# An Enhanced Deep Feature Representation for Person Re-identification

Shangxuan Wu Ying-Cong Chen Xiang Li An-Cong Wu
Jin-Jie You Wei-Shi Zheng*
Intelligence Science and System Lab, Sun Yat-sen University, China
Guangdong Provincial Key Laboratory of Computational Science, China
alanwsx@outlook.com, chyingc@mail2.sysu.edu.cn, lixiang651@gmail.com
wuancong@mail2.sysu.edu.cn, youjinjie9@gmail.com, wszheng@ieee.org

## Abstract

*Feature representation and metric learning are two critical components in person re-identification models. In this paper, we focus on the feature representation and claim that hand-crafted histogram features can be complementary to Convolutional Neural Network (CNN) features. We propose a novel feature extraction model called Feature Fusion Net (FFN) for pedestrian image representation. In FFN, back propagation makes CNN features constrained by the hand-crafted features. Utilizing color histogram features (RGB, HSV, YCbCr, Lab and YIQ) and texture features (multi-scale and multi-orientation Gabor features), we get a new deep feature representation that is more discriminative and compact. Experiments on three challenging datasets (VIPeR, CUHK01, PRID450s) validates the effectiveness of our proposal.*

## 1. Introduction

Person re-identification aims at matching people from different views under surveillance cameras, which has been studied extensively in the past five years. To address the re-identification problem, existing methods exploit either cross-view invariant features [9, 7, 27, 19, 14, 33, 12, 20, 18] or cross-view robust metrics [4, 5, 12, 17, 33, 23, 3, 28, 34, 25].

Recently, Convolutional Neural Network (CNN) have been adopted in person re-identification, *e.g.* [16, 1, 26, 10]. Deep Learning provides a powerful and adaptive approach to handle computer vision problems without excessive handcraft on image features. The back propagation algorithm dynamically adjusts the parameters in CNN, which unifies both feature extraction and pairwise comparison process in a single network.

However, in real-world person re-identification, a person's appearance often undergoes large variations across non-overlapping camera views, due to significant changes in view angle, lighting, background clutter and occlusion (see Fig. 1). Hand-crafted concatenation of different appearance features, *e.g.* RGB, HSV colorspaces and LBP descriptor, which are designed to overcome cross-view appearance variations in re-identification tasks, sometimes would be more distinctive and reliable.

(a) VIPeR (b) CUHK01 (c) PRID450s

Figure 1: Sample images from VIPeR, CUHK01 and PRID450s datasets. Images on the same column represent the same person.

In order to effectively combine hand-crafted features and deeply learned features, we investigate the combination and complementary of a multi-colorspace hand-crafted features (ELF16) and deep features extracted from CNN. A deep feature fusion Network (FFN) is proposed in order to use hand-crafted features to regularize the CNN process so as to make the convolution neural networks extract features complementary to hand-crafted features. After extracting features by our FFN, traditional metric learning methods can be applied to boost the performance. Experimental results on three challenging person re-identification datasets (VIPeR, CUHK01, PRID450s) demonstrate the effectiveness of our new features. A significant improve-

*Corresponding author

ment of Rank-1 matching rate is achieved as compared to state-of-the-art methods (8.09%, 7.98% and 11.2%) on the three datasets. In a word, we show that hand-crafted features could improve the extraction process of CNN features in FFN, achieving a more robust image representation.

## 2. Related Works

**Hand-crafted Features.** Color and texture are two of the most useful characteristics in image representation. For example, HSV and LAB color histograms are used to measure the color information in the image. LBP histogram [22] and Gabor filter describe the textures of images. Recent papers use a combination of different features to produce more effective features [27, 9, 7, 9, 32, 33, 20].

Recently, features specifically designed for person re-identification significantly boost the matching rate. Local descriptors encoded by Fisher Vectors (LDFV) [19] build descriptors on Fisher Vector. Color invariants (ColorInv) [14] use color distributions as the sole cue for good recognition performance. Symmetry-driven accumulation of local features (SDALF) [7] proves that symmetry structure of segments can improve the performance significantly, and an accumulative method of features provides robustness to image distortions. Local maximal occurrence features (LOMO) [18] analyzes the horizontal occurrence of local features and maximizes the occurrence to stably represent re-identification images.

**Deep Learning.** Convolutional Neural Network has been widely used in many computer vision problems, but only a few papers concern deep learning on person re-identification.

Li *et al*. first proposed deep filter pairing nerual network (FPNN) [16] which used patch-matching layer and max-out pooling layer to handle pose and viewpoint variant. FPNN was also the first work to employ deep learning on person re-identification problems. Ahmed *et al*. improved deep learning architecture by specifically designing cross-input neighbourhood difference layer [1]. Later, the deep metric learning in [26] used "siamese" deep neural structure and a cosine layer to deal with big variations of person images. Hu *et al*. proposed deep transfer metric learning (DTML) [10], which transfers cross-domain visual knowledge into target datasets.

