This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Spatial Pyramid-Based Statistical Features for Person Re-Identification: A Comprehensive Evaluation

Jianlou Si, Honggang Zhang, *Senior Member, IEEE*, Chun-Guang Li, *Member, IEEE*, and Jun Guo

***Abstract*—Person re-identification (Re-Id) across nonoverlapping camera views is one of challenging problems in surveillance video analysis. The difficulties in person Re-Id mainly come from the large appearance variations caused by camera view angle, human pose, illumination, and occlusion. Recently, extensive efforts have been cast into addressing this problem by developing invariant features or discriminative distance metrics. However, there is still a lack of systematic evaluations on the pipeline for feature extraction and combination. In this paper, we propose a spatial pyramid-based statistical feature extraction framework as a unified pipeline of feature extraction and combination for person Re-Id, and systematically evaluate the configuration details in feature extraction and the fusion strategies in feature combination. Extensive experiments on benchmark datasets demonstrate the critical components in feature extraction. Moreover, by combining multiple features, our proposed approach can yield state-of-the-art performance. It should be mentioned that our approach achieves rank 1 matching rate of 45.8% on dataset VIPeR and 61.5% on dataset CUHK01, respectively.**

***Index Terms*—Multiple kernel local Fisher discriminant analysis (mkLFDA), person re-identification (Re-Id), spatial pyramid-based statistical features.**

I. Introduction

NOWADAYS, visual surveillance cameras are widely deployed in airports, train stations, and other important venues. The captured surveillance videos contain important cues for the public security (e.g., [1]–[4]). Thus, automatically verifying the identity of a pedestrian from nonoverlapping surveillance camera views is increasingly becoming one of the most critical tasks in video analysis. This is the problem termed as person re-identification (Re-Id) [1].

One commonly used approach to tackle person Re-Id task is to formulate it as an image matching or verification problem, where matching each instance from the probe set (captured from one camera view) against all from the gallery set (captured from another disjoint camera view) (see Fig. 1). Unfortunately, the large variations in person appearance due to view angle, pose, illumination, and occlusion make the accurate matching quite difficult. Therefore, extensive efforts have been cast into designing cross-view invariant features or metrics, e.g., [5]–[13].

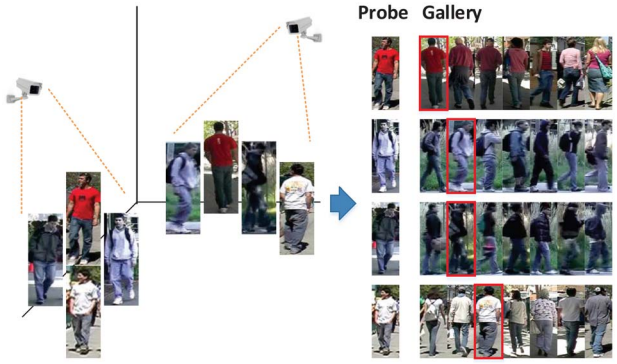


Fig. 1. Person Re-Id. (Left) Images captured from disjoint camera views. (Right) Each line contains a probe image, and the corresponding ranked gallery set, where the true match is marked in red box.

On one hand, it has been verified that the local statistical features, such as color or oriented gradients histogram [14], are effective for person Re-Id, e.g., [5]–[7] and [15]. These works try to either aggregating local features to global representation, e.g., [6] and [16], or mapping features from one camera view to another, e.g., [17] and [18]. On the other hand, multiple visual traits (e.g., color, texture, and spatial structure) are jointly used to describe the individual appearance, via either concatenating different descriptors, e.g., [8] and [11], or weighting different distance metrics, e.g., [5] and [19]. While these works have improved the Re-Id performance, however, there is still a lack of comprehensive evaluations on the detailed configurations of feature extraction and the strategies for features combination.

In this paper, we propose a unified pipeline, as illustrated in Fig. 2, for extracting and combining multiple statistical features for person Re-Id. To be more specific, we extract five types of spatial pyramid-based statistical features, including spatial pyramid-based color histogram (spHist), spatial pyramid-based histogram of oriented gradient (spHOG),

Manuscript received July 5, 2016; revised November 7, 2016; accepted December 12, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61601042 and Grant 61402047, in part by the 111 project under Grant B08004, and in part by the Beijing Natural Science Foundation under Grant 4162044. This paper was recommended by Associate Editor K. Huang. *(Corresponding author: Chun-Guang Li.)*

The authors are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: sijianlou@bupt.edu.cn; zhhg@bupt.edu.cn; lichunguang@bupt.edu.cn; guojun@bupt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2645660

2168-2216 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

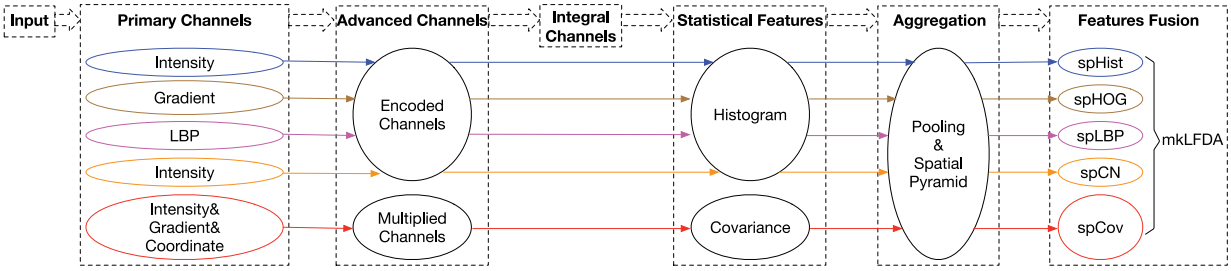


Fig. 2. Pipeline of our proposed framework.

spatial pyramid-based local binary pattern (spLBP), spatial pyramid-based color names (spCNs), and spatial pyramid-based covariance feature (spCov), and combine them via multiple kernel local Fisher discriminant analysis (mkLFDA). Moreover, we conduct comprehensive experiments on four benchmark datasets to evaluate the critical components in detailed configurations of feature extraction and also the effectiveness of different strategies for combining multiple features. In addition, we show in experiments that our framework can yield state-of-the-art performance on benchmark datasets when properly combining multiple features.

The main contributions of this paper are highlighted as follows.

1) We propose a unified pipeline for extracting and combining spatial pyramid-based statistical features for Re-Id.
2) We conduct extensive evaluations on detailed configurations in feature extraction and on different strategies for feature combination.
3) Experimental results on four benchmark datasets demonstrate that our proposed approach is comparable or even surpassing the state-of-the-art performance.

The rest of this paper is organized as follows. In Section II, we review the related works. In Section III, we present the detailed pipeline of our framework. Extensive experimental evaluations are provided in Section IV. We conclude with discussion in Section V.

II. Related Work

In person Re-Id, appearance changes caused by camera configuration, human pose, and photographic environment lead to large intraclass/interclass variations and make the problem difficult to tackle. Therefore, extensive efforts have been paid to address this difficulty. Roughly, the existing works can be divided into two categories.

1) Constructing invariant and discriminative representation, e.g., [5]–[7], [16], and [19]–[32].
2) Learning discriminative distance metric, e.g., [7]–[12], [15], [18], and [33]–[41].

Almost all of these works make their own contributions upon local statistical features. Zhao *et al.* [6], [16] weighted the local statistical features by human saliency, and many researchers [26]–[28] proposed to learn a robust higher-level attribute representation from the basic local statistical features. In the metric or subspace-based methods, usually, local statistical features are used as the appearance representation and a more discriminative metric or projection space is learned to further improve the Re-Id performance, e.g., Zheng *et al.* [12], [37] formulated person Re-Id as a relative distance comparison learning problem, Xiong *et al.* [15] introduced kernel trick into linear metric models to improve the model performance.

In addition, there exist a number of works dealing with person Re-Id from other perspectives. For example, dictionary learning and sparse representation were used to cope with the appearance transformation across camera views and obtained a more robust feature subspace [17], [42]–[47], patches correspondence between image pairs was learned to alleviate the spatial misalignment [48], the space of feature transformation was invested to reidentify individuals [39], [49], and recently, deep learning was also introduced to boost the Re-Id performance [29], [30]. Besides, there are also some interesting works for Re-Id from miscellaneous ways, e.g., via partial-body [42], gait information [50], super-resolution [46], [51], multishot [52], [53], cross-domain [54], [55], camera-network [56], or dealing with large-scale dataset [57], [58].

Although these existing works have improved the Re-Id performance significantly, they have not exploited the full potential of the representation due to lacking of efforts to systematically analyzing local statistical features themselves. Nevertheless, making a better use of feature representation is a crucial component to further improve the performance.

In the past decade, feature extraction framework based on spatial pyramid matching (SPM) [59], [60] and integral channel features (ChnFtrs) [61] have achieved great successes in image categorization and object detection tasks, respectively. However, neither of them is suitable for Re-Id due to the following two characteristics of this task.

1) Person Re-Id is a fine-grained classification problem to classify each person with the corresponding identity, so the feature representation needs to be more discriminative than that extracted through SPM for category classification.
2) The feature for Re-Id should also be invariant across camera views, so more attentions need to be paid on constructing robust higher-level features instead of registered primary image channels of ChnFtrs. Thus, Re-Id needs a specific framework.

In this paper, we propose a unified pipeline, which integrates SPM and ChnFtrs, to construct spatial pyramid-based

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 3. Examples of image channels. (a) Original image. (b)–(e) Primary image channels. (f) and (g) Advanced image channels.

statistical features for person Re-Id, and then combine the extracted multiple features by mkLFDA.

III. Proposed Framework

In this section, we present a unified pipeline to extract and combine five spatial pyramid-based statistical features for person Re-Id. For clarity, we illustrate the flowchart of the pipeline in Fig. 2. The pipeline consists of five successive stages: 1) generating primary channels; 2) constructing advanced channels; 3) extracting local region statistical features; 4) features aggregation; and 5) features combination.

A. Generating Primary Image Channels

Image channels can be regarded as the feature maps of the original image, where each output pixel is computed from the input pixels in the corresponding spatial region. The primary image channels are usually derived from single channel images with simple process. Specifically, given a single channel image $\mathbf{I} \in \mathbb{R}^{H\times W}$, where H and W denote the height and width of image $\mathbf{I}$, respectively. The primary image channels can be denoted as $\mathbf{C} = \Omega(\mathbf{I}) \in \mathbb{R}^{H\times W}$, where Ω is a simple transform, e.g., linear filtering, identity transformation, and nonlinear or pixel-wise transformations as in [61].[1] These channels carry on sufficient information to compute more powerful discriminative features. In this paper, we use four primary channels: 1) intensity channels; 2) coordinate channels; 3) gradient channels; and 4) local binary pattern (LBP) [62]. For completeness, we give a brief review on each of these primary channels.

1) Intensity Channels: Identity transformation is the simplest way to generate primary image channels. These channels contain all details of the original image. In Fig. 3(b), we display the gray and HSV color intensity channels, separately.

