Person Re-Identification Accuracy Improvement by Training a CNN with the New Large Joint Dataset and Re-Rank

Rykhard Bohush[1], Sviatlana Ihnatsyeva[1], Sergey Ablameyko[2,3]
[1] *Polotsk State University, Novopolotsk, Belarus*
[2] *Belarusian State University, Minsk, Belarus*
[3] *United Institute for Informatics Problems of NAS of Belarus, Minsk, Belarus*
rbohush@gmail.com

Abstract. The paper is aimed to improve person re-identification accuracy in distributed video surveillance systems based on constructing a large joint image dataset of people for training convolutional neural networks (CNN). For this aim, an analysis of existing datasets is provided. Then, a new large joint dataset for person re-identification task is constructed that includes the existing public datasets CUHK02, CUHK03, Market, Duke, MSMT17 and PolReID. Testing for re-identification is performed for such frequently cited CNNs as ResNet-50, DenseNet121 and PCB. Re-identification accuracy is evaluated by using the main metrics Rank, mAP and mINP. The use of the new large joint dataset makes it possible to improve Rank1 mAP, mINP on all test sets. Re-ranking is used to further increase the re-identification accuracy. Presented results confirm the effectiveness of the proposed approach.

Key words: convolution neural network, PolReID, re-identification, large-scale dataset, re-rank.

1. Introduction

Due to the widespread use of intelligent video surveillance systems, the task of re-identifying a person in a distributed camera system is an urgent task. In general, re-identification is the task of identifying the person you are looking for in a different place or time using distributed video surveillance systems in a city [34]. Such a system extracts features of the analyzed person and compares them with features of other persons in the existing dataset. Finding and highlighting effective distinguishing features manually is a long and time-consuming process, and for re-identification task, due to the ambiguity of appearance from several angles, lighting variations, different camera resolutions, occlusions, it would take an irrationally large amount of time [7]. That is why the re-identification task remained unresolved for a long time. The recent achievements in the field of deep learning, in particular, the development of convolutional neural networks (CNN) allowed to perform the re-identification process with sufficiently high accuracy and in a reasonable time.

Despite the successful application of CNN for this task, a large number of problems still have to be faced when developing a re-identification system. Feature extraction is a difficult process in ReID system, due to the ambiguity of people appearance from various angles, lighting variations, different camera resolutions, occlusions.

Machine GRAPHICS & VISION 31(1/4):93–109, 2022. DOI: 10.22630/MGV.2022.31.1.5 .

CNN-based ReID systems are highly dependent on the size and quality of the datasets used for training. It is obvious that the dataset must contain a large amount of various data, and the larger the dataset, the more accurate the re-identification result will be. It is also desirable to use a dataset that will have the maximum similarity to the data that the re-identification algorithm will have to work with.

Also, the data for training and testing should be independent and identically distributed. However, experiments show that such models are well suited for a training set and will perform poorly in an invisible domain [2].

This leads to the main problem: how to increase the training set without using additional data? One of the ways to do it is to add augmentation tools such as reflection vertically or horizontally, rotation, changes in brightness and contrast, color fluctuations and others to the existing dataset. The use of data augmentation enables to diversify the training set. With a small original dataset size, these transformations may also not be enough to obtain satisfactory re-identification results. Increasing training sample size and diversity can also be achieved by combining existing datasets.

So, the success of solving re-identification problem depends a lot from datasets that will be used for network training. That's why we detail consider this task here.

In this paper, we analyze the existing datasets and discuss a problem of increasing a data volume for person re-identification task. Approaches for increasing images number in datasets are described. Based on this analysis, we propose a new large joint dataset for person re-identification task. Our dataset includes the existing public datasets CUHK02, CUHK03, Market, Duke, MSMT17 and our collected PolReID. Testing for re-identification was performed for such famous CNNs as ResNet-50, DenseNet121 and PCB. Re-identification accuracy was evaluated by using the main metrics Rank, mAP and mINP. The use of the new large joint dataset allowed us to improve Rank1 mAP, mINP on all test sets.

2. Overview of datasets

It is known that if the training and test samples were obtained under the same video surveillance conditions, then the re-identification accuracy is higher than if these conditions are different. Such a problem is called domain shift [39], a partial solution of which can be the several datasets combination [12, 14, 22]. This is increasing the variety of training examples and allows to extract features that are more resistant to changes in the domain.

Various re-identification systems types use different datasets. For research systems, datasets are used that consist of individual person images – bounding boxes. The largest scale of them are CUHK01 [19], CUHK02 [18], CUHK03 [20], Market-1501 [41], DukeMTMC-ReID [25], MSMT17 [32]. For re-identification systems that use temporal information, sets include of people images obtained from several consecutive frames, Tab. 1. Comparative table of datasets for re-identification.

Dataset	Number of cameras	Number of persons	Bounding boxes
PRW	6	932	34 304
CUHK-SYSU	6	8 432	96 143
MARS	6	1 261	1 191 003
LPW	3,4,4	2 731	592 438
Market1501	6	1 501	32 217
CUHN01	2	971	3 884
CUHN02	10	1 816	7 264
CUHN03	6	1 360	13 164
MSMT17	15	4 101	126 441
VIPeR	2	632	1 264
PRID	2	934	24 541
QMUL iLIDS	2	119	476
Airport	6	9 651	39 902
CUHK-PEDES	–	13 003	80 412
ICFG-PEDES	–	4 102	52 522
LR-PRID	2	100	200
LR-VIPeR	2	632	1 264
SYSU-MM01	6	491	38 271
RegDB	2	412	8 240
PKU-Sketch	2	200	400

tracklets. Tracklets are contained in such datasets as LPW [27], MARS [40]. More complex re-identification systems assume that person must first be detected in the image and only then identified. For such systems datasets are used, consisting of frames from video surveillance cameras, for example PRW [42], CUHK-SYSU [35]. Comparison of datasets is shown in Table 1.

