Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE221122

Estimation CNN-Based Person Re-Identification Accuracy in Video Using Different Datasets

Shiping YE [a,b], Svetlana IHNATSYEVA [c,1], Rykhard BOHUSH [b,c,1] , Chaoxiang CHEN [a,b], Sergey ABLAMEYKO[b,d,e]
[a] *Zhejiang Shuren University, Hangzhou, China*
[b] *International Science and Technology Cooperation Base of Zhejiang Province: Remote Sensing Image Processing and Application, Hangzhou 310000, China*
[c] *Euphrosyne Polotskaya State University of Polotsk, Novopolotsk, Belarus*
[d] *Belarusian State University, Minsk, Belarus*
[e] *United Institute for Informatics Problems of NAS of Belarus, Minsk, Belarus*

Abstract. The paper analyses the problem of person re-identification accuracy in distributed video surveillance systems by using various datasets for training convolutional neural networks (CNN). After analysis we constructed large joint image dataset of people consisting of CUHK02, CUHK03, Market-1501, DukeMTMC-ReID, MSMT17 and our collected PolReID. PolReID includes 52035 images for 657 people Experimental results on assessing re-identification accuracy based on the main metrics Rank and mAP are presented. The research was carried out for the most widely used CNNs in re-identification, such as ResNet-50, DenseNet121 and PCB. We show that the constructed large dataset allowed us to improve Rank1, mAP for all test sets.

Keywords. ReID system, PolReID, joint dataset, cross domain, CNN

1. Introduction

One of the main problems in video control for people movement is how to identify the same person in another time and in another place. This problem is called person re-identification (person ReID). The ReID system already knows features of controlled person and compares them with features newly appeared persons in video from other cameras.

In many surveillance videos, due to the camera resolution and shooting angle, it is usually impossible to get a very high-quality image of the face. When facial recognition fails, ReID becomes a very important alternative technology.

Re-identification is not an easy task and has many problems. ReID has a very important function, which is to use multiple cameras. Video cameras can have dissimilar resolutions, shooting at different times - different degrees of illumination, different

1 Corresponding Authors: S.A. Ihnatsyeva, R. Bohush, Euphrosyne Polotskaya State University of Polotsk, 211440, Belarus, Novopolotsk, Blokhin st. 29.

camera positions will give different backgrounds, and this leads to the existence of such a problem as domain shift.

People appearance may change during movement, or different people may appear similar. There is also the problem of occlusion. At some points, a person part can be hidden by other people or landscape elements. During ReID process, it is necessary to analyse a big set of data and each of datasets is a separate domain.

ReID has been studied in academia for many years, but it hasn't made a huge breakthrough with the development of deep learning in recent years. One of the first attempts to solve ReID task was done by using by the random erasing method [1]. Random erasing is a method to increase dataset by adding images, in which an arbitrary image fragment is randomly deleted, which is filled with zero or random values. This method improves the algorithm's occlusions resistance.

Widely used recently deep neural networks allowed to reach a good success in the person re-identification problem. Especially when data for training and testing are independent and identically distributed. However, these methods still have problems in an invisible domain [2].

One of the approaches to increase the stability of the ReID system is to use a dataset that will have the maximum similarity with the data with which the re-identification algorithm will have to work. The correspondence between people on different frames is established based on the analysis of the spatial coordinates of faces and people, as well as their CNN features, using the Hungarian algorithm. However, facial features are not always available. To solve the task without facial features, it is necessary to significantly increase the training dataset, which would include a huge number of identifiers and their images.

In this paper, we move on this way and propose to increase the training dataset. To do it, we first analysed CNN re-identification for existed datasets: Market-1501, DukeMTMC-ReID and MSMT17. Three different CNN architectures were used in experiments: DenseNet-121, ResNet-50, PCB. Then, we used own PolReID dataset and new Joint Large Dataset. The performed experiments show that newly constructed dataset allows to improve all re-identification metrics.

2. The General Scheme of Person Re-Identification of a Person in Frames from Several Surveillance Cameras

It is possible to distinguish close-world re-identification systems using ready-made data sets for training and testing, and open-world systems, in which the image gallery is constantly updated with new frames [3]. Close-world systems are usually used for research purposes and the data set consists of a limited number of video sequences or images obtained from several surveillance cameras. The data in such sets is annotated and prepared in advance, the request is present in the gallery. In open-world systems, a data set is used that changes over time, as new recordings from surveillance cameras become available, restrictive frameworks need to be generated in real time, training data needs to be annotated. Such systems are the closest to real conditions.

