Person Re-identification on Mobile Devices Based on Deep Learning

MIAOMIAO ZHU ^{*†}	Member,	SHENGRONG GONG [‡]	Non-member
ZHENJIANG QIAN [‡]	Non-member,	SEIICHI SERIKAWA [†]	Member
LIFENG ZHANG [†]	Member		

(Received November 29, 2020, revised January 24, 2021)

Abstract: As an important supplement to face recognition, person re-identification can solve the cross-camera and cross-scene pedestrian recognition and retrieval, determining whether a specific pedestrian in the image or video sequence. At present, many person re-identification related researches carried out the verification and evaluation with experiments on large person re-identification datasets such as Market-1501, DukeMTMC-reID, MSMT17, and CUHK03. This research aims to propose a complete real available person re-identification with mobile devices. Based on the existing deep-learning person re-identification research, they combine with the actual application scenarios on the premise of analyzing the technical feasibility, which aims to connect pedestrian detection with person re-identification to perform real-time pedestrian detection and query. In this process, not only can the features extracted by pedestrians be reused, but the research on person re-identification can be applied effectively, such as tracking criminals and searching for missing children.

Keywords: Deep-learning, CNN, Pedestrian detection, ReID, Intelligent security, Mobile Devices,

1. Introduction

With the rapid development of computer science and technology, artificial intelligence research and application are gradually expanding. Every country in the world has paid much attention and support to the construction of a safe city. One of the most important parts is installing a large number of video surveillance cameras to quickly and accurately obtain real-time information in the scene. According to statistics, until September 2019, China had built the world's largest video surveillance network with more than 200 million video cameras.

Urban video surveillance equipment installation provides a sufficient guarantee for a safe society, especially in traffic management and criminal case investigations. However, due to the vast number of videos and images captured by cameras distributed in different regions, once a case needs to be resolved, the workload is enormous if only relying on traditional ways of human to locking and analyzing the target under the multi-camera video surveillance network, to obtain information such as location, time and other information. Moreover, it is easy to miss the information because of various limitations, which will affect the efficiency and accuracy significantly. Then, how to effectively use the face and pedestrian images captured by these cameras for real-time detection and judgment and effectively make use of it

further has become an important research topic in computer vision.

At present, face recognition has been effectively used on many occasions. However, in many practical application scenarios, due to the environment's effects such as resolution, viewpoint, illumination, etc., most surveillance cameras are difficult to obtain high-quality face images, just a pedestrian image. At this time, person re-identification, as a useful supplement to face recognition, can recognize pedestrians based on information such as wear, pose, hairstyle, etc., and has become important research after face recognition technology.

Simultaneously, apart from the cameras in the surveillance network, mobile devices such as mobile phones are currently the popular means used by people to obtain videos and images. Face recognition based on mobile devices has been effectively applied on many occasions, such as photo classification and real-time face authentication unlock and login in APP, and so on.

In this paper, based on the existing deep learning of person re-identification research, combining with the actual application scenarios, the paper proposes a complete process for person re-identification with mobile devices. It aims to combine pedestrian detection and person re-identification to perform real-time pedestrian detection and query.

2. Related work

Person re-identification(ReID), originates from multi-camera tracking and is used to determine whether the pedestrians in different images or videos captured in non-overlapping views belong to the same person. The implementation process mainly includes two parts: pedestrian de-

* Corresponding: zhu.miaomiao234@mail.kyutech.jp

[†] Department of Electrical and Electronic Engineering, Kyushu Institute of Technology,
1-1 Sensui-cho, Tobata-ku, Kitakyushu, Fukuoka, Japan 804-8550

[‡] School of Computer Science and Engineering, Changshu Institute of Technology,
No. 99, 3rd South Ring Road, Changshu, Jiangsu, China 215500

Figure 1: Typical flow chart for person re-identification.

tection and identification. A typical flow chart is shown in Fig. 1.

2.1 Deep learning In summary, deep learning, as an implementation algorithm of machine learning, is a learning process in which deep neural networks are used for the feature representation. In other words, it can automatically generate a model by training a dataset containing a large amount of data instead of manually extracting features by machine learning. In short, it has a strong ability to express the features of the target and belongs to unsupervised learning. In 2006, as a representative in the field of machine learning, Geoffrey Hinton *et al.* [1] [2] introduced deep learning to the field of artificial intelligence, mainly for images, speeches, natural language processing, etc., all of which are difficult to extract features.

