Data-driven pedestrian re-identification based on hierarchical semantic representation

Cheng, K., Xu, F., Tao, F., Qi, M. and Li, M.

Data-driven Pedestrian Re-identification based on Hierarchical Semantic Representation

Journal:	*Concurrency and Computation: Practice and Experience*
Manuscript ID	CPE-17-0411.R1
Editor Selection:	Special Issue Submission
Wiley - Manuscript type:	Special Issue Paper
Date Submitted by the Author:	09-Nov-2017
Complete List of Authors:	Cheng, Keyang Xu, Fangjie Tao, Fei Qi, Man
Keywords:	pedestrian re-identification, deep learning, attribute learning, CAEs

Data-driven Pedestrian Re-identification based on Hierarchical Semantic Representation *Concurrency and Computation: Practice and Experience*†

Keyang Cheng[1], Fangjie Xu[1],Fei Tao[1] and Man Qi[2]

[1]*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, China.*
[2]*School of Law, Criminal Justice and Computing, Canterbury Christ Church University, Canterbury, CT1 1QU,UK.*

SUMMARY

Limited number of labeled data of surveillance video causes the training of supervised model for pedestrian re-identification to be a difficult task. Besides, applications of pedestrian re-identification in pedestrian retrieving and criminal tracking are limited because of the lack of semantic representation. In this paper, a data-driven pedestrian re-identification model based on hierarchical semantic representation is proposed, extracting essential features with unsupervised deep learning model and enhancing the semantic representation of features with hierarchical mid-level 'attributes'. Firstly, CNNs, well-trained with the training process of CAEs, is used to extract features of horizontal blocks segmented from unlabeled pedestrian images. Then, these features are input into corresponding attribute classifiers to judge whether the pedestrian has the attributes. Lastly, with a table of 'attributes-classes mapping relations', final result can be calculated. Under the premise of improving the accuracy of attribute classifier, our qualitative results show its clear advantages over the CHUK02, VIPeR and i-LIDS data set. Our proposed method is proved to effectively solve the problem of dependency on labeled data and lack of semantic expression. And it also significantly outperforms the state of the art in terms of accuracy and semanteme.

KEY WORDS: pedestrian re-identification; deep learning; attribute learning; CAEs

1. INTRODUCTION

With the development of artificial intelligence and the popularization of high-definition surveillance camera in daily life, people's awareness of security has been increasingly strengthened, and the monitoring of pedestrians video has been applied in many aspects of our lives. As one of the most important technology in monitoring, pedestrian re-identification is widely applied in security domains like monitoring of public, prediction of crime, investigation of case and so on. Pedestrian re-identification is the technology to re-identify an individual who once appeared in one camera then appears in another non-overlapping one, or an individual who appears in the same camera at different times. Feature extraction and classification design are two major tasks in pedestrian re-identification.

Conventional features like color, texture, edge, have been widely used to describe pedestrians, and sometimes these complementary features are mixed to overcome the difficulty of recognition results from the variations of viewing angle and light condition [8, 14]. However, all features should

*Correspondence to: School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, China. E-mail: kycheng@ujs.edu.cn

not share the same weight, and it's necessary to set different weights for features. Zhao et al. proposed an unsupervised salience learning method to learn the inter-salience of each feature patch, and measured the similarity of two pictures with the salience, which was proved to be able to improve the matching ratio of re-identification [41]. Aiming for objects' unsteady inter-salience properties in person re-identification, [21] proposed a new algorithm, by fusing inter-salience and intra-salience based on Zhao's work. Spatial information of features' distribution is another useful cue. [67, 70] segment pedestrian pictures into patches for features extracting. There are different ways for segmentation, such as, horizontal stripe segmentation [70], triangle segmentation, concentric ring segmentation, and local patches segmentation [67]. Besides spatial information, human's biological characteristic of being symmetrical is utilized in [13], two descriptors named local weighted CIELAB histogram and salient region features are chosen to build person appearance statistical characteristics, and the histogram is based on a vertical symmetry axis of pedestrian torso and legs as the center according to the local symmetry. Features extracted by these methods are usually discriminative and representative, having achieved an amount of success in pedestrian re-identification. However, as a non-rigid object, a human can't be described by a simple feature or a mixture of several feature sets. In machine learning, data-driven features are better to describe pedestrians and more discriminative for classification.

Deep learning, a method of learning features by translating raw data into higher level features or more abstract representations through simple non-linear transformation, has been popular in the field of machine learning [31]. Compared with conventional neural network, deep learning performs better in achieving high-level features through unsupervised layered-model, in terms of the purpose of getting better representation of images. Another challenge for pedestrians re-identification is that there is a limited number of labeled data in surveillance video actually, which affects the training of supervised model significantly. Fortunately, unsupervised learning model of deep learning can be trained from unlabeled data and optimize parameters automatically, which brings new dawn to pedestrians re-identification .

Inputting features into classification can get the direct mapping of low-level features to classes, however, in pedestrian re-identification, getting a large amount of training samples of each pedestrian is not realistic, so the direct mapping of low-level features to classes is not suitable for pedestrians re-identification. Besides, digit based low-level features which is lack of semanteme, can't be understood and described by human, limiting the application of pedestrians re-identification. In the case of crime searching, witnesses describe the crime with appearances like hairstyle, clothes, and bags. These appearance characters are mid-level features called attributes, which are discriminative and semantic to represent pedestrians. Attribute learning is a method to add attributes between low-level features and classes, making pedestrian re-identification more practical in video surveillance. Attributes in research are usually designed artificially. For example, [30] and [29] selected subsets of 28 attributes and 15 attributes respectively from a human expert defined larger set of people attributes [55]. For flowers recognition, [6] defined 27 attributes for the data set of 17 species of flowers. Besides defining artificially, many researches also mining attributes automatically. [56] automatically discovered and learned new attributes that permit successful discrimination through a pair-wise learning process. And [54, 71] introduced MTL-LOREA and MLCNN methods, utilizing both low level features and semantic/data-driven attributes, for pedestrian re-identification. Cheng et al. [7] proposed an approach to mine both human understandable and discriminative attributes based on data driving with the assist of naming the candidate attributes by an annotator. In some particular cases, more detailed attributes are needed instead of some simple appearance description. For approaching the problem of identifying target suspects or finding missing people in many practical applications, [4] mined fine-grained attributes of clothes from online shopping stores. And [24, 43, 53] proposed relative attributes for providing more descriptive information to the images. Ulteriorly, as the local features don't have much positive effects on learning global attributes, [59] introduced a sparse feature preservation (SFP) method to preserve the most important features on the learning of each attribute model.

In this paper, we discuss how to take the advantages of data-driven deep learning and rich semantic attribute learning. In section 3, we review the related works in this research area.Section

3 discusses the combination of deep learning model and attribute classifiers, and introduces the related CNN, CAE and attributes of pedestrian re-identification respectively. And then , we bring forward hierarchical attributes for pedestrian re-identification, and explain how to use hierarchical attributes in two different conditions. In section 4, we show the experiment results of our models and comparison with other pedestrian re-identification methods.

