Video-based Person Re-identification by Deep Feature Guided Pooling

Youjiao Li[1,2], Li Zhuo[1,3], Jiafeng Li[1], Jing Zhang[1], Xi Liang[1], Qi Tian[4]
[1]Signal & Information Processing Lab, Beijing University of Technology, Beijing, China
[2]College of Computer Science and Technology, Shandong University of Technology, Zibo, China
[3]Collaborative Innovation Center of Electric Vehicles in Beijing, Beijing, China
[4]Department of Computer Science, University of Texas at San Antonio, TX 78249, USA
liyoujiao@emails.bjut.edu.cn; zhuoli@bjut.edu.cn; lijiafenga@163.com; zhj@bjut.edu.cn; liangxi627@163.com; qitian@cs.utsa.edu

Abstract

Person re-identification (re-id) aims to match a specific person across non-overlapping views of different cameras, which is currently one of the hot topics in computer vision. Compared with image-based person re-id, video-based techniques could achieve better performance by fully utilizing the space-time information. This paper presents a novel video-based person re-id method named Deep Feature Guided Pooling (DFGP), which can take full advantage of the space-time information. The contributions of the method are in the following aspects: (1) PCA-based convolutional network (PCN), a lightweight deep learning network, is trained to generate deep features of video frames. Deep features are aggregated by average pooling to obtain person deep feature vectors. The vectors are utilized to guide the generation of human appearance features, which makes the appearance features robust to the severe noise in videos. (2) Hand-crafted local features of videos are aggregated by max pooling to reinforce the motion variations of different persons. In this way, the human descriptors are more discriminative. (3) The final human descriptors are composed of deep features and hand-crafted local features to take their own advantages and the performance of identification is promoted. Experimental results show that our approach outperforms six other state-of-the-art video-based methods on the challenging PRID 2011 and iLIDS-VID video-based person re-id datasets.

1. Introduction

Person re-identification (re-id) aims to match pedestrians across different non-overlapping camera views. It has been playing an increasingly key role in various intelligent surveillance and security systems. In recent years, a large amount of attention has been paid on solving the problem of person re-id in both academic research and industry. However, person re-id is an extremely challenging task due to various complicated factors such as the variations in pose, lighting, and camera viewpoint. To tackle the above problems, many efforts have been made in person re-id, and most of them rely on still images [1-5]. However, information in a still image is limited, which makes the accuracy of person re-id is far from satisfactory. Recently, many researchers have begun to use video for person re-id [6-8]. Compared with image-based person re-id, video-based techniques can achieve better performance by fully utilizing the rich space-time information to extract more robust appearance features.

In spite of the above advantages, there are still some particular difficulties in the process of video-based person re-id. Firstly, rich video information comes along with a variety of sophisticated noise. How to extract the features robust to noise is critical to the improvement of identification performance. Secondly, motion information is a kind of important discriminative cues for identifying different persons in videos. How to effectively exploit the motion information is still an open question. Lastly, different persons may be similar in both motion and appearance (as showed in Figure 1). How to distinguish these persons is the key to improve the person re-id performance. Based on the aforementioned observations and motivations, an efficient video-based person re-id method named Deep Feature Guided Pooling (DFGP) is proposed in this paper to make full use of the vivid information in the significant amount of video frames. Its contributions are in the following aspects:

(1) PCA-based convolutional network (PCN) [9], a light-weight and effective deep learning network, is introduced to extract deep features of video frames. Deep features are then aggregated by average pooling to obtain deep feature vectors of person. The vectors are utilized to guide the generation of human appearance features, which makes the appearance features more robust to severe noise in videos.

(2) Hand-crafted local features of videos are aggregated by max pooling to reinforce the motion variations of differ-

Figure 1: The appearance and movement information of two different pedestrian are similar.

ent persons. In this way, the human descriptors are more discriminative.

(3) The final human descriptors are composed of deep features and hand-crafted local features to take their own advantages for promoting the identification performance.