2) Coordinate Channels: A pixel in an image contains both intensity and location information. In this paper, we take the coordinates of each pixel to compose coordinate channels. Although it seems useless for discriminating individuals by itself, it could make great contributions to construct spatial-keeping features when combined with other channels.

[1]It also should be pointed out that, to maintain the size of the image channels unchanged, we pad the original image with the mirror pixels of the boundaries.

3) Gradient Channels: Gradient in an image is to describe the directional intensity changes in a local region, which characterizes the object shape or texture. In this paper, we use a simple gradient kernel $\mathbf{k} = [-1, 0, 1]^T$ to extract four types of gradient features, i.e., horizontal gradient amplitude ($\mathbf{C}_{|G_x|} = |\mathbf{I}\otimes\mathbf{k}^T|$), vertical gradient amplitude ($\mathbf{C}_{|G_y|} = |\mathbf{I}\otimes\mathbf{k}|$), gradient magnitude ($\mathbf{C}_{|G_{xy}|} = \sqrt{(\mathbf{I}\otimes\mathbf{k})^2 + (\mathbf{I}\otimes\mathbf{k}^T)^2}$), and gradient orientation ($\mathbf{C}_{G_\theta} = \text{atan2}(\mathbf{I}\otimes\mathbf{k}, \mathbf{I}\otimes\mathbf{k}^T) + \pi$), where $\otimes$ denotes discrete convolution and atan2(., .) calculates the four quadrant arctan.

4) LBP Channel: LBPs are binary sequences determined by the signs of the intensity difference between the central pixel and its neighboring pixels, which are invariant to monotonic transformation of intensity image. In this paper, we take a quantized LBP image as a primary image channel, which is generated by sliding a mask window of 3×3 over the image followed by quantization.

To gain some intuition of different primary channels, we illustrate them in Fig. 3(b)–(e).

B. Constructing Advanced Image Channels

Advanced channels are constructed over a set of primary channels for fast calculation of statistical features. Specifically, given a set of primary channels $\mathcal{C} = \{\mathbf{C}^{(m)} \in \mathbb{R}^{H\times W}, m = 1, \ldots, M\}$, the advanced channels are generated as $\mathcal{C}_\mathcal{A} = \Omega(\mathcal{C}) \in \mathbb{R}^{H\times W\times N}$, where the bold Ω denotes a mapping function which corresponds to encoding or multiplying operation. For clarity, we displayed some examples of advanced image channels in panels (f) and (g) of Fig. 3.

1) Encoded Image Channels: Given an image with M primary channels and a predefined codebook $\mathbf{V} = \{\mathbf{v}_n \in \mathbb{R}^M, n = 1, \ldots, N\}$, the M-dimensional feature vector $\mathbf{f}_{ij}$ at each pixel (i, j) is encoded with respect to the codebook as N coefficients $\{a_{ij}^n, n = 1, \ldots, N\}$. Then, the nth encoded channel is generated as $\mathbf{C}^{(n)}(i, j) = a_{ij}^n$. This type of advanced channels is used for the purpose of fast computing histogram-like features.

According to the encoding strategy, we sort the encoding methods into hard encoding [e.g., histogram encoding (HE)] and soft encoding [e.g., kernel codebook encoding (KCE), linear interpolation encoding (LIE), and salient color encoding (SCE)].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

a) Histogram encoding: Each feature vector $\mathbf{f}_{ij}$ is encoded as an N-dimensional coefficient vector as follows:

$$a_{ij}^n = 1 \quad \text{if} \quad n = \arg \min_{n' \in \{1,\ldots,N\}} \left\| \mathbf{f}_{ij} - \mathbf{v}_{n'} \right\|^2 \tag{1}$$

otherwise $a_{ij}^n = 0$. This means that only the coefficient corresponding to the group to which $\mathbf{f}_{ij}$ belongs is assigned to 1.

b) Kernel codebook encoding: KCE is based on kernel density estimation, in which each feature vector $\mathbf{f}_{ij}$ is encoded as

$$a_{ij}^n = k\left(\mathbf{f}_{ij}, \mathbf{v}_n\right) / \sum_{l=1}^{N} k\left(\mathbf{f}_{ij}, \mathbf{v}_l\right) \tag{2}$$

where the kernel function is $k(\mathbf{f}, \mathbf{v}) = \exp(-(\gamma/2)||\mathbf{f} - \mathbf{v}||^2)$.

c) Linear interpolation encoding: When encoding each primary image channel (i.e., $M = 1$) separately, feature f_{ij} can also be encoded as

$$a_{ij}^n = \max\left(0, 1 - \left|f_{ij} - v_n\right|/b\right) \tag{3}$$

where b is the interbin distance. The linear interpolation strategy can be easily extended to 2-D or 3-D space, which leads to bilinear or trilinear version used in [14].

d) Salient color encoding: When jointly quantizing the three-channel color space into a discrete color name space, SCE can also be adopted in [23]. At first, the color feature space (denoted as $\mathbb{F}$) is uniformly divided into $32 \times 32 \times 32$ cubes $\{\mathbb{F}_c, c = 1, \ldots, 32768\}$, where each cube covers 512 color values, i.e., $\mathbb{F}_c = \{\mathbf{f}^{(l)}, l = 1, \ldots, 512\}$. Then the probability of assigning each $\mathbf{f}_{ij} \in \mathbb{F}_c$ to a color name $\mathbf{v}_n$ is defined as

$$a_{ij}^n = p(\mathbf{v}_n|\mathbb{F}_c), \quad \text{for } \mathbf{f}_{ij} \in \mathbb{F}_c \tag{4}$$

where

$$p(\mathbf{v}_n|\mathbb{F}_c) = \sum_{l=1}^{512} p\left(\mathbf{v}_n|\mathbf{f}^{(l)}\right) p\left(\mathbf{f}^{(l)}|\mathbb{F}_c\right) \tag{5}$$

in which if $\mathbf{v}_n \in \text{KNN}(\mathbf{f}^{(l)})$, then

$$p\left(\mathbf{v}_n|\mathbf{f}^{(l)}\right) = \frac{\exp\left(\dfrac{-\left\|\mathbf{v}_n - \mathbf{f}^{(l)}\right\|^2}{\frac{1}{k-1}\sum_{\mathbf{v}_p \neq \mathbf{v}_n} \left\|\mathbf{v}_p - \mathbf{f}^{(l)}\right\|^2}\right)}{\sum_{q=1}^{k} \exp\left(\dfrac{-\left\|\mathbf{v}_q - \mathbf{f}^{(l)}\right\|^2}{\frac{1}{k-1}\sum_{\mathbf{v}_s \neq \mathbf{v}_q} \left\|\mathbf{v}_s - \mathbf{f}^{(l)}\right\|^2}\right)} \tag{6}$$

otherwise $p(\mathbf{v}_n|\mathbf{f}^{(l)}) = 0$, and

$$p\left(\mathbf{f}^{(l)}|\mathbb{F}_c\right) = \frac{\exp\left(-\alpha\left\|\mathbf{f}^{(l)} - \boldsymbol{\mu}_c\right\|^2\right)}{\sum_{t=1}^{512} \exp\left(-\alpha\left\|\mathbf{f}^{(t)} - \boldsymbol{\mu}_c\right\|^2\right)} \tag{7}$$

where k refers to the number of nearest neighbors, and $\boldsymbol{\mu}_c$ is the mean color vector of $\mathbb{F}_c$.

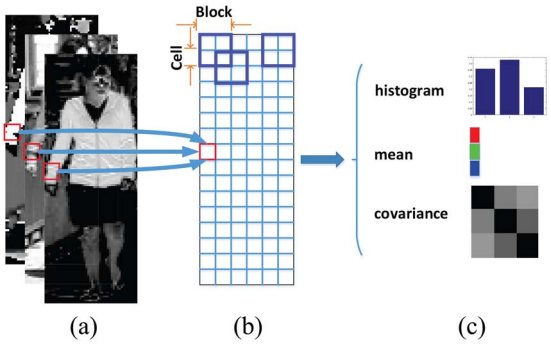


Fig. 4. Illustration of local statistical features extraction. Three types of statistical features, i.e., histogram, mean vector, and covariance matrix, are computed by using the local sums on different image channels. (a) Image channels. (b) Dense grid. (c) Statistical features.

2) Multiplied Image Channels: Multiplied channels are multiplication of any two primary channels. Specifically, given a set of primary channels $\mathcal{C} \in \mathbb{R}^{H \times W \times M}$, the multiplied channel can be computed as

$$\mathbf{C}^{(m_1,m_2)} = \mathbf{C}^{(m_1)} \odot \mathbf{C}^{(m_2)} \tag{8}$$

where $\mathbf{C}^{(m_1)}, \mathbf{C}^{(m_2)} \in \mathcal{C}$, and $\odot$ denotes element-wise multiplication. Similar to encoded channels, multiplied image channels are used to facilitate the calculation of covariance feature within a local region.

C. Extracting Local Statistical Features

Owning to the primary channels and advanced channels, the statistical features (e.g., histogram, mean vector, and covariances) in each region of interest (ROI) can be extracted effectively. Given a set of channels, the sums over each ROI is calculated from each channel separately at the same position, then they are combined to construct a particular type of statistical feature. Specifically, four histogram-like features (i.e., spHist, spHOG, spLBP, and spCN) and one covariance feature (i.e., spCov) are extracted. The ROIs are defined as cells, which are dense rectangles generated by gridding the input channels, as illustrated in Fig. 4. After resizing each image into 128×48, the cell size is set as 4×4 for spHist and spCN, and 8×8 for others in our framework.

To speed up the local feature extraction, we also introduce integral channels as intermediate channels for fast calculation of region sums. Each pixel in the integral image can be calculated quickly by summing up all the pixels inside a rectangle bounded by the upper left corner of the input image and the pixel of interest. For an input channel $\mathbf{C}$, its integral channel is defined as

$$\mathbf{C}_{\text{Intg}}\left(i', j'\right) = \sum_{i \leq i', j \leq j'} \mathbf{C}(i, j). \tag{9}$$

Equipped with the integral channels, the local statistical features can be calculated more efficiently. Let $\mathcal{C}_{\text{Intg}} \in \mathbb{R}^{H \times W \times M}, \mathcal{C}_{\text{Intg}'} \in \mathbb{R}^{H \times W \times N}, \mathcal{C}_{\text{Intg}''} \in \mathbb{R}^{H \times W \times M \times M}$ be the integral image sets from primary channels, encoded channels, and multiplied channels, respectively. For simplicity, we also denote the vector $\mathbf{p}_{i,j} = \mathcal{C}_{\text{Intg}}(i, j, :), \mathbf{q}_{i,j} = \mathcal{C}_{\text{Intg}'}(i, j, :),$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

and $\mathbf{E}_{i,j} = \mathcal{C}_{\text{Intg}''}(i, j, :, :)$. Then, the mean vector over a cell at $\{(i', j'), (i'', j'')\}$ is computed as

$$\mathbf{u}_{(i',j';i'',j'')} = \frac{\left(\mathbf{p}_{i'',j''} + \mathbf{p}_{i'-1,j'-1} - \mathbf{p}_{i'-1,j''} - \mathbf{p}_{i'',j'-1}\right)}{S}$$

the histogram over a cell is computed as

$$\mathbf{h}_{(i',j';i'',j'')} = \left(\mathbf{q}_{i'',j''} + \mathbf{q}_{i'-1,j'-1} - \mathbf{q}_{i'-1,j''} - \mathbf{q}_{i'',j'-1}\right)$$

and the covariance matrix [63] is calculated as

$$\mathbf{O}_{(i',j';i'',j'')} = \frac{1}{S-1}\big[\mathbf{E}_{i'',j''} + \mathbf{E}_{i'-1,j'-1} - \mathbf{E}_{i'-1,j''} - \mathbf{E}_{i'',j'-1} - S\mathbf{u}_{(i',j';i'',j'')}\mathbf{u}_{(i',j';i'',j'')}{}^{T}\big]$$

where $S = (i'' - i' + 1)(j'' - j' + 1)$.

