2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Person Re-Identification Using CNN Features Learned from Combination of Attributes

Tetsu Matsukawa, Einoshin Suzuki
Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University, Japan
Email: {matsukawa, suzuki}@inf.kyushu-u.ac.jp

***Abstract*—This paper presents fine-tuned CNN features for person re-identification. Recently, features extracted from top layers of pre-trained Convolutional Neural Network (CNN) on a large annotated dataset, *e.g.*, ImageNet, have been proven to be strong off-the-shelf descriptors for various recognition tasks. However, large disparity among the pre-trained task, *i.e.*, ImageNet classification, and the target task, *i.e.*, person image matching, limits performances of the CNN features for person re-identification. In this paper, we improve the CNN features by conducting a fine-tuning on a pedestrian attribute dataset. In addition to the classification loss for multiple pedestrian attribute labels, we propose new labels by combining different attribute labels and use them for an additional classification loss function. The combination attribute loss forces CNN to distinguish more person specific information, yielding more discriminative features. After extracting features from the learned CNN, we apply conventional metric learning on a target re-identification dataset for further increasing discriminative power. Experimental results on four challenging person re-identification datasets (VIPeR, CUHK, PRID450S and GRID) demonstrate the effectiveness of the proposed features.**

I. Introduction

Person re-identification is a matching task of person images captured from different and dis-joint camera views. A typical procedure in person re-identification is composed of the following two steps. Firstly, discriminative and robust appearance features are extracted from person images [1], [2], [3]. Then a learned metric using training data with correct matching pairs is applied to increase discriminative power of the extracted features [2], [4], [5], [6], [7].

Recently, deep learning has been applied in person re-identification [8], [9], [10]. Deep learning methods in the person re-identification unify the feature extraction and the distance metric learning processes into one framework. Although deep learning requires a large amount of annotated data to obtain high performances, the amount of available training data in person re-identification is often insufficient.

Recent seminal researches show that the neural activations of top layers of a pre-trained Convolutional Neural Network (CNN) on a large annotated dataset, *e.g.*, ImageNet, can be used as strong off-the-shelf feature descriptors [11], [12], [13]. This approach requires a large number of annotated data only for the auxiliary task of the feature extraction, and conventional metric learning methods can be applied with relative smaller training data for the target task. Though the CNN is trained on ImageNet classification, extracted CNN features exhibit remarkably high performance on a diverse range of recognition tasks [11], [12], [13].

Fig. 1. Examples of combination-attributes. To explicitly learn features to distinguish different attribute combinations, we treat each of the different attribute combinations as a different class for CNN fine-tuning.

However, large disparity exists among person images and object categories of ImageNet, which limits the performance of the CNN features for person re-identification. Re-training a pre-trained CNN for another task is called *fine-tuning*, which transfers the knowledge of pre-training data and significantly improves the performance on another task [14], [15].

Recognizing person images by semantic attributes, such as *gender*, *clothing type* and *carrying object*, is another emerging task in surveillance research. The pedestrian attribute recognition is applicable to retrieval of person images by textual description, *e.g.*, eye-witness [16]. Also, it has been used as additional label information to assist person re-identification [17], [18]. Researchers have been building large-scale datasets for this task [19], [20], [21].

In this paper, we conduct a fine-tuning of CNN features on a pedestrian attribute dataset to bridge the gap of ImageNet classification and person re-identification. There are several advantages to focus on pedestrian attributes. First, attributes are often easy to be labeled by a human annotator compared to person identity when a large number of persons are in similar appearances. Second, a large number of training samples can be collected per attribute because different people have common attributes. Finally, since the attribute information adds additional constraints to improve the person re-identification accuracies, attribute datasets can be potentially combined with datasets which are annotated by person identities.

The annotated attribute labels in a pedestrian attribute dataset are often coarse and many people share the same attributes. Although feature descriptors are required to be discriminative enough to distinguish different persons, CNN tries to classify different people who have a common attribute into the same class. Therefore, the discriminative power of CNN features solely fined-tuned on pedestrian attributes is

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

typically insufficient.