The structure of this paper is as follows. Section 2 introduces the related work. Section 3 explains the proposed method in detail. Section 4 presents an extensive comparison with state-of-the-art methods on two publically available datasets, and the analysis of our method is also presented in this section. In Section 5, the final conclusions are given.

2. Related work

Since the person re-id problem was proposed by Everingham *et al*. [10] in 2004, person re-id for still images has been extensively studied. The existing methods can be normally divided into two categories, feature representation and distance metric learning. An effective feature representation should be robust and discriminative to complicated environment. Distance metric learning should be learned to map the raw features into a new space, which is more discriminative to large variations of person images across different views. Since 2012, as deep learning study gradually goes deeper, it is constantly applied to person re-id, which has achieved superior performance over traditional methods.

2.1. Deep learning in image-based person re-id

In general, there are three ways in this category.

(1) The feature representation and distance metric learning are incorporated into an integrated framework and jointly optimized through deep learning model, which can achieve excellent performance. Li *et al*. [11] proposed a Filter Pairing Neural Network (FPNN) network. The FPNN network begins with a convolutional layer, max pooling next, and then a patch-matching layer that matches the filter responses across two views. Ahmed *et al*. [12] proposed an Improved Deep Learning Architecture (IDLA), with the use of cross-input neighborhood differences and patch summary features to learn the cross-view relationships of the features. The siamese model [13] using image pairs are employed in these methods. An important reason for using the siamese network is that the person re-id datasets are usually small and cannot take full advantages of deep learning.

(2) The feature representation is composed of deep feature and hand-crafted feature, and distance metric learning is used for similarity measure [14-16]. Zheng *et al*. [14] proposed a query-adaptive fusion method to evaluate the performance of different features. CNN [17] features combined with five kinds of hand-crafted features are used for person re-id. Wu *et al*. [16] proposed a feature fusion net to fuses CNN feature and ELF feature [18]. The extraction of CNN features is constrained by back propagation. The fusion feature is combined with the traditional metric learning method to complete the re-id process.

(3) Deep features are directly connected with metric learning. With the introduction of large scale person re-id datasets, satisfactory performance could be achieved by using the deep features alone. In [19], an ID-discriminative embedding is learned through CNN features firstly, and then a confidence weighted similarity metric is used for similarity measurement.

2.2. Deep learning in video-based person re-id

In video-based person re-id, the study of deep learning has only recently launched [20-22], in which the rich space-time information can be used. Methods in this category typically learn video frame features in an end-to-end manner, and involve a max/average pooling step to aggregate multiple frame-level features. In [20], video sequences are taken as training samples to train a CNN model with softmax loss. Some other works focus on utilizing temporal information in videos. In [21], deep features are extracted from video sequence through two symmetry CNN model, and then fed into a recurrent layer to capture the temporal information. In [22], hand-crafted low-level features are fed into several Long Short-Term Memory (LSTM) modules and the LSTM can remember and propagate previously good features and forget newly input inferior ones.

3. Deep feature guided pooling method

By utilizing space-time information, a video-based person re-id method named Deep Feature Guided Pooling method (DFGP) is proposed in this work. The framework of DFGP is showed in Figure 2. The proposed method contains three steps. Firstly, a PCA-based Convolutional Network (PCN) is used to train a lightweight deep network model which takes video sequences with different views as input.