In general, any re-identification system implies the presence of several CCTV cameras K (Figure 1). All video sequences received from cameras are fed to a detector, with the help of which bounding frames with images of people are extracted from individual frames, which are placed in the gallery.

Research datasets often contain already extracted restrictive frameworks, and in this case it is assumed that the process of detecting people and forming a gallery is done in advance. For each detected person, a descriptor is formed and placed in the feature table. Feature extraction in the case of a close-world research system is performed in advance, in the case of an open-world system – in real time. For each incoming request, a descriptor is also generated, and a search is performed in the feature table by ranking table according to established similarity criteria. The best matches are given as a re-identification result. Depending on the selected system type, re-identification result can be a ranked list of images that most match the request, or a video sequence on which the identified persons are marked.

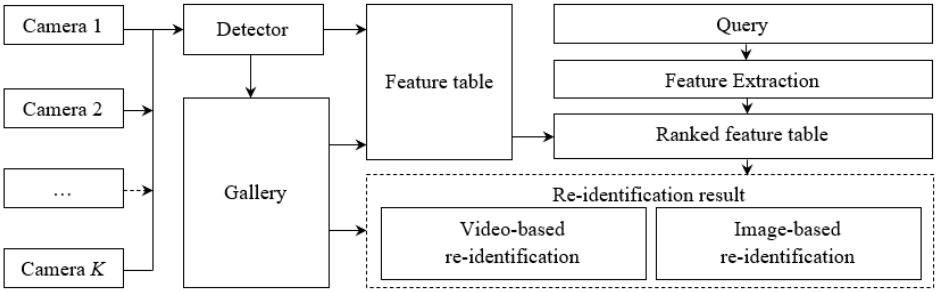

Figure 1. Person re-identification system general scheme

3. Datasets for Re-Identification

3.1. Existed Datasets

Existed data sets for re-identification differ in size, composition, and number of cameras used. Data sets can contain individual frames entirely, as in the PRW [4] and CUHK-SUSU [5] datasets, or rectangular fragments cut from these frames, the so-called bounding boxes containing only the image of a person. Such data sets may contain sets of bounding boxes obtained from several consecutive frames, which are called tracklets, as for example in the MARS [6], LPW [7] data sets, etc. Also, datasets may contain restrictive frames obtained from individual frames taken with a certain time interval, such as Market-1501 [8], CUHN01 [9], CUHN02[10], CUHN03 [11] VIPer [12], etc. Depending on the type of re-identification algorithm, the corresponding data set is used.

Images for data sets can be obtained under different conditions, shooting can be carried out outdoors (Market-1501 [8], LPW [7], PRID [13]) or indoors (QMUL iLIDS [14], Airport [15]) with a different number of cameras and the way they are located. For example, in the CUHN01 dataset, images for each person are obtained from two cameras whose viewing areas do not intersect. In CUHN02, five such pairs of cameras are used, and in CUHN03, images are obtained from six cameras, but for each person, restrictive frames are provided from only two cameras out of six. The VIPeR data set was formed based on images obtained from two outdoor surveillance cameras, and only one image from each camera is used for each person. Three different locations were used in the formation of the LPW, and three cameras were installed at the first location, and four

cameras were installed at the other two locations. The data sets PRW, Market1501 and MARS were obtained in the same place near a supermarket at Tsinghua University from the same six cameras and differ only in the way the data is presented: whole frames, bounding boxes with a human image, and tracklets.

The accuracy of the re-identification algorithm during training is influenced by the size, variety and quality of the training sample. The formation of a data set for training and testing is a time-consuming and expensive process from the point of view of remuneration. In addition, one should keep in mind such a problem as domain shift [16], when there is a significant decrease in the accuracy of re-identification when using the system in conditions stylistically different from the training sample. A partial solution to the problem may be to combine different data sets, which is considered in [17], including from the required domain [18].

The most common and large datasets are Market-1501, DukeMTMC-ReID, CUHK02, CUHK03, MSMT17.

