




Multi-Depth Deep Similarity Learning for Person Re-Identification

Amir Sezavar¹ , Hassan Farsi^{2*} , Sajad Mohamadzadeh³ 

^{1,2,3}Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran
E-mail: hfarsi@birjand.ac.ir

Received: May 21, 2022

Revised: July 01, 2022

Accepted: July 04, 2022

Abstract— Detecting same people in different surveillance cameras, named person re-identification, has become a challenging and critical task in image processing. Since surveillance images usually have low resolution and different viewpoints, matching persons on them is still difficult. In this paper, a proposed method for person re-identification is introduced based on exploring similarity in different depth layers of convolutional neural network (CNN). To this end, after determining each person as a category for training CNN, optimum filters are obtained to find the best discriminative feature maps based on them. Smoothed discriminative features (SDF) are defined to compute similarity between persons. Experimental results, performed on CUHK01 database, demonstrate that the proposed method outperforms state-of-the-art feature extraction methods for person re-identification.

Keywords— Person re-identification; Image retrieval; Deep learning.

1. INTRODUCTION

Nowadays, using surveillance cameras in different aspects of life is developing sharply. Every day, millions of image streams of persons who are crossing in front of these cameras are saved, thus developing algorithms for detecting specific persons among this huge amount of data is needed. In detailed definition, person re-identification (Re-id) is finding a person who has been seen in a query camera between other non-overlapped cameras [1]. Each camera may capture hundreds of persons and hundred cameras may be in the way of one person. Therefore, Re-identifying and detecting person by human is difficult and time consuming specially in big cities. So, robust and fast automatic systems are needed to re-identify and track query person among thousands of images. Main applications of person Re-id are in security monitoring, crime prediction and controlling, health caring systems and urban security systems.

Although some machine learning techniques - especially deep learning - have achieved high accuracy on image processing, there are some challenges in person Re-id that limits the efficiency of these algorithms. Identification model should be robust against viewpoint and distance variation because people are not in the same direction and distance in different cameras. Low resolution and small scale are also another challenge for extracting good features from surveillance images. Also, different light scenes, rain, snow and fog can make Re-id more difficult. Therefore, a reliable model should be able to extract robust features from each frame in order to re-identify persons in the mentioned situations correctly. Query person is usually known as probe frame and other frames of people are in gallery set. Based on the number of available frames for each person, Re-id is categorized to single-shot (when one frame is available) and multi-shot (at least two frames are available for each person either in gallery set or probe) [1, 2]. In Fig. 1, a sample of person Re-id problem is illustrated.

* Corresponding author

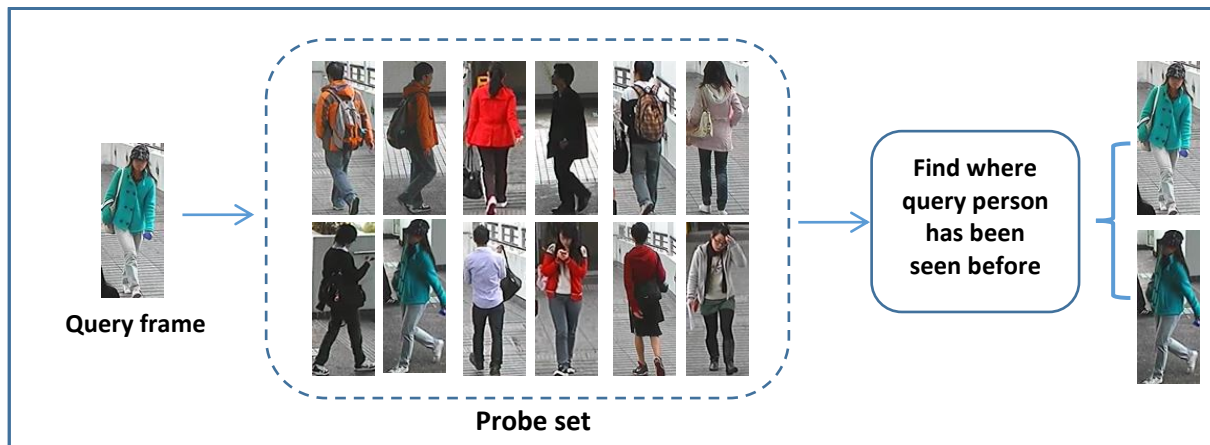


Fig. 1. Person Re-id sample system.

