Learning Invariant Color Features for Person Reidentification

Rahul Rama Varior, *Student Member, IEEE*, Gang Wang, *Member, IEEE*, Jiwen Lu, *Senior Member, IEEE*, and Ting Liu, *Student Member, IEEE*

***Abstract*— Matching people across multiple camera views known as person reidentification is a challenging problem due to the change in visual appearance caused by varying lighting conditions. The perceived color of the subject appears to be different under different illuminations. Previous works use color as it is or address these challenges by designing color spaces focusing on a specific cue. In this paper, we propose an approach for learning color patterns from pixels sampled from images across two camera views. The intuition behind this work is that, even though varying lighting conditions across views affect the pixel values of the same color, the final representation of a particular color should be stable and invariant to these variations, i.e., they should be encoded with the same values. We model color feature generation as a learning problem by jointly learning a linear transformation and a dictionary to encode pixel values. We also analyze different photometric invariant color spaces as well as popular color constancy algorithm for person reidentification. Using color as the only cue, we compare our approach with all the photometric invariant color spaces and show superior performance over all of them. Combining with other learned low-level and high-level features, we obtain promising results in VIPeR, Person Re-ID 2011, and CAVIAR4REID data sets.**

***Index Terms*— Person re-identification, illumination invariance, photometric invariance, color features, joint learning.**

I. Introduction

MATCHING pedestrians across multiple CCTV cameras have gained a lot of interest in recent years. Despite several attempts by many researchers, [1]–[3], it largely remains challenging mainly due to the following reasons. First, the images are captured under different lighting conditions. Therefore the perceived color of the subject appears to be different with respect to the illumination. Second, from surveillance cameras, no biometric aspects are available [2]. Third, most often, the surveillance cameras will be of lower resolution [1]. Figure 1 shows some examples of images from different datasets.

Fig. 1. Some examples from CAVIAR4REID, VIPeR and Person Re-ID 2011 datasets. Images clearly show the appearance changes due to the environmental variables - Illumination, Shading, Camera view angle. **Best viewed in color**.

Modern person re-identification systems primarily focus on two aspects. (1) A feature representation for the probe and gallery images and (2) a distance metric to rank the potential matches based on their relevance. In the first category, majority of the works concentrate on designing low level features. Since each of the features capture different aspects of the images, usually a combination of these features are used to obtain a richer signature. In the second category, the person re-identification is formulated as a ranking or a metric learning problem. The proposed work belongs to the first category in which the focus is on learning robust feature representations, specifically invariant color based features.

Color based features have been proven to be an important cue for person re-identification [1]. An interesting insight on the importance of color features was demonstrated in an experiment conducted by Gray and Tao [4]. They used AdaBoost for giving weights to the most discriminative features and observed that, over 75% of the classifier weights were given to color based features. These observations support the fact that color has to be given much more attention than other handcrafted features (i.e., engineered features with selected statistical measures and parameters) based on shape, texture and regions. But due to the illumination variations across the camera views, the perceived color of same parts for a particular person appear to be different. Taking this observation into consideration and as validated from our experiments, we suggest that using color features *as it is*, i.e., the RGB, HSV or YUV color histograms will not be adequate to achieve

Manuscript received February 15, 2015; revised August 3, 2015 and November 9, 2015; accepted December 8, 2015. Date of publication February 18, 2016; date of current version June 7, 2016. This work was supported in part by the Rapid-Rich Object Search Laboratory within the Interactive Digital Media Strategic Research Programme through the National Research Foundation, Singapore, in part by the Singapore Ministry of Education Tier 2 under Grant ARC28/14, and in part by the Agency for Science, Technology and Research, Singapore, within the Science and Engineering Research Council under Grant PSF1321202099. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Pierre-Marc Jodoin. *(Corresponding author: Gang Wang.)*

R. Rama Varior, G. Wang, and T. Liu are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: rahul004@e.ntu.edu.sg; wanggang@ntu.edu.sg; liut0016@e.ntu.edu.sg).

J. Lu is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: lujiwen@mail.tsinghua.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/TIP.2016.2531280

a stable and invariant color representation. Hence we propose a multilayer feature learning framework that learns invariant color features from raw pixel values as opposed to histogram or other color based handcrafted features.

The proposed framework aims at learning stable color feature representations for images from both camera views by transforming the pixels to an invariant space. Since the images are captured under different unknown lighting conditions, modeling accurate transformations for each of these illumination settings is practically impossible. Existing algorithms to achieve color-constancy rely on strong assumptions about the statistics of color distribution in the image. Hence, these algorithms can project the image only to an approximate color constant space. Further, it should be noted that, no prior information regarding the illumination is given in any of the existing datasets for person re-identification. Therefore, to obtain invariant color features, we propose a feature learning framework to explore the structures and patterns inherent in the image pixels. In the learning step, we take advantage of pairs of pixel values from image pairs and enforce an invariance condition while jointly learning a transformation and encoding scheme. An auto-encoder based framework transforms the 3-dimensional RGB pixel values to a higher dimensional space first and encode them using a dictionary which maps these pixels to an invariant space. These encoded values are pooled over a region and concatenated to form the final representation of an image. This framework can be extended to a multilayer structure for learning complementary features at a higher levels.

Experiments were conducted on a synthetic color constancy dataset [5] and publicly available person re-identification datasets such as VIPeR [6], Person Re-ID 2011 [7] and CAVIAR4REID [2]. From the results, it can be inferred that (1) by learning invariant color features, significant improvement can be achieved in the results when compared with the traditional color histograms and other handcrafted features; and (2) when combined with other types of learned low-level (i.e., the first order information at the pixel level) and high-level features (i.e., more abstract concepts that usually refer to attributes or object classes), it can achieve promising results in several benchmark datasets.

In summary, the contributions of our work are as follows.

- We propose a novel approach for learning inter camera invariant color feature representations from the pixels sampled from matching pair of images. In contrast with the previous works such as color histograms and normalized color spaces, our approach is more robust and efficient in representing the color features.
- We propose a joint learning framework which solves the coupled problem of learning a linear auto-encoder transformation and a dictionary to encode invariant color based features.
- We show that color as a single cue can bring a good performance that beats several hand-engineered features designed for person re-identification and when combined with other types of learned low-level and high-level features, it can achieve promising performance in several challenging datasets.

The rest of this paper is organized as follows. Section II reviews some of the related works in color constancy, person re-identification and feature learning. Section III describes the motivation and the major contributions of this work. Section IV describes the framework for learning the invariant color features. In section V, we demonstrate the experimental evaluation of our method and compare with the other competing methods for person re-identification. In section VI, we perform an analysis of the obtained results and section VII concludes this paper.

II. Related Work

A. Color Constancy

The human perceptual system has the ability to ensure that the perceived color of an object remains relatively constant even under varying illumination [8]. Land and McCann proposed the Retinex theory [9] to explain this perceptual effect. The central idea behind retinex algorithm is that, the human visual system functions with three independent cone systems peaking on the long, middle and short wavelengths of the visible spectrum and the images formed by these receptors are compared to generate the color sensation [9]. In one of the pioneering works [10], Forsyth proposed the CRULE and MWEXT algorithms to achieve color constancy in Mondriaan world images (i.e., flat and frontally presented image consisting of numerous colored patches) by estimating the illuminant based on the information obtained from images such as reflectances and possible light sources. For a detailed overview of the color-constancy algorithms derived from the Retinex theory and [10], we refer the reader to [11]–[13].

All of the aforementioned works and the derived works are based on strong assumptions since the color constancy is an under-constrained problem [10], [13]. For example, in [10], the main assumption is constrained gamuts, i.e., the limited number of image colors which can be observed under a specific illuminant. Several other assumptions were based on the distribution of colors that are present in an image (e.g., White-Patch, Gray-World and Gray-Edge). In the Gray-World algorithm [14], the major assumption is that, the average reflectance in a scene is achromatic. In the White-Patch algorithm [15], the main assumption was that, the maximum response in the RGB-channels corresponds to a scene point with perfect reflectance. The majority of these works vary in their assumptions and therefore, no color constancy algorithm can be considered as universal.

Color features also gained a lot of interest in object recognition. In one of the earliest works, Swain and Ballard [16] identified that color histograms were stable representations over change in views. Funt and Finlayson [17] used ratios of colors from neighbouring location to achieve some illumination invariance. Gevers and Smeulders [18] analysed different color spaces to achieve invariance to a substantial change in viewpoint, object geometry and illumination. But it was observed that the object recognition accuracy degrades substantially for all of the color spaces with a change in illumination color. The invariance property achieved by them has become the basis of several works in tracking, segmentation [19] and image retrieval [20].

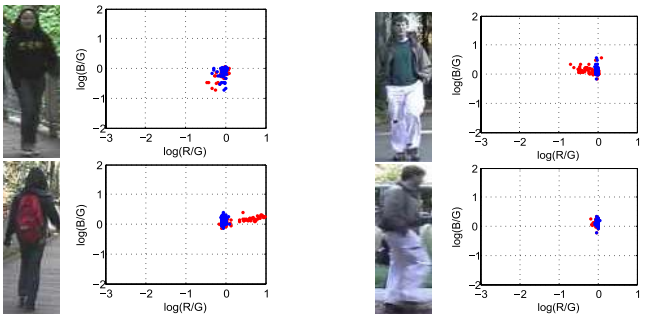


Fig. 2. Two examples to illustrate the difference in color clouds for the same person in different views due to appearance change in the second view. Red dots indicate the $\log\frac{R}{G}$ vs $\log\frac{B}{G}$ values of pixels sampled from upper part of the body and blue dots indicate those for the lower part of the body. In (a), due to the presence of additional object (bag), the shape of the color clouds vary significantly. In (b), the color of the upper body has changed across the views which result in a different shape for the corresponding color clouds as shown in the plot. **Best viewed in color**.

A Diagonal Matrix Transform (DMT) is the basis of majority of the works [1], [18], [21], [22] on color constancy. But the DMT has been proven to be a suboptimal method for achieving color constancy by West and Brill [23] and D'Zmura and Lennie [24]. To improve the performance of the DMT, spectral sharpening [25] derived for each camera can be incorporated. Zaidi [26] proposed a two parameter affine transformation to generate an illumination independent color descriptors. Funt and Lewis [27] observed no improvement in the affine model compared to the DMT with spectral sharpening. However, Finlayson *et al.* [28] suggest that a generalized diagonal transformation is sufficient to achieve color constancy. Color Eigenflows [29] provides a different approach by developing a statistical linear model that describes how colors change jointly under typical photic parameter changes. However, in this paper, the main aim is to develop a stable *feature representation* for color features which is robust to varying lighting conditions. We would like to point out that the proposed method does not aim at estimating the illuminant or correcting the original image captured under a different illumination.