These deep methods combine feature extraction and image-pair classification into a single CNN network. Pairwise comparison and symmetry structures are widely used among them, which could be inheritances of traditional metric learning methods [9, 7, 27, 19, 14, 33, 12, 20, 18, 34, 25]. Since pairwise comparison is form to learn the deep neural network, it is demanded to form quite a lot of pairs for each probe image and perform deep convolution on these pairs. Compared to these works, our FFN is not based on pairwise input but directly extracts deep features on a single image, so that our deep architecture can be followed by any conventional classifiers, while existing deep learning works cannnot.

## 3. Methodology

### 3.1. Network Architecture

We use our modification of convolutional neural network (Feature Fusion Net, FFN) to learn new features. The network architecture is shown in Fig. 2. Our Feature Fusion Network consists of two parts. The first part deals with traditional convolution, pooling and activation neurons for input images; the second part processes additional hand-crafted feature representations of the same image. These two sub-networks are finally linked together to produce a full-fledged image description, so the second part will regularize the first part during learning. Finally, our new feature (4096D vector) is extracted from the last Full Convolution Layer (Fusion Layer) of FFN.

### 3.2. CNN Features

The upper part of Fig. 2 describes a traditional process of convolution and pooling. Every convolution layer is followed by a pooling layer and a local response normalizaion (LRN) layer [13], except for the $3^{rd}$ layer. Finally, the output of the $5^{th}$ pooling layer is a 4096D vector, which we regarded as CNN Features.

Most re-identification models regard CNN as a whole binary classifier with direct image input like DeepReID [16] and Ahmed's Improved Deep Re-id Model [1]. However, the work in [6] inspires us to come up with strong reason for taking the convolution layer as a feature extractor. One major characteristic of Re-identification images are whole-body images under different camera views. Most of the body parts could be found in all the camera views, but suffer from serious malposition, distortion and misalignment. The convolution in CNN allows part displacement and visual changes to be alleviated in higher-level convolution layers. Multiple convolution kernels provide different descriptions for pedestrian images. In addition, pooling and LRN layers provide nonlinear expression of corresponding description, which significantly reduces the overfitting problem. These layers contribute to a stable Convolution Neural Network that could be applied to new datasets (See Section 4 for detailed training process).

### 3.3. Hand-crafted Features

The lower part of Fig.2 extracts conventional hand-crafted features widely used in person re-identification. In this work, we employ the Ensemble of Local Features (ELF) [9] and is improved in [32, 33]. It extracts RGB, HSV and YCbCr histograms of 6 horizontal stripes of in-