Note that extracting local statistical features from the original integral channels as in (9) suffers from a spatial aliasing problem which is caused by the pixels near the cell boundaries.[2] To alleviate this shortcoming, we introduced a spatial trilinear interpolation step (as in [14]), i.e., by preconvolution on the input channels with a predefined kernel as in [64]. In this paper, the convolution kernel $\mathbf{K}$ for 4×4 or 8×8 cell-based feature is defined as

$$\frac{1}{16}\begin{bmatrix}1 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 1\end{bmatrix}$$

or

$$\frac{1}{256}\begin{bmatrix}1 & 2 & 3 & 4 & 3 & 2 & 1\\ 2 & 4 & 6 & 8 & 6 & 4 & 2\\ 3 & 6 & 9 & 12 & 9 & 6 & 3\\ 4 & 8 & 12 & 16 & 12 & 8 & 4\\ 3 & 6 & 9 & 12 & 9 & 6 & 3\\ 2 & 4 & 6 & 8 & 6 & 4 & 2\\ 1 & 2 & 3 & 4 & 3 & 2 & 1\end{bmatrix}$$

where the weights are distributed according to the distance between the position of pixels and their neighbors. Then the convoluted integral channels are generated as

$$\mathbf{C}_{\text{Intg}}(i', j') = \sum_{i\le i', j\le j'} \mathbf{C}_{\text{Cov}}(i, j), \text{ where } \mathbf{C}_{\text{Cov}} = \mathbf{C} \otimes \mathbf{K}.$$

Remark 1: To eliminate the variations caused by illumination changes and cluttering background, we take a local contrast normalization for the extracted cell-based features. Similar to [14], we group each 2×2 cells into a larger half-overlapping block and normalize each block separately, as illustrated in Fig. 4(b). The grouping is performed by concatenating each cell-based feature for spHOG and by averaging for others. Specifically, we consider five different schemes as follows: 1) ℓ_1-norm, $\mathbf{x} \to \mathbf{x}/(\|\mathbf{x}\|_1 + \epsilon)$; 2) ℓ_1-sqrt, $\mathbf{x} \to \sqrt{\mathbf{x}/\|\mathbf{x}\|_1 + \epsilon}$; 3) ℓ_2-norm, $\mathbf{x} \to \mathbf{x}/(\|\mathbf{x}\|_2 + \epsilon)$; 4) ℓ_2-clip, limiting the maximum values following the ℓ_2-norm; and 5) ℓ_1^2-norm, $\mathbf{x} \to \mathbf{x}/(||\mathbf{x}||_1^2 + \epsilon)$, where $\epsilon > 0$ is a tiny constant. For covariance features, we normalize these block-based covariance matrices $\mathbf{O}_{\text{block}}$ as $\mathbf{O}_{\text{block}} \to \text{diag}(\mathbf{O}_{\text{block}})^{-(1/2)}\mathbf{O}_{\text{block}}\text{diag}(\mathbf{O}_{\text{block}})^{-(1/2)}$, then we stack the upper triangular part of the covariance matrix into a feature vector due to symmetry. After that, the covariance features can be normalized as vectorial features.

[2]When doing encoding, a naive distribution scheme such as voting the nearest codeword would result in aliasing effects. Similarly, when extracting cell-based local features, pixels near the cell boundaries would produce aliasing along spatial dimensions. Such aliasing effects can cause sudden changes in the computed feature vector.

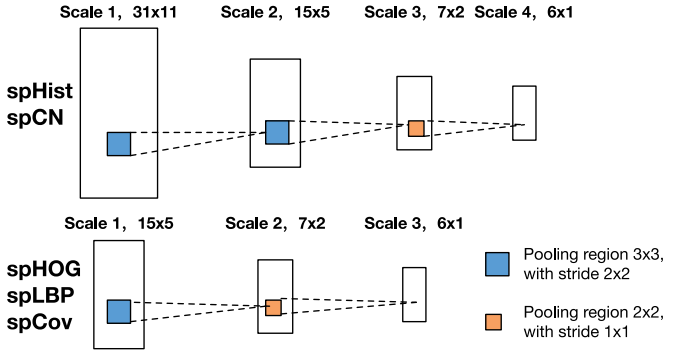


Fig. 5. Multiscale pooling. In this paper, the input images are resized to 128×48, so after local normalization there are 31×11 block features for spHist and spCN, and 15×5 block features for others.

D. Features Aggregation

1) Pooling: In our framework, we extract spatial pyramid features by multiscale pooling, as illustrated in Fig. 5. Specifically, after pooling, spHist and spCN consist of features extracted from $31 \times 11 + 15 \times 5 + 7 \times 2 + 6 \times 1 = 436$ blocks, and other features are extracted from $15\times5+7\times2+6\times1 = 95$ blocks. We evaluate two pooling methods: 1) average pooling and 2) max pooling.

2) Spatial Pyramid: As in SPM framework, the final features are aggregated in a spatial pyramid way by combining the multiscale features. For simplicity and computational efficiency, we directly stack the multiscale features as a concatenated vector. Moreover, normalization is also applied on both the single-scale features and the final concatenated features.

To evaluate spatial pyramid-based features, we compare two components: 1) spatial scale and 2) normalization method. To be more specific, we compare multiscale features to single-scale features, and evaluate two normalization methods (i.e., ℓ_1-norm and ℓ_2-norm).

For clarity, we illustrate the data articulation flow by different processing operations in the feature extraction pipeline in Fig. 6. The default settings of extracting the spatial pyramid-based statistical features, i.e., spHist, spHOG, spLBP, spCN, and spCov, and also settings of extracting the original features, i.e., color histogram (Hist), HOG [14], LBP [65], salient color names (SCNs) [23] and covariance matrix (Cov) [63] are listed in Table I.

E. Features Combination Based on mkLFDA

1) Kernel Local Fisher Discriminant Analysis: kLFDA [66] aims to simultaneously maximize the interclass separability and minimize the intraclass variance while preserving the local neighborhood structure. The optimization objective is as

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

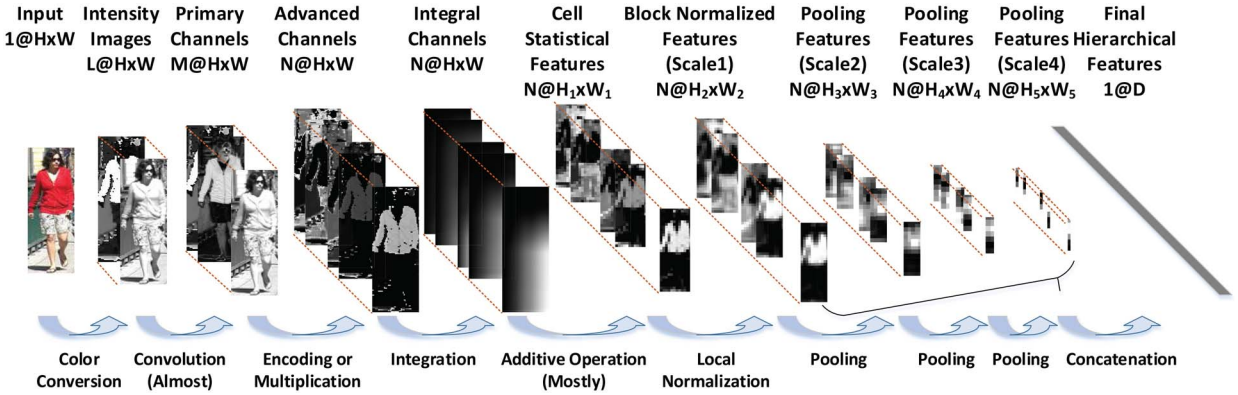


Fig. 6. Data flow of our feature extraction pipeline. The texts on top indicate the data types, and the texts in bottom show the corresponding operations.

TABLE I
CONFIGURATION SETTINGS FOR DIFFERENT FEATURES EXTRACTION

Feat	Configuration Settings					
	Input	$\mathcal{C}_{\mathcal{A}}$	$\mathcal{C}_{Intg}$	Contrast Normalize	Pooling	Spatial Scale & Normalize
spHist	HSV	KCE, $\gamma = 1/12$	Conv	ℓ_2-norm	Max	Multi, ℓ_1
spHOG	YUV	KCE, $\gamma = \pi/6$	Conv	ℓ_1^2-norm	Avg	Multi, ℓ_2
spLBP	YUV	HE	Conv	ℓ_2-norm	Max	Multi, ℓ_2
spCN	RGB	SCE	Conv	ℓ_2-norm	Max	Multi, ℓ_2
spCov	Gray, HSV, YUV, LAB	MULT	Orig	ℓ_2-norm	Avg	Multi, ℓ_2, +Mean
Hist	HSV	HE	Orig	ℓ_2-norm	—	Single, ℓ_1
HOG	Gray	LIE	Conv	ℓ_2-norm	—	Single, ℓ_2
LBP	Gray	HE	Orig	ℓ_1-sqrt	—	Single, ℓ_1
SCN	RGB	SCE	Orig	ℓ_2-norm	—	Single, ℓ_2
Cov	Gray	MULT	Orig	ℓ_2-norm	—	Single, ℓ_2

follows:

$$\max_{\mathbf{A}} \operatorname{tr}\left(\mathbf{A}^T\widetilde{\mathbf{S}}^{(b)}\mathbf{A}\right)/\operatorname{tr}\left(\mathbf{A}^T\widetilde{\mathbf{S}}^{(w)}\mathbf{A}\right) \tag{10}$$

where $\operatorname{tr}(\cdot)$ is the trace of a matrix, $\mathbf{A} \in \mathbb{R}^{s\times d}$ is the projection matrix, and s is the data size. $\widetilde{\mathbf{S}}^{(b)}$ and $\widetilde{\mathbf{S}}^{(w)}$ denote the between-class and within-class local scatter matrix, respectively, in which $\widetilde{\mathbf{S}}^{(b)} = (1/2)\sum_{i,j=1}^{s}\widetilde{\mathbf{W}}_{ij}^{(b)}(\mathbf{k}_i - \mathbf{k}_j)(\mathbf{k}_i - \mathbf{k}_j)^T \in \mathbb{R}^{s\times s}$ and $\widetilde{\mathbf{S}}^{(w)} = (1/2)\sum_{i,j=1}^{s}\widetilde{\mathbf{W}}_{ij}^{(w)}(\mathbf{k}_i - \mathbf{k}_j)(\mathbf{k}_i - \mathbf{k}_j)^T \in \mathbb{R}^{s\times s}$, where $\mathbf{k}_i = [\kappa(\mathbf{x}_1, \mathbf{x}_i), \ldots, \kappa(\mathbf{x}_s, \mathbf{x}_i)]^T \in \mathbb{R}^s$, $\widetilde{\mathbf{W}}^{(b)}$ and $\widetilde{\mathbf{W}}^{(w)}$ denote the weight matrices of the between-class and within-class local adjacency graph, respectively.