Extracting features from images plays a significant role in person Re-id. Despite hand-crafted features, deep learning based feature extraction has become more successful since it learns how to extract features in a hierarchical structure multilayer neural networks [3]. Although deep learning methods extract high-level detailed features from input, we claim that in person Re-id it is needed to compare both high level and local low level features. That is because color and local shapes play significant roles in distinguishing between different persons. Therefore, in this paper, the proposed method for person Re-id stands on comparing persons by local features from multi-depth layers.

The rest of the paper is organized as follows: some new researches in person Re-id by focusing on deep learning are reviewed in section 2. The proposed multi-depth features for similarity computing is introduced in section 3 where details of the pipeline is described. Section 4 discusses the results, details of implementation and evaluation metrics. The paper is concluded in section 5 in addition to some future recommendations.

2. RELATED WORKS

Person Re-id methods can be categorized in two types; the first one focuses on feature extraction and the second aims to create novel similarity measurements. Several researches have been done to extract hand-crafted features from each person. Munaro et al. introduced a model based on skeletal points of person [4]. It uses RGBD sensors to extract color and depth of pixels by concatenating color features and texture features using SIFT and SURF methods. In another work, a novel color descriptor is proposed based on salient color name (SCNCD) in which, evaluating similarity is done based on color names [5]. Other researches used handcrafted features based on color of pixels such as color based ranking aggregation (CBRA) [6] and color-based re-ranking [7]. The main difference between these methods and the proposed method is that, by using convolutional neural network (CNN), the model learns how to extract reliable features for each person where features are not same for all images. Therefore, it causes improvement in performance when feature extraction is adaptive based on input frame.

A metric learning approach was introduced by Liao et al. in cross-view quadratic discriminant analyzing model named XQDA that learns whether two input images are the same or different [8]. By combining representation learning and metric, a deep filter pairing

was introduced by Li et al. [9] in which, a CNN was used to extract features of pair images, following by some convolution layers to check either same or different persons. In recent years, a system to overcome occlusion was proposed by Chen et al. in which, occlusion-aware mask network (OAMN) was proposed [10]. They could improve performance against occlusion and body part loss. Another novelty in person Re-id was proposed by Khan et al. by combining deep CNN and autoencoders. To this end, they used pre-trained CNN to extract features of two parts of images, up and down, and then they used autoencoder to reduce the size of features. By considering tradeoff between complexity and accuracy, they provided an efficient model for re-id in smart cities [11]. These researches are some samples of salient works in person Re-id that have tried to solve some challenging tasks. Beside these approaches, there is also a need to work more on feature extraction to describe persons more accurately. Using well-trained CNN to extract features can help overcoming occlusion because during training, some neurons are randomly dropout and model learns how to extract features and classify them even with missing some parts of data

3. METHODOLOGY

In this section the proposed method is described in details. The first part is training a CNN as a baseline in order to learn to extract the best feature maps. The second part is exploring different layers to find the best filters which represent different persons more distinct in addition to similar persons. Finally the whole model is summarized.

3.1. Baseline Network

As deep learning models have grown up successfully in different machine learning tasks, they have been used in many image processing tasks especially in recent years [12, 13]. Among deep models, CNNs - which learn to extract features through hierarchical convolutional layers - are preferred for 2D and 3D inputs [3]. Among some well-known architectures such as AlexNet [14], VGGnet [15] and ResNet [16], VGG16 is selected for baseline based of input database and number of convolutional layers. In order to train CNN, after preparing the dataset - as will be described in section 4 - frames of each person is considered as a different category and the problem is considered as a classification task. Therefore, number of classes is equal to the number of individuals. Because deep CNNs have large number of parameters, they need huge amount of data for training. When training data is not enough, the model may perform well on training data but loses performance on test data. On this situation, the model is overfitted. To avoid overfitting, data augmentation technique is used before training. Parameters and training details will be explained in the next section. After the model is well-trained (by getting maximum accuracy and minimum cost function for training and data validation) , it is used as a baseline for the next step, exploring the best filters.