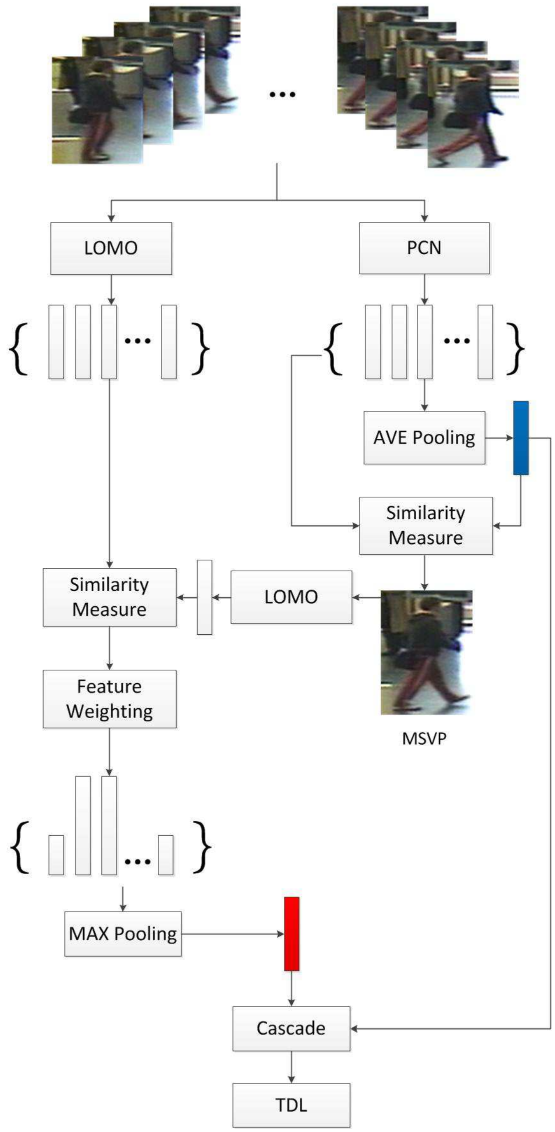

Figure 2: Flow chart of the proposed DFGP. The video feature extraction is a four-step process. (1) For all image sequences of a person video, two kinds of feature are extracted: PCN and LOMO. (2) The set of PCN features are aggregated by average pooling to generate a feature vector marked in blue. This feature vector is further used to find the MSVF. (3) The set of LOMO features are weighted according to their similarity to the LOMO feature of MSVF. The weighted LOMO features are then aggregated by max pooling to generate a feature vector marked in red. (4) The two feature vectors are concatenated together to stand for the final representation of the video.

For all image sequences of a person video, PCN deep features are extracted through the trained model. Next, the set of features of a video are aggregated by average pooling to generate a deep feature vector. In a video, the frame whose deep feature is mostly similar to the pooling result is found by k-nearest neighbor algorithm. The frame is denoted as Maximally Stable Video Frame (MSVF). Secondly, hand-crafted features, Local Maximal Occurrence (LOMO) [3], are extracted from the same video sequence and weighted according to their similarity to MSVF. The set of features are aggregated by max pooling to generate a hand-crafted feature vector. The hand-crafted feature vector is cascaded with the deep feature vector to form the final person descriptor. Lastly, Top-push Distance Learning model (TDL) [8] is employed to learn the distance metric. By now, the process of re-id is accomplished.

Next, the implementation details of the proposed method are introduced.

3.1. Deep feature extraction with PCN

Deep learning can abstract semantics features to provide more invariance to intra-class variability. PCANet [23] has been successfully applied into many applications [24, 25], such as image classification, face recognition and action detection, *etc*. Especially in the face recognition, PCANet copes well with distortion, deformation and occlusion. However, PCANet feature dimension will exponentially grow with increasing number of samples. PCN is an improvement of PCANet. PCN introduces pooling layer to feature extraction stage to build robustness to small distortions and further reduce the resolution of feature maps.

The basic architecture of PCN can be seen in Figure 3 [9], which is composed of three stages: the first two cascaded feature extraction stages and the last stage of nonlinear output. Each feature extraction stage consists of a convolutional layer and a max/average pooling layer.

3.1.1. The first feature extraction stage in PCN. Assume that there are N training images of size $m \times n$. In each image, a patch of size $k_1 \times k_2$ at every k pixel locations is taken. All the patches are collected, vectorized and then combined into a matrix of $k_1 \times k_2$ rows and $\left(\left\lceil \frac{m-k_1}{k} \right\rceil + 1\right) \times \left(\left\lceil \frac{m-k_2}{k} \right\rceil + 1\right)$ columns. For the matrix X_i corresponding to the *ith* input image I_i, patch mean is subtracted from each patch to generate a matrix $\overline{X}_i$. Once matrices for all input images are structured in this way, they are assembled to form a matrix $\mathbf{X} = [\overline{X}_1, \overline{X}_2, ..., \overline{X}_N]$.