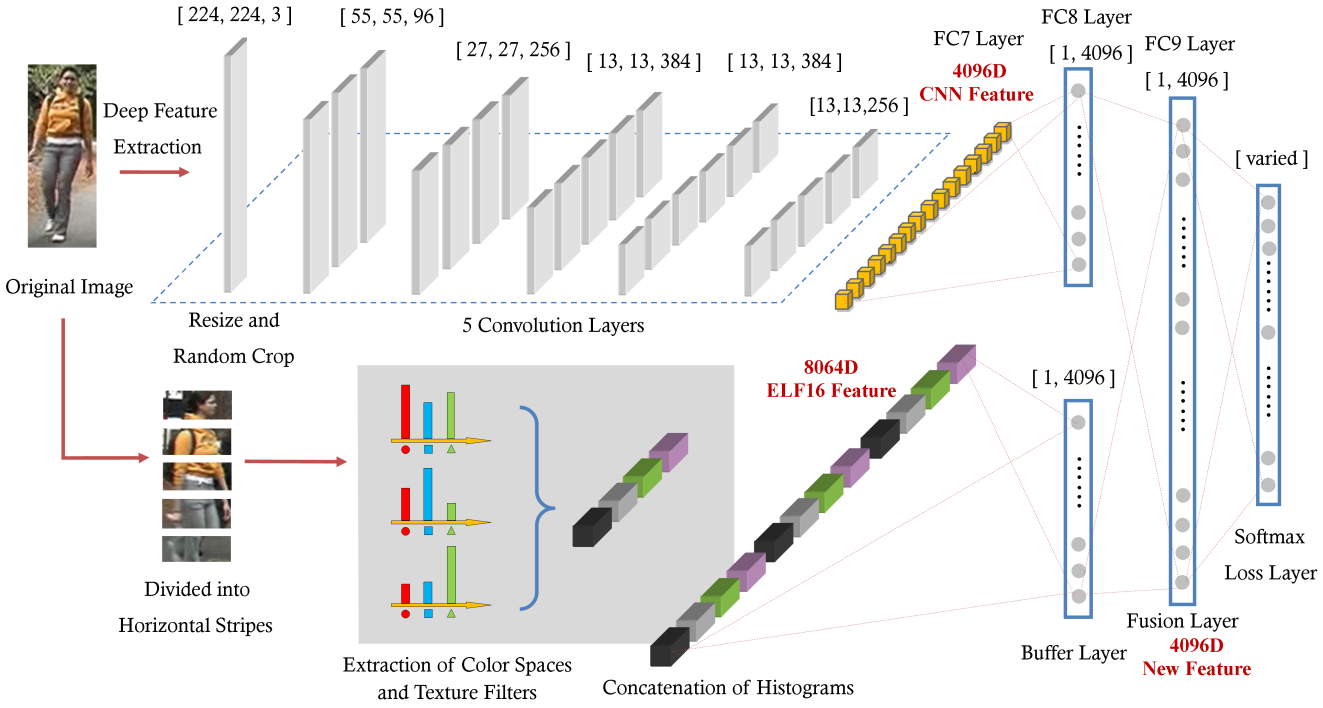



Figure 2: Fusion Feature Net (FFN) for ELF16 features and CNN features.

put image. Also, 8 Garbor filters and 13 Schmid filters are applied to get corresponding texture information.

We modify ELF feature by improving the color space and stripe division [3]. Input image is equally partitioned into 16 horizontal stripes, and our featurs are composed of color features including RGB, HSV, LAB, XYZ, YCbCr and NTSC and texture features including Gabor, Schmid and LBP. A 16D histogram is extracted for each channel and then normalized by $L_1$-norm. All histograms are concatenated together to form a single vector. In this work, we denote the above type of hand-crafted features as ELF16.

### 3.4. Proposed New Features

We aim to jointly map CNN features and hand-crafted featues to a unitary feature space. A feature fusion deep neural network is proposed in order to use hand-crafted features to regularize CNN features so as to make CNN extract complementary features. In our framework, by using back propagation, the parameters of the whole CNN network could be affected by hand-craft features. In general, as a results of the fusion, the regularized CNN features output by our proposal network should be more discriminative than both CNN features and the employed hand-crafted features.

**Fusion Layer and Buffer Layer.** Our Fusion Layer uses full connection to provide self-adaptation on person re-identification problems. Both ELF16 Features and CNN Features are followed by a 4096D-output full connection layer (Buffer Layer), which provides buffer for the fusion action. Buffer Layer is essential in our architecture, since it bridges the gap between two features with huge difference, and guarantees the convergence of FFN.

If the input of Fusion Layer is

$$\boldsymbol{x} = [\boldsymbol{ELF16}, \boldsymbol{CNN\_Features}], \tag{1}$$

then the output of this layer is computed by:

$$\boldsymbol{Z}_{Fusion}(\boldsymbol{x}) = h(\boldsymbol{W}_{Fusion}^{T}\boldsymbol{x} + \boldsymbol{b}_{Fusion}), \tag{2}$$

where $h(\cdot)$ denotes the activation function. The ReLU and dropout layers are adopted, with a dropout ratio 0.5. According to back propagation algorithm, parameters of $l^{th}$ layer after a new iteration are written as:

$$\boldsymbol{W}_{new}^{(l)} = \boldsymbol{W}^{(l)} - \alpha[(\frac{1}{m}\Delta\boldsymbol{W}^{(l)}) + \lambda\boldsymbol{W}^{(l)}], \tag{3}$$

$$\boldsymbol{b}_{new}^{(l)} = \boldsymbol{b}^{(l)} - \alpha[\frac{1}{m}\Delta\boldsymbol{b}^{(l)}], \tag{4}$$

where parameters $\alpha$, $m$ and $\lambda$ are set under the guidance of [2].

Existing deep re-identification networks for person re-identification adopt Deviance Loss [26] or Maximum Mean Discrepancy [1] as loss function. But we aim at extracting

deep features on every image effectively rather than performing pairwise comparison through a deep neural network. Therefore, softmax loss function is applied in our model, and intuitively speaking a more discriminative feature representation should result in lower softmax loss as well. For a single input vector $\boldsymbol{x}$ and a single output node $j$ in the last layer, the loss could be calculated by:

$$p(y=j|\boldsymbol{x};\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}_j^T \boldsymbol{x}}}{\sum_{k=1}^{n} e^{\boldsymbol{\theta}_k^T \boldsymbol{x}}}. \tag{5}$$

The last layer of our network is designed to minimize the cross-entropy loss:

$$J = -\sum_{k=1}^{n} \boldsymbol{p}_k \log \boldsymbol{p}_k, \tag{6}$$

in which the number of output node $n$ varies on different training sets as described in Section 4.