To address this problem, we focus on combinations of attributes for grouping similar people. For example, there are many people wearing *sweater*, but the people who wear *red sweater* and *jeans* would be limited and thus such an attribute combination represents more person specific information. Based on this observation, we treat combinations of multiple attributes as different classes, which we call combination-attributes (Fig. 1). We then conduct a fine-tuning of CNN features by minimizing a loss function for classifying the combination-attributes. This auxiliary task forces to classify more person specific information, and thus more discriminative features can be learned within CNN. It is worth noting that the proposed combination-attribute labels require no manual effort for the annotator once the basic attribute labels are given. The major contributions of this paper are;

- We show that the fine-tuning on the pedestrian attribute dataset largely improves the performance of CNN features for person re-identification.
- We propose a loss function for classifying combination attributes to increase discriminative power of CNN features.

II. Related Works

Although deep learning is actively researched on person re-identification [8], [9], [10], most of their focuses are the design of a new architecture for matching person images. Because they require a large number of annotated samples, their performances are still lower than traditional hand-crafted features and metric learning approach on small sampled datasets [3].

Paisitkriangkrai *et al.* used CNN features in their metric ensemble approach [6]. They observed that CNN features perform poorer than hand-crafted features and suspected that pre-trained CNN on ImageNet regards color information less important. Wu *et al.* argued that hand crafted histogram features often perform well and complementary used the CNN features with hand crafted features [22]. For unsupervised settings, Hu *et al.* conducted cross dataset re-identification using CNN features trained on a different dataset [23]. None of the previous works on supervised re-identification showed that CNN features alone can perform competitive to hand crafted features.

Recently, CNN has been adopted also in pedestrian attribute recognition [24], [25], [26]. Although, Zhu *et al.* [25] applied the learned CNN for person re-identification, they used the attribute prediction scores, *i.e.*, the output layer. It is known that the upper layers of CNN are more sensitive to semantics, while intermediate layers are specific to low-level patterns, such as color and gradients [13]. Use of upper or intermediate layers rather than the output layer is a common and effective practice of CNN features [11], [12], [13].

There are several works that use interactions of multiple attributes for attribute recognition [27], [28]. While their objective is improving attribute recognition accuracies themselves, we introduce combination attributes to emerge discriminative features in CNN for improving person re-identification.

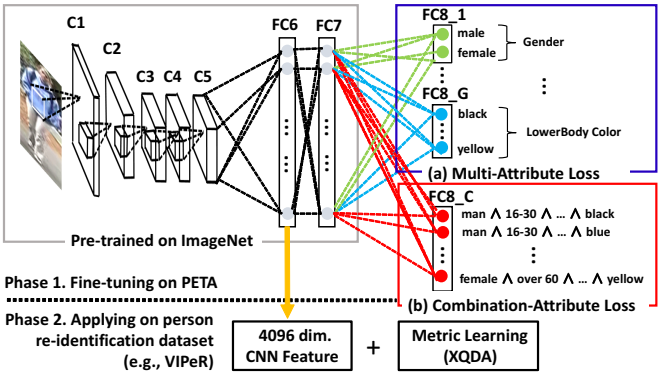


Fig. 2. The proposed CNN features. For a CNN fine-tuning, new fully connected layers and softmax loss layers for classifying multi-attributes and combination-attributes are attached to the FC7 layers of `AlexNet`.

III. Learning CNN Features using Classification of Attribute Combinations

A. Overview

We outline the our approach in Fig. 2. We use the CNN architecture designed by Krizhevsky *et al.* (`AlexNet`) [29]. The CNN is composed of five successive convolutional layers (C1,...,C5) and three fully connected layers (FC6,...,FC8). Pooling layers are applied to the first, second and fifth convolutional layers. The CNN is initialized with the pre-trained model on 1.2M images for classifying 1,000-class classification of ImageNet. Our approach consists of two phases: a CNN fine-tuning on an auxiliary dataset (Phase 1) and a feature extraction on a target re-identification dataset (Phase 2).

Phase 1. We conduct fine-tuning on Pedestrian Attribute (PETA) dataset [20], which is annotated by multiple attribute labels in several groups (Sec.IV), so that CNN jointly learns two auxiliary tasks. (a) One task is classification of multiple attributes within each of attribute groups. For this task, we attach FC8 layers in each of which the number of output nodes is equal to the number of multiple attribute labels in a group. (b) The other task is classification of combination-attributes. For each sample in the PETA dataset, we assign a combination attribute label, which is an index of the frequent attribute combinations on the PETA dataset. For this task, we attach an additional FC8 layer, in which the number of output nodes is equal to the number of combination-attribute labels. Using backpropagation, parameters of CNN are optimized to minimize loss functions for both tasks.

