Unsupervised Person Re-identification via Discriminative Exemplar-level and Patch-level Feature Fusion

Zhiping Lai\textsuperscript{1,a}, Meibin Qi\textsuperscript{1,b}, Cuiqun Chen\textsuperscript{1,c} and Jianguo Jiang\textsuperscript{1,d}

\textsuperscript{1}DSP Laboratory, School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China

\textsuperscript{a}zplai_hfut@163.com; \textsuperscript{b}qimeibin@163.com; \textsuperscript{c}chencuiqun_hfut@163.com; \textsuperscript{d}jgjiang@hfut.edu.cn

Abstract. The majority of existing person re-identification (re-ID) approaches adopt supervised learning pattern, which require large amount of labeled data to train models. However, due to the high cost of marking by hand, they are limited to be widely used in reality. On the other hand, due to the difference of the camera angle, there are many variations in pedestrian postures and illumination. It is known that extracting discriminative features is pretty effective to solve the problem of person re-ID. Therefore, we propose to fuse exemplar-level features and patch-level features to obtain more distinguishing pedestrian image features for unsupervised person re-ID. Firstly, We carefully design exemplar-level and patch-level feature learning framework (EPFL). The skeleton frame adopts bicomponent branch, one branch is used to learn the global features of pedestrian images, the other is used to learn local features. Then, the global features at the example level and local features at the patch level are fused, thus the discriminative pedestrian image features can be obtained. Furthermore, feature memory bank (FMB) is introduced to facilitate the calculation of the similarity between pedestrian images on unlabeled dataset. We carry on our proposed method on two frequently-used datasets, namely, Market-1501 and DukeMTMC-reID dataset. Experimental results clearly demonstrate the advantage of the proposed approach for unsupervised person re-ID.

1. Introduction

Given a query, person re-identification (re-ID) aims at matching the same person across non-overlapping camera views, which can be extensively treated as a sub-problem of image retrieval. At present, most of person re-ID approaches are on the basis of learning supervised models, which demand substantial labels of data. However, it is not only expensive but also difficult to manually label training data.

Designing unsupervised models has become a common strategy for this problem. We can make use of Traditional hand-crafted features [1,2,3] for unsupervised person re-ID. Recently, some unsupervised re-ID approaches take the cross-domain learning for reference. They utilize the labeled data of the source domain and unlabeled data of target domain simultaneously [4,5,6,7,8,9,10,11,12]. Zhong et al. [11] employ exemplar memory technique for the target domain, which can learn invariant network to transfer knowledge. Wei et al. [9] put forward PTGAN to reduce the domain gaps for person re-ID. However, the performance of these models cannot make us feel very satisfied.
The primary reason is diggin the discriminative identity knowledge becomes very challenging in case of lacking pairwise labels of pedestrians.

Therefore, we devise a novel frame called EPFL, which is utilized for unsupervised person re-ID. The framework consists of two branches, which are used to mutually learn exemplar-level patch-level feature. These two features represent global features and local features, respectively. For the branch named ELF, we extract the global features from the whole images to represent identities of different pedestrians. In order to calculate the similarity between pedestrian images on unlabeled dataset, we also adopt a mechanism of memory [11] to store features of the training set in each batch. For another branch PLF, we embed a patch feature module (PFM) to obtain different patch features, namely 3 patch features, 4 patch features and 6 patch features. Finally, the global example-level features and local patch-level features are concatenated, thus more distinguishing pedestrian image features can be acquired.

2. The Proposed Method

2.1. Framework

As shown in figure 1, we carefully design exemplar-level and patch-level feature learning framework (EPFL). This framework includes two branches, one for global exemplar-level features (ELF) and the other for local patch-level features (PLF).

![Figure 1. The framework of EPFL. The ResNet-50[13] is used as CNN backbone. The green channel represents ELF to extract global example-level features through GAP. Feature memory bank is exploited to memory features. The blue channel represents PLF to obtain different local patch features by PFM.](image)

2.2. Exemplar-level features

For the branch of ELF, we extract global features with the help of applying global average pooling (GAP). The feature memory bank (FMB) \( V^* = \{v_j^*\}_{j=1}^{N} \) is used to store the features of all training images where. Then the probability of \( x_i \) being recognized as i-th class is

\[
p(i|x_i) = \frac{\exp\left(\beta f(x_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\beta f(x_j)/\tau\right)}
\]
where \( \tau \) as the temperature parameter is used to balance the scale of distribution and \( N \) denotes the number of the images during the training process. Following [11], we set \( \tau = 0.05 \) in this paper.

