which has been studied extensively in the computer vision community. However, as shown in Fig. 1, people tend to vary their poses a lot unless they are asked not to do so. Surveillance cameras are also intended to watch a large area. Hence it is unrealistic to assume that clear human faces can always be viewed. Therefore, for practical surveillance, face recognition alone is insufficient to guarantee the human identification accuracy. As a approach complementary to the face recognition approach, human identification based on clothing colors is studied in this paper. Taking most of the surveillance scenarios into consideration, we assume that people do not change their clothing in a short time duration, Noticing the symmetry characteristic of human clothing, we conclude that the colors of most clothing are view-angle insensitive.

Human re-identification is a very challenging problem for the following reasons. Among surveillance cameras, color calibrations are not always the

Figure 1: Sample images from VIPeR dataset: (a)(b) are from ID 302, (c)(d) from ID 188, and (e)(f) from ID 358.

same. Also, the illuminations of different camera views usually varies a lot. In addition, the pose variations of people make it infeasible to expect identical appearances among cameras. Many works have addressed these difficulties and can be summarized into the following three groups: (i) appearance-based methods, (ii) color calibration or color transfer function estimation between cameras, and (iii) inter-camera relationship modeling.

For appearance-based methods, various features have been proposed to represent clothing colors and textures (Bird et al., 2005; Gheissari et al., 2006; Wang et al., 2007; Gray and Tao, 2008; Lin and Davis, 2008; Hamdoun et al., 2008; Schwartz and Davis, 2009; Kuo et al., 2010; Bak et al., 2010; Alahi et al., 2010; Berdugo et al., 2010; Bazzani et al., 2010; Farenzena et al., 2010; Hirzer et al., 2011). Among them, Farenzena et al. (Farenzena et al., 2010) combined weighted HSV histograms, Maximally Stable Color Regions (MSER), and Recurrent High-Structured Patches. The former two features are used to represent clothing colors, and the last one is for texture description. In the feature extraction process, according to the symmetric and asymmetric axes, the human body is divided into sub-regions to deal with pose variations using an algorithm called the Symmetric Driven Accumulation of Local Features (SDALF). More recently, Bak et al. (Slawormir et al., 2011) proposed to use mean Riemannian covariance, which consists of covariance matrices of RGB colors, corresponding gradient magnitudes, and orientations from multiple shots.

These state-of-the-art appearance-based algorithms achieve high accuracy through extraction of multiple features. More specifically, color histograms in various color spaces and local descriptors are combined. These redundant feature extractions contribute higher accuracy at the expense of computational efficiency. Among these features, clothing color-based features are exploited extensively, because they are robust to pose variation. For the color histogram matching, traditional distances are used in various color spaces. However, since the color calibrations from different cameras are different, matching directly with traditional distances may lead to bias. So the estimation of color relationships between cameras, *i.e.* the algorithm group (ii) mentioned above, has been studied in (Javed et al., 2008; Prosser et al., 2008b; Prosser et al., 2008a; Gilbert and Bowden, 2006). Among these algorithms, the brightness transfer function (BTF) method achieves quite good performance. One disadvantage of BTF is that the BTF between each pair of cameras has to be estimated. If we have N cameras, then $N(N-1)/2$ BTFs must be estimated. The computational complexity of the algorithm w.r.t number of cameras is $O(N^2)$, which is not practical for surveillance systems with many cameras.

For camera networks, if people walk away from one camera to another, and the entrance/exit times of each camera can be modeled statistically. For instance, a person exiting from a camera usually appears in another camera within a certain elapsed time. Based on this concept, inter-camera relationship modeling methods, categorized as algorithm group (iii), have been proposed in (Javed et al., 2008; Huang and Russell, 1997; Pasula et al., 1999; Song and Roy-Chowdhury, 2007). The disadvantage of this scheme is that such methods assume correct correspondence between people and people walking in almost the same elapsed time between cameras. Such assumptions may not be true in many practical situations. It also requires people correspondences for all possible pairs of cameras. Thus these algorithms also have the computational complexity of $O(N^2)$, where N is the camera number.

Considering the facts listed above, we believe that the clothing color-based human re-identification is one of the most promising approaches. Thus improving the color matching for this approach is very important. For color matching problems, however, the importance of distance metrics has not been emphasized enough: only simple distance metrics have been exploited. Therefore in this paper, we propose a method for learning optimal distance metrics between color histograms from different cameras. The proposed method reduce the complexity of the BTF from $O(N^2)$ to $O(1)$, since it does not assume that the camera configuration is known. It is robust to differences of color calibration between cameras as well. To obtain better accuracy, we apply nonlinear kernel functions to learn a nonlinear distance metric. We experimentally validated the approach based on the combination of large margin component analysis (LMCA) and the Jensen-Shannon kernel. The proposed method improves the identification accuracy for clothing color matching approach. Therefore, the proposed method and many previous works are complementary.