Phase 2. We use the CNN features on a target person re-identification dataset. For each image on the dataset, we extract a 4,096 dimensional feature vector from the first fully connected layer (FC6) of the learned CNN. We then apply metric learning on the target dataset to increase discriminative power for person re-identification.

Since for the Phase 2, we use an existing metric learning [2], the rest of this section describes the details of the Phase 1.

B. Multi-Attribute Classification Loss

The training dataset consists of N person images. Each image is annotated for G attribute groups, *e.g.*, *Gender*, *Age*,

Luggage and *UpperBody Clothing*. For each group, we have $K^{(g)}$ attributes, *e.g., male* and *female* in the *Gender* group, *black*, *white* and *yellow* in the *UpperBody Color* group. Let $D = \{(\boldsymbol{x}_i, \boldsymbol{l}_i^1, ..., \boldsymbol{l}_i^G)\}_{i=1}^N$ be the dataset where $\boldsymbol{x}_i$ is the i-th image and $\boldsymbol{l}_i^g = (l_{i,1}^g, ..., l_{i,K^{(g)}}^g)$ is its attribute label vector for the g-th attribute group. The label $l_{i,k}^g$ takes $l_{i,k}^g \in \{0, 1\}$. $l_{i,k}^g = 1$ and 0 respectively represents the presence and the absence of the k-th attribute of group g in $\boldsymbol{x}_i$.

By using a multi-attribute classification loss function, the CNN is trained to predict attribute labels in each of G attribute groups. In this paper, we assume that each image can have only one attribute in each group. We consider a $K^{(g)}$ class multi-class classification problem for each g-th group.

It is reported that sharing the CNN parameters for classifying multiple attributes improves performances of attribute recognition [24], [25], [26] [1]. Following this strategy, we share the CNN parameters and add G fully connected layers, each of them classifies attributes in each group (Fig. 2 (a)).

For each g-th attribute group, we minimize the softmax loss function. In the pedestrian attribute dataset, the number of training samples per attribute is often largely imbalanced. In such a case, CNN largely degrades performances. To handle such imbalanced training labels, we use the following weighted cross entropy loss defined by

$$L^g = -\frac{1}{N^g} \sum_{i=1}^{N} \sum_{k=1}^{K^g} \frac{l_{i,k}^g \log p_{i,k}^g}{N_{k(i)}^g}, \quad g = 1, ..., G, \qquad (1)$$

where N^g is the number of training samples in the g-th attribute group, and $N_{k(i)}^g$ is the number of training samples of k-th attribute that the i-th sample has in the g-th group. The probability $p_{i,k}^g$ is modeled by a softmax function applied to the outputs of the FC8 layer for the g-th attribute group. Let $o_{i,k}^g$ denote the k-th output for $\boldsymbol{x}_i$, then the softmax function is defined by

$$p_k^g = \frac{\exp(o_{i,k}^g)}{\sum_{k'=1}^{K^{(g)}} \exp(o_{i,k'}^g)}. \qquad (2)$$

C. Combination-Attribute Classification Loss

We focus on the combinations of attributes to group the people who commonly have more fine grained appearance information. For obtaining the combination-attribute labels, the consideration of combinations is required for only among different attribute groups because each attribute in the group is mutually exclusive. One would like to combine only discriminative subsets from all G attribute groups. However, there are many possible subsets, *i.e.*, ${}_GC_r$ combinations, where r is the number of attribute groups to be combined. For simplicity, we use the combinations involving all G attribute groups. In this case, there are $K^{(C')} = K^{(1)} \times K^{(2)} \times \cdots \times K^{(G)}$ different attribute combinations. We treat each combination as a different class in the classification loss function for the fine tuning.

[1]Note that we solves a $K^{(g)}$-class classification for each attribute group g, while previous works solve $K^{(g)}$ binary classifications. Comparing these two settings is beyond the scope of this paper.

