Person re-identification using compound descriptor and invisible region replacement

Sviatlana Ihnatsyeva
Faculty of Information Technology
Euphrosyne Polotskaya State University of Polotsk
Novopolotsk, Belarus
ignateva604@gmail.com

Rykhard Bohush
Faculty of Information Technology
Euphrosyne Polotskaya State University of Polotsk
Novopolotsk, Belarus
r.bogush@psu.by

***Abstract*—In this paper we proposed person re-identification algorithm using compound descriptor that includes global and local features for the top, middle, and bottom of the person figure. Local areas are formed based on the person figure key points coordinates. If there are not enough visible points, the area is recognized as invisible and feature vector corresponding component is replaced by an average value for the k-nearest neighbors. Testing was performed on datasets for re-identification Market-1501, DukeMTMC-ReID, MSMT17, PolReID1077. Our algorithm allows us to increase accuracy re-identification for metric Rank1 by 8 - 51% and for metric mAP by 28 - 97% relative to the baseline.**

Keywords— convolution neuron networks, PolReID1077, occlusion

I. INTRODUCTION

To person re-identification by their appearance by convolutional neural networks (CNN) a feature vector is formed that characterizes the image as a whole. Global descriptor includes features that are distinctive for a person and characteristics of the background, illumination level, camera resolution, and other interfering factors. Quite often occlusions occur when another object covers a human figure part. This object features mix with the person feature, and worsen re-identification accuracy.

Various approaches and methods are used to solve occlusion problem. For example, in [1, 2], to increase CNN resistance to occlusions augmentation is used. This is increasing the network generalizing ability due to regularization effect achieved by adding occlusions to training data. In [3] for re-identification, local features are used that characterize the surroundings near the human figure key points. In algorithm viewed an image local feature as a graph node and proposed an adaptive direction graph convolutional layer to pass relation information between nodes. The proposed layer can automatically suppress the message passing of meaningless features by dynamically learning direction and degree of linkage. In [4] when searching for matches in the gallery for the same person on other images, the visible part that is not on the request act as an interfering factor, since there are no signs of this area in the request. It is proposed to remove the "extra" part and compare only visible areas. A semantic-driven model is introduced that first learns to extract features in different areas and output semantic probability maps of the visibility of different body parts. Probability maps are then combined with global object to extract local features of visible parts for images. In [5] the image is divided into 6 horizontal sections and the key points are defined. If the image fragment does not contain key points, then it is considered invisible and is not considered. In [6] it's proposed to estimate the human pose, and it is noted that many algorithms use information about the pose both in training and in testing, which leads to inference complexity. To achieve high accuracy while preserving low inference complexity, it's proposed a network named Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD), where the pose information is exploited to regularize the learning of semantics aligned features but is discarded in testing. In paper [7] Feature Completion Transformer (FCFormer) to implicitly complement the semantic information of occluded parts in the feature space is presented. Specifically, Occlusion Instance Augmentation (OIA) is proposed to simulate real and diverse occlusion situations on the holistic image. These augmented images not only enrich the amount of occlusion samples in the training set, but also form pairs with the holistic images. To obtain rich occlusion samples an Occlusion Instances Library (OIL) is building that contains 17 classes of occlusion samples obtained from the COCO [8] and Occluded-duke [9] training sets.

In the majority of considered algorithms, the signs hidden parts are not taken into account, while some important distinguishing feature can be blocked. Therefore, we propose to use a compound descriptor for a human body and replace the features of the missing image fragment with the averaged value for the k-nearest neighbors.

