Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230832

Evaluation of deep learning-based algorithms on person re-identification

Qiaochu Xu

Department of Statistical Science, University College London, London, UK

qiaochu.xu.22@ucl.ac.uk

Abstract. At present, CCTVs and multifunctional cameras have been well equipped everywhere in our daily lives. However, most of the public sectors especially the department of public safety are still using the traditional approach, analysing these data by humans. This research presents a summary of the current technologies and algorithms based on deep learning in person Re-identification (Re-id), which aims to recognize a particular individual in several camera views. This paper firstly described the current challenges in using deep learning for person Re-id. Then, a brief introduction of standard person Re-id including feature learning and metric learning is given. Furthermore, the open world Re-id problem is examined from three aspects: person re-id with occlusions, unsupervised learning-based method, and cross-modal person re-Id. The author then evaluated the regularly used datasets and analysed the performance of the algorithms from recent times. Finally, some potential future directions of research for Re-id are proposed.

Keywords: DNN, person re-id.

1. Introduction

IHS Markit estimates that by the end of 2021, more than one billion security cameras will have been installed worldwide. This means people have experienced intensive surveillance in recent days. However, most of the departments that preserved and controlled these data, especially the department of public safety, they are still analyzing these data in a traditional approach, by human force. Except for the human cost, the most critical problems by human analysis are time-consuming and the challenge of inaccuracy. As in some specific scenarios, our security sector needs a real-time response to incoming events. Thus, constructing an intelligent surveillance system is of vital significance.

Person Re-identification (Re-id) was established with the advancement of the computer vision community. The person Re-id aims to identify a specific person among various cameras on different occasions. Generally speaking, a complete Person Re-identification includes three parts: person detection, person tracking, and person retrieval. This paper will discuss the most critical part which is the person retrieval.

As the technology in Re-identifying people in different cameras, there exist many challenges different from other computer vision technologies:

- Scarce resolutions: CCTV usually provides images or videos with a scarce resolution.

DOI: 10.54254/2755-2721/6/20230832

- Different standards: Due to the fact that the cameras are not equipped to the same standard, thus the cross-camera images might experience the problems of different illumination, distinct background, and different imaging style.
- Target incomplete: In some scenarios, the target person might be covered (e.g.covered by vehicles or other people).
- Limited training data: Compared with the datasets used in other computer vision projects, the datasets of person Re-id is relatively small. In addition, the current method of labelling is mainly by people and it is difficult to identify two strangers with a similar appearance.

The remaining section of the paper will be grouped as follows, in section 2, the standard person Re-id method is given which includes the current algorithms in feature learning and metric learning. Section 3 will focus on the open world person Re-id method and introduce several algorithms in solving the problem of person Re-id with targets covered, some approaches related to unsupervised learning, and algorithms related to cross-modal Re-id. Section 4 listed some datasets in person-Re-id and evaluates the performance of recent algorithms. In the last section, the author proposed some future possible directions for this research.

2. Close-world re-identification

The person Re-id has been split into two parts: close-world case and open-world case. The close-world person Re-id is the problem under the ideal conditions without much influence on the illumination, different angles of view. The close-world algorithms mainly focus on people's feature representation and similarity metrics. In this section, some algorithms related to feature learning and deep metric learning are given.

2.1. Feature learning

This section briefly introduces some algorithms based on supervised feature learning which is a technique that can make a system automatically generate the representations needed for feature detection and classification from raw data.

2.1.1. Image based feature learning. In order to extract one's features from images, traditional methods usually focus on one's appearance. These approaches usually used the perceptual principles of symmetry and asymmetry to identify the position of one's head, torso, and legs and used weight distance to leverage the complementary information of other body parts. However, people's pose is unpredictable in real scenarios, and features that coded artificially are not as precise as the real one. With the resurgence of deep learning, people are gradually interested in methods related to Convolutional Neural Networks (CNN). [1] presented a (PDC) model which can simultaneously learn the representations of the whole body and individual body parts. This model is composed mostly of two subnets: a feature embedding subnet (FEN) and a feature weighting subnet (FWN). FEN first extracted the feature from 6 body parts and fused it with FWN to generate global representations. In general, CNNs collect obvious features within the first few layers and construct deeper abstract concepts, such as pairings, at the middle levels. Similarly, when compared with the high-level feature, the mid-level features become particularly important and informative. Hence, [2] provides a gating function that extracts local patterns from the mid-level and amplifies the local similarities throughout the last few layers in order to propagate more pertinent network patterns to the higher layers.