Berwick and Lee [21] proposed a log-chromaticity color space to achieve specularity, illumination color and illumination pose invariance. A recent work [1] makes use of the log-chromaticity color space to achieve color-constancy upto translation for person re-identification. The assumption in [1] is that the shape of the color cloud is sufficiently preserved in the $\log\frac{R}{G}$, $\log\frac{B}{G}$ space. But this assumption is violated in many real world images. The color cloud is formed based on sampled observations from upper body and lower part of the body. Fig.2 shows some examples of those observations from the VIPeR dataset. Relying on the color-clouds can be error prone since in a different view, the upper part of the body may have different colors for the same subject. In this paper, feature learning techniques are used to discover a color constant space in contrast to methodologies based on assumptions.

B. Person Re-Identification

Research in person re-identification has taken a giant leap in the recent years. As mentioned in section I, existing works focus on the different steps that need to be taken for dealing with this problem. The majority of the works [2]–[4], [30], [31] predominantly focus on the first step, i.e., designing features based on texture, color, shape, regions and interest points. Since the primary focus of this work is on color based feature design, a complete evaluation of all of the aforementioned features is beyond the scope of this paper. To obtain the global chromatic content, most of the works use the color histogram features in the RGB, HSV or YUV space. These color spaces do not possess the property of illumination invariance. In addition to a weighted HSV histogram, the Maximally Stable Color Regions (MSCR) are also used in [30] to obtain the per-region color displacement.

The Salient Color Name based color descriptor was proposed for person re-identification in [32]. Each pixel was represented as a vector of 16 color names and the proposed color based descriptor was combined with several other standard color spaces to achieve photometric invariance. In [33], different color models were evaluated for person re-identification and a new color space, $[g1, g2, g3]$ was proposed based on the log-chromaticity color space which was used in [1]. In contrast to the above methods, the proposed work aims at learning a feature representation and encoding scheme which can extract stable structures and patterns inherent in the pixels. In addition to that, the proposed method relies on the regularities and patterns in the data and it does not depend directly on the diagonal model or diagonal-offset model which is the basis of several illumination correction algorithms.

In another relatively closer work, Porikli [34] and Javed *et al.* [35] proposed a Brightness Transfer Function (BTF) to find a transformation that maps the appearance of an object in one camera view to the other. But it should be noted that, the system has to be re-trained each time when the illumination changes. In addition to that, the method adopts normalized histograms of object brightness values for the BTF computation. Therefore, a pixel level correspondence cannot be achieved. A weighted BTF that combines different BTFs computed on several low-level features was proposed in [36] for person re-identification.

It is important to note that all the aforementioned features for person re-identification are handcrafted focusing on specific cues. However, in this work we propose a model to learn color based features for a pair of camera to achieve invariance across the two different views based on the intuition that the features should be stable and robust to varying lighting conditions. To the best of our knowledge, this is the first work that focuses on learning of low-level inter-camera invariant color features using feature learning techniques for person re-identification. Since the framework is based on learning from the data, the scope of this work can be beyond person re-identification.

C. Feature Learning

Recent researches have shown a growing interest in unsupervised feature learning methods such as auto-encoders [37], [38], sparse coding [39] and Deep Belief Nets [40] since they can be generalized to a larger extent.

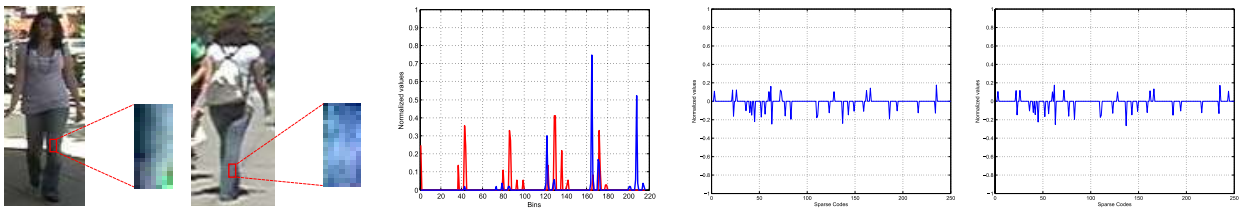


Fig. 3. (a) and (b) Training image pairs from view A and view B and patches sampled from them. (c) Overlapping RGB histogram (216 dimensional) of the extracted patches. Red for patch from view A image and blue for the one from view B. (d) and (e) shows the result of the encoding using our approach. It can be seen that, using our formulation, the obtained encoding is very close to each other for corresponding pixels in the sampled patch. We learn the transformation and dictionary jointly so that the encoding is same for the two sampled patches. **Best viewed in color.**

Since no handcrafted feature can be considered as universal, learning relations from data can be advantageous.

Modeling complex distributions and functions have been a bottleneck in machine learning. Recent studies in deep learning indicate that such deep architectures can efficiently handle these challenges and have shown that better generalization can be obtained. Several successful algorithms have been proposed [40]–[44] to train large networks such as deep belief networks and stacked auto-encoders. The intuition behind using large networks is that, to learn a complex function that computes the output from input, automatically learning features at multiple layers of abstraction can help to a large extend [44]. Additionally, biological evidences substantiate that in the visual cortex, recognition happens at multiple layers [45].

Feature learning algorithms have been proposed for person re-identification in [46]–[48]. Yi *et al.* [46] propose a deep metric learning framework based on Convolutional Neural Networks for person re-identification. Li *et al.* [47] proposed a Filter Pairing Neural Network for handling geometric and photometric transformation across views. To achieve cross-view invariance, Zhao *et al.* [48] proposed to learn mid-level filters. However, all the above works focus on learning edges at different orientations in the first layer and higher level patterns in the further layers. Learned color features have been previously proposed for image retrieval [49], [50] and visual tracking [51]. However, they do not consider any invariant projections for the pixels. In [49], a color codebook of a set of colors is constructed by selecting and clustering characteristic colors from real-world images. Further a color histogram is computed from the color palette. But, the representation is not obtained at a pixel level and they do not consider any invariance which is necessary for a color constant representation. This paper addresses the problem of learning color features from data. We also propose a new method to learn an invariant transformation and encoding simultaneously at a pixel level to obtain a color constant representation.

Encoding of the local descriptors is an extensive topic of research. The bag of features model (BoF) [52], spatial pyramid matching (SPM) [53] have been pioneering among these algorithms and has achieved state-of-the-art in many benchmark datasets. An extension of the SPM has been proposed in [39] where they make use of sparse coding to achieve a better quantization of the local descriptors and they have shown state-of-the-art results. Following the work, several sparse coding algorithms such as the LCC [54] and the LLC [55] have been proposed. Sparse coding has also been proven to be effective in several other computer vision applications such as visual tracking [56] and action recognition [57]. In the proposed work, we make use of a sparse coding approach as used in [39].

III. Motivation

The importance of color features for person re-identification has been proven in [4] and several works [1], [32], [33] have addressed the photometric variations which is a major challenge in person re-identification. In the existing literatures addressing the photometric variations of the images, histograms in specially designed color spaces are computed. These color spaces have the property of photometric invariance to some extent. But the existing methods for achieving color constancy are based on several assumptions about the statistics of color distribution, surfaces and its reflectance properties. Hence, a histogram representation in such a *weakly* corrected color space will not be robust enough to achieve invariance.

Hand-engineered features focus on particular cues and add more complexity to the system. Ideally, for each illumination setting, a linear transformation is required to transform it to a canonical illuminant space. Since there is no prior information regarding the illumination for any of the images, estimating accurate linear transformation for all different illumination settings is practically impossible. As it is tedious to compute or learn multiple linear transformations, we adopt one of the prominent illumination correction methods [58] as a preprocessing step to transform the images into a *weak* illumination invariant space. Figure 3 shows an example of a pair of images corrected by the aforementioned method and a pair of patches sampled from the corresponding parts of the body. It can be observed that, even in the corrected space, the color histogram (as shown in figure 3 (c)) does not yield an invariant representation.

A robust representation should capture a certain amount of *information* which are the stable structures and patterns in the observed input. Though pixel values are affected by different illumination settings, it can be reasonably assumed that there exists a space where the color patterns are invariant to these variations in illumination. Input pixels captured under an unknown illuminant will lie away from a color constant

space and the objective of color constancy algorithms is to project the pixels back to the color constant space. Therefore as mentioned before, an existing color constancy algorithm is initially used to bring the pixel values *closer* to the color constant space. Since the existing color constancy algorithms are weak correction algorithms, the pixel values require further processing to achieve a robust representation which is stable under varying lighting conditions. Therefore, we propose to learn a transformation or a set of filters to extract stable structures and patterns which bring pixels of corresponding colors to be closer to each other. Additionally, previous works [28], [59] suggest that, to achieve color constancy, linear transformations are sufficient unless camera auto-gain control and transducer non-linearities are taken into consideration. Therefore, we use a linear auto-encoder with an invariance constraint to discover such invariant patterns from the data with the objective that in the transformed space, the representation for pixels of same colors should be as close as possible.

Many researchers have empirically found that encoding schemes for quantizing the local descriptors are essential for good performance. Sparse coding [39], [54] has been found to achieve state-of-the-art performance for many classification problems. In this work, a sparse coding technique is adopted to encode each pixel by a set of dictionary elements or codebooks. It was proven that sparse coding can be used to obtain robust representations that can capture the salient properties of the descriptors [39]. Additionally, biological evidences [60] show the plausibility of sparse coding principle in encoding the sensory inputs in mammalian cortex.

For a consistency between the feature mapping and sparse coding, a joint learning framework [61], [62] was used to obtain a transformation and the codebook simultaneously while enforcing the final encoded representation of each pixel belonging to the same color to be same. As validated from our experiments, we observe that the joint learning framework helps to obtain a robust representation and boost the performance significantly. Sparse auto-encoders with an invariance constraint can be considered as an alternative to the proposed approach. While the proposed approach tries to approximate the mapped features with a set of sparse codes by selecting a few of the dictionary atoms, the deviation of the average activations from a very small value is penalized in sparse auto-encoders. This introduces certain amount of sparsity to the extracted patterns. But as pointed out in [44], learning multiple layers of relatively simpler functions can learn a better function approximation to map the input to output compared to a single layer mapping. Therefore, the feature learning and encoding scheme is kept separate in the proposed approach instead of enforcing multiple constraints over the single feature mapping scheme. More details are given in the experimental analysis section.

IV. Proposed Methodology

As mentioned in the previous sections, the appearance of color changes across camera views due to a stark change in illumination. Fig.3 shows an example of pair of patches corrected by a color constancy algorithm which still appears to be of different colors. It can be seen from the figure 3 that the histograms of these corrected patches appear to be very different. Finlayson et al. [25] have shown that the diagonal model is an accurate model to achieve color-constancy for narrow-band (sensitive to single wavelength) imaging sensors. The diagonal model states that the R, G, B values for a pixel under canonical illumination can be obtained by individually scaling the observed R, G and B values of the same pixel under an unknown illumination. But in practice, such narrow band sensors do not exist and the diagonal model is thus considered as an approximate model to correct the images for its illumination changes.