Problem (10) can be solved by the generalized eigenvalue problem as $\widetilde{\mathbf{S}}^{(b)}\mathbf{A} = \lambda\widetilde{\mathbf{S}}^{(w)}\mathbf{A}$, where the optimal $\mathbf{A}_* \in \mathbb{R}^{s\times d'}$ is composed of the d' leading eigenvectors corresponding to the d' largest eigenvalues.

2) Multiple Kernel Local Fisher Discriminant Analysis: Instead of designing hand-crafted kernel for the input data, MKL automatically learns an ensemble kernel $\kappa(\cdot,\cdot)$ over a given set of kernels $\kappa^{(p)}(\cdot,\cdot)$. In this paper, the ensemble kernel function is defined as

$$\kappa\left(\mathbf{x}, \mathbf{x}'\right) = \sum_{p=1}^{P}\beta_p\kappa^{(p)}\left(\mathbf{x}, \mathbf{x}'\right) \tag{11}$$

where $\beta_p \geq 0$ for $p = 1, \ldots, P$. Different kernels correspond to the similarities defined by using different functions or using inputs from different representations. Then, by replacing the Fisher discriminant ratio in (10) with its equivalent quadratic transformation, we formulate mkLFDA as

$$\min_{\mathbf{A},\boldsymbol{\beta}} \operatorname{tr}\left(\mathbf{A}^T\widetilde{\mathbf{S}}_{\boldsymbol{\beta}}^{(w)}\mathbf{A}\right), \text{ s.t. } \operatorname{tr}\left(\mathbf{A}^T\widetilde{\mathbf{S}}_{\boldsymbol{\beta}}^{(b)}\mathbf{A}\right) = 1 \text{ and } \boldsymbol{\beta} \geq \mathbf{0} \tag{12}$$

where

$$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_P]^T \in \mathbb{R}^P \tag{13}$$

$$\mathbb{K}^{(i)} = \left[\mathbf{k}_i^{(1)}, \ldots, \mathbf{k}_i^{(P)}\right] \in \mathbb{R}^{s\times P} \tag{14}$$

$$\widetilde{\mathbf{S}}_{\boldsymbol{\beta}}^{(w)} = \sum_{i,j=1}^{s}\frac{1}{2}\widetilde{\mathbf{W}}_{ij}^{(w)}\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right)\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right)^T$$

$$\widetilde{\mathbf{S}}_{\boldsymbol{\beta}}^{(b)} = \sum_{i,j=1}^{s}\frac{1}{2}\widetilde{\mathbf{W}}_{ij}^{(b)}\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right)\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right)^T. \tag{15}$$

3) Optimization Algorithm: Instead of optimizing $\mathbf{A}$ and $\boldsymbol{\beta}$ simultaneously, we follow the iterative optimization strategy in [67], i.e., alternately updating one variable with the other fixed.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

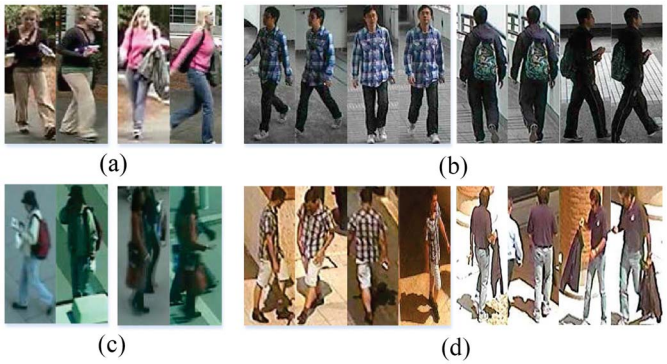


Fig. 7. Examples of person images from different datasets. (a) VIPeR. (b) CUHK01. (c) PRID2011. (d) 3DPeS.

With $\boldsymbol{\beta}$ fixed, mkLFDA retrogresses to kLFDA as in (10), and can be solved as a generalized eigenvalue problem. With $\mathbf{A}$ fixed, mkLFDA can be reformulated as a nonconvex quadratically constrained quadratic programming problem, which can be solved efficiently via its semidefinite programming relaxation [67] as

$$\begin{aligned} \min_{\boldsymbol{\beta},\mathbf{B}} \quad & \operatorname{tr}\left(\widetilde{\mathbf{S}}_{\mathbf{A}}^{(w)}\mathbf{B}\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(\widetilde{\mathbf{S}}_{\mathbf{A}}^{(b)}\mathbf{B}\right) = 1, \ \boldsymbol{\beta} \succeq \mathbf{0}, \text{ and } \begin{bmatrix} 1 & \boldsymbol{\beta}^T \\ \boldsymbol{\beta} & \mathbf{B} \end{bmatrix} \succeq 0 \end{aligned} \tag{16}$$

where

$$\begin{aligned} \widetilde{\mathbf{S}}_{\mathbf{A}}^{(w)} &= \sum_{i,j=1}^{s} \frac{1}{2}\widetilde{\mathbf{W}}_{ij}^{(w)}\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right)^T \mathbf{A}\mathbf{A}^T\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right) \\ \widetilde{\mathbf{S}}_{\mathbf{A}}^{(b)} &= \sum_{i,j=1}^{s} \frac{1}{2}\widetilde{\mathbf{W}}_{ij}^{(b)}\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right)^T \mathbf{A}\mathbf{A}^T\left(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}\right). \end{aligned} \tag{17}$$

IV. Experiments

In this section, we provide extensive evaluations on the configurations of the spatial pyramid features, the different strategies for combining multiple features, and also the performance comparison with the state-of-the-art methods.

A. Data Sets and Experimental Protocols

1) Data Sets Description: We choose four publicly available benchmark datasets for person Re-Id, including VIPeR [68], CUHK01 [69], PRID2011 [70], and 3DPeS [71]. Each image is resized into a fixed size 128×48. Some example images are shown in Fig. 7.

VIPeR contains 632 pedestrian image pairs with obvious viewpoint and illumination changes from two cameras in an academic campus. Most pairs have at least 90° intraclass angular variation.

CUHK01 is composed of 971 pedestrian images from two cameras. Each individual has two images per camera.

PRID2011 consists of 385 individuals from camera A and 749 individuals from camera B, where only the first 200 pedestrians appear in both camera views.

3DPeS contains numerous video sequences taken from a real outdoor surveillance scenario with eight camera views. It totally consists of 1011 images of 192 individuals, where each individual has 2–26 images.

2) Experimental Settings: We randomly and evenly divide each dataset, except PRID2011, into training set and testing set. In PRID2011, we use 100 of the 200 pedestrians appear on both camera views for training and use the rest 100 pedestrians with the extra of 549 pedestrians which only appear on camera B for testing.

In the testing phase, the single-shot matching mechanism is adopted, and the matching result is recorded using the cumulative match characteristic (CMC) performance curves. Each experiment is repeated ten trials and the average accuracy is recorded.

In our experiment, we apply kLFDA [66] as the metric learning model to evaluate the features extracted with different parameter settings, and as the base model to constitute the final multiple kernel learning model. The dimension of each spatial pyramid feature is reduced to 300 by PCA, except for PRID2011, which is reduced to 100. In kLFDA, we use Gaussian kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-(||\mathbf{x} - \mathbf{x}'||^2/2\sigma^2)\right)$. As we do not normalize features after PCA, the Euclidean distance between each two feature vectors is around 100. Thus, the bandwidth parameter σ is set as 100 for all kinds of features. The number of the nearest neighbors for computing the affinity matrix is set to 2 for CUHK01 and 1 for others, and the regularization parameter for class scatter matrix is set to 0.01.

B. Details in Spatial Pyramid Statistical Features

In our framework, we construct spHist, spHOG, and spLBP from each single color channel, separately, and then concatenate them into the whole multiple color channel features.

For spHist, the visual vocabulary consists of the centrals of eight equant histogram bins in each intensity channel, and the whole spHist feature has a length of $8 \times 436 \times 3$.

For spHOG, we use 18 orientation bins which are spaced evenly over 0 to 2π. Moreover, the concatenation grouping strategy is used to compute block features and the dimension of each block normalized histogram is extended to 72. Therefore, a $72 \times 95 \times 3$-D spHOG will be extracted for each color image.

For spLBP, we use 3×3 neighborhood to compute LBP, and thus obtain 59 uniformed patterns. The final spLBP has a length of $59 \times 95 \times 3$.

For spCN, the color distribution is described by quantifying the 3-D color space jointly into N color names. We use 16 color names as in [23]. Therefore, the total dimension of an spCN feature is 16×436.