However, since the aforementioned approaches learn the local feature representation from relevant section just once, the performance may still be affected by factors like brightness difference and occlusion if the acquired local feature lack robustness. Besides CNNs, researchers also learn from the human way of thinking. [3] proposed a comparative attention network (CAN) based on the attention mechanism dynamically generating discriminative representations in a periodic manner of catching and matching an individual's photos for automatically locating the most recognized aspects of people. Due to the problem of poor stability and robustness of the feature representation with single granularity, and the deepening of the network depth will lead to the gradual loss of image feature

DOI: 10.54254/2755-2721/6/20230832

information, designing a multi-granularity deep feature fusion representation method has become the mainstream method to improve the ability of person feature representation. Moreover, another network architecture based on multi-directional attention networks called HydraPlus-Net (HP-Net) [4] achieves a better performance. This network trains multi-scale attention-enhanced features for high granularity human retrieval tasks and captures the attention not only at low-level but also integrate more abstract and complex semantic attention. In addition, image feature segmentation is a common method in local feature collection, which generally slices the feature mapping through a deep neural network into a various of exclusive predefined strip or regions and performs local feature learning independently for each region.

2.1.2. Video based feature learning. The main difference between video-based feature learning and image-based feature learning is that the Image-based method lacks spatial and temporal cues. In addition, the raw data acquired from CCTV or Surveillance systems are videos. The most popular or commonly used method for Video-based feature learning is the Recurrent Convolutional Network which is a kind of artificial neural network with an infinite impulse response and the connection between each node form a cycle that allows the output to influence some subsequent inputs. [5] proposed a novel recurrent neural network (RCN) that uses CNN which contains a recurrent final layer so that information can flow across time frames. In addition, they employ temporal pooling to generate guise patterns for the entire sequence.

In order to make optimal use of the temporal and spatial data, [6] present a CNN based on a temporal attention model (TAM) to jointly learn features and metrics. This model applies a temporal recurrent layer to obtain features temporally and gives variable weights to each frame in an image stream. Thus, this model can pay attention to more specific and discriminative images. When comparing the similarities between two images, they integrate the surrounding information to achieve a better metric by the spatial recurrent model (SRM). Similar to image-based feature learning, some researchers begin developing models based on attention mechanisms after persistently focusing on selecting more informative time steps, more discriminative temporal cues, and more efficient similarity measures. To determine the similarity of feature representations of video pairs, [7] presented a (SCAN) model, applying an unchangeable attention module and a generalized similarity measurement module. SCAN first uses a shared CNN to acquire frame-level features and then self-attention Subnetworks and collaborative attention subnetworks generate temporal representations.

2.2. Distance Metric Learning (ML)

Distance metric learning (ML) may be used to measure the similarity of two objects, and is also commonly employed in Person Re-id tasks. The ML method generally learns the similarity of multiple images and constructs the loss function according to the difference between image labels and similarity. The similarity between photographs of the same person in the person retrieval problem is greater than the resemblance among images of different walkers. By this property, metric learning method is widely used in this topic. This section starts by describing some usual loss functions, and then, reviews some methods based on metric learning.

Supervised contrastive loss: The most commonly used loss function in SNN is the contrastive loss function proposed by Yann LeCun. The expression of supervised contrastive loss function is

$$L_c = \frac{1}{2N} \sum_{I=1}^{N} [yd^2 + (1-y)(\alpha - d)_+^2] \tag{1}$$

where d stands for the euclidean distance of two objects,

$$y = \begin{cases} 1, & label\ is\ matched \\ 0, & label\ not\ matched \end{cases} \tag{2}$$

Triplet loss is another loss function that is extensively employed in person re-identification with expression given in equation (3). This requires three input images, namely the Anchor, the Positive, and the Negative.

DOI: 10.54254/2755-2721/6/20230832

$$L_T = \frac{1}{N_T} \sum_{m=1}^{N_T} \max\{d_{a,p}^m - d_{a,n}^m, \alpha\} \quad (3)$$

where $\alpha < 0$, N_T is the number of input images, $d_{a,p}^m$ is the distance of the positive, $d_{a,n}^m$ is the distance of the negative.

In Person Re-id, the similarities of the images of different pedestrians might be high due to illumination, change of background and change of body gesture. [8] conclude a global loss term based on higher order statistics that investigates the overall structure of the embedding space in order to optimize the local loss. The [8] summarized how to employ an integrated architecture that combined both deep and shallow neural networks to spontaneously collect feature and similarity metrics and to tune the neural networks via extensive triplet sampling.