As mentioned in [58], an illumination invariant representation can be obtained by normalizing the image channel-wise with the ℓ_2 norm of the respective channels. Effectively, it cancels the diagonal transformation dependency of the pixels and gives a weak approximation of the canonical illuminant space. Eventhough this technique does not help to achieve complete illumination invariance, we use ℓ_2 norm as a pre-processing step in our approach. Patches sampled from the images are pre-processed using ℓ_2 norm. We use a linear auto-encoder to transform the pixels into a rich higher dimensional space. These transformed pixel values are encoded using a sparse coding technique to obtain more robust representations. The objective behind the encoding is that the encoded values for corresponding pixels should be the same (or very close). To achieve consistency between the linear transformation and the encoding, we adopt a joint learning strategy to optimize the linear auto-encoder transformation and dictionary learning simultaneously. The parameters are updated alternatively to find the optimal mapping and encoding for the pixel values. An illustration of our approach is shown in figure 4. The detailed flow of our approach is given in the subsequent sections.

A. Training Patch Collection

To train the system, patch pairs were extracted from the training image pairs manually, since no ground truth on patch correspondence was provided. One of the main constraints with any learning based approach for color constancy is that it is very challenging to create a dataset of real-world images with all possible colors captured under all possible illumination settings and light intensities. Therefore, as a selection principle, we carefully choose patch pairs so that they are distributed among different colors under varying illumination settings i.e. the difference in their colors range from zero (or almost no change in their visual appearance) to a maximum observable change. This is to ensure that no bias is introduced in the sampled dataset at the training stage. From each of the training image pairs in each dataset, we sample 3 pairs of patches. Additionally, the number of samples per image pair can be reduced if the dataset is large. Fig.3 shows an example of pair of patches sampled from the VIPeR dataset. Once the patch pairs were extracted, we apply ℓ_2 norm based correction and sample pixel pairs randomly from these patch pairs. These sampled pixels undergo standard normalization and are used as the input training data for our system. In the experiments section, we have given an analysis on the number of training pixel pairs required.

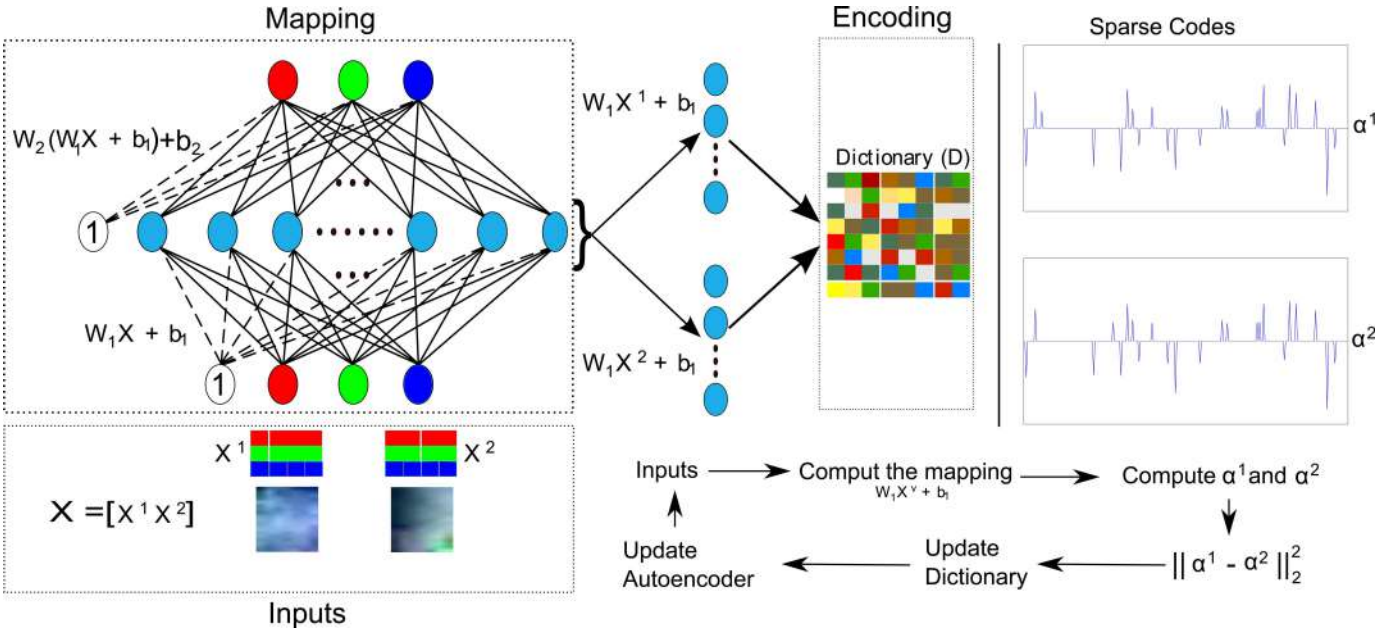


Fig. 4. Flow diagram of the joint learning procedure to train the mapping and encoding. Inputs - pixels are sampled randomly from the extracted patches. $\mathbf{X}^1$ and $\mathbf{X}^2$ correspond to the sampled pixels from patches extracted from corresponding parts of the matching image pairs which should be of same color. Mapping-The auto-encoder maps the sampled inputs into a 60 dimensional space which gives the higher dimensional representations for $\mathbf{X}^1$ and $\mathbf{X}^2$. Based on the dictionary $\mathbf{D}$, these two sets of higher dimensional responses are encoded and two sets of sparse codes, α^1 and α^2 are obtained for $\mathbf{X}^1$ and $\mathbf{X}^2$ respectively. Based on the error of encoding, $||\alpha^1 - \alpha^2||_2^2$, the dictionary is updated and the auto-encoder is updated. Further, the sparse codes are obtained for the inputs as mentioned above and this process is repeated until convergence. Thus, the joint learning strategy is adopted to learn the transformation $\mathbf{W_1}, \mathbf{b}_1, \mathbf{W_2}, \mathbf{b}_2$ and $\mathbf{D}$. **Best viewed in color**.

B. Objective Formulation

Once the training data is prepared, the objective function for learning invariant features is formulated with an invariance constraint for the feature representation obtained from pixel pairs across the two views. The objective is formulated by making use of an auto-encoder and sparse encoding technique. As mentioned in section III, auto-encoder with invariance constraint is capable of learning filters that can capture stable structures and patterns in the data. At the same time, taking the performance into consideration, a sparse encoding is also applied to the descriptors to represent them compactly. Mathematically:

$$\underset{\mathbf{W_1},\mathbf{W_2},\mathbf{D},\alpha_1,\alpha_2}{\text{minimize}} \ \ell_{ae} + \ell_{sc} + \varepsilon_{en} + \Omega, \tag{1}$$

$$\ell_{ae} = \frac{1}{m}\sum_{\upsilon=1}^{2}\sum_{i=1}^{m}\left\|\left(\mathbf{W_2}^T\left(\mathbf{W_1}^T\mathbf{x}_i^\upsilon + \mathbf{b}_1\right) + \mathbf{b}_2\right) - \mathbf{x}_i^\upsilon\right\|_2^2 \tag{2}$$

$$\ell_{sc} = \frac{\beta}{m}\sum_{\upsilon=1}^{2}\sum_{i=1}^{m}\left\|\left(\mathbf{W_1}^T\mathbf{x}_i^\upsilon + \mathbf{b}_1\right) - \mathbf{D}\alpha_i^\upsilon\right\|_2^2 \tag{3}$$

$$\varepsilon_{en} = \frac{\gamma}{m}\sum_{i=1}^{m}\left\|\alpha_i^1 - \alpha_i^2\right\|_2^2 \tag{4}$$

$$\Omega = \lambda\left(||\mathbf{W_1}||_F^2 + ||\mathbf{W_2}||_F^2\right) + \rho\,\|\mathbf{D}\|_F^2 + \frac{\eta}{m}\sum_{\upsilon=1}^{2}\sum_{i=1}^{m}\left(\sum_{k=1}^{d}\sqrt{(\alpha_{i(k)}^\upsilon)^2 + \delta}\right) \tag{5}$$

where ℓ_{ae} is the auto-encoder loss function, ℓ_{sc} is the loss due to the sparse encoding, ε_{en} is the error of the encoded values and Ω is the regularization term to avoid learning trivial values and to enforce sparsity.

The first term ℓ_{ae} denotes the reconstruction error of each sampled pixel. $\mathbf{x}_i^\upsilon \in \mathbb{R}^{3\times 1}$ denotes the 3 dimensional R, G, B value of the i^{th} pixel from view υ. In this paper, $\upsilon = 1$ refers to view A images and $\upsilon = 2$ refers to view B images. m is the total number of pixels sampled from each views, i.e. i ranges from 1 to m. $\mathbf{W_1} \in \mathbb{R}^{3\times h}$ is the linear transformation matrix that transforms each of the pixels into a h dimensional space and $\mathbf{b}_1 \in \mathbb{R}^{h\times 1}$ is the bias term. Similarly, $\mathbf{W_2} \in \mathbb{R}^{h\times 3}$ is the transformation of the higher dimensional space into the original 3 dimensional space and $\mathbf{b}_2 \in \mathbb{R}^{3\times 1}$ is the bias term. Let us denote the collection of all pixels from both views by $\mathbf{X}$. This means, $\mathbf{X} = [\mathbf{X}^1 \ \ \mathbf{X}^2] \in \mathbb{R}^{3\times 2m}$, where $\mathbf{X}^1 = [\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_m^1] \in \mathbb{R}^{3\times m}$ and $\mathbf{X}^2 = [\mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_m^2] \in \mathbb{R}^{3\times m}$ are the R, G, B values of m randomly sampled pixels from patches extracted from view A images and the corresponding pixels from the patches extracted from view B images respectively. As mentioned in section II-A, previous works [28], [59] have proven that, a linear transformation is sufficient to transform images under an unknown illuminant to images under the canonical illuminant. We borrow this intuition into our work and therefore we use a linear auto-encoder.

The second term, ℓ_{sc} is the encoding term where the linearly transformed pixel values are encoded by a Dictionary. $\mathbf{D} \in \mathbb{R}^{h\times d}$ are the basis vectors (Dictionary or Codebook) to encode each of the transformed pixel values in the h dimensional space where d is the number of such learned basis vectors or dictionary atoms. $\alpha_i^\upsilon \in \mathbb{R}^{d\times 1}$ denotes the final

encoded sparse representation for the input pixel $\mathbf{x}_i^v$. β denotes the penalty of the sparse coding loss (ℓ_{sc}). Similar to $\mathbf{X}$, let us denote the sparse codes of $\mathbf{X}^1$ and $\mathbf{X}^2$ by α^1 and α^2. Therefore, the final encoding, $\alpha = [\alpha^1 \quad \alpha^2] \in \mathbb{R}^{d\times 2m}$ is the sparse representation for $\mathbf{X}$.