For spCov, we take more color information into the basic feature set by using multiple intensity channels generated from different color spaces, and concatenate them as follows:

$$\left[x, y, |\mathrm{G}_x|, |\mathrm{G}_y|, \mathrm{L}, \mathrm{A}, \mathrm{B}, \mathrm{H}, \mathrm{S}, \mathrm{V}_{hsv}, \mathrm{Y}, \mathrm{U}, \mathrm{V}_{yuv}\right] \tag{18}$$

where $\mathrm{L}, \mathrm{A}, \mathrm{B}, \mathrm{H}, \mathrm{S}, \mathrm{V}_{hsv}, \mathrm{Y}, \mathrm{U},$ and V_{yuv} denotes different intensity channels. Thus, a 13×13 matrix descriptor can be computed from each cell. In addition, the mean vector is also used, and thus an spCov feature of $(91 + 13) \times 95$ is obtained for each input image.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
TOP r MATCHING ACCURACY ON VIPER USING SPHIST WITH DIFFERENT PARAMETER SETTINGS

Configurations	Rank1	Rank5	Rank10	Rank20	Rank50
(a) **spHist**	35.19	66.30	80.38	**91.23**	98.20
(b) **Hist**	29.49	57.15	71.14	84.15	96.39
(c) RGB	15.66	41.27	57.41	73.70	90.82
(d) YUV	24.78	54.81	69.62	83.58	94.27
(e) LAB	26.30	55.16	70.22	83.83	95.35
(f) KCE, $\gamma = \frac{1}{24}$	34.49	62.66	75.63	87.66	96.84
(g) KCE, $\gamma = \frac{1}{6}$	33.89	65.51	79.05	89.81	98.10
(h) LIE	34.91	65.79	78.58	90.38	97.59
(i) HE	31.93	62.28	75.82	88.04	97.06
(j) Orig	35.70	65.92	79.02	89.59	98.07
(k) No normalization	31.52	63.29	77.15	89.18	98.01
(l) ℓ_1-norm	34.62	67.18	80.51	90.76	**98.23**
(m) ℓ_1-sqrt	**36.74**	**68.20**	80.63	90.66	97.94
(n) ℓ_1^2-norm	35.66	67.85	**80.92**	90.47	98.13
(o) ℓ_2-clip	34.65	67.25	80.38	90.38	98.10
(p) Avg	34.84	67.12	79.97	90.85	98.07
(q) Multi, ℓ_2	35.22	66.27	80.22	90.85	98.13
(r) Single, ℓ_1	35.54	65.98	78.92	89.81	98.10

TABLE III
TOP r MATCHING ACCURACY ON VIPER USING SPHOG WITH DIFFERENT PARAMETER SETTINGS

Configurations	Rank1	Rank5	Rank10	Rank20	Rank50
(a) **spHOG**	26.58	**56.96**	**71.39**	**83.67**	**95.66**
(b) **HOG**	5.16	16.65	26.90	39.94	60.76
(c) Gray	5.92	20.32	30.76	46.08	68.39
(d) HSV	23.92	50.92	63.86	78.20	92.34
(e) RGB	9.49	27.75	40.38	56.96	80.82
(f) LAB	24.18	53.07	67.91	81.96	95.06
(g) KCE, $\gamma = \frac{\pi}{18}$	23.54	52.78	66.90	80.66	94.37
(h) KCE, $\gamma = \frac{\pi}{2}$	22.53	50.38	65.70	79.81	93.86
(i) LIE	23.04	52.12	66.55	80.54	94.40
(j) HE	16.52	40.89	55.60	71.87	88.45
(k) Orig	23.83	53.77	68.99	82.59	94.91
(l) No normalization	17.44	41.84	55.35	71.77	89.59
(m) ℓ_1-norm	23.48	52.69	66.30	79.08	92.82
(n) ℓ_1-sqrt	23.45	52.06	65.92	79.59	92.63
(o) ℓ_2-norm	24.24	53.07	67.88	79.75	92.53
(p) ℓ_2-clip	23.07	52.44	66.93	79.49	93.48
(q) Max	24.46	53.67	68.99	82.72	94.62
(r) Multi, ℓ_1-norm	**27.25**	55.89	70.22	83.54	94.91
(s) Single, ℓ_2-norm	21.20	48.45	63.54	78.70	93.23

TABLE IV
TOP r MATCHING ACCURACY ON VIPER USING SPLBP WITH DIFFERENT PARAMETER SETTINGS

Configurations	Rank1	Rank5	Rank10	Rank20	Rank50
(a) **spLBP**	20.32	**49.97**	**64.08**	**78.99**	**94.08**
(b) **LBP**	4.21	14.21	22.82	37.78	60.95
(c) Gray	4.49	17.37	29.21	45.13	69.46
(d) HSV	13.58	36.46	51.99	68.13	88.99
(e) RGB	9.72	29.62	43.83	60.98	83.77
(f) LAB	16.61	41.61	56.87	73.20	90.35
(g) Orig	19.18	47.37	61.80	76.93	93.58
(h) No normalization	14.18	38.77	54.78	72.47	92.15
(i) ℓ_1-norm	18.32	45.19	60.32	77.15	92.63
(j) ℓ_1-sqrt	18.01	44.91	59.62	75.44	91.77
(k) ℓ_1^2-norm	19.18	46.74	62.50	78.35	93.54
(l) ℓ_2-clip	18.45	45.22	61.74	76.84	93.04
(m) Avg	19.53	47.78	62.41	78.01	93.77
(n) Multi, ℓ_1-norm	**20.47**	48.29	62.63	78.07	93.39
(o) Single, ℓ_2-norm	19.11	45.47	60.92	76.27	92.47

C. Evaluation on Configuration Details of Five Spatial Pyramid Statistical Features

While specific spatial pyramid-based statistical feature can be extracted via the proposed framework, different configurations may change the performance dramatically. Thus, it is necessary to systematically evaluate the detailed configurations and exploit the optimal one for person Re-Id.

In this paper, we compare different configurations step by step following the feature extraction pipeline, and conduct extensive experiments on the commonly used benchmark dataset VIPeR. For fair comparison, all configuration parameters are set as the same as default in Table I except the one on which we are evaluating. Concretely, we adjust the settings of input color spaces (RGB, HSV, YUV, LAB, or Gray), encoding methods (HE, KCE, LIE, or SCE), integral channel types (original or convolutional), local contrast normalizations (ℓ_1-norm, ℓ_1-sqrt, ℓ_2-norm, ℓ_2-clip, or ℓ_1^2-norm), pooling methods (max or average pooling), number of scales (single-scale or multiscale) and global normalizations (ℓ_1-norm or ℓ_2-norm). The comparison results of using different features are listed in Tables II–VI, respectively.

1) Input and Primary Channels: The color spaces and the primary channels contain the raw visual information, which is crucial to constructing higher-level statistical features for Re-Id. For spHist [Table II(a) versus (c)–(e)], using HSV color space yields a better illumination invariant color histogram and improves the rank 1 matching rate of 19.53%, 10.41%, and 8.89% than RGB, YUV, and LAB, respectively. For spHOG [Table III(a) versus (c)–(f)] and spLBP [Table IV(a) versus (c)–(f)], due to the color information is more crucial than the texture for Re-Id task, extracting the gradient or LBP features from color space can attain higher accuracy than only from the gray intensity space, and that is also the main reason of the remarkable performance improvement comparing with the original HOG and LBP, which only reach 5.16% and 4.21% at rank 1. As the color names are defined explicitly for RGB and HSV, e.g., a red pixel is $[1, 0, 0]^T$ in RGB space and is $[0, 1, 1]^T$ in HSV, mapping the pixel intensity into the color names space is more proper in these color spaces, and the higher accuracy [Table V(a) versus (c)–(e)], i.e., 20.13% and 21.17%, can be achieved, respectively. For spCov [Table VI(a) versus (c)], by including the color information into the inputs and primary channels (as shown in Table I), the rank 1 accuracy is promoted even 30.57% comparing with the gray version. This confirms that using color information and enriching the diversity of the primary channels is an effective way to enhance the discriminability of the features for Re-Id task.

2) Encoding Methods: The way to generate the advanced channels is critical in constructing the higher-level features. It is well known that using the encoded or multiplied channels facilitates the computing of the histogram or covariance features. Unfortunately, histogram achieves good robustness via quantifying the primitive image characters into discrete bins with the inevitable quantization error. The original color histogram utilizes the hard voting to encode. In our experiments, using the kernel codebook encoding or linear interpolation method [Table II(a) and (h) versus (i)], the rank 1 matching rate can get 3.26% or 2.98% improvement. Although the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
TOP r MATCHING ACCURACY ON VIPER USING SPCN WITH DIFFERENT PARAMETER SETTINGS

Configurations	Rank1	Rank5	Rank10	Rank20	Rank50
(a) **spSCN**	20.13	48.96	**63.99**	**78.42**	90.92
(b) **SCN**	19.02	46.49	61.11	75.41	88.99
(c) HSV	**21.17**	**49.34**	63.83	77.53	**92.63**
(d) YUV	20.06	44.94	59.43	74.08	89.65
(e) LAB	14.56	37.25	52.88	68.73	86.74
(f) KCE, $\gamma = \frac{1}{10}$	12.25	31.30	43.83	57.85	75.47
(g) KCE, $\gamma = \frac{1}{2}$	15.60	40.70	54.84	71.90	89.59
(h) KCE, $\gamma = 1$	14.97	37.31	51.27	67.69	85.73
(i) HE	16.61	41.52	55.28	70.09	86.68
(j) Orig	20.13	48.51	63.39	77.91	90.79
(k) No normalization	16.27	43.54	58.80	74.94	88.96
(l) ℓ_1-norm	18.89	47.25	61.20	76.74	89.94
(m) ℓ_1-sqrt	20.70	48.64	63.20	77.75	91.46
(n) ℓ_1^2-norm	19.97	47.94	63.07	77.97	90.16
(o) ℓ_2-clip	19.56	47.59	61.96	76.23	89.49
(p) Avg	20.28	48.35	62.72	76.42	90.00
(q) Multi, ℓ_1-norm	20.16	48.48	63.67	78.32	90.98
(r) Single, ℓ_2-norm	19.49	45.82	61.08	76.17	89.75

TABLE VI
TOP r MATCHING ACCURACY ON VIPER USING SPCOV WITH DIFFERENT PARAMETER SETTINGS

Configurations	Rank1	Rank5	Rank10	Rank20	Rank50
(a) **spCov**	**34.91**	66.80	79.15	**89.87**	98.20
(b) **Cov**	3.01	10.47	17.82	27.41	47.50
(c) Gray	4.34	14.94	23.89	37.34	61.17
(d) No normalization	18.77	47.47	63.23	80.06	94.78
(e) ℓ_1-norm	33.10	66.23	79.18	**89.87**	**98.29**
(f) ℓ_1^2-norm	33.42	64.27	77.91	88.45	97.34
(g) Max	34.46	65.63	79.08	88.99	97.85
(h) Without Mean	30.38	61.46	74.72	86.39	96.39
(i) Multi, ℓ_1-norm	33.73	**67.15**	**79.27**	88.99	98.20
(j) Single, ℓ_2-norm	31.80	63.04	77.56	88.54	97.50

original HOG feature uses the linear interpolation to diminish the quantization error, it can be further optimized by taking advantage of the kernel codebook method and achieves 3.54% improvement at rank 1 [Table III(a) versus (i)]. However, it is important to note that inappropriate parameter settings for kernel codebook encoding will lead to suboptimal performance. For example, we observe 3.57% or 1.42% accuracy drop at rank 20 when set $\gamma = 1/24$ or $1/6$ for spHist [Table II(a) versus (f) and (g)], and 3.04% or 4.05% accuracy drop at rank 1 when set $\gamma = \pi/18$ or $\pi/2$ for spHOG [Table III(a) versus (g) and (h)]. For spCN [Table V(a) versus (f)–(i)], the salient color encoding method proposed in [23] achieves the best performance 20.13% at rank 1 compared with the hard encoding and kernel codebook encoding. These confirm that, the choice of the encoding methods plays an important role in the histogram features extraction.