In summary the contribution of this paper is two-fold: (a) the use of nonlinear distance metric learning is proposed to achieve better accuracy compared with conventional simple distance metrics, and (b) the combination of LMCA and the Jensen-Shannon kernel function is proposed. The basic idea of the proposed method is illustrated in Fig. 2(b) and compared with the conventional methods shown in Fig. 2(a). In contrast to the conventional methods that match

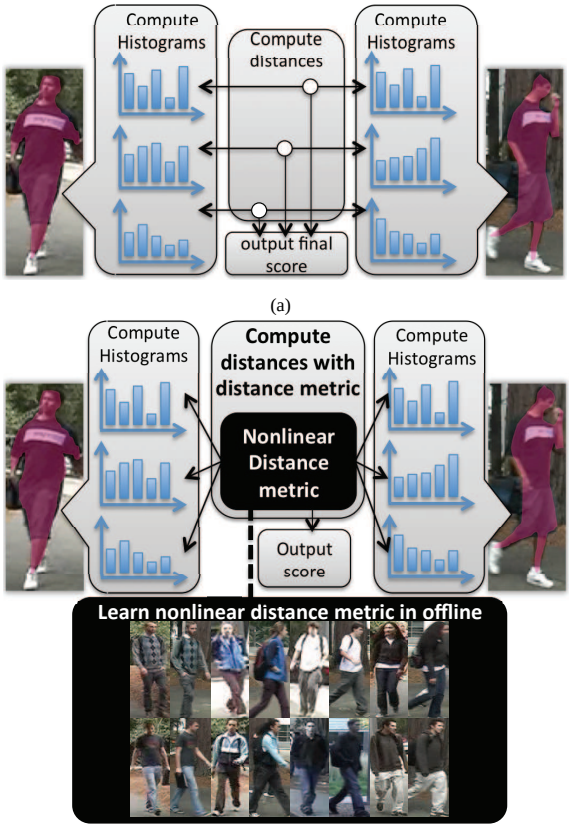

Figure 2: (a) conventional and (b) proposed color matching schemes.

clothing color histograms in a simple distance metric, the proposed framework computes the distance with a learned optimal distance metric.

The remainder of this paper is organized as follows. In Section 2, the proposed method is described in depth with a review of some basic algorithms. In Section 3 the experimental results are presented to validate the proposed approach. Finally in Section 4, concluding remarks are given.

2 PROPOSED METHOD

In this paper, we assume that the human region in a camera view is given by a human detector or background subtraction algorithm. The training datasets from many surveillance cameras under different conditions and corresponding subject labels are assumed available for metric learning. As mentioned above, different cameras under different conditions cause different color calibration. Hence the direct use of histograms affects the identification accuracy adversely. In contrast, the optimal distance metric is learned with a training dataset in the proposed work.

To learn the optimal metric, color histograms $X = \{x_i; i = 1,\ldots,n\}$ are firstly computed for the training dataset, where n is the total number of training samples in the dataset. Then using histograms X and corresponding labels Y, the optimal distance metric is learned. By including most of the possible variations in practical situations in the training dataset, we expect this approach to work in real applications. During the registration process, every time people enter the fields of views of cameras, they are registered. For each person, color histograms m_c are extracted as models. Denoting the number of people registered as C, $M = \{m_c; c = 1,\ldots,C\}$ are obtained. In the re-identification process, color histograms of test images, $T = \{t_k; k = 1,\ldots,K\}$ are obtained, where K is the number of images to be matched. Histograms M and T are matched using the learned distance metric.

In the proposed method, the large margin criterion is used to learn the optimal distance metric. Nonlinear projection $\phi(x_i)$ is used to project the input histograms onto a higher dimensional space. For nonlinear projection, several types of kernel functions are investigated; the final results are shown in experimental results section.

2.1 Clothing Color Histograms

Traditionally, color histograms are extracted in the color spaces including RGB HSV, Lab, and YCbCr. We use the HSV joint histograms in the HSV color space because they showed better accuracy than other color spaces according to our experiments. Since clothing colors generally do not vary drastically among front, back, right, and left views, for viewpoint invariant matching, vertical combination of such clothing colors as the upper and lower body colors are robust in many cases. Based on this observation, Bird et al. (Bird et al., 2005) divided the entire human region into several vertically segmented regions to obtain regional histograms. Following this scheme, in this paper, the human region is segmented vertically into P pieces, and for each sub-region $p(p = 1,\ldots,P)$, HSV joint histograms $x_{ip} \in \mathcal{R}^{b_h \times b_s \times b_v}$ are computed, where b_h, b_s, b_v are the number of bins in the H, S and V color channels and x_{ip} is vectorized from the 2D joint histogram. Then to describe all the human region features, these histograms in each region are concatenated as $x_i = \{x_{i1},\ldots,x_{iP}\} \in \mathcal{R}^D$ and normalized so that $\sum_i |x_i| = 1$, where $D = b_h \times b_s \times b_v \times P$. We apply this procedure for training, model, and test images to obtain color histograms X, M, and T, respectively.

