UDC 004.932

ANALYSIS PERSON RE-IDENTIFICATION METHODS

S. IHNATSYEVA, R. BOHUSH
Polotsk State University, Belarus

Currently, deep neural networks are used to solve a large number of applied problems, among which the re-identification problem can be distinguished. Person re-identification is a search for interest person in frames obtained from several non-overlapping video cameras. The growing computing power of computers and the growing demand for intelligent video surveillance make this task urgent.

Introductio. The widespread use of video surveillance systems leads to the need to automate the process of detection and re-identification. Re-identification implies that when processing input data, each person is assigned a unique identifier, and when this person meets on frames received from another camera, or from the same one, but after a period of time, the system must recognize him. On such shots, it is not always possible to recognize a person by face, due to different viewing angles, which leads to the need to take into account other distinctive features, for example, such as height, physique, clothing, hairstyle, etc.

For research purposes, re-identification systems of a closed-world are used. This means that homogeneous, sufficiently annotated data is used, and the image-query of the searched person exists in the gallery of the dataset. In real situations, open-world systems are used, when the input data can be heterogeneous, the bounding boxes with the person image must be generated in real time, the data often does not contain annotations, and the dataset is open and can change for time.

Re-identification task is accompanied by such difficulties as: low quality images, different viewing angles, overlaps, background heaps, illumination, etc. Systems that assume use in real conditions are additionally complicated by the fact that the dataset in which it is necessary to search has significant sizes and may constantly change, data does not have annotations, weather conditions may change, the way people move, change in appearance (hats, outerwear, glasses, bags, and other items).

Recent developments in person re-identification systems are aimed at finding solutions to existing problems, and this article conducts modern approaches research and analysis.

1. The main re-identification systems problems

In the first case, there are several people number, and you need to establish a correspondence between these people and the images in the gallery. In the second case, there is a request, and you need to establish whether this person is found in the database gallery, and, if so, find him. In general, the re-identification task can be described as establishing a correspondence between a query (or queryes) and people's images in gallery from database. Features are extracted from the query-image, these features are processed using various methods and algorithms, and the system produces a person retrieval result by comparing the extracted features with objects features located in gallery. The general re-identification algorithm scheme is shown in Figure 1.

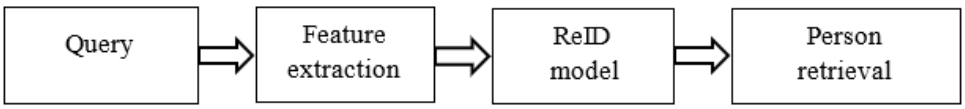

Fig. 1. – The re-identification algorithm general scheme

To extract features on the image in modern re-identification systems, a pre-trained neural network is used, which is re-trained on the dataset with which the re-identification system is to work. This approach allows you to efficiently extract a large features number. To improve training and extracted features subsequent interpretation, various re-identification models are used, with different approaches and algorithms aimed at solving re-identification problems. The main problems are the following:

1.Camera angle of view. Depending on the position in relation to the camera, the object will look completely different. Images taken from the side, from above, from behind, in front, at an angle will have completely different features for the same person, it is very difficult to distinguish something in common. In different frames, the face may be absent, the visible attributes of clothing and objects in the hands may change. For example, a backpack on the back: from the back it looks like a rectangle that obscures person part, from the side and top - a rectangle next to the person, in front - two stripes separating the arms from the body. A person easily uses a backpack as an additional feature, while this can complicate the task for an automated system.

2. **Pose's variability**. A person can take on a very large number of different positions in space. He can sit, stand, lie, raise his arms up, spread his legs, bend forward, to the side, and many other positions. And it can be extremely difficult for the system to find a common feature between a person who sits on a bench and, for example, stands on his hands.

3. **Lighting**. The color perception is highly dependent on the lighting degree at different day times, under different weather conditions, in the artificial lighting presence. A person intuitively feels the difference, while the computer vision system requires additional algorithms to solve this problem.

4. **Resolution**. Different CCTV cameras may differ in the received frames quality. Old video surveillance systems may be with resulting images low quality, which doesn't allow taking into consider small details and objects on a person's clothing, may give blur, which does not allow to clearly define the boundaries between a person and the background.

5. **Background clutters**. Re-identification task is often accompanied by a complex background presence, and the bounding boxes containing people images may contain separate other people bodies parts caught in the frame, a colorful background, objects that merge in color with clothes, hair. It also complicates the task of feature extraction.

6. **Occlusion**. The people in the video are moving and can often be obscured by other people or landscape elements. Different person parts can be hidden at different times. For example, if cameras are installed on the street, then the lower part can be hidden by benches, cars, curbs, the middle part can be blocked by a fence, the upper part - by road signs or signpost. It is impossible to predict at what point in time and how exactly a person will be partially hidden behind other objects, which makes it impossible to consider the signs of only a certain part.

