A New Model for Person Reidentification Using Deep CNN and Autoencoders

A. Sezavar, H. Farsi*, S. Mohamadzadeh

Department of Electrical and Computer Engineering, University of Birjand, Iran

PAPER INFO	ABSTRACT
	Person re-identification (re-id) is one of the most critical and challenging topics in image processing and artificial intelligence. In general, person re-identification means that a person seen in the field of view of one camera can be found and tracked by other non-overlapped cameras. Low-resolution frames, high occlusion in crowded scene, and few samples for training supervised models make re-id challenging. This paper proposes a new model for person re-identification to overcome the noisy frames and extract robust features from each frame. To this end, a noise-aware system is implemented by training an auto-encoder on artificially damaged frames to overcome noise and occlusion. A model for person re-identification is implemented based on deep convolutional neural networks. Experimental results on two actual databases, CUHK01 and CUHK03, demonstrate that the proposed method performs better than state-of-the-art methods.

INTRODUCTION

Person re-identification in video sequences is one of the most critical and challenging issues in video and image processing. Nowadays, due to the growing use of surveillance and control cameras in public places, the importance of recognizing people is more noticeable than ever. Person re-identification (re-id), in general, means that a person seen in the viewing angle of a camera can be found and tracked in other cameras that do not overlap with the primary camera. Naturally, in crowded public places, the number of people in the camera's field of view is vast, and therefore the target person (query) must be identified among all these people. In other words, person re-id means comparing the person seen in the search camera with a gallery of candidates seen in non-overlapped cameras [1]. Figure 1 indicates the concept of person re-id systems.

Although image processing techniques and applications have made significant progress in recent years, some challenges in person re-id encourage researchers to work on it. For example, in a surveillance system, a person who leaves the frame of a camera will be seen at least in another camera, which may have a different distance and angle than the first camera [2, 3]. Also, distances between the person and the cameras are high, and biometric features of individuals cannot be extracted well. Meanwhile, the frames recorded by these cameras have low quality. Therefore, the importance of personal re-id systems is felt more than ever.

Person re-id has many applications in industry. Among the most important of them, it can be mentioned security and surveillance applications in the city, airports, stores, and other places that are controlled by surveillance cameras. Also, health monitoring systems, long-tracking systems, and crime prevention are other critical applications of person re-identification. Every re-id system consists of two main parts, feature extraction, and similarity measurement. Researchers have developed different methods for feature extraction or metric learning (which will be noted in the next section). The mainstream of the proposed work in this paper is based on extracting concluded and efficient features based on the deep autoencoder and reducing re-id noises, such as occlusion, using another type of autoencoder. By combining these two parts, the proposed pipeline is defined. The main reason that CNN can extract robust features is that it simultaneously learns feature extraction and

*Corresponding Author Email: hfarsi@birjand.ac.ir (H. Farsi)

classification at an end-to-end optimization process despite traditional machine learning methods. Therefore, optimum weights and coefficients are calculated by minimizing the loss function. In our method, the model is forced to extract features even if the frame is noisy and damaged by combining CNN and auto-encoder and creating random arbitrary noises and distortions. Also, the main advantage of the proposed model is that it can repair distorted and noisy frames to overcome noise and occlusions, which are the main challenges of personal re-id.

The rest of the paper is organized as follows. Related works and some state-of-the-art methods are noticed in section 2. Section 3 describes the proposed method, numerical results, databases, and metrics are referred to in section 4, and finally, section 5 concludes the paper.

The Focus of the research on personal re-id is the two categories of extracting the appropriate feature and creating a similarity measure. Also, recently, some researchers emphasized feature extraction and similarity learning simultaneously. Feature extraction and similarity metric learning have been done using either supervised or unsupervised learning, and extensive research has been conducted in each of the above categories. Munaro et al. [4] proposed a method for person re-id, which finds the key points of the body skeleton in order to extract features from crucial points. RGBD sensors were used to extract color and depth features. Also, texture features were extracted using SIFT and SURF methods, and the model used combined color and texture features for person re-id.

