Multi-Pose Learning based Head-Shoulder Re-identification

Jia Li[1,2], Yunpeng Zhai[3], Yaowei Wang[4]*, Yemin Shi[2], Yonghong Tian[2]
[1]School of Electronic and Computer Engineering,
Shenzhen Graduate School at Peking University, Shenzhen, China
[2]National Engineering Laboratory for Video Technology, Peking University, Beijing, China
[3]School of Information and Communication Engineering, BUPT, China
[4]School of Information and Electronics, Beijing Institute of Technology, China

Abstract

The whole body of person is probably invisible in video surveillance because of occlusion and view angles (such as in crowded public places), on which occasion conventional person re-identification (i.e., whole-body based Re-ID) approaches may not work. To address this problem, we propose a novel deep pairwise model based on multi-pose learning (MPL) which aims at head-shoulder part instead of the whole body. The proposed method explicitly tackles pose variations by learning an ensemble verification conditional probability distribution about relationship among multiple poses. To facilitate the research on this problem, we contribute three head-shoulder datasets based on CUHK03, CUHK01 and VIPeR. Experiments on these datasets demonstrate that our proposed method achieves the state-of-the-art performance.

1. Introduction

Person re-identification (Re-ID) aims to identify a person of interest across non-overlapping camera views. Because of its great potentials for public security and video surveillance applications, person Re-ID has drawn more and more attention in recent years. Despite decades of studies, most of existing Re-ID approaches rely on the image of person's whole body, while, which are probably not available by reason of occlusion and viewpoint variations in real-world surveillance (such as in subway station or inside public transportation).

To address the problem of no access to whole body images in crowded conditions, one feasible solution is to design a model aiming at person head-shoulder part instead of the whole body. Unfortunately, as we know there is no method focusing solely on head-shoulder part. On one hand, existing whole-body based person Re-ID methods cannot make sense in crowded conditions for lack of available whole-body images. On the other hand, these approaches also don't work on head-shoulder part since they cannot take advantage of its special characteristics. As shown in Fig. 1, head-shoulder part is more sensitive to pose variations because it has less information than the whole body.

Figure 1. Examples of pose variation on head-shoulder images and whole-body images. Head-shoulder part is more sensitive to pose variation since the whole body has more details.

In this paper, we propose a novel deep multi-pose learning framework aiming at head-shoulder part to solve person Re-ID problem in crowded conditions. Inspired by deep pairwise re-id systems, we further consider that relationships between image pairs with different poses may have different patterns. For instance, front-front pose pair images have more details than back-back pose pairs. As a result, we simultaneously split data into groups by pose pairs and train similarity classifier for each. Specifically, we collect person head-shoulder data cropped from benchmarks, and then group the training data by predicting their pose using a trained naive pose classifier. For each group, we model the relationship between the image pair and learn its similarity probability by a specific branch in a CNN model. Motivated

*Corresponding author: Yaowei Wang (yaoweiwang@bit.edu.cn)

0-7695-6354-6/18/$31.00 ©2018 IEEE
DOI 10.1109/MIPR.2018.00057

by Scott et al.[12] , we take advantage of the pose prediction to bootstrap the train phase. Moreover, since the naive pose classifier may get wrong prediction and sometimes poses are ambiguous, we learn an ensemble conditional probability of n pose pair classes and pair similarity in the end of our CNN model.

The main contributions can be summarized as follows:

- We present a head-shoulder based person Re-ID method aiming at surveillance in crowded public places where the whole body is always not visible.
- We propose a novel deep multi-pose learning based model for head-shoulder Re-ID. It builds a conditional probability distribution model to leverage the different relationships of different pose-pairs mode.
- To facilitate the research on head-shoulder Re-ID, we contribute three head-shoulder datasets.