The contributions of our model are as follows: (1)We introduce a method combining deep learning and attribute learning for pedestrians re-identification; (2) We bring forward hierarchical attributes and show how to utilize hierarchical attributes in pedestrians re-identification; (3) We demonstrate that our methods outperform the state-of-the-art and achieve good result in attribute absence problem.

2. RELATED WORK

2.1. Attribute-based Modeling

Attribute-based modeling has come to be a very important method in the computer vision research recently [25]. In comparison with the performances of recognition based on high-level features and on low-level features [57, 62], attributes can give benefits of various mid-level representations for describing objects, scenes and actions [12]. As we know, traditional classification approaches such as supervised learning and unsupervised learning that both aim at projecting the low-level features which are produced onto a basis set defined by the assumptions of the particular model (e.g. maximization of variance, likelihood or sparsity) [23]. But they lack semantic expression ability, so that they are not easy to be understood by human. However, attribute learning not only focuses on better semantically representing data instances by projecting them onto a basis set defined by domain-specific axes, but also helps semantic attribute prediction [45].

Recent work in this area has also examined that the exploitation of the constantly growing semantic attributes can be efficiently utilized in pedestrian re-identification to bridge the so-called ”semantic gap” between extractable low-level feature representations and high-level semantic understanding of the visual objects and improve the accuracy of the recognition when the semantic attributes are constructed to a proper hierarchy [19]. Remarkably, Wang Z and Ye M et al. give a way employing attributes to deal with person re-identification problem [58, 63].

The learning and usage of semantic attribute representations benefit for visual recognition [5]. For example, just a single pair of images to be used for target recognition seems as a challenging case of one-shot learning in person re-identification. Attribute features contribute to the modeling not only encode the color and shape information of low-level features, but also give the semantics to the features [10]. That is to say, mid-level representations bring more significant improvements than low-level features [6, 11]. The data-driven method based on hierarchical semantic representation effectively combines the raw data representation and semantic representation to give the pedestrian attribute a better interpretation [40, 65]. Therefore, using the hierarchical approach to represent human interaction can optimize the human-labeled attribute-profiles, which are more specified and constrained [27, 38, 52].

2.2. Deep Learning Methods

Nowadays, deep learning methods have been popularly applied to address problems with its unsupervised learning characteristics [1, 2, 15]. For example, convolutional Neural Network(CNN), one of the deep learning model has been used to solve large number of data and achieved state-of-the-art results on recognition with high precision [28]. Moreover, this method enables to effectively leverage data-driven, highly representative and layered hierarchical attribute features from sufficient training data to advance recognition [20, 61]. Krizhevsky et al. and Chu M et al. show that CNN is a good features generator and selector, which is a novel data-driven method of unsupervised learning [9, 16, 26].

Deep learning methods are great means of getting features, which were adopted by Sermanet et al. to localize, detect and recognize humans [48, 49]. Zeiler et al. made the convolutional network

visualizing and to be easily understood for recognition [64]. Park E et al. combined the features of deep learning model showing remarkable results [37, 44]. Lin W et al proposed to represent an activity by a combination of category components [36,50]to enhance the recognition in the condition of lacking training data.Ouyang et al. also combined many of these deep learning networks for person re-identification [42]. Furthermore, Caffe presented by Jia et al in ACM 2014, which is one of the most outstanding convolutional architecture provides a good platform for person re-identification [22].

3. PROPOSED ALGORITHM

3.1. Combine Deep Learning with Attribute Learning

Firstly, This paper proposes a novel method Combining Deep Learning and Attribute Learning for Pedestrians Re-identification(CDA-PR), which combines deep learning with attribute learning, taking both advantages of them to achieve higher accuracy rate and better practicability. CDA-PR consists of two major modules: features extraction and attributes classification. In features extraction, we take the structure of Convolutional Neural Network(CNN) to construct the model, and use another deep learning model Convolutional Auto-Encoder(CAE) to pre-train the CNN with unlabeled pedestrian pictures. The whole image is obviously unsuitable for part-based attributes classification. Considering human's body structure and attributes we used, we segment images into 5 horizontal blocks with different sizes. Accordingly, there're 5 different CNNs for these blocks, trained independently and extracting features of different parts. Features of these blocks are then input into corresponding attribute classifiers to get attribute-related probabilities. And with a table of 'attribute-classes mapping relations', the final result can be calculated. The overall architecture of CDA-PR is shown in Figure 1. We reform the feature maps of one sample generated from the last layer of CAEs into a vector, and perform a sparse operation on it, so each sample relates to one sparse feature vector. The table of 'attribute-classes mapping relations' is got from statistics, while each element of it means that the probability of one pedestrian has the attribute a_k belongs to class y_j.

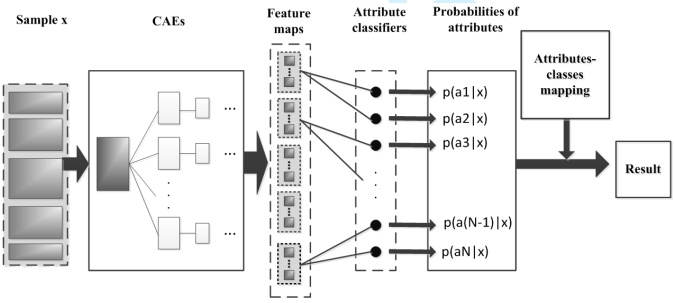


Figure 1. The structure of CDA-PR. Well trained CNN generates 5 sets of output feature maps from input pictures with 5 blocks, and a group of attribute classifiers, containing several Support Vector Machines (SVMs), output probabilities p(a_k|x) of related attributes.

3.2. CNN Learning with CAE

CNN is popular in image recognition, LeCun et al. constructed a classical model LeNet-5 [3] for handwritten digit recognition, and achieved a significant result. However, in handwritten digit recognition, each sample is labeled with annotation, making the training of CNN much more easily by computing the error between the output and annotation. When re-identifying pedestrians, there are few labeled pedestrian's samples for supervised training. For this reason, we pre-train each layer of the CNN with CAE. Actually, the structure just like Stacked Convolutional Auto-Encoders with a series of feature maps in each layer. CAE is a special neural network for handling image data [47],

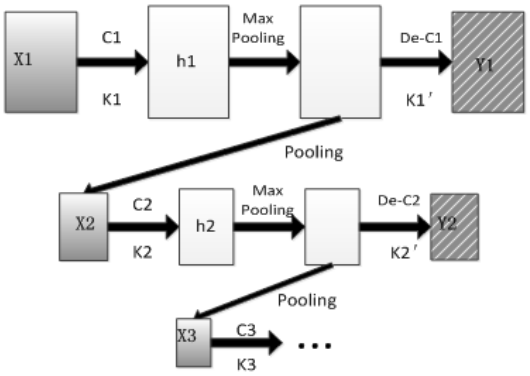


Figure 2. The process of learning features by CAEs. Each layer includes the process of convolution, max-pooling and de-convolution.

not like simple Auto-Encoder or stacked Auto-Encoder, unable to manage the pooling operation of 2-dimension images and generate a lot of parameters to be computed, which will reduce the efficiency of the system. The process of learning features by CAE is shown in Figure 2. Input image X of each layer generates a set of feature maps h, with each map generated from a different convolutional kernel. Each map represents one type of feature of the input image. Feature maps h are subsampled by max-pooling windows, unchanging the size, and then be decoded into the reconstruction Y of input image X. According to the error between Y and X, parameters can be updated to the best state by the classical BP(Back Propogation) algorithm.