The third term, ε_{en} denotes the encoding error of the sparse codes of pairs of pixels from the two views. This term enforces the invariance and ensures that the learned filters and encoding scheme are capable of extracting stable and invariant features. It takes ℓ_2 norm of the error of the corresponding pixels from two views. γ is the penalty for the encoding error term (ε_{en}) in the total cost. The final term, Ω ensures that the learned parameters are not trivial. The last term in the right hand side of Ω ensures that the computed codes are sparse. For a simple and straightforward gradient based optimization, we replace $\left|\alpha_i^v\right|_1$, the ℓ_1 norm with an approximation that can smooth it at the origin, $\sum_{k=1}^{d}\sqrt{(\alpha_{i(k)}^v)^2+\delta}$. Here, $\alpha_{i(k)}^v$ indicates the k^{th} coefficient of the sparse codes α_i^v. Therefore, $\left|\alpha_i^v\right|_1$ is approximated by taking the sum of the smooth approximation for all the coefficients of α_i^v. These smooth approximations have been widely used in sparse coding literature [63], [64]. δ is infinitesimally small (1×10^{-4}). λ, ρ and η are the penalties for the respective regularization terms on the right hand side of Ω.

C. Optimization

The optimization is done alternatively between α, $\mathbf{D}$ and $(\mathbf{W_1}, \mathbf{b}_1, \mathbf{W_2}, \mathbf{b}_2)$. We use L-BFGS gradient based optimization procedure to update these values. The gradients with respect to each of the terms are given in Appendix A. Initially, α^1 and α^2 are updated based on their gradients while keeping $\mathbf{D}$ and $(\mathbf{W_1}, \mathbf{b}_1, \mathbf{W_2}, \mathbf{b}_2)$ fixed. Then $\mathbf{D}$ is updated keeping α^1, α^2 and $(\mathbf{W_1}, \mathbf{b}_1, \mathbf{W_2}, \mathbf{b}_2)$ fixed. Finally, keeping α^1, α^2 and $\mathbf{D}$ fixed, $(\mathbf{W_1}, \mathbf{b}_1, \mathbf{W_2}, \mathbf{b}_2)$ are updated together.

Theoretically, the gradient based optimization is simple for the above objective functions. But for faster convergence and a good optima, it requires a bit of finesse. With that in consideration, practically, good initializations are required for $\mathbf{W_1}$, $\mathbf{W_2}$ and $\mathbf{D}$. We give the objective function for the initializations in Appendix B. Initialization helps to achieve faster convergence as well as a better local optima. Joint learning is done until convergence. Equation (19) for L_{sc} in Appendix B was minimized using the SPAMS toolbox [65].

To compute the feature representation for each pixel in an image, we first use $\mathbf{W_1}$ and $\mathbf{b}_1$ to transform it into a higher dimensional space and use the learned dictionary $\mathbf{D}$ to compute the sparse codes. Therefore, the dimensionality of the features for each of the input image will be $M \times N \times d$ where $M \times N \times 3$ is the input image dimensions. Figure 3 (d) and (e) show the representation obtained by us using the joint learning framework. It can be visually observed that the error of encoding for the pixel sampled from the two patches is much less than the histogram representation of the patch.

D. Multi-Layer Framework

The same formulation can be extended to a multilayer framework so that the final representation of the patches sampled from the image pairs are close to each other. To achieve this objective, we encode all the pixels in the sampled patches using the method mentioned in the above section and adopt a max pooling scheme over the 2×2 regions to get the representation of a patch. The *max* pooling strategy makes the representation slightly translation invariant. We further adopt the same formulation in equation (1) to obtain the transformation and the dictionary. Once the representation is obtained at the second layer, max pooling is done over 4×4 regions. For simplicity, we keep the dimensions h and d same for the second layer. We observed that further increasing the number of layers did not give any significant advantage over the performance. Also considering the complexity of the approach, we use only two layers for our model.

E. Parameters

The formulation contains several parameters such as the weight penalties of the cost function terms, the dimensions of the linearly transformed space and the number of dictionary atoms. All the parameters were empirically determined by a 2-fold cross-validation on the training data of VIPeR dataset and kept same for others. The training data was initially split into equal halves and the parameters were varied within a certain range as indicated by the magenta, red and blue 'dot' (•) marks on the graphs of figure 5. Initially, the number of pixels for training (the amount of training data) was varied and the validation performance is as shown in figure 5a. Further, we kept the number of training pixel pairs as $100k$ i.e. in our experiments, $m = 100k$. Further, we conducted experiments with the auto-encoder alone by enforcing the invariance constraint. The initialization in the Appendix B, equation (18) is used to learn the mapping. The dimensions of the new space, h is kept as a parameter. This is analogous to learning a certain number of filters in auto-encoders [37], [66] based on image patches. The experiment was conducted with different values for h. The obtained cross-validation performance is shown in figure 5b. It can be seen that the performance does not improve much after $h = 60$. Therefore, we fix $h = 60$ for the auto-encoder mapping.

The dimensions of the final sparse codes were also tuned in a similar fashion. We adapted the sparse coding framework with the invariance condition and performed a joint optimization with the auto-encoder mapping. The performance variation at different dimensions is as shown in figure 5c. It can be seen that at $d = 250$, the performance is at its peak and decreases when we increase the dimensions further. Thus we fixed the sparse code dimension as 250 for all the experiments. The whole framework was then extended to a second layer as mentioned in section IV-D with the same parameters for the auto-encoder and sparse codes dimensions and the final performance is as shown in table II. The other parameters were also tuned as shown above for the dimensions. The obtained parameters are as follows. $\beta = 1$, $\gamma = 0.1$, $\lambda = 3 \times 10^{-3}$, $\rho = 0.01$ and $\eta = 0.01$.

V. Experiments

To validate our algorithm, we first conducted experiments on a synthetic color constancy dataset. Further, experiments

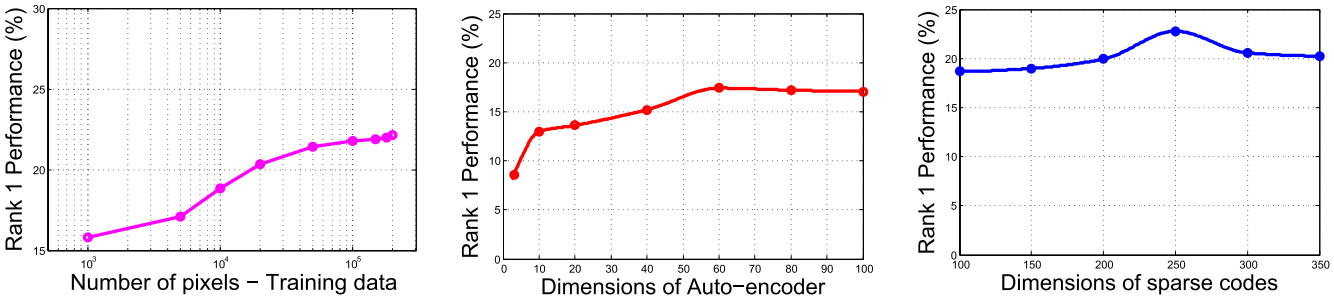


Fig. 5. Graph that shows the performance variation on the VIPeR dataset at Rank 1 with respect to the dimensions of the auto-encoder and sparse codes. The Y-axis shows Rank 1 Performance in %. (a) The performance at Rank 1 in % increases as the dimension of mapped output from the auto-encoder increases until 60. Further it does not give much gain with increase in dimensionality. (b) The performance at Rank 1 in % increases as the dimension of final sparse codes increases until 250. Further it can be seen that the performance slightly drops. (c) Graph that shows the performance variation on the VIPeR dataset at Rank 1 with respect to the number of pixel pairs sampled as the training data. Performance does not improve significantly when the sampled pixel pairs is more than $100k$.

were conducted on publicly available person re-identification datasets such as VIPeR [6], Person Re-ID 2011 [7] and CAVIAR4REID [2]. The characteristics of these three datasets are ideal for the evaluation of the proposed Jointly Learned Color Features (JLCF) since the images were captured from two cameras under varying environments such as indoor and outdoor, bright and dark illumination and different view angles. All the experiments are done without using a mask for removing the background. Below, we list the baseline approaches we compare with.

1) **Hist**: We compare our approach with the 3D histogram generated in RGB, HSV and YUV spaces. The histograms are computed from the images without any pre-processing for illumination changes. We use 6 bins for each of the channels so that the representation is 216 dimensional which is close to the 250 dimensional space proposed in this work. Image is divided into 8×8 blocks with a stride of 4 and for each of those blocks, we compute the histogram. We refer to the histogram representations as RGBHist, HSVHist and YUVHist.
2) **cHist**: cHist corresponds to the histogram of the images in the weakly corrected (ℓ_2 norm based correction) space. As mentioned for **Hist**, the number of bins for each of the channels and the size of the image blocks are kept same. The representation based on the corrected color space is referred to as cRGBHist, cHSVHist and cYUVHist.
3) rg**Hist**: rgHist corresponds to the histogram in the rg space for the corrected images. rg color channels are one of the first photometric invariant color channels proposed. The rg space corresponds to

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B} \tag{6}$$

The image is divided into blocks of 8×8 and for each of the blocks, we compute histogram of 16 bins for each channel. The final representation for each block will be 256 dimensional.

4) **Opponent**: The opponent color space is invariant to specularity. It can be computed by

$$\begin{bmatrix} O^1 \\ O^2 \\ O^3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7}$$

Image is divided into 8×8 blocks and the final histogram representation is computed as mentioned in [67].
5) **C**: The C color space adds photometric invariance to shadow shading to the opponent color space. It is computed by normalizing the Opponent descriptor by the intensity.

$$C = \begin{bmatrix} \frac{O^1}{O^3} & \frac{O^2}{O^3} & O^3 \end{bmatrix}^T \tag{8}$$

For 8×8 blocks, the histogram is computed as mentioned above for the opponent color space.
6) **Independent Learning**: The learning strategy will be optimizing the auto-encoder transformation first and then obtaining the sparse codes of the transformed pixel values without joint learning. The objective given in appendix A for initialization is used to find the optimal auto-encoder transformation. After obtaining the transformation, dictionary is learned using ℓ_{sc} in equation 1.
7) **JLCF without Invariance(JLCF WI)**: To show that the learning invariant features are important, we develop the representation without the encoding error term, i.e., excluding the $\|\alpha_i^1 - \alpha_i^2\|_2^2$ term in equation (1).

The final representation of an image in each of the color space is obtained by concatenating the histogram of each blocks in the image.