Another method for person re-id was presented by Matsukawa et al. [5] for defining a descriptor based on the hierarchical distribution of pixel features. This descriptor improves the hierarchical covariance one, which has been well used in image classification applications. Since two orders of Gaussian distribution are used, the descriptor is called Gaussian of Gaussian (GOG), and the color and texture features of the images are extracted simultaneously. Researchers in another work introduced a feature extraction method called Local Maximal Occurrence (LOMO) and a new metric learning method named Cross-view Quadratic Discriminant Analysis (XQDA). In the feature extraction section, the presented method tries to extract robust features against changes in the brightness and direction of the person. In order to create stability against direction changes, the occurrence of horizontal local features is checked, and the probability of occurrence of these local features is maximized in the selected feature vector. Therefore, the maximum horizontal local information in the image will be in the feature vector, and the changes in this feature will be less with the change in the person's direction [6]. Some other research also has been done on handcrafted features based on color, such as Color Based Ranking Aggregation (CBRA) [7] and color-based re-ranking [8].

Beyond deep learning methods for re-id, another model is presented that tries extracting features and training similarity measures simultaneously using convolutional neural networks. According to the authors, the first-person recognition method uses deep learning [9]. Convolutional neural networks have been used to learn appropriate features from images automatically. Also, the similarity learning stage is trained and tuned by the same networks. In another method recently proposed for recognizing a person, researchers have presented a model that uses the characteristics of a person's ranking in the ranking list of the gallery [10]. By computing similarity from different layers of deep convolutional neural networks, a model is proposed by Sezavar et al. [11] to find the best filter in each layer that better describes the similarities. Although other methods were introduced before to use these features, they are based on data post-processing, and in the training phase, the network does not have access to the features of the ranking list. The authors declare that the paper is the first method to show that the appearance feature and the ranking background information can be simultaneously optimized for training to extract more distinctive features and achieve better recognition performance.

MATERIALS AND METHODS

The proposed method consists of two main parts: quality improvement based on auto-encoder and feature extraction for person re-id based on deep CNNs. Before going on to details, it is necessary to review auto-encoders and CNN.

Auto-encoder refers to a multilayer neural network that aims to reconstruct input through encoder and
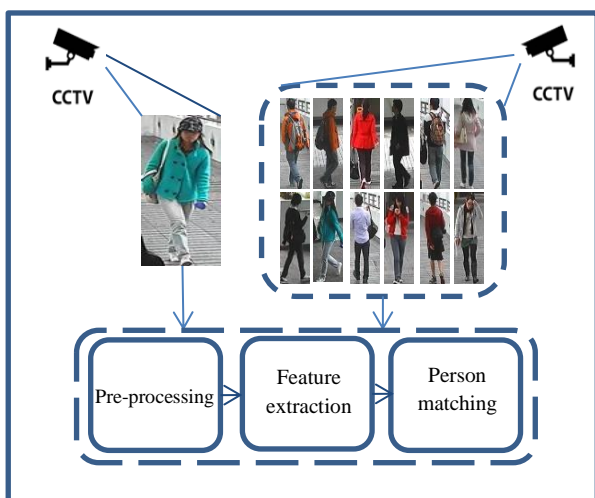

Figure 1. Concept of person re-identification

decoder layers [12]. In the encoder, the input image (or signal) is fed to the network, and a small vector is produced by extracting features and down-sampling. Then, the decoder tries to reconstruct the input by feeding the encoded feature vector during a supervised learning process. Auto-encoders have been used in many tasks, such as image processing for feature extraction and denoising [13-15].