On the other hand, although we treat each exemplar image as a class, there are actually several images belong to the same identity. It cannot be avoided of the changes of postures and viewpoints in reality. In order to make full use of these positive examples during the training process, it is necessary to learn a robust re-ID model. Considering this trait, each image and its corresponding nearest neighbors are encouraged to be close as far as possible. Suppose that the training sample \( x_i \) is in its \( k \) neighbors, and we assign the probability weight of \( x_i \) belonging to \( j \) class as

\[
\omega_{ij} = \begin{cases} 
  \frac{1}{k} & j \neq i, \forall j \in k \\
  1 & j = i
\end{cases}
\]

(2)

Therefore, the loss function of ELF in our model can be written as

\[
L_{ELF} = \frac{1}{n} \sum \sum \omega_{ij} \log p(j | x_i)
\]

(3)

where \( n \) represents the number of images trained in a batch. When \( i = j \), we can make the samples more similar to itself and far away from other samples by optimizing equation (3). When \( i \neq j \), the network is optimized by leading \( x_i \) to be close to its neighbors to overcome the changes in pedestrian postures, viewpoints under different cameras. Following [11], we also set \( k = 6 \).

2.3. Feature memory bank

For the sake of reducing the computational complexity of extracting features from data during each training batch, we use FMB to save the up-to-date output of 2048 dimensional global features for each image. During the back-propagation, for each \( V^i_j \), we update it during training on the unlabeled dataset through equation (4)

\[
V_{js} = \begin{cases} 
  (1-\lambda) \times V_{j,s-1} + \lambda \times V_{js} & t > 0 \\
  V_{js} & t = 0
\end{cases}
\]

(4)

where \( t \) is the training epoch and the hyper-parameter \( \lambda \in [0,1] \) controls the updating rate of \( V_{js} \), \( V_{js} \) is the up-to-date feature. We first initialize the features saved in FMB, and then keep updating batch-by-batch by using equation (4) during training on the unlabeled dataset which is different from [11].

2.4. Patch-level feature

For the branch of PLF, we embed a patch feature module (PFM). PFM is composed of PGN[14] and CNN. PGN can automatically generate \( M \) patches from the feature map. It is known that different patches of the same image are located in different spatial areas and their semantic information is also different. Then the independent CNN branches are used to extract the features of different patches. Because the PGN is learnable and thus it can adaptively adjust the locations of the patches to find more effective patches. Therefore, in order to retain the pedestrian information as much as possible and remove the background information, we divide person images into 3 patches as shown in figure 2. It is easily observed that these 3 patches roughly correspond to the head, upper body and lower body of pedestrians. Meanwhile, since the upper and lower body of pedestrians contain a lot of information, we also generate 4 patches and 6 patches from feature maps, thus more discriminative local information can be acquired. Because there is overlap between different patches, so the information will be complementary. For the sake of pulling similar patches together and pushing the dissimilar patches away in feature space, we enforce PEDAL [14] for constraint. The loss function can be formulated as follows
where $\sigma$ is the scaling number, $g_m(l_i)$ represents the feature of $m$-th patch, $K^m$ is $k$ nearest patches of $g_m(l_i)$ which is set to 10 following [14]. Particularly, we use a similar FMB to store patch features in a batch. $v^m_j$ represents patch features saved in FMB.

\[
L_{\text{pedal}}^m = -\log \frac{\sum_{i\neq K^m} \exp \left(- \frac{\sigma}{2} P g_m(l_i) - v^m_i P \right)}{\sum_{j=1, j\neq i}^N \exp \left(- \frac{\sigma}{2} P g_m(l_i) - v^m_j P \right)}
\]  

(5)

where $\sigma$ is the scaling number, $g_m(l_i)$ represents the feature of $m$-th patch, $K^m$ is $k$ nearest patches of $g_m(l_i)$ which is set to 10 following [14]. Particularly, we use a similar FMB to store patch features in a batch. $v^m_j$ represents patch features saved in FMB.