II. PERSON RE-IDENTIFICATION ALGORITHM

In order to solve the occlusions problem during re-identification an algorithm is proposed based on a compound descriptor including the feature vector $f_{global}\ (I^{(Y)})$ for the image $I^{(Y)}$, where Y is person identifier, and three local descriptors for top $f_{p1}(I^{(Y)})$, bottom $f_{p2}(I^{(Y)})$ and middle $f_{p3}(I^{(Y)})$ parts of the human silhouette. In this case the CNN input receives image packets and M_p mask coordinates that select local areas. If there are two persons in the image, then the interest object is the one whose bounding box area is larger. To select local areas and determine the overlaps presence, CNN is used, which determines a bounding box coordinate that describes the person boundaries silhouette and selects 17 key points for nose, left and right eyes, ears, shoulders, elbows, wrists, hips, knees and ankles. A point is considered visible if $T_{point\ i} \geq 0.5$, where $T_{point\ i}$ is the network confidence degree, i is the serial number of one of the 17 points. Local areas coordinates are determined based on the coordinates of key points and their visibility.

To determine the local area p_1 visibility, four of the five conditions must be met: at least one key point on the face is visible, at least one shoulder, elbow, wrist or hip is visible. The region p_2 is assumed to be occlusions free if at least one hip-predicting point is discernible. Visibility p_3 is determined by two of three the conditions fulfillment: $T_{point\ i} \geq 0.5$ for at least one the hips, knees or ankles predicted point.

For each image $I^{(Y)}$ binary masks are formed that highlight visible local areas. Each image entering the CNN for re-identification can be described as $(I^{(Y)}, I_{p1}{}^{(Y)}, I_{p2}{}^{(Y)}, I_{p3}{}^{(Y)})$, where:

$$I_{p1}{}^{(Y)} = (I^{(Y)} \circ M_{p1}), \quad (1)$$

$$I_{p2}^{(Y)} = (I^{(Y)} \circ M_{p2}), \quad (2)$$

$$I_{p3}^{(Y)} = (I^{(Y)} \circ M_{p3}), \quad (3)$$

$M_p \in \{0,1\}^{W \times H}$ is a binary mask that defines the fragment location to be discarded and replaced with a reduced image from the package; W and H are the image width and height; $\circ$ – element-wise multiplication.

For each person image a compound vector consists of four components:

$$f_{gen}(I) = \{f_{global}(I^{(Y)}), f_{p1}(I_{p1}^{(Y)}), f_{p2}(I_{p2}^{(Y)}), f_{p3}(I_{p3}^{(Y)})\}. \quad (4)$$

Descriptor corresponding component is considered invalid and equal to zero if one or more local areas are considered to be occluded by other objects. For all n gallery images, a feature vectors table $T_f = \{f_{gen}(I_1^{(Y)}), f_{gen}(I_2^{(Y)}), \ldots, f_{gen}(I_n^{(Y)})\}$ formed. Table T_f is ranked by the cosine distance metric with the query feature vector. For all images that have one or more local regions an overlap, the invalid feature vector components are replaced by the corresponding local descriptor $f_p^k(I_p^{(Y)})$ average value for k-nearest neighbors. For each real component $f_{gen}(I^{(Y)})$, k-nearest ones are found, from which k_1 best ones are selected, those that are closest to most of the real feature vector components.

When training model, images are transmitted to the input of the CNN in batch:

$$R_p = \begin{Bmatrix} I_1^{(Y_1)}, I_{p1_1}^{(Y_1)}, I_{p2_1}^{(Y_1)}, I_{p3_1}^{(Y_1)} \\ I_2^{(Y_2)}, I_{p1_2}^{(Y_2)}, I_{p2_2}^{(Y_2)}, I_{p3_2}^{(Y_2)} \\ \ldots, \\ I_B^{(Y_1)}, I_{p1_B}^{(Y_1)}, I_{p2_B}^{(Y_1)}, I_{p3_B}^{(Y_1)} \end{Bmatrix}, \quad (5)$$

where B – batch size.