Taking the structure of CNN in paper [39] for reference, we construct a 5-layer network for each horizontal blocks, including an input layer, 3 hidden layers and an output layer. In the first hidden layer, we set 20 output feature maps as set h1, with convolutional kernel size of 5×5. We set 50 output feature maps as set h2 and 5×5 kernel size for the second layer. The third layer has 100 output feature maps and the kernel size is still 5×5. Max-Pooling window is designed to be 2×2, and maps keep the size after max pooling, but the pixels are set to 0 except the max pixel in the window. Maps' size changes to be half after pooling operation, with those pixels value 0 being removed. The last layer is output layer, transforming all feature maps into a vector. So each input image is represented by 5 feature vectors after being extracted by CNN.

3.3. Attribute Learning of Pedestrian Re-identification

3.3.1. Attributes of Pedestrian

Attributes are semantic description of pedestrians' appearance. Different part of body has different attributes, for example, head related attributes are 'bald', 'wear hat', 'long hair', 'short hair', etc., clothes related attributes are 'long sleeve', 'short sleeve', 'dress', 'skirt' , 'wear coat', etc., shoes related attributes are 'single-color shoes', 'multi-color shoes', 'high-heeled shoes', 'sandal', 'boots', etc., belongings related attributes are 'backpack', 'single-shoulder bag', 'handbag', etc.. In particular environment, attributes should be designed carefully to match the condition. Take pedestrians in airport and railway station as an example, they are more likely to have attributes like 'suitcase' and 'backpack', which are not so common in markets. We firstly design 17 existing and distinguishing attributes for the pedestrians in VIPeR data set. Specifically, they are 'wear hat', 'not long hair', 'long hair', 'long sleeve', 'short sleeve', 'sleeveless', 'wear coat', 'coat texture', 'has logo', 'has bag', 'take goods', 'short bottoms', 'long bottoms', 'bottom texture', 'sandal', 'shoes' and 'boots'. For more suitable and distinguishing, these attributes we designed are both easy to be recognized by human and machine. We can distinguish the differences between 'long sleeve', 'short sleeve' and 'sleeveless' easily by observing the length of the gussets, while the computer can recognize them through judging the location of the boundary of two colors of arm and gusset. We excluded those attributes which are easy for human to recognize but difficult or imprecise

for computers, like ‘male’ and ‘female’, ‘young’ and ‘old’, because these attributes bring more uncertainty for identification. We design an SVM classifier for each attribute, training with positive samples, labeled with 1, and negative samples, labeled with -1, Figure 3 shows a part of images of VIPeR.

(a) (b)

Figure 3. images in (a) have the attribute ‘back pack’, and images in (b) have the attribute ‘long hair’.

3.3.2. Hierarchical Attributes

In terms of objective laws, things in the world are in hierarchical structures, for example, animals can be classified into birds, dogs, cats and so on, and be more subdivided, birds can be classified into different kinds like dove, sparrow, canary and so on. Accordingly, human also perceive things hierarchically. We distinguish categories and then tell differences of those fine-grained ones [66]. When in the case of criminal description, witness usually remembers rough appearances of the criminal, and can’t think of details, like the texture of coat, but fine-grained appearance description can help narrow down the retrieval range in most instances.

Attributes of pedestrian have the hierarchical structure as well, and here we utilize it in pedestrians re-identification. We designed fine-grained attributes for several coarse-grained attributes according to the data sets we use. Of course in the practical application, they should be more detailed. We allocate ‘bald’, ‘short hair’ and ‘plate hair’ these three fine-grained attributes for coarse-grained attribute ‘not long hair’, fine-grained ‘shoulder-length hair’, ‘dishevelled hair’ and ‘ponytail’ for attribute ‘long hair’, fine-grained ‘non-stripe’, ‘dense stripes’ and ‘sparse stripes’ for ‘coat texture’, fine-grained ‘trousers’ and ‘long dress’ for ‘long bottom’, fine-grained ‘backpack’ and ‘single-shoulder bag’ for ‘has bag’, and fine-grained ‘handbag’, ‘take things’ and ‘take suitcase’ for attribute ‘take goods’. Figure 4 illustrates these fine-grained attributes with their coarse-grained attributes. We find that attributes of the same coarse-grained attribute usually look similar but have differences between each other. Coarse-grained attributes make a rough classification and then fine-grained attributes make a particular classification. Therefore, there are 17 coarse-grained attributes and 16 fine-grained attributes totally. We name our model as hierarchical attributes CDA-PR_HR(Combining Deep Learning and Attribute Learning for Pedestrians Re-identification with Hierarchical Attributes).

3.3.3. Attribute Learning

For supervised methods, a large amount of samples are needed when training models, but in the real world, it’s impossible to gather enough samples of each pedestrian. Attribute learning can overcome the difficulty of being lack of training samples, because we just need to judge whether the pedestrian has the attribute, instead of who he is. For the reason that attributes can be shared with different pedestrians, each attribute classifier can be trained with a lot of samples owing a same attribute, which is enough to generate well-trained classifiers.

After designing the attributes for pedestrian re-identification, we allocate an SVM for each attribute, and features extracted by CAEs are firstly input into 17 coarse-grained attribute SVMs, and then input into 16 fine-grained attribute SVMs. Before training one SVM, samples owing the attribute are labeled with 1 as positive ones and others labeled with -1 as negative ones. We select 9/10 of these positive and negative images as training samples to train the SVMs. The classifier output a posterior probability about owing the attribute, represented as $p(a_k|x)$. The prior probabilities of attributes are obtained by statistics, forming 2 probabilistic tables named

Figure 4. Samples of coarse-grained attributes and their fine-grained attributes. Fine-grained attributes of the same coarse-grained attributes look similar but have difference between each other.