Convolutional Neural Network (CNN) is a powerful type of deep network in which, instead of using neurons and calculating the linear combination of inputs, the result of convolution of the filters in each layer and the inputs is transferred to the subsequent layers. When our inputs are two-dimensional (such as an image), the concept of two-dimensional convolution becomes meaningful, and the network is made of filters whose coefficients can be trained. Also, CNN is one of the most important deep learning methods in which multiple layers are trained effectively and is commonly used in machine vision applications. Different parts of the image are given as input to hierarchical layers, and in each layer, by applying digital filters, significant features are extracted from the image [16-18].

The proposed method has three main duties: frame repairing for denoising and occlusion overcoming, feature extraction, and person re-identifying. Details of each part are explained in the following subsections.

Because detected frames in person re-id usually have low resolution and are noisy. In crowded scenes, occlusion occurs. This part is implemented to overcome occlusion and repair the frame by repairing occluded pixels. To this end, an auto-encoder is implemented and trained on randomly noised and damaged frames. Assume that the input frame of the person is, which is an RGB frame. We define Artificial Damaged Frame (ADF) as follows:

$$f' = \Omega(f + n) \tag{1}$$

which f' is the ADF, n is noise distribution (is supposed to be different kinds with different variances), and Ω is the operator that indicates the random patch pixel damaging. After creating the ADF, it is fed to the auto-encoder as an input frame, while the desired output is the main frame, f. During the training process, filter weights are updated until the mean squared error between the desired output and ADF is minimum.

$$E = \sum_i \sum_j \sqrt{(D(E(f'(i,j)) - f(i,j))^2} \tag{2}$$

in which, $D(.)$ and $E(.)$ denote decoder and encoder performing, respectively. The concept of this step is illustrated in Figure 2.

For the next step, a CNN is implemented for feature extraction. To this end, VGGNet [19] (demonstrated as an influential CNN in literature) is considered the backbone, and some fully-connected layers are added to tune the network for our work. The network is trained as a classification task where each person is considered a separate category. During the training process, the weights of convolution layers will be tuned until the classification error become minimum. The output of each convolution and fully connected layer is calculated by Equations (3) and (4).

$$C = Max\left(0, \sum_{i=1}^{K} \sum_{j=1}^{K} p(i,j) \times h(i,j)\right) \tag{3}$$

$$fc = Max\left(0, \sum_{j=1}^{K} w_i \times n_i\right) \tag{4}$$

where p refers to the input frame for each layer and h indicates filter's kernel. Also, in the training phase and for each layer, neuron weights (w_i) and input values (n_i) are
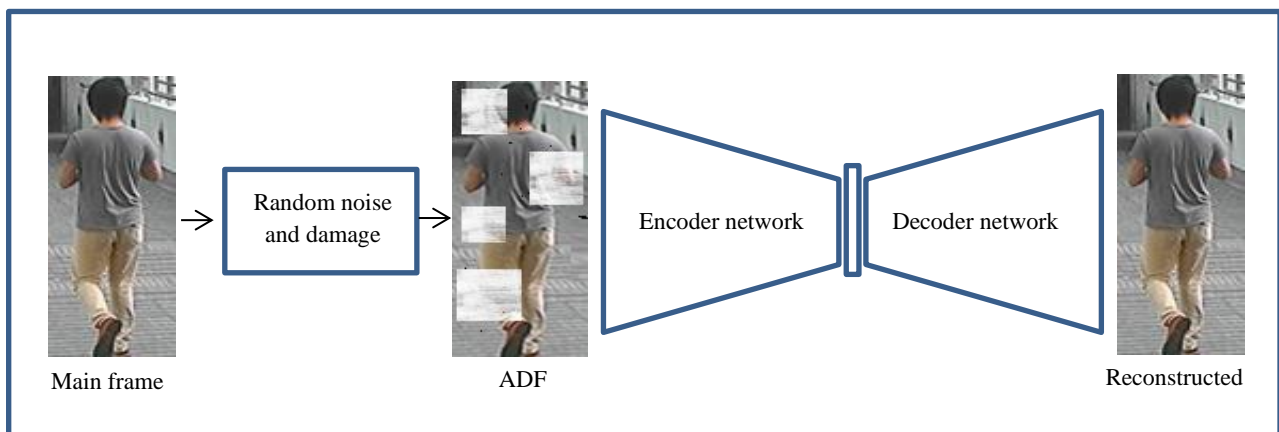

