IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 3, June, 2023, pp. 685-691
Journal homepage: http://iieta.org/journals/isi

Real-Time Person Re-Identification Using Omni-Scale Feature Learning Network and Yolov5: A Comparative Study

Sundus A. Abdul Hussien*, Ali A. Abed

Department of Computer Engineering, University of Basrah, Basrah 61, Iraq

Corresponding Author Email: engpg.sundus.audah@uobasrah.edu.iq

https://doi.org/10.18280/isi.280318

Received: 31 January 2023
Accepted: 21 May 2023

Keywords:
Person Re-Id, OSNet, Yolov5, MARS, custom-created dataset

ABSTRACT

Video-based person re-identification seeks to match video footage of an individual across non-overlapping multi-camera systems in real-time, probing for instances of the same identity appearing at different locations and times. The critical process in video-based person re-identification involves feature aggregation from the video track. This study introduces a method utilizing a convolutional neural network model named Omni-Scale Feature Learning Network (OSNet) for video-based re-identification. The performance of this method is evaluated on the large-scale MARS dataset and compared with other network models. Furthermore, a novel approach using You Only Look Once version 5 (Yolov5) is proposed for the first time for image, video, and real-time person detection and re-identification. This approach was trained on a custom-created dataset, gathered from two cameras capturing multiple identities of Computer Engineering students at the University of Basrah. The proposed method yielded promising results, with a re-identification accuracy of 80%. The aim of this work is to establish a real-time person re-identification system using the Yolov5 algorithm, and to contrast its performance with that of the OSNet.

1. INTRODUCTION

Person Re-Identification (Re-Id) endeavors to match individuals across non-overlapping multi-camera systems, captured at diverse locations and times. This field has garnered considerable interest in recent years due to its extensive applications in surveillance, public security, and criminal investigation. Although deep learning methods have been demonstrated to be effective in the person Re-Id process, the endeavor remains fraught with challenges due to variable backgrounds, human pose variations, occlusions of body regions, changes in viewpoints, and lighting conditions. The process of feature aggregation, in particular, has been a focal point in many works.

In the initial stages of Re-Id research, image-based systems were favored over video-based ones, largely due to the dearth of comprehensive video datasets. However, with the recent proliferation of video-based Re-Id datasets, such as those in [1-3], video-based Re-Id techniques have experienced substantial advancement. Videos, as compared to images, possess advantageous qualities that prove beneficial for Re-Id tasks. For instance, a video sequence, composed of multiple frames, retains crucial temporal information, offering vital motion cues between frames. Additionally, distinct frames from each video provide varied visual instances for each identity, thereby enriching sample diversity. Furthermore, while image-based methods primarily exploit visual features, video-based Re-Id techniques target a blend of visual and temporal features. The latter approach extracts discriminative features for Re-Id tasks that are more resilient to Re-Id problems. However, the adoption of video-based approaches, while resulting in more discriminative feature embeddings, also escalates the complexity of Re-Id tasks. The fusion operation, the simplest method for creating video features by combining frame-level features, disregards the temporal dimension of frame sequences. Additionally, comparing samples becomes challenging due to variations in video frame rates and durations. Lastly, certain outlier frames can mislead the process of learning robust video representations, as not all frames contain discriminative information.

The YOLO (You Only Look Once) framework [4], has been deployed in numerous research studies for object detection. However, YOLO has not been utilized as a standalone tool for personal re-identification. This study proposes a novel deep learning-based Yolov5 framework [5], devised for constructing a real-time, video-based person re-identification system. This system is capable of identifying individuals who have been previously registered in the dataset.

The remainder of this paper is structured as follows: Section 2 reviews the most prominent and relevant methods, providing a survey of related works. Section 3 delves into the specifics of video-based datasets, provides the research background, and outlines the concepts of OSNet and Yolov5 architectures. The experiments and their corresponding results are detailed in Sections 4 and 5, respectively. Finally, Section 6 offers a conclusion, summarizing the primary findings, addressing limitations, and proposing directions for future research.

2. RELATED WORKS

Video-based person re-identification (Re-Id) can be viewed as an extension of image-based Re-Id. However, in this case, the learning algorithm is fed with pairs of video sequences as opposed to pairs of images. A critical stage of this process is the manner in which temporal features from the video are fused.

In the study conducted by Wang et al. [2], a method of selecting discriminative representations of spatial-temporal features was proposed. This involved choosing frames with the minimum or maximum flow energy, as determined by the fields of optical flow. Striving to fully utilize temporal information, McLaughlin et al. [6] developed a convolutional neural network (CNN) for the extraction of features from every frame. Subsequently, recurrent neural networks (RNNs) were employed to consolidate the temporal information among frames. The average of the outputs of RNN cells was adopted to summarize the resulting features.

In another study [7], RNNs were similarly utilized to code videos into sequences of features with the final hidden state serving as a representation of the video. A Quality Aware Network (QAN) was designed, which used a weighted average attention mechanism to aggregate temporal features [8]. This mechanism generated attention scores from the feature maps at the frame level.

Several other studies [9-11], introduced an attention module at different levels of the CNNs backbone to refine frame-level representations of a person, considering the contextual temporal relations among frames. This resulted in the development of STE-NVAN (Spatially and Temporally Efficient Non-Local Attention Network), M3D (Multi-scale 3D Convolution Network), and COSAM (Co-segmentation Inspired Attention Networks) respectively.

A hierarchical co-attention module was introduced by dividing frames into more than one granularity, with the goal of capturing discriminative features from multiple semantic levels [12]. Building on the success of 3D CNNs [13, 14], Liao et al. [15] proposed a method for aggregating temporal dependency features in a sequence of video frames.

The YOLO framework, primarily used as a tool for object detection, has been deployed in numerous studies. For instance, multiple face recognition was achieved [16], where a system was validated and tested using Yolov5x with a dataset of 7378 enhanced photos. This demonstrated an impressive inference rate of 0.817 seconds and the capability to recognize up to 20 faces simultaneously. YOLOv3 and TensorFlow Lite object detection platforms were used for automatic face mask detection through live infrared (IR) camera monitoring [17]. To date, YOLO has not been independently utilized for re-identification tasks.

The proposed method in this paper aims to employ YOLO for identifying individuals in a multi-camera system, using person re-identification datasets (MARS and a custom-created dataset). These datasets, consisting of full-body images captured from different angles, provide a robust basis for training the system.

3. VIDEO-BASED DATASETS AND BACKGROUND

Person re-id datasets have witnessed growth in number and scale. This growth has a large effect on deep models, which need a large number of samples to reach effective training. The sample variety presents great solutions for challenges in re-id such as lighting, occlusion, variation in pose, intra-class disparity, inter-class disparity, and background clutter. These solutions allow deep models for learning appearance generalization. The examples of video datasets: PRID [1] contains videos from 2 cameras for 983 identities, ILIDs_VID [2] has 300 identities with 300 videos collected from 2 cameras, MARS [3] is a large-scale dataset which contains 1261 identities and 20715 videos from 6 cameras and it is counted as the state-of-the-art dataset, and DukeMTMC_Video Re-Id [18] contains 1812 identities.

3.1 OSNet architecture

OSNet formed by assembling the light bottleneck layers with no pressure to dedicate the blocks at different depths (stages) of the network unlike the border bottleneck shown in Figure 1. Based on a multi-stream design, OSNet has similarities to ResNeXt [19] and Inception [20], but it also has significant distinctions. 1) The multi-stream architecture in OSNet is dependent on the exponent's scale-incremental