'coarse-grained attributes-classes mapping' and 'fine-grained attributes-classes mapping'. Figure 5 shows an example about 10 attributes mapping to 10 classes, and each element of it means the probability of one sample belonging to the class y_j when it owing the attribute A_k(or a_i), representing with $p_1(y_j|A_k)$(or $p_2(y_j|a_i)$). A_k represents a coarse-grained attribute, while a_i represents a fine-grained attribute. We can calculate the posterior probability of the class according to the formula of conditional probability:

$$p_1(y_j \mid x_t) = \sum_{k=1}^{N} p_1(y_j \mid A_k) p_1(A_k \mid x_t) \tag{1}$$

$$p_2(y_j \mid x_t) = \sum_{i=1}^{M} p_2(y_j \mid a_i) p_1(a_i \mid x_t) \tag{2}$$

Here $p_1(y_j|x_t)$ and $p_2(y_j|x_t)$ represent the probabilities belonging to class y_j respectively according to the coarse-grained attributes $\{A_1,A_2,\cdots A_k,\cdots A_N\}$and fine-grained attributes $\{a_1,a_2,\cdots a_l,\cdots a_m\}$. Formula (3) computes the final probability based on the two partial probabilities with different weight ω_1 and ω_2, which represent the importance of coarse-grained and fine-grained attributes. If the witness can remember fine-grained attributes about the criminal, we set ω_1=0.6 and ω_2=0.4, for the reason that the coarse-grained attributes are the basis for re-identification and they should be allocated with higher weight. The weights make up it when the coarse-grained classifier makes a miss-recognition of one coarse-grained attribute but the fine-grained classifier recognizes correctly. And when the witness has none or few memory about the fine-grained attributes, ω_2 will be set as 0.

$$p(y_j \mid x_t) = \omega_1 p_1(y_j \mid x_t) + \omega_2 p_2(y_j \mid x_t) \tag{3}$$

In the end, we select the class with the maximum posterior probability as the prediction of the sample:

$$\hat{y} = \arg\max_{y_j} p(y_j \mid x_t) \tag{4}$$

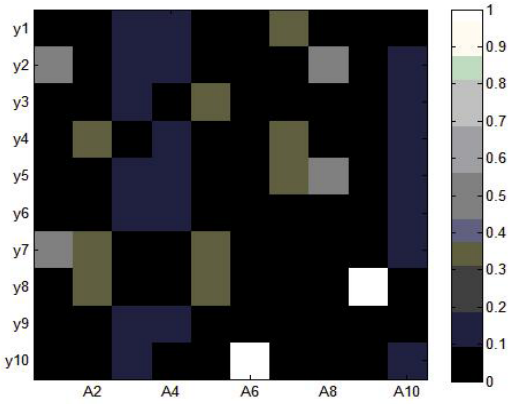
4. EXPERIMENTS AND DISCUSSION

4.1. Data Sets and Conditions

We selected three challenging data sets to validate our model, the CUHK02 data set [32], VIPeR data set introduced by Gray et al. [17], and i-LIDS data set [69].

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
y1	0	0	0.14	0.17	0	0	0.33	0	0	0
y2	0.5	0	0.14	0.17	0	0	0	0.5	0	0.14
y3	0	0	0.14	0	0.33	0	0	0	0	0.14
y4	0	0.33	0	0.17	0	0	0.33	0	0	0.14
y5	0	0	0.14	0.17	0	0	0.33	0.5	0	0.14
y6	0	0	0.14	0.17	0	0	0	0	0	0.14
y7	0.5	0.33	0	0	0.33	0	0	0	0	0.14
y8	0	0.33	0	0	0.33	0	0	0	1	0
y9	0	0	0.14	0.17	0	0	0	0	0	0
y10	0	0	0.14	0	0	1	0	0	0	0.14

(a)



(b)

Figure 5. Example of the‘attributes-classes mapping’ probability table. (a) illustrates the probabilities of 10 attributes mapping to 10 classes, and (b) visualizes these probabilities with different colors. The lighter color means a larger probability(white color represents 1), and the darker color means a small probability (black represents 0).

CUHK02 CUHK02 is a person re-identification data set with five camera view settings, containing 7264 images of 1816 different pedestrians. Each image is scaled to 160×60 pixel size, and in this experiment, to unify the size of input images, we re-scale them to 128×48 pixel size. This data set is usually used for evaluating re-identification algorithms under different camera view transforms, and here it is used to pre-train the CNNs to avoid overfitting for that it has enough samples for training parameters of the model.

VIPeR VIPeR is comprised of 1264 images of 632 pedestrians, each pedestrian has 2 images from two cameras with different viewpoints, poses and lighting conditions. Images of VIPeR are uniformly scaled to 128×48 pixel size. 632 pedestrians corresponding to 632 classes, and in experiments, classes of images are distinguished with numbers from 1 to 632. We took the cross-validation method mentioned in paper [18], dividing VIPeR data set into 10 portions, taking one portion for testing where there is no sample owing the same attributes, and the other 9 portions for training. We repeated the experiment for 10 times and took the average of them as the result.

i-LIDS i-LIDS contains 479 images of 119 pedestrians captured from non-overlapping cameras. Each pedestrian has 2 to 7 images, with different viewpoints, poses, lighting conditions, and some are also subject to occlusion. We scaled the images to 128×48 pixel size the same with CUHK02 and VIPeR. For transfer learning with zero training examples, we selected 10 pedestrians with each pedestrian owing 4 images to test our model. Classes of pedestrians of i-LIDS are also distinguished with numbers from 1 to 10.

For the reason that VIPeR has the largest amount of data, and the most diversiform attributes, we design attributes for pedestrian and train the CNNs on VIPeR data set. We pre-processes the dataset in the data preparation stage. The method based on the unsupervised CNN model and the pedestrian attribute involves the dataset including the pre-training dataset and the trimming dataset (target

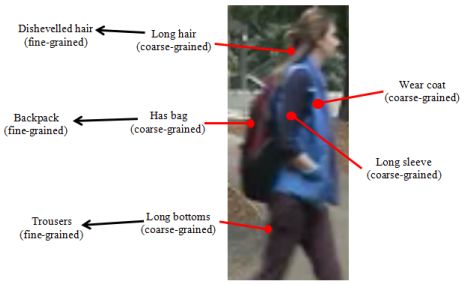


Figure 6. The sample has 5 coarse-grained attributes, 'long hair', 'wear coat', 'long sleeves', 'has bag' and 'trousers', and 3 fine-grained attributes, 'dishevelled hair', 'backpack and 'trousers'. These 8 attributes are labeled with 1 and others are labeled with -1.

dataset). Because different data sets have different image sizes and different data distributions, the data are uniformly processed before feature extraction and classification. The CNN model uses pre-training data sets to do pre-training, and then uses the target data set to fine tuning the parameters. We labeled all images with 17 coarse-grained attributes and 16 fine-grained attributes, here we take one image for example, Figure 6 shows the attributes of one pedestrian.