3.2. Filter Exploring

After training baseline, we visualize feature maps of convolution layers in different depths to analyze better distinguishing filters. To this end, for different images of same persons and for different images of different persons, the Euclidian distance between local heat maps in feature spaces are computed. In order to better find similarities in frames of

same persons, before computing distance, each feature map is smoothed by low pass filter as described in Eq. (1) to improve appearance.

$$B(x, y) = \sum_{i=1}^l \sum_{j=1}^l F(x, y) \cdot k(x - i, y - j) \tag{1}$$

where k is a kernel matrix with length $l \times l$ and all amounts to $1/l^2$, B is the blurred feature map and F is the visualized feature map. The blurred feature map shows local heatmaps' similarities and is considered as a local feature. To find optimum filters of multi-layers, same person pairs and different person pairs are considered. First, for same persons, after feeding frames to trained network, distance between all smoothed filters are computed as described in Eq. (2). After that, this distance is computed for different persons using Eq. (3).

$$d_{si} = \sqrt{\sum_{i=1}^w \sum_{j=1}^h (B_{1si}(x, y) - B_{2si}(x, y))^2} \tag{2}$$

$$d_{di} = \sqrt{\sum_{i=1}^w \sum_{j=1}^h (B_{1di}(x, y) - B_{2di}(x, y))^2} \tag{3}$$

where d_{si} is the distance between i -th filters of same persons, d_{di} is the distance between i -th filters of different persons, w and h are weight and height of the frame, B_{1si} and B_{2si} are i -th blurred feature maps of the same person 1 and 2, while B_{1di} and B_{2di} denote i -th blurred feature maps for different person 1 and 2. After that, for each layer, filter which maximizes the difference between d_{si} and d_{di} is calculated as optimum filter for the specific layer:

$$f_o = \operatorname{argmax}_i (d_{di} - d_{si}) \tag{4}$$

in which, f_o denotes the optimum filter for the layer. By concatenating optimum filters of different layers, final features vector is calculated to use in re-identification. Block diagram of the proposed method is illustrated in Fig. 2 and the algorithm for training and test steps is shown in algorithm 1 (Fig. 3).