Subsequently, Convolutional Neural Networks (CNNs), as one of the representative algorithms of deep learning, has a strong ability of representation learning (the process of convolution is also the process of feature extraction). In 1998, Yann Lecun *et al.* [3] proposed LeNet-5 model and applied it to the recognition of handwritten characters. In 2012, in order to prove the potential of deep learning again, Geoffrey Hinton *et al.*, the author of [1] [2], participated in ImageNet image recognition competition for the first time. The CNN network AlexNet they built had won first place twice. The accuracy rate of the network is nearly 10% higher than that of the second place. The details of AlexNet can refer to [4]. Until then, due to its high accuracy, CNN had gradually attracted the attention of more and more researchers in the field of computer vision, such as image classification, object detection, image segmentation, image annotation, image generation, and so on.

To sum up, in the field of computer vision, deep learning is the means, CNN is the method and feature learning is the purpose. In other words, a good feature representation plays a key role in the accuracy of the final algorithm.

2.2 Object detection Object detection is to determine a specific target (object) from a static image or dynamic video sequence. While obtaining the category information, it also needs to be get its position and size, that is, Detection = Classification + Localization. Detection, classifica-

Figure 2: The major developments and milestones of deep learning based object detection techniques in recent years, especially after 2012.

tion and segmentation are the three core topics in the field of computer vision.

In recent years, with the research of deep learning, especially since the CNN achieved great success in classification tasks in 2012, due to its strong representation ability, deep learning has also become the best choice for image feature extraction in object detection. Figure 2 lists the major developments and milestones of deep learning based object detection techniques in recent years, especially after 2012.

Current state-of-the-art object detectors with deep learning can be mainly divided into two major categories [5]: one is two-stage detectors, such as R-CNN, a pioneering two-stage object detector proposed by Girshick *et al.* in 2014, and its variants Fast R-CNN, Faster R-CNN, and Mask R-CNN, etc., the other is one-stage detectors, such as Joseph Redmon *et al.* developing a real-time detector called YOLO (You Only Look Once) in 2016, and its variants YOLO9000, YOLOv3, YOLOv4 [6] and PP-YOLO [7], and so on.

It can be seen that since the CNN is used for object detection, the accuracy and speed of the object detection algorithm have been greatly improved in the short term and have been applied in many fields, including face detection in face recognition technology, pedestrian detection in person re-identification technology and vehicle detection in driverless technology and so on.

2.3 Person Re-identification As an instance-level recognition problem and a fundamental task in distributed multi-camera surveillance, person re-identification originates from pedestrian tracking and aims to match people appearing in different non-overlapping camera views. As we know, Zafed *et al.* [8] proposed the term “person re-identification” in 2005, which is the first research on multi-camera tracking. It can be observed from Fig. 1 ReID can be divided into image-based ReID and video-based ReID. The research process of image-based ReID is shown in Fig. 3. First, input a pedestrian image (Probe image), and then identify the pedestrian from the gallery (Gallery images). These images are usually obtained by manual labeling or detection algorithms from one or more surveillance networks. In summary, it can be summarized into four steps: preparing data, training, feature extraction, and evaluating results.

Figure 3: The research process of image-based ReID.

Table 1: Statistics of some commonly used datasets.

Dataset	Years	Cameras	Provider	Identities
VIPeR	2007	2	California	632
GRID	2009	8	QMUL	1,025
CUHK03	2014	10	CUHK	1,467
Market-1501	2015	6	Tsinghua	1,501
DukeMTMC-reID	2017	8	Duke	1,812
MSMT17	2018	15	Peking	4,101

Table 2: Basic description of common dataset

Dataset	Identities(T-Q-G)*	Images(T-Q-G)
CUHK03	767-700-700	7365-1400-5332
Market-1501	751-750-751	12936-3368-15913
DukeMTMC-reID	702-702-1110	16522-2228-17661
MSMT17	1041-3060-3060	30248-11659-82161

*(T:train, Q:query, G:gallery)

Figure 4: Number of papers on ReID that are published in top-tier computer vision conferences in recent years.

At present, a number of datasets for image-based ReID have been released, VIPeR[9], the first dataset used for ReID was published in 2007. CUHK03[10], the first dataset large enough for training deep learning models was released in 2014. Besides, the commonly used datasets including GRID[11], Market1501[12], DukeMTMC-reID[13] and MSMT17[14]. The basic description is shown in Table 1 and Table 2.

(a) Illumination (b) Obstruction (c) Viewpoint (d) Pose

Figure 5: Sample images in the DukeMTMC-reID dataset. Note the wide range of viewpoint, obstruction, pose and illumination changes.