2.2 Distance Metric Learning

Many supervised/unsupervised distance metric learning algorithms have been proposed (Yang, 2006). We adopt a supervised learning algorithm in this work. Among supervised algorithms, linear discriminant analysis (LDA) is a major algorithm. However, LDA has some limitations. For example, it cannot be applied when there are not enough data to estimate intra-class scatter or insufficient classes to make the between class scatter matrix be non-singular. On the other hand, not only the linear distance metric but also the nonlinear distance metric using kernel functions have been proposed to improve accuracy. Among them, the support vector machine (SVM) has gained much popularity due to its good performance. However, since it was originally formulated as a binary classification problem, it cannot be applied directly to multi-class problems. To extend the binary classification to the multi-class case, we can discriminate matching scores that share the same labels and those between different labeled data. In the practice of this scheme, the number of differently labeled pairs often becomes extremely larger than that of the pairs sharing the same labels, causing the unbalanced training problem. To avoid this, sampling is often adopted from a large number of differently labeled pairs. However, there is no guarantee that appropriate data can always be sampled, which may cause over-fitting. For these problems, Prosser et al.(Prosser et al., 2010) proposed ensemble rankSVM to solve this problem as ranking problem. Although it shows good performance, it is noted that the tuning of this algorithm is computationally expensive(Zheng et al., 2011). To improve the ensemble RankSVM further, recently Zheng et al.(Zheng et al., 2011) proposed "Probabilistic Relative Distance Comparison" (PRDC), which is one of variants of distance metric learning. This algorithm is promising, however, training process is still computationally expensive. On the other hand, while a number of distance metric learning have been proposed, the large margin nearest neighbor (LMNN) (Weinberger and Saul, 2009) is viewed as one of the best methods (Kulis, 2010) in terms of accuracy. To further improve LMNN and alleviate the difficulties in processing high dimensional data, Torresani et al. proposed large margin component analysis (LMCA) (Torresani and Lee, 2007). Furthermore, the algorithm was kernelized to learn the distance metric in higher dimensional and nonlinear space for better performance. In this paper we use it to learn the large margin distance metric. In the following description, we briefly review the algorithm and describe how to apply it to the re-identification problem.

Linear LMCA:
Given color histogram features $X = \{x_i; i = 1,\ldots,n\} \in \mathcal{R}^{D\times n}$ and corresponding labels $Y = \{y_i; i = 1,\ldots,n\} \in \{0,1\}^n$, LMCA minimizes the following loss function $\varepsilon(L)$:

$$\varepsilon(L) = \sum_{ij}\eta_{ij}||L(x_i - x_j)||^2 + c\sum_{ijl}\eta_{ij}(1-y_{il}) \cdot h(||L(x_i - x_j)||^2 - ||L(x_i - x_l)||^2 + 1), \quad (1)$$

where $L \in \mathcal{R}^{d\times D}$ is a linear projection for X, $\eta_{ij} \in \{0,1\}$ takes 1 iff x_j and x_i shares the same label ($y_i = y_j$) and x_j is the k-nearest neighbor of x_i, $c > 0$ is an appropriate balancing parameter, $y_{il} \in \{0,1\}$ is a variable that takes 1 iff $y_i = y_l$, and $h(s)$ is a hinge function that is defined as $h(s) = \max(s,0)$. The first term minimizes projected distances between the data pairs that share the same labels to encourage the invariance property, and the other term is for discrimination, which makes projected distances between the data pairs with the same labels and those with different labels distant with unit distance 1. The hinge function gives a loss for the invasive data and does not affect those data that have enough margins. Optimizing L by gradient descent based on this objective function, optimal linear discriminative projection can be obtained. Here the distance metric can be given by L^TL:

Kernel LMCA:
For nonlinear projection $\phi(x_i)$, the inner product is expressed as $k(x_i,x_j) = \phi_i^T\phi_j$. In subsequent description, the following notation is used: $\phi_i = \phi(x_i)$C$\Phi = [\phi_1,\ldots,\phi_n]^TCk_i = \Phi\phi_i = [k(x_1,x_i),\ldots,k(x_n,x_i)]$. Using nonlinear projection, the loss function in the projected space is expressed as:

$$\varepsilon(L) = \sum_{ij}\eta_{ij}||L(\phi_i - \phi_j)||^2 + c\sum_{ijl}\eta_{ij}(1-y_{il}) \cdot h(||L(\phi_i - \phi_j)||^2 - ||L(\phi_i - \phi_l)||^2 + 1). \quad (2)$$