Then, the eigenvectors of XX^T are computed. And the ones corresponding to the L_1 largest eigenvalues are selected as the convolutional filters. The learned filters with the input images are convoluted to get L_1 feature maps. The feature maps are divided into several non-overlapping pooling regions to apply average pooling. Thus, feature maps with reduced resolution are generated.

3.1.2. The second feature extraction stage in PCN. At the second stage, similar process with stage 1 is performed. The

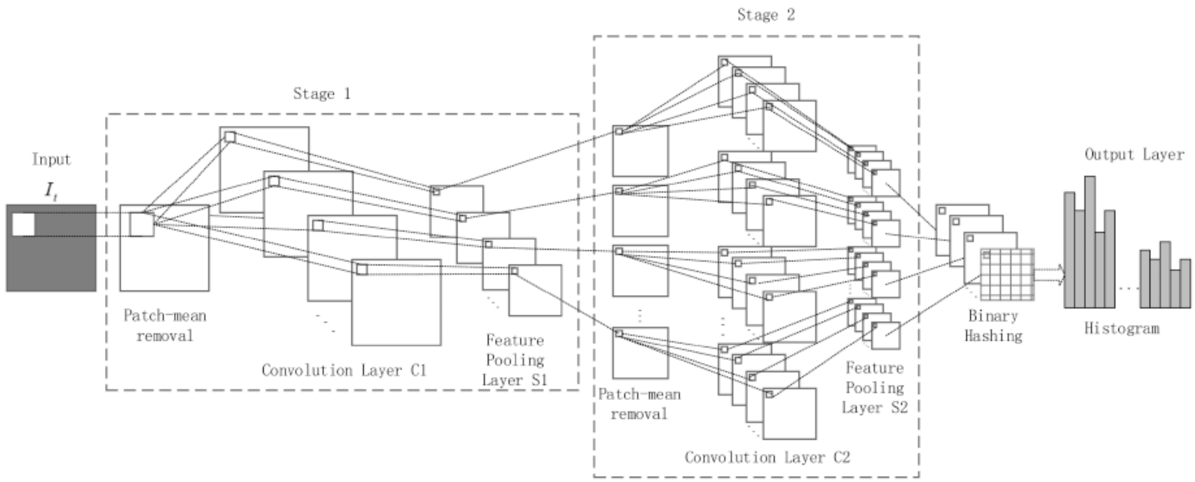

Figure 3: Basic structure of a three-stage PCN.

pooled feature maps in stage 1 are treated as the original input to the stage 2. And a matrix $\mathbf{Y}=[\overline{Y}_1,\overline{Y}_2,...,\overline{Y}_N]$ is obtained. The number of convolutional filters in this stage is L_2. Finally $L_1 \times N \times L_2$ feature maps are produced in total. Also, average pooling is applied to the feature maps.

3.1.3. The output stage in PCN. In the output stage, the output of stage 2 is binarized and the block-wise histograms are computed to form final representations of the input image.

Parameters involved in the PCN model include the image patch size k_1, k_2, the convolutional filter number L_1, L_2, the number of stages and the block size for histograms. Specifically, the image size is set to 64×128, the patch size 10×10, the stage number 2, the filter number $L_1=20$, $L_2=10$, and the block size 20×20.

Video sequences are passed through the trained PCN to produce deep features. Let $v=v_1, v_2... v_T$ be a video sequence of length T, where v_i is the *ith* image. For all frames in *v*, deep features are aggregated by average pooling to generate a 4096-dimension feature vector, denoted as:

$$F_{PCN}=\frac{1}{T}\sum_{i=1}^{T}f(i) \qquad i=1,2,...,T \tag{1}$$

where $f(i)$ is the PCN feature of v_i.