Datasets CUHK02 and CUHK03 were generated on the University campus of Hong Kong. Each image for each person is obtained from two cameras. CUHK02 had two cameras at five different locations on campus, while CUHK03 had cameras installed in pairs at three locations. CUHK02 includes 7264 images for 1816 persons, CUHK03 has 13 164 bounding boxes for 1360 IDs. Images are not divided into test and training sets, and it is proposed to perform testing for images of 100 randomly selected persons, and use the rest for training. A small number of video surveillance scenes, the images number for each person, the random nature of the test images choice in each experiment can be attributed to the shortcomings of these datasets, which were partially taken into account when forming the Market-1501 dataset. Frames from six video surveillance cameras were used to form it located near the supermarket at Tsinhua University. The dataset includes 32 668 images for 1501 people. In the test sample for 750 persons, there are 19 732 bounding boxes, including 2739 distractors, and the remaining images are assigned to the training sample. Distractors in Market-1501 include examples where the detector mistook some objects for believable person images. Additionally, upon request, 500 000 distractors can be obtained under the same conditions as the labeled images.

DukeMTMC-ReID is a subset of the Duke Multi-Target Multi-Camera (MTMC) dataset generated from video 85 minutes at 60 frames per second. Eight CCTV cameras were used, located on the Duke University campus. Duke MTMC is designed for multi-tracking and based on it is marked up for re-identification tasks. The algorithm is trained on 16 522 images of 702 persons and 17 661 of 702 other identities for testing.

Market-1501 and DukeMTMC-ReID are based on the frame received from outdoor security cameras, which limits the variety of lighting options to natural light only. This was taken into account when developing the MSMT17 dataset. For its formation, frames from twelve outdoor surveillance cameras and three cameras installed indoors were used. Video recording was carried out at different day times for four days. Person number in MSMT17 is 4101, for which there are 126 441 images. The training sample includes 32 621 images for 1041 people. The test sample includes 93 820 bounding boxes for 3060 persons.

For training and testing of heterogeneous re-identification systems special datasets are used, where text (CUHK-PEDES [17], ICFG-PEDES [5]), a low-resolution image (LR-PRID [21], LR-VIPeR [15]), an image from an infrared camera (SYSU-MM01 [33], RegDB [23] or drawing (PKU-Sketch [24]).

The CUHK-PEDES dataset [17] combines five existing datasets, such as CUHK03 [20], Market-1501 [41], dataset from [36], VIPeR [8] and CUHK01 [19], and each image is annotated with two text descriptions in English received from crowdsourcing employees, which can be used as a query. The text contains information about the appearance of a person, his actions, and poses. Each text description contains an average of 23.5 words. Another dataset for heterogeneous re-identification systems by text description is the ICFG-PEDES dataset [5], which contains an average of 37.2 words with a more detailed description of appearance than CUHK-PEDES, and is formed based on the MSMT17 dataset [32].

The datasets LR-PRID [21], LR-VIPeR [15] are formed on the basis of images from the datasets PRID [10] and VIPeR [8], respectively, and for each person there is a pair of images, one of which has a low resolution and the other high. These datasets are used for heterogeneous re-identification systems with images of different resolutions.

SYSU-MM01 [33] was obtained from two infrared and four RGB cameras. It contains 15 712 images from an infrared camera and 22 559 color images for 491 people. RegDB dataset [23] includes 10 color images taken during the day and 10 thermal images from a night IR camera for 412 people. Both sets are used in heterogeneous re-identification systems with images from infrared and RGB cameras.

In [24], a dataset is proposed, including two images from different cameras for two hundred people, and one sketch for a person. To create sketches, volunteers were involved, who described the people appearance to five artists, to train an open-world cross-modality re-identification system. Sketch drawn according to the description is used to search for a person whose photo is not in this system. Another dataset for open-world re-identification systems is the MPR Drone [16], which differs from traditional sets in that a flying drone camera takes people images. Since only one camera is used, the whole set consists of two parts. The first part is marked up for 113 610 detected bounding rectangles, and the second contains raw frames for the first part.

Paper [6] presents a large unlabeled LUPerson dataset for unsupervised training of re-identification systems, which was created using more than seventy thousand street videos from various cities, which includes more than four million images for two hundred thousand people.

3. Re-identification for different algorithms and datasets

As we know the accuracy of a person re-identification in a distributed video surveillance system is largely determined by the number and variety of the image database that is used in CNN training. An images set obtained under the same conditions in the same video surveillance system is called a domain. Such a set can be divided into two parts, the first one is used for training (source domain), and the second one (target) is used for testing. The target domain is invisible if it differs from the source. Each image in the dataset is affected by a combination of factors, including camera resolution, the same background, lighting conditions, and the person appearance.

Training on a single domain allows providing sufficiently high accuracy only for this system. However, when using such a trained CNN for another domain that is unknown (invisible) to the system, the re-identification accuracy will be significantly lower. Thus, if datasets obtained under different conditions were used for training and testing, then there is a problem of domain transfer (domain change) for real ReID systems. An increase in accuracy can be provided by the search for new methods and algorithms, as well as combining several datasets.

