UNSUPERVISED PERSON RE-IDENTIFICATION USING RELIABLE AND SOFT LABELS

Jun Sun and Cheolkon Jung

School of Electronic Engineering, Xidian University, Xian, Shaanxi 710071, China
zhengzk@xidian.edu.cn

ABSTRACT

In this paper, we propose unsupervised person re-identification (ReID) using reliable and soft labels. We provide unsupervised person ReID to consider unknown pedestrian queries. We update ResNet model for person ReID based on reliable and soft labels. First, we perform unsupervised clustering on person images under different cameras, and select samples based similarity between images and cluster centers. Then, we conduct re-clustering for the selected samples and assign labels to them, i.e. reliable labels. We get probability of the unselected samples, i.e. soft labels. Finally, we update ResNet model for person ReID using reliable and soft labels. Experiments on Market-1501 and DukeMTMC-ReID demonstrate that the proposed method outperforms state-of-the-arts for unsupervised person ReID in terms of the cosine distance and accuracy.

Index Terms— Person re-identification, reliable labels, ResNet, soft labels, unsupervised clustering

1. INTRODUCTION

Person ReID aims at spotting a person of interest in other cameras [1]. It is a challenging task in computer vision due to the changes of scale, illumination, viewpoint angle, and pose. It is a core technology for video surveillance applications such as cross-camera tracking [2], multi-camera event detection [3], and person retrieval [3]. Up to now, a lot of outstanding results in person ReID have been achieved by researchers. Most researchers use supervised learning for person ReID. Ahmed et al. [4] presented a deep convolutional architecture with layers specially designed to address the problem of person ReID. Lin et al. [5] proposed a consistent-aware deep learning (CADL) approach for person ReID in a camera network. Sun et al. [6] proposed the optimization of the deep representation learning process based on Singular Vector Decomposition (SVD) to de-correlate the network for person ReID. Zhao et al. [7] exploited pairwise salience distribution relationship between pedestrian images and solved the person ReID problem by a saliency matching strategy. Zhang et al. [8] presented a deep mutual learning (DML) strategy for person ReID. Although supervised learning has made significant progress, person ReID may not have such a large label data for training. In practical situations, unsupervised person ReID is needed. Kordrov et al. [9] proposed dictionary learning-based sparse coding for unsupervised person ReID based on a graph Laplacian regularisation term. Fan et al. [10] proposed progressive unsupervised learning for person ReID to transfer pre-trained networks to unknown pedestrian data. Zhao et al. [11] proposed adjacency constrained patch matching for person ReID to build dense correspondence between images pairs and learn human saliency in an unsupervised manner. Peng et al. [12] presented a multi-task dictionary learning method which was able to learn dataset-shared and target-data-biased representation for person ReID. However, since they performed unsupervised person ReID on small datasets, their performance was limited in large-scale dataset. When a supervised person ReID model such as ResNet [13] is applied to unknown data, it causes severe performance degradation. Thus, unsupervised person ReID is required in a real environment.

In this paper, we propose unsupervised person ReID using reliable and soft labels. We introduce reliable and soft labels into ResNet model update to transfer the model to unlabeled dataset. We adopt camera viewpoint-based unsupervised clustering for person ReID to be robust to viewpoint change. First, we fine-tune ResNet model on an irrelevant dataset and use it as an initial model. We perform unsupervised clustering on person images under different cameras to consider viewpoint change. We select samples close to cluster centers and assign labels to them, i.e. reliable labels. For unselected samples, we assign its probability to them, i.e. soft labels. Finally, we update ResNet parameters based on reliable labels and soft labels. Fig. 1 illustrates the flow diagram of the proposed method. Compared with existing methods, main contributions of the proposed method are as follows:

- We provide an unsupervised person ReID framework based on reliable and soft labels.
- We introduce the reliable and soft labels into person ReID to update ResNet model for unknown people data.

This work was supported by the National Natural Science Foundation of China (No. 61872280) and the International S&T Cooperation Program of China (No. 2014DFG12780).