By this way, the model can learn how to map those visually similar patches closer so as to mine more visual consistent clues for these similar patches. However, it is not uncommon for different pedestrians to dress similar clothes in the actual monitoring system. So pulling the similar patches together may make it ineffective to distinguish the similar patches of different identities. Therefore, the features of 3 patches, 4 patches and 6 patches are concatenated respectively to provide image-level guidance. Since the identity information is not simply determined by one patch, more importantly in the combination of different patches. Simultaneously, we leverage IPFL[14] to further exploit latent discriminative information. The definition of the triplet-based loss function IPFL as follows,

\[
L_{\text{ipfl}}^m = \max \left\{ P g(l_i) - p_i, P - P g(l_i) - n_i, P + \text{mar}, 0 \right\}
\]  

(6)

where mar is margin of the IPFL, $g(l_i)$ represents the cascaded patch features. $p_i, n_i$ denotes positive and negative sample features, respectively.

The PLF loss of our model can be formulated as

\[
L_{\text{plf}} = L_{\text{ipfl}}^3 + \alpha L_{\text{ipfl}}^4 + \beta \left( \frac{1}{3} \sum_{a=1}^3 L_{\text{pedal}}^a + \frac{1}{4} \sum_{b=1}^4 L_{\text{pedal}}^b + \frac{1}{6} \sum_{c=1}^6 L_{\text{pedal}}^c \right)
\]  

(7)

Where $\beta$ controls the weight of the PEDAL. Following[14], the weight $\beta$ is set to 2.
2.5. The Final Loss of the Network
For the branch of ELF and PLF, we combine the ultimate loss for the network, which can be expressed as,

\[ L = (1-\alpha)L_{ELF} + \alpha L_{PLF} \]  \hspace{1cm} (8)

where \( \alpha \in [0,1] \) is utilized to balance the relative importance of the ELF loss and the PLF loss.

3. Experiment Result

3.1. Datasets

**Datasets**: Two widely person re-ID datasets are adopted as experimental evaluations, namely, Market-1501[15] and DukeMTMC-reID [16]. Market-1501 offers 1,501 identities of 32,668 pedestrian images. Each identity is photographed by up to 6 cameras in a university campus. For DukeMTMC-reID, it includes 1,404 identities of 36,411 person images. These pedestrian images are caught by 8 camera views which are not in the overlapped region of a university. For person re-ID, we evaluate the performance of our approach by the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP).

3.2. Implementation

We adopt ResNet-50 as the backbone which is pre-trained on ImageNet [17]. Meanwhile, we only remove the last fully-connected layer and keep the other layers of ResNet-50. We resize the input image to \( 384 \times 128 \) and exploit SGD to optimize the model. We perform random cropping, scaling, rotation, brightness, contrast, and saturation of an image. For Market-1501, the scaling number \( \sigma \) is put to 30. For DukeMTMC-reID, the scaling number \( \sigma \) is placed to 10. The learning rate is initialized to 0.0001, and then every 50 epochs decreased by 0.1. \( \alpha \) is put into 0.3.

3.3. Performance Comparison

**Market-1501**: We observe that our approach outperforms the compared methods from table 1. Specifically, our rank-1 accuracy has reached 70.5% and mAP attained 43.6% on Market-1501. When we compare our approach with the hand-crafted feature representation, it is clearly that we achieve an improvement by a large margin. This is also higher than the baseline [14] which is also the current best result by 2% in rank-1 accuracy. The comparisons indicate that fusing exemplar-level and patch-level features make our re-ID work better.

**DukeMTMC-reID**: As can be seen from table 1, our approach obtains the best performance among the compared methods with rank-1 = 74.7%, mAP = 54.3% on DukeMTMC-reID dataset. Specifically, our result outperforms the hand-crafted features by +56.2% in rank-1 accuracy. The rank-1 accuracy is 2.7% higher than the current best result PAUL. Therefore, we put forward the EPFL framework is awfully effective.

| Methods Ours(EPFL) | Market-1501 Rank-1 | Rank-5 | Rank-10 | mAP | DukeMTMC-reID Rank-1 | Rank-5 | Rank-10 | mAP |
|-------------------|--------------------|--------|---------|-----|----------------------|--------|---------|-----|
| LOMO[1]           | 70.5               | 83.6   | 87.6    | 43.6| 74.7                 | 84.2   | 86.6    | 54.3|
| BoW[15]           | 35.8               | 52.4   | 60.3    | 14.8| 17.1                 | 28.8   | 34.9    | 8.3 |
| UMDL[12]          | 34.5               | 52.6   | 59.6    | 12.4| 18.5                 | 31.4   | 37.6    | 7.3 |
| PTGAN[9]          | 38.6               | -      | 66.1    | -   | 27.4                 | -      | 50.7    | -   |
| PUL[18]           | 45.5               | 60.7   | 66.7    | 20.5| 30.0                 | 43.4   | 48.5    | 16.4|
| SPGAN[8]          | 51.5               | 70.1   | 76.8    | 22.8| 41.1                 | 56.6   | 63.0    | 22.3|
| TJ-AIDL[6]        | 58.2               | 74.8   | 81.1    | 26.5| 44.3                 | 59.6   | 65.0    | 23.0|
| HHL[7]            | 62.2               | 78.8   | 84.0    | 31.4| 46.9                 | 61.0   | 66.7    | 27.2|