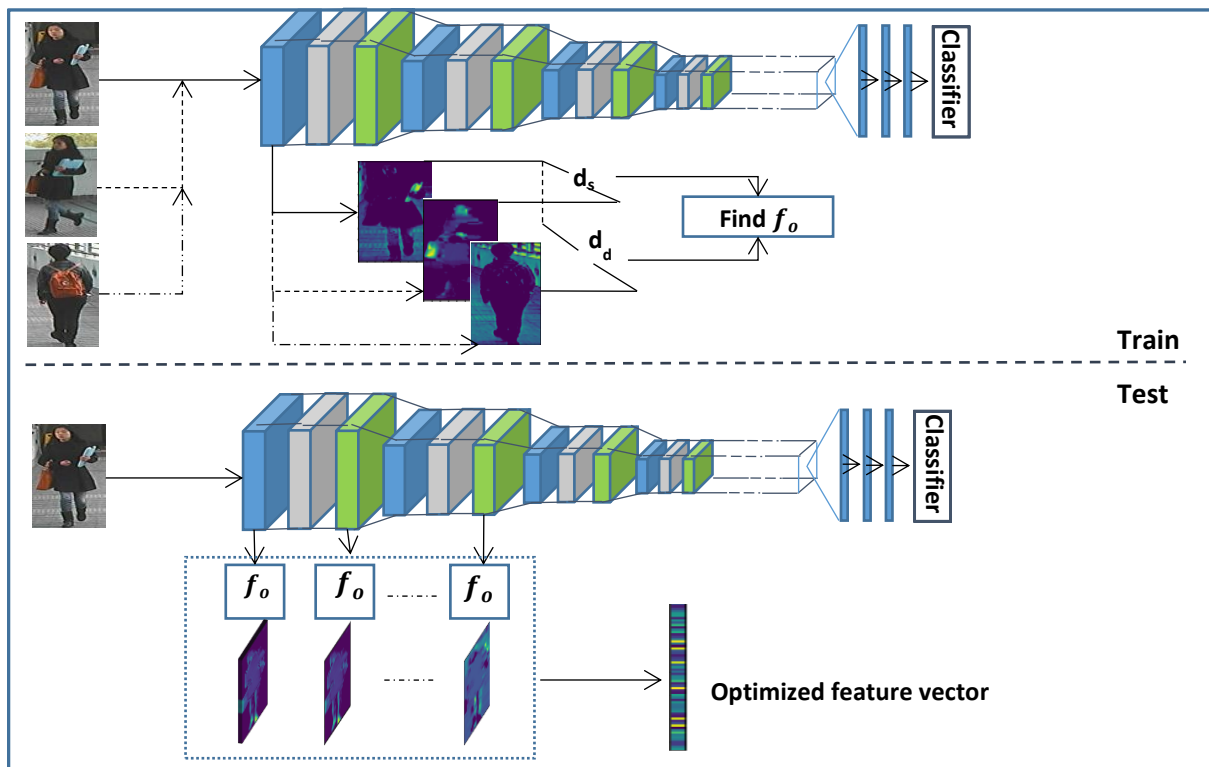


Fig. 2. Structure of the proposed method.