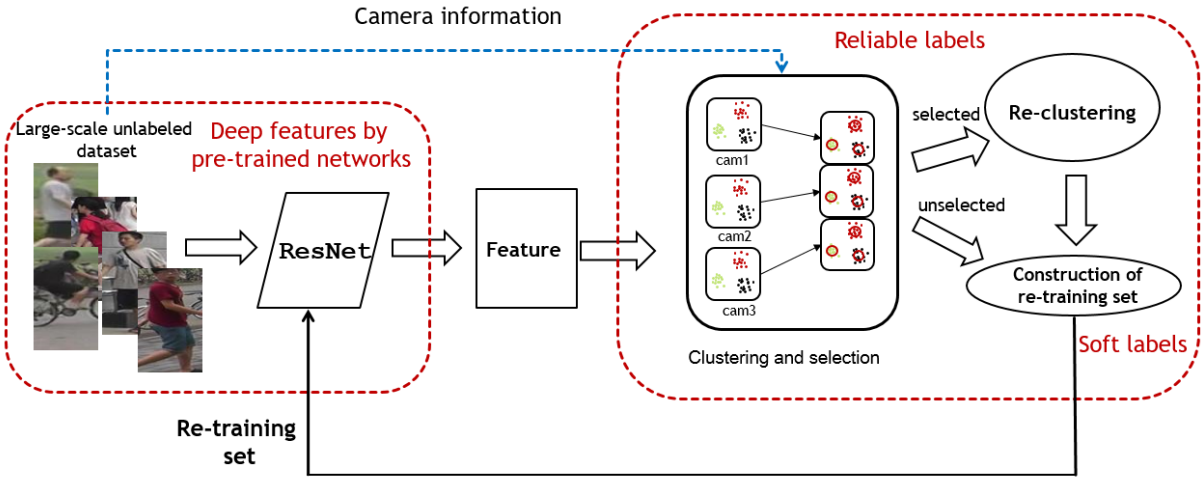

Fig. 1. Flow diagram of the proposed unsupervised person ReID.

- We use camera viewpoint-based clustering for person ReID to deal with the viewpoint change problem.

2. PROPOSED METHOD

2.1. Camera-Based Clustering

Progressive unsupervised learning (PUL) [10] have used iterations to update convolutional neural network (CNN) model and clustering results for unknown people data by: 1) Pedestrian clustering and 2) fine-tuning for CNN. Although the clustering results make the model better, PUL does not fully consider the information of unlabeled dataset. This is because PUL only uses a threshold to select reliable labels without using unselected images and considering camera information. Assume that we have an unlabeled training dataset $X = \{X_a, X_b, X_c\}$, $X_a = \{x_1^a, x_2^a, ..., x_{N_a}^a\}$,$X_b = \{x_1^b, x_2^b, ..., x_{N_b}^b\}$,$X_c = \{x_1^c, x_2^c, ..., x_{N_c}^c\}$, where X_k represents the k-th camera, x_j^i represents the j-th image from camera i, and N_k represents the total number of images in the k-th camera. We use the pre-trained model to extract the features of X as follows:

$$F = \left\{f_1^a, f_2^a, ...f_{N_a}^a, f_1^b, f_2^b, ...f_{N_b}^b, f_1^c, f_2^c, ..., f_{N_c}^c\right\} \quad (1)$$

$$f_i^k = \phi(x_i^k) \quad (2)$$

where f_i^k represents features extracted by the pre-trained model of i-th images in the k-th camera, N_k represents the total number of images in the k-th camera, and ϕ represents the pre-trained model. We use F to do clustering by optimizing:

$$\min_{\mu_1,\mu_2,...,\mu_{M_k}} \sum_{i=1}^{N_k}\sum_{j=1}^{M_k} \left\|f_i^k - \mu_j^k\right\| \quad , \ k \in \{a,b,c\} \quad (3)$$

where μ_j represents the j-th cluster center of the k-th camera, N_k represents the total number of images in the k-th camera set, and M_k represents the number of identities in the k-th camera. Thus, we get the clustering results as follows:

$$U^k = \left\{\mu_1^k, \mu_2^k, ...\mu_{M_k}^k\right\} \quad , \quad k \in \{a,b,c\} \quad (4)$$

Then, we calculate the similarity matrix S based on a threshold λ to select reliable images as follows:

$$S^k = \left(U^k \times F^k\right) V^k \quad , \quad k \in \{a,b,c\} \quad (5)$$

where F^k represents the features in the k-th camera and V^k indicates whether this point is selected or not. If $S^{i,j} > \lambda$, the image is selected, which is set $v^{i,j}$ to 1, else set to 0. In this work, we set λ to 0.85 empirically. Finally, we use the selected images as a new set for re-clustering as follows:

$$\min_{\mu_1,\mu_2,...,\mu_M} \sum_{i=1}^{N}\sum_{j=1}^{M} \left\|f_i - \mu_j\right\| \quad (6)$$

where f_i represents the feature of the selected images, μ_j represents the cluster centers, N represents the total number of the selected images, and M represents the number of identities.