Formally, for each i-th sample, we construct $K^{(C')}$ dimensional attribute combination indicator $l_i'^C$ where each dimension $l'^C_{i,(k_1,...,k_G)}$ is given by

$$l'^C_{i,(k_1,...,k_G)} = \begin{cases} 1 & \text{if } (l_{i,k_1}^1 = 1) \wedge ... \wedge (l_{i,k_G}^G = 1), \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

In this $K^{(C')}$ dimensional indicator, only one dimension in each i-th sample can be 1, and this dimension corresponds to the combination label of different attributes.

In general, a training dataset has imbalanced labels and several combinations are rare in the dataset. We discard the combinations that satisfy $\sum_{i=1}^N l'^C_{i,(k_1,..,k_G)} \leq N_{\min}$ from $l'^C_{i,(k_1,...,k_G)}$. By reshaping the label indicator, we obtain $K^{(C)} (\leqq K^{(C')})$ dimensional combination-attribute label vector $\boldsymbol{l}_i^C = (l_{i,1}^C, ..., l_{i,K^{(C)}}^C)$.

We add one fully connected layer for the $K^{(C)}$ dimensional classification problem and minimize the softmax loss function with the weighted cross entropy loss in Eqs.(1) and (2) (Fig. 2 (b)). We denote the loss of combination attributes by L^C.

When relevant attributes to the combination are missing, the combination labels are undefined. Further, we discard rare combinations. Therefore, some label information is missing in the combination-attributes and hence not used in the fine-tuning. To circumvent this problem, we jointly minimize the combination-attribute classification and multi-attribute classification loss functions (Fig. 2 (a) and (b)). The total loss function for our fine-tuning is defined as follows

$$L = \alpha L^C + (1 - \alpha) \frac{1}{G} \sum_{g=1}^{G} L^g, \qquad (4)$$

where $0 \leq \alpha \leq 1$ is a parameter to control the contribution of combination-attribute and multi-attribute classification losses.

Backpropagation is used to learn the parameters of the CNN [2]. Since the lower layers are shared for each attribute, the sum of the losses coming from all attributes are propagated to optimize the lower layers of the CNN.

IV. Settings for Fine-tuning

A. Pedestrian attribute dataset

We use Pedestrian Attribute (PETA) dataset [20] which is the largest public dataset for pedestrian attribute recognition. The dataset consists of 19,000 images with 61 annotated attributes. The images in the PETA dataset are extracted from the 10 public datasets for person re-identification. Since our objective is to learn transferable CNN features from different datasets, we conduct fine-tuning on different datasets from the dataset for evaluating re-identification performances. For example, when evaluating the performance of VIPeR dataset in Sec.IV, we exclude VIPeR dataset from the PETA dataset and the remaining 9 datasets are used for fine-tuning.

From all the annotated attributes, we manually selected subsets of attributes and made 7 groups of mutually exclusive

[2]Not all samples have labels in all classification problems. If the label is missing, such data are ignored when minimizing the loss function.

TABLE I
GROUP OF MUTUALLY EXCLUSIVE ATTRIBUTES.

Group (g)	Attributes	$K^{(g)}$
Gender	male, female	2
Age	less 15, 15-30, 31-45, 46-60, over 60	5
Luggage	backpack, other, folder, luggage case nothing, plastic bags, suitcase	7
UpperBody Clothing	sweter, tshit, suit, jacket, no sleeve, other	6
UpperBody Color	black, blue, brown, green, grey orange, pink, purple, red, white, yellow	11
LowerBody Clothing	suit, shorts, shirt skirt, long skirt trousers, hot pants, jeans, capri	8
LowerBody Color	black, blue, brown, grey, pink, red, white, yellow	8

attributes ($G = 7$); *Gender*, *Age*, *Luggage*, *UpperBody Clothing*, *UpperBody Color*, *LowerBody Clothing* and *LowerBody Color* (Table I). Attribute groups related to *Footwear*, *Hair* and *Accessory* are not used since they are too localized in an image. In addition, rare attributes that are annotated in less than 10 persons are not used. There are several persons who have more than two labels in the each attribute group, *e.g.*, *black* and *white* in *UpperBody Color* group. Since such a case is rare and we randomly labeled by one of the attribute labels.