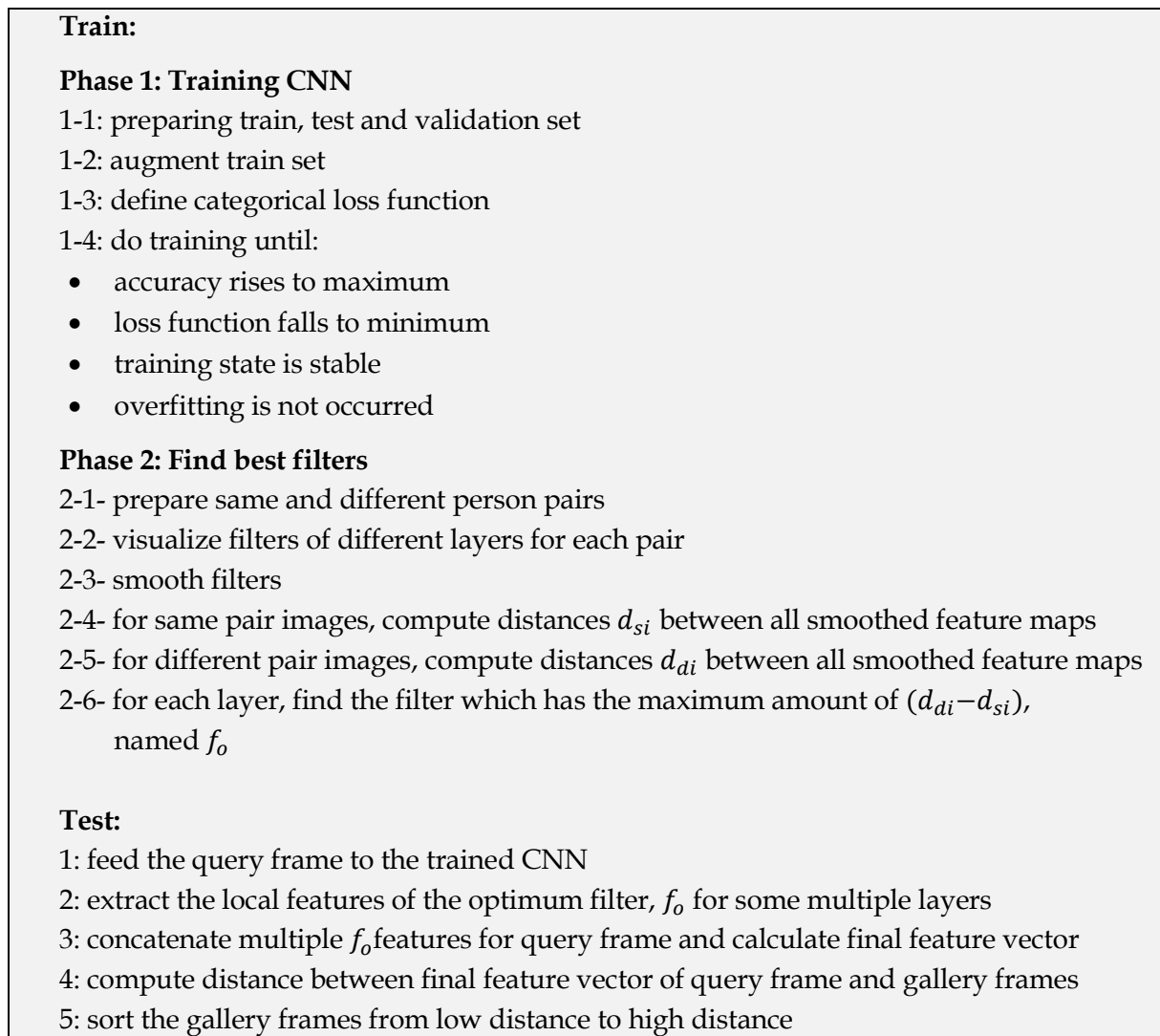


Fig. 3. The algorithm of training and testing the proposed method for person Re-id.

4. NUMERICAL RESULTS

As a baseline deep network, VGG16 is considered and trained on a famous Re-id database named CUHK01 [17]. It consists of 971 people walking on different streets and four frames are available from two disjoint cameras on different views for each person. Therefore, a total of 3884 images are available on CUHK01. The database is divided into two sets; one of them consists of 487 people for training baseline and the rest for testing. The testing set is also divided to two subsets, gallery and probe. The training was done without overfitting by tuning hyper-parameters of the model. Training and validation accuracies and loss functions during the training phase are illustrated in Fig. 4.

Next step after training baseline is to find optimum filters which discriminate well between persons in different layers of CNN. To this end, different batches of three frames are selected in each of which, two frames belong to the same person and one frame related to another person. For each batch, after extracting feature maps and smoothing them, d_{di} and d_{si} were computed and optimum filter was calculated based on Eq. (4). A sample of this process for three frames after simulation is demonstrated in Fig. 5.

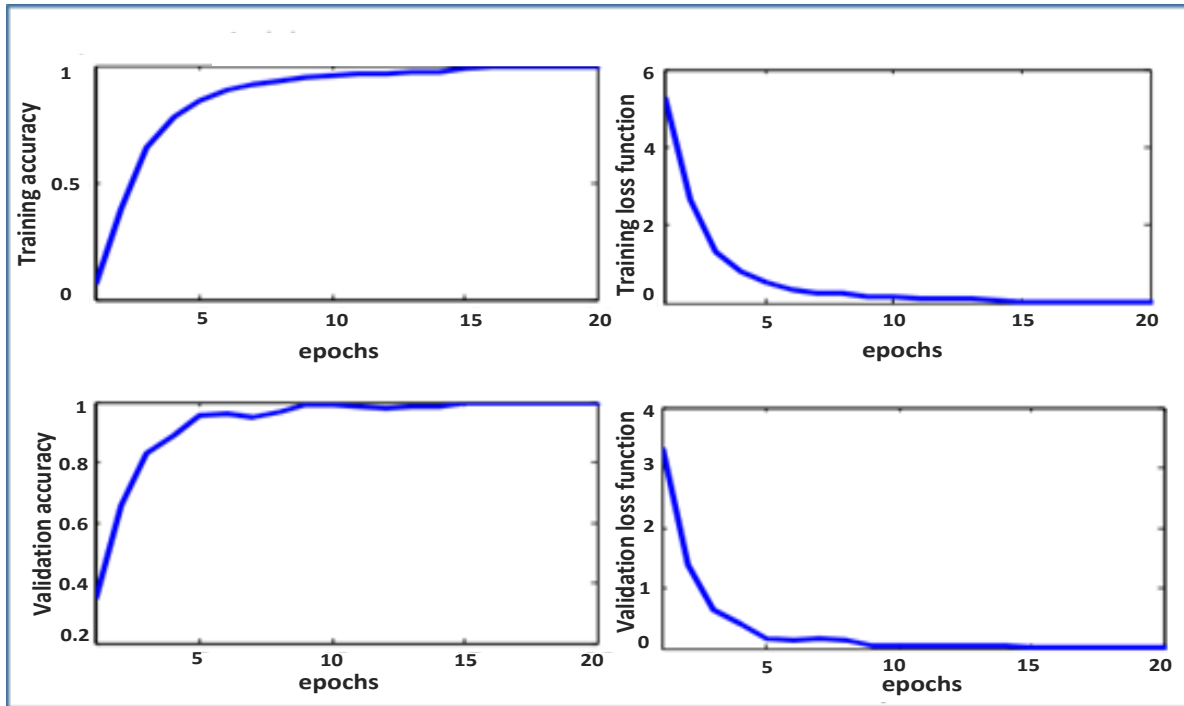


Fig. 4. Training and validation accuracies and loss function on CUHK01.