2.2. Reliable and Soft Labels

For the selected images, we obtain M labels from the clustering results by Eq. (6), called reliable labels. Reliable labels are virtual labels on the sample related to the clustering result, and have no relation with real labels. For the unselected images, we assign the probability to them, called soft labels. We first consider the distance among cluster centers $C = \{\mu_1, \mu_2, ..., \mu_M\}$ and unselected images

Table 1. Performance comparison on Market-1501 dataset in terms of cumulative matching scores at Rank1, Rank5 and Rank10. Bold numbers represent the best performance.

Method	Rank1	Rank5	Rank10
Bow	35.80%	52.40%	60.30%
LOMO	27.20%	41.60%	49.10%
UMDL	34.50%	52.60%	59.60%
PUL	42.69%	57.93%	64.87%
Ours	**44.86%**	**59.56%**	**66.27%**

Table 2. Performance comparison on DukeMTMC-ReID dataset in terms of cumulative matching scores at Rank1, Rank5 and Rank10. Bold numbers represent the best performance.

Method	Rank1	Rank5	Rank10
Bow	17.1%	28.8%	34.9%
LOMO	12.3%	21.3%	26.6%
UMDL	18.5%	31.4%	37.6%
PUL	30.0%	43.4%	48.5%
Ours	**32.84%**	**45.32%**	**53.27%**

$F' = \{f_1, f_2, ..., f_{N'}\}$ as follows:

$$S' = CF' \tag{7}$$

where F' represents the l2 norm features of the unselected images, and N' represents the total number of the unselected images. We use the similarity as the input for softmax function. To make it adjust to the person ReID problem, we use relaxation parameters to control as follows:

$$prob = \exp\left(\frac{s_i}{t}\right) / \sum_{i=1}^{M} \exp\left(s_i/t\right) \tag{8}$$

where s_i is the i-th column of the similarity matrix S', M is the total number of cluster centers, t is the relaxation parameter. If t is small, the distribution is sharp; otherwise, smooth. Here, we set t to 0.02.

2.3. ResNet Model Update

Based on reliable and soft labels, we construct re-training dataset and update ResNet model for person ReID. Finally, we obtain new parameters for ResNet model on unknown people data.

3. EXPERIMENTAL RESULTS

We evaluate the proposed method on Market-1501 [14] and DukeMTMC-ReID [15] which have camera viewpoint information with large-scale. Market-1501 is collected in front of a supermarket in Tsinghua University using six cameras. It contains 32,668 annotated bounding boxes of 1,501 identities. The dataset employs Deformable Part Model (DPM) [16] as the pedestrian detector. The dataset is split into three parts: 12,936 images of 751 identities for training, 19,732 images of 750 identities for gallery in the test stage, and 3,368 images of 750 identities for query in the test stage. DukeMTMC-ReID is a subset of the multi-target multi-tracking dataset [17] for image-based ReID. It contains 36,411 images of 1,812 identities captured by 8 cameras. We split the dataset into three parts like Market-1501: 16,522 images of 702 identities for training, 17,661 images of 1,110 identities for gallery in the test stage, 2,228 images of 702 identities for query in the test stage. We evaluate Rank1, Rank5, Rank10 accuracy for the two datasets, and all experiments follow the strategy of single query. We use pre-trained ResNet-50 model [13] on ImageNet as the initial CNN model. Followed PUL [10], we update the fully-connected layer to adapt to different datasets and insert a dropout layer before the fully-connected layer. All images are resized to 224×224. We randomly shift images horizontally and vertically within 45 pixels and rotate with 20 degrees. We set the batch size as 16. In the feature extraction stage, we use the output of the average-pooling layer of ResNet-50 as the feature. In the clustering stage, we use k-means clustering and l2-normalization of features to compute similarity. To keep the consistency with PUL [10], we set the number of clusters k to 750 for Market-1501 and 700 for DukeMTMC-ReID. In the testing stage, we use the output of the average-pooling layer of the trained ResNet-50 as the feature and utilize cosine distance to measure the similarity. For experiments, we use a PC with Intel-Xeon E5-2640 v3 2.6GHz CPU, 32GB RAM and Nvidia TITAN X 12GB.