4.2. Preprocessing

To make the features extracted from CNN more relevant to the attributes, all images from above two data sets are segmented into 5 horizontal blocks. We segment images according to pedestrian's body structure and attributes we designed, Figure 7(a) takes an example of the 5 blocks. From top to bottom, the first horizontal block takes the vertical 1 to 36 pixels(the whole images are resized to 128×48 pixels), ensuring include the head and hair of pedestrians, corresponding to attributes 'wear hat', 'not long hair' and 'long hair', the second block takes the 14 to 73 pixels, including the upper body, corresponding to attributes 'wear coat', 'long sleeve', 'short sleeve', 'sleeveless', 'coat texture', 'has logo' and 'has bag', the third block takes the 36 to 104 pixels, including partial upper body and partial lower part of the body, corresponding to attributes 'take goods', the fourth block takes the 60 to 112 pixels, including legs, corresponding to attributes 'short bottoms', 'long bottoms' and 'bottom texture', and the fifth block take the 101 to 128 pixels, corresponding to the attributes 'sandal', 'shoes' and 'boots'. There are some overlaps among the 5 horizontal blocks as some attributes share the same part of body. Figure 7(b) shows image blocks of several pedestrians.

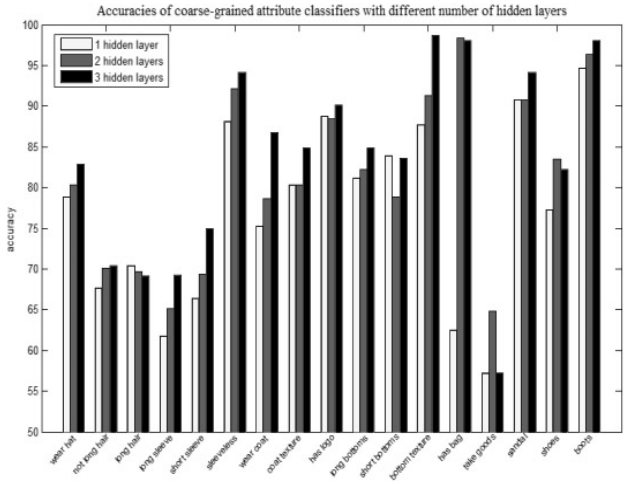
(a) (b)

Figure 7. Pedestrian images are segmented into 5 horizontal blocks according to pedestrian's body structure and attributes we design. Block1 includes the head and hair of pedestrians, block2 includes the upper body, block3 includes partial upper body and partial lower part of the body, block4 includes legs, and block5 includes the feet of pedestrian.

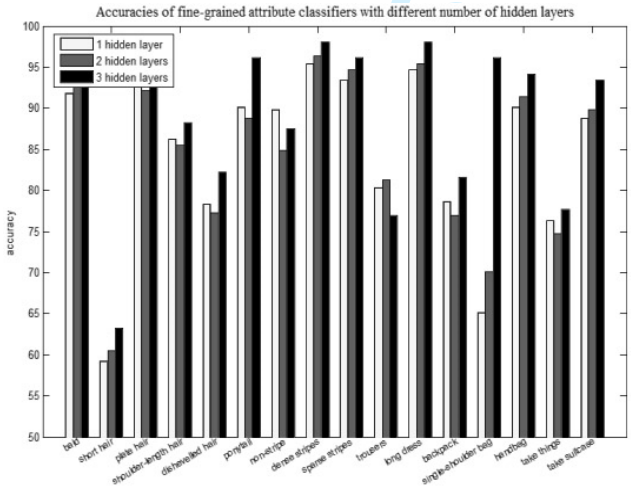
4.3. Results and Discussion

4.3.1. Accuracies of Attribute Classifiers with Different Number of Hidden Layers

We extract feature maps through different number of hidden layers(1 to 3 layers) to find the most suitable structure, and for each structure we all extract 100 final features maps for attribute classifiers. For example, we extract 100 feature maps through structure with only one hidden layer, and extract 100 feature maps through structure with 3 hidden layers as well. Figure 8 illustrates the accuracies of coarse-grained and fine-grained attribute classifiers got through 1 to 3 hidden layers, and from it we can see that features extracted from 3 hidden layers can achieve higher accuracies on most attributes, like 'shot sleeve', 'wear coat' and 'bottom texture' in coarse-grained attributes and 'dishevelled hair', 'ponytail' and 'single-shoulder bag' in fine-grained attributes. We didn't do experiment on structure with 4 hidden layers for the reason that feature maps after 4 hidden layers are too small to represent any information. The experiment result shows that structure with 3 hidden layers is better for the pedestrian re-identification on VIPeR.



(a)



(b)

Figure 8. Accuracies of attribute classifiers with different number of hidden layers. (a) illustrates the result of 17 coarse-grained attribute classifiers and (b) illustrates the result of 16 fine-grained attribute classifiers. Accuracies with 3 hidden layers, represented by the red line, perform better than other two in most of the attributes.

4.3.2. Weights of Layers

Figure 9 shows the visualization of 5 blocks' filters learned in the first convolution layer, and

each 2 rows represent different blocks of the pedestrian. (a) demonstrates the filters for the red channel of the RGB pedestrian image, while (b) demonstrating the filters for the green channel and (c) demonstrating the filters for the blue channel.We set 20 filters for the first hidden layer, and train the filters on the CUHK02 data set. Here, we randomly select 10 from all 20 filters of each block for exhibition. The filters show some simple point and line in the visualization for capturing low level features of the pedestrian.

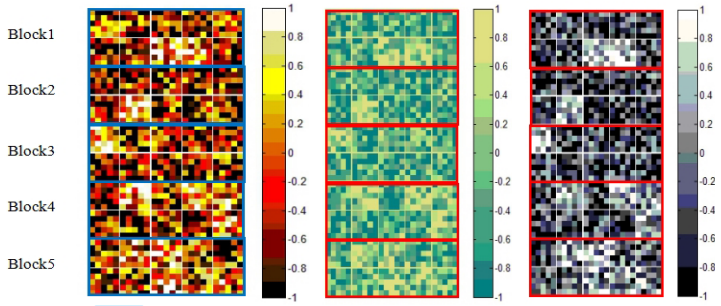


Figure 9. Visualization of 5 blocks' weights learned in the first convolution layer. From (a) to (c) are filters for red, green and blue these 3 color channels of the RGB pedestrian image, and we randomly select 10 from all 20 filters of each block.