Loss function L_{R_p} value is calculated for each batch:

$$L_{R\_p} = \lambda_{global} \cdot S_{global} \cdot E((I^{(Y)})^{out}) +$$

$$+ \lambda_{p1} \cdot S_{p1} \cdot E((I_{p1\_i}^{(Y)})^{out}) +. \quad (6)$$

$$+ \lambda_{p2} \cdot S_{p2} \cdot E((I_{p2\_i}^{(Y)})^{out}) + \lambda_{p3} \cdot S_{p3} \cdot E((I_{p3\_i}^{(Y)})^{out}),$$

where λ_{global}, λ_{p1}, λ_{p2}, λ_{p3} – coefficients that determine influence degree each fragment on the loss function; S is the square of the fragment; $(I^{(Y)})^{out} = f_{global}(I^{(Y)})$ is the features vector for entire image; Y_i is its identifier, $(I_p^{(Y)})^{out} = f_p(I_p^{(Y)})$ is the local features vector; E– cross-entropy loss function.

To increase the diversity of the training sample, the augmentation method [1] we used, which employ color exclusion and pixel shift vertically and horizontally, replacing the fragment with a reduced copy of another image from the batch. The first two transformations are applied to randomly selected images and the last one to some batches.

If augmentation was used when batch forming and the image has replaced fragments $({}^{aug}I^{(Y)})$, then:

$$L_{R\_p\_aug} = \lambda_{global} \cdot S_{global} \cdot (E((I^{(Y)})^{out}) \cdot \lambda +$$

$$+ E\,(({}^{aug}I^{(Y\_mini)})^{out}) \cdot (1 - \lambda)) + \lambda_{p1} \cdot \Sigma_{p1} \cdot E((I_{p1\_i}^{(Y)})^{out}) + \quad (7)$$

$$+ \lambda_{p2} \cdot S_{p2} \cdot E((I_{p2\_i}^{(Y)})^{out}) + \lambda_{p3} \cdot S_{p3} \cdot E((I_{p3\_i}^{(Y)})^{out}),$$

where $({}^{aug}I^{(Y)})^{out} = f({}^{aug}I^{(Y)})$ – image feature vector with replaced fragment, *Y_mini* – person identifier on small copy image.

After passing through all training sample images, the loss function value for the epoch is calculated:

$$L_{epoch} = L_{epoch\_aug} = \frac{\sum \left(L_{R\_p} + L_{R\_p\_aug}\right) \cdot B}{S_{size}^{train}}, \quad (8)$$

where B – batch size, S_{size}^{train} – train sample size. Hidden layers weights w_{hidden} is changed based on L_{epoch} value:

$$w_{classifier} = w_{classifier} - \eta \cdot \frac{\partial L_{epoch}}{\partial w_{classifier}} \quad (9)$$

and classification layer $w_{classifier}$:

$$w_{hidden} = w_{hidden} - 0.1 \cdot \eta \cdot \frac{\partial L_{epoch}}{\partial w_{hidden}}, \quad (10)$$

where η - learning rate.

To reduce the loss function value, augmentation is applied only at the preliminary training stage, and fine tuning is performed on the original images. Loss function for the epoch is calculated as:

$$L_{epoch} = \frac{\sum L_{R\_p} \cdot B}{S_{size}^{train}}. \quad (11)$$

III. Training CNN and Experiments

A. Training model

For re-identification used [10] as baseline which was supplemented by a proposed algorithm based in compound descriptor. Hyperparameters for training are presented in Table 1. The training was carried out on a personal computer with characteristics: Intel Core i5 3.11 GHz, 16 Gb RAM, Nvidia GeForce RTX-3060 6 Gb.

TABLE I. Hyperparameters for training

Backbone network	ResNet-50
Datasets	Market-1501, DukeMTMC-ReID, MSMT17, PolReID1077
Learning rate	0.07; after 40 epochs: 0,007
Batch size	16
Epoch	80

To assess the re-identification accuracy we used metrics is Rank1 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. Metric mAP is estimates the mean value of the average precision for all queries and is calculated as:

$$mAP = \frac{1}{Q}\sum_{i=1}^{Q} AP_i, \qquad (12)$$

where AP is the average precision (the domain below the *precision/recall* curve). Here, *precision* =TP/(TP+FP), TP is the number of true positive query predictions, FP is the number of false positive query predictions, *recall*=TP/(TP+FN) is sensitivity, and FN is the number of false negative query predictions.