B. Setup of fine-tuning

We implement our method in the `Caffe` framework [30]. The input layer of the `AlexNet`[3] is 227×227 pixels. Following the previous works [11], [26], we resize all training images into 256×256 pixels and randomly crop 227 × 227 sub-windows. For test time, we deterministically resize the all input images to 227×227 pixels. We follow the instruction of `Caffe`; we start the last fully connected layer from random weights, and all of the CNN parameters except the last layer start from the pre-trained `AlexNet`. We increase the learning rates of fully connected layers (FC6, FC7 and FC8) 10 times larger than the convolutional layers. We conduct the fine-tuning with batch size 256. We perform data augmentation by horizontal mirroring and random cropping. The initial learning rate is set to $\gamma = 0.0001$ and we decrease the learning rate by every 20,000 iterations as $\gamma_{new} = 0.1 * \gamma$. The fine-tuning typically takes 50,000 iterations to coverage (about 4 hours on a NVIDIA GTX TITAN X GPU).

V. EXPERIMENTS

A. Setup

We evaluate the performance of the fine-tuned CNN features using four person re-identification datasets; VIPeR [31], CUHK01 [32], PRID450S [5] and GRID [33]. VIPeR contains 632 person image pairs captured at outdoor with different viewpoints and illumination conditions. CUHK01 contains 971 person image pairs captured on a university campus. PRID450S contains 450 image pairs captured by two different surveillance cameras. GRID contains 250 image pairs captured on underground station and includes additional 775 images that do not belong to the person of 250 image pairs.

We follow the experimental condition with the *single shot* settings which are commonly used in previous works [31], [3], [22]. Specifically, we randomly divide each of the datasets into training and test sets containing half of the available individuals. The number of probe images is equal to the gallery images in all datasets. Note that for GRID dataset, we add additional 775 images into the gallery set. The evaluation procedure is repeated 10 times and we report the average Cumulative Matching Characteristic (CMC) curves.

We extract a 4,096 dimensional feature vector from the FC6 layer of CNN and normalize the L2 norm of the feature vector. We apply Cross view Quadratic Discriminant Analysis (XQDA) [2] for metric learning. The XQDA simultaneously learns a discriminant metric and low dimensional subspace and its latent dimension can be automatically tuned.

B. Performance Analysis on VIPeR

We analyze the parameters of the fine-tuning on VIPeR dataset. As default settings, we use parameter $\alpha = 0.5$, threshold $N_{\min} = 5$, the number of attribute groups for combine $r = 7$, the FC6 layer for feature extraction, and iteration number 50,000.

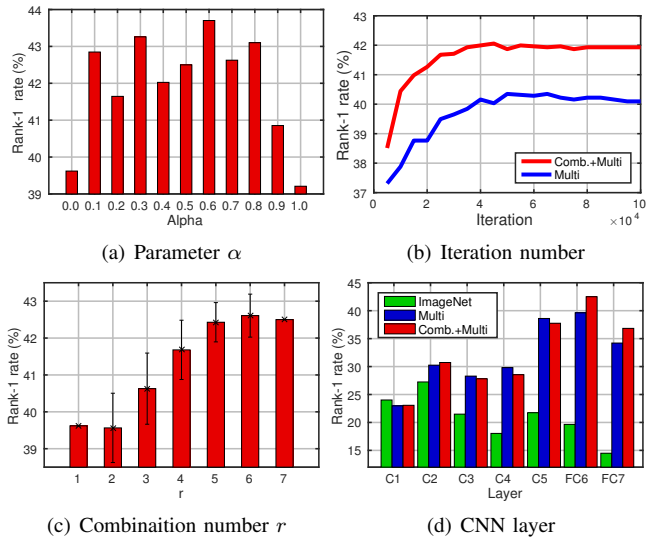
Parameter α. Fig. 3(a) shows the performances with varying α. When the combination-attribute loss is not used for the fine-tuning ($\alpha = 0$), the rank-1 rate is 39.6%. When the fine-tuning is conducted only using the combination-attribute loss ($\alpha = 1$), the rank-1 rate is 39.2%. When $0.1 \leq \alpha \leq 0.9$, the rank-1 rates are better than $\alpha = 0$ or 1 by 1.2-4.1%. These results validates our addition of combination-attribute loss in the fine-tuning.