A. Synthetic Color Constancy Dataset

The efficiency of the proposed feature representation was tested on a synthetic color constancy dataset which was used in [5]. The dataset contains 7 sets of images of indoor objects captured in a laboratory setting under varying illuminations and its original image. For more details on the settings under

TABLE I

PERFORMANCE (AVERAGE PRECISION) BASED ON THE PIXEL-WISE AND HISTOGRAM BASED COMPARISON ON THE SYNTHETIC DATASET WITH DIFFERENT QUERIES. (C - C COLOR SPACE, OPP. - OPPONENT COLOR SPACE, CN - COLOR NAMES, GW - GREY WORLD CORRECTION, SG - SHADES OF GREY CORRECTION, GE - GREY EDGE CORRECTION, ℓ_2 - NORM - CORRECTION BASED ON THE ℓ_2 NORM OF THE PIXELS, CEF - COLOR EIGEN FLOWS) (a) PIXEL-WISE COMPARISON. (b) HISTOGRAM COMPARISON

(a)

Algorithm	Raw RGB	*rg*	C	Opp.	CN	GW	max RGB	SG	GE	ℓ_2 norm	CEF	JLCF	JLCF No ℓ_2	SAE-60	SAE - 250
Query 1	0.499	**0.684**	0.273	0.185	0.192	0.383	0.473	0.480	0.283	0.413	0.436	0.451	0.453	0.445	0.442
Query 2	0.037	0.042	0.028	0.032	0.048	0.033	0.041	0.032	0.042	0.034	0.198	**0.258**	0.183	0.079	0.079
Query 3	0.169	0.204	0.193	0.240	0.229	0.183	0.162	0.175	0.134	0.182	0.265	**0.365**	0.344	0.241	0.241
Query 4	0.125	0.128	0.135	0.216	0.209	0.130	0.126	0.125	0.147	0.128	0.160	**0.217**	0.206	0.128	0.127
Query 5	0.192	0.105	0.189	**0.395**	0.118	0.277	0.204	0.284	0.339	0.275	0.192	0.268	0.307	0.199	0.199
Query 6	0.221	**0.776**	0.591	0.252	0.277	0.248	0.257	0.236	0.562	0.263	0.114	0.440	0.438	0.217	0.220
Query 7	0.337	0.135	0.122	0.113	0.442	0.326	0.367	0.312	0.129	0.321	0.307	**0.463**	0.441	0.240	0.239
mAP	0.226	0.296	0.219	0.205	0.216	0.226	0.233	0.235	0.234	0.231	0.239	**0.352**	0.339	0.221	0.221

(b)

Algorithm	Raw RGB	*rg*	C	Opp.	CN	GW	max RGB	SG	GE	ℓ_2 norm	JLCF	JLCF No ℓ_2	SAE-60	SAE - 250
Query 1	0.124	0.170	0.241	0.150	0.278	0.137	0.133	0.138	0.438	0.136	**0.553**	0.513	0.509	0.477
Query 2	0.049	0.037	0.028	0.034	0.035	0.048	0.042	0.043	0.034	0.046	**0.150**	0.112	0.036	0.033
Query 3	0.126	0.210	0.185	0.159	0.168	0.121	0.125	0.122	0.125	0.121	**0.428**	0.296	0.253	0.252
Query 4	0.148	0.304	0.117	0.165	0.198	0.157	0.137	0.144	0.134	0.153	**0.369**	0.231	0.178	0.165
Query 5	0.479	0.336	0.208	0.379	0.240	0.413	0.418	0.482	**0.505**	0.453	0.317	0.175	0.243	0.253
Query 6	0.716	0.230	0.497	0.468	0.224	0.698	0.716	0.712	0.614	**0.727**	0.528	0.381	0.302	0.307
Query 7	0.108	0.255	0.141	0.141	0.208	0.115	0.108	0.120	0.097	0.121	0.445	**0.869**	0.430	0.394
mAP	0.250	0.220	0.203	0.214	0.193	0.241	0.240	0.252	0.278	0.251	**0.399**	0.368	0.279	0.269

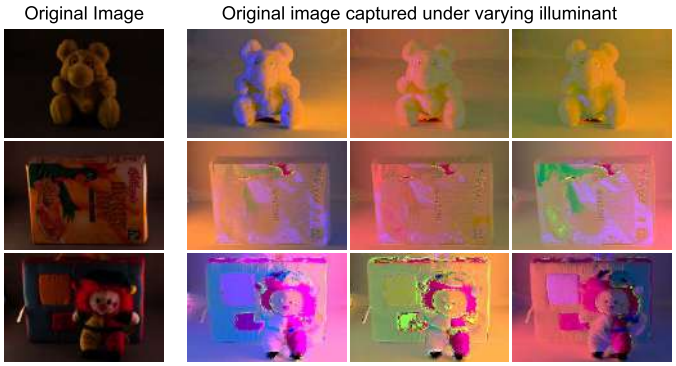


Fig. 6. Sample images shown from the synthetic dataset for color constancy experiments. Images are captured under a laboratory setting under different illuminants. In the retrieval problem, the original image is given as the query and the matching images under these different illumination is expected to be retrieved.

which the images were captured, we refer the reader to [5]. Some example images are shown in figure 6. To compare the algorithms, we formulate a retrieval problem by using the 7 original images as queries and the mean average precision (mAP) over the entire set of queries is reported as the performance measure. We adopted two different methods to compute the matching score.

1) Pixelwise Comparison: The images were corrected using different color constancy algorithms and the total pixel wise ℓ_1 distance with respect to the query was computed. This distance value is sorted to find the top ranking matches. For the proposed method, only the first layer features were used for the comparison. All the pixels were transformed as well as encoded by using the transformation and dictionary learned from the patches sampled from VIPeR dataset. The same patches were used for computing the color model for color eigenflows as well.
2) Histogram Comparison: This is similar to the above except that the ℓ_1 distance is computed for the histograms of the corrected images. For the proposed method, we do a max-pooling over 8×8 regions on the features obtained for pixel-wise comparison. We do not compare the color eigenflows model here as the "flow" of one image to another was based on the pixel-wise error between the original and the reconstructed images.

The results are shown in table I. From the results, it can be seen that the proposed approach outperforms the other color-constancy algorithms and designed color spaces in both cases (for 4/7 cases in pixel-wise comparison and 5/7 cases for histogram based comparison). The mean average precision for the proposed method is the best among all the methods for both the methods. It should be noted that, the proposed approach without any initial correction is also capable of extracting stable features as indicated by the performance on table I.

B. Person Re-Identification Datasets

For person re-identification datasets, first we do the comparison with the baselines approaches. Keeping the settings same for all of the baselines, we use LADF [68] metric learning framework for all the comparisons. For baseline 6, 7 and the proposed JLCF, the matching score of first and second layer is combined to form the final score. The results are reported based on the Cumulative Matching Characteristics(CMC) [69]. Each of the datasets, experimental settings and their evaluations are given in detail in the following subsections.

Even though invariant color features are important, to achieve a good performance, color features alone are insufficient. Therefore, complementary features such as learned low-level and high-level features are added and for comparison with state-of-the-art results. The following learned features are used for fusion with our method.

1) **AE**: Single layer auto-encoder features are learned from patches of size $8 \times 8 \times 3$. 400 filters are learned and the filter response of patches for an image are pooled over 8×8 regions. Vectorizing this representation gives the final feature for a single image. The features learned from a single layer auto-encoder are gabor-like edges.
2) **CNN**: The imagenet pre-trained model of Caffe [70] which follows the architecture in [71] is used to obtain high-level features. The dimensionality of the obtained feature is 4096.

The matching scores obtained from the metric learning framework for each of these features are combined to obtain the final scores for identifying the best match. For each query, we first obtain the gallery matching scores for JLCF and rescale it from 0 to 1. Similarly, we obtain matching scores in the range of 0 to 1 for AE and CNN features. These scores can be summed up to obtain the final matching scores. No weights were used to amplify the *influence* of the matching scores of a particular type of feature.

1) VIPeR Dataset: VIPeR [6] is the most popular and challenging dataset to evaluate Person Re-Identification. The dataset contains 632 pedestrians from arbitrary viewpoints under varying illumination conditions and have relatively low resolution. The images are normalized to a size of 128×48. We use the same settings as mentioned in [31] and [68] for the evaluation and the dataset is split into equal halves which leads to images of 316 individuals for training and testing.

Table II shows the performance comparison of our approach with different baseline methods and photometric invariant color spaces. Experimental results suggest that the encoding based on the joint learning helps to achieve good performance. Table II also shows our comparison with the independent learning strategy as well as the joint learning framework without the invariance term, ε_{en} in equation 1. Additionally, to prove that the proposed work is efficient in extracting stable patterns even without the ℓ_2 norm based correction, we added an experiment by learning the transformation and encoding parameters without the initial ℓ_2 norm based pre-processing. Results in table II shows that the JLCF without pre-processing (JLCF No ℓ_2) performs much better than any other histogram based approaches in corrected images. This proves that the proposed approach can extract robust feature representations which are still stable, though not as good as the one developed with the ℓ_2 norm preprocessing.

TABLE II

PERFORMANCE COMPARISON OF DIFFERENT BASELINES AND PHOTOMETRIC INVARIANT COLOR SPACES ON THE VIPeR, PERSON RE-ID 2011 AND CAVIAR4REID DATASETS. PROPOSED JOINTLY LEARNED COLOR FEATURES (JLCF) OUTPERFORM ALL THE BASELINES. (a) VIPeR. (b) PERSON RE-ID 2011. (c) CAVIAR4REID

(a)

Method	Rank 1	Rank 5	Rank 10	Rank 15
RGBHist	7.59	27.53	43.35	54.74
HSVHist	11.07	33.23	48.42	60.44
YUVHist	1.90	10.44	18.35	24.05
cRGBHist	7.59	27.53	43.67	54.74
cHSVHist	12.34	38.29	52.53	65.19
cYUVHist	3.48	13.60	22.46	28.16
rgHist	1.90	7.59	16.46	22.78
C Color Space	7.59	26.27	40.50	48.73
Opponent	7.91	25.31	38.29	43.67
Independent Learning	20.25	47.78	64.24	76.89
JLCF WI	19.94	48.73	65.50	75.00
JLCF No ℓ_2	23.42	50.63	64.87	72.78
Proposed JLCF	**26.27**	**51.90**	**67.09**	**78.17**

(b)

Method	Rank 1	Rank 10	Rank 20	Rank 50
RGBHist	0.8	12.3	25.2	43.7
HSVHist	0.1	7.0	9.9	20.7
YUVHist	0.9	6.6	13.3	28.7
cRGBHist	3.0	13.2	24.5	46.0
cHSVHist	1.0	9.2	17.9	34.0
cYUVHist	1.5	7.1	16.4	30.0
rgHist	0.8	3.6	6.6	17.2
C Color Space	8.1	32.1	40.4	58.9
Opponent	2.8	23.4	33.6	52.6
Independent Learning	8.4	39.3	52.3	67.0
JLCF WI	11.0	39.4	54.1	69.2
Proposed JLCF	**17.8**	**48.1**	**60.1**	**70.9**

(c)

Method	Rank 1	Rank 5	Rank 10	Rank 20
RGBHist	19.79	65.47	81.89	98.94
HSVHist	18.73	56.42	76.00	93.68
YUVHist	17.05	51.36	68.84	94.52
cRGBHist	24.21	**68.42**	84.84	97.47
cHSVHist	20.21	59.36	77.47	94.94
cYUVHist	22.73	53.26	72.42	96.84
rgHist	12.84	43.36	68.84	92.63
C Color Space	22.52	56.63	72.21	94.73
Opponent	25.05	58.10	71.78	90.52
Independent Learning	25.47	66.52	86.526	99.36
JLCF WI	26.68	65.47	82.73	98.31
Proposed JLCF	**32.63**	67.15	**87.57**	**99.36**