The PCN feature can represent the global information of people, which makes it robust to occlusion and deformation. The subsequent averaging pooling can enhance this effect.

Based on F_{PCN}, the most stable frame in *v* denoted as Maximally Stable Video Frame (MSVF) could been found. The overall MSVF search method is given in Algorithm 1.

Next，the MSVF is employed to guide the extraction of hand-crafted features.

Algorithm 1 MSVF search method

1: **Input:** all the PCN features $f(i)$ of image sequences in *v* with length T
2: **for** i=1 to T **do**
Calculate the difference $D(i)$ between $f(i)$ and F_{PCN}.
Determine the location j^* whose $D(j^*)$ is minimal.
3: **end for**
4: **Output** j^*

3.2. Hand-crafted feature extraction with LOMO

Hand-crafted local features have been verified to be effective to person re-id. The Local Maximal Occurrence (LOMO) feature is a local feature specifically designed for person re-id. LOMO is robust to the variations in lighting and viewpoint. Therefore, the final human descriptors combine LOMO features with PCN features to take their own advantages in this work.

LOMO feature consists of two key parts, HSV histogram and Scale Invariant Local Ternary Pattern (SILTP) [26] histogram. LOMO uses sliding windows to decompose person image. Within each sub-window, a SILTP texture histogram and a HSV color histogram are extracted. Each histogram bin can be seen as the occurrence probability of one pattern in a sub-window. Finally, the local occurrence probability can be maximized at the same horizontal location to obtain final LOMO feature vector.

In our experiments, the sub-window size is set to 10×10 and the overlapping step 0.5. And a 26960-dimensional feature vector is generated. The LOMO feature vector of the *ith* image is denoted as $g(i)$. All the LOMO features of video sequence in *v* are combined to form a feature sets

$g = \{g_1, g_2, ..., g_T\}$. Also, LOMO feature of MSVF is abstracted and denoted as g_{MSVF}.

The MSVF is considered as the most stable frame in a person video. Frames similar to MSVF are stable enough as well. On the contrary, the rest frames are much likely to suffer from background clutter and occlusion. Therefore, features in g similar to g_{MSVF} gain more weight to improve the robustness of person descriptor. Average threshold method is used to measure the similarity between g_{MSVF} and $g(i)$ of g, which is simplified as follows:

$$Threshold = \frac{1}{T}\left(\sum_{i=1}^{T} | g_{MSVF} - g(i) |\right) \tag{2}$$

If $g(i)$ is greater than the threshold, the weight of $g(i)$ is set to 0.5, otherwise 1.5:

$$\beta_i = \begin{cases} 0.5 & g(i) \geq Threshold \\ 1.5 & g(i) < Threshold \end{cases} \tag{3}$$

Then a novel feature sets is generated, denoted as $\breve{g} = \{\beta_1 g_1, \beta_2 g_2, ..., \beta_T g_T\}$. The features in $\breve{g}$ are aggregated by max pooling to obtain a 26960-dimensional feature vector, which represents local feature of a person video:

$$F_{LOMO} = \max(\breve{g}) \tag{4}$$

Motion variations of different person are reserved by max pooling. In this way, the person descriptors are more discriminative. F_{LOMO} is cascaded with F_{PCN} to obtain the final person descriptor, *i.e.* $F_{DFGP} = [F_{PCN}, F_{LOMO}]$, totally with 31056 dimension.

3.3. Metric learning with TDL

Top-push Distance Learning model (TDL) is a typical metric learning method for video-based person re-id. TDL uses a more stringent constraint on top-rank matching, so as to make the model more effective. Specifically, TDL is not the distance between positive sample pairs and all relative negative sample pairs, but the minimum distance, as indicated as follows:

$$D(\vec{x}_i, \vec{x}_j) + \rho < \min_{y_k \neq y_i} D(\vec{x}_i, \vec{x}_k), \qquad y_i = y_j, \tag{5}$$

where D is the distance between features, $\vec{x}$ is the feature vector with label y, and ρ is a slack parameter. In this work, ρ is set to 1.