3. Open-world person re-identification

Open-world person re-id is closer to the real scenario that in an uncertain spatial there might be partial overlap between large amounts of cross-camera data and the candidate dataset not including all the identities which might appear. If a new person's identity is detected, the model will extract the features and add them to the person identification dataset. In the open-world Person Re-id scene, there exists plenty of influencing factors e.g., brightness, viewpoint, body gesture, occlusion, and also face a problem of retrieval speed caused by massive data. This section will first introduce some algorithms used in addressing the problem of Re-id with occlusions, and then describe the current unsupervised learning-based method and some possible approaches in cross-modal Re-id.

3.1. Occluded person-re-id

The study of the problem of occluded person retrieval mainly focuses on how to acquire feature representation and appropriately metric the similarity when part of pedestrian information is occluded. [9] proposed a CNN framework to work out re-id with occlusion which is the first attempt to define the person retrieval with occlusion problem. The framework they proposed is called the Attention Framework of Person Body (AFPB), consisting of two components. Firstly, the occlusion simulator will create an occlusion on a full-body person image to mimic the occluded image. Then, the Multi-task losses compel the neural network to deduce if a sample comes from the body data distribution as a whole or the data with occlusion, and utilize that information to develop feature representation for the occluded problem. [10] proposed a (IGOAS) architecture. This network first generates hierarchical occlusion data from easy to hard and then uses an occlusion suppression module to focus less on the irrelevant background and pay attention to the foreground.

3.2. Unsupervised learning

In the practical application scenarios, most of the targets that need to be retrieved are unlabeled data, and the labeling work is extremely difficult. However, the majority of the existing deep network approaches require extensive labeled data for supervised training to perform superior. Consequently, some unsupervised people retrieval approaches are presented to overcome the data labelling challenge.

3.2.1. Single-Domain Unsupervised learning. The majority of current unsupervised re-identification research focuses on cross-domain analysis. This study aims to transfer the pre-trained model from the source domain (SD) that usually contains labeled data to the target domain (TD). A cluster-based method called Bottom-Up Clustering (BUC) [11] jointly optimizes a CNN and the identity of different samples. This model first uses the repelled loss to utilize the similarity inside each cluster, and then diversity regularization terms to permit a balanced number of images in each cluster. In addition, cluster-based method with the objective of learning special projection for each viewpoint through simultaneously learning the asymmetric metric and obtaining optimal classification.

3.2.2. Cross-domain Unsupervised learning. The majority of current unsupervised re-identification research focuses on cross-domain analysis. This study aims to transfer the pre-trained model from the

DOI: 10.54254/2755-2721/6/20230832

SD that usually contains labeled data to the TD. Generally, since there exists a certain distinction between the SD and TD, which might cause unwanted results. The ideal cross-domain method will eliminate this problem by narrowing the gap between these domains solving the cross-domain problem by learning the domain-invariant feature. [11] proposed a (PUL) model which is an iterative process consisting of two components. They first apply transfer learning using another CNN pre-trained on an externally labeled dataset to obtain the feature representation of their unlabeled training set and then, fine-tuning the network based on the selected training sample. They then repeat iterating this process to achieve a stronger model.

3.2.3. GAN (Generative adversarial networks). Although there are some unsupervised learning algorithms, all of these methods require a hypothesis that the label in each domain is the same. However, most Re-id datasets carry different identity information, and therefore, it cannot apply the above unsupervised methods directly. Generative Adversarial Networks (GAN) include a generative network and a discriminative network. The generative network will continuously generate fake samples for the discriminative network to evaluate and it is an iterative training process improving both networks. The core idea of applying GAN based model is to allow images of people to transfer across different datasets without changing the identity but they change other attributes such as background, illumination, etc. [10] present a Multi-camera Transfer GAN (CTGAN) which allows to transfer images across domain to domain by using only one model. In addition, they utilized a method called Mixed Selective Convolution Descriptor Aggregation, which filtered out background noise while preserving the useful depth descriptors.

3.3. Multi-modal person re-id

In reality, the visible light camera collects the modal data, but the image quality will be blurred. Scarce resolution, inconsistent brightness, and other problems will affect the quality of visible light cameras to capture pedestrian images, causing low retrieval precision in real applications. Since infrared (IR) cameras are widely equipped in recent years, some believe integrating RGB images with IR images is required. (DZP) structure for developing domain-specific structure automatically in OneStream networks suited for RGB-IR Re-ID tasks was presented in 2017. Furthermore, depth maps are also used in person re-identification. Many methods combined the RGB, depth, and thermal images in the re-id system to extract the color information, biometrics, and local structure information from the RGB image, the depth map, and thermal images respectively. Besides all the image-based methods, the text description is also a crucial identification approach. [12] present an (ARLTM) architecture to transfer targets into semantic level integrated both linguistic information and visual information. However, the description provided by the witness cannot always be correct and complete which will influence the final results.