We compare the performance of the proposed method with those of Bow [14], LOMO [18], UMDL [12], PUL [10]. Table 1 provides performance comparison on Market-1501 in terms of cumulative matching scores at Rank1, Rank5 and Rank10. The proposed method performs better than the other unsupervised ones. Rank 1 accuracy of the proposed method reaches 44.86% in Rank1 performance. Table 2 shows performance comparison on DukeMTMC-ReID in terms of cumulative matching scores at Rank1, Rank5 and Rank10. Fig. 3 shows person retrieval results. Due to the limited performance of unsupervised person ReID, persons with similar clothing and body size cause wrong detection as shown in the first and fourth rows. In the second and fifth rows, it can be observed that the proposed method successfully adapts to viewpoint change. Even if we have only back information of persons, we get correct matching results including other viewpoint images.

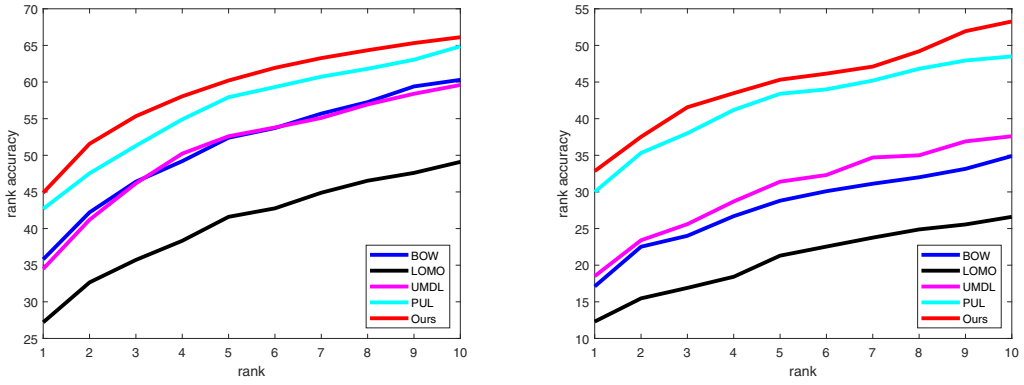

Fig. 2. Cumulative match characteristic (CMC) curves. (a) Market-1501. (b) DukeMTMC-reID.

Fig. 3. Retrieval results. Left: Query. Right: Retrieval results from Rank1 to Rank10. Red boxes indicate correct matching results.

4. CONCLUSION

In this paper, we have proposed unsupervised person ReID using reliable and soft labels. We have updated ResNet model using camera viewpoint-based unsupervised clustering. First, we have obtained reliable and soft labels by camera viewpoint-based unsupervised clustering on person images. Reliable labels are assigned to the selected images whose distance is close to the cluster centers, while soft labels are the probability of the unselected images. Experimental results demonstrate that the proposed method achieves Rank1 accuracy of 44.86% in Market-1501 and 32.84% in DukeMTMC-ReID in terms of cosine distance as well as outperforms state-of-the-arts for unsupervised person ReID. In the future, we will combine discriminative loss (e.g. triplet loss or structured loss) with classification loss for network training to get better features for person ReID.

5. REFERENCES

[1] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *arXiv:1610.02984*, 2016.

[2] Xiaogang Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013.

[3] Change Loy Chen, Tao Xiang, and Shaogang Gong, "Multi-camera activity correlation analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1988–1995.

[4] Ejaz Ahmed, Michael Jones, and Tim K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[5] Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 3820–3828.

[7] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *Proceedings of the IEEE Conference on Computer Vision*, 2014, pp. 2528–2535.

[8] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] Elyor Kodirov, Tao Xiang, and Shaogang Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *Proceedings of the British Machine Vision Conference*, 2015, pp. 44.1–44.12.

[10] Hehe Fan, Liang Zheng, and Yi Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 14, no. 4, pp. 83, 2017.

[11] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

[12] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1306–1315.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE Conference on Computer Vision*, 2015, pp. 1116–1124.

[15] Zhendong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 3774–3782.

[16] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[17] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.