As shown in Fig. 4, according to the statistics on the number of papers on ReID that are published in top-tier computer vision conferences(CVPR, ICCV/ECCV), it can be seen that ReID has attracted more and more attention and has become a topical research area in computer vision in recent years. For example, in 2020, there were 34 and 18 papers published on CVPR and ECCV respectively. Especially in 2014, Wei Li *et al.*[10] introduced CNN to ReID for the first time, using the FPNN network(Filter paring neural network) to train and getting True or False by comparing two patches. At the same time, ReID also faces several challenges. For example, compared with the sample images in the DukeMTMC-reID dataset illustrated in Fig. 5, we can see that due to the changes of camera viewing conditions, the intraclass (instance/identity) variations are typically big. For instance, some people in Fig. 5 carry a backpack, the view change across cameras(frontal to back) brings large appearance changes in the backpack area, which makes matching the same person difficult. At the same time, we must admit that the current data volume is still far from satisfactory due to the difficulties in collecting across-camera matched person images. The related researches have been done in previous papers[15].

2.4 Kivy First released in 2011, Kivy is an open source Python library for rapid development of applications that makes use of innovative user interfaces, such as multi-touch APPs. It only needs to write a set of codes and can run on major desktops and mobile platforms, including Linux, iOS, Android, Windows, and macOS. However, although Kivy is a cross-platform, if we want to run the python code on different platforms, we only need to use the packaging tool project-bulldozer for packaging python into an executable program corresponding to the platform.

3. Proposed Approach

Based on the analysis of the key technologies used above, including deep learning, object detection, and person re-identification, this section mainly introduces the overall design and partial implementation of person re-identification on mobile devices based on deep learning.

3.1 Overall design In this paper, applying the object detection algorithm and a novel lightweight deep ReID CNN, a person-identification system based on mobile devices is built by Kivy, an open source Python library. In

Figure 6: Person re-identification based on mobile devices.

other words, it aims to combine pedestrian detection with person re-identification to perform real-time pedestrian detection and query. The overall implementation process of this system is shown in Fig. 6. It can be clearly seen that the system mainly includes two parts, pedestrian detection, and person re-identification.

(a) Pedestrian detection

The system receives the pedestrian images from mobile devices through real-time pedestrian detection algorithm, crop these images into a uniform size to form the candidate pedestrian library (Gallery images).

(b) Person re-identification

Use appropriate ReID models to extract the pedestrian features of the candidate image library (Gallery images) and the existing image library to be queried (Probe images), and calculate the similarity between them by distance metric. Finally, rank the results, make the candidate images with a smaller distance (or higher similarity) as query results and output them in the terminal.

Compared with the image-based ReID as shown in Fig. 3, the main difference between image-based ReID and ReID based on mobile devices is that the former is mainly carried out through experimental verification and evaluation on large existing ReID datasets, such as CUHK03, Market-1501, DukeMTMC-reID, and MSMT17. In the process of deep learning, these datasets are divided into training sets (Train) and testing sets (Query, Gallery). Pedestrian images are usually obtained by manually labeling or detection algorithm from a lot of surveillance cameras with non-overlapping fields of view. Currently, pedestrian detection and identification are independent. However, the latter's testing sets, especially the candidate pedestrian library (Gallery images), are pedestrian images captured in real time by mobile devices. detection and identification are performed simultaneously.

3.2 Techniques and methods Based on the above analysis of system requirements and key technologies, the implementation technology proposed in this paper is shown

Figure 7: Complete process for ReID on mobile devices.

in Fig. 7. The whole system is mainly composed of the following four parts.

The first part is input. In the early stage of the system design, crop the images to be retrieved and queried to a uniform size to form an image library to be queried (Probe images) as the system input.

The second part is pedestrian detection. As described in the Section 2.2, currently there are many high performance object detection algorithms to choose. Pedestrian detection, as one of the multiple categories to be detected in the object detection task, aims to detect all pedestrians in a given image or video. After the comparison and analysis, YOLOv3[16], as one of the most popular real-time object detection algorithms, has high detection effects and speed. For example, it is 1000 times faster than R-CNN and 100 times faster than Fast R-CNN. Therefore, we use the YOLOv3 as the initial model that pre-trained on the COCO dataset, fine-tune based on the latest technology. As shown in Fig. 8, these cropped images form the candidate pedestrian library (Gallery images), the filename contain some information, such as time, location, and camera number, and so on.

The third part is person re-identification. As an instance-level recognition problem, person ReID relies on discriminative features, which not only capture different spatial scales but also encapsulate an arbitrary combination of multiple scales. After comparing with many works, we found that BNNeck [17] is the best performance acquired by global features in all person re-identification works. At the same time, none of the existing person re-identification models addresses Omni-scale feature learning, Kaiyang Zhou *et al.*[18] proposed a novel deep ReID CNN, termed Omni-Scale Network(OSNet), it can well capture the local discriminative features and also it is the key to overcome the

Figure 8: Form the candidate pedestrian library (Gallery images) through real-time pedestrian detection.

challenges of the ReID work. In the part of ReID model, it is mainly carried out by combining BNNeck and OSNet methods.