Domain adaptation approaches were considered in various papers. In [31] and [14] training results are shown for different datasets composition. Analysis of results shows that an increase in the training set has a positive effect on the re-identification accuracy. Thus, in [31], adding MSMT17 to the synthetic dataset used as a training sample allows increasing the re-identification accuracy by 12.7% in mAP for DukeMTMC-ReID. Data inclusion from target domain into training sample allows increasing Rank1 for MSMT17 by more than 45% in [31]. Similarly, in [14], the use of target domain images for training allows increasing mAP for Market-1501 from 33.9% to 82.3% and for DukeMTMC-ReID from 33.6% to 73.2%. Maximum accuracy in mAP among the considered algorithms was achieved using the IDM algorithm [3,4] for the Market-1501 and MSMT17 datasets, which became possible using the features of intermediate domains generated during the learning process.

The JVTC approach [38] and the UnrealPerson synthetic dataset as training proved to be the most effective when using DukeMTMC-ReID as the target domain. Table 2 shows an efficiency of cross-domain person re-identification for different datasets.

Tab. 2. Efficiency of person re-identification for different algorithms and datasets.

Algorithm	Year	Training dataset	Metrics	Testing Dataset		
			(%)	Market	Duke	MSMT17
Open-ReID [31]	2020	RandPerson [31]	mAP	28.8	27.1	6.3
			Rank1	55.6	47.6	20.1
		RandPerson [31] +MSMT17	mAP	35.8	39.8	36.8
			Rank1	62.3	61.0	65.0
SNR [14]	2020	Market	mAp	84.7	33.6	-
			Rank1	94.4	55.1	-
		Duke	mAP	33.9	72.9	-
			Rank1	66.7	84.4	-
		Market+ Duke+CUHK+ MTMC17	mAP	82.3	73.2	-
			Rank1	93.4	85.5	-
NRMT [39]	2020	Market	mAP	–	62.3	19.8
			Rank1	–	78.1	43.7
		Duke	mAP	72.2	–	20.6
			Rank1	88.0	–	45.2
CBN [38]	2021	UnrealPerson [38]	mAP	54.3	49.4	15.3
			Rank1	79.0	69.7	38.5
JVTC [38]	2021	UnrealPerson [38]	mAP	80.2	75.2	34.8
			Rank1	93.0	88.3	68.2
QAConv [30]	2022	ClonedPerson [30]	mAP	21.8	–	18.5
			Rank1	22.6	–	49.1
IDM [3]	2022	Market	mAp	–	73.2	40.2
			Rank1	–	85.5	69.9
		Duke	mAP	85.3	–	40.5
			Rank1	94.2	–	69.5
		MSMT17	mAP	85.2	73.6	-
			Rank1	94.1	84.6	-

4. New large joint ReID dataset

At the first stage, we significantly increased the PolReID dataset [12] from 5609 images for 54 people to 52 035 bounding boxes for 657 people [13]. The training set includes 397 identities (32 516 bounding boxes), test sample – 259 IDs (19519 images). Examples of images are shown in Fig. 1. Statistical information for PolReID composition is given in Tab. 3.

To form PolReID, video sequences were obtained, the total duration of which was 8 hours 1 minute 53 seconds. Cameras in different locations from several angles took each person images. Various camera number from two to nine carried out video surveillance. In total, 839 ways of placing cameras with different characteristics (resolution, frames number per second) were used to create the set. Video surveillance was carried out under dissimilar weather conditions for four seasons. The cameras were installed indoors with natural and artificial lighting, outdoors for all day. The presence of various attributes in people in the form of bags, packages, backpacks, briefcases, grocery baskets, scarves, hats,

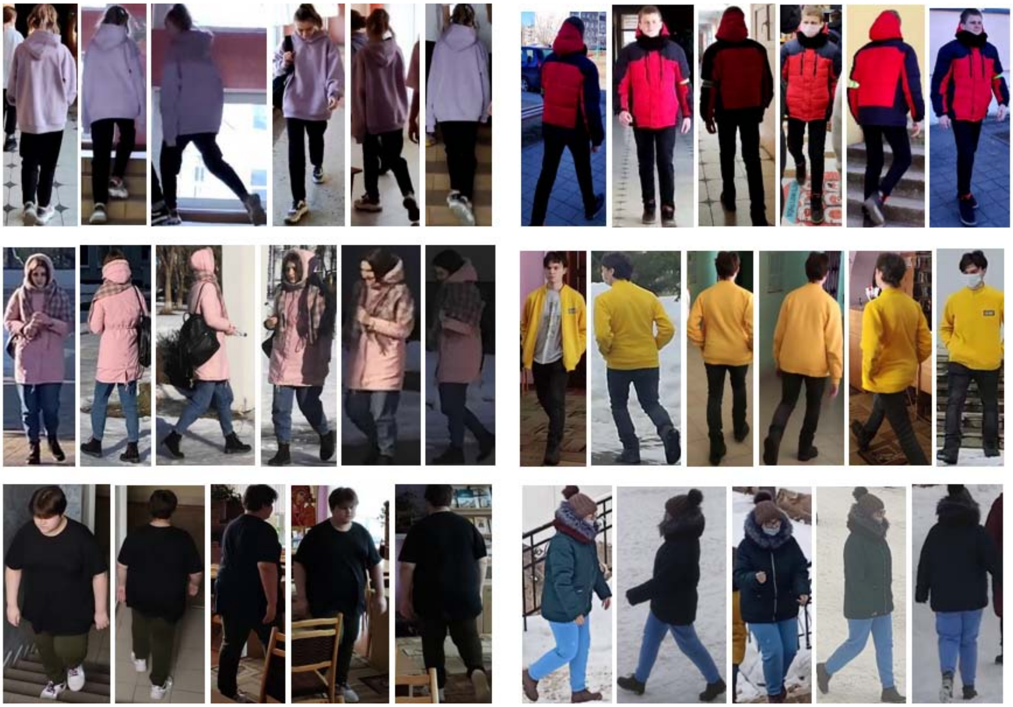

Fig. 1. Some images from PolReID dataset.