3) Integral Channels: Integrated channels are introduced as the intermediate channels to speed up the rectangular-based statistical features extraction. By using the convoluted integrated channels, the spatial aliasing problem can be reduced. Compared with the features extracted via the original integrated channels, the matching rate can be improved slightly via the convoluted channels (Tables II(a) versus (j), III(a) versus (k), IV(a) versus (g), and V(a) versus (j)], e.g., the rank 5 accuracy is boosted 0.38%, 3.19%, 2.7%, and 0.45% for spHist, spHog, spLBP, and spCN, respectively. Another observation is that using convoluted integrated channels has more effects on the Re-Id performance when extracting 8×8 cell-based texture features than 4×4 cell-based color features.

4) Local Contrast Normalizations: The illumination or background varies across the images. Thus, the local contrast normalization is necessary to obtain stable local features. We evaluate two different grouping strategies and five different local normalization schemes (as in Remark 1) in our experiments. For spHOG extraction [Table III(a) versus (m)–(p)], the ℓ_1^2-norm normalization scheme improves the rank 1 accuracy from 23.07% to 26.58%. For other features extraction [Tables II(a) versus (l)–(o), IV(a) versus (i)–(l), V(a) versus (l)–(o), and VI(a) versus (e) and (f)], different schemes make only nuanced performance fluctuation within the range of 1.5% to 2.3%. However, without the local contrast normalization, the accuracy will drop significantly [Tables II(a) versus (k), III(a) versus (l), IV(a) versus (h), V(a) versus (k), and VI(a) versus (d)], e.g., the rank 1 accuracy decrease from 35.19% to 31.52% for spHist, from 26.58% to 17.44% for spHOG, from 20.32% to 14.18% for spLBP, from 20.13% to 16.27% for spCN, and from 34.91% to 18.77% for spCov.

5) Multiscale Pooling and Multiscale Features Concatenation: Pooling and multiscale features concatenation are crucial for constructing hierarchical features, and make the representation more robust to the local transformation and local misalignment. In some feature vectors, the salient (or prominent) components make greater contributions to the identification task, so the max pooling leads to better performance, e.g., spHOG gets 2.12% improvement at rank 1 using max pooling than average pooling [Table III(a) versus (q)]. However, these improvements are not that significant [Tables II(a) versus (p), IV(a) versus (m), V(a) versus (p), and VI(a) versus (g)]. We also compare each statistical feature of the single-scale to the multiscale. According to the experimental results [Tables II(a) versus (r), III(a) versus (s), IV(a) versus (o), V(a) versus (r), and VI(a) versus (j)], the features from multiscale yield better performance. Specifically, on the rank 5, matching rate is improved 0.32%, 8.51%, 4.50%, 3.14%, and 3.76% with respect to spHist, spHOG, spLBP, spCN, and spCov, separately. However, different normalization strategies, e.g., ℓ_1-norm or ℓ_2-norm, on each spatial resolution features for concatenation have little effect on the final performance [Tables II(a) versus (q), III(a) versus (r), IV(a) versus (n), V(a) versus (q), and VI(a) versus (i)].

In addition, for spCov, it is important to mention that including the mean vector (i.e., first-order statistics) into the covariance feature (i.e., second-order statistics) boosts the feature performance significantly, i.e., from 30.38% to 34.91% at rank 1 [Table VI(a) versus (h)].

D. Comparison to the Original Features

We compare the performance of the spatial pyramid statistical features extracted in the proposed framework to the original features on four benchmark datasets, where all the parameter settings are the same as the configurations in Table I. Although the five specific sp-features have different degrees of effect on the performance improvement for Re-Id task, e.g., on VIPeR [Fig. 8(a)], spHist and spCov achieve 35.19% and 34.19%

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

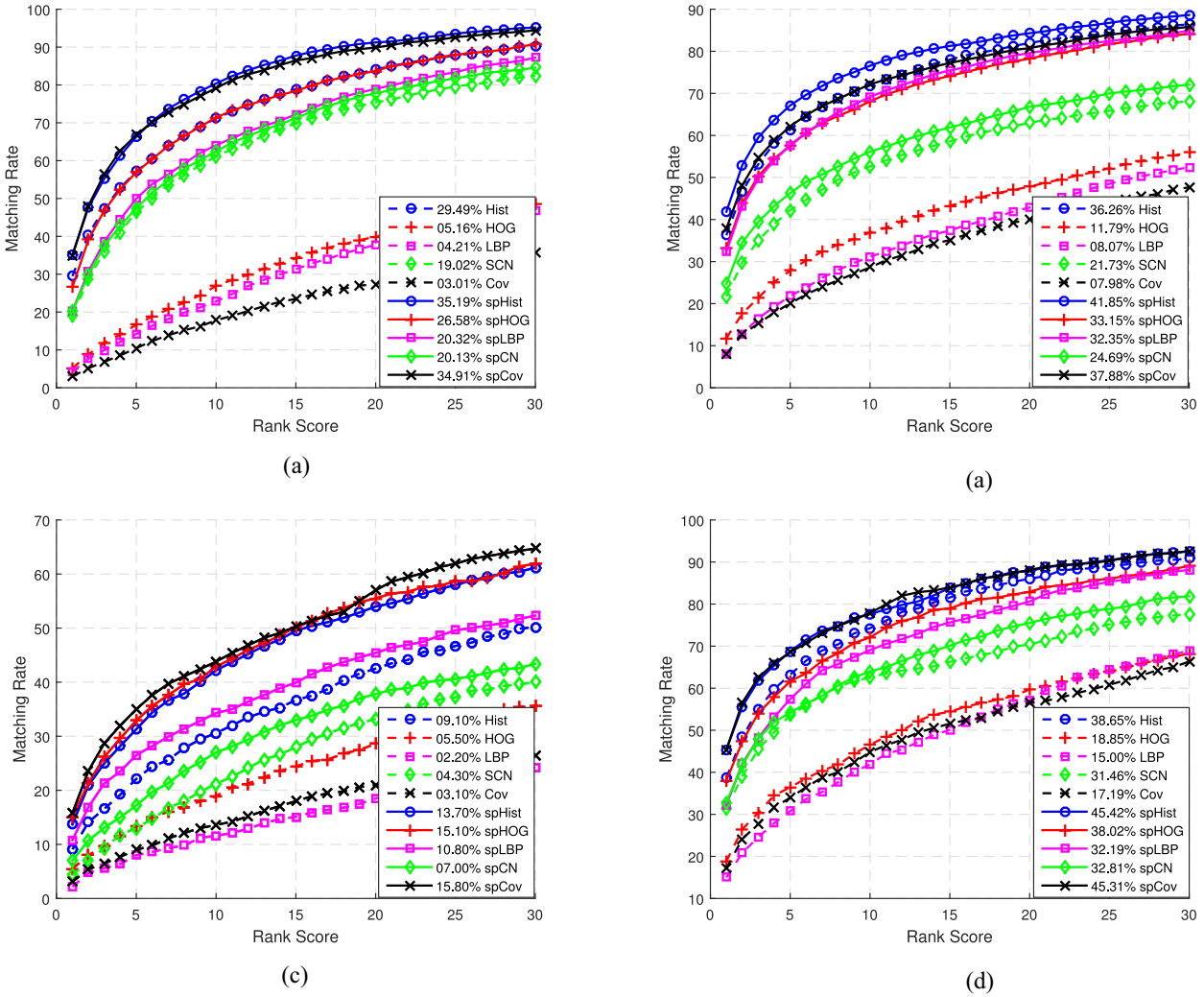


Fig. 8. Results on different datasets reported as averaged CMC curves. (a)–(d) Comparisons between spatial pyramid statistical features and their corresponding original features on four benchmarks are shown, respectively.

rank 1 matching rate, but spLBP and spCN only achieve 20.32% and 20.13%, or even the relative merits of the features vary across different datasets, e.g., spHist achieves the best accuracy than other features on VIPeR [Fig. 8(a)], but only achieves the moderate accuracy on PRID2011 [Fig. 8(c)], all the experiment results in Fig. 8 show a consistent trend that the sp-features perform significantly better than the original ones. Be specific, compared to the original features, spHist improves 6.71% at rank 1 matching accuracy on 3DPeS, spHOG improves 21.42% matching rate on VIPeR, spLBP improves 24.28% accuracy on CUHK01, spCN improves 2.7% accuracy on PRID2011, and spCov even boosts the matching rate at rank 1 with 31.9% on VIPeR.

To summarize, the performance gain of our proposed spatial pyramid-based statistical features comes from multiple aspects of our unified framework. Briefly, we use: 1) abundant input primary channels to enhance the discriminability; 2) appropriate encoding methods to minimize the information loss; 3) proper contrast normalization to smooth the local variations; and 4) spatial pyramid hierarchical strategy to alleviate the spatial misalignment.

E. Evaluation on Different MKL Methods

To obtain a better Re-Id accuracy, we use mkLFDA to combine multiple features extracted in the proposed unified framework. The optimal ensemble kernel is learned over a kernel set consists of all the linear combination of nine basis kernels [as (11)], including: five kernels from Gaussian kernel function on all the features and four kernels from RBF χ^2 kernel function $\kappa(\mathbf{x}, \mathbf{x}') = \exp\left((\sum_i((2x_i x_i')/(x_i + x_i'))/2\sigma^2)\right)$ on histogram features. Note that, when using RBF χ^2 kernel, the PCA step is ignored and the normalization parameter σ is empirically set to 1 for spHist and 10 for other features. The Re-Id accuracy of each kernel and the ensemble kernel are shown in Fig. 9. In addition, we compare the performance of ensemble kernel learned via mkLFDA to three different MKL methods (i.e., arithmetic average method, geometric average method, and kernel alignment method) and kLFDA with feature concatenation. Also we conduct a comparative experiment by adopting CNN feature extracted via AlexNet, which has been pretrained on ImageNet and fine-tuned on each dataset, respectively.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