The gradient of the loss function becomes

$$\frac{\partial\varepsilon(L)}{\partial L} = \sum_{ij}\eta_{ij}L(\phi_i - \phi_j)(\phi_i - \phi_j)^T + c\sum_{ijl}\eta_{ij}(1-y_{il}) \cdot h'(s_{ijl})L[(\phi_i - \phi_j)(\phi_i - \phi_j)^T - (\phi_i - \phi_l)(\phi_i - \phi_l)^T], \quad (3)$$

where

$$s_{ijl} = ||L(\phi_i - \phi_j)||^2 - ||L(\phi_i - \phi_l)||^2 + 1. \quad (4)$$

Considering the parameterizations of L as $L = \Omega\Phi$, Eq. (3) yields

$$\frac{\partial \varepsilon(L)}{\partial L} = 2\Omega \sum_{ij} \eta_{ij} [E_i^{(k_i-k_j)} - E_j^{(k_i-k_j)}]\Phi + 2c\Omega \sum_{ijl} \eta_{ij}(1-y_{il})h'(s_{ijl}) \cdot [E_i^{(k_i-k_j)} - E_j^{(k_i-k_j)} - E_i^{(k_i-k_l)} + E_l^{(k_i-k_l)}]\Phi, \quad (5)$$

where

$$s_{ijl} = \{||\Omega(k_i-k_j)||^2 - ||\Omega(k_i-k_l)||^2 + 1\}, \quad (6)$$

$E_i^v = [0,\ldots,0,v,0,\ldots,0]$ is a $n \times n$ matrix that only takes value v at the i-th column, and $h'(s)$ is a differential of $h(s)$. To avoid the discontinuity of $h'(s)$ around 0, approximation by a smooth hinge function was proposed(Rennie and Srebro, 2005). Using Eq. (5), the steepest descent update rule is given by

$$L \leftarrow L - \lambda \frac{\partial \varepsilon(L)}{\partial L}. \quad (7)$$

However in Eq. (5), all the terms in $\frac{\partial \varepsilon(L)}{\partial L}$ have Φ and $L = \Phi\Omega$ by assumption, and Φ in Eq. (7) can be removed. Hence the update rule is reduced to the following simple gradient descent update of Ω:

$$\Omega \leftarrow \Omega - \lambda\Gamma, \quad (8)$$

$$\Gamma = 2\Omega \sum_{ij} \eta_{ij} [E_i^{(k_i-k_j)} - E_j^{(k_i-k_j)}] + 2c\Omega \sum_{ijl} \eta_{ij}(1-y_{il})h'(s_{ijl}) \cdot [E_i^{(k_i-k_j)} - E_j^{(k_i-k_j)} - E_i^{(k_i-k_l)} + E_l^{(k_i-k_l)}]. \quad (9)$$

By assumption, projection onto higher dimensional space can be obtained easily:

$$L\phi_q = \Omega\Phi\phi_q = \Omega k_q. \quad (10)$$

Here the distance metric can be represented as $k_q^T \Omega^T \Omega k_q$.

Since this algorithm automatically selects data that fall within the margin through the learning process, we expect better generalization compared to the methods that sample fixed pairs in advance, which is a simple extension of the binary classification problem to the multi-class problem.

2.3 Kernel Functions

In this section, kernel functions that are suitable for matching two distributions $a, b \in \mathcal{R}[0,1]^D$ are investigated, such as normalized histograms or probability distributions.

Histogram intersection kernel:
One popular distance between histograms is the histogram intersection. The histogram intersection kernel, which satisfies Mercer's condition, was proposed in (Odone et al., 2005; Grauman and Darrell, 2005):

$$k(a,b) = \sum_i \min(a_i, b_i). \quad (11)$$

χ^2 kernel:
Another popular distance between histograms is the χ^2 distance. Here, the χ^2 kernel function, satisfies Mercer's condition as well (Zhang et al., 2006; Fowlkes et al., 2004). It is defined as:

$$k(a,b) = \exp\left\{ -\frac{\left(1 - \sum_i \frac{(a_i-b_i)^2}{\frac{1}{2}(a_i+b_i)}\right)}{\sigma^2} \right\}. \quad (12)$$

Bhattacharyya kernel:
Hellinger distance $\frac{1}{2}\sum_i \left(\sqrt{a_i} - \sqrt{b_i}\right)^2$ is effective between two probability distributions. Here a and b are normalized histograms such that $\sum_i a_i = 1$, $\sum_i b_i = 1$, and thus the distance yields $1 - \sum_i \sqrt{a_i b_i}$, which is called the Bhattacharyya distance. Based on the Bhattacharyya distance, the following Bhattachayya kernel is derived that satisfies Mercer's theorem (Jebara and Kondor, 2003):