Table 1. Performance(%) on Market-1501 and DukeMTMC-reID dataset.
3.4. Effect of Major Components

We perform detailed analysis to evaluate the effectiveness of each component in our method on DukeMTMC-reID dataset.

Effect of ELF: To investigate the effectiveness of the branch of ELF, ablation studies are conducted in table 2. We can see that the “PAUL+ELF” outperforms the baseline “PAUL”. Specifically, the rank-1 accuracy is raised from 72.0% to 72.9% on DukeMTMC-reID dataset. The major cause is that the global exemplar-level features exploited can maximize the diversity over the exemplars while maintain the similarity within each exemplar.

Effect of PLF: We train the model only with the branch of PLF to validate its effectiveness. Table 2 shows that the result of “PAUL+PLF” is better than the baseline PAUL. The rank-1 accuracy is increased by 1.2%. This is because fusing different patch features of the same image can mine more latent discriminative information.

| Method          | Rank-1 | Rank-5 | Rank-10 | mAP  |
|-----------------|--------|--------|---------|------|
| PAUL (baseline) | 72.0   | 82.7   | 86.0    | 53.2 |
| Baseline+ELF    | 72.9   | 82.9   | 86.5    | 53.6 |
| Baseline+PLF    | 73.2   | 82.9   | 85.8    | 54.0 |
| Baseline+ELF+PLF| 74.7   | 84.2   | 86.6    | 54.3 |

4. Conclusion

In this paper, we carefully design feature learning framework (EPFL) based on exemplar-level and patch-level for unsupervised person re-ID. ELF is exploited to learn global features of each example. Meanwhile, PLF is employed to learn local features. The global features from the whole images represent identities of different pedestrians and the local features show the detailed information. Fusing the global example-level features and local patch-level features, thus the discriminative pedestrian image features can be obtained. Specifically, for the sake of reducing the extra computation cost, we introduce the feature memory bank, and thus the accuracy can also be significantly improved. Experimental results verify the validity of our proposed approach on two common large-scale datasets, namely, Market-1501 and DukeMTMC-reID dataset.

Acknowledgments

This work is greatly supported by Natural Science Foundation of China (No. 61771180 and No. 61876056) a Anhui Province of Key Research Development Project (No. 1704d0802183). Thank you very much for encouragement.

References

[1] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence re-presentation and metric learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [ C ]. Boston, USA: 2015. 2197-2206.
[2] Wu J, Jiang J, Qi M, et al. Independent metric learning with aligned multi-part features for video-based person re-identification[J]. Multimedia Tools and Applications, 2019.
[3] Cuiqun, Chen, Meibin, et al. Optimizing Multi-Granularity Region Similarity for Person Re-Identification[J]. IEEE Access, 2019.
[4] Zhao R, Ouyang W, and Wang X. Unsupervised salience learning for person re-identification. In CVPR, 2013.
[5] Kodirov E, Xiang T, and Gong S. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In BMVC, 2015.
[6] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In CVPR, 2018.

[7] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero-and homogeneously. In ECCV, 2018.

[8] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In CVPR, 2018.

[9] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In CVPR, 2018.

[10] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong and Jian-Huang Lai. Unsupervised Person Re-identification by Soft Multilabel Learning. In CVPR, 2019.

[11] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification. In CVPR, 2019.

[12] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In CVPR, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[14] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based Discriminative Feature Learning for Unsupervised Person Re-identification. In CVPR

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

[16] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proc. IEEE Int. Conf. Comp. Vis., pages 3754–3762, 2017.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[18] H. Fan, L. Zheng, C. Yan, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018.

[19] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A Bottom-up Clustering Approach to Unsupervised Person Re-identification. In AAAI, 2019.