Tab. 3. PolReID composition.

Characteristic		Number of persons
Gender	Male	440
	Female	217
Age	18–30 years	524
	Over 30	133
Season	Summer	95
	Autumn and Spring	274
	Winter	288
Shooting conditions	Indoors	340
	Outdoors	214
	Outdoors and	103
	indoors	103
Mask availability	Masked	177
	Without mask	447
	Masked and unmasked	33

glasses, folders for papers, notebooks, phones, headphones, food and drinks changes over time, which makes it possible to study and take into account minor appearance changes. The person detection on frames and the formation of bounding boxes were performed using the YOLOv4 CNN [1,29].

At the second stage, to increase training sample size and diversity, several common datasets were combined, such as CUHK02, CUHK03, Market-1501, DukeMTMC-ReID, MSMT17 and PolReID.

When merging several datasets, one has to face such difficulties as various file names in sets, sundry names and location of directories. In this regard, when adding each new set to the training sample, a separate software implementation of the renaming process all files and placing them in the appropriate folders was developed. The training set includes images from such directories as "train" and "train_all" (Fig. 2). Directory "train_all" includes also instance validation images to determine the accuracy in the training process. Each file has the same name format, and a character "_" separates each value. For example, in Fig. 2 two images of a person are shown. According to the first file name 00007_c01s06_028546_04.jpg, we can say that this is a person whose ID is 00007, obtained from frame 28546 of the sixth video sequence from the first surveillance camera. In total, four different people were present in this frame. For the second file, whose name is 00007_c02s03_071052_01.jpg, we can say that this is an image of the same person, but it was taken from frame number 71 052 in the third sequence from the second surveillance camera. The total images number in the training set of the joint dataset was 115 956 bounding boxes for 6174 people.

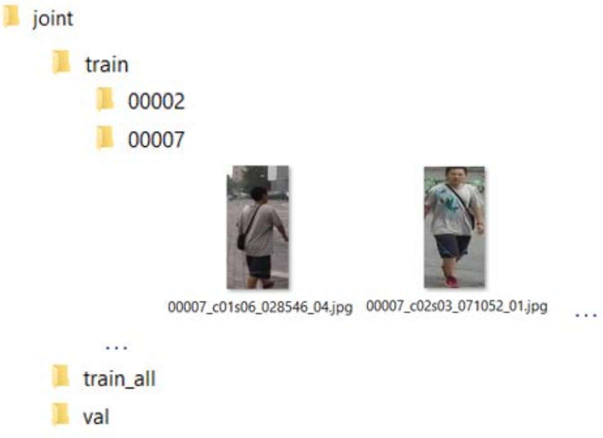

Fig. 2. The structure of the joint training dataset.

5. Metrics for re-identification

Choice of metrics is critical task for evaluating re-identification results. The most common group of metrics is RankN, which includes Rank1, Rank5, Rank10, and mAP.

RankN group characterizes ranking quality and shows the percentage of queries number for which the correct result was among the first N results. Accordingly, Rank1 metric shows the queries percentage for which the first candidate image ID matches the query ID. If $N = 5$, then Rank5 shows the queries percentage for which among the first five given candidate images there was a correct solution. For the first ten candidate images are considered. To calculate RankN, the number sum ratio queries for which the correct solution was found among the first returned results to the total queries number Q is determined:

$$\mathrm{RankN} = \frac{\sum K_{i,N}}{Q}\,, \tag{1}$$

where i – query number, $K_{i,N}$ – i-th query for which the correct solution was found among the first N returned results.

Metric mAP estimates the mean value of the average precision for all queries and is calculated with the formula:

$$\mathrm{mAP} = \frac{1}{Q}\sum_{i=1}^{Q} \mathrm{AP}_i\,, \tag{2}$$

where Q – the total number of queries, AP – average precision defined as the area under the precision-recall curve, where $\mathrm{pr} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ – precision, TP – number of true positive identifications or simply true positive, FP – false positive, rc $= \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ – recall, FN – false negative.

In re-identification systems, it is a priority that the correct predictions are at the beginning of ranked list and have as few false predictions as possible. It should be noted that RankN and mAP metrics do not reflect the difficulty of finding correctly identified person images for a request. With the same Rank metrics for different requests, the AP accuracy for them may differ. To take into account the search for the hardest match correct predictions, mINP (mean Inverse Negative Penalty) metric is proposed in [37]. Analysis of this metric makes it possible to eliminate the dominance of light matches that affect the Rank and mAP metrics. Additional metrics are introduced for calculation mINP: Negative Penalty (NP) assigned for incorrect predictions for the i-th query and reducing correct re-identification probability if the most difficult match is incorrectly found and Inverse Negative Penalty (INP) – the reciprocal for NP. Growth NP indicates an improvement in system performance. mINP characterizes the average INP value for all requests and is calculated as:

$$\mathrm{mINP} = \frac{1}{Q}\sum_i (1 - \mathrm{NP}_i) = \frac{1}{Q}\sum_i \left(1 - \frac{R_i^{\mathrm{hard}} - |G_i|}{R_i^{\mathrm{hard}}}\right) = \frac{1}{Q}\sum_i \frac{|G_i|}{R_i^{\mathrm{hard}}}, \tag{3}$$

where Q – total number of queries, $\mathrm{NP}_i = \frac{R_i^{\mathrm{hard}} - |G_i|}{R_i^{\mathrm{hard}}}$ – Negative Penalty, R_i^{hard} – position of hardest correct prediction, $|G_i|$ – total correct predictions number for the query.