### 3.5. How do Hand-crafted Features Influence the Extraction of CNN Features?

If the parameters of the network are influenced by the ELF16 features $\tilde{\boldsymbol{x}}$, *i.e.*, the gradient of the network parameters are adjusted according to $\tilde{\boldsymbol{x}}$, then ELF16 features in the lower part of FFN could make CNN features more complementary with it, since the final objective of FFN is to make our features more discriminative in different images.

Denote CNN features (in FC7 layer) as $\boldsymbol{x}$ and ELF16 features as $\tilde{\boldsymbol{x}}$, Denote the weight connecting the $j^{th}$ node in $n^{th}$ layer and the $i^{th}$ node in $(n+1)^{th}$ layer as $\boldsymbol{W}_{ij}^n$. Let $\boldsymbol{Z}_j^n = \sum_j \boldsymbol{W}_{ji}^{n-1} \boldsymbol{a}_i^{n-1}$ where $\boldsymbol{a}_i^{n-1} = h(\boldsymbol{Z}_i^{n-1})$. Denote

$$\boldsymbol{\delta}_i^n = \frac{\partial J}{\partial \boldsymbol{Z}_i^n}. \tag{7}$$

Note that $\boldsymbol{Z}_j^8 = \sum_j \boldsymbol{W}^{n-1} \boldsymbol{x}_i^{n-1}$. We show that by using back propagation, $\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_{ij}^7}$ is influenced by $\tilde{\boldsymbol{x}}$. In this way, CNN Features learn its parameters which will form features complementary to the ELF16 features $\tilde{\boldsymbol{x}}$. Note that

$$\frac{\partial J}{\partial \boldsymbol{W}_{ij}^7} = \boldsymbol{x}_j \boldsymbol{\delta}_i^8, \tag{8}$$

where

$$\boldsymbol{\delta}_i^8 = (\sum_j \boldsymbol{W}_{ji}^8 \boldsymbol{\delta}_j^9) h'(\boldsymbol{Z}_i^8), \tag{9}$$

$$\boldsymbol{\delta}_j^9 = (\sum_k \boldsymbol{W}_{kj}^9 \boldsymbol{\delta}_k^{10}) h'(\boldsymbol{Z}_j^9). \tag{10}$$

$\boldsymbol{\delta}_j^9$ is influenced by $\tilde{\boldsymbol{x}}$ in two ways. Firstly,

$$\boldsymbol{Z}_k^9 = \sum_j \boldsymbol{W}_{kj}^8 \boldsymbol{a}_j^8 + \sum_j \tilde{\boldsymbol{W}}_{kj}^8 \tilde{\boldsymbol{a}}_j^8, \tag{11}$$

where

$$\tilde{\boldsymbol{a}}_j^8 = h(\sum_i \tilde{\boldsymbol{W}}_{ji}^7 \tilde{\boldsymbol{x}}_i). \tag{12}$$

In other words, the information in ELF16 features $\tilde{\boldsymbol{x}}$ could propagate through $h'(\boldsymbol{Z}_j^9)$, and thus the convolution filters of Deep Feature Extraction part would adapt itself according to $\tilde{\boldsymbol{x}}$. Secondly, the output of softmax loss layer is influenced by $\tilde{\boldsymbol{x}}$ during the forward propagation process, and thus $\boldsymbol{\delta}_k^{10}$ is also influenced by $\tilde{\boldsymbol{x}}$.

## 4. Settings for Feature Fusion Network

### 4.1. Training Dataset

Market-1501 is a multi-shot person re-identification dataset recently reported by [31]. It consists of 38195 images from 1501 identities, which is the largest public person re-identification dataset available. We trained our Feature Fusion Network on Market-1501, and used it to extract features in Section 5.