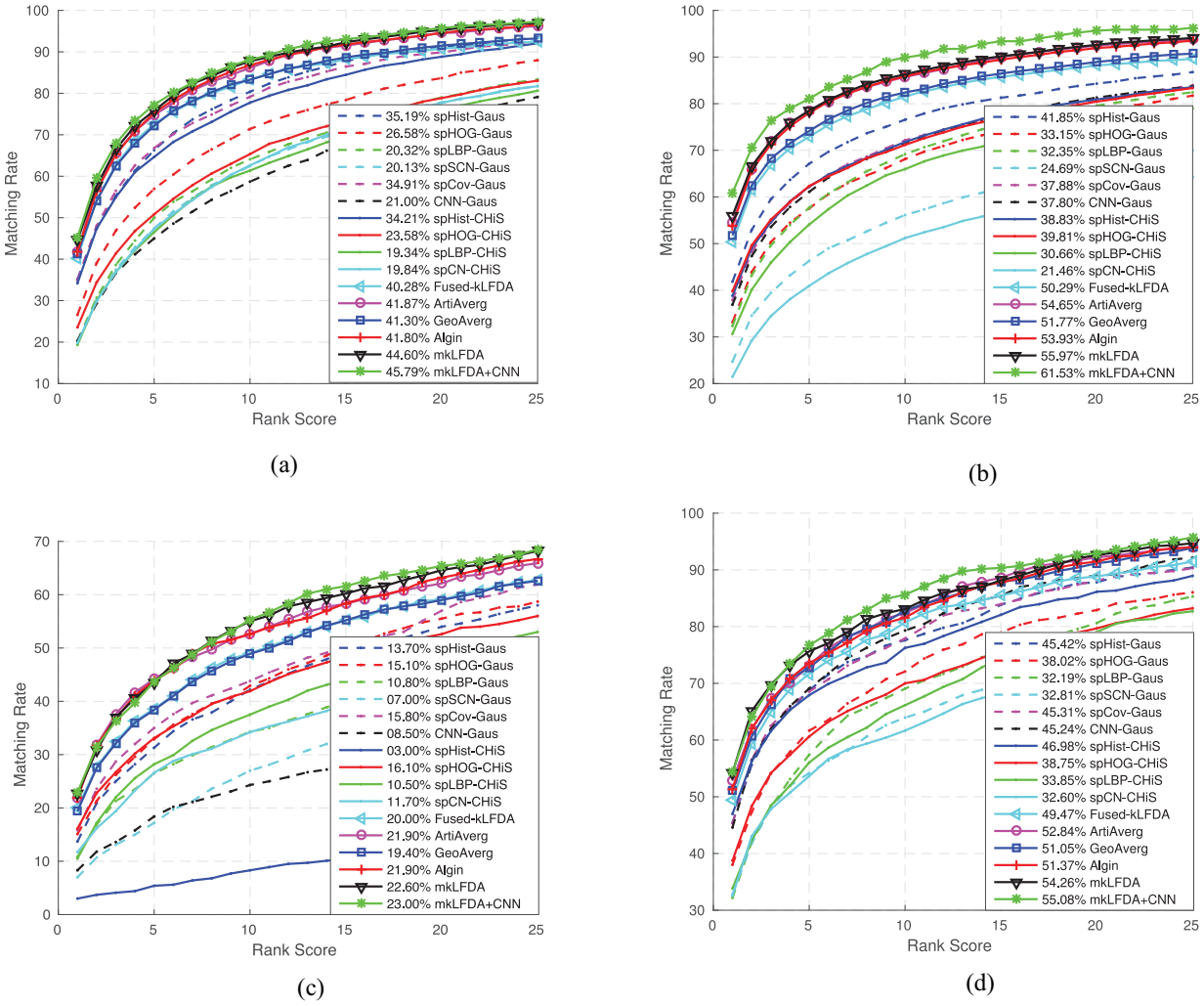


Fig. 9. Results on different datasets reported as averaged CMC curves. (a)–(d) Comparisons between basis kernels and ensemble kernels on four benchmarks are shown, respectively.

1) Arithmetic Average Method: The ensemble kernel is simply constructed by the arithmetic average of all kernels, i.e., $\kappa(\mathbf{x}, \mathbf{x}') = (1/P)\sum_{p=1}^{P}\kappa^{(p)}(\mathbf{x}, \mathbf{x}')$.

2) Geometric Average Method: The ensemble kernel is simply constructed by the geometric average of all kernels, i.e., $\kappa(\mathbf{x}, \mathbf{x}') = \sqrt[P]{\prod_{p=1}^{P}\kappa^{(p)}(\mathbf{x}, \mathbf{x}')}$.

3) Kernel Alignment Method: Define the kernel alignment between two kernels as $A(\kappa_i, \kappa_j) = (\langle \mathbf{G}_i, \mathbf{G}_j\rangle_F / \sqrt{\langle \mathbf{G}_i, \mathbf{G}_i\rangle_F \langle \mathbf{G}_j, \mathbf{G}_j\rangle_F})$, where $\mathbf{G}$ is the corresponding Gram matrix. The combination weight of each kernel is determined by its kernel alignment with respect to the target kernel $\hat{\kappa}$, i.e., $\beta_p = (A(\kappa^{(p)}, \hat{\kappa}) / \sum_{p=1}^{P} A(\kappa^{(p)}, \hat{\kappa}))$. In this paper, we set the element value in the target Gram matrix as 1 if the corresponding data pair is in the same class, and 0 otherwise.

4) mkLFDA: In our experiment, instead of initialize $\mathbf{A}$, we initialize the parameter $\mathbf{A}\mathbf{A}^T$ as an identity matrix. Although we are not able to provide the theoretical guarantee for convergence of mkLFDA, we observe that the alternating algorithm converges rapidly in only a few iterations in experiments.

According to the experimental results on the four benchmark datasets as in Fig. 9, we observe that the ensemble kernel improves the performance significantly, even just using the simplest average combination. Compared to the other three MKL algorithms and kLFDA, mkLFDA obtains a better Re-Id accuracy. Particularly, the rank 1 matching rate is boosted to 44.60% on VIPeR, to 55.97% on CUHK01, to 22.60% on PRID2011, and to 54.26% on 3DPeS, respectively. We also observe that CNN features perform comparably with our hand-crafted features, but can further boost the performance when fused with our approach. For example, by adopting the CNN feature into our mkLFDA framework, the rank 1 matching rate on dataset CUHK01 has been improved from 56.0% to 61.5%.

F. Comparison to the State-of-the-Art

To demonstrate the effectiveness of the proposed framework on person Re-Id task, we also compare our approach to many existing methods. These methods can be roughly divided into six groups.

1) Multifeature fusion (Ensemble [72], LateFusion [73], and ELF [74]).
2) Feature extraction/learning (MidFilter [22], MTL [26], Transfer [27], SCNCD [23], ExplicitPoly [21],

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VII
TOP r MATCHING ACCURACY ON VIPER USING OUR APPROACH AND STATE-OF-THE-ART METHODS

Methods	Rank1	Rank5	Rank10	Rank20	Rank50
mkLFDA+CNN	**45.8**	**77.3**	**88.4**	**95.7**	**99.9**
mkLFDA	44.6	75.8	87.7	95.3	**99.5**
Ensemble [72]	**45.9**	**77.5**	**88.9**	**95.8**	**99.5**
LateFusion [73]	30.2	51.6	62.4	73.8	-
ELF [74]	12.0	31.0	41.0	58.0	-
MidFilter [22]	43.4	73.0	85.0	93.7	-
MTL [26]	42.3	72.2	81.6	89.6	-
Transfer [27]	41.6	71.9	86.2	95.1	-
SCNCD [23]	37.8	68.5	81.2	90.4	97.0
ExplicitPoly [21]	36.8	70.4	83.9	91.7	97.8
SalMatch [6]	30.2	52.0	65.5	79.2	-
Salience [16]	26.7	50.7	62.4	76.4	-
ColorInv [19]	24.2	-	57.1	69.7	87.0
ViewInv [20]	21.4	45.9	62.6	79.7	-
eBiCov [75]	20.7	42.0	56.2	68.0	-
SDALF [5]	19.9	38.9	49.4	65.7	92.2
Attribute [28]	17.4	39.0	50.8	-	86.4
MLAPG [33]	40.7	-	82.3	92.4	-
XQDA [7]	40.0	-	80.5	91.1	-
KernelML [15]	36.1	68.7	81.3	91.1	-
RMLLC [34]	31.3	62.1	75.3	86.7	-
LADF [9]	30.0	65.0	79.0	91.0	98.0
LAFT [18]	29.6	-	69.3	-	96.8
MtMCML [35]	28.8	59.3	75.8	88.5	-
RPLM [36]	27.0	-	69.0	83.0	95.0
FuncSpace [39]	25.8	-	69.6	83.7	95.1
LFDA [11]	24.2	-	67.1	-	94.1
KISSME [8]	22.0	-	68.0	-	93.0
PCCA [10]	19.6	48.9	64.9	80.3	-
RDC [12]	18.3	42.7	57.8	72.4	-
ImprDeep [29]	34.8	63.0	76.0	-	-
CPDL [45]	34.0	64.2	77.5	88.6	-
ISR [44]	27.0	-	61.0	72.0	94.1
SSCDL [17]	25.6	53.7	68.1	83.6	-
MirrorRep [49]	43.0	75.8	87.3	94.8	-
CSL [48]	34.8	68.7	82.3	91.8	96.2
CompTemp [31]	24.0	47.0	60.0	75.0	-

TABLE VIII
TOP r MATCHING ACCURACY ON CUHK01 USING OUR APPROACH AND STATE-OF-THE-ART METHODS

Methods	Rank1	Rank5	Rank10	Rank20	Rank50
mkLFDA+CNN	**61.5**	**81.7**	**90.0**	**96.3**	**99.1**
mkLFDA	**56.0**	**78.5**	**86.3**	**92.6**	**97.4**
Ensemble [72]	53.4	76.4	84.4	90.7	96.4
MidFilter [22]	34.3	55.1	65.0	74.9	-
Transfer [27]	31.5	52.5	65.8	77.6	-
SalMatch [6]	28.5	46.0	56.0	-	-
ImprDeep [29]	47.5	72.0	80.0	-	-
DeepReid [30]	27.9	64.0	77.0	88.0	-
MirrorRep [49]	40.4	64.6	75.3	84.1	-

TABLE IX
TOP r MATCHING ACCURACY ON PRID2011 USING OUR APPROACH AND STATE-OF-THE-ART METHODS

Methods	Rank1	Rank5	Rank10	Rank20	Rank50
mkLFDA+CNN	**23.0**	**43.7**	**55.4**	**65.4**	**78.9**
mkLFDA	**22.6**	**43.5**	**55.0**	64.6	77.5
Ensemble [72]	17.90	39.0	49.0	62.0	-
MidFilter [22]	12.5	23.9	30.7	36.5	51.6
MTL [26]	18.0	37.4	50.1	**66.6**	**82.3**
SalMatch [6]	4.9	17.5	26.1	33.9	47.8
RPLM [36]	15.0	-	42.0	54.0	70.0
PCCA [10]	3.5	10.9	17.9	27.1	45.0
PRDC [37]	4.5	12.6	19.7	29.5	46.0

TABLE X
TOP r MATCHING ACCURACY ON 3DPES USING OUR APPROACH AND STATE-OF-THE-ART METHODS

Methods	Rank1	Rank5	Rank10	Rank20	Rank50
mkLFDA+CNN	**55.1**	76.9	**86.2**	**93.0**	**99.7**
mkLFDA	54.3	75.6	83.1	92.6	**99.3**
Ensemble [72]	53.3	77.0	85.0	92.0	-
kernelML [15]	54.0	**77.7**	86.0	92.4	-
LFDA [11]	33.4	-	70.0	-	95.1
PCCA [10]	41.6	70.5	81.3	90.4	-
rPCCA [15]	47.3	75.0	84.5	91.9	-
CSL [48]	**57.9**	**81.1**	**89.5**	**93.7**	-

TABLE XI
TIME COST ANALYSIS OF OUR PROPOSAL ON VIPER

Feature Extraction Phase					Matching Phase	
spHist	spHOG	spLBP	spCN	spCov	training	testing
∼90ms	∼80ms	∼260ms	∼60ms	∼150ms	∼20min	∼260ms

SalMatch [6], Salience [16], ColorInv [19], ViewInv [20], eBiCov [75], SDALF [5], and Attribute [28]).