INP enables to evaluate all correct matches finding complexity. The larger this value, the better the system is at finding all people with the same ID. Accordingly, one should strive to reduce NP and reduce the distance from ranking list beginning to position of the most difficult image to search for, which may be incorrectly identified.

6. Training and testing

To test re-identification task in our experiments, we used the algorithm from [43, 44]. Training on datasets was carried out for 60 epochs at learning rate of 0.05 and batch size of 32. During the learning process from 30 to 50 epochs fluctuations have been observed around in the Top1 error minimum. Top1 error rate indicates how many times the CNN has predicted the correct label with the highest probability. Therefore, to get as close as possible to the minimum of the function after epoch 40, the learning rate should be decreased by a factor of 0.1. Figure 3 shows Top1 error graphs during training the re-identification model.

The work [26] considered only the effect of reducing the learning rate on the value of the loss function. Experimental results for the re-identification accuracy for several datasets and CNN are presented in Table 4. Three CNN DenseNet-121 [11], ResNet-50 [9], PCB [28] have been used for feature extraction.

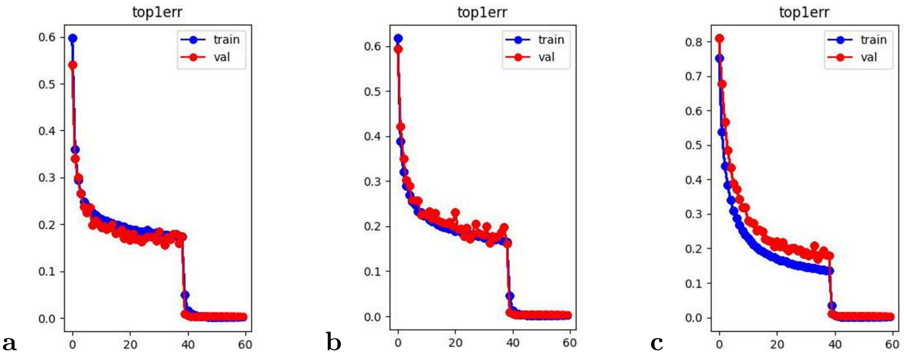

Fig. 3. Top1 error graphs of training and validation. (**a**) For DenseNet-121; (**b**) For ResNet-50; (**c**) For PCB.

Tab. 4. Experimental results for different datasets and CNN.

Dataset for train		Dataset for test								
		Market-1501			DukeMTMC-ReID			PolReID		
	Metrics	DenseNet	ResNet	PCB	DenseNet	ResNet	PCB	DenseNet	ResNet	PCB
Market -1501	Rank1(%)	88.9	83.3	92.7	37.2	30.6	40.4	63.7	57.6	62.6
	mAP(%)	73.0	71.2	77.7	20.2	15.9	22.2	34.6	29.4	35.3
	mINP	0.41	0.39	0.4	0.03	0.02	0.03	0.06	0.04	0.05
Duke MTMC -ReID	Rank1(%)	49.2	43.9	55.1	81.5	79.0	84.9	74.2	67.9	72.2
	mAP(%)	21.7	18.9	25.9	64.8	62.4	70.3	43.4	37.2	40.8
	mINP	0.03	0.02	0.03	0.27	0.25	0.30	0.08	0.06	0.07
MSMT 17	Rank1(%)	54.2	48.5	55.5	55.6	50.8	54.4	83.6	79.7	86.4
	mAP(%)	26.4	22.8	25.7	34.5	30.8	33.3	58.1	52.9	60.6
	mINP	0.05	0.04	0.04	0.06	0.06	0.06	0.13	0.11	0.14
PolReID	Rank1(%)	49.7	39.7	47.4	54.8	47.6	51.1	93.9	91.2	94.2
	mAP(%)	24.7	18.2	21.1	33.2	27.2	29.2	77.4	74.2	83.3
	mINP	0.05	0.03	0.03	0.06	0.04	0.03	0.27	0.25	0.34
Presented Dataset	Rank1(%)	**94.1**	92.1	93.1	**86.5**	84.2	86.4	95.3	94.1	**95.4**
	mAP(%)	**83.3**	80.6	81.6	**74.0**	71.2	73.9	83.8	80.9	**84.7**
	mINP	**0.57**	0.51	0.49	**0.39**	0.35	0.37	**0.35**	0.32	0.35

Experiments had two stages. At the first stage, unlike [26] CNN training is carried out not only on Market-1501, DukeMTMC-ReID, MSMT17, but also on PolReID dataset, and accuracy is assessed using three metrics as mAP, Rank1, and mINP. The tests were performed for different domains. The obtained results show that re-identification accuracy decreases significantly for invisible domains. The maximal values of Rank1, mAP and mINP metrics for cross-domain re-identification have been obtained by training on dataset MSMT17 and testing on PolReID, Rank1 = 86.38%, mAP = 60.62%, mINP = 0.142.

This can be explained by the fact that MSMT17 includes the largest number of people images, which were obtained under different conditions (Tab. 1). PolReID and MSMT17 include data obtained by indoor and outdoor surveillance cameras, by using various lighting and weather conditions, etc.