4. Experiments

To evaluate the proposed method, extensive experiments were conducted on two publicly available datasets for video-based person re-id: the PRID 2011 dataset [27] and the iLIDS-VID dataset [7]. The two datasets both contain two cameras. The iLIDS-VID dataset was published after PRID 2011. Compared with the PRID 2011 dataset, iLIDS-VID is more orderly and challenging.

The PRID 2011 dataset consists of video pairs recorded from two static surveillance cameras, as shown in Figure 4(a). 385 persons were recorded in camera A, and 749 persons in camera B. Among all persons, only 200 persons were recorded in both cameras. Each video is comprised of 5 to 675 image frames, with an average of 100. To guarantee the effective length of the video, the widely adopted experimental protocol is to select 178 persons with more than 27 frames. This strategy is adopted in our experiments. The PRID 2011 dataset was captured in spacious outdoor scenes with clean background and rare occlusions, and the variations of pedestrian motion are small.

The iLIDS-VID dataset was captured in an airport arrival hall under a CCTV network, as shown in Figure 4(b). It contains 600 videos of 300 people. Each video is comprised of 23 to 192 frames, with an average of 73. The challenges of this dataset lie in clothing similarities, lighting and viewpoint variations, complicated background and occlusions.

In our experiments, both datasets were randomly divided into training set and testing set by half. The procedure was repeated 10 times and the average results are reported in this work.

Table 1 reported the comparison of the proposed DFGP method with the six state-of-the-art video-based person

(a) PRID 2011

(b) iLIDS-VID

Figure 4: Video pair of the same person under different camera views.

Method	PRID 2011				iLIDS-VID			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
SDALF [30]	5.2	20.7	32	47.9	6.3	18.8	27.1	37.3
Salience [31]	25.8	43.6	52.6	62	10.2	24.8	35.5	52.9
RPRF [32]	19.3	38.4	51.6	68.1	14.5	29.8	40.7	58.1
SRID [28]	35.1	59.4	69.8	79.7	24.9	44.5	55.6	66.2
DVDL [29]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
Color&LBP+DVR [7]	37.6	63.9	75.3	88.3	34.5	56.7	67.5	77.5
DFCP	**51.6**	**83.1**	**91**	**95.5**	**34.5**	**63.3**	**74.5**	**84.4**

Table 1. Comparison with six state-of-the-art methods on PRID 2011 and iLIDS-VID datasets.

re-id methods [8] on PRID 2011 and iLIDS-VID datasets,including SRID [28], DVDL [29], SDALF [30], Salience[31], RPRF [32], and Color&LBP+DVR [7]. SDALF and Salience focus on building general and effective person feature representation in complex scenes. In the above two methods, the local information is an important component of the pedestrian descriptor. By employing HOG3D [33] space-time feature, Color&LBP+-DVR can fully utilize the rich information in videos to achieve better performance, which makes it the baseline method in video-based person re-id. RPRF, SRID and DVDL focus on designing efficient metric learning method. And random forest, block sparsity and dictionary learning are respectively introduced to the learning step.

The experimental results in table 1 show clearly that the matching performance of the proposed DFGP method on both datasets is better than other methods in general. Compared with the best-performing Color&LBP+DVR method in the six methods, DFGP improved the average matching rate by 9.7%. Especially on Rank-5, the average matching rate increased by 13.6%. The above results indicate that PCN deep feature employed in DFGP is more robust to video noise and outperforms space-time features like HOG3D. This makes the DFGP method effectively eliminate the interference of complicated environment and quickly locate the right matching.

Moreover, DFGP outperformed other methods much better on PRID 2011than on iLIDS-VID. A possible reason lies in the minor motion difference of people in PRID 2011. And DFGP could enhance the motion information in videos to amplify the inter-class variation. Another reason lies in the serious identity ambiguity issue in iLIDS-VID, which means the person bounding box contains a non-unique identity. Figure 5 illustrates an identity ambiguity example which was wrong matched in our method. In this situation, a pedestrian may interact with others frequently. But the motion relationship of different person is absent in DFGP.