2. Related Work

2.1. Person Re-identification

Most existing person Re-ID methods include two stages: learning a new discriminative feature representation for the input image [3, 19, 18, 16], and learning an improved similarity metric for comparing features across images [11, 15]. In recent years, several Re-ID CNN architectures [2, 4, 13, 14, 17] have been proposed and they achieve phenomenal results on several benchmarks (CUHK03[9], CUHK01[8]) . The first Siamese CNN architecture used for Re-ID was proposed in [17] for metric learning. It consists of three independent convolutional networks for different overlapping regions and features are combined by using a cosine similarity as the connection function. E. Ahmed *et al.* [2] proposed an element that computes cross-input neighborhood differences, capturing local relationships between pair inputs. A recent work [14] also attempts to model the cross-view relationships by jointly learning subnetworks to extract the single image along with the cross image representations. These learned models aim at the whole body cannot be directly applied to head-shoulder data. Different from works all above, our proposed deep network solely focuses on head-shoulder part, which no more needs the whole-body images.

There are some person Re-ID methods in crowds. S. M. Assari *et al.*[5] models multiple personal, social and environmental constraints on human motion across cameras, which requires the information of relationship between cameras and human motion trajectory. Deep Filter Pairing Neural Network (FPNN) was introduced in [9] to jointly handle misalignment, photometric and geometric transformations, occlusion and cluttered background. None of these methods consider the influence of pose variation or utilize head-shoulder part to solve the problem of person Re-ID in crowds.

2.2. Multi-pose learning

Multi-pose learning method is a solution for face detection and recognition. In [1] and [10], a face image is processed by several pose-specific deep CNN models to generate multiple pose-specific features using 3D rendering. These multiple pose-specific features help reduce the effect of pose variation. However, it's hard to build a 3D model of head-shoulder part because of its high deformability.

3. Methodology

3.1. Network Architecture

Given an pair of images as input, our goal is to determine whether they belong to the same person. We regard the head-shoulder Re-ID problem as binary classification and take image pair as input. We pair the images in training data set with label 1 for positive (image pair belong to same person) or 0 for negative (image pair belong to different people). Fig. 2 illustrates our network architecture. Motivated by [2], we designed a pairwise-CNN network which takes two images as input. The network can be divided into two parts. First we use two sharing-weights branches to generate discriminative descriptors of two images. To use the model pretrained by ImageNet, we use GoogleNet as the first part of our network and take layer "pool5/7x7_s1" as its output. In our work, we use the product of two features to roughly represent the relationship between two pair-input images.

Secondly, we have two ways to learn the similarity. Simply, a fully connected layer and a softmaxlayer are concatenated to learn the similarity directly. We test it as baseline in section 4. Moreover, in multi-pose learning way, we build six branches corresponding six specific pose-pairs mode(left/right-left/right, left/right-front, left/right-back, front-front, front-back, back-back). Each branch includes a fully connected layer to reduce the number of dimensions, each of which concatenates another fully connected layer. The top six fully connected layers are used for learning the probability of similarity (pair input images belongs to same person or not in case of specific pose-pair mode) . On the other hand, two more fully layers as pose-pair mode classifier are concatenated to A$\times$B. And finally a proposed ensemble conditional probability loss function is used to learn similarity. It will be introduced in section 3.2.