At the second stage, a large joint set that included CUHK02, CUHK03, Market-1501, DukeMTMC-ReID, PolReID, MSMT17 was used for training. The training and test samples do not overlap. This approach allowed increasing re-identification accuracy for all three metrics. Maximum values were obtained for PolReID: Rank1 = 95.41%, mAP = 84.74%, mINP = 0.345. In [12], the joined training set was also used, and despite the fact that its size was larger (8 690 IDs/537 109 images), the results in the current experiment turned out to be better for similar test sets. This is primarily due to the fact that in [12] the LPW dataset [27] consisting of tracklets was included in the training set. A large images number with similar features led to an uneven distribution of the training examples variety.

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

Another key difference from [26] is that re-ranking is performed to further improve the accuracy of re-identification [45]. The k-inverse feature vector is computed for the image. To calculate it, k-inverse nearest neighbors are used by the Jaccard distance:

$$d(p, g_i) = 1 - \frac{|R^*(p,k) \cap R^*(g_i,k)|}{|R^*(p,k) \cup R^*(g_i,k)|}, \tag{4}$$

where $R^*(p,k)$ and $R^*(g_i,k)$ – k-reciprocal nearest neighbors, p – query, g – gallery image.

Therefore, it is of interest to study this approach for the used CNN and datasets. However, PCB conducts uniform partition on the conv-layer for learning part-level features. It does not explicitly partition the images.

After passing through the network, the feature vector enters the classification layer. During testing, pieces are concatenated to form the final descriptor of the input image.

Tab. 5. Experimental results for person re-identification with re-rank.

Dataset for train		Dataset for test					
		Market-1501		DukeMTMC-ReID		PolReID	
	Metrics	DenseNet	ResNet	DenseNet	ResNet	DenseNet	ResNet
Market-1501	Rank1(%)	91.75	90.83	44.30	36.49	68.65	60.27
	mAP(%)	86.52	84.92	33.57	25.80	46.09	39.53
	mINP	0.709	0.675	0.094	0.064	0.125	0.092
DukeMTMC-ReID	Rank1(%)	53.15	47.68	85.55	83.03	77.92	72.20
	mAP(%)	32.95	27.46	80.82	78.11	56.41	48.65
	mINP	0.098	0.068	0.567	0.526	0.155	0.116
MSMT17	Rank1(%)	58.58	52.02	61.89	57.99	87.27	83.24
	mAP(%)	39.23	33.18	51.06	46.33	71.31	65.96
	mINP	0.146	0.112	0.188	0.159	0.250	0.211
PolReID	Rank1(%)	53.95	43.71	61.98	54.26	94.44	91.86
	mAP(%)	35.25	25.80	49.44	41.83	85.23	82.42
	mINP	0.115	0.066	0.162	0.121	0.403	0.390
Presented Dataset	Rank1(%)	**94.80**	94.12	**88.73**	87.34	**96.45**	95.57
	mAP(%)	**92.01**	90.41	**86.03**	84.04	**90.85**	88.77
	mINP	**0.833**	0.788	**0.663**	0.613	**0.495**	0.457

This makes the CNN sensitive to the content of each part and leads to significant errors when searching for the k-reciprocal nearest neighbors to the query image.

The increase in mINP indicates that image number at rank list top has increased. The growth of this metric indicates the positive impact of this approach on re-identification (Tabs. 4 and 5).

Table 5 shows re-rank tests results only for the DenseNet-121 and ResNet-50 re-rank. The most efficient CNN using re-rank based on test results is DenseNet-121. Training on the joint dataset increased the accuracy scores for PolReID Rank1 = 96.45%, mAP = 90.85%, mINP = 0.495.

In [14] for the jointed set including Market-1501, DukeMTMC-ReID, CUHK and MSMT17 research was performed: for Market-1501 Rank1 = 93.4%, mAP = 82.3%, for DukeMTMC-ReID metrics Rank1 = 85.5, mAP = 73.2.

Our approach for Market-1501 dataset increases to Rank1 = 94.80%, mAP = 92.01% and for DukeMTMC-ReID to Rank1 = 88.73%, mAP = 86.03% (Tab. 5). Thus, the proposed approach makes it possible to obtain improved accuracy for people re-identification.

7. Conclusion

Convolutional neural networks are now widely used for video person re-identification task. However, to reach a good result, the networks must be appropriately trained. The success of solving re-identification task depends a lot on datasets that are used for training. That is why this task was analyzed in detail in our present paper.

We analyzed various existed datasets and their size and composition. A number of experiments with different size of datasets have been performed. Then, we considered a problem of forming a large dataset and propose a large joint dataset for person re-identification. This dataset includes the existed datasets CUHK02, CUHK03, Market, Duke, MSMT17 and our collected PolReID. The obtained unified dataset includes 6174 identifiers and 115 956 images.

The proposed new large dataset allows to improve re-identification metrics on all test sets. For further increasing the re-identification accuracy, we used re-rankings. The most efficient CNN using re-rank after training by presented dataset based on test results is DenseNet-121. We achieve for Market-1501 dataset 94.80% on Rank1, 92.01% on mAP, 0.833 on mINP and for DukeMTMC-ReID 88.73% on Rank1, 86.03% on mAP, 0.833 on mINP = 0.663 after re-rank. Training on the joint dataset and re-rank increased the accuracy scores for PolReID to Rank1 = 96.45%, mAP = 90.85%, mINP = 0.495.

References

[1] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao. YOLOv4: Optimal speed and accuracy of object detection. arXiv, 2020. arXiv:2004.10934. doi:10.48550/arXiv.2004.10934.

[2] S. Bąk and P. Carr. One-shot metric learning for person re-identification. In *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2017)*, pages 1571–1580, Honolulu, HI, USA, 21-26 Jul 2017. doi:10.1109/CVPR.2017.171.