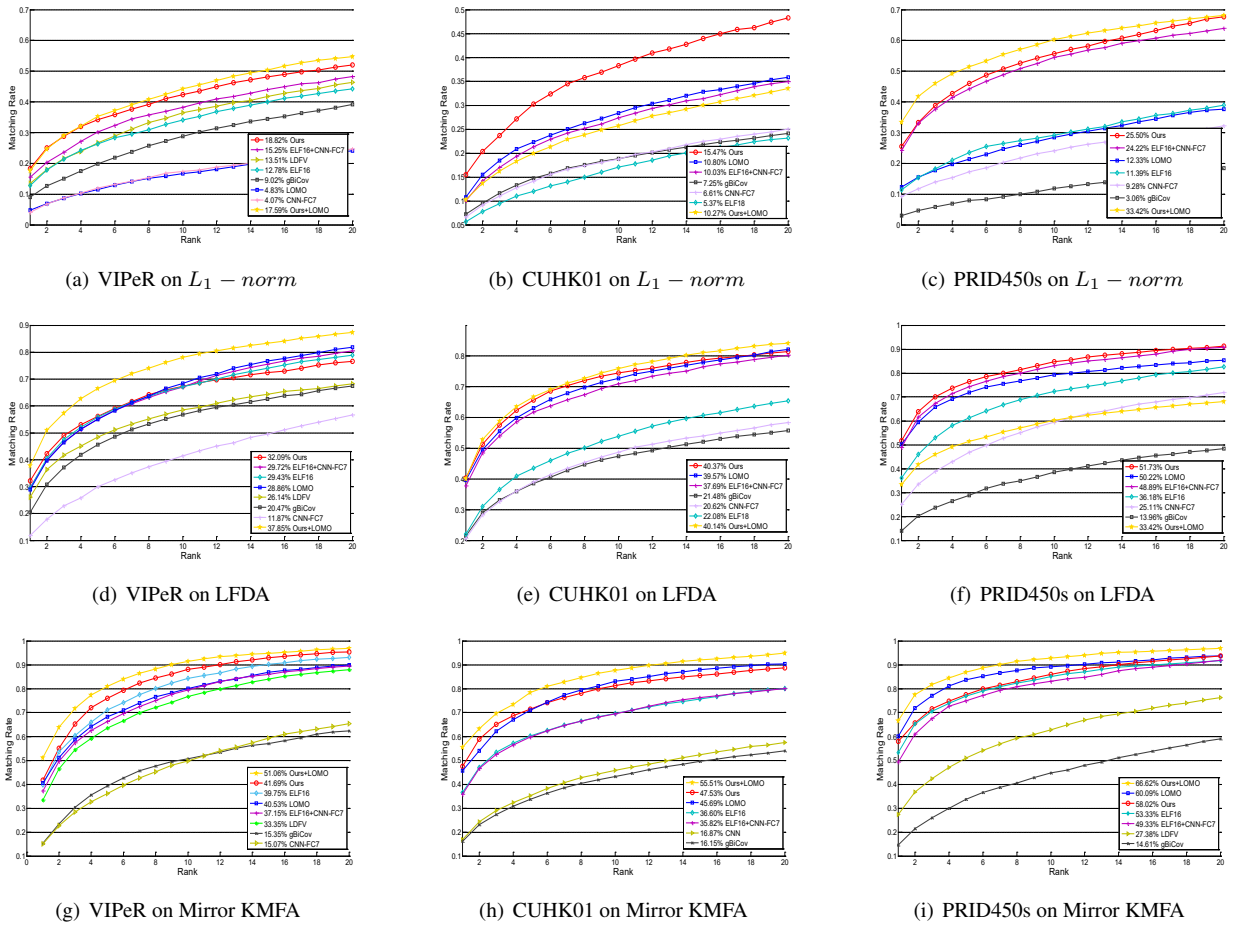
### 4.2. Training Strategies

Our training strategy applied mini-batch stochastic gradient descent (SGD) for faster back propagation and smoother convergence [2]. In each iteration of training phase, 25 images form a mini-batch and were forwarded to softmax loss Layer. The initial learning rate $\gamma = 1e-5$, which is significantly smaller than most of other CNN models. Every 20000 iterations the learning rate decreased by $\gamma_{new} = 0.1 * \gamma$. We finetuned our network based on ImageNet [13] model provided by [11]. Our FFN model took 50000 iterations to converge (about 4 hours on a Tesla K20m GPU). In order to improve the adaptation of our model, we further use difficult samples to finetune the network.

Hard negative mining [1] gives us a logical way to emphasize difficult samples in CNN. This training strategy is originally designed to balance the positive and negative samples in pairwise comparison for person re-identification. We applied this strategy to our Feature Fusion Network as well. About 12000 images of 630 IDs were wrongly-labeled by the previous network, and were manually picked out for further finetuning. We replaced the last softmax loss layer with less output nodes and continued to finetune our model on these difficult samples, with lower learning rate $(1e-6)$ and fewer iterations (about 10000). The whole training process took about 5-6 hours to converge to a tolerable training loss (about 0.05 typically).

## 5. Experiments

This section evaluated our new features in different perspectives. We presented extensive experimental results on three benchmark datasets in order to clearly demonstrate the effectiveness of our features.




(a) VIPeR on $L_1 - norm$ (b) CUHK01 on $L_1 - norm$ (c) PRID450s on $L_1 - norm$

(d) VIPeR on LFDA (e) CUHK01 on LFDA (f) PRID450s on LFDA

(g) VIPeR on Mirror KMFA (h) CUHK01 on Mirror KMFA (i) PRID450s on Mirror KMFA

Figure 3: CMC Curves of three datasets. $L_1 - norm$, LFDA and Mirror KMFA were used to evaluate the features. The yellow CMC Curves in the last row indicates the final model we used in Section 5.4.