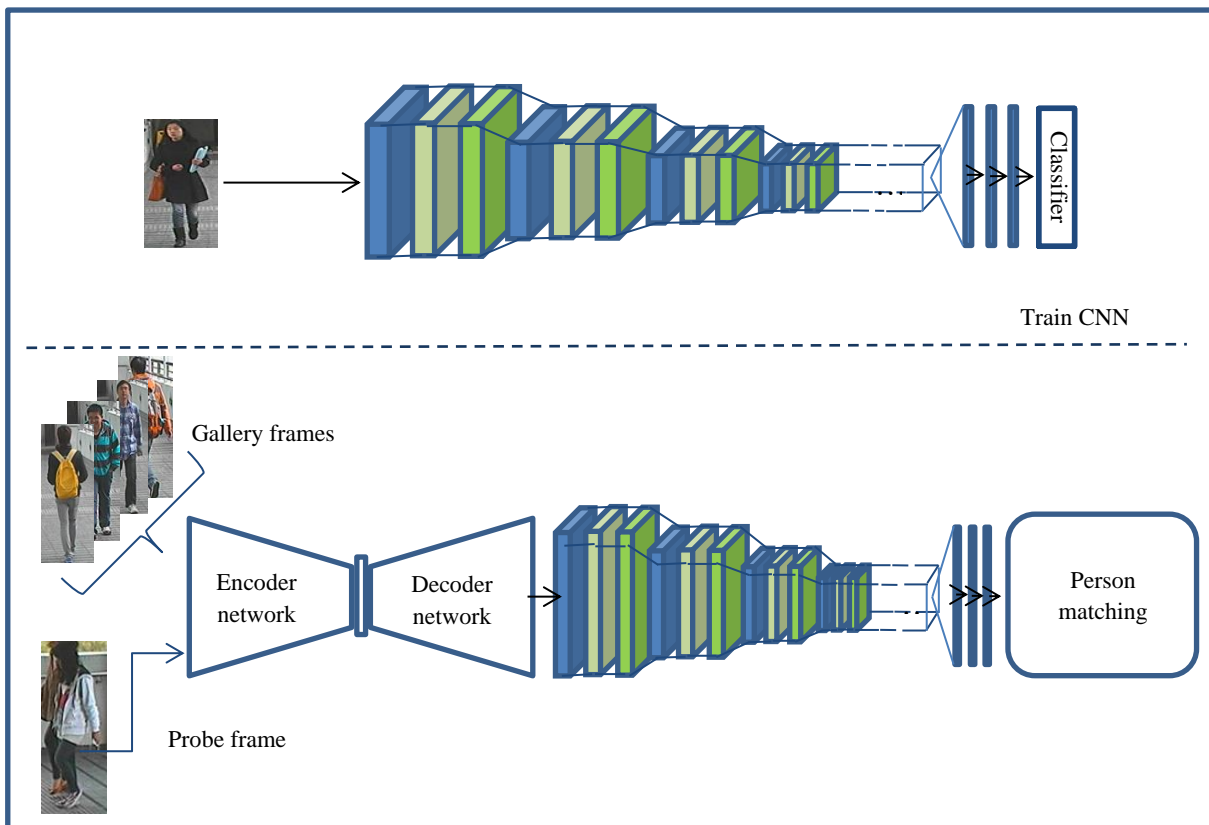**Figure 2.** The concept of ADH and auto-encoder training to reconstruct frame

Figure 3. The structural diagram of the proposed method

tuned to classify the input frame to a specific person. Therefore, it means that the network is learned well to extract features of each frame. We extract the last fully-connected features of each frame as a feature vector for person re-id. Then, the Euclidean distance between these feature vectors of each gallery frame and probe frame is calculated and sorted to find the nearest person. The structural diagram of the proposed method is shown in Figure 3