$$k(a,b) = \sum_i \sqrt{a_i b_i}. \quad (13)$$

Jeffrey divergence kernel:
In human re-identification based on clothing color histogram matching, one key point is that the robust matching between histograms is affected heavily by noises. Due to the big appearance changes, the shape of the color histograms are often altered, which complicates robust matching. Thus we cannot rely on histogram shape matching, while Gauss kernel or other correlation-based kernels strongly rely on histogram shape. For comparing and matching such uncertain information, one popular way is an information-based measure. For this problem, the Kullback-Leibler divergence (KLD) $H(a||b)$ is an effective measure of the differences between two probabilistic pieces of information. However, due to its asymmetry($H(a||b) \neq H(b||a)$) the KLD cannot be used directly as a kernel function. Symmetric Jeffrey divergence (JD) $H(a||b) + H(b||a)$(Jeffreys, 1946) is used to define the Jeffrey divergence kernel function as $\exp(-\mathrm{JD}/\sigma^2)$. The definition is given by

$$k(a,b) = \exp\left\{ -\frac{\sum_i \left(a_i \log\frac{a_i}{b_i} + b_i \log\frac{b_i}{a_i}\right)}{\sigma^2} \right\} = \prod_i \left\{ \left(\frac{b_i}{a_i}\right)^{\frac{a_i}{\sigma^2}} \left(\frac{a_i}{b_i}\right)^{\frac{b_i}{\sigma^2}} \right\}, \quad (14)$$

where σ is another parameter that provides tuning flexibility. We don't have proof for the positive definiteness of this kernel function yet, even though experiments show that it is always positive definite.

Jensen-Shannon kernel:
Another way to circumvent KLD's asymmetry is the Jensen-Shannon divergence (JSD)(Lin, 1991). Not only it is symmetric but it also has other numerous desirable properties. As shown in (Huang et al., 2005), it is numerically more stable than KLD or JD and provides a theoretical upper bound in terms of the variational distance; no general upper bound exists for KLD or JD. Therefore the JSD-based kernel function, which satisfies Mercer's theorem, is proposed (Chan et al., 2004). To give one more parameter σ for flexibility, an exponential function is used as $\exp(-\mathrm{JSD}/\sigma^2)$. The resulting kernel function is

$$k(a,b) = \exp\left\{-\frac{\sum_i \left(\frac{a_i}{2}\log\frac{2a_i}{a_i+b_i} + \frac{b_i}{2}\log\frac{2b_i}{a_i+b_i}\right)}{\sigma^2}\right\}$$
$$= \prod_i \left\{\left(\frac{a_i+b_i}{2a_i}\right)^{\frac{a_i}{2\sigma^2}} \left(\frac{a_i+b_i}{2b_i}\right)^{\frac{b_i}{2\sigma^2}}\right\}. \quad (15)$$

This kernel is positive definite (Chan et al., 2004). Even though a variety of kernel functions exists, this kernel has not received enough attention yet.

To investigate the effectiveness of this kernel for our problem, a preliminary experiment using synthetic data generated using Gaussian mixture model is conducted. As shown in experimental section, we use HSV joint histograms in the HSV color space. In HSV joint histograms, hue, saturation and value components can be easily shifted circularly by illumination changes or color calibration differences. To simulate data pairs from the same person, a pair of circularly shifted and original distributions are used. For data pairs from different persons, mutually different distributions are used. An example of data pair is illustrated in Fig.3. 500 samples of data pair is used to draw a distribution of distances by each kernel. The result can be shown in Fig.4. The distances between data pairs with circular shift should be small, while those between data pair with different distributions should be large. To quantify the effectiveness, the ratio of mean distance for simulated intra-person pairs against that for simulated different individual pairs are computed. The results can be seen in Table.1, where the mean distance for simulated intra-person pairs is denoted by m_{shift} and that for simulated different individual pairs m_{diff}. The larger the ratio is, the better. Viewing these results, we can see that the proposed Jensen-Shannon kernel function is the most appropriate for our HSV color space based matching problem.

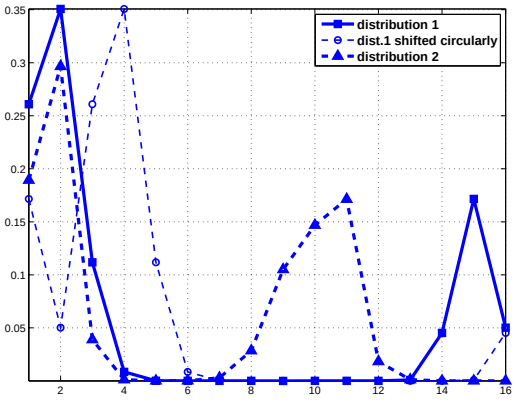

Figure 3: Synthetic test distributions; the bold line and the regular dotted line indicate a pair of color histograms of a person from different cameras, which simulates circularly shifted hue histogram in HSV color space, and the bold line and the bold dotted line indicate a pair of color histograms from different person.