Market-1501 has 32,668 images for 1501 IDs. 12936 bounding boxes for 751 people are used for training and 19732 images for 750 people for testing. There are also 2793 distractors in the dataset [8]. Duke MTMC-ReID dataset was generated using 8 surveillance cameras located on Duke University campus. It consists of 36411 bounding boxes. The dataset is divided into a training sample (16522 images for 702 person) and a test sample (17661 bounding rectangles for 702 identity) [19]. The CUHK02 dataset includes images for 1816 people from five pairs of cameras. From each pair of cameras, 4 images were obtained for 971, 306, 107, 193 and 239 people. The cameras were located on the campus of the Chinese University of Hong Kong (CUHK). When using this dataset, the authors ask for the privacy students CUHK [10]. At the same university, the CUHK03 dataset was generated, which included 5 bounding boxes from 2 angles for 1467 people. The use of this dataset is possible only by agreement with the authors for academic research [11]. When generating the MSMT17 dataset, 3 indoor and 12 outdoor surveillance cameras were used. Data collection was carried out at different times under different weather conditions in the morning, noon and afternoon for four days. There are 126441 images in MSMT17 for 4101 people [20].

In [18], when forming the joint dataset, the LPW dataset was used, consisting of 2731 people in 592438 images. For each identifier, there are several tracklets received from different CCTV cameras. Because tracklet is consecutive frames set, then images included in it have minor differences. This leads to a decrease in the diversity of the training sample, and an uneven distribution of data in the entire joint dataset.

3.2. Building New Large Joint Dataset

At the first step, we developed our own PolReID dataset and at the second step we combined new joint dataset consists of Market-1501, DukeMCMT-ReID, CUHK02, CUHK03, MSMT17 and PolReID.

For the construction of PolReID [21] dataset, video sequences received from volunteers were used. To form a set of images for each person, from 2 to 10 cameras located in different locations, and from 1 to 9 video sequences from each camera were used. Images were taken in more than 700 locations, outdoors under different weather conditions and seasons (summer, autumn and winter) and indoors, under natural and artificial lighting of different intensities. YOLOv4 was used to form the bounding boxes.

For each person in PolReID there are images with occlusions. Each person is presented from different angles. In total there are 52035 images for 657 people.

For training in PolReID, 32516 bounding boxes (398 IDs) are used, for testing - 19519 bounding boxes (259 ID). PolReID includes images for 440 men and 217 women; for 524 people aged 18 to 30, and 133 people over 30. Images for 340 people were obtained indoors, for 214 people – from external surveillance cameras, and 103 people were recorded by both internal and external surveillance cameras. 210 people have a mask on their face, 33 of them are recorded on some cameras without a mask. Filming was carried out in summer for 95 people, in winter – for 288, in spring and autumn – for 274. Examples of images are shown in Fig. 2.

Figure 2. Some images from PolReID dataset

At the second stage, we combine the existing datasets with PolReID. Thus, the combined new joint dataset consists of Market-1501, DukeMCMT-ReID, CUHK02, CUHK03, MSMT17 and PolReID. When merging data sets, their structure should be taken into account. Merging a dataset consisting of images of people at different points in time should not be merged with a dataset consisting of tracklets. This will lead to an uneven training sample, in which for some people there will be a large number of similar images. The ability of the CNN to identify the most distinguishing features of people may also be reduced.

In addition, the more different shooting conditions the used data set contains, the more stable the trained model will be to different test data. Thus, the Market-1501 and Duke-MTMC-ReID data sets were formed by outdoor surveillance cameras, while MSMT17 and PolReID were formed by both external and internal cameras. During the formation of MSMT17, video surveillance was carried out at different times of the day under various weather conditions. In PolReID, in addition to various shooting times and weather conditions, images obtained at different times of the year from a large number

of cameras are also presented. Combining such data sets into one allowed us to form a training sample that was not only large, but also quite diverse.

The training sample of the joint dataset consists of 115,956 images for 6174 persons. The CUHK02 and CUHK03 datasets are included in full, and Market-1501, DukeMTMC-ReID, MSMT17 and PolReID, according to the train and test split proposed in the source documents.

4. Re-ID Accuracy Results for Different Datasets and CNN

Three different CNN architectures were used for feature extraction, such as DenseNet-121, ResNet-50, PCB. The algorithm from paper [22] was used for re-identification in our experiments.

Training was carried out for 60 epochs at learning rate of 0.05 and batch size of 32. For different datasets and CNN architectures, during the learning process from 30 to 50 epochs, fluctuations are observed around loss function minimum. Therefore, to get as close as possible to the minimum of the function, learning rate decreases by a factor of 0.1 after 40 epochs. The positive effect for the model convergence is shown in Fig.3.