As shown in Figure 3, the trained auto-encoder on the ADF is used before feeding frames to the CNN. It causes improving frame quality and repairs damaged and occluded pixels. After that, extracted features from trained CNN are used to compare and match a similar person. The main novelty of the proposed method consists of two parts. The first one is that our model has an encoder-decoder in its backbone to make it noise-aware and repair noisy frames, which is a critical challenge in person re-id. To our best of knowledge, there have been no end-to-end models in the literature to do this for personal re-id. Also, our algorithm to create arbitrary distorted frames based on what is needed for a person's re-id helps in the training phase to make the model noise-aware. The second is training CNN to classify frames (by considering each person as a different class) and use features to compare frames to determine whether they are the same or different.

RESULTS AND DISCUSSION

In order to evaluate the proposed method, two famous databases, CUHK01 [20] and CUHK03 [21], are used. CUHK01 comprises 971 people, each with four frames from two non-overlapped cameras. Also, CUHK03 is made of 1467 people; for each, more than five frames are available, captured by five different non-overlapped cameras. For each dataset, half of the people are considered for the training step, and the remaining persons are used for the test, divided by gallery and probe set. Details of datasets and each set are shown in Table 1.

For evaluation, rank metrics are used. After computing the similarity between query frame and gallery frames, a ranked list is created by sorting similar frames decreasing. Rank-1 is the most challenging metric and means the probability of assigning the correct frame corresponding to the probe frame in the first retrieve. If the first frame in the ranked list is related to the query frame, Rank-1 is equal to 1 for this frame. Overall Rank-1 for the database is reported by averaging Rank-1 for all query frames. Rank-k also means the probability of being in the first k frames corresponding to the query frame. These metrics have been well-used in the literature for examining re-id systems. The training states for CNN and feature extraction is demonstrated in Figure 4. It consists of training and validation accuracies and loss functions

per epoch. It can be seen that the model is trained well on training frames. Also, converging curves of training and validation losses show that the model is generalizable for the new data.

The numerical results for the proposed method on CUHK01 can be found in Table 2, in which, Rank-k for K=1, 5, 10, and 20 are computed and compared with some state-of-the-art models. Also, these results are computed for CUHK03, and comparison results are shown in Table 3.

It can be seen from Table 2 that the proposed method achieved 92.1% rank-1 and performed better than other methods. Also, in other ranks, the proposed method stands upper than other methods, and this is because of using an auto-encoder for improving frames and using deep CNN for extracting features. Table 3 shows that Rank-1 equals 94.4% for the proposed method, which is 1.9% better than the best model in state-of-the-art methods. Also, by looking at performance in other ranks, it can be found that the proposed model performs well compared to other models on CUHK03.

In order to evaluate better, Rank-1 accuracies for different artificial noises for both databases are shown in Figure 5, with and without using the ADH-trained auto-encoder. As observed, we can see the improvement in

performance by forcing the model to learn from noisy and randomly damaged inputs. This improvement in accuracy is because the proposed model is noise-aware and can repair frames and extract robust features from each

Table 2. Numerical results and comparison on CUHK01

Method	Rank-1	Rank-10	Rank-20
KCVDCA [22]	47.8%	83.4%	89.9%
TCP [23]	53.7%	91.0%	93.3%
LOMO-XQDA[6]	49.2%	84.2%	90.8%
GOG-Lab [5]	52.16%	-	77.94%
SGL-DGDnet [24]	70.9%	93.5%	95.9%
FCDF [25]	48.14%	81.68%	92.47%
EMD [31]	69.38%	96.84%	-
MuDeep [30]	79.01%	99%	99.5%
PSD[26]	83.2%	97.1%	98.8%
The proposed	92.1%	97.5 %	99.3%