The search for key points is performed by YOLOv7Pose [11]. ResNet-50 [12] is used to extract features. Fig. 1 shows heatmaps examples that allow you to visualize which image area has a greater influence on decision the network. It can be seen from figure that more local features are distinguished after ResNet-50 third level, than after fourth. So, in Fig. 1a, CNN identifies several local areas, which characterized the person figure. These areas are the head, shoulders, hips and feet. While after fourth network level (Fig. 1b), only one part is distinguished, and as can be seen in the figure. Therefore, we embed classification layer after third level (Fig. 2), while reducing convolution layers number, but increasing re-identification accuracy. The effectiveness has also been confirmed experimentally. For the Market-1501 dataset, the re-identification accuracy when using our algorithm and pre-trained Resnet-50 from the third layer was 90.17 in the Rank1 metric and 76.95 for mAP, while from the fourth: 87.98 and 73.71 in metrics Rank1 and mAP respectively.

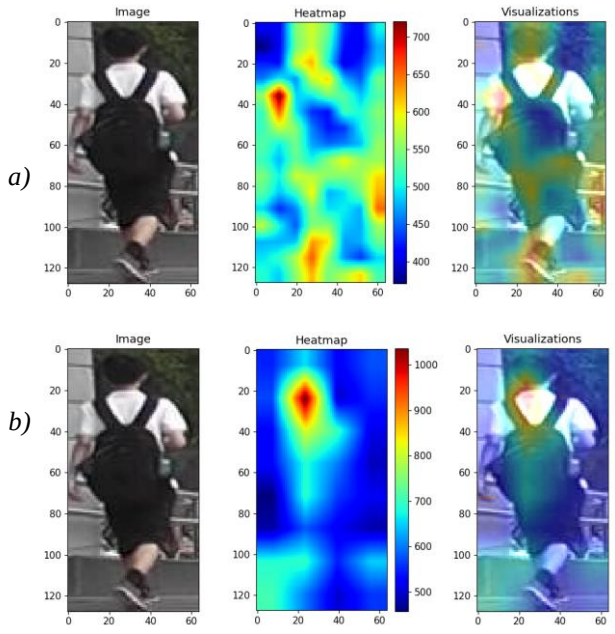

Fig. 1. Visualization heatmaps a) - after the third layer of the ResNet-50, b) - after the fourth layer of the ResNet-50

In proposed algorithm for each component the k-nearest neighbors feature vectors are searched independently of each other. Thus, a situation may arise that for different components the closest images will be different. Therefore, among the selected nearest neighbors, we look for those images whose local regions turned out to be the closest in most cases to k_1 nearest neighbors. Increasing *k* can lead to a large number of incorrect vectors among the nearest neighbors and this will add noise to the average component for feature vector invisible part. While an insufficient samples number means that the k_1 best nearest neighbors will not be optimally determined. The *k*-nearest and k_1-best nearest neighbors' number to replace the feature vector component of invisible fragments was determined experimentally. For testing, a pre-trained ResNet-50 model, augmentation [1] and the proposed algorithm were used. The values of *k* and k_1 were changed in steps of 2. The results of the experiments are presented in table 2.