4. Datasets and commonly used algorithms

4.1. Datasets

Deep learning is highly reliant on a vast amount of training data. Currently, the most open datasets for Person Re-id are image datasets including Market-1501, CHUK01, DukeMTMC-ReId, CHUK-03, VIPeR, PRID-2011 and this section gives a brief introduction of these datasets:

- Viewpoint Invariant Pedestrian Recognition (VIPeR) [10]: The VIPeR dataset uses two outdoor cameras catching 632 people from various angles and illumination situations.
- Market-1501 [10]: This dataset, published in 2015, contains 1501 identities and 32668 images collected by five standard cameras and one low-resolution camera.
- Duke Multi-Tracking Multi-Camera ReIdentification (DukeMTMC-ReId) [11]: This data collection represents a subsection of the DukeMTMC. The dataset is comprised of high-definition films from 8 cameras capturing 16522 photos from 702 identities.

DOI: 10.54254/2755-2721/6/20230832

- CHUK01 [10]: This dataset contains 971 identities with two cameras that took 2 images for each person.
- CHUK03 [11]: This dataset contains 1467 identities and 19732 images taken from 5 different cameras.
- PRID2011 [10]: Multiple human trajectories captured by two distinct static surveillance cameras are included in the dataset with 200 individuals appear in both perspectives. 178 of the 200 individuals have made more than 20 appearances.

Table 1. The information of datasets.

Dataset	Image/Video	Published Time	Cameras	Identities	Image/Video
VIPeR	Image	2007	2	632	1264
CHUK01	Image	2012	2	971	3884
CHUK03	Image	2014	10	1467	14096
Market-1501	Image	2015	6	1501	32668
DukeMTMC-ReID	Image	2017	8	1812	36411
PRID-2011	Video	2011	2	200	400
iLIDS-VID	Video	2014	2	300	600
MARS	Video	2016	6	1261	20715

4.2. Image based algorithm

The table1 below illustrates some recent image-based person Re-identification algorithms. Most of the algorithm has been examined on the CUHK-03, DukeMTMC-reID, and Market-1501 datasets. Besides the algorithm mentioned in section 2.1.1, (InSTD) [10], identity-guided human semantic parsing approach (ISP) [11] and Pyramid-Net [13].

Table 2. Image-based algorithms.

Method	Published Time	Description	Market-1501		CUHK03		DukeMTMC-reID	
			R-1	mAP	R-1	mAP	R-1	mAP
MG	2016	Matching Gate function	76.04	48.45	68.1	88.1	-	-
CAN	2017	Attention Mechanism	60.3	35.9	77.6	-	-	-
PDC	2017	Local feature	84.14	63.4	88.7	-	-	-
HP-Net	2017	Metric learning	76.9	-	91.8	-	-	-
Pyramid-Net	2019	Pyramid model	95.7	88.2	76.9	78.9	89.0	79.0
ISP	2020	Clustering local feature	95.3	88.6	76.5	74.1	89.6	80.0

DOI: 10.54254/2755-2721/6/20230832

Table 2. (continued).

InSTD	2021	Local feature	97.6	90.8	-	-	95.7	89.1

The supervised image-based method overall has shown an improvement since 2016. In 2021, the accuracy of InSTD achieved 97.6% rank-1(R1) accuracy on the Market-1510 dataset with 90.8% mean average precision (mAP). Among these algorithms, most of them have applied local feature learning especially using the attention mechanism. Some of the algorithms, such as HP-Net, also applied metric learning and comparative loss, optimizing the performance of algorithms. Therefore, how to obtain more strong and distinguishable local representation of pedestrians is an important future research direction.

4.3. Video-based algorithms

Table 2 shows the performance of video-based approaches from 2016-2020. Similarly, the performance of these video-based approaches is assessed using Rank-1 and Rank-5 accuracy on the three datasets MARS, PRID-2011, and iLID-VID. In addition, QAN [10], ASTA-Net [10], RGST [10] and GLTR [10] are added for comparison. The GLTR based on the attention mechanism has achieved 95.3% rank-1(R-1) accuracy and 100% rank-5(R-5) accuracy on the PRID-2011 dataset in 2019, pushing the video-based method to a higher level. Metric learning-based methods such as QAN also achieve great performance with 90.3% rank-1 and 98.2% rank-5 accuracy (table 3). Although these innovations are astounding, they also increase the difficulty of subsequent study. Future research will target broader more complicated scenarios and more realistic datasets for potential breakthroughs. Nevertheless, larger scale video data will necessitate more computational power, which are likely increasing the demands on hardware devices. In addition, how to deploy video re-id to a comprehensive surveillance tracking system will be a crucial aspect of future study.