The last part is output. This is most difficult in this system, that is, how to integrate pedestrian detection and person re-identification via Kivy.

3.3 Evaluation In order to evaluate the ReID method, Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are two widely used measurements. Recently, Mang Ye *et al.*[19] introduced a new evaluation metric (mINP) for ReID, indicating the cost for finding all the correct matches, which provides an additional criteria to evaluate the ReID system for real applications. The final output results are mainly based on those three common evaluation. Take the gallery images sorted in the top K as the output result on the terminal of the mobile device and set warning information based on the ratio of correct results. The early warning information is set according to the ratio of the correct results. If the similarity of the detected pedestrian features after a distance measurement exceeds a certain threshold, an alarm will ring on the mobile device terminal. Regarding reproducibility, the mobile device used only needs to be a smart phone and has a certain amount of memory.

As shown in Fig. 9(a), some sample person re-identification results by the method of OSNet on Markket-1501. The images in the first column are the query images used to match with manually trimmed pedestrians in the second column existing dataset (gallery), while the retrieved images are sorted according to the similarity scores from left to right in the third column. The green font means correct match, the red font and the images in rectangles indicate the false match. However, as shown in Fig. 9(b), in the real-world scenarios, the goal of person re-identification on mobile devices is to find a target person in a gallery of images or videos which come from multiple scenarios. Comparison with traditional ReID research, our purposed approach is to search a person from the whole scene images or videos instead of matching them with manually cropped pedestrians in the existing dataset.

(a) Traditional ReID research, match with manually cropped pedestrians in the existing datasets.

(b) Person Re-identification on Mobile Devices, search a person from the whole scene images or videos

Figure 9: Comparison between traditional ReID research and ReID on Mobile Devices.

3.4 Application scenarios The purpose of this system design is to effectively combine pedestrian detection and person re-identification. Besides, make person re-identification technology more useful and practical, instead of just testing and verifying some existing datasets, such as tracking criminals online and searching for missing children. Further designed and improved system can be embedded into each user's mobile phone by the public security department. Once a person takes a photo with a specific pedestrian (for example, fugitives and missing children) into the lens occasionally in the usual camera process, the mobile terminal and the system background will give information prompts at the same time, including the specific location of the person involved. Because people acquire photos iffy, the privacy and infringement issues are not involved. The pictures can be collected early, and build a query library of images (Probe images) for subsequent person re-identification. Finally, complete the online query.

4. Discussion and Future Work

For the research of this topic, the preliminary work mainly involves requirements analysis and key technology study. As mentioned above, this design includes mostly three parts: pedestrian detection, person re-identification, and Kivy. At present, pedestrian detection and APP interface design have been completed. The next part includes the following aspects: (1) Construct a person re-recognition model and extract the features of pedestrians in all images (Probe image and Gallery images); (2) Calculate the similarity between Probe image and Gallery images by distance metric; (3) Find the best match and output the real-time result based on the calculation of similarity distance; (4) Kivy, designing a person re-identification app based on mobile terminal devices for practical scenarios, such as tracking criminals and searching for missing children.

5. Conclusions

Person re-identification is a useful supplement to face recognition. Because of its important security and surveillance applications, it has been drawing lots of attention from both academia and industry. It is not difficult to find that face recognition has played an important role in the criminal investigation. For example, in China, many criminals were arrested by face recognition in 2019, such as Yuwu Xie and Mei Li, et al. This paper proposes a complete process for person re-identification with mobile devices, which aims to combine pedestrian detection with person re-identification to perform real-time pedestrian detection and query. In this process, not only can the features extracted by pedestrians be reused, but the research on person re-identification can be applied effectively, such as tracking criminals and searching for missing children. It is worth pointing out that, benefiting from the development of deep learning in recent years, face recognition, pedestrian detection, and person re-identification will be applied to more scenarios, the effective combination of these techniques will become more powerful. It will raise the cognitive level of artificial intelligence to a new stage.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61972059, Grant 61702055, and Grant 61773272 and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191474 and Grant BK20191475.