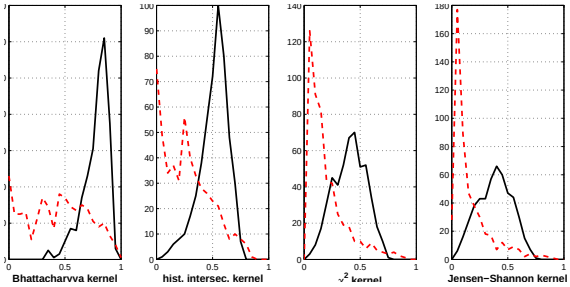

Figure 4: Distance distribution in each kernel function; dotted line indicates a distribution of distances between data pairs with different color distributions, the other one indicates a distribution of those between shifted and original color distributions.

Table 1: Ratios of mean distance between shifted and original data pairs against that between different data pairs.

	Bhat.	HI	χ^2	JS
$\frac{m_{shift}}{m_{diff.}}$	1.8059	1.9113	2.0271	2.3351

2.4 Matching

Normalized color histograms are projected onto optimized nonlinear space, as shown by Eq. (10), which yields $\Omega k(X,M)$, $\Omega k(X,T)$, where Ω is learned by Eq. (8), and X, M, and T are color histograms computed from training dataset, model images, and test images, respectively. Here matching can be done as:

$$s_{ck} = f(\Omega k(X,m_c), \Omega k(X,t_k)), \quad (16)$$

where $f(\cdot,\cdot)$ is an arbitrary similarity function. Estimated identity $\hat{\omega}_k$ for input data t_k is given by taking

Figure 5: Subregion division.

(a) Correct match: ID 126

(b) Correct match: ID 351

(c) False match: ID 006 & 135

(d) False match: ID 078 & 107

Figure 6: Successful and failure cases by the proposed method.

the maximum for similarity scores s with respect to registered models $M = \{m_c; c = 1,\ldots,C\}$:

$$\hat{\omega}_k = \max_c(s_{ck}). \tag{17}$$

Although any similarity function can be used as f in Eq. (16), in the following experiments, we used correlation.

3 EXPERIMENTS

We conducted experiments using the Viewpoint Invariant PEdestrian Recognition (VIPeR) dataset by Gray et al. (Gray and Tao, 2008) to show the effectiveness of the proposed method. This dataset is constructed to evaluate viewpoint invariant human re-identification algorithms. Several examples from VIPeR can be seen in Fig. 1. Due to the pose and illumination variations, the clothing colors are significantly different between the two images of the same person. Furthermore, in the dataset, the camera labels are not available to indicate from which camera the images were taken. Since estimating such color calibration functions as BTF is impossible, this dataset is quite difficult for human re-identification.

In this dataset, two images for each of 632 subjects are included: one for the model and one for matching. The training dataset is constructed by randomly selecting 200 individuals from the 632 people. Since there are two images for each subject, the training dataset has 400 images. The remaining 432 individuals are used for evaluation. In all experiments, through ten cross validations, we estimated the mean accuracy and the corresponding standard deviations. In the human body extraction, if an automatic segmentation algorithm was employed, the result would be affected by its accuracy. To avoid this and evaluate only the effect by difference of distance metrics, the human body region is segmented manually. In practical situations, this can be done relatively easily by background subtraction, for example.

Table 2: CMC in comparison.

	CMC (1)	CMC (10)
Baseline (Euclidean dist.)	8.7±0.8	24.2±1.0
Baseline (NCC)	10.8±1.0	28.6±0.8
Baseline (hist. intersec.)	13.2±1.1	35.3±1.0
Baseline (Bhat. dist.)	17.2±0.8	39.2±1.0
LMCA with lin.	13.0±1.0	37.3±2.4
LMCA with Gauss(5.0)	15.0±1.2	43.8±1.9
LMCA with hist. intersec.	17.1±2.4	51.1±2.3
LMCA with χ^2	19.6±1.5	53.6±1.5
LMCA with Bhat.	18.8±1.3	50.9±1.5
LMCA with JD	16.6±0.1	53.2±1.3
LMCA with JS	**20.5**±1.5	**55.7**±1.5

For feature extraction, we used HSV joint histogram in the HSV color space. We used five bins for HS channels and three bins for V channel for quantization. For division into sub-regions, the best accuracy was given by segmentation into eight regions, as illustrated in Fig. 5. The blue lines represent the boundaries of each region and color histograms are computed from the red areas in each sub-region. Concatenating these histograms in each image, we obtained $5 \times 5 \times 3 \times 8 = 600$ dimensional features.