TABLE III

PERFORMANCE COMPARISON OF SOME OTHER COLOR BASED SIGNATURES DEVELOPED FOR PERSON RE-IDENTIFICATION ON THE VIPeR DATASET. HIST(2 PARTS) [1] REFERS TO THE HISTOGRAM IN THE LOG-CHROMATICITY COLOR SPACE. FOR SALIENT COLOR NAMES [32], THE PERFORMANCE OF THE RGB COLOR SPACE ALONE IS SHOWN FOR A FAIR COMPARISON WITH OURS. PROPOSED JOINTLY LEARNED COLOR FEATURES (JLCF) OUTPERFORMS THE OTHER METHODS

Method	Rank 1	Rank 5	Rank 10	Rank 15
Hist(2 Parts) [1] with LADF [68]	6.65	24.37	38.29	46.20
Hist(2 Parts) [1]	15.54	32.18	41.77	47.28
Salient Color Names (RGB Space) [32]	20.7	47.2	60.6	68.8
Proposed JLCF	**26.27**	**51.90**	**67.09**	**78.17**

TABLE IV

PERFORMANCE COMPARISON OF SOME OF THE ALTERNATIVES TO THE PROPOSED APPROACH. SAE - *D* REFERS TO THE SPARSE AUTO-ENCODER WITH THE FINAL DIMENSION KEPT AS *D*

Method	Rank 1	Rank 5	Rank 10	Rank 15
SAE - 60	10.13	28.80	40.19	50.95
SAE - 250	8.86	25.63	39.56	49.68
JLCF (First Layer only)	**22.78**	**49.37**	**65.82**	**77.21**

Table III shows the comparison of the proposed JLCF with other color based features proposed for person re-identification. Same experimental settings used for the other baseline methods were used for conducting this experiment. In [1], histogram intersection was used to compute the distance between two image representations for *Hist (2 Parts)*. Hist (2 Parts) signature is a 10 bin histogram computed over the log-chromaticity color space for the upper part and the lower part of the body. In table III, we also report the matching results by plugging Hist (2 Parts) with LADF metric learning framework for a fair comparison with the proposed method. It can be observed that the proposed JLCF outperforms both the results obtained by the Hist(2 Parts) color signature.

An evaluation of the state of the art algorithms using a single method is given in table VI. It can be seen that our method achieves the best result among the different algorithms. To achieve the state of the art result, we combined the features from [68] in addition to the AE and CNN features so that we can get a richer signature. The combination was done in the same way as done for the AE and CNN features. However, using color as a single cue, we achieve comparable results with several state-of-the-art methods based on multiple handcrafted features.

2) Person Re-ID 2011: The Person Re-ID 2011 dataset [7] consists of images extracted from multiple person trajectories recorded from two different, static surveillance cameras. Images from these cameras contain a sheer difference in illumination, background, camera characteristics and a significant view point change. Multiple images per person are available in each camera view. There are 385 person trajectories from one view, 749 from the other and 200 people appear in both views. For more details regarding the dataset, we refer the reader to [7] and [72]. For our experiments, images of 100 individuals appearing in both views were used for training the system.

In our experiments we do a multi-shot re-id with the same settings as mentioned in [72] and compare our results with the baselines as well as the state-of-the-art results. Table II shows the comparison of our approach with the baselines and photometric invariant color spaces. Table VI shows the comparison of our approach with state-of-the-art results in Person Re-ID 2011 dataset. As shown in the results, our method clearly outperforms the baselines, different photometric invariant color spaces and when combined with AE and CNN features, it outperforms the state-of-the-art results.

3) CAVIAR4REID: CAVIAR4REID [2] is a person re-id evaluation dataset which was extracted from the well known caviar dataset for evaluation of people tracking and detection algorithms. It is a relatively smaller dataset which contains 72 pedestrian images(50 of them in both camera views and the remaining 22 with one camera only) taken from a shopping mall in Lisbon. Re-identification in this dataset is challenging due to a large variation in the resolution, illumination, occlusion and pose changes.

The experimental settings are kept the same as in [1] which leads to 25 identities (10 images per individual) for training and testing. Table II shows that the proposed JLCF signatures are performing significantly better than the baseline color features and different photometric invariant color spaces. We also compare with other standard approaches by combining JLCF with AE and CNN features and the results are reported in table VI. Similar to the VIPeR dataset, we observed that lack of an optimal score combining mechanism affects the performance at rank 1. However, it should be noted that we achieve the best results at higher ranks.

VI. ANALYSIS

In this section, we give the analysis of our approach and compare it with the baseline methods for color features.

A. Number of Training Pixels

Since the proposed algorithm is a feature learning based model, the amount of training data (specifically, number of pixel pairs sampled) affects the performance of the system. Figure 5a shows the variation of rank 1 performance of the system with respect to the number of pixel pairs sampled. It can be seen that the performance does not improve significantly when the number of pixel pairs sampled is more than 10^5. The collected training patches were distributed across different colors sampled under varying illumination. Therefore, the sampled pixel pairs should be uniformly distributed among them to avoid any bias to a particular color or a specific illumination setting. As the number of training pixels are reduced, the amount of data corresponding to different illumination settings are reduced and the resulting learned model becomes inferior.

B. JLCF vs Designed Color Spaces

Color histograms are representations which capture the color distribution. However, the difference in illumination can

TABLE V

PERFORMANCE COMPARISON WITH DIFFERENT STANDARD COLOR CONSTANCY ALGORITHMS ON THE VIPeR, PERSON RE-ID 2011 AND CAVIAR4REID DATASETS. PROPOSED JLCF OUTPERFORMS THE HISTOGRAM BASED ENCODING ON ALL THE CORRECTED COLOR SPACES

Dataset ⇒	VIPeR				Person Re-ID 2011				CAVIAR4REID			
Method ⇓	r=1	r=5	r=10	r=15	r=1	r=10	r=20	r=50	r=1	r=5	r=10	r=20
GW - RGB	8.4	25.6	43.2	50.7	2.8	12.4	24.3	46.1	25.2	64.8	83.2	98.2
maxRGB - RGB	8.4	26.3	42.4	50.3	2.7	13.1	24.5	44.3	26.4	63.4	83.7	98.4
SG- RGB	7.9	26.3	41.2	51.6	3.1	13.7	25.6	46.8	26.7	64.2	83.9	97.1
GE- RGB	8.1	27.2	40.9	52.3	3.1	12.9	26.1	46.0	27.1	63.7	84.4	98.2
ℓ_2 - RGB	7.6	27.5	43.7	54.7	3.0	13.2	24.5	46.0	24.2	**68.4**	84.8	97.5
GW - HSV	13.4	39.4	52.6	63.3	3.8	10.7	18.1	34.6	22.2	49.2	66.2	92.2
maxRGB - HSV	12.8	36.9	53.1	61.7	0.6	4.8	9.2	20.8	16.2	43.6	65.2	91.4
SG- HSV	11.3	40.1	53.7	62.1	1.2	11.3	17.5	36.2	16.4	45.5	68.4	89.5
GE- HSV	11.7	30.2	39.6	49.2	0.8	5.3	9.2	18.5	18.7	44.3	59.9	92.2
ℓ_2 - HSV	12.3	38.3	52.5	65.2	1	9.2	17.9	34.0	20.2	59.4	77.5	95.5
GW - YUV	2.6	11.3	22.1	26.9	0.6	8.2	11.2	28.6	17.3	52.3	71.2	92.9
maxRGB - YUV	2.2	8.7	15.4	23.1	3.0	8.5	12.9	28.2	16.7	51.2	72.6	95.4
SG- YUV	3.2	12.4	19.1	21.7	1.4	7.9	12.7	33.2	17.2	49.7	71.1	96.1
GE- YUV	3.4	9.2	15.1	25.1	3.6	12.6	18.2	34.2	20.9	46.8	69.7	93.2
ℓ_2 - YUV	3.5	13.6	22.5	28.2	1.5	7.1	16.4	30.0	22.7	53.3	72.4	97.0
JLCF	**26.3**	**51.9**	**67.1**	**78.2**	**17.8**	**48.1**	**60.1**	**70.9**	**32.6**	67.2	**87.6**	**99.4**

cause a significant change in the appearance which is caused by the variation in the RGB pixel values. Therefore, without the illumination correction, the histogram will not be a robust representation for color images. Using ℓ_2 norm, we do a correction for each of the images and then obtain the histogram in such a corrected space, the **cHist**. Our approach uses a weak illumination correction algorithm and learn an optimal transformation to encode the pixel values in such a way that pixels corresponding to same patches are close enough. The other photometric invariant color spaces can be considered as handcrafted features addressing specific cues as mentioned in section V. Table I and II shows the comparisons of our method with the baseline approaches and it can be seen that JLCF clearly outperforms most of the representations for color constancy dataset and all of them in person re-identification. This is due to the fact that the ℓ_2 norm based correction which is inspired from the diagonal model is a weak illumination correction due to the strong assumptions. The comparison also shows that learned invariant color features are better than handcrafted color features. From the results in table I and II, it can be seen that even though **"JLCF No ℓ_2"** is inferior to the performance achieved by using the ℓ_2 norm based correction it is better than the histogram representations for corrected images. This indicates that the representation obtained even without any ℓ_2 norm based pre-processing can still capture the stable patterns in the pixels.

In addition to table II, we also report some additional results based on the histograms computed on the original images corrected by the standard color-constancy algorithms in table V.

C. JLCF vs Sparse Auto-Encoder

To prove that the proposed approach is more efficient than the sparse auto-encoder, we conducted experiments with the sparse auto-encoder on the synthetic color constancy dataset as well as the VIPeR dataset. Results reported in table I and IV indicate that the proposed architecture is more efficient than sparse auto-encoders. As explained in section III, multiple layers of relatively simpler functions can learn better approximations to the intended function compared to a single layer mapping with multiple constraints. For the proposed objective function, even though the auto-encoder and sparse coding framework is in a single layer, they can be considered as two separate functions which are individually pre-trained (initializations) and fine tuned. Therefore, it results in a better local optima and has better generalization capability than the sparse auto encoder framework with invariance and sparsity constraints.

D. JLCF vs Independent Learning

As mentioned in section III, to be consistent with each other, the linear auto-encoder transformation and dictionary for sparse coding must be learned jointly. As it can be seen from the results in table II, the joint learning improves the performance significantly for all the datasets. This is due to the fact that, for the encoding of each pixel, an optimal dictionary which can give same representation for pixels of same color has to be learned together with the linear transformation.

E. Proposed JLCF vs JLCF Without Invariance

To prove that the invariance term in the objective function is necessary to capture the required stable patterns, we conducted experiments with and without the color constancy term (ε_{en}) and report the results in table II. It can be seen that the invariant encoding improves the performance significantly. This is due to the fact that, without the encoding error term, the objective function merely encodes the pixels in a new space without any correction for the varying lighting conditions. But it should also be noted that the sparse encoding of linearly transformed pixel values results in a much better representation than histograms.