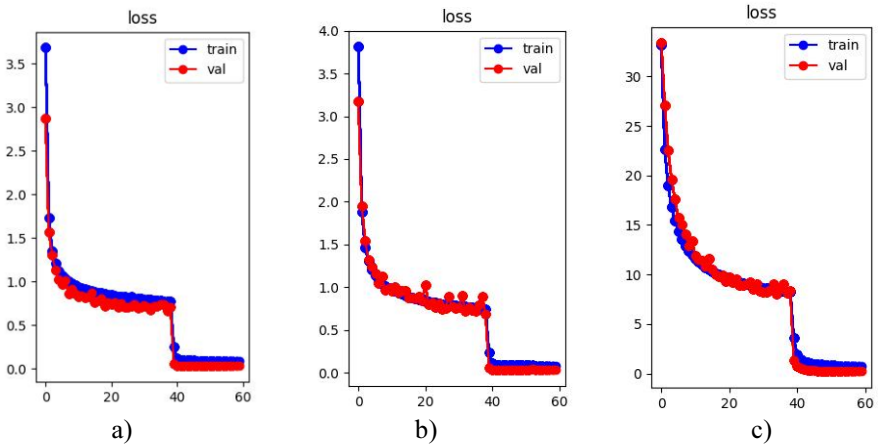

Figure 3. Loss function graph during training re-identification model. a) – loss function with backbone DenseNet-121 on joint training sample, b) - loss function with backbone ResNet-50 on joint training sample, c) loss function with backbone PCB on joint training sample.

Choosing training dataset for the re-identification accuracy is presented in Table 1. At the first stage experiments CNN was trained on one of the Market-1501, DukeMTMC-ReID and MSMT17 datasets. The tests were performed for different domains. The results show that re-identification accuracy decreases significantly for invisible domains. The maximal values of Rank1 and mAP metrics for cross-domain re-identification have been obtained by training on dataset MSMT17 and testing on PolReID, Rank1 = 86.38, mAP = 60.62. First, this can be explained by the fact that MSMT17 includes the largest number of people images, which were obtained under different environments. Secondly, during the PolReID collected, as well as for MSMT17, indoor and outdoor surveillance cameras, different lighting conditions, weather conditions, and times of day were used.

In the second stage experiments, a joint set including CUHK02, CUHK03, Market-1501, DukeMTMC-ReID , PolReID , MSMT17 was used for training. The training and test samples do not overlap. This approach allowed to increase re-identification accuracy for all metrics, maximum values were obtained for PolReID Rank1 = 95.41, mAP = 84.74.

PCB is most effective the source and target domains match, as well as for PolReID and Market1501 for cross-domain re-identification. DenseNet-121 is most effective for DukeMTMC-ReID, also for Market-1501 and DukeMTMC-ReID when trained on a jointed dataset.

Table 1. Experimental result

Dataset for test	**CNN**	**Dataset for train**							
		Market-1501		**DukeMTMC-ReID**		**MSMT17**		**Joint Dataset**	
		Metrics							
		Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
Market 1501	DN*	88.86	73.01	49.23	21.71	54.22	26.40	94.09	83.34
	RN**	83.33	71.16	43.88	18.68	48.49	22.81	92.12	80.62
	PCB	92.70	77.69	55.05	25.89	55.53	25.74	93.14	81.62
Duke MTMC-ReID	DN	37.21	20.18	81.51	64.81	55.61	34.51	86.45	74.00
	RN	30.57	15.86	79.04	62.40	50.76	30.84	84.20	71.19
	PCB	40.44	22.23	84.87	70.30	54.35	33.26	86.36	73.86
MSMT17	DN	12.72	03.92	19.84	5.94	70.53	40.99	76.73	51.13
	RN	9.24	2.68	15.04	4.32	65.71	36.56	72.05	45.64
	PCB	11.06	3.10	16.49	4.57	70.42	42.81	73.87	48.17
PolReID	DN	63.66	34.55	74.21	43.44	83.64	58.09	95.25	83.82
	RN	57.61	29.39	67.85	37.16	79.69	52.91	94.12	80.89
	PCB	62.61	35.31	72.20	40.80	86.38	60.62	95.41	84.74

DN* - DenseNet-121, RN - ResNet-50.

5. Conclusion

Currently, convolutional neural networks are used in human re-identification systems. When using CNN, the training sample is important. It is clear that the success of solving re-identification problem depends on a lot from data sets that will be used for network training. That's why we detail considered this task here.