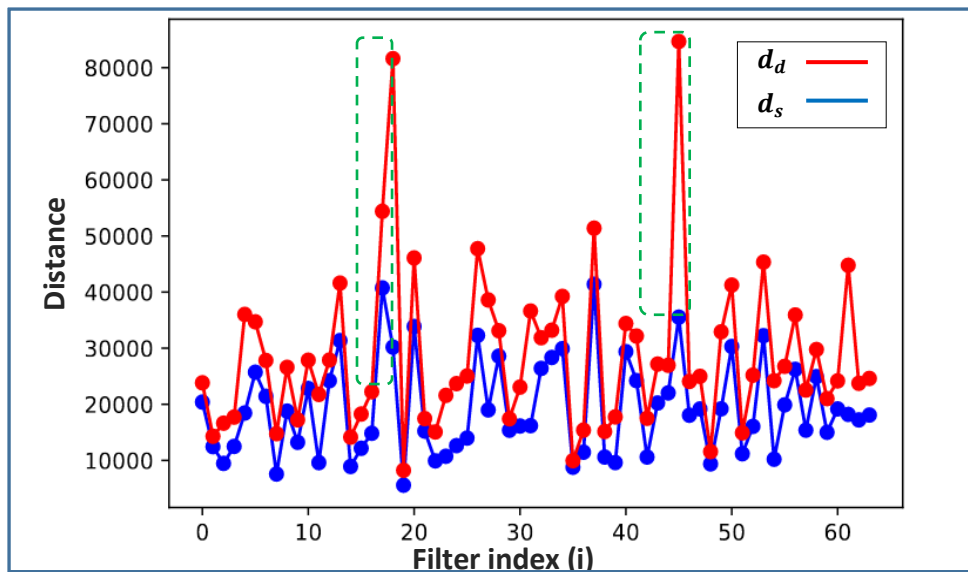


Fig. 5. Finding optimum filter for a batch of three frames.

It can be seen from Fig. 5 that for the sample batch, when filter related to $i=46$ is used, distance between same frames (d_s) and different frames (d_d) stand on maximum margin. Therefore, for this batch, this filter of this layer can discriminate better between different persons and can be considered as f_0 . The next f_0 filter for this layer is located in $i=19$. Frames of this sample batch and their relevant visualized f_0 are demonstrated in Fig. 6 in which, reshaped histograms of visualized f_0 are illustrated. To this end, each feature map is divided into three parts (top, middle and bottom) and histogram of each part is plotted. Then the reshaped histogram is created by putting them together in one diagram with three peaks. These histograms are shown to compare better between f_0 feature maps since a difference is sensible in the middle peak of histograms.

In order to compare numerical efficiency of the proposed model with some different feature extraction methods, ranking accuracy of Re-id model is calculated. In this metric, persons who are identified by the model will be sorted from most to least similar and a ranked list is calculated. In this list, Rank-k denotes the probability of existence of relevant person in k-th place of ranked list. The proposed method is implemented on Python using Tensorflow and Keras [18] Libraries on GeForce1080 GPU and 16 GB of RAM. The learning rate of training was set to 0.001 and the size of blurring filter was considered as 20. Experimental results for k=1, 10 and 20 are illustrated in Table 1.