[3] Y. Dai, J. Liu, Y. Sun, et al. IDM: An intermediate domain module for domain adaptive person re-ID. In *Proc. 2021 IEEE/CVF Conf. Computer Vision (ICCV 2021)*, pages 11844–11854, Montreal, QC, Canada, 10-17 Oct. doi:10.1109/ICCV48922.2021.01165.

[4] Y. Dai, Y. Sun, J. Liu, et al. Bridging the source-to-target gap for cross-domain person re-identification with intermediate domains. ArXiv, 2022. arXiv:2203.01682v1. doi:10.48550/arXiv.2203.01682.

[5] Z. Ding, C. Ding, Z. Shao, and D. Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv, 2021. arXiv:2107.12666v2. doi:10.48550/arXiv.2107.12666.

[6] D. Fu, D. Chen, J. Bao, et al. Unsupervised pre-training for person re-identification. In *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2021)*, pages 14745–14754, Nashville, TN, USA, 20-25 Jun 2021. doi:10.1109/CVPR46437.2021.01451.

[7] S. Gong and T. Xiang. Person re-identification. In *Visual Analysis of Behaviour: From Pixels to Semantics*, pages 301–313, London, 2011. Springer. doi:10.1007/978-0-85729-670-2_14.

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. 10th IEEE Int. Workshop on Performance Evaluation of Tracking and*

Surveillance (PETS 2007), Sep 2007. https://www.researchgate.net/publication/228345677_Evaluating_appearance_models_for_recognition_reacquisition_and_tracking.

[9] K. He, X. Zhang, Sh. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. doi:10.1109/cvpr.2016.90.

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conf. Image Analysis (SCIA 2011)*, volume 6688 of *Lecture Notes in Computer Science*, pages 91–102, Ystad, Sweden, 23-25 May 2011. doi:10.1007/978-3-642-21227-7_9.

[11] G. Huang, Zh. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi:10.1109/CVPR.2017.243.

[12] S. Ihnatsyeva, R. Bohush, and Ablameyko. Joint dataset for CNN-based person re-identification. In *Proc. 15th Int. Conf. Pattern Recognition and Information Processing (PRIP 2021)*, pages 33–37, Minsk, Belarus, 21-24 Sep 2021. United Institute of Informatics Problems, NAS Belarus, Minsk. https://elib.psu.by/handle/123456789/28586.

[13] SvetlanaIgn (S. Ihnatsyeva). PolReID. GitHub, Sep 2022. https://github.com/SvetlanaIgn/PolReID. [Accessed 1 Dec 2022].

[14] X. Jin, C. Lan, W. Zeng, et al. Style normalization and restitution for generalizable person re-identification. In *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2020)*, pages 3140–3149, Seattle, WA, USA, 13-19 Jun 2020. doi:10.1109/cvpr42600.2020.00321.

[15] X. Jing, X. Zhu, F. Wu, et al. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. *IEEE Transactions on Image Processing*, 26(3):1363–1378, 2015. doi:10.1109/TIP.2017.2651364.

[16] R. Layne, T. M. Hospedales, and S. Gong. Investigating open-world person re-identification using a drone. In Agapito L. et al., editors, *Computer Vision – Proc. European Conf. Computer Vision Workshops (ECCVW 2014)*, volume 8927, Part III of *Lecture Notes in Computer Science*, pages 225–240, Zurich, Switzerland, 6-7 Sep 2014. doi:10.1007/978-3-319-16199-0_16.

[17] S. Li, T. Xiao, H. Li, et al. Person search with natural language description. In *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2017)*, pages 5187–5196, Honolulu, HI, USA, 21-26 Jul 2017. doi:10.1109/CVPR.2017.551.

[18] W. Li and X. Wang. Locally aligned feature transforms across views. In *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2013)*, pages 3594–3601, Portland, OR, USA, 23-28 Jun 2013. doi:10.1109/CVPR.2013.461.

[19] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In K. M. Lee et al., editors, *Computer Vision – Proc. 11th Asian Conf. Computer Vision (ACCV 2012)*, volume 7724 of *Lecture Notes in Computer Science*, pages 31–44, Daejeon, Republic of Korea, 5-9 Nov 2012. doi:10.1007/978-3-030-58555-6_14.

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2014)*, pages 152–159, Columbus, OH, USA, 23-28 Jun 2014. doi:10.1109/CVPR.2014.27.

[21] X. Li, W. Zheng, X. Wang, et al. Multi-scale learning for low-resolution person re-identification. In *Proc. 2015 IEEE Int. Conf. Computer Vision (ICCV 2015)*, pages 3765–3773, Santiago, Chile, 7-13 Dec 2015. doi:10.1109/ICCV.2015.429.

[22] C. Luo, C. Song, and Z. Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In A. Vedaldi et al., editors, *Computer Vision – Proc. European*

Conf. Computer Vision (ECCV 2020), volume 12360 of *Lecture Notes in Computer Science*, pages 224–241, Glasgow, United Kingdom, 23-28 Aug 2020. doi:10.1007/978-3-030-58555-6_14.

[23] T. D. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. doi:10.3390/s17030605.

[24] L. Pang, Y. Wang, Y. Song, et al. Cross-domain adversarial feature learning for sketch re-identification. In *Proc. 26th ACM Int. Conf. Multimedia (MM '18)*, pages 609–617, Seoul, Republic of Korea, 22-26 Oct 2018. doi:10.1145/3240508.3240606.

[25] E. Ristani, F. Solera, R. S. Zou, et al. Performance measures and a data set for multi-target, multi-camera tracking. In G. Hua et al., editors, *Computer Vision – Proc. European Conf. Computer Vision Workshops (ECCVW 2020)*, volume 9914 of *Lecture Notes in Computer Science*, pages 17–35, Amsterdam, The Netherlands, 8-16 Oct 2016. doi:10.1007/978-3-319-48881-3_2.