Iteration number. Fig. 3(b) reports the rank-1 rate per iteration number of the fine-tuning. The performance of person re-identification is evaluated by CNN features every 5,000 iterations. The performances using both the combination-attribute and the multi-attribute losses for fine-tuning consistently outperform those of only using the multi-attribute loss.

Combination number r. Fig. 3(c) shows the rank-1 rates when different subsets of 7 attribute groups are used for combination labels. All the 7 attribute groups are used for multi-attribute loss, and only the combination-attribute loss is changed. For each r, we learn CNNs for all the possible subsets in the ${}_7C_r$ combinations and report their means and standard deviations. $r = 1$ represents that the combination attribute loss is not used. As to increase r, performances increase and they saturate around $r = 5, 6$ and 7. This may be because while the increase of the combination attributes makes the attribute-combination label to be more person specific information, it decreases generalization ability because higher r reduces the number of training images per each combination-attribute label.

CNN layers. Fig. 3(d) reports the rank-1 rates of different CNN layers. For each feature vector of all layers, the extracted feature vectors are L2 normalized and XQDA metric learning is adopted. It can be seen that the lower layers yield better results when fine-tuning is not conducted. This is because the higher layers are more sensitive to semantic information [13] and there is large disparity among VIPeR and ImageNet in the important semantic information. By conducting fine-tuning, the performance of all layers except the first layer improves

[3]https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet



(a) Parameter α (b) Iteration number

(c) Combinaition number r (d) CNN layer

Fig. 3. Performance analysis of the CNN features on VIPeR dataset.

drastically. The performance of the FC6 layer is higher than the FC7 layer. This may be because the person re-identification needs to distinguish persons who have the same attributes but have different appearances. The low level information contained in the FC6 layer would be also useful.

C. Performance Comparison

We then compare the performance of the proposed features on the four datasets and show the results in Table II and Fig. 4.

CNN features. Table II (a) shows the comparison of CNN features. For all features, the L2 norm normalization and XQDA metric learning are applied. FT-CNN (Comb.+Multi) is the proposed Fine-tuned CNN and FT-CNN (Multi) is the CNN fine-tuned with only the multi-attribute loss functions. CNN (ImageNet) is the pre-trained CNN on ImageNet. The fine-tuning of CNN features on pedestrian attribute dataset largely improves the performance of CNN features, the rank-1 rates of FT-CNN (Multi) are better than those of CNN (ImageNet) by 19.9%, 16.3%, 17.8% and 16.4% on VIPeR, CUHK01, PRID450S and GRID, respectively. The attribute combination labels further improve the performance of FT-CNN (Multi) by 2.9%, 2.0%, 2.4% and 0.6%, respectively on the datasets.

The FT-CNN (Person) shows the results when the person identity labels are used for fine-tuning on the PETA dataset. Since for most of the persons in the PETA dataset, the number of training samples per person is less than 3 and the number of images per each class is not sufficiently large. Therefore, our attribute based fine-tuning performs better than the learning with person identity labels in the PETA dataset.

The Feature Fusion Net (FFN) [22] uses the FC9 layer, which is more upper layer than the FC6 layers of our CNN. FFN also contains ELF16 features in its representation. The extracted features of FFN were downloaded from the authors' Web page, and we applied XQDA metric learning. Our CNN features achieve better rank-1 rates than FFN by 10.7%, 14.4% and 6.6% on VIPeR, CUHK01 and PRID450S datasets, respectively. FFN is trained on Market-1501 dataset [34] which contains 38,195 images with 1,501 person identity labels, which are more discriminative information than attribute labels. When we performed fine-tuning of our CNN on this dataset with person labels, the rank-1 rates on VIPeR, CUHK01, PRID450S and GRID were 46.3%, 49.9%, 56.5% and 24.3%, respectively. We speculate that the use of lower layer (FC6) in our features largely contributes to outperforming FFN.

State-of-the-art. Table II (b) lists performances of several state-of-the art methods. Currently, the Gaussian of Gaussians (GOG) descriptor with XQDA metric learning achieves the best results [3]. Although the performances of FT-CNN features are no better than those of GOG, they significantly outperform the second best descriptor Local Maximal Occurrence (LOMO) [2] feature on GRID dataset and yield slightly lower performances on other datasets. On VIPeR and PRID450S, FT-CNN exhibits largely better performances than the improved deep learning architecture [10]. This might be because these datasets lack enough training samples for deep learning, whereas our CNN is fine-tuned on a large auxiliary dataset and we can apply metric learning, which works well with a relatively small sampled training set.