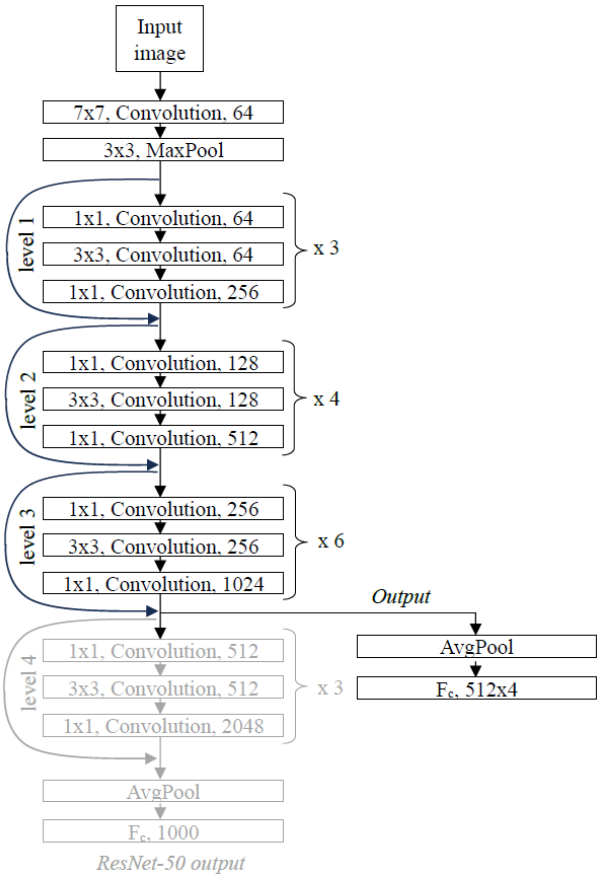

Fig. 2. ResNet-50 architecture modification for our algorithm

TABLE II. INFLUENCE OF THE VALUE OF K- AND K1-NEAREST NEIGHBORS ON THE RE-IDENTIFICATION ALGORITHM ACCURACY

k	Metrics	k_1				
		1	**2**	**4**	**6**	**8**
2	Rank1 mAP	89.99 81.31	–	–	–	–
4	Rank1 mAP	90.02 81.32	92.96 83.20	–	–	–
6	Rank1 mAP	90.05 81.37	93.02 83.24	92.99 82.95	–	–
8	Rank1 mAP	90.11 81.38	**93.02** **83.25**	93.02 83.11	92.81 82.85	–
10	Rank1 mAP	90.11 81.38	92.96 83.22	92.99 83.09	92.90 83.03	92.78 82.77
12	Rank1 mAP	90.17 81.38	92.90 83.17	92.90 83.08	92.81 83.07	92.61 82.04

As can be seen from the table 2, the highest accuracy is achieved at k = 8 and k_1 = 2. Replacing a feature vector fragment for image invisible area with the averaged value for the k-nearest neighbors is performed only at the testing stage. When training the CNN, this component is 0 and not taken into account when comparing and calculating the loss function.

B. Experimental result

Table 3 shows testing the algorithm on the Market-1501 [13], DukeMTMC-ReID [14], MSMT17 [15], and PolReID1077 [16] datasets results. We also estimated the network training time for each experiment. Four experiments were performed for each dataset:

1) For re-identification we used baseline [10] and ResNet-50 in a standard configuration without preliminary training on ImageNet. The obtained values were considered as reference values.

2) Based on the proposed algorithm and a stripped-down version of ResNet-50 without preliminary training on ImageNet re-identification was performed.

3) For re-identification the proposed algorithm was used using a compound descriptor, replacement of invisible fragments and a stripped-down version of ResNet-50 without preliminary training on ImageNet. During training, a two-stage training technology was used, in which augmentation [1] was applied during the first 45 training epochs.

4) In this case, an approach similar to the previous one was used, but ResNet-50 is pre-trained on ImageNet.