3) Metric learning (MLAPG [33], XQDA [7], KernelML [15], RMLLC [34], LADF [9], LAFT [18], MtMCML [35], RPLM [36], FuncSpace [39], LFDA [11], KISSME [8], PCCA [10], RDC [12], and PRDC [37]).
4) Deep learning (ImprDeep [29] and DeepReid [30]).
5) Dictionary learning and sparse representation (CPDL [45], ISR [44], and SSCDL [17]).
6) Others (MirrorRep [49], CSL [48], and CompTemp [31]).

The matching rates are directly cited from tables or figures provided in the relevant literatures.

We present the comparison results in Tables VII–X, while marking the first (bold red) and second (bold blue) best performance out. We observe that our approach obtains comparable performance with respect to the state-of-the-art methods on all the four datasets, and even surpasses some of them.

Specifically, VIPeR has been tested with dozens of methods so far. Our approach achieves the second-best result on it, with only a narrow gap with the best one. On the medium dataset CUHK01, we get state-of-the-art result compared to all the others, including multifeature fusion and deep leaning methods, e.g., 8.1% and 14.0% rank 1 accuracy have been improved separately. Besides, if the gallery and probe images are selected randomly from the dataset without considering the camera label, CPDL [45] can get a 59.47% rank 1 accuracy on CUHK01. While under the same partition scheme, our approach can easily get a higher rank 1 accuracy as 62.59%. On the more practical dataset PRID2011, our method still achieves state-of-the-art performance, and boosts the performance significantly, e.g., 5.1% improvement at rank 1. As the images in dataset 3DPeS are very irregular and there does not exist a single stable mapping between pairwise images, using only kLFDA in our method did not catch up with state-of-the-art. However, our approach is still comparable with or better than other feature fusion, feature extraction/learning, and deep learning-based methods.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

G. Analysis of Time Cost

Although variant spatial pyramid statistical features and features fusion strategies are used in the framework, the execution efficiency is still acceptable for Re-Id task. Table XI shows that, taking VIPeR for example, the most time consuming phase is training mkLFDA model over the whole training dataset.[3] Once the model is trained, each image can be reidentified quickly within 1 second.

V. Conclusion

We have proposed and evaluated a unified framework to utilize spatial pyramid-based statistical features for person Re-Id. In this framework, three types of image statistical characteristics are computed conveniently, including histogram distribution, mean vector, and covariance matrix. Specifically, we have implemented five spatial pyramid based statistical features, i.e., spHist, spHOG, spLBP, spCN, and spCov, and combined them via mkLFDA. Extensive evaluations on benchmark datasets have shown that using abundant input primary channels, appropriate encoding methods, proper contrast normalization schemas, and spatial pyramid-based hierarchical features aggregation are all beneficial to improve the performance. Experiment results have demonstrated that our approach could catch up with state-of-the-art or even significantly improve the best performance on some practical Re-Id scenarios. We hope that this paper could provide useful guidelines for researchers and practitioners to deploy well performed Re-Id system for real world applications.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions for further improve the quality of this paper.

References

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*, vol. 1. London, U.K.: Springer, 2014.

[2] S. E. Budge, J. A. Sallay, Z. Wang, and J. H. Gunther, "People matching for transportation planning using texel camera data for sequential estimation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 3, pp. 619–629, May 2013.

[3] J. García, A. Gardel, I. Bravo, J. L. Lázaro, and M. Martínez, "Tracking people motion based on extended condensation algorithm," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 3, pp. 606–618, May 2013.

[4] V. Bruni and D. Vitulano, "An improvement of kernel-based object tracking based on human perception," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1474–1485, Nov. 2014.

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

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

[7] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 2197–2206.

[3]Our code is implemented using MATLAB in a Linux system, and the running time is recorded on an ordinary PC with 8GB memory and 3.10 GHz basic frequency.

[8] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2288–2295.

[9] Z. Li *et al.*, "Learning locally-adaptive decision functions for person verification," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 3610–3617.

[10] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2666–2672.

[11] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 3318–3325.

[12] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[13] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust face recognition for uncontrolled pose and illumination changes," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 1, pp. 149–163, Jan. 2013.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, San Diego, CA, USA, 2005, pp. 886–893.

[15] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. ECCV*, Zürich, Switzerland, 2014, pp. 1–16.

[16] R. Zhao, W. L. Ouyang, and X. G. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 3586–3593.

[17] X. Liu *et al.*, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 3550–3557.

[18] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 3594–3601.

[19] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[20] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1095–1108, May 2015.

[21] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 1565–1573.

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

[23] Y. Yang *et al.*, "Salient color names for person re-identification," in *Proc. ECCV*, Zürich, Switzerland, 2014, pp. 536–551.

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

[25] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. CVPR*, New York, NY, USA, 2006, pp. 1528–1535.

[26] C. Su *et al.*, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 3739–3747.

[27] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 4184–4193.

[28] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. BMVC*, 2012, pp. 24.1–24.11.

[29] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. CVPR*, 2015, pp. 3908–3916.

[30] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 152–159.

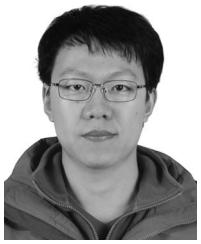
[31] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human re-identification by matching compositional template with cluster sampling," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 3152–3159.

[32] J. Si, H. Zhang, and C.-G. Li, "Person re-identification via region-of-interest based features," in *Proc. VCIP*, Valletta, Malta, 2014, pp. 249–252.

[33] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 3685–3693.

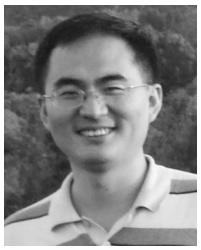
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[34] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting listwise similarities," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4741–4755, Dec. 2015.
[35] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.
[36] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. ECCV*, Florence, Italy, 2012, pp. 780–793.
[37] W.-S. Zheng, S. G. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, 2011, pp. 649–656.
[38] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. BMVC*, Aberystwyth, U.K., 2010, pp. 21.1–21.11.
[39] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Re-identification in the function space of feature warps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1656–1669, Aug. 2015.
[40] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based kiss metric learning," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 242–252, Feb. 2015.
[41] J. Si, H. Zhang, and C.-G. Li, "Regularization in metric learning for person re-identification," in *Proc. ICIP*, Québec City, QC, Canada, 2015, pp. 2309–2313.
[42] W.-S. Zheng *et al.*, "Partial person re-identification," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 4678–4686.
[43] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 4516–4524.
[44] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, Aug. 2015.
[45] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proc. IJCAI*, Buenos Aires, Argentina, 2015, pp. 2155–2161.
[46] X.-Y. Jing *et al.*, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 695–704.
[47] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised ℓ_1 graph learning," in *Proc. ECCV*, Amsterdam, The Netherlands, 2016, pp. 178–195.
[48] Y. Shen *et al.*, "Person re-identification with correspondence structure learning," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 3200–3208.
[49] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. IJCAI*, 2015, pp. 3402–3408.
[50] L. Wei, Y. Tian, Y. Wang, and T. Huang, "Swiss-system based cascade ranking for gait-based person re-identification," in *Proc. AAAI*, Austin, TX, USA, 2015, pp. 1882–1888.
[51] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 3765–3773.
[52] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. ECCV*, Zürich, Switzerland, 2014, pp. 688–703.
[53] W.-S. Zheng, S. G. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2650–2657.
[54] A. J. Ma, J. Li, P. C. Yuen, and P. Li, "Cross-domain person reidentification using domain adaptation ranking SVMs," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1599–1613, May 2015.
[55] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 3567–3574.
[56] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *Proc. ECCV*, Zürich, Switzerland, 2014, pp. 330–345.
[57] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 1116–1124.
[58] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.
[59] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, New York, NY, USA, 2006, pp. 2169–2178.
[60] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. BMVC*, 2011, pp. 76.1–76.12.
[61] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. BMVC*, London, U.K., 2009, pp. 91.1–91.11.
[62] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[63] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.
[64] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. ICCV*, Kyoto, Japan, 2009, pp. 32–39.
[65] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ICME*, Florence, Italy, 2010, pp. 1469–1472.
[66] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.
[67] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.
[68] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. Workshop PETSW*, vol. 3. no. 5, Rio de Janeiro, Brazil, 2007, pp. 41–47.
[69] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. ACCV*, Daejeon, South Korea, 2012, pp. 31–44.
[70] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. SCIA*, Ystad, Sweden, 2011, pp. 91–102.
[71] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPeS: 3D people dataset for surveillance and forensics," in *Proc. ACMMW*, 2011, pp. 59–64.
[72] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 1846–1855.
[73] L. Zheng *et al.*, "Query-adaptive late fusion for image search and person re-identification," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 1741–1750.
[74] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, Marseilles, France, 2008, pp. 262–275.
[75] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. BMVC*, Guildford, U.K., 2012, pp. 57.1–57.11.



Jianlou Si received the B.E. degree from Jilin University, Jilin, China, in 2012. He is currently pursuing the Ph.D. degree with the Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing, China.

His research interests include computer vision, pattern recognition, and machine learning.



Honggang Zhang (SM'12) received the B.S. degree from Shandong University, Jinan, China, in 1996, and the master's and Ph.D. degrees from the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1999 and 2003, respectively.

He was a Visiting Scholar with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, from 2007 to 2008. He is currently an Associate Professor and the Director of Web Search Center, BUPT. His current research interests include image retrieval, computer vision, and pattern recognition. He has published over 60 technical papers in his fields.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Chun-Guang Li (S'05–M'11) received the B.E. degree in telecommunication engineering from Jilin University, Jilin, China, in 2002, and the Ph.D. degree in signal processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2007.

He is currently an Associate Professor with the School of Information and Communication Engineering, BUPT. From 2011 to 2012, he visited the Visual Computing Group, Microsoft Research Asia, Beijing. From 2012 to 2013, he visited the Vision, Dynamics, and Learning Lab, Johns Hopkins University, Baltimore, MD, USA. His current research interests include statistical signal processing and machine learning.

Dr. Li is a member of the ACM and CCF.

Jun Guo received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1982 and 1985, respectively, and the Ph.D. degree from Tohuku Gakuin University, Sendai, Japan, in 1993.

He is currently a Professor and the Vice President of BUPT. His current research interests include pattern recognition theory and application, information retrieval, content-based information security, and network management. He has published over 100 technical papers in his fields.