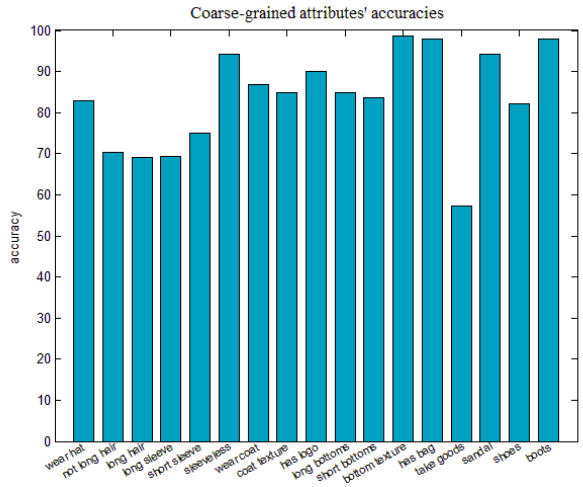
4.3.3. Accuracies of Coarse-grained and Fine-grained Attributes

Accuracies of all attribute classifiers with 3 hidden layers are shown in Figure 10, (a) shows the 17 coarse-grained attributes' accuracies and (b) shows the 16 fine-grained attributes' accuracies. Almost all accuracies of these attributes range from 60% to nearly 100%, and achieve mean values of 83.5% about coarse-grained attributes and 88.8% about fine-grained attributes. In coarse-grained attributes, low accuracies distribute in attributes like 'not long hair', 'long hair', 'long sleeve' and 'take goods'. We explain this situation for 3 reasons: (1)Images in VIPeR have a complex background, especially in the upper body region of the pedestrian, with trees and buildings. Segregating the background of the images will improve the performance; (2)Attributes about human hair are difficult to judge, especially for females with different hair styles; (3)It is tough for attribute classifier 'take goods' when pedestrians take phones, cups or other small things in low-resolution images. In fine-grained attributes, 'short hair' has a little lower accuracy for the reason that pedestrian with ponytail and plated hair looks like short hair in the front.

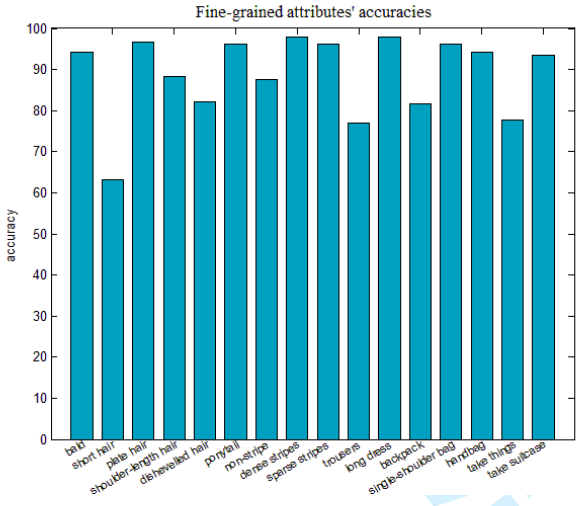
Attributes we designed, including coarse-grained and fine-grained attributes, are partly different from [30], but share some same or similar attributes of it. They are 'wear coat', 'has logo', 'handbag', 'backpack', 'shorts(short bottoms)' and 'trousers'. [30] selects conventional color and texture features as the input of SVM classifiers, color features including 8 channels of RGB, HSV and YCbCr, 21 texture features are exported from luminance channel. We compare the accuracies of the SVM classifiers about the 6 similar attributes and their accuracy on VIPeR data set in Table 1. Our model outperforms OAR in most of the 6 attributes: our model achieves a 90.1% accuracy in attribute 'has logo', about 30% higher than OAR, and nearly 40% improvement in attribute 'handbag'. Besides, in attributes 'wear coat', 'backpack', 'shorts(short bottoms)', our model also performs better than OAR, except in attribute 'trousers'.

4.3.4. Re-identification Rates of CDA-PR and CDA-PR_HA

We test our CDA-PR and CDA-PR_HA on 45 pedestrians of VIPeR, and the correctly matched rank is obtained. Rank-k recognition rate is the expectation of the matches at rank-k, and the Cumulative Match Characteristic(CMC) curve is the cumulated values of recognition rate at all ranks. Table 2 illustrates the rates of our CDA-PR model, CDA-PR with hierarchical attributes and other 7 models. Since Zhao et.al. [67], Umeda et.al. [56], Layne et.al. [30], KLFDA [60], Salience [68], Svmml [35] and RankBoost [51] have published their results on VIPeR, we take



(a)



(b)

Figure 10. Accuracies of all attribute classifiers. (a) illustrates the accuracies of 17 coarse-grained attribute classifiers, and (b) illustrates accuracies of 16 fine-grained attribute classifiers.

Table I. Comparison of accuracies about 6 arribute classifiers of 2 model, OAR[9] and our CDA-PR, and their mean accuracy on VIPeR data set.

Attributes	Methods	
	OAR	OURS
Wearcoat	69.7	86.8
Haslogo	60.8	90.1
Handbag	54.5	94.1
Backpack	68.6	81.6
Shorts (short bottoms)	76.1	83.6
Trousers	78.0	77.0
Meam	68.0	85.5

Table II. Comparison of top ranked matching rate between our CDA-PR method, CDA-PR_HA and other state-of-art results for the VIPeR dataset.

	Rank-1	Rank-5	Rank-10	Rank-20	Rank-25
CDA-PR	0.68	0.96	0.98	0.99	0.99
CDA-PR_HA	0.84	0.98	0.99	1	1
Zhao et.al. [67]	0.27	0.51	0.62	0.76	-
Umeda et.al. [56]	0.18	0.32	0.41	-	0.59
KLFDA [60]	0.32	0.66	0.78	0.91	-
Layne et.al. [30]	0.21	0.42	0.55	-	0.72
Salience [68]	0.3	0.52	-	-	-
Svmml [35]	0.3	0.63	0.77	0.88	-
RankBoost [51]	0.24	0.46	0.56	0.69	-

these results for comparison directly.And the later mentioned four results summed up by [50]. It is observed that our 2 performances based on model CDA-PR and CDA-PR_HA outperform all the three approaches with a high advantage. In particular, rank 1 matching rate is around 68% for CDA-PR and 84% for CDA-PR_HA, versus 27% for approach of Zhao et.al., 18% for approach of Umeda et.al. and 21% for approach of Layne et.al. The matching rate at rank 10 is around 98% for CDA-PR and 99% for CDA-PR_HA, versus 62% for approach of Zhao et.al., 41% for approach of Umeda et.al. and 55% for approach of Layne et.al.

CDA-PR_HA achieves better accuracy than CDA-PR for the reason that fine-grained attributes supply more clue for re-identification when some pedestrians share the same coarse-grained attributes. Figure 11 illustrates matching situations of 4 samples with single coarse-grained attributes and combination of coarse-grained and fine-grained attributes. (a) illustrates the ranking of top 10 pedestrians with single coarse-grained attributes and (b) illustrates the ranking of top 10 pedestrians with hierarchical attributes. The ranking of matching pedestrian, marked with red rectangle, in (b) is usually higher than that in (a). For example, the third sample achieves a fifth ranking with single coarse-grained attributes, but achieves a second ranking with hierarchical attributes.

Figure 11. Examples of Person Re-identification on VIPeR using CDA-PR and CDA-PR_HA. In each row, the left-most image is the test sample, (a) and (b) illustrate the top 10 matched pedestrians with single coarse-grained attributes and hierarchical attributes respectively. Highlighted red boxes are used to mark the correctly matched pedestrians.