Table 3. Video-based algorithms.

Method	Published Year	Description	MARS		PRID-2011		iLIDS-VID	
			R-1	R-5	R-1	R-5	R-1	R-5
RCN	2016	RNN+CNN	-	-	70.0	90.0	58.0	84.0
TAM&SRM	2017	Attention Mechanism	70.6	90.0	79.4	94.4	55.2	86.5
QAN	2017	Triplet loss	-	-	90.3	98.2	68.0	86.8
SCAN	2019	Temporal Representation	87.2	95.2	95.3	99.0	88.0	96.7
GLTR	2019	Attention Mechanism	-	-	95.5	100.0	86.0	98.0
ASTA-Net	2020	Attention Mechanism	90.4	97.0	96.4	100	88.1	98.6
RGST	2020	Attention Mechanism	89.4	96.9	93.7	99.0	86.0	98.0

5. Conclusion

Currently, both image- and video-based methods have made outstanding achievements in close-world case. Nevertheless, in the study of generalized Person Re-id, there still exists some challenges that need to be tackled, and these challenges might be the future research directions:

1) The occlusion problem is one of the biggest challenges. Not only in Re-id but other computer vision fields such as facial verification also regard it as problematic.

DOI: 10.54254/2755-2721/6/20230832

2) Unsupervised learning: Although a large amount of data can be acquired from cameras, image annotation and labeling are still challenging work and hard to accomplish in reality. Thus, unsupervised methods are of vital importance in address this problem. The current unsupervised approaches are based on learning invariant features, and a performance gap still exists with the supervised methods.

3) Cross-modal Re-id: Most current methods focus on RGB image identification. However, this image is frequently influenced by illumination and low resolutions. The cross-modal Re-id can integrate infrared images, the depth map, and even text images, and it is more consistent with the device diversity and complexity of intelligent surveillance systems in real life.

References

[1] Su, C., Li, J., Zhang, S., Xing, J., Gao, W., & Tian, Q. Pose-driven deep convolutional model for person re-identification. 2017， Int. Conf. compute. Vis. 3960-3969.

[2] Varior, R. R., Haloi, M., & Wang, G. Gated siamese convolutional neural network architecture for human re-identification. 2016 Euro. Conf. Com. Vis. 791-808.

[3] Liu, H., Feng, J., Qi, M., Jiang, J., & Yan, S. End-to-end comparative attention networks for person re-identification. 2017, IEEE Trans. Image Proc., 26(7), 3492-3506.

[4] Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis.2020, Int. Conf. compute. Vis. 350-359.

[5] McLaughlin, N., Del Rincon, J. M., & Miller, P. Recurrent convolutional network for video-based person re-identification. 2019, Int. Conf. compute. Vis. 1325-1334.

[6] Zhou, Z., Huang, Y., Wang, W., Wang, L., & Tan, T. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. 2017 Int. Conf. compute. Vis. Pat. Rec. 4747-4756.

[7] Zhang, R., Li, J., Sun, H., Ge, Y., Luo, P., Wang, X., & Lin, L. Scan: Self-and-collaborative attention network for video person re-identification. 2019 IEEE Trans. Image Proc., 28(10), 4870-4882.

[8] Lu, J., Chen, X., Luo, M., & Wang, H. Person Re-Identification Research via Deep Learning. 2020, La. Opt. Pro. 57(16), 160003.

[9] Zhuo, J., Chen, Z., Lai, J., & Wang, G. Occluded person re-identification. 2018 Int Conf Mult Expo. 1-6.

[10] Zhao, C., Qi, D., Dou, S., Tu, Y., Sun, T., & Bai, S. Key technology for intelligent video surveillance: a review of person re-identification. 1979, In Sci. Sin Inf., 51(12).

[11] Wei, W., Yang, W., Zuo, E., Qian, Y., & Wang, L. Person re-identification based on deep learning-An overview. 2021, J. Vis. Com. Image Rep. 103418.

[12] Xia, D.,Guo, F.,Liu, H. & Xia, Y. Review on Research Progress of Open-World Person Re-identification. 2021 J Data Acc. Proc. (03),449-467.

[13] Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Ji, R. Pyramidal person re-identification via multi-loss dynamic training. 2022, Int. Conf. compute. Vis. Pat. Rec. 8514-8522.