Table 3. Numerical results and comparison on CUHK03

Method	Rank-1	Rank-10	Rank-20
EMD[31]	52.09%	91.78%	-
MuDeep [30]	75.34%	99.5%	99.5%
PSD [26]	91.8%	99.1%	99.6%
EBb [27]	92.5%	98.9%	99.5%
SpindleNet[28]	88.5%	98.6%	99.2%
HGD-ResNet [29]	88.5%	98.4%	99.2%
The proposed model	94.4%	99.1%	99.65%

Table 1. Details of datasets and splitting sets

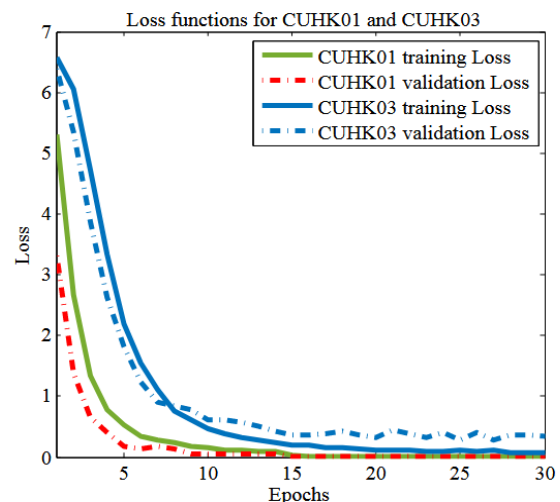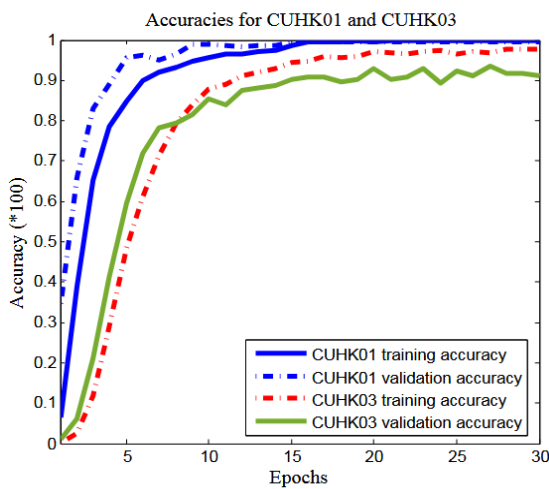| Dataset	#N person	#N frame per person	#N person For train and validation	#N person for train and validation
CUHK01	971	4	485	486
CUHK03	1467	7-10	734	733

Figure 4. Training accuracies and loss functions for training and validation data on CUHK01 and CUHK03

person. Some examples of visual results of the proposed method are illustrated in Figure 6, where query frames and some ranked frames are illustrated.
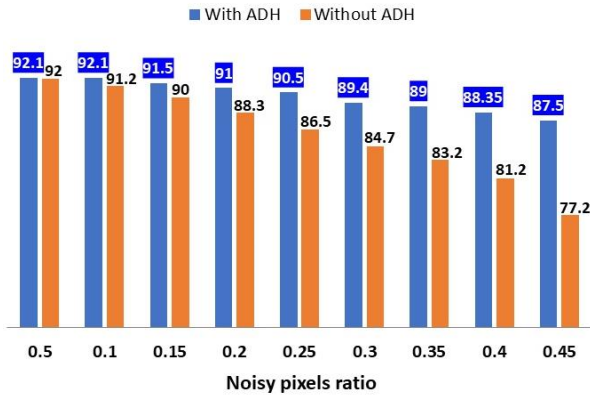

Figure 5. Comparing Rank-1 for different noises with and without ADH trained auto-encoder

Query frame   Ranked frame

Figure 6. Visual outputs of the proposed model ranked frames for three different query frames. A green marker in each row marks the correct re-identified frame.