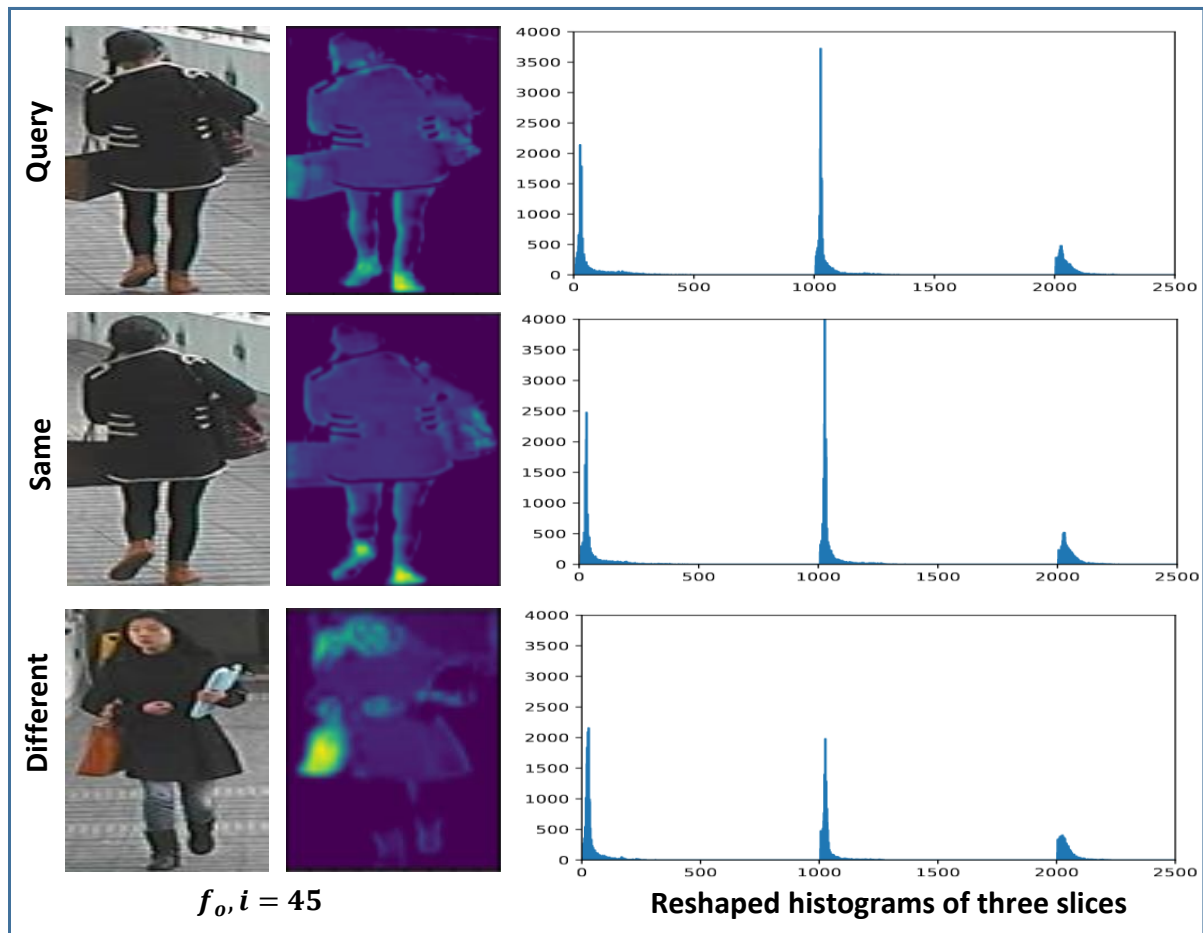


Fig. 6. A sample batch of three frames, query, same and different person and visualizing the f_o , calculated optimized feature map, and reshaped histogram for each optimized feature map.

Table 1. Numerical results of the proposed method and other state-of-the art methods on CUHK01.

Method	Rank-1	Rank-10	Rank-20
FCDF [19]	48.14%	81.68%	92.47%
LOMO-XQDA [8]	49.2%	84.2%	90.8%
KCVDCA [20]	47.8%	83.4%	89.9%
TCP [21]	53.7%	91.0%	93.3%
GOG-RGB [22]	51.96%	-	76.29%
GOG-Lab [22]	52.16%	-	77.94%
MLGD [23]	54.5%	83.5%	90.5
Proposed	55.2%	89.01%	94%

Based on Table 1, it can be seen that the proposed method achieved 55.2% average accuracy of rank-1 improving 1.5% of TCP [21]. Another method for extracting robust features for Re-id is GOG which achieved 51.96% accuracy of rank-1. In the 10-th place of ranked list, TCP performs a little better achieving 91% for rank-10 while the proposed method performs 89.01% accurate in rank-10. In rank-20, the proposed method stands at top of other methods by achieving 94% of accuracy. The average accuracy of the proposed method shows that our features are robust and the model can be used generally for person Re-id tasks.