We investigated the training dataset size and composition effect on the re-identification accuracy. We carried out a number of experiments with different size of dataset to solve re-identification task. We showed the main re-identification metrics for Market-1501, DukeMTMC-ReID and MSMT17 datasets and for ResNet-50, DenseNet121 and PCB CNN.

We considered the problem of forming a training sample for neural network re-identification algorithms and proposed a joint dataset consisting of CUHK02, CUHK03, Market-1501, DukeMTMC-ReID, MSMT17 and our collected PolReID. The built unified dataset includes 6174 identifiers and 115 956 images. The developed large dataset allowed us to improve all re-identification metrics for all tests.

Funding

This research was funded by Ministry of Science and Technology of the People's Republic of China (grant number G2021016001L, G2021016002L, G2021016028L, G2022016010L).

References

[1] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. Random Erasing Data Augmentation. AAAI, 2020.
[2] Bąk S, Carr P. One-Shot Metric Learning forPerson Re-identification. IEEE Conf. on Computer Visionand Pattern Recognition (CVPR); 2017: 1571-1580
[3] Ye, S., Bohush, R., Chen, C., Zakharova, I., Ablameyko, S. Person Tracking and Re-Identification in Video for Indoor Multi-Camera Surveillance Systems. Pattern Recognition and Image Analysis, 2020., Vol. 30, № 4 827–837.
[4] Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q. Person Re-identification in the Wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017: 3346-3355
[5] Xiao T, Li S, Wang B, Lin L, Wang X. Joint Detection and Identification Feature Learning for Person Search. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017: 3376-3385
[6] Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable Person Re-identification: A Benchmark. 2015 IEEE International Conference on Computer Vision (ICCV); 2015: 1116-1124.
[7] Song G, Leng B, Liu Y, Hetang C, Cai S. Region-based Quality Estimation Network for Large-scale Person Re-identification. AAAI. Source: https://arxiv.org/abs/1711.08766
[8] Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable Person Re-identification: A Benchmark. 2015 IEEE International Conference on Computer Vision (ICCV); 2015: 1116-1124.
[9] Li W, Zhao R, Wang X. Human Reidentification with Transferred Metric Learning. Proceedings of the 11th Asian conference on Computer Vision *(ACCV)*; 2012.
[10] Li W, Wang X. Locally Aligned Feature Transforms across Views. *2013 IEEE Conference on Computer Vision and Pattern Recognition*; 2013: 3594-3601.
[11] Li W, Zhao R, Xiao T, Wang X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*; 2014: 152-159.
[12] Gray D, Brennan S, Tao H. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance; 2007
[13] Hirzer M, Beleznai C, Roth PM, Bischof H. Person Re-identification by Descriptive and Discriminative Classification. *SCIA*. Lecture Notes in Computer Science; 2011: 91-102.
[14] Zheng W, Gong S, Xiang T. Associating Groups of People. *BMVC*; 2009. DOI: 10.5244/C.23
[15] Karanam S, Gou M, Wu Z, Rates-Borras A, Camps OI, Radke RJ. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence;* 2019: *41;* 523-536.
[16] Zhao F, Liao S, Xie G, Zhao J, Zhang K, Shao L. Unsupervised Domain Adaptation with Noise Resistible Mutual-Training for Person Re-identification. ECCV 2020. Lecture Notes in Computer Science; 2020: 526-544.
[17] Jin X, Lan C, Zeng W, Chen Z, Zhan L. Style Normalization and Restitution for Generalizable Person Re-Identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020: 3140-3149
[18] Ihnatsyeva S, Bohush R, Ablameyko S. Joint Dataset for CNN-based Person Re-identification. Pattern Recognition and Information Processing (PRIP'2021) Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus / United Institute of Informatics Problems of the National Academy of Sciences of Belarus. Minsk, 2021: 33–37.
[19] Ristani E, Solera F, Zou RS, Cucchiara R, Tomasi C. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. 2016. Source: https://arxiv.org/abs/*1609.01775*.
[20] Wei L, Zhang S, Gao W, Tian Q. Person Transfer GAN to Bridge Domain Gap for Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition;* 2018: 79-88
[21] PolReID, Source: https://github.com/SvetlanaIgn/PolReID
[22] Person reID baseline pytorch. Source: https://github.com/layumi/Person_reID_baseline_pytorch