We used a cumulative match characteristic (CMC) curve to evaluate the matching performance. CMC (n) represents the probability that correct matches appear in the top n-th. CMC (1) and CMC (10), which represent the probability that correct matches are always placed first and within top 10 respectively, are shown in Table 2. In the table, each cell includes estimated mean value $\pm$ corresponding standard deviation. For comparisons, such basic techniques as simple Euclidean distance, normalized correlation, histogram intersection, and Bhattacharyya distance are listed as baselines. In the LMCA evaluation, in addition to linear LMCA, we tested various kernel LMCAs, including Gauss, histogram intersection, χ^2, Bhattacharyya, Jeffrey divergence, and Jensen-Shannon kernels. The combination of LMCA and the Jensen-Shannon kernel gave the best result. On the other hand, for the

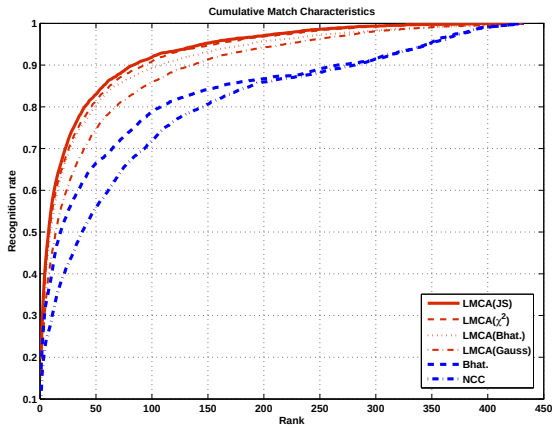

Figure 7: CMC curve.

Gauss kernel, the accuracy was degraded compared to the simple distances, which means it is unsuitable for matching normalized histograms. Moreover, the result from LMCA plus the Jensen-Shannon kernel is better than that from the PRDC(Zheng et al., 2011), which can achieve 12.64% for CMC(1) and 44.28% in CMC(10) in equivalent data setup described in their paper. Considering in (Zheng et al., 2011) 29 different feature channels such as Schmid and Gabor filter for RGB, YCbCr, HSV color spaces are used, the proposed method which uses only HSV joint histograms is much more efficient and probably faster in re-identification process.

The summarizing CMC curve is shown in Fig. 7. The proposed method (LMCA with the Jensen-Shannon kernel) is the best for all ranks. Some successful and failure examples by the proposed method are shown in Figs. 6(a)-(d). Although the appearances are quite different, for pairs in (a) and (b) from the same ID, the proposed method effectively absorbed the differences. However, pairs in (c) and (d) from different IDs are falsely ranked first by the proposed method. For the images in the pair (c), roughly speaking, the only difference is the facial skin color, and most of body region is the same color. This kind of cases is still critical in the proposed method. The other failure case in the pair (d), the only difference is the upper body color. In addition, the upper body region is including the same color. Presumably it is due to the limitation of using histogram without spatial information. Some extent of spatial ambiguity is required for viewpoint invariance, so it is a trade-off problem. Although seeking better features is outside the scope of this paper, more representative features would give better accuracy using the proposed method for these cases.

Finally we experimented with SDALF(Farenzena et al., 2010), which is one of the current state-of-the-art algorithm. The results are shown in Table 3. All the experimental setups are the same as mentioned above. The first three lines show the accuracy when only the color features (MSCR, weighted HSV histograms, and their combination) are exploited. Since the proposed algorithm can be seen as an alternative to weighted histogram matching, we used our proposed method with MSCR. The results are in the last line. The SDALF performance drastically improved by only replacing the histogram matching part from the original weighted histogram matching to the proposed nonlinear distance metric learning-based histogram matching.

Table 3: Combination with state-of-the-art algorithm.

	CMC (1)	CMC (10)
(a) SDALF (wHSVhist)	9.8±0.5	28.2±1.7
(b) SDALF (MSCR)	7.8±0.7	23.2±0.8
(a) + (b)	14.7±1.1	41.1±1.6
LMCA(JS) + (b)	**21.3**±1.2	**57.7**±2.6

The result shows the top level accuracy on VIPeR with the advantage of only using color features and so fast processing. This suggests the effectiveness of the proposed method.

The kernel LMCA learning took about 80 [sec] with 400 training datasets[1]. Thus given a sufficient number of training datasets, the algorithm can learn optimal distance metrics relatively quickly. Compared to the color calibration between cameras (such as BTF) or inter-camera relationship modeling that requires optimization for each pair of cameras, in the proposed method, optimization can be done just once at setup. Thus while BTF has computational complexity of $O(N^2)$ w.r.t number of cameras N, that of the proposed method is just $O(1)$. Therefore the proposed method is especially easy to implement when a large number of cameras are employed.

4 CONCLUSIONS

In this paper, we study the human re-identification problem based on clothing colors. The re-identification approach based on distance metric learning is validated experimentally. The combination of LMCA and the Jensen-Shannon kernel provided the best accuracy in our experiments.

The proposed method does not assume that the camera configuration is known. Our approach is different from the brightness transfer function estimation

[1]The other experimental environments included Matlab unoptimized code on Mac OS X, Core2Duo 2.2 GHz with 2GB memory.

or inter-camera relationship modeling, where some amount of the exact human correspondences in each pair of cameras are needed. Therefore the advantage of the proposed approach lies in the computational efficiency, which becomes obvious when a large number of cameras are deployed.