Fusion with a hand crafted feature. Table II (c) lists the performances when CNN features are combined with hand crafted features. We simply concatenated FT-CNN and LOMO features and applied XQDA metric learning. Previously, the FFN combined with LOMO and Mirror KMFA [7] achieved the best performances [22]. Since our CNN features outperforms FFN in Table II (a), the combination of FT-CNN and LOMO achieves better results than FFN and LOMO. The Mirror KMFA metric learning uses a kernel embedding of feature vectors, which requires more computational cost. Besides, XQDA metric learning works on the original feature space. Our method achieves the best results even combined with simpler metric learning.

VI. Conclusion

We have proposed to conduct CNN fine-tuning using a new loss function for classifying combinations of pedestrian attributes. The proposed method improves the discriminative ability of attribute-based CNN features with no additional cost to the annotator. Experimental results on four challenging person re-identification datasets demonstrated the high performance gain by conducting fine-tuning on a pedestrian attribute datasets and the effectiveness of the proposed combination-attribute loss function was confirmed. As a result, CNN features achieved competitive performances to well-designed hand-crafted descriptors.

For further improvements of the CNN features, we plan to increase the number of training samples by combining person re-identification datasets and pedestrian attributes datasets. We will also investigate the combination of classification losses for person identity and attribute combinations labels.

Acknowledgment

This work was supported by JSPS KAKENHI 15K16028.

TABLE II

PERFORMANCE COMPARISON WITH DIFFERENT METHODS. CMC@RANK-r(%). (a): CNN FEATURES, (b): STATE-OF-THE-ART RESULTS, (c): COMBINATIONS OF CNN FEATURES AND HAND-CRAFTED FEATURES. * INDICATES THE RESULTS OBTAINED BY US.

	Methods	Reference	VIPeR				CUHK01				PRID450S				GRID			
			r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
(a)	FT-CNN (Comb.+Multi) + XQDA	**Ours**	**42.5**	**72.0**	**83.0**	**92.0**	**46.8**	**71.8**	**80.5**	**88.2**	**58.2**	**83.5**	**90.0**	**94.3**	**25.2**	**45.5**	54.1	64.6
	FT-CNN (Multi) + XQDA	**baseline**	39.6	69.4	81.5	90.6	44.8	71.1	79.6	87.9	55.8	81.2	89.2	93.8	24.6	43.9	**54.2**	**65.2**
	FT-CNN (Person) + XQDA	**baseline**	37.9	67.6	78.5	88.4	44.0	68.3	77.8	86.2	56.4	81.0	88.4	93.2	23.9	43.1	52.9	63.0
	FFN + XQDA	WACV2016* [22]	31.8	61.0	73.7	85.3	32.4	55.9	66.5	76.6	51.6	78.0	86.0	93.1	-	-	-	-
	CNN (ImageNet) + XQDA	**baseline**	19.7	44.5	58.1	72.9	28.5	52.3	63.6	74.9	38.0	63.4	75.3	85.5	8.2	21.1	29.8	39.5
(b)	GOG_{Fusion}+XQDA	CVPR2016 [3]	**49.7**	**79.7**	**88.7**	**94.5**	**57.8**	**79.1**	**86.2**	**92.1**	**68.4**	**88.8**	**94.5**	**97.8**	**24.7**	**47.0**	**58.4**	**69.0**
	LOMO+XQDA	CVPR2015 [2]	40.0	-	80.5	91.1	50.0	75.3	83.4	89.5	61.4	83.9	91.0	95.3	16.6	-	41.8	52.4
	Improved Deep	CVPR2015 [10]	34.8	63.6	75.6	84.5	47.5	72.1	80.5	88.5	34.8	63.7	76.2	81.9	-	-	-	-
	SCNCD	ECCV2014 [1]	37.8	68.5	81.2	90.4	-	-	-	-	41.6	68.9	79.4	87.8	-	-	-	-
	DALF	ICPR2014 [4]	35.4	62.0	73.5	84.1	-	-	-	-	-	-	-	-	18.1	37.3	46.2	59.8
(c)	FT-CNN (Comb.+Multi) + LOMO + XQDA	**Ours**	**52.1**	79.6	89.2	95.0	**62.3**	**83.7**	**90.0**	**94.3**	**71.5**	**90.6**	**94.7**	**97.5**	**29.1**	**49.4**	**59.0**	**69.4**
	FFN + LOMO + Mirror KMFA	WACV2016 [22]	51.1	**81.0**	**91.4**	**96.9**	55.5	78.4	83.7	92.6	66.0	86.8	92.8	97.0	-	-	-	-
	FFN + LOMO + XQDA	WACV2016* [22]	47.8	76.7	86.7	93.6	59.8	81.6	87.9	93.6	68.8	88.6	93.5	97.2	-	-	-	-
	MetricEnsemble (CH,SIFT,LBPs,CNN)	CVPR2015 [6]	45.9	77.5	88.9	95.8	53.4	76.4	84.4	90.5	-	-	-	-	-	-	-	-