Our model achieved an accuracy about 67.8% of CDA-PR and 83.3% of CDA-PR_HA in final pedestrian re-identification on VIPeR, and an accuracy of 35% of CDA-PR and 40% of CDA-PR_HA on i-LIDS data set with zero training samples. We also compared with other 7 pedestrians re-identification approaches. These include 4 attribute-based methods, OAR [30],

Table III. Comparison of 7 models' accuracies on VIPeR data set and i-LIDS data set. CDA-PR and CDA-PR_HA outperform all 7 models.

Methods	data sets	
	VIPeR	i-LIDS
CDA-PR	67.8	35
CDA-PR_HA	83.3	40
OAR[13]	21.4	
Umeda et.al. [56]	18.0	
SDC_knn [30]	26.3	
SDC_ocsvm [30]	26.7	
Ahmed et.al. [1]	34.8	
AIR [29]	-	11.5
W.AIR [29]	-	16.5

Umeda's method [56], SDC_knn and SDC_ocsvm [30], and a method based on deep learning introduced by Ahmed et.al [1]. Table 3 illustrates accuracies of these 7 models, from which we learned that our model is the second best, just lower than model proposed by Ahmed et.al. [1] takes an improved deep learning method and pre-train the model on CUHK01 [33] and CUHK03 data sets [34], achieving a 34.8% rank-1 accuracy on VIPeR. Although the model of Ahmed performs better than other 4 approaches, the character of being lack of semanteme limits its application in real life. [29] validated their AIR and weighted AIR models on i-LIDS data set, achieving zero-shot transfer learning re-identification rate as 11.5% and 16.5%, lower than our CDA-PR and CDA-PR_HA models. Experiments prove that our model based on deep learning and attribute learning significantly improves the accuracy while strengthening the semantic representation of pedestrians re-identification. And hierarchical attributes help the model to achieve higher accuracy and make it more flexible and practical in application.

4.3.5. Performance on Attribute Absent

We also test the performance of our CDA-PR method on images with attribute absent, usually caused by the variances of pose and illumination and sometimes occlusion. For VIPeR and i-LIDS data sets, the major absent attributes are bags, textures, logos (because of the pose of pedestrians) and shoes (because of the angle of the camera). We select 8 samples according to these four kinds of attribute absent problems from data sets, each allocated 10 images. Figure 12 gives 8 pair of examples of the 4 kinds of attribute absent. Each example represents one pedestrian with 2 images, the left image has the attribute marked with highlighted red boxes and the right image is lack of the attribute because of different illumination, poses of pedestrian, or angles of camera. In experiment, 4 sets of 40 pedestrian samples are tested, each set includes 5 pairs of pedestrians and each pair includes one sample of attribute absence and another of non-absence. Table 4 illustrates the final re-identification rates of these samples with different kinds of attribute absences or non-absences. In particular, 'A/5'in table means that A samples are re-identified correctly in 5 samples of this kind of attribute absence or non-absence. CDA-PR achieves a rate of 55% on samples with attribute absence and a rate of 60% on samples with attribute non-absence, and CDA-PR_HA achieves a rate of 80% on samples with attribute absence and a rate of 85% on samples with attribute non-absence. This is a good record for problem of attribute absence.'Bag absence' has little affection on the rate of final result for the reason that straps of bags also help re-identifying.

Table IV. Final rates of attribute absence and non-absence with CDA-PR and CDA-PR_HA.

	CDA-PR		CAD-PR_HA	
	Absent	**Non-absent**	**Absent**	**Non-absent**
Logo absent	0/5	1/5	2/5	3/5
Bag absent	5/5	5/5	5/5	5/5
Texture absent	3/5	2/5	4/5	4/5
Shoes absent	3/5	4/5	5/5	5/5
Rate(%)	55	60	80	85

Figure 12. Samples of 4 kinds of attribute absent. Each kind of attribute absent shows 2 pair of examples, and in each example, the right image is the sample that be lack of the attribute marked with highlighted red boxes in the left image.

5. CONCLUSION

Pedestrians re-identification is a task facing a great challenge of occlusion and variations of viewpoints, poses and illumination. In this paper, we propose a model based on deep learning and attribute learning for pedestrians re-identification, taking the advantages of both feature extraction of deep learning and semantic representation of attribute learning. Experiments show that our model greatly improves the accuracy of re-identification and has a good practicability in application. And hierarchical attribute we proposed is proved to improve the accuracy and practicability of the pedestrian re-identification in practical application. Attributes are crucial for pedestrians re-identification. Our next step is to get better attributes automatically from data to represent pedestrians. [46].

ACKNOWLEDGMENT

This research is supported by the Natural Science Foundation of China No.61602215, the science foundation of Jiangsu province No.BK20150527, China State Scholarship Fund No.201608320098 and International Postdoctoral Exchange Fellowship Program No.201653.

REFERENCES

1. E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.
2. Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 35(8):1798–1828, 2013.
3. Y. Boureau, F. R. Bach, Y. Lecun, and J. Ponce. Learning mid-level features for recognition. pages 2559–2566, 2010.
4. Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. pages 5315–5324, 2015.
5. K. Cheng, K. Hui, Y. Zhan, and M. Li. Sparse representations based distributed attribute learning for person re-identification. *Multimedia Tools and Applications*, pages 1–23, 2017.
6. K. Cheng and X. Tan. Sparse representations based attribute learning for flower classification. *Neurocomputing*, 145(18):416C426, 2014.
7. K. Cheng, Y. Zhan, and M. Qi. Al-ddcnn a distributed crossing semantic gap learning for person re-identification. *Concurrency & Computation Practice & Experience*, 29, 2016.
8. L. Cheng. *Research on key technology of person re-identification*. PhD thesis, Beijing University of Posts and Telecommunications, 2013.
9. M. Chu and N. Thuerey. Data-driven synthesis of smoke flows with cnn-based feature descriptors. 36(4), 2017.
10. E. G. Danaci and N. Ikizler Cinbis. A comparison of low-level features for visual attribute recognition. In *Signal Processing and Communications Applications Conference*, pages 2038–2041, 2015.
11. E. G. Danaci and N. Ikizler-Cinbis. Low-level features for visual attribute recognition: An evaluation. *Pattern Recognition Letters*, 84:185–191, 2016.
12. E. Ergul, m. Karayel, MehErgul2016UNSUPERVISED, O. Timu?, and E. Kiyak. Unsupervised feature learning for mid-level data representation. 12:51–79, 03 2016.
13. C. Fan, Y. Chen, L. Cao, and Y. Miao. Person re-identification based on visual perceptual model. *Computer Engineering & Applications*, 2016.
14. M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. pages 2360–2367, 2010.
15. I. Filkovic, Z. Kalafatic, and T. Hrkac. Deep metric learning for person re-identification and de-identification. In *International Convention on Information and Communication Technology, Electronics and Microelectronics*, pages 1360–1364, 2016.
16. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
17. D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 41–47, 2007.
18. D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision - ECCV 2008, European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings*, pages 262–275, 2008.
19. H. He, T. Watson, C. Maple, J. Mehnen, and A. Tiwari. A new semantic attribute deep learning with a linguistic attribute hierarchy for spam detection. In *International Joint Conference on Neural Networks*, pages 3862–3869, 2017.
20. S. Hoochang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285, 2016.
21. Z. H. Huo and Y. Chen. Person re-identification based on multi-salience fusion. *Opto-Electronic Engineering*, 42(9):41–47, 2015.
22. Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan. Caffe: Convolutional architecture for fast feature embedding. pages 675–678, 2014.
23. G. Khodabandelou, C. Hug, R. Deneckre, and C. Salinesi. *Supervised vs. Unsupervised Learning for Intentional Process Model Discovery*. Springer Berlin Heidelberg, 2014.
24. D. J. Kim, D. Yoo, S. Im, N. Kim, T. Sirinukulwattana, and I. S. Kweon. Relative attributes with deep convolutional neural network. In *International Conference on Ubiquitous Robots and Ambient Intelligence*, pages 157–158, 2015.
25. D. J. Kim, D. Yoo, S. Im, N. Kim, T. Sirinukulwattana, and I. S. Kweon. Relative attributes with deep convolutional neural network. In *International Conference on Ubiquitous Robots and Ambient Intelligence*, pages 157–158, 2015.
26. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.
27. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.
28. M. Langkvist, L. Karlsson, and A. Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42(1):11–24, 2014.
29. R. Layne, T. Hospedales, and S. Gong. Person re-identification by attributes. In *BMVC*, volume 2, page 8, 2012.
30. R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. *Person Re-Identification. Advances in Computer Vision and Pattern Recognition*, pages 93–117, 2014.
31. Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
32. W. Li and X. Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.
33. W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44, 2012.
34. W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