TABLE VI

PERFORMANCE COMPARISON OF DIFFERENT STATE-OF-THE-ART RESULTS ON THE VIPeR, PERSON RE-ID 2011 AND CAVIAR4RED DATASETS. PROPOSED JOINTLY LEARNED COLOR FEATURES (JLCF) COMBINED WITH AE AND CNN CAN ACHIEVE PROMISING RESULTS IN THE FOLLOWING DATASETS. WE HAVE SPLIT THE RESULTS OF VIPeR DATASET INTO TWO TABLES WITH THE FIRST TABLE CONSISTING OF UNSUPERVISED METHODS AND THE TABLE BELOW CONSISTING OF SUPERVISED TECHNIQUES. (a) VIPeR. (b) PERSON RE-ID 2011. (c) CAVIAR4REID

(a)

Method	Rank 1	Rank 5	Rank 10	Rank 15
Unsupervised Methods				
ELF [4]	12.00	41.50	59.50	68.00
SDALF [30]	19.87	38.89	49.37	58.46
eBiCov [3]	20.66	42.00	56.18	63.11
CPS [2]	21.84	44.00	57.21	65.18
PatMatch [31]	26.90	47.46	62.34	73.41
Supervised Methods				
WBTF [36]	21.99	46.84	59.97	70.00
SalMatch [73]	30.15	52.31	65.53	73.41
Mid-level Features [74]	29.11	52.34	65.95	73.92
LADF [68]	29.43	63.29	76.27	83.23
VWCM [75]	30.70	62.97	75.95	81.01
CMWCE [33]	37.6	68.1	**81.3**	**87.4**
Salient Color names [32]	37.8	68.5	81.2	87.0
Deep ML [46]	28.2	59.3	73.5	81.2
Proposed JLCF	26.27	51.90	67.09	78.17
Proposed JLCF + AE + CNN + [68]	**38.00**	**70.25**	81.01	87.03

(b)

Method	Rank 1	Rank 10	Rank 20	Rank 50
Descr. Model [7]	4	24	37	56
RPML [72]	15	42	54	70
Proposed JLCF	17.8	48.1	60.1	70.9
Proposed JLCF + AE + CNN	**21.5**	**49.0**	**62.7**	**73.0**

(c)

Method	Rank 1	Rank 5	Rank 10	Rank 20
HPE [76]	9.7	33.2	55.6	76.3
LF [77]	36.1	51.2	88.6	97.5
LADF [68]	29.64	62.01	78.52	94.23
SSCDL [78]	**49.1**	80.2	93.5	97.9
Proposed JLCF	32.63	67.15	87.57	99.36
Proposed JLCF + AE + CNN	45.89	**80.84**	**94.10**	**100.00**

F. Zero-Shot Transfer

To study whether the learned dictionary and transformation matrix can be generalized, experiments were conducted by doing a zero-shot transfer, i.e., by training on one dataset and testing on other datasets without any supervised training or fine-tuning on the target dataset. Specifically, the transformation matrix and the dictionary was trained on VIPeR dataset and testing was done on Person Re-ID 2011 and CAVIAR4REID datasets. Results are reported in table VII. Even though the performances does not match the performances obtained while training on the same dataset, it outperforms the histogram based representation on other specially designed color spaces (comparing table II and VII). We have also given a comparison of our zero shot transfer method from VIPeR to Person Re-ID 2011 with a doman transfer re-identification algorithm [78] in Table VIII. In [78], transfer learning is performed by calibration from source domain with minimal annotation on the target domain. It can be observed that the proposed system can outperform [78] without any label information in the domain. In the proposed system, a set of filters as well as dictionary elements were learned during the training phase and these filters can be generalized. Similar observations can be seen in several learning based methodologies. For example, CNN architectures trained on one dataset can be generalized successfully to other datasets [79].

TABLE VII

CROSS DATASET PERFORMANCE AT DIFFERENT RANKS ON THE PERSON RE-ID 2011 AND CAVIAR4REID DATASETS. TRAINING IS DONE ON VIPeR DATASET WITH $m = 100k$ PIXEL PAIRS

Person Re-ID 2011				CAVIAR4REID			
r=1	r=10	r=20	r=50	r=1	r=5	r=10	r=20
9.5	29.6	44.5	60.8	27.79	61.89	82.95	99.16

TABLE VIII

PERFORMANCE COMPARISON OF OUR ZERO-SHOT TRANSFER METHOD TO A DOMAIN TRANSFER RE-IDENTIFICATION (DTR) ALGORITHM [78] ON THE PERSON RE-ID 2011 DATASET. IT CAN BE OBSERVED THAT THE PROPOSED ALGORITHM WITH BLIND TRANSFER CAN ACHIEVE BETTER GENERALIZATION. THE RESULTS FOR [78] ARE TAKEN FROM THE CMC CURVE GIVEN IN THE LITERATURE

Method	Rank 1	Rank 5	Rank 10	Rank 15
DTR [79]	2.5	15.8	21.5	28.4
JLCF	**9.5**	**29.6**	**44.5**	**60.8**

G. Comparison With State-of-the-Art Methods

To compare with the state-of-the-art results, the matching scores obtained using JLCF is fused with the matching scores obtained using AE and CNN features. The proposed JLCF, with color features alone in the RGB color space, outperforms several recent methods such as [31], [48], and [72] at Rank 10 and 15 and exhibit comparable performances at lower ranks. All the methods mentioned in table VI that outperforms the proposed JLCF, works based on an ensemble of complementary color and texture features. It should be noted that even though color features are important for person re-identification, color alone would not be sufficient for a good performance. Therefore, to add complementary information, we combine learned low-level and high-level features to the proposed JLCF and conducted experiments. By a combination of these fea-

tures, we obtained state-of-the-art results as shown in table VI. The results indicate that the proposed color feature (JLCF) is complementary to these texture features. To differentiate the comparisons between supervised and unsupervised techniques, we have split the results in table VI (a) into two subtables. The first subtable shows the results of unsupervised methods and the second subtable shows the results of supervised methods.

VII. Conclusion

In this paper, we propose a novel framework for learning invariant color features to handle illumination and other lighting condition changes across two camera views. In contrast to the previous works based on auto-encoders and sparse coding, we combine them to learn a robust encoding jointly by forcing similar colors to be close to each other. We have evaluated the proposed approach on a synthetic color-constancy dataset as well as publicly available person re-identification datasets and demonstrated superior performance compared to several standard color-constancy algorithms as well as specially designed color spaces. By combining with other types of learned low-level and high-level features, we achieve promising results in several benchmark datasets.

Appendix A
Derivatives of the Objective Function

Let the loss function be

$$L = \ell_{ae} + \ell_{sc} + \varepsilon_{en} + \Omega \tag{9}$$

The partial derivative of the loss function w.r.t α_i^v can be given as

$$\begin{aligned}\frac{\partial L}{\partial \alpha_i^1} &= -\frac{2\beta}{m} \times \mathbf{D}^T \times (\mathbf{W_1}^T \mathbf{x}_i^1 + \mathbf{b_1} - \mathbf{D}\alpha_i^1) \\ &\quad + \frac{\eta}{m}\begin{bmatrix} \frac{\partial \Omega_{sp}}{\partial \alpha_{i(1)}^1} & \cdot\cdot & \frac{\partial \Omega_{sp}}{\partial \alpha_{i(k)}^1} & \cdot\cdot & \frac{\partial \Omega_{sp}}{\partial \alpha_{i(d)}^1} \end{bmatrix}^T \\ &\quad + \frac{2\gamma}{m}(\alpha_i^1 - \alpha_i^2) \end{aligned} \tag{10}$$

$$\begin{aligned}\frac{\partial L}{\partial \alpha_i^2} &= -\frac{2\beta}{m} \times \mathbf{D}^T \times (\mathbf{W_1}^T \mathbf{x}_i^2 + \mathbf{b_1} - \mathbf{D}\alpha_i^2) \\ &\quad + \frac{\eta}{m}\begin{bmatrix} \frac{\partial \Omega_{sp}}{\partial \alpha_{i(1)}^2} & \cdot\cdot & \frac{\partial \Omega_{sp}}{\partial \alpha_{i(k)}^2} & \cdot\cdot & \frac{\partial \Omega_{sp}}{\partial \alpha_{i(d)}^2} \end{bmatrix}^T \\ &\quad - \frac{2\gamma}{m}(\alpha_i^1 - \alpha_i^2) \end{aligned} \tag{11}$$

where $\Omega_{sp} = \left(\sum_{k=1}^{d} \sqrt{(\alpha_{i(k)}^v)^2 + \delta}\right)$ and [...] indicates the vector containing the derivatives of Ω_{sp} w.r.t each dimension of (α_i^v). The derivative of Ω_{sp} w.r.t each dimension k, of (α_i^v) can be given as

$$\frac{\partial \Omega_{sp}}{\partial \alpha_{i(k)}^v} = \frac{(\alpha_{i(k)}^v)}{\sqrt{(\alpha_{i(k)}^v)^2 + \delta}} \tag{12}$$

The partial derivative of the loss function w.r.t $\mathbf{D}$

$$\frac{\partial L}{\partial \mathbf{D}} = -\frac{2\beta}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}(\mathbf{W_1}\mathbf{x}_i^v + \mathbf{b_1} - \mathbf{D}\alpha_i^v) \times (\alpha_i^v)^T + 2\rho\mathbf{D} \tag{13}$$

The partial derivative of the loss function w.r.t $\mathbf{W_1}, \mathbf{W_2}, \mathbf{b_1}, \mathbf{b_2}$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{W_1}} &= \frac{2\beta}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}\mathbf{x}_i^v \times ((\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) - \mathbf{D}\alpha_i^v)^T \\ &\quad + \frac{2}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}((\mathbf{W_2}^T(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) + \mathbf{b_2}) - \mathbf{x}_i^v) \\ &\quad \times (\mathbf{x}_i^v)^T \times \mathbf{W_2}^T + 2\lambda\mathbf{W_1} \end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{W_2}} &= 2\lambda\mathbf{W_2} + \frac{2}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) \\ &\quad \times ((\mathbf{W_2}^T(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) + \mathbf{b_2}) - \mathbf{x}_i^v)^T \end{aligned} \tag{15}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{b_1}} &= \frac{2}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}\mathbf{W_2} \times ((\mathbf{W_2}^T(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) + \mathbf{b_2}) - \mathbf{x}_i^v) \\ &\quad + \frac{2\beta}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}((\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) - \mathbf{D}\alpha_i^v) \end{aligned} \tag{16}$$

$$\frac{\partial L}{\partial \mathbf{b_2}} = \frac{2}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}((\mathbf{W_2}^T(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) + \mathbf{b_2}) - \mathbf{x}_i^v) \tag{17}$$

Appendix B
Initializations

Initializations for $(\mathbf{W_1}, \mathbf{W_2}, \mathbf{b_1}, \mathbf{b_2})$ and $\mathbf{D}$ can be obtained by solving the objective functions L_{ae} and L_{sc} respectively.