CONCLUSION

Person re-identification is still a challenging and vital task in image processing. This paper proposed a new method for person re-identification by combining auto-encoders and deep CNNs. The auto-encoder was trained on artificially damaged frames to improve frame quality and overcome noise and occlusion. For extracting features, a model based on convolutional neural networks was designed in order to extract robust features for person re-

identification. Experimental results on two famous datasets, CUHK01 and CUHK03, demonstrated that the proposed method performs better than state-of-the-art methods. Using an unsupervised method and combining it with generative models to overcome insufficient data for training can be considered in future work to improve the model.

REFERENCES

[1] Gou, M., Wu, Z., Rates-Borras, A., Camps, O. and Radke, R. J., 2018. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3), pp.523-536. Doi: 10.1109/TPAMI.2018.2807450

[2] Gong, S., Cristani, M., Loy, C. C., and Hospedales, T. M., 2014. The re-identification challenge. In *Person re-identification*. Springer, London, pp. 1-20.

[3] Yaghoubi, E., Kumar, A., and Proença, H., 2021. Sss-pr: A short survey of surveys in person re-identification. *Pattern Recognition Letters*, 143, pp.50-57. Doi: 10.1016/j.patrec.2020.12.017

[4] Munaro, M., Ghidoni, S., Dizmen, D. T., and Menegatti, E., 2014. A feature-based approach to people re-identification using skeleton keypoints. *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5644-5651. Doi: 10.1109/ICRA.2014.6907689

[5] Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y., 2016. Hierarchical gaussian descriptor for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363-1372.

[6] Liao, S., Hu, Y., Zhu, X., and Li, S. Z., 2015. 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.

[7] de Carvalho Prates, R. F., and Schwartz, W. R., 2015. CBRA: Color-based ranking aggregation for person re-identification. *IEEE International Conference on Image Processing (ICIP)*, pp. 1975-1979. Doi: 10.1109/ICIP.2015.7351146

[8] Mortezaie, Z., Hassanpour, H., and Beghdadi, A., 2021. A Color-Based Re-Ranking Process for People Re-Identification: Paper ID 21. *9th European Workshop on Visual Information Processing (EUVIP)*, pp. 1-5. Doi: 10.1109/EUVIP50544.2021.9484056

[9] Li, W., Zhao, R., Xiao, T., and Wang, X., 2014. Deepreid: Deep filter pairing neural network for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152-159.

[10] Wu, G., Zhu, X., and Gong, S., 2022. Learning hybrid ranking representation for person re-identification. *Pattern Recognition*, 121, pp.108239. Doi: 10.1016/j.patcog.2021.108239

[11] Sezavar, A., Farsi, H., and Mohamadzadeh, S., 2022. Multi-Depth Deep Similarity Learning for Person Re-Identification. *Jordan Journal of Electrical Engineering. All rights reserved-Volume*, 8(3), pp. 279-287. Doi: 10.5455/jjee.204-1653115709

[12] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A., 2008. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096-1103. Doi: 10.1145/1390156.1390294

[13] Tripathi, M., 2021. Facial image denoising using AutoEncoder and UNET. *Heritage and Sustainable Development*, 3(2), pp.89-96. Doi: 10.37868/hsd.v3i2.71

[14] Saad, O. M., and Chen, Y., 2020. Deep denoising autoencoder for seismic random noise attenuation. *Geophysics*, 85(4), pp. 367-376. Doi: 10.1190/geo2019-0468.1

[15] Shang, Z., Sun, L., Xia, Y., and Zhang, W., 2021. Vibration-based damage detection for bridges by deep convolutional denoising autoencoder. *Structural Health Monitoring*, 20(4), pp. 1880-1903. Doi: 10.1177/1475921720942836

[16] Arel, I., Rose, D. C., and Karnowski, T. P., 2010. Deep machine learning-a new frontier in artificial intelligence research [Research frontier]. *IEEE Computational Intelligence Magazine*, 5(4), pp. 13-18. Doi: 10.1109/MCI.2010.938364

[17] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S., 2016. Deep learning for visual understanding: A review. *Neurocomputing*, 187, pp. 27-48. Doi: 10.1016/j.neucom.2015.09.116