### 5.1. Datasets and Experiment Protocols

Our test was based on three publicly available datasets: VIPeR [8], CUHK01 [15] and PRID450s [24]. Each of our datasets was presented in two disjoint camera views, with significant misalignment, light change and body part distortion. Table 1 briefly introduces these three datasets. Also, some sample images of these datasets are shown in Fig. 1.

In each individual experiment, we randomly selected half of the identities as training set, and the other half as testing set. Training set was used to train projection matrix $\boldsymbol{W}$ (in metric learning methods). Testing set used $\boldsymbol{x}' = \boldsymbol{W}^T\boldsymbol{x}$ to get the final projection of $\boldsymbol{x}$ and measures the distance between a pair of input images. For the reliability and stableness of our results, each experiment was repeated 10 times and the average Rank-i accuracy rate was computed. Cumulative Matching Curve (CMC) was also provided in Fig. 3, providing a more intuitional comparison between different algorithms.

We applied single-shot protocol in our experiment, that is during testing phase, one image was chosen from View2 as probe and all the images in View1 were regarded as the gallery. For CUHK01 specifically, which has 2 images of the same person in one camera view, we randomly chose one image of each identity as the gallery.

Mirror Kernel Marginal Analysis (KMFA), proposed by [3], provides a high-performance metric learning algorithm on person re-identification. This method was adopted in Section 5.3.2, with chi-square kernel embedded and parameters set to the optimal according to [3].

### 5.2. Features

Six feature extraction approaches were evaluted in our experiments for comparison, including LDFV [19], gBiCov [20], ImageNet [13] CNN features, LOMO features [18], ELF16 features and our proposed features[1]. For ImageNet CNN features alone, FC7 layer data, which produced the highest accuracy rate in our tests, was chosen in

[1]Our proposed features are available at http://isee.sysu.edu.cn/resource

| | VIPeR | CUHK01 | PRID450s |
|---|---|---|---|
| No. of images | 1264 | 3884 | 900 |
| No. of identities | 632 | 971 | 450 |
| No. of images in training set | 316 | 485 | 225 |
| No. of camera views | 2 | 2 | 2 |
| No. of images per view per ID | 1 | 2 | 1 |

Table 1: Re-identification datasets used in our experiments.

our experiments. Local Maximal Occurrence Representation (LOMO) is another high-performance feature representation specifically designed for re-identification problem. LDFV features were evaluated only on VIPeR dataset due to its copyrights of code. All images were resized to 224×224 for our feature extraction.

In order to demonstrate the effectiveness of our new feature, two compound features (ELF16+CNN-FC7 and Ours+LOMO) were also added for the comparison. ELF16+CNN-FC7 denotes the concatenation of normalized CNN-FC7 feature to ELF16 feature. Ours+LOMO denotes the concatenation of our new features and normalized LOMO features.

All of these features were extracted and evaluated in its default dimension (see Table 5).

## 5.3. Evaluations on Features

### 5.3.1 Unsupervised Method

Fig. 3 (a)-(c) shows the performance of our features compared to other features on $L_1 - norm$, evaluating an algorithm's capability in an original and unsupervised perspective. Our features significantly outperformed other standalone features (see Fig.3 (a)-(c)), suggesting that raw information provided by our feature is more accurate for representing re-identification images in most cases.

ELF16+CNN-FC7 features performed the second-best and outperformed both ELF16 and CNN-FC7, which provides supports on our assumption that traditional feature and CNN features are complementary. Also, our new features significantly outperformed ELF16+CNN-FC7, which may be bause of the following two reasons:

- CNN features in our network were trained to be complementary to the traditional features, while in ELF16+CNN-FC7, the CNN features are simply cascaded with ELF16 features, which may not be optimal.
- The use of Buffer Layer and Fusion layer could automatically tune the weights for each feature, and makes the fused feature perform much better.

LOMO features were specifically designed to describe person re-identification images. However, it ranked the seventh on VIPeR and the third on CUHK01, which is not stable enough for $L_1 - norm$.

### 5.3.2 Metric Learning Methods

To demonstrate the maximal effectiveness of our image description, we put it into two metric learning methods: LFDA [23] and Mirror KMFA [3], along with other widely-used features. We used each of the features to learn distance metric between each probe image and gallery set.In this experiments, we evaluated their capability on supervised metric learning methods.

Fig. 3 (d)-(i) shows the CMC curves on three datasets, with Rank-1 identification rate labeled on each feature type. Note that LDFV performed badly using chi-square kernel, so we adopted Mirror MFA without kernel trick in the comparison.