7. **Appearance**. People can be dressed the same, or very similar, for different reasons. This can be a specialized uniform, such as a school uniform, or the same uniform for consultants in a shop. This greatly complicates the re-identification task, and does not allow considering only clothes as features. An extremely difficult problem is the problem of changing a person's appearance. When movement, a person can take off or put on a jacket, hat, hood. Even if the process is on camera, it is difficult for the system to determine that this is the same person, since the extracted features will have great a distance.

8. **Spatial-temporal distribution**. Person's movement trajectory, like speed, is extremely unpredictable. A person can walk slowly, then quickly, stop, turn around and go in the opposite direction, get on a bicycle or scooter and ride, run, and then sit on a bench and sit. This also affects the accuracy of re-identification.

9. **Heterogeneous data**. Sometimes, re-identification task comes down to finding a person, when the query is not an image, but a video sequence, a verbal description, a drawing, a face photograph or a frame obtained from a night vision camera, an infrared camera, while the database contains full-length images. Such a system should be use special means to solve such a problem.

2. Feature extraction

To person re-identification from an query-image in gallery, it is necessary to highlight the person characteristic features in the image, and deep convolutional neural networks are used to extract features. Networks such as ResNet-50, DenseNet-121, PCB, etc. are often used as the backbone network.

ResNet (Residual network) is a deep convolutional neural network used for feature extraction. A characteristic particularity is the use of residual blocks, that is, data from the convolutional layer output is transferred to the next layer input, and at the same time is transmitted through several layers, after which the data is combined using a shortcut connection. ResNet has modifications that differ in the convolutional layers number and can have a depth of 18, 34, 50, 101 or 152 layers. Increasing the network depth usually improves the accuracy of the algorithm, but leads to a decrease in the speed of feature extraction. Most often, researchers choose an average depth of 50 layers, thereby providing a balance between speed and accuracy.

DenseNet (Densely Connected Convolutional Network) - A neural network consisting of dense blocks, in which each convolutional layer is connected to each next layer in block, that is, each convolutional block layer receives features from all previous layers as input, and in turn , transfers them to all subsequent layers. Feature maps after each layer are concatenated. DenseNet can be 121, 169, 201 or 264 layers deep. Most commonly used DenseNet-121.

PCB (Part-based Convolution baseline) is an add-on that can be applied to any basic convolutional network without fully connected layers for classification. The idea is that an image is fed to the neural network input. At the convolutional level, this image is divided into parts, and for each part, the features are extracted and analyzed separately, which allows you to extract more complete information about the human body parts. [1]

Some researchers, when developing a re-identification method, test different convolutional networks as a backbone network. For example, in article [2], the authors conduct tests of their algorithm using ResNet-50, DenseNet-121 and PCB as the backbone network, and note that the mean average precision metric (mAP) is 72.2%, 76.9% and 82.8%, respectively, with the same other parameters for Duke-MTMC-ReID dataset. In [3], a research was also carried out on the backbone network choice influence on the algorithm accuracy, and experiments showed for ResNet-50 mAP = 32.58%, and for DenseNet-161 mAP = 42.25% for the proposed algorithm. These and other similar experiments show that the choice of the underlying network has an influence on the productivity and accuracy of the re-identification algorithm. Metrics such as mAP, CMC, rank1, rank5 and rank10 are used to evaluate re-identification algorithms.

3. ReID models

There is no one-size-fits-all system that can effectively address and address all re-identification problems. Some problems are partially solved using dataset augmentation, which is used to re-train the backbones neural network, and some using special algorithms.

The simplest and low cost way to increase system resistance is to augment the dataset. The more variable the training dataset, the more reliable the trained system will be. The simplest ways to enlarge a dataset are to rotate, flip horizontally or vertically, resize, change the brightness, contrast of the image, and other similar manipulations. It is used by most re-identification systems and can increase the system's resistance to changes in the person position relative to a CCTV camera, to a difference in the quality of the images obtained and to a difference in illumination.

To increase resistance to occlusions, the "random erasing" method is used - an arbitrary fragment is randomly removed from the image and filled with null or random values. The application of this method gave good value of the increase in mAP, for example, for Market-1501 using ResNet50 as the basic network, the mAP increased from 65.60 to 71.31 [4]. Due to its good performance and ease of productivity, the method has become widespread.

The occlusion problem is considered by researchers in [5], and the authors propose to consider the signs of individual parts of the body. First, key points are determined with using CNN, then local features are extracted for each point, i.e. a person is considered not as a whole, but in parts. Each key point is considered as a node in the graph, and the relationship between the nodes is determined. The final step is matching the topological information with local features and predicting the similarity between the two images. Consideration of each body part separately also allows to improve the system's resistance to pose variations taken by a person.