[18] Sezavar, A., Farsi, H., and Mohamadzadeh, S., 2019. Content-based image retrieval by combining convolutional neural networks and sparse representation. *Multimedia Tools and Applications*, 78(15), pp. 20895-20912. Doi: 10.1007/s11042-019-7321-1

[19] Simonyan, K., and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*. Doi: 10.48550/arXiv.1409.1556

[20] Li, W., Zhao, R., and Wang, X., 2013. Human reidentification with transferred metric learning. *Asian Conference on Computer Vision*. Springer, Berlin, Heidelberg. pp. 31-44. Doi: 10.1007/978-3-642-37331-2_3

[21] Li, W., Zhao, R., Xiao, T., and Wang, X., 2014. Deepreid: Deep filter pairing neural network for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 152-159.

[22] Chen, Y. C., Zheng, W. S., Lai, J. H., and Yuen, P. C., 2016. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(8), pp. 1661-1675. Doi: 10.1109/TCSVT.2016.2515309

[23] Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N., 2016. 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.

[24] Cheng, D., Gong, Y., Chang, X., Shi, W., Hauptmann, A., and Zheng, N., 2018. Deep feature learning via structured graph Laplacian embedding for person re-identification. *Pattern Recognition*, 82, pp. 94-104. Doi: 10.1016/j.patcog.2018.05.007

[25] Fayyaz, M., Yasmin, M., Sharif, M., Shah, J. H., Raza, M., and Iqbal, T., 2020. Person re-identification with features-based clustering and deep features. *Neural Computing and Applications*, 32(14), pp.10519-10540. Doi: 10.1007/s00521-019-04590-2

[26] Liu, Y., Zhang, Y., Coleman, S., Bhanu, B., and Liu, S., 2020. A new patch selection method based on parsing and saliency detection for person re-identification. *Neurocomputing*, 374, pp. 86-99. Doi: 10.1016/j.neucom.2019.09.073

[27] Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J. and Wang, X., 2018. Eliminating background-bias for robust person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 5794-5803.

[28] Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S. and Tang, X., 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1077-1085.

[29] Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y., 2019. Hierarchical gaussian descriptors with application to person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9), pp. 2179-2194. Doi: 10.1109/TPAMI.2019.2914686

[30] Qian, X., Fu, Y., Jiang, Y.G., Xiang, T. and Xue, X., 2017. Multi-scale deep learning architectures for person re-identification. *Proceedings of the IEEE International Conference on Computer Vision,* pp. 5399-5408.

Persian Abstract

چکیده

شناسایی مجدد افراد (re-id) یکی از موضوعات مهم و چالش برانگیز در پردازش تصویر و هوش مصنوعی است. به طور کلی، شناسایی مجدد شخص به این معنی است که فردی که در میدان دید یک دوربین دیده می‌شود، توسط دوربین‌های دیگر غیرهم‌پوشانی پیدا و ردیابی می‌شود. فریم‌های با وضوح پایین، انسداد بالا در صحنه‌های شلوغ، و نمونه‌های کمی برای آموزش مدل‌های تحت نظارت، شناسایی مجدد را چالش‌برانگیز می‌سازد. این مقاله یک مدل جدید برای شناسایی مجدد افراد برای غلبه بر فریم های پر سر و صدا و استخراج ویژگی های قوی از هر فریم پیشنهاد می کند. برای این منظور، یک سیستم آگاه از نویز با آموزش یک رمزگذار خودکار روی فریم‌های آسیب‌دیده مصنوعی برای غلبه بر نویز و انسداد پیاده‌سازی می‌شود. مدلی برای شناسایی مجدد افراد بر اساس شبکه های عصبی کانولوشنال عمیق پیاده سازی شده است. نتایج تجربی روی دو پایگاه داده واقعی، CUHK01 و CUHK03، نشان می‌دهد که روش پیشنهادی بهتر از روش‌های پیشرفته عمل می‌کند.