REFERENCES

Alahi, A., Vandergheynst, P., Bierlaire, M., and Kunt, M. (2010). Cascade of descriptors to detect and track objects across any network of cameras. *CVIU*, 114(6):624–640.

Bak, S., Corvee, E., Brémond, F., and Thonnat, M. (2010). Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In *Proc. of AVSS*.

Bazzani, L., Cristani, M., Perina, A., Farezena, M., and Murino, V. (2010). Multiple-shot Person Re-identification by HPE signature. In *Proc. of ICPR*.

Berdugo, G., Soceanu, O., Moshe, Y., Rudoy, D., and Dvir, I. (2010). Object Reidentification in Real World Scenarios Across Multiple Non-overlapping Cameras. In *Proc. of Euro. Sig. Proc. Conf.*

Bird, N. D., Masoud, O., Papanikolopoulos, N. P., and Isaacs, A. (2005). Detection of Loitering Individuals in Public Transportation Areas. *IEEE Trans. on ITS*, 6(2):167–177.

Chan, A., Vasconcelos, N., and Moreno, P. (2004). A family of probabilistic kernels based on information divergence. *Univ. of California, San Diego, Tech. Rep.*

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In *Proc. of CVPR*.

Fowlkes, C., Belongie, S., Chung, F., and Malik, J. (2004). Spectral grouping using the Nyström method. *IEEE Trans. on PAMI*, 26(2):214–225.

Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person Reidentification Using Spatiotemporal Appearance. In *Proc. of CVPR*.

Gilbert, A. and Bowden, R. (2006). Tracking Objects Across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity. In *Proc. of ECCV*.

Grauman, K. and Darrell, T. (2005). The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. In *Proc. of ICCV*.

Gray, D. and Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Proc. of ECCV*.

Hamdoun, O., Moutarde, F., Stanciulescu, B., and Steux, B. (2008). Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proc. of ICDSC*.

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person Re-identification by Descriptive and Discriminative Classification. In *Scandinavian Conference on Image Analysis*, pages 91–102.

Huang, T. and Russell, S. (1997). Object identification in a Bayesian context. In *Proc. of Joint Conf on AI & IJCAI*.

Huang, X., Li, S. Z., and Wang, Y. (2005). Jensen-shannon boosting learning for object recognition. In *Proc. of CVPR*.

Javed, O., Shafique, K., Rasheed, Z., and Shah, M. (2008). Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *CVIU*, 109(2):146–162.

Jebara, T. and Kondor, R. (2003). Bhattacharyya and Expected Likelihood Kernels. In *Proc. of Comp. Learn. Theory*.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *A Math. and Physical Sciences*, 186(1007):453–461.

Kulis, B. (2010). ICML 2010 Tutorial on Metric Learning.

Kuo, C.-H., Huang, C., and Nevatia, R. (2010). Inter-camera Association of Multi-target Tracks by On-Line Learned Appearance Affinity Models. In *Proc. of ECCV*.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. on Info. Theory*, 37(1):145–151.

Lin, Z. and Davis, L. S. (2008). Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance. In *Proc. of ISVC*.

Odone, F., Barla, A., and Verri, A. (2005). Building kernels from binary strings for image matching. *IEEE Trans. on Image Proc.*, 14(2):169–180.

Pasula, H., Russel, S. J., Ostland, M., and Ritov, Y. (1999). Tracking many objects with many sensors. In *Proc. of IJCAI*.

Prosser, B., Gong, S., and Xiang, T. (2008a). Multi-camera Matching under Illumination Change Over Time. In *Proc. of Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*.

Prosser, B., Gong, S., and Xiang, T. (2008b). Multi-camera Matching using Bi-Directional Cumulative Brightness Transfer Functions. In *Proc. of BMVC*.

Prosser, B., Zheng, S., Gond, S., and Xiang, T. (2010). Person Re-Identification by Support Vector Ranking. In *Proc. of BMVC*.

Rennie, J. D. M. and Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *In proc. of ICML*.

Schwartz, W. R. and Davis, L. S. (2009). Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proc. of Brazil. Symp. on Comp. Graph. and Image Proc.*

Slawormir, B., Corvee, E., Brémond, F., and Thonnat, M. (2011). Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In *Proc. of AVSS*, pages 179–184.

Song, B. and Roy-Chowdhury, A. K. (2007). Stochastic Adaptive Tracking In A Camera Network. In *Proc. of ICCV*.

Torresani, L. and Lee, K.-c. (2007). Large margin component analysis. *NIPS*.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P. (2007). Shape and Appearance Context Modeling. In *Proc. of ICCV*.

Weinberger, K. Q. and Saul, L. K. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *JMLR*, 10:207–244.

Yang, L. (2006). Distance metric learning: A comprehensive survey. *Michigan State Univ. Tech. Report*.

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2006). Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *IJCV*, 73(2):213–238.

Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person Re-identification by Probabilistic Relative Distance Comparison. In *Proc. of CVPR*.