L_{ae} can be solved by

$$\begin{aligned}&\underset{\mathbf{W},\mathbf{b}}{\text{minimize}}\ L_{ae} \\ L_{ae} &= \frac{1}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}\left\|\left(\mathbf{W_2}^T\left(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}\right) + \mathbf{b_2}\right) - \mathbf{x}_i^v\right\|_2^2 \\ &\quad + \sum_{i=1}^{m}\left\|(\mathbf{W_1}^T\mathbf{x}_i^1 + \mathbf{b_1}) - (\mathbf{W_1}^T\mathbf{x}_i^2 + \mathbf{b_1})\right\|_2^2 \\ &\quad + \lambda(||\mathbf{W_1}||_F^2 + ||\mathbf{W_2}||_F^2) \end{aligned} \tag{18}$$

and L_{sc} can be solved by

$$\begin{aligned}&\underset{\mathbf{D},\alpha}{\text{minimize}}\ L_{sc} \\ L_{sc} &= \frac{1}{m}\sum_{v=1}^{2}\sum_{i=1}^{m}||(\mathbf{W_1}^T\mathbf{x}_i^v + \mathbf{b_1}) - \mathbf{D}\alpha_i^v||_2^2 + \lambda||\alpha_i^v||_1. \end{aligned} \tag{19}$$

References

[1] 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.

[2] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–6.

[3] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Machive Vis. Conf. (BMVC)*, 2012, pp. 57.1–57.11.

[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 262–275.

[5] A. Gijsenij, R. Lu, and T. Gevers, "Color constancy for multiple light sources," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 697–707, Feb. 2012.

[6] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, Oct. 2007, pp. 1–7.

[7] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandin. Conf. Image Anal. (SCIA)*, 2011, pp. 91–102.

[8] J.-M. Morel, A. B. Petro, and C. Sbert, "Fast implementation of color constancy algorithms," *Proc. SPIE*, vol. 7241, p. 724106, Jan. 2009.

[9] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.

[10] D. A. Forsyth, "A novel algorithm for color constancy," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 5–35, Aug. 1990.

[11] A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational color constancy: Survey and experiments," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2475–2489, Sep. 2011.

[12] M. Ebner, *Color Constancy*. New York, NY, USA: Wiley, 2007.

[13] S. D. Hordley, "Scene illuminant estimation: Past, present, and future," *Color Res. Appl.*, vol. 31, no. 4, pp. 303–314, Aug. 2006.

[14] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, Jul. 1980.

[15] E. H. Land, *The Retinex Theory of Color Vision*. New York, NY, USA: Scientific America, 1977.

[16] M. J. Swain and D. H. Ballard, "Indexing via color histograms," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 1990, pp. 390–393.

[17] B. V. Funt and G. D. Finlayson, "Color constant color indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 5, pp. 522–529, May 1995.

[18] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognit.*, vol. 32, no. 3, pp. 453–464, Mar. 1999.

[19] T. Gevers, "Robust segmentation and tracking of colored objects in video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 776–781, Jun. 2004.

[20] T. Gevers and A. W. M. Smeulders, "PicToSeek: Combining color and shape invariant features for image retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 102–119, Jan. 2000.

[21] D. Berwick and S. W. Lee, "A chromaticity space for specularity, illumination color- and illumination pose-invariant 3-D object recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jan. 1998, pp. 165–170.

[22] G. Healey and D. Slater, "Global color constancy: Recognition of objects by use of illumination-invariant properties of color distributions," *J. Opt. Soc. Amer. A*, vol. 11, no. 11, pp. 3003–3010, 1994.

[23] G. West and M. H. Brill, "Necessary and sufficient conditions for Von Kries chromatic adaptation to give color constancy," *J. Math. Biol.*, vol. 15, no. 2, pp. 249–258, 1982.

[24] M. D'Zmura and P. Lennie, "Mechanisms of color constancy," *J. Opt. Soc. Amer. A*, vol. 3, no. 10, pp. 1662–1672, 1986.

[25] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Spectral sharpening: Sensor transformations for improved color constancy," *J. Opt. Soc. Amer. A*, vol. 11, no. 5, pp. 1553–1563, 1994.

[26] Q. Zaidi, "Identification of illuminant and object colors: Heuristic-based algorithms," *J. Opt. Soc. Amer. A*, vol. 15, no. 7, pp. 1767–1776, 1998.

[27] B. V. Funt and B. C. Lewis, "Diagonal versus affine transformations for color correction," *J. Opt. Soc. Amer. A*, vol. 17, no. 11, pp. 2108–2112, 2000.

[28] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Color constancy: Generalized diagonal transforms suffice," *J. Opt. Soc. Amer. A*, vol. 11, no. 11, pp. 3011–3019, 1994.

[29] E. G. Miller and K. Tieu, "Color eigenflows: Statistical modeling of joint color changes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001, pp. 607–614.

[30] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2360–2367.

[31] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3586–3593.

[32] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 536–551.

[33] Y. Yang, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Color models and weighted covariance estimation for person re-identification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 1874–1879.

[34] F. Porikli, "Inter-camera color calibration by correlation model function," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2003, pp. II-133–II-136

[35] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 26–33.

[36] A. Datta, L. M. Brown, R. Feris, and S. Pankanti, "Appearance modeling for person re-identification using weighted brightness transfer functions," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 2367–2370.

[37] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[38] Z. Zuo and G. Wang, "Learning discriminative hierarchical features for object recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1159–1163, Sep. 2014.

[39] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1794–1801.

[40] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[41] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[42] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 873–880.

[43] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007, p. 153.

[44] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[45] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, "A quantitative theory of immediate visual recognition," *Prog. Brain Res.*, vol. 165, pp. 33–56, 2007.

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for practical person re-identification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2014, pp. 34–39.

[47] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 152–159.

[48] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 144–151.

[49] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1437–1440.

[50] R. Rama Varior and G. Wang, "A data-driven color feature learning scheme for image retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1334–1338.

[51] T. Liu, R. Rama Varior, and G. Wang, "Visual tracking using learned color features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1976–1980.

[52] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 524–531.

[53] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2169–2178.

[54] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 2223–2231.

[55] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.

[56] T. Liu, G. Wang, L. Wang, and K. L. Chan, "Visual tracking via temporally smooth sparse coding," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1452–1456, Sep. 2015.

[57] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 1–1, Dec. 2015.

[58] M. S. Drew, J. Wei, and Z.-N. Li, "Illumination-invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Jan. 1998, pp. 533–540.

[59] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Color constancy: Enhancing von Kries adaptation via sensor transformations," in *Proc. IS&T/SPIE's Symp. Electron. Imag., Sci. Technol.*, 1993, pp. 473–484.

[60] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current Opinion Neurobiol.*, vol. 14, no. 4, pp. 481–487, Aug. 2004.

[61] M. A. Ranzato, Y. Boureau, and Y. Le Cun, "Sparse feature learning for deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 1185–1192.

[62] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham, "Multi-modal unsupervised feature learning for RGB-D scene labeling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 453–467.

[63] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 801–808.

[64] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient ℓ_1 regularized logistic regression," in *Proc. Nat. Conf. Artif. Intell.*, 2006, p. 401.

[65] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 689–696.

[66] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.

[67] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[68] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3610–3617.

[69] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," in *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[70] Y. Jia. (2013). *Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding*. [Online]. Available: http://caffe.berkeleyvision.org/.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[72] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 780–793.

[73] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2528–2535.

[74] Z. Zhang, Y. Chen, and V. Saligrama, "A novel visual word co-occurrence model for person re-identification," in *Proc. Eur. Conf. Comput. Vis. Workshop Vis. Surveill. Re-Identificat. (ECCV Workshop)*, 2014, pp. 122–133.

[75] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 898–903, May 2012.

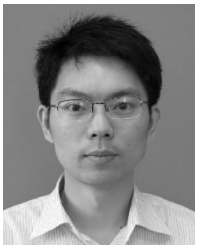
[76] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3318–3325.

[77] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3550–3557.

[78] R. Layne, T. M. Hospedales, and S. Gong, "Domain transfer for person re-identification," in *Proc. ACM/IEEE Int. Workshop Anal. Retr. Tracked Events Motion Imag. Stream (ARTEMIS)*, 2013, pp. 25–32.

[79] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Jun. 2014, pp. 806–813.

Rahul Rama Varior received the B.Tech. degree from the College of Engineering Trivandrum, University of Kerala, India, in 2010. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. From 2010 to 2012, he was a Digital Design and Verification Engineer with Wipro Technologies, India. His current research interests include computer vision, pattern recognition, and machine learning.

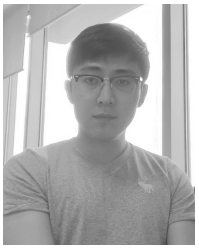


Gang Wang received the Ph.D. degree from the University of Illinois at Urbana–Champaign, in 2010. He has been an Assistant Professor with Nanyang Technological University, Singapore, since 2010. From 2010 to 2014, he was a Research Scientist with the Advanced Digital Sciences Centre, Singapore. He interned in numerous companies, including Microsoft, Kodak, and Adobe. He has received a number of research awards, including the Harriett & Robert Perry Fellowship CS/AI Award MMSP Top 10 Percent Paper Award, and the PREMIA Silver Paper Award. He is an Associate Editor of *Neurcomputing*.

Jiwen Lu (M'11–SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2003, 2006, and 2012, respectively.

He was a Research Scientist with the Advanced Digital Sciences Center, Singapore, from 2011 to 2015. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, machine learning, multimedia, and biometrics. He has authored/co-authored over 120 scientific papers in these areas, where more than 40 papers are published in the IEEE TRANSACTIONS journals and top-tier computer vision conferences. He serves as an Associate Editor of *Pattern Recognition Letters*, *Neurocomputing*, the IEEE ACCESS, and the IEEE BIOMETRICS COUNCIL NEWSLETTERS, and an Elected Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is/was the Workshop Co-Chair for ACCV16, the Special Session Co-Chair for VCIP'15, and an Area Chair for WACV'16, ICB16, ICME'15, and ICB'15, respectively. He has given tutorials at several international conferences, including CVPR'15, FG'15, ACCV'14, ICME'14, and IJCB'14.

Dr. Lu was a recipient of the First-Prize National Scholarship and the National Outstanding Student Award from the Ministry of Education of China in 2002 and 2003, the Best Student Paper Award from the Pattern Recognition and Machine Intelligence Association of Singapore in 2012, and the Top 10% Best Paper Award from IEEE International Workshop on Multimedia Signal Processing in 2014, and the National 1000 Young Talents Plan Program in 2015, respectively.



Ting Liu received the B.S. degree in optical information from Shandong University, Jinan, China, in 2010, and the M.S. degree in optical engineering from Tianjin University, Tianjin, China, in 2013. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and deep learning.