5. CONCLUSIONS

In this paper, a new method for person re-identification is proposed which aims to calculate similarities between different layers of CNN. To this end, after training a deep baseline on database, different batches of input frame is fed to the model where each batch consists of a query frame, same frame and different frame. After that, a system is defined to find optimum filters in each layer by maximizing the difference between same and different smoothed feature maps. By concatenating optimized feature vectors of different depth of CNN, a robust feature vector is created for each person that will be used to compute similarity between different people. We obtain that by computing optimum filters - that maximize the distance between different persons and minimize it between same persons - the performance of the model gets better. Therefore, it is concluded that finding optimum filters in each layer of CNN can help in extracting more reliable features, and that concatenating optimum features grows up the accuracy of person Re-id system. Experimental results on a famous Re-id database, CUHK01, show that the proposed method achieves reliable performance compared to new models in person Re-id.

REFERENCES

- [1] S. Gong, M. Cristani, C. Loy, T. Hospedales, "The re-identification challenge," in *Person Re-Identification. Advances in Computer Vision and Pattern Recognition*, Springer: London, pp. 1-20, 2014.
- [2] I. Masi, G. Lisanti, F. Bartoli, A. Del Bimbo, *Person Re-Identification: Theory and Best Practice*, Tutorial at BTAS, 2015.
- [3] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. Lew, "Deep learning for visual understanding: a review," *Neurocomputing*, vol. 187, pp. 27-48, 2016.
- [4] M. Munaro, S. Ghidoni, D. Dizmen, E. Menegatti, "A feature-based approach to people re-identification using skeleton keypoints," *2014 IEEE International Conference on Robotics and Automation*, pp. 5644-5651, 2014.
- [5] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S. Li, "Salient color names for person re-identification," *European Conference on Computer Vision*, pp. 536-551, 2014.
- [6] R. de Carvalho Prates, W. Schwartz, "CBRA: color-based ranking aggregation for person re-identification," *2015 IEEE International Conference on Image Processing*, pp. 1975-1979, 2015.
- [7] Z. Mortezaie, H. Hassanpour, A. Beghdadi, "A color-based re-ranking process for people re-identification paper ID 21," *2021 9th European Workshop on Visual Information Processing*, pp. 1-5, 2021.
- [8] S. Liao, Y. Hu, X. Zhu, S. Li, "Person re-identification by local maximal occurrence representation and metric learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197-2206, 2015.

- [9] W. Li, R. Zhao, T. Xiao, X. Wang, "Deepreid: deep filter pairing neural network for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152-159, 2014.
- [10] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, R. Ji, "Occlude them all: occlusion-aware attention network for occluded person re-id," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11833-11842, 2021.
- [11] S. Khan, T. Hussain, A. Ullah, S. Baik, "Deep-ReID: deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance," *Multimedia Tools and Applications*, pp. 1-22, 2021.
- [12] A. Sezavar, H. Farsi, S. Mohamadzadeh, "Content-based image retrieval by combining convolutional neural networks and sparse representation," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20895-20912, 2019.
- [13] X. Zhao, R. Zhang, J. Wu, P. Chang, "A deep recurrent neural network for air quality classification," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 9, no. 2, pp. 346-354, 2018.
- [14] A. Krizhevsky, I. Sutskever, G. Hinton. "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [15] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv Preprint ArXiv:1409.1556*, 2014.
- [16] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.
- [17] W. Li, R. Zhao, X. Wang, "Human reidentification with transferred metric learning," *Asian Conference on Computer Vision*, pp. 31-44, 2012.
- [18] F. Chollet, *Keras*, 2015. <<https://keras.io/>>
- [19] M. Fayyaz, M. Yasmin, M. Sharif, J. Shah, M. Raza, T. Iqbal, "Person re-identification with features-based clustering and deep features," *Neural Computing and Applications*, vol. 32, no. 14, pp.10519-10540, 2020.
- [20] Y. Chen, W. Zheng, J. Lai, P. Yuen, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1661-1675, 2016.
- [21] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335-1344, 2016.
- [22] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, "Hierarchical gaussian descriptors with application to person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2179-2194, 2019.
- [23] D. Vishwakarma, S. Upadhyay, "A deep structure of person re-identification using multi-level gaussian models," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 4, no. 4, pp. 513-521, 2018.