The results clearly show the outstanding performance of our proposed features, as it exceeded all the stand-alone features in VIPeR and CUHK01. Compared to ELF16 and CNN-FC7 features alone, our new features yielded much better results. Also, the simple concatenation of these two features (ELF+CNN-FC7) could not represent the image as good as ours, and it indicates the necessity of Fusion Layer in the proposed FFN.

| Rank | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| Our Model | **51.06** | **81.01** | **91.39** | **96.90** |
| Deep Feature Learning[6] | 40.50 | 60.80 | 70.40 | 84.40 |
| LOMO+XQDA [18] | 40.00 | 67.40 | 80.51 | 91.08 |
| Mirror KMFA($R_{\chi^2}$) [3] | 42.97 | 75.82 | 87.28 | 94.84 |
| mFilter+LADF [30] | 43.39 | 73.04 | 84.87 | 93.70 |
| mFilter [30] | 29.11 | 52.10 | 67.20 | 80.14 |
| SalMatch [28] | 30.16 | 52.31 | 65.54 | 79.15 |
| LFDA [23] | 24.18 | 52.85 | 67.12 | 78.96 |
| LADF [17] | 29.34 | 61.04 | 75.98 | 88.10 |
| RDC [33] | 15.66 | 38.42 | 53.86 | 70.09 |
| KISSME [12] | 24.75 | 53.48 | 67.44 | 80.92 |
| LMNN-R [5] | 19.28 | 48.71 | 65.49 | 78.34 |
| PCCA [21] | 19.28 | 48.89 | 64.91 | 80.28 |
| $L_2 - norm$ | 10.89 | 22.37 | 32.34 | 45.19 |
| $L_1 - norm$ | 12.15 | 26.01 | 32.09 | 34.72 |

Table 2: Top Matching Rank on VIPeR (Sorted by the proposed time).

Our proosed features are always better than LOMO features. Since LOMO emphasizes on HSV and SILTP histograms, it performed better on PRID450s, which was undergoing specific lighting conditions. But on other datasets, our new features are still better than LOMO features.

The concatenation of these two features (Ours+LOMO) has a strong discriminative ability and outperformed all other features on Mirror KMFA. This indicates that our deep learning methods is complementary to LOMO features. Thus, we regard this 31056D mix features as the final image representation in Mirror KMFA person re-identification model.

| Rank | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| Our Model | **55.51** | **78.40** | **83.68** | **92.59** |
| Mirror KMFA($R_{\chi^2}$) [3] | 40.40 | 64.63 | 75.34 | 84.08 |
| Ahmed's Deep Re-id [1] | 47.53 | 72.10 | 80.53 | 88.49 |
| mFilter [30] | 34.30 | 55.12 | 64.91 | 74.53 |
| SalMatch [28] | 28.45 | 45.85 | 55.67 | 67.95 |
| DeepReID [16] | 27.87 | 64.01 | 82.50 | 87.36 |
| ITML [4] | 15.98 | 35.22 | 45.60 | 59.81 |
| eSDC [29] | 19.67 | 32.72 | 40.29 | 50.58 |
| LFDA [23] | 22.08 | 41.56 | 53.85 | 64.51 |
| KISSME [12] | 14.02 | 32.20 | 44.44 | 56.61 |
| LMNN-R [5] | 13.45 | 31.33 | 42.25 | 54.11 |
| $L_2 - norm$ | 5.63 | 16.00 | 22.89 | 30.63 |
| $L_1 - norm$ | 10.80 | 15.51 | 37.57 | 35.57 |

Table 3: Top Matching Rank on CUHK01(Sorted by proposed time).

| Rank | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| Our Model | **66.62** | **86.84** | **92.84** | **96.89** |
| Mirror KMFA($R_{\chi^2}$) [3] | 55.42 | 79.29 | 87.82 | 93.87 |
| Ahmed's Deep Re-id [1] | 34.81 | 63.72 | 76.24 | 81.90 |
| ITML [4] | 24.27 | 47.82 | 58.67 | 70.89 |
| LFDA [23] | 36.18 | 61.33 | 72.40 | 82.67 |
| KISSME [12] | 36.31 | 65.11 | 75.42 | 83.69 |
| LMNN-R [5] | 28.98 | 55.29 | 67.64 | 78.36 |
| $L_2 - norm$ | 11.33 | 24.50 | 33.22 | 43.89 |
| $L_1 - norm$ | 25.50 | 25.33 | 51.73 | 53.07 |

Table 4: Top Matching Rank on PRID450s (Sorted by proposed time)

### 5.4. Comparison with State-of-the-Art

This experiment compared overall performance between state-of-the-art person re-identification model and ours. Our model is based on Mirror KMFA, using the concatenation of our new features and normalized LOMO features (Ours+LOMO).