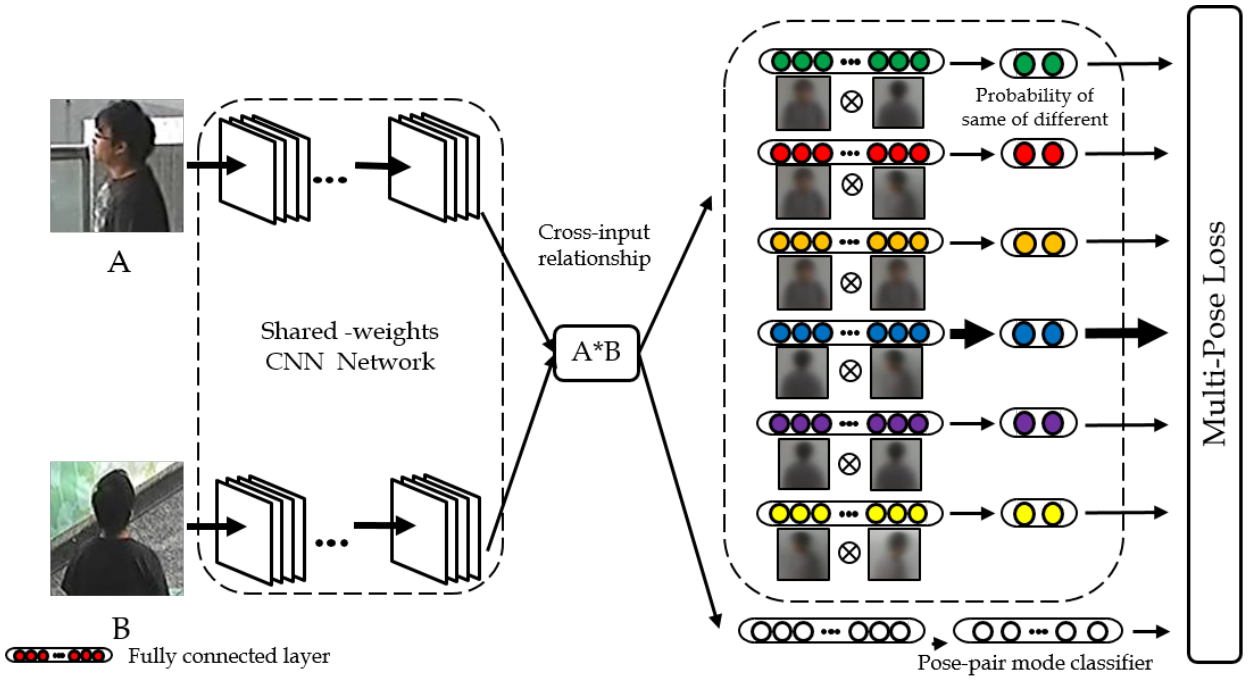

Figure 2. The architecture of proposed multi-pose learning based head-shoulder Re-ID model. Each pair input will go through all branches even though just one branch is perfect matched. Multi-pose loss of learning conditional probability gives heavier weight to the matched branch in back propagation.

Figure 3. Examples of head-shoulder images which are hard to judge their poses.

3.2. Deep Multi-pose Learning

Given training data (**x**, y*,z*, $\boldsymbol{f_1}$, $\boldsymbol{f_2}$), where $\boldsymbol{f_1}$, $\boldsymbol{f_2}$ indicate two descriptors output from our network of two input images. **x** indicates the representation of the relationship of that two descriptors. $\mathbf{x} = \boldsymbol{f_1} \times \boldsymbol{f_2}$. y* denotes the verification labels $\in$ 0, 1(0 indicates different people, 1 indicates the same person). z* denotes the pose-pair labels $\in$ 1,...,K(left-left, left-front,etc). Since sometimes it is hard to distinguish the pose of a person, such as shown in Fig. 3, we use a conditional probability form to describe the relationship of the two input images. The probability can be computed as following:

$$p(y^* = j|\mathbf{x}) = \sum_i p(y^* = j|z^* = i, \mathbf{x}) p(z^* = i|\mathbf{x}) \quad (1)$$

The model is trained by maximizing the cross entropy between the verification label $\overline{y}$ and the model prediction given by Eqn. 1. The cost function to minimize is

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{n=1}^{N} log \sum_{i=1}^{K} p(y_n^* = \overline{y}_n | z_n^* = i, \mathbf{x}_n) p(z_n^* = i|\mathbf{x}_n) \quad (2)$$

where N is the number of training samples. And $p(z_n^* = i|\mathbf{x}_n)$ indicates the probability that relationship $\mathbf{x}_n$ of two images belong to the i-th pose-pair mode, $p(y_n^* = \overline{y}_n|z_n^* = i, \mathbf{x}_n)$ indicates the correct predicted probability that whether the two images belong to the same person in case of the i-th pose-pair mode (In our MPL network, we use several specific classifiers for different pose-pair mode to learn these probability distribution). We simply refer $p(z_n^* = i|\mathbf{x}_n)$ as Z_i^n and refer $p(y_n^* = j|z_n^* = i, \mathbf{x}_n)$ as Y_{ji}^n.

Denote$\{\alpha_i^n|i = 1...K\}$ as the outputs of pose-pair mode's classifiers. $\{\beta_{ji}^n|i = 1...K, j = 1, 2\}$ as the outputs of K verification classifiers.

We assume $\{Y_{ji}^n\}$ obey index distribution. We use softmax function to calculate and normalize the probabilities. Y_{ji}^n and Z_i^n would be written as:

$$Z_i^n = \frac{\alpha_i^n}{\sum_t^K \alpha_t^n}, \quad Y_{ji}^n = \frac{exp(\beta_{ji}^n)}{\sum_t^2 exp(\beta_{ti}^n)} \tag{3}$$

By combining Eq.2 and Eq.3, we formulate our model as following optimization problem:

$$min_\theta \mathcal{L}^{id} = -\frac{1}{N}\sum_{n=1}^{N} log \sum_i^K Y_{g_n i}^n Z_i^n \tag{4}$$

Where g_n means the true verification label of the n-th input pair. The problem can be optimized using gradient descent. The partial derivatives of the inputs α_i^n are:

$$\frac{\partial L}{\partial \alpha_i^n} = \frac{\partial L}{\partial Z}\frac{\partial Z}{\partial \alpha_i^n} = \sum_j^d \frac{\partial L}{\partial Z_j^n}\frac{\partial Z_j^n}{\partial \alpha_i^n} \tag{5}$$

$$\begin{aligned} \frac{\partial Z_j^n}{\partial \alpha_i^n} &= \frac{\partial}{\partial \alpha_i^n}\left(\frac{\alpha_j^n}{\sum_t^K \alpha_t^n}\right) \\ &= \frac{1\{i=j\}}{\sum_t^K \alpha_t^n} - \frac{\partial Z_j^n}{\sum_t^K \alpha_t^n} \end{aligned} \tag{6}$$

$$\begin{aligned} \frac{\partial L}{\partial Z_j^n} &= -\frac{\partial}{\partial Z_j^n}\left(log \sum_i^K Y_{g_n i}^n Z_i^n\right) \\ &= -\frac{Y_{g_n j}^n}{\sum_i^K Y_{g_n i}^n Z_i^n} \end{aligned} \tag{7}$$

and similarly β_{ji}^n:

$$\frac{\partial L}{\partial \beta_{ji}^n} = \begin{cases} \frac{Z_i^n Y_{ji}^n}{\sum_t^K Y_{g_n t}^n Z_t^n}(Y_{g_n i}^n - 1) & , j = g_n \\ \frac{Z_i^n Y_{ji}^n}{\sum_t^K Y_{g_n t}^n Z_t^n}(Y_{g_n i}^n) & , j \neq g_n \end{cases} \tag{8}$$

Also we learn the probability distribution of pose-pair mode which is defined as following:

$$\begin{aligned} P_i^n &= \frac{exp(\alpha_i^n)}{\sum_t^K exp(\alpha_t^n)} \\ min_\theta \mathcal{L}^{pose} &= -\frac{1}{N}\sum_{n=1}^{N} log P_{h_i}^n \end{aligned} \tag{9}$$

where h_i indicates the pose-pair mode label of sample i.

By combining Eq. 4 and Eq. 10, we formulate our MPL model as the following optimization problem:

$$min_\theta \mathcal{L} = \mathcal{L}^{pose} + \gamma \mathcal{L}^{id} \tag{10}$$

where $\gamma(\gamma > 0)$ is a regularization parameter. In our experiment γ is set to 0.3 in order to enhance identification ability of our model.

Figure 4. Examples of head-shoulder dataset cropped from VIPeR and CUHK03.

4. Experiments

4.1. Datasets and settings

In order to demonstrate the effectiveness of our new framework, we conduct a set of experiments to compare with state-of-art person Re-ID methods. Since there is not any head-shoulder based person Re-ID dataset, we collect our datasets from public person Re-ID datasets(CUHK01, CUHK03, VIPeR in this work) and resize the images to 224×224. We automatically cropped the head-shoulder part from the raw data as shown in Fig. 4.

CUHK03 [9] contains 13164 images of 1360 individuals collected on the CUHK campus. Authors in [9] provide both manually labeled bounding boxes and automatically detected bounding boxes obtained by running a detector. All the experiments presented in this paper use the labeled bounding boxes. All images with varying size are resized as 384×128 and then cropped the top 128×128 pixels to build our CUHK03 head-shoulder dataset. Following the splitting settings provided in [9], evaluation is conducted 20 times with 1260 identities for training and 100 identities for testing, and finally generate the average results.

CUHK01 [9] contains 3884 images of 971 individuals collected on the CUHK campus. All images are resized as 384×128 and then cropped the top 128×128 pixels to build our CUHK01 head-shoulder dataset. Following the splitting settings provided in[8], evaluation is conducted 20 times with randomly choosing 871 identities for training and 100 identities for testing, and finally generate the average results.

VIPeR [6] contains 1264 images of 632 individuals (two images per individual) captured using 2 cameras. Images are all scaled to 128×48. We cropped the top 48×48 pixels of each image and resize it to 128×128 to build our VIPeR head-shoulder dataset. We randomly split the data set into half, 316 individuals for training and 316 for testing with no

Dataset	VIPeR	CUHK03	CUHK01
labeled/ all training images	100/1264	1000/13164	400/3884

Table 1. The number of pose training data

Method	Head-shoulder	
	Rank 1	Rank 5
KISSME	9.0	30.7
kLFDA	31.3	62.9
DGD	45.6	75.7
Ours - Baseline	**47.0**	**85.0**
Ours - With MPL	**50.0**	**86.0**

Table 2. Performance comparison of state-of-the-art algorithms on CUHK03 dataset. The proposed baseline Pairwise-CNN beats other methods tested on head-shoulder CUHK03 dataset. The proposed pairwise-CNN with MPL obtains better results than baseline.

overlapping on person identities. The process is repeated 10 times, and the averaged performance is reported as the final result. On account of the much smaller number of images and identities compared to other datasets, we fine-tune the model trained by CUHK03 for testing VIPeR dataset.

Pose Labeling: We divide the training data into 3 groups (profile, front, back) according to view angle. Specifically, part of face can be seen in a head-shoulder image of pose left-right. Almost the whole face can be seen in front pose images while no face in back pose images.

Since the huge number of training data, it's hard to label them manually. In this work, our model uses a form of probability to represent pose so that we don't need 100% accurate pose labels. We just labeled a small number of the training data and then use them to train a Support Vector Machine classifier to generate the pose label of each unlabeled training image (Details of training data number is in Table 1).

All images are normalized to 60×60 pixels and we extract HOG descriptors of all images in each pose training group. Specifically, we extract HOG descriptor with 4×4 pixel cells and 6 orientations. Finally we get a 4950 dimension descriptor of each image as input of SVM classifier.

4.2. Results and Discussion

As far as we know, it's the first work of Re-ID based on person head-shoulder part. No existing result or model can be compared directly, so we compare our proposed method with other state-of-art Re-ID methods in two ways. In the first way, we compare our model with other public models both trained and tested on head-shoulder dataset. We retrained other models using the same code and settings as they published. More details are discussed later. In the second way, we compare our model trained and tested on head-shoulder data with other public models trained and tested on the whole-body data.

Method	Head-shoulder	
	Rank 1	Rank 5
KISSME	10.7	24.2
kLFDA	43.5	73.5
DGD	56.6	78.2
Ours - Baseline	**60.0**	**88.0**
Ours - With MPL	**63.9**	**90.7**

Table 3. Performance comparison of state-of-the-art algorithms on CUHK01 dataset. Trained on head-shoulder CUHK01 data, our approach obtains better results than others.

In the first way, we compare our method to KISSME[7], kLFDA[11], and a deep Re-ID method DGD[14] on our head-shoulder dataset. Table 2, Table 3, Table 4 show the results. We retrained and tested KISSME, kLFDA, DGD models using our three proposed head-shoulder dataset. While testing kLFDA, we use same kernels and other settings as proposed in kLFDA. Finally, we choose the best results. While testing DGD, we resize head-shoulder images to 64*160, same as in DGD. We also use the same network setting to guarantee the fairness of comparison.

It shows that results of KISSME and kLFDA have low accuracy rate at rank-1 on head-shoulder dataset. It proves that existing whole-body based Re-ID models don't work on head-shoulder part. These results demonstrate that our method has better performance than others including deep methods.

All previous methods have low results on head-shoulder VIPeR dataset because of the low resolution of VIPeR data. Head-shoulder part for recognition lacks much information of a person. But our approach can take advantage of pose information and achieve better performance of 16.2 at rank-1. It demonstrates our method is more suitable for head-shoulder based Re-ID.

In the second way, we compare our method against KISSME, eSDC[19], SDALF[3], and two deep Re-ID method FPNN, IDLA[2] on CUHK01. Table 5 shows the results. Even though head-shoulder images have less information than whole-body images, our model beats others, and nearly catches up with the deep re-id models. It proves that our method is useful for Re-ID in crowded places.

All of these experiments show our method with multi-

Method	Head-shoulder	
	Rank 1	Rank 5
KISSME	10.9	32.9
DGD	7.5	25.4
Ours - Baseline	**14.1**	**37.9**
Ours - With MPL	**16.2**	**40.5**

Table 4. Performance comparison of state-of-the-art algorithms on head-shoulder VIPeR dataset. The proposed pairwise-CNN with MPL obtains better results than baseline at various ranks.

Method	Whole-body
	Rank 1
SDALF	9.9
eSDC	22.83
KISSME	29.4
FPNN	27.8
IDLA	65.0
Ours - Baseline	**60.0**
Ours - With MPL	**63.9**

Table 5. Performance comparison of state-of-the-art algorithms on CUHK01 dataset. Trained on head-shoulder CUHK01 data, our approach obtains better results than others.

pose learning gets better accuracy than the baseline. It proves that the MPL model benefits from the pose information and can tackle pose variations. In conclusion, results demonstrate that the proposed method can take advantage of head-shoulder information effectively and achieve the state-of-the-art performance..

5. Conclusion

This paper proposes a novel deep pairwise model based on multi-pose learning (MPL) aiming at surveillance in crowded public places where the whole body of person is always not available. To facilitate the research on head-shoulder Re-ID, we build three head-shoulder datasets. Experiments on these datasets demonstrate that our proposed method achieves the state-of-the-art performance.

Acknowledgement. This work is partially supported by grants from the National Key R&D Program of China under grant 2017YFB1002401, the National Natural Science Foundation of China under contract No. U1611461, No. 61390515, No. 61425025, No. 61633002 and No. 61471042.

References

[1] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harelc, and T. Hassnera. Face recognition using deep multi-pose representations. In *WACV*, 2016.

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

[3] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. In *CVPR*, 2010.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identication by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[5] e, H. Idrees, and M. Shah. Human re-identification in crowd videos using personal, social and environmental constraints. In *ECCV*, 2016.

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

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

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

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

[10] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016.

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

[12] R. Scott, L. Honglak, A. Dragomir, S. Christian, E. Dumitri, and R. Andrew. Training deep neural networks on noisy labels with bootstrapping. arXiv:1412.6596, 2014.

[13] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identication. In *ECCV*, 2016.

[14] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identication. In *CVPR*, 2016.

[15] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identication using kernel based metric learning methods. In *ECCV*, 2014.

[16] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Li. Salient color names for person re-identification. In *ECCV*, 2014.

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

[18] Z. Zhang, Y. Chen, and V. Saligrama. A novel visual word co-occurrence model for person re-identication. In *ECCV Workshop*, 2014.

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