In [6], the image is also considered in parts, however, strict horizontal separation (PCB) [1] is used and the Pose-guided Visible Part Matching (PVPM) method is proposed. The algorithm consists of two main components: pose-guided attention (PGA) and pose-guided visibility predictor (PVP). PVP - evaluates the corresponding parts characteristics in positive image pairs and generates pseudo-labels for the invisible parts. And the PGA module is used to obtain differencial features for each image part.

Attention schemes are widely used in re-identification algorithms to improve feature learning. For example, in [7], the RGA (Relation aware global attention) module is proposed, which combines the local and global attention methods and allows compactly covering structural information for the image as a whole and local information about a person's appearance. In [8], a video-based re-identification method is proposed, which also uses attention scheme. The neural network extracts features, and the attention module is used to compute a weighted sum between several consecutive frames. Схемы внимания довольно широко применяются в алгоритмах повторной идентификации для улучшения изучения признаков.

An effective solution to increasing the dataset and improving the model learning ability is to use the Generative Adversarial network (GAN). A neural network is used to generate syntetic images based on existing ones [9]. There are many different uses for the GAN. For example, in [10], a re-identification algorithm is proposed, which implies application in real scenarios, which are complicated by various factors that degrade the image quality. And to train the network to extract reliable features, GAN is used to generate images with varying degrees and degradation types. This allows the system to be more resistant to changes in weather conditions, lighting and other factors that degrade image quality. Researchers in [11] propose using GANs to generate person images assuming different pose variations.

In [12], instead of generating images to augment the dataset, it is proposed to share the re-identification and data generation learning process throughout the learning process. The algorithm consists of two modules, generative (for generating images) and discriminative (for training). Two image comparisons are introduced: self-identification (i.e. generated by an image with a similar appearance of a person) and cross-identification (i.e., an image is generated, i.e. based on photographs of people with different appearance) and generated images are entered online in training, where features are extracted and analyzed in detail. This allows for more efficient use of the generated data.

In [13], a variant of solving the difference in illumination problem is proposed, and the use of a synthetic set 100 virtual people consisting, presented with different illumination degrees is proposed.

Some researchers suggest using auxiliary information to improve the re-identification quality. In [2], to improve the re-identification accuracy, a two-stream system is proposed, which, in addition to visual features, also uses spatial-temporal information. That is, the surveillance cameras location, the distance between them is taken into account, and the time it takes to get the frames is estimated. Visual features are extracted using CNN, and spatio-temporal information is provided with each image in the form of data about the camera number and the frame time. After receiving metrics two types, the metrics are combined and analyzed, after which a decision is made. Since a person takes some time to overcome the distance, it is possible to exclude unlikely options, even if there is visual similarity.

In [14], a two-stage re-identification learning algorithm is proposed - intra-camera and inter-camera similarity calculations. First, similarity metrics are computed based on the extracted features for the images taken from each camera separately. Different cameras create different pseudo-labels. At the second stage, each sample is considered as a new feature vector, which eliminates the mismatch in the distribution between cameras and creates more reliable pseudo-labels.

Table 1 shows the mAP / rank1 values for some state-of-the-art re-identification algorithms for different datasets.

Table 1. – Comparison results of several state-of-the-art algorithms for person re-identification

Alhorithm	mAP / rank1			
	Market1501	DukeMTMC-ReID	CUHK03	MSMT17
stReID [2]	86.7 / 97.2	82.8 / 94	-	-
PTL [3]	87.34 / -	79.16 / -	-	-
RGA[7]	74.5 / 79.6	-	77.4 / 81.1	57.5 / 80.3
DG-Net[12]	83.0 / 94.8	74.8 / 86.6	61.1 / 65.6	52.3 / 77.2
IICS [14]	73.9 / 90.1	66.2 / 80.8	-	31.9 / 62.6

As can be seen from Table 1, the algorithm accuracy is influenced not only by the selected algorithm, but also by the dataset that was used for training and testing.

For research purposes, when developing algorithms for person re-identification, prepared marked and annotated data sets are most often used. To test re-identification algorithms, datasets such as Market-1501, Duke MTMC-ReID, MSMT17, CUHK03 are often used - containing bounding boxes with persons images, where each person can be represented from different angles, obtained from different video cameras. The dataset can also be represented by a video sequence, for example, datasets PRID-2011, MARS.

Market-1501 dataset for person re-identification was collected in front of a supermarket at Tsinhua University. Images were taken from 5 high-resolution video cameras and one low-resolution video camera. The dataset includes 32,668 handcrafted bounding boxes for 1501 individuals. The training uses 12,936 bounding boxes for 751 identity, and 19,732 images in the gallery for 750 people to test the re-identification algorithm.