Table 2-4 summarize some of the highest performance models on VIPeR, CUHK01 and PRID450s, including LOMO+XQDA [18], Mirror KMFA [3], Ahmed's Improved Deep ReID [1] and Mid-level Filter [30]. Our model can beat them by about 10% in Rank-1 matching rate.

Three Deep Learning methods (DeepReID [16], Ahmed's Deep Re-id [1], Ding's Deep Feature Learning [6])are specifically listed in Table 3 and 4. All of them modified CNN for pairwise comparison, and employed unique layers to match two views of the input images.

In comparison, our model regards CNN as a feature extractor, while adopting metric learning to calculate relative distance of different images. This not only contributes to the improvement in accuracy, but also enables us to use larger datasets in CNN training process. Our model also clearly exceeded their performance on CUHK01 (7.98%) and PRID450s (11.2%).

| Feature | Extraction Time | Default Dimension |
|---|---|---|
| gBiCov | 13.6152s | 5940 |
| LOMO | 0.2610s | 26960 |
| ELF16 | 0.5720s | 8064 |
| CNN-FC7 | 0.1773s | 4096 |
| Ours (with ELF16) | 0.1769s+0.5720s | 4096 |

Table 5: Average time of extracting features of a single 64x128 image (Evaluated on a 2.00GHz Xeon CPU with 16 cores).

### 5.5. Running Time

We evaluate the running time of these feature-extraction algorithms, as shown in Fig. 5. The reporeted time is the average feature extraction time for a single $48 \times 128$ image on VIPeR dataset (in its default dimension). Note that we have included the time of extracting ELF16 features in the last row. bottou2012stochastic It can be seen that our Fusion Feature Network is even faster than some of the hand-crafted methods (such as gBiCov), which breaks the stereotype of huge and clumsy Convolutional Neural Network. Also, most of the time was spent on the extraction of ELF16 features. Compared to LOMO features, our features have much lower dimension, and will perform faster in the metric learning step followed. With a balance between the speed and dimensional complexity, our Feature Fusion Network can be easily applied to actual use. Besides, compared to other CNN-based models, our FFN does not need to fine-tune on target datasets, which makes it faster to apply.

## 6. Conclusion

In this paper, we have presented a novel and effective way of feature extraction for person re-identification called Feature Fusion Network (FFN). This model jointly utilizes both CNN feature and hand-crafted features, including RGB, HSV, YCbCr, Lab, YIQ color feature and Gabor texture feature. It could adjust the weights of these information automatically with the back propagation process of Neural Network. Also, we have proved that FFN regularizes the CNN process so as to make CNN focus on extracting complementary features. Experiments on three challenging person re-identification datasets (VIPeR, CUHK01, PRID450s) show the effectiveness of our learned deep features. By using Mirror Kernel Marginal Fisher Analysis (KMFA), our proposed features significantly outperform the state-of-the-art person re-identification models on these three datasets by 8.09%, 7.98%, and 11.2% (in Rank-1 accuracy rate), respectively.

## Acknowledgements

This research was partly supported by Guangdong Provincial Government of China through the Computational

Science Innovative Research Team Program, and partially by Natural Science Foundation of China (Nos. 61472456, 61522115, 61573387), Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, and the GuangDong Program (No. 2015B010105005).

## References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE CVPR*, 2015.

[2] L. Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*. 2012.

[3] Y.-C. Chen, W.-S. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *IJCAI*, 2015.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[5] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*. 2011.

[6] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE CVPR*, 2010.

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE PETS Workshop*, 2007.

[9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008.

[10] J. Hu, J. Lu, and Y.-P. Tan. Deep transfer metric learning. In *IEEE CVPR*, 2015.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[12] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE CVPR*, 2012.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[14] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE TPAMI*, 35(7):1622–1634, 2013.

[15] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE CVPR*, 2014.

[17] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE CVPR*, 2013.

[18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE CVPR*, 2015.

[19] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012.

[20] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *IVC*, 32(6):379–390, 2014.

[21] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE CVPR*, 2012.

[22] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002.

[23] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE CVPR*, 2013.

[24] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. 2014.

[25] X. Wang, W.-S. Zheng, X. Li, and J. Zhang. Cross-scenario transfer person re-identification. *IEEE TCSVT*, PP(99):1–1, 2015.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *IEEE ICPR*, 2014.

[27] Y. Zhang and S. Li. Gabor-lbp based region covariance descriptor for person re-identification. In *IEEE ICIG*, 2011.

[28] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE ICCV*, 2013.

[29] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE CVPR*, 2013.

[30] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE CVPR*, 2014.

[31] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, J. Bu, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE ICCV*, 2015.

[32] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE CVPR*, 2011.

[33] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE TPAMI*, 35(3):653–668, 2013.

[34] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE TPAMI*, PP(99):1–1, 2015.