TABLE III. PROPOSED ALGORITHM ACCURACY ON THE MARKET-1501, DUKEMTMC-REID, MSMT17 AND POLREID DATASETS USING RESNET-50

Approach to training	**Metrics**	**Datasets**			
		Market-1501	**DukeMTMC -ReID**	**MSMT17**	**PolReID1077**
Baseline without pretrained	Rank1 mAP Training time, m	83.19 61.08 152	72.85 52.53 153	49.54 24.83 322	88.94 65.58 368
Our algorithm	Rank1 mAP Training time, m	86.67 69.49 183	77.83 59.33 250	54.87 28.31 473	89.29 65.27 405
Our algorithm and augmentation	Rank1 mAP Training time, m	90.26 76.38 204	80.97 65.20 250	63.05 36.08 499	93.83 76.50 387
Our algorithm, augmentation and pretrained CNN	Rank1 mAP Training time, m	**93.02** **83.25** 233	**84.92** **71.34** 240	**74.85** **48.89** 574	**95.85** **83.82** 372

The experimental results show that the use of the proposed algorithm for person re-identification from images makes it possible to increase the accuracy relative to the baseline in the Rank1 metric by 12% and mAP by 36% for the Market-1501 dataset. For DukeMTMC-ReID, the accuracy increased by 17%, 36% and 112% in the Rank1 and mAP metrics, respectively. For MSMT17 the accuracy increased for Rank1 by 51%, mAP by 97%. From Table 3 it is obvious that the metrics for PolReID1077 database have also become better for Rank1 by 8% and mAP by 28%.

IV. CONCLUSION

To reduce the occlusions impact on the re-identification accuracy, an algorithm is presented that uses a four-component feature vector in which invisible regions are replaced by k-nearest neighbors' average values. Compound descriptor includes the feature vector for the all person image and three local descriptors for top, bottom and middle parts of the human silhouette. The use of the proposed approach together with augmentation [1] made it possible to increase the accuracy relative to the baseline by 8 - 51% in the Rank1 metric and by 28 - 97% mAP for four datasets.

REFERENCES

[1] S.A.Ihnatsyeva., R.P. Bohush, " Improving person re-identification based on two-stage training of convolutional neural networks and augmentation", Informatika [Informatics], 2023, vol. 20, no. 1, pp. 40−54 (In Russian).

[2] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang**,** "Random Erasing Data Augmentation", AAAI Conference on Artificial Intelligence, 2017.

[3] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, P. 6448-6457

[4] Q. Yang, P. Wang, Z. Fang, Q. Lu. "Focus on the Visible Regions: Semantic-Guided Alignment Model for Occluded Person Re-Identification", Sensors (Basel, Switzerland), Vol. 20., 2020

[5] Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y., Zheng, W. "Learning to Know Where to See: A Visibility-Aware Approach for Occluded Person Re-identification", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021 P. 11865-11874.

[6] K. Zheng, C. Lan, W. Zeng, J. Liu, Zh. Zhang, Zh.-J Zha. "Pose-Guided Multi-Granularity Feature Learning for Occluded Person Re-Identification." Proceedings of 2022 the 12th International Workshop on Computer Science and Engineering. 2020

[7] T. Wang, H. Liu, W. Li, M. Ban, T. Guo, Y. Li. "Feature Completion Transformer for Occluded Person Re-identification". 2023. *ArXiv, abs/2303.01656*.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European Conference on Computer Vision (ECCV), 2014, pp. 740–755.

[9] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 542–551

[10] Person reID baseline pytorch. URL: https://github.com/layumi/Person_reID_baseline_pytorch

[11] Wang, C., Bochkovskiy, A., Liao, H.M. "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors", 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, P. 7464-7475.

[12] He, Kaiming, X. Zhang, Shaoqing Ren and Jian Sun. "Deep Residual Learning for Image Recognition,", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016

[13] L. Zheng, L. Shen, L. Tian, Sh. Wang, J. Wang, Q. Tian. "Scalable Person Re-Identification: A Benchmark," IEEE Int. Conf. on Computer Vision (ICCV), pp. 1116-1124, 2015

[14] Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. *ECCV Workshops*. 2016

[15] 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. 2017, P. 79-88

[16] S. A. Ihnatsyeva, R. P. Bohush, "Training Sample Formation for Convolution Neural Networks to Person Re-Identification from Video". Doklady BGUIR. 2023, 21 (3), p. 87–95. (in Russian)