[26] Y. Shiping, S. Ihnatsyeva, R. Bohush, C. Chen, and S. Ablameyko. Estimation CNN-based person re-identification accuracy in video using different datasets. In C.-H. Chen et al., editors, *Applied Mathematics, Modeling and Computer Simulation*, volume 30 of *Advances in Transdisciplinary Engineering*, pages 978–985. IOS Press, 2022. doi:10.3233/ATDE221122.

[27] G. Song, B. Leng, Y. Liu, et al. Region-based quality estimation network for large-scale person re-identification. *Proc. AAAI Conf. Artificial Intelligence*, 32(1):7347–7354, 2018. doi:10.1609/aaai.v32i1.12305.

[28] Y. Sun, L. Zheng, Y. Yang, et al. Beyond part models: Person retrieval with refined part pooling. In V. Ferrari et al., editors, *Computer Vision – Proc. European Conf. Computer Vision (ECCV 2017)*, volume 11208, Part IV of *Lecture Notes in Computer Science*, pages 501–518, Munich, Germany, 8-14 Sep 2018. doi:10.1007/978-3-030-01225-0_30.

[29] `Tianxiaomo`. Pytorch-YOLOv4. GitHub, 2020. `https://github.com/Tianxiaomo/pytorch-YOLOv4`. [Accessed 1 Dec 2022].

[30] Y. Wang, X. Liang, and S. Liao. Cloning outfits from real-world images to 3D characters for generalizable person re-identification. In *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2022)*, pages 4890–4899, New Orleans, LA, USA, 18-24 Jun 2022. doi:10.1109/CVPR52688.2022.00485.

[31] Y. Wang, S. Liao, and L. Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proc. 28th ACM Int. Conf. Multimedia (MM '20)*, Seattle, WA, USA, 12-16 Oct 2020. doi:10.1145/3394171.3413815.

[32] L. Wei, S. Zhang, W. Gao, and Tian Q. Person transfer GAN to bridge domain gap for person re-identification. In *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2018)*, pages 79–88, Salt Lake City, UT, USA, 18-23 Jun 2018. doi:10.1109/CVPR.2018.00016.

[33] A. Wu, W. Zheng, H. Yu, et al. RGB-infrared cross-modality person re-identification. In *Proc. 2017 IEEE Int. Conf. Computer Vision (ICCV 2017)*, pages 5390–5399, Venice, Italy, 22-29 Oct 2017. doi:10.1109/ICCV.2017.575.

[34] D. Wu, S.-J. Zheng, X.-P. Zhang, et al. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019. doi:10.1016/j.neucom.2019.01.079.

[35] T. Xiao, S. Li, B. Wang, et al. Joint detection and identification feature learning for person search. In *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2017)*, pages 3376–3385, Honolulu, HI, USA, 21-26 Jul 2017. doi:10.1109/CVPR.2017.360.

[36] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. Xiaogang Wang: personal web page, 2016. `http://www.ee.cuhk.edu.hk/~xgwang/PS/paper.pdf`.

[37] M. Ye, J. Shen, G. Lin, et al. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2020. doi:10.1109/TPAMI.2021.3054775.

[38] T. Zhang, L. Xie, L. Wei, et al. UnrealPerson: An adaptive pipeline towards costless person re-identification. In *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2021)*, pages 11501–11510, Nashville, TN, USA, 20-25 Jun 2021. doi:10.1109/CVPR46437.2021.01134.

[39] F. Zhao, S. Liao, G. Xie, et al. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In A. Vedaldi et al., editors, *Computer Vision – Proc. European Conf. Computer Vision (ECCV 2020)*, volume 12356 of *Lecture Notes in Computer Science*, pages 526–544, Glasgow, United Kingdom, 23-28 Aug 2020. doi:10.1007/978-3-030-58621-8_31.

[40] L. Zheng, Z. Bie, Y. Sun, et al. MARS: A video benchmark for large-scale person re-identification. In B. Leibe et al., editors, *Computer Vision – Proc. European Conf. Computer Vision (ECCV 2016)*, volume 9910 of *Lecture Notes in Computer Science*, pages 868–884, Amsterdam, The Netherlands, 11-14 Oct 2016. doi:10.1007/978-3-319-46466-4_52.

[41] L. Zheng, L. Shen, L. Tian, et al. Scalable person re-identification: A benchmark. In *Proc. 2015 IEEE Int. Conf. Computer Vision (ICCV 2015)*, pages 1116–1124, Santiago, Chile, 7-13 Dec 2015. doi:10.1109/ICCV.2015.133.

[42] L. Zheng, H. Zhang, S. Sun, et al. Person re-identification in the wild. In *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2017)*, pages 3346–3355, Honolulu, HI, USA, 21-26 Jul 2017. doi:10.1109/CVPR.2017.357.

[43] Z. Zheng, X. Yang, Z. Yu, et al. Joint discriminative and generative learning for person re-identification. In *Proc. 2019 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2019)*, pages 2133–2142, Long Beach, CA, USA, 15-20 Jun 2019. doi:10.1109/CVPR.2019.00224.

[44] `layumi` (Z. Zheng). Person_reID_baseline_pytorch. GitHub. `https://github.com/layumi/Person_reID_baseline_pytorch`. [Accessed 1 Dec 2022].

[45] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2017)*, pages 3652–3661, Honolulu, HI, USA, 21-26 Jul 2017. doi:10.1109/CVPR.2017.389.