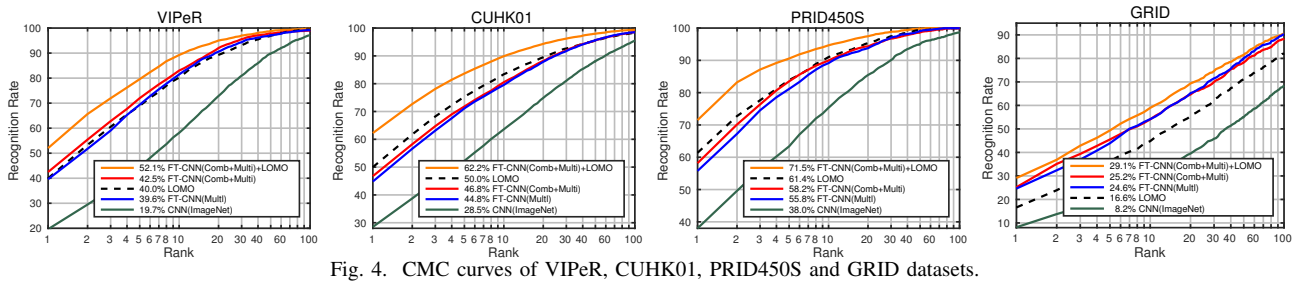


Fig. 4. CMC curves of VIPeR, CUHK01, PRID450S and GRID datasets.

REFERENCES

[1] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. ECCV*, 2014.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015.

[3] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proc. CVPR*, 2016.

[4] T. Matsukawa, T. Okabe, and Y. Sato, "Person re-identification via discriminative accumulation of local features," in *Proc. ICPR*, 2014.

[5] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," *Person Re-Identification*, 2014.

[6] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. CVPR*, 2015.

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

[8] D. Yi, Z. Lei, S. Liao, and S. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, 2014.

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

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

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014.

[12] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proc. CVPR Workshop*, 2014.

[13] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE Trans. of PAMI*, vol. 38, pp. 1790–1802, 2016.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. CVPR*, 2014.

[15] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014.

[16] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, "Attribute-based people search: Lessons learnt from a practical surveillance system," in *Proc. ICMR*, 2014.

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

[18] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attribute assisted person reidentification," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, pp. 869–878, 2015.

[19] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *Proc. ICCV Workshop*, 2013.

[20] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. ACMMM*, 2014.

[21] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv:1603.07054*, 2016.

[22] S. Wu, Y.-C. Chen, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. WACV*, 2016.

[23] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li, "Cross dataset person re-identification," in *Proc. ACCV Workshop*, 2014.

[24] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. ACPR*, 2015.

[25] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label CNN based pedestrian attribute learning for soft biometrics," in *Proc. ICB*, 2015.

[26] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," in *Proc. ICCV*, 2015.

[27] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Improving pedestrian attribute classification by weighted interactions from other attributes," in *Proc. ACCV Workshop*, 2014.

[28] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi, "Mix and match: Joint model for clothing and attribute recognition," in *Proc. BMVC*, 2015.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012.

[30] 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 preprint arXiv:1408.5093*, 2014.

[31] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008.

[32] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. ACCV*, 2012.

[33] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *IJCV*, vol. 90, pp. 106–129, 2010.

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015.