Figure 5: An example of wrong matching on iLIDS-VID.

Therefore, the performance of re-id can be further improved if tracking interacting objects methods are involved [34-36].

To sum up, the PCN feature can abstract the global high-level feature in video, and LOMO feature can generate the local low-level feature. These two kinds of complementary features make the re-id system perform remarkable.

5. Conclusions

In this work, a novel and effective way for video-based person re-id called deep feature guided pooling method is proposed. This model jointly utilizes both PCN deep feature and LOMO hand-crafted features. The extraction of hand-crafted features is guided by deep features. Motion information of different persons is enhanced by feature weighting and pooling. The pedestrian descriptors combined by deep features and hand-crafted features are robust and discriminative enough to work well in complex scenes. Experimental results on two public datasets demonstrate the effectiveness of the proposed method.

Acknowledgment

The work in this paper is supported by the National Natural Science Foundation of China (No. 61531006, No. 61372-

149, No. 61370189, and No. 61471013), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No.CIT&TCD20150311, CIT&TCD201404043), the Beijing Natural Science Foundation (No.4142009, No.4163071), the Science and Technology Development Program of Beijing Education Committee (No.KM201410005002, No. KM2015100050-04), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality.

References

[1] H. Wang, S. Gong, X. Zhu, and T. Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016.

[2] S. Wang, L Duan, Y. Zhao, and J. Dong. A Simple Deep Feature Representation for Person Re-identification. In *BIC-TA*, 2016.

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

[4] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv:1701.07717*, 2017.

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

[6] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *arXiv preprint arXiv:1701.08398*, 2017.

[7] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[8] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016.

[9] Y. Gan, J. Liu, J. Dong, and G. Zhong. A PCA-based convolutional network. *arXiv preprint arXiv:1505.03703*, 2015.

[10] M. Everingham and A. Zisserman. Automated Person Identification in Video. In *CIVR*, 2004.

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

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

[13] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. S¨ackinger, and R. Shah. Signature verification using a “siamese” time delay neural network. In *NIPS*, 1993.

[14] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query adaptive late fusion for image search and person re-identification. In *CVPR*, 2015.

[15] Y. Li, L. Zhuo, X. Hu, and J. Zhang. A combined feature representation of deep feature and hand-crafted features for person re-identification. In *PIC*, 2016.

[16] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016.

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

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

[19] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *CVPR*, 2017.

[20] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[21] N. McLaughlin, J. M. D. Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[22] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016.

[23] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma. PCANet: A simple deep learning baseline for image classification?. *IEEE TIC*, 24(12):5017-5032, 2015.

[24] S. Wang, L. Chen, Z. Zhou, X. Sun, and J. Dong. Human fall detection in surveillance video based on PCANet. *MTA*, 75(19):11603-11613, 2016.

[25] F. Gao, J. Dong, B. Li, and Q. Xu. Automatic change detection in synthetic aperture radar images based on PCANet. *IEEE GRSL*, 13(12):1792-1796, 2016.

[26] S. Liao, G. Zhao, V. Kellokumpu, M. Pietik¨ainen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *CVPR*, 2010.

[27] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*. 2011.

[28] S. Karanam, Y. Li, and R. Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPR Workshop*, 2015.

[29] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

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

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

[32] Y. Li, Z. Wu, and R. J. Radke. Multi-shot re-identification with random-projection-based random forests. In *WACV*, 2015.

[33] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[34] X. Wang, E. Turetken, F. Fleuret, and P. Fua. Tracking interacting objects using intertwined flows. *IEEE TPAMI*, 38(11):2312-2326, 2016.

[35] A. Maksai, X. Wang, and P. Fua. What players do with the ball: a physically constrained interaction modeling. In *CVPR*, 2016.

[36] X. Wang, E. Turetken, F. Fleuret, and P. Fua. Tracking interacting objects optimally using integer programming. In *ECCV*, 2014.