References

- [1] G. E. Hinton, S. Osindero and Y. Teh, "A fast learning algorithm for deep belief nets", *Neural Computation*, Vol.18, No.7, pp.1527-1554, 2006. DOI: 10.1162/neco.2006.18.7.1527
- [2] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, Vol.313, No.5786, pp.504-507, 2006. DOI: 10.1126/science.1127647
- [3] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, Vol.86, No.11, pp.2278-2324, 1998. DOI: 10.1109/5.726791
- [4] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems 25 (NIPS)*, vol.1, pp.1097-1105, 2012.
- [5] X. Wu, D. Sahoo and S. C. H. Hoi, "Recent Advances in Deep Learning for Object Detection", *Neurocomputing*, Vol.396, pp.39-64, 2020. DOI: 10.1016/j.neucom.2020.01.085
- [6] A. Bochkovskiy, C. Wang and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection", *arXiv:2004.10934*, 2020.
- [7] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding and S. Wen, "PP-YOLO: An Effective and Efficient Implementation of Object Detector", *arXiv:2007.12099*, 2020.
- [8] W. Zajdel, Z. Zivkovic and B. J. A. Krose, "Keeping Track of Humans: Have I Seen This Person Before?", *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp.2081-2086, 2005, DOI: 10.1109/ROBOT.2005.1570420
- [9] D. Gray, S. Brennan and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking", *In Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol.3, No.5, 2007.
- [10] W. Li, R. Zhao, T. Xiao and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.152-159, 2014. DOI: 10.1109/CVPR.2014.27
- [11] C. C. Loy, T. Xiang and S. Gong, "Multi-camera activity correlation analysis", *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1988-1995, 2009. DOI: 10.1109/CVPR.2009.5206827
- [12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable person re-identification: A benchmark", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.1116-1124, 2015. DOI: 10.1109/ICCV.2015.133
- [13] Z. Zheng, L. Zheng and Y. Yang, "Unlabeled Samples Generated by Gan Improve the Person Re-identification Baseline in vitro", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.3774-3782, 2017. DOI: 10.1109/ICCV.2017.405
- [14] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.79-88, 2018. DOI: 10.1109/CVPR.2018.00016
- [15] M. Zhu, S. Gong, Z. Qian and L. Zhang, "A Brief Review on Cycle Generative Adversarial Networks", *In The 7th IIAE International Conference on Intelligent Systems and Image Processings (ICISIP)*, pp.235-242, 2019. DOI: 10.12792/icisip2019.046
- [16] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement", *arXiv:1804.02767*, 2018.
- [17] H. Luo, Y. Gu, X. Liao, S. Lai and W. Jiang, "Bag of Tricks and A Strong Baseline for Deep Person Re-identification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.1487-1495, 2019. DOI: 10.1109/CVPRW.2019.00190
- [18] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-Scale Feature Learning for Person Re-Identification", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.3701-3711, 2019. DOI: 10.1109/ICCV.2019.00380

- [19] M. Ye, J. Shen, G. lin, T. Xiang, L. Shao and S. C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook", *arXiv:2001.04193v1*, 2020.

Miaomiao Zhu (Member) was born in Shaanxi, China, 1981. She received a M.S. degree in engineering from Xi'an University of Architecture and Technology, china, in 2008. Since 2018, she has been a Ph.D. candidate in Electrical Engineering and Electronics from Kyushu Institute of Technology, Japan. Hers current research interests include Image Processing, Image Enhancement and Wireless sensor network.

Shengrong Gong (Non-member) received a Ph.D degree in engineering from Beihang University, in 2001, is a professor at Changshu Institute of Technology, China. His current research interests include Computer vision and Machine learning. He is a member of CCF and IEEE.

Zhenjiang Qian (Non-member) was born in Changshu, China, 1982. He received a Ph.D degree in engineering from Nanjing University, china, in 2013, is an associate professor at Changshu Institute of Technology, China. His current research interests include cyber-physical system and theorem proving. He is a member of CCF and IEEE.

Seiichi Serikawa (Member) was born in Kumamoto Japan on June, 1961. He received a B.S. degree and M.S. degree in electrical engineering from Kumamoto University in 1984 and 1986. He received a Ph.D. degree in electronic engineering from Kyushu Institute of Technology in 1994. Currently, he is a professor of Dept. of electrical engineering and electronics in Kyushu Institute of Technology, Japan. His current research interests include computer vision, sensors and robotics. He is a member of IEICE, IEEE, IPSJ, IIAE.

Lifeng Zhang (Member) received the B.S. degree in Electronic Engineering from Southeast University in 1994. He received M.S. and Ph.D. degrees in Electrical Engineering from Kyushu Institute of Technology, in 1999 and 2001, respectively. Currently, he is a professor at Kyushu Institute of Technology, Japan. His current research interests include computer vision, image processing, biometrics authentication and remote sensing system. He is a member of IIAE, IEEE, and IEICE.