35. 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 Conference on Computer Vision and Pattern Recognition*, pages 3610–3617, 2013.
36. W. Lin, Y. Shen, J. Yan, M. Xu, J. Wu, J. Wang, and K. Lu. Learning correspondence structures for person re-identification. *IEEE Transactions on Image Processing*, 26(5):2438–2453, 2017.
37. W. Lin, M. T. Sun, R. Poovendran, and Z. Zhang. Activity recognition using a combination of category components and local models for video surveillance. *IEEE Transactions on Circuits & Systems for Video Technology*, 18(8):1128–1139, 2008.
38. J. Markowitz, A. C. Schmidt, P. M. Burlina, and I. J. Wang. Hierarchical zero-shot classification with convolutional neural network features and semantic attribute learning. In *Fifteenth Iapr International Conference on Machine Vision Applications*, pages 194–197, 2017.
39. J. Masci, U. Meier, D. An, J. Schmidhuber, and rgen. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59, 2011.
40. C. J. Mattingly, R. Boyles, C. P. Lawler, A. C. Haugen, A. Dearry, and M. Haendel. Laying a community-based foundation for data-driven semantic standards in environmental health sciences. *Environmental Health Perspectives*, 124(8):1136–1140, 2016.
41. W. F. Michal. Orientation histograms for hand gesture recognition. *Mitsubishi Electric Research Labs*, pages 296–301, 1995.
42. W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship in pedestrian detection. In *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pages 3222–3229, 2013.
43. D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, pages 503–510, 2011.
44. E. Park, X. Han, T. L. Berg, and A. C. Berg. Combining multiple sources of knowledge in deep cnns for action recognition. In *IEEE Winter Conference on Application of Computer Vision*, pages 1–8, 2016.
45. P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.
46. A. Przybyek. *Systems Evolution and Software Reuse in Object-Oriented Programming and Aspect-Oriented Programming*. Springer Berlin Heidelberg, 2011.
47. J. L. Qu, C. F. Du, Y. Z. Di, F. Gao, and C. R. Guo. Research and prospect of deep auto-encoders. *Computer and Modernization*, (9):128–134, 2014.
48. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *Eprint Arxiv*, 2013.
49. P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun. Pedestrian detection with unsupervised multi-stage feature learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.
50. Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *IEEE International Conference on Computer Vision*, pages 3200–3208, 2015.
51. V. Shet, S. Khamis, and C. H. Kuo. Person re-identification using semantic color names and rankboost. In *IEEE Workshop on Applications of Computer Vision*, pages 281–287, 2013.
52. B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 801–808, 2011.
53. Y. Souri, E. Noury, and E. Adelimosabbeb. Deep relative attributes. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.
54. C. Su, F. Yang, S. Zhang, and Q. Tian. Multi-task learning with low rank attribute embedding for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3739–3747, 2015.
55. N. T. People analysis cctv investigator handbook. *Home Office Centre of Applied Science and Technology*, 2(3):1–35, 2011.
56. T. Umeda, Y. Sun, G. Irie, K. Sudo, and T. Kinebuchi. *Attribute Discovery for Person Re-Identification*. Springer International Publishing, 2016.
57. R. N. J. Veldhuis, A. M. Bazen, W. Booij, and A. J. Hendrikse. A comparison of hand-geometry recognition methods based on low- and high-level features. *Stw/nwo/dutch Ministry of Economic Affairs*, 2017.
58. Z. Wang, R. Hu, Y. Yu, C. Liang, and W. Huang. Multi-level fusion for person re-identification with incomplete marks. *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1267–1270, 2015.
59. H. Wu, X. Kong, H. Yang, and Y. Li. Sparse feature preservation for relative attribute learning. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 385–390, 2016.
60. F. Xiong, M. Gou, O. Camps, and M. Sznaier. *Person Re-Identification Using Kernel-Based Metric Learning Methods*. Springer International Publishing, 2014.
61. Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. pages 1798–1807, 2015.
62. W. Yang, X. Yin, and G. S. Xia. Learning high-level features for satellite image classification with limited labeled samples. *IEEE Transactions on Geoscience and Remote Sensing*, 53(8):4472–4482, 2015.
63. M. Ye, C. Liang, Z. Wang, Q. Leng, J. Chen, and J. Liu. Specific person retrieval via incomplete text description. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 547–550, 2015.
64. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
65. C. Zhang, R. Li, Q. Huang, and Q. Tian. Hierarchical deep semantic representation for visual categorization. *Neurocomputing*, 2017.
66. X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking neural activations for fine-grained recognition. *IEEE Transactions on Multimedia*, PP(99):1–1, 2017.
67. R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

68. R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision*, pages 2528–2535, 2014.
69. W. S. Zheng, S. Gong, and T. Xiang. Associating groups of people. *Active Range Imaging Dataset for Indoor Surveillance*, 2009.
70. W. S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 35(3):653–668, 2013.
71. J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics*, pages 535–540, 2015.