The Duke MTMC-ReID dataset is an extension of the Duke MTMC dataset acquired in March 2014 from the Duke University campus. The images were taken from 8 cameras located between the buildings, and include 36,411 bounding boxes. 16,522 images for 702 identity are used to train the re-identification algorithm. The remaining 17,661 bounding boxes for 702 people are used for testing.

MSMT17 is a large re-identification dataset from 12 outdoor and 3 indoor video cameras located on campus. The survey was conducted on different days and at different times of the day, and includes images for 4 days with different weather conditions, in the morning, noon, and afternoon. Total dataset contains 126,441 bounding boxes for 4,101 people.

CUHK03 - the dataset consists of 13,164 images for 1,360 pedestrians from 6 video surveilance cameras at Chinese University Hong Kong. During the formation of this dataset, part of the data was handcraft annotated, while the other part was generated using an automatic detector.

Data sets such as PRID-2011, MARS can be used to test person re-identification algorithms based on video sequence analysis:

MARS (Motion Analysis and Re-identification set) is the Market-1501 dataset extension, and is also obtained on the Tsinhua University campus from six CCTV cameras. The dataset includes 20,715 tracklets for 1261 person, and 3,248 distractor tracklets. Tracklets are sequences of images for each of the pedestrians captured by at least 2 different cameras. There are 1,191,003 images in total, of which 509,914 bounding boxes for 625 identity are used for training, and 681,089 bounding boxes for 636 people are used for testing (12,180 tracklets).

The PRID (Person ReID) dataset was created at the Austrian Institute of Technology. 2 video cameras were used to form the dataset. Cameras was located on different sides of the building. Images from different cameras have different degrees of illumination and viewing angles. The dataset consist 475 tracklets for 385 person captured by the first camera and 856 tracklets with 749 person's images captured by the second camera. The first 200 tracklets contain pedestrians who are in the field of view of both cameras, the rest are recorded either only by the first or only by the second camera.

Conclusion. The investigation showed that in order to develop an effective re-identification system, attention should be paid to the choice of the image database, the neural network used and the re-identification model. The more diverse and large the dataset, the more robust the system being developed will be. The accuracy and speed

of re-identification is also influenced by the backbone neural network choice. However, the key moment that determines the re-identification system itself is the choice of methods and approaches that improve the study of the objects features. Modern systems use various combinations of such approaches, multi-stage ReID models, auxiliary information. It seems promising to use GAN, attention schemes, and additional data, such as spatial-temporal information, which is provided with each frame. The use of various methods for expanding the dataset can increase the model's resistance to occlusions, changes in lighting, changes in the person pose in the frame. Further research will help to select the most promising methods for developing an effective system for person re-identification.

REFERENCES

1. Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, Shengjin Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)", in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480-496.
2. Wang, G., Lai, J., Huang, P., Xie, X, "Spatial-Temporal Person Re-identification", 2019. *URL:* https://arxiv.org/pdf/1812.03282v1.pdf.
3. Yu, Zhengxu [et al.], "Progressive Transfer Learning for Person Re-identification", 2019. *URL:* https://arxiv.org/pdf/1908.02492v1.pdf.
4. Zhong, Zhun, L. Zheng, Guoliang Kang, Shaozi Li and Y. Yang. "Random Erasing Data Augmentation.", 2020. *URL:* https://arxiv.org/pdf/1708.04896.pdf.
5. Guan'an Wang [et al.], "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6449-6458.
6. Shang Gao, Jingya Wang, Huchuan Lu, Zimo Liu, "Pose-Guided Visible Part Matching for Occluded Person ReID" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11744-11752.
7. Zhizheng Zhang, [et al.], "Relation-Aware Global Attention for Person Re-Identification", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3186-3195.
8. Pathak, P., Amir Erfan Eshratifar and M. Gormish. "Video Person Re-ID: Fantastic Techniques and Where to Find Them.", 2020. *URL:* https://arxiv.org/pdf/1912.05295v1.pdf.
9. Zheng, Zhedong [et al.], "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro.", *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3774-3782.
10. Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, Liang Li, "Real-World Person Re-Identification via Degradation Invariance Learning" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14084-14094.
11. Qian, X. [et al.], "Pose-Normalized Image Generation for Person Re-identification", 2018. *URL:* https://arxiv.org/pdf/1712.02225v4.pdf.
12. Zheng, Zhedong [et al.], "Joint Discriminative and Generative Learning for Person Re-Identification.", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2133-2142.
13. Slawomir Bak, Peter Carr, Jean-Francois Lalonde, "Domain Adaptation through Synthesis for Unsupervised Person Re-identification" Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 189-205.
14. Xuan, Shiyu and Shiliang Zhang. "Intra-Inter Camera Similarity for Unsupervised Person Re-Identification.", *URL:* https://arxiv.org/pdf/2103.11658.pdf.