Center-Level Verification Model for Person Re-identification

Ruochen Zheng, Yang Chen, Changqian Yu, Chuchu Han, Changxin Gao(✉), and Nong Sang

Key Laboratory of Ministry of Education for Image Processing and Intelligent Control, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China
{m201772447,cgao}@hust.edu.cn

Abstract. In past years, convolutional neural network is increasingly used in person re-identification due to its promising performance. Especially, the siamese network has been widely used with the combination of verification loss and identification loss. However, the loss functions are based on the individual samples, which cannot represent the distribution of the identity in the scenario of deep learning. In this paper, we introduce a novel center-level verification (CLEVER) model for the siamese network, which simply represents the distribution as a center and calculates the loss based on the center. To simultaneously consider both intra-class and inter-class variation, we propose an intra-center submodel and an inter-center submodel respectively. The loss of CLEVER model, combined with identification loss and verification loss, is used to train the deep network, which gets state-of-the-art results on CUHK03, CUHK01 and VIPeR datasets.

Keywords: Center-level · Intra-class variation · Inter-class distance

1 Introduction

Person re-identification (re-id), which aims at identifying persons at non-overlapping camera views, is an active task in computer vision for its wide range of applications. Because of the interference caused by different camera views, lighting conditions and body poses, many traditional approaches are proposed to solve these problems from two categories: feature extracting [10,12,14,21] and metric learning [8,12,16,20]. With the development of deep learning and the emergence of large datasets, deep neural network shows impressive performance in re-id [1,6,15,17,23]. The verification loss and triplet loss are widely used in deep learning. The verification loss [1,6,15,17,23] can be divided into two forms according to loss function differences: contrastive loss and cross-entropy loss. Both of them punish the dissimilarity of the same person and the similarity of the different persons. And the triplet loss [3–5,13,16] embeds space to make data points with the same label closer than the data points with different labels.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11256, pp. 98–107, 2018.
https://doi.org/10.1007/978-3-030-03398-9_9

Note that, both verification loss and triplet loss only take sample-level loss as consideration. However, the sample-level loss is not quite appropriate to deep leaning based method. Because mini batch is the common strategy adopted in both verification loss and triplet loss, in the training stage of deep learning. In a batch, only one image or several images are randomly selected in a camera for one identity, which cannot represent the real distribution of the image sets of the identity.

Recurrent neural network (RNN) [15,17] provide a possible solutions for this problem by establishing a link between frames. However, temporal sequences are needed for RNN model in re-id task, so RNN can only work in video sequence. For the image set, center loss [18], which models a class as a center, may provide a simple yet effective way to address this problem. It is effective to punish intra-class variation by center loss. For each class, the center loss is calculated with the samples and the center, the center will be recorded and updated during training stage. Therefore, to some extent, the center can be considered as a representation of the distribution of the corresponding class. [9] has applied center loss on person re-identification. However, it only pays attention to reducing the intra-class variation, ignoring the inter-class distance. We argue that an effective constraint for inter-class distance will further boost the performance.

Motivated by center loss [18], this paper introduces a new architecture named Center-LEvel VERification (CLEVER) model for the siamese network, to overcome the shortcoming of sample-level loss. For each person identity, we take its center as the simple representation of its distribution. Based on the centers, we propose to simultaneously reduce intra-class variation and enlarge inter-class distance, by using *intra-center loss* and *inter-center loss* respectively, as shown in Fig. 1. Similar with the contrastive loss, a margin for the distances of different centers is set to limit the minimum inter-class distance, the distance less than the margin will be punished as inter-center loss. Moreover, by taking center-level as consideration, the combination of CLEVER model, identification loss and verification loss performs better than only combining identification loss and verification loss.

In summary, our contributions are two-fold: (1) We propose a center-level verification (CLEVER) model based on siamese network, which can both reduce intra-class variation and enlarge inter-class distance. (2) We show competitive results on CUHK03 [11], CUHK01 [10] and VIPeR [7], proving the effectiveness of our method.

2 Related Work

In this section, we describe previous works relevant to our method, including methods based on loss function models on person re-identification and methods trying to reduce the intra-class variation and enlarging inter-class distance.

Many works adopt the combination of identification loss and verification loss to train a network. Verification loss can be divided into cross-entropy form and contrastive loss according to differences of loss function. Cross-entropy form

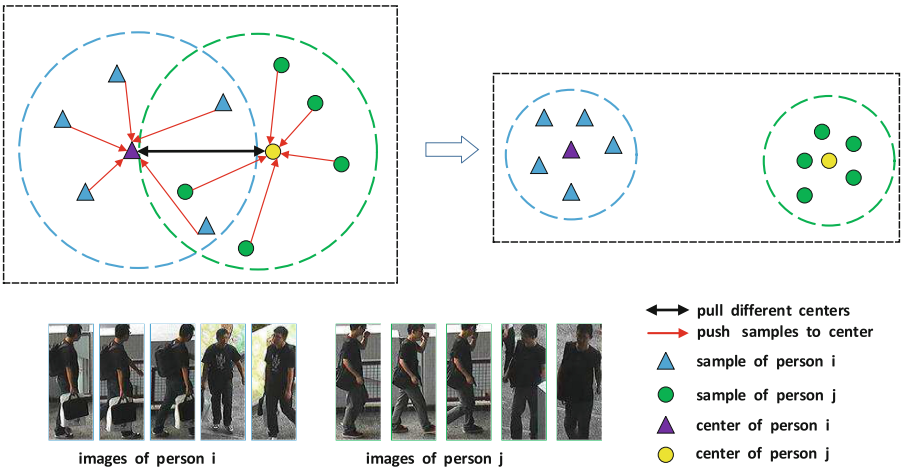

Fig. 1. Illustration of our motivation. Our CLEVER model makes a discriminate separation between two similar persons, by pushing images to their corresponding center and pulling their centers away.

adopts softmax layer to measure the similarity and dissimilarity of image pairs. [6,23] adopts the form of cross-entropy loss, combining with identification loss in their network. Different from cross-entropy loss, contrastive loss [15,17] form owns a margin to get a definite separation between positive pairs and negative pairs. However, both cross-entropy form and contrastive loss pay attention on sample-level, ignoring the real distribution of the whole image set.

Another loss function associated with our model is center loss. [18] adopts combination of center loss and softmax loss on face recognition task. And [9] applies center loss on the person re-id task to reduce intra-class variation. However, the neglect of constraint on inter-class distance limits the performance of these tasks.

The approach closest to our CLEVER model in motivation is the method [24,25]. Both of the methods concentrate on reducing the intra-class variation and enlarging inter-class distance. However, the two methods and area of concern are different from our CLEVER model. [24] pays attention on "image to video retrieval" problem with dictionary learning method, [25] tries to solve video based ReID with metric learning method. Our CLEVER model bases on 'image to image' ReID with deep learning method.

3 Our Approach

In this section, we present the architecture of our CLEVER model, as shown in overview. The CLEVER model has two main components: intra-center submodel and inter-center submodel. Intra-center submodel pushes samples to its corresponding center, while inter-center pulling different centers away. Specially,

we take the form of image pairs as input to the siamese network. The images from two cameras with same identity, termed as positive pairs, are taken as input to intra-center submodel. In contrast, inter-center submodel adopts negative pairs, which represent images of different identities. In this section, we first introduce intra-center submodel and then inter-center submodel. The combination of sample-level will be presented at last.

3.1 Intra-center Model

In intra-center submodel, positive pairs are taken as the input of network. The distances between center and positive pairs will be punished by intra-center loss as follows:

$$L_{intra} = \frac{1}{2m}\sum_{i=1}^{m}(\|x_{i1} - c_{y_i}\|_2^2 + \|x_{i2} - c_{y_i}\|_2^2) \tag{1}$$

where x_{i1} and x_{i2} are the features extracting from images of identity y_i. And c_{y_i} is the center y_i corresponding. Specially, the center is updated as:

$$\frac{\partial L_c}{x_{i1}} = x_{i1} - c_{y_i} \tag{2}$$

$$\frac{\partial L_c}{x_{i2}} = x_{i2} - c_{y_i} \tag{3}$$

$$\triangle c_j = \frac{\sum_{i=1}^{m}\delta(y_i = k)\cdot(2\cdot c_k - x_{i1} - x_{i2})}{1 + \sum_{i=1}^{m}\delta(y_i = k)} \tag{4}$$

$$c_k^{t+1} = c_k^t - \alpha \cdot \triangle c_k^t \tag{5}$$

where $\sum_{i=1}^{m}\delta(y_i = k)$ counts the number of pairs that belong to class k in a batch. The value of α, which ranges from 0 to 1, could be seen as learning rate of centers. The main difference of our inter-center submodel and center loss is that we adopt positive-image pair as input. Our method benefits from taking positive image pairs to update center simultaneously, so that we can learn a center closer to real center of image set. We conduct experiments to prove the effectiveness of this strategy.

However, intra-center model only cares about reducing the intra-class variation, the combination of identification loss still shows a weak ability to distinguish similar but different identities, which often occur in person re-identification task. Therefore, we propose inter-center loss to enlarge the distances of different classes in Sect. 3.2.

3.2 Inter-center Model

In the case of small intra-class variation based on intra-center submodel, we propose an inter-center submodel, which limits the minimum distances between different centers to pull different classes away. The inter-center distances less than margin will be punished by inter-center loss as follows:

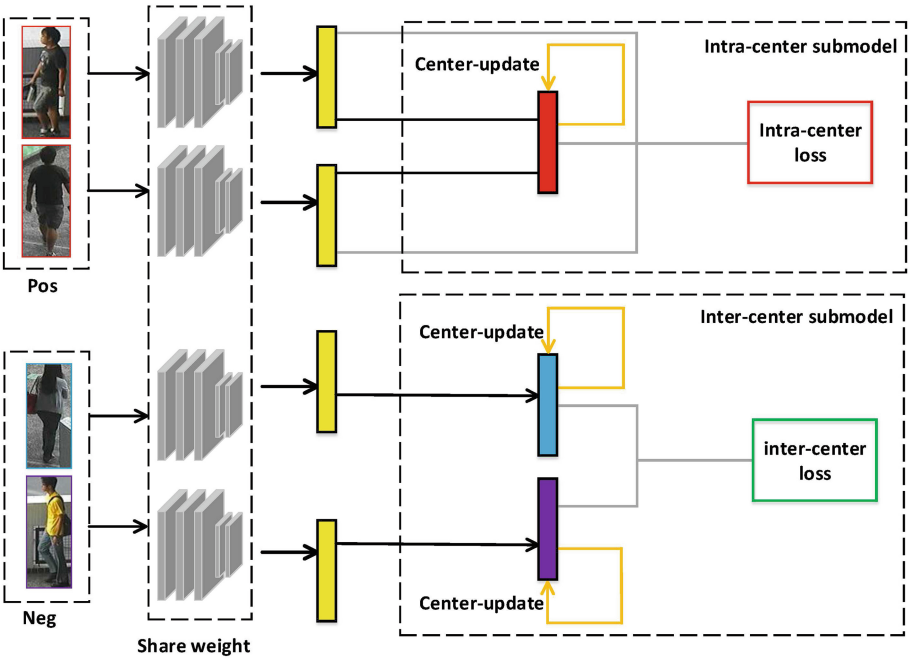

Fig. 2. An overview of the proposed CLEVER architecture. It contains intra-center submodel and inter-center submodel.

$$L_{inter} = \frac{1}{m} \sum_{j=1}^{m} \max(0, d - \|c_{y_{j1}} - c_{y_{j2}}\|_2^2) \tag{6}$$
$$y_{j1} \neq y_{j2}$$

where $\|c_{y_{j1}} - c_{y_{j2}}\|_2^2$ is the squared Euclidean distance between the center of c_{j1} and c_{j2}. And d plays a role as margin of the distances, m is the number of pairs in a batch. Negative pairs will be taken as input for inter-center submodel. They will also participate in the update of their corresponding centers.

3.3 Joint Optimization of Center-Level and Sample-Level

By setting the weight of center loss and inter-center loss. The center-level loss function can be formulated as follows:

$$L_{CLEVER} = \beta \cdot L_{intra} + \gamma \cdot L_{inter} \tag{7}$$

where β and γ control the balance of two terms. Our center-level loss function has the similar form to contrastive loss of image-level, thus it can be seen as the verification loss of center level.

The architecture of our center-level model is showed in Fig. 2. For the intra-center submodel, images of same identity coming from two cameras will be randomly selected as a positive pair for input. The two images of different camera will jointly update the corresponding center, which makes the operation more efficient and accuracy. Negative pairs will also update their corresponding centers in inter-center submodel. The architecture of center-level is also capable with image-level, which makes it possible for combining verification loss based on sample-level with our center-level model. Therefore, the final loss function could be formulated as follows:

$$L = L_I + L_V + L_{CLEVER} \tag{8}$$

where L_I is the identification loss coming from siamese network of two cameras, L_V is the verification loss, which adopts the cross-entropy loss form for it is more concise. The verification loss plays a role as dividing hard samples, which is very helpful for training the network.

Table 1. Results on CUHK03 using the single-shot setting. The results of several different combinations of components are listed. [9] offers code of * "IV", we adopt the code and get a slightly different result. Here we report the result we get.

Method	rank1	rank5	rank10
baseline IC [9]	80.20	96.10	97.90
CLEVER(intra only)+I	81.45	96.25	98.00
baseline IV*	81.90	95.30	97.75
CLEVER(intra only)+IV	83.10	96.35	98.40
CLEVER(inter only)+IV	81.45	95.30	97.80
CLEVER+I	82.00	96.45	98.45
CLEVER+IV	84.85	97.15	98.25

4 Experiment

4.1 Datasets

We conduct our experiments on CUHK03, CUHK01 and VIPeR datasets. CUHK03 contains 13164 images of 1360 identities. It provides two settings, one is annotated by human and the other one is annotated from deformable part models (DPM). We will evaluate our model on the bounding boxes detected by DPM, which is closer to practical scenarios. Following the conventional experimental setting, 1160 persons will be used for training and 100 persons for testing. The results of single shot will be reported. CUHK01 contains 971 identities with two camera views, and each identity owns two images. VIPeR contains 632 identities with two camera views, each identity owns one image. For the CUHK01 and VIPeR datasets, we randomly divide the individuals into two equal parts, with one used for training and the other for testing. Both CUHK01 and VIPeR adopt single-shot setting.

4.2 Implementation Details

We set [9] as our baseline. A CNN that contains only nine convolutional layers and four max pooling layers is proposed in [9], for more detail about structure can be found in [9]. Each image is resized to 128×48 to adjust to convolution network. Note that, smaller inputs make feature maps smaller, and shallower networks have fewer parameters, which makes the depth network easier to apply to real-world scenarios. Before training, the mean of training images will be subtracted from all the images.

For the hyper parameters setting, the batch size is set to 200, 100 images for positive pairs and the other for negative pairs. α is set as 0.5, β and γ are set as 0.01 and 0.008, respectively. The value of d is set as 250. The number of training iterations is 25k, the initial learning rate is 0.001, decayed by 0.1 after 22k iterations. For the value of centers, we uniformly initialize them with zero vector with the same size as features.

For the experiment on CUHK03, we follow the protocol in [11], all experiments are repeated 20 times with different splitting of training and testing sets, the results will be averaged to ensure stable results. For the CUHK01 and VIPeR datasets, we conduct experiment following the set of [6]. The model will be pre-trained on CUHK03 [11] and Market1501 [22] at first. Then we fine-tune it on CUHK01 and VIPeR. The experiment will be repeated with 10 random splits. To evaluate the performance of our methods, the Cumulative Matching Characteristic curve (CMC) will be used. The CMC curves represents the number of true matching in first k ranks.

4.3 Effectiveness of Each Component

We evaluate the effectiveness of the components of the CLEVER model on CUHK03 dataset. The results are shown in Table 1. For the abbreviations for different combinations, the combination of identification loss and verification loss is called "IV". "IC" means the combination of identification loss and center loss. Taking identification loss only is called "I". From Table 1, we can see the combination of our "CLEVER" model and "IV" gets best performance, it achieves 84.85% rank-1 accuracy, obtaining 2.95% improvement on "IV", which proves the effectiveness of our CLEVER model. The strategy of pair image input gets proved on the comparison between "IC" and "CLEVER(intra only)+I". The accuracy of rank-1 obtains 1.25% improvement.

Another interesting result comes from the contrast experiment of verification loss. We replace "CLEVER+IV" by "CLEVER+I", the accuracy drops 2% in rank-1 accuracy, which prove the importance of verification loss. We analyzes that verification loss can serve as verifying the hard samples, which is helpful for training. The validity of the inter-center submodel can be verified from the comparative experiments of "CLEVER+IV" and "CLEVER(intra only)+IV". By setting a minimum distances among different centers, the model obtain 1.75% improvement.

4.4 Comparison with the State of the Arts

Table 2 summarizes the comparison of our method with the state-of-the-art methods. It is obvious that our method performs better than most of approaches above, which proves competitiveness of our method. It should be noted that "CNN Embedding" [23] and "Deep Transfer" [6] uses ImageNet data for pre-training, but we get higher rank-1 accuracy than them on CUHK03 datasets without ImageNet pretraining. In CUHK01 and VIPeR datasets, "Deep Transfer" gets best performance for its advantage of taking ImageNet data, and our method still show competitive results.

Table 2. Comparison with state-of-the-art methods on CUHK03 (detected), CUHK01 and VIPeR datasets using the single-shot setting.

Dataset	CUHK03			CUHK01			VIPeR		
Method	rank1	rank5	rank10	rank1	rank5	rank10	rank1	rank5	rank10
Siamese LSTM [17]	57.30	80.10	88.30	-	-	-	42.40	68.70	79.40
CNN Embedding [23]	83.40	97.10	98.70	-	-	-	-	-	-
GOG [14]	67.30	91.00	96.00	57.80	79.10	86.20	49.70	**79.70**	88.70
MCP-CNN [4]	-	-	-	53.70	84.30	91.00	47.80	74.70	84.80
Ensembles [16]	62.10	89.10	94.30	53.40	76.30	84.40	45.90	77.50	**88.90**
CNN-FRW-IC [9]	82.10	96.20	98.20	70.50	90.00	94.80	50.40	77.60	85.80
IDLA [1]	54.74	86.50	94.00	47.53	71.50	80.00	34.81	63.32	74.79
Deep Transfer [6]	84.10	-	-	**77.00**	-	-	**56.30**	-	-
DGD [19]	80.50	94.90	97.10	71.70	88.60	92.60	35.40	62.30	69.30
Quadruplet+MargOHNM [2]	75.53	95.15	**99.16**	62.55	83.44	89.71	49.05	73.10	81.96
CLEVER+iv	**84.85**	**97.15**	98.25	70.90	**90.86**	**94.92**	52.33	79.41	88.53

4.5 Discussions on CLEVER Model

The sample-based approaches in past years pay attention to optimizing the network by controlling the distance between individuals. However, such a strategy cannot effectively use the information of the global distribution in each comparison, because only two or three images are utilized in the comparison process. Our method records the center information based on sample level, and the center information can be seen as the representation of the global information. The significance of the existence of the center is not only to control the intra-class variation, but also to limit the distance between different classes. Our approach proves the effectiveness of this strategy in person re-identification.

5 Conclusion

In this paper, we have proposed a center-level verification model named CLEVER model for person re-identification, to handle the weakness of the sample-level models. The loss function of the CLEVER model is calculated by the samples and their centers, which to some extent represent the corresponding distributions. Finally, we combine the proposed center-level loss and the sample-level loss, to simultaneously control the intra-class variation and inter-class distance. The control of center improves the generation ability of network, which has outperformed most of the state-of-the-art methods on CUHK03, CUHK01 and VIPeR.

Acknowledgements. This work was supported by National Key R&D Program of China (No. 2018YFB1004600), the Project of the National Natural Science Foundation of China (No. 61876210), and Natural Science Foundation of Hubei Province (No. 2018CFB426).

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
2. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the CVPR, vol. 2 (2017)
3. Chen, W., Chen, X., Zhang, J., Huang, K.: A multi-task deep network for person re-identification. In: AAAI, vol. 1, p. 3 (2017)
4. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)
5. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognit. **48**(10), 2993–3003 (2015)
6. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244 (2016)
7. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, pp. 1–7. Citeseer (2007)
8. Hirzer, M.: Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2288–2295. IEEE Computer Society (2012)
9. Jin, H., Wang, X., Liao, S., Li, S.Z.: Deep person re-identification with improved embedding. arXiv preprint arXiv:1705.03332 (2017)
10. Li, W., Wang, X.: Locally aligned feature transforms across views. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3594–3601. IEEE (2013)

11. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
12. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
13. Liu, J., et al.: Multi-scale triplet CNN for person re-identification. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 192–196. ACM (2016)
14. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical Gaussian descriptor for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1363–1372 (2016)
15. McLaughlin, N., del Rincon, J.M., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1325–1334. IEEE (2016)
16. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1855 (2015)
17. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 135–153. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_9
18. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
19. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1249–1258. IEEE (2016)
20. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248 (2016)
21. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 144–151 (2014)
22. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
23. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person reidentification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **14**(1), 13 (2017)
24. Zhu, X., Jing, X.-Y., Wu, F., Wang, Y., Zuo, W., Zheng, W.-S.: Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image. In: AAAI, pp. 4341–4348 (2017)
25. Zhu, X., Jing, X.-Y., You, X., Zhang, X., Zhang, T.: Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. IEEE Trans. Image Process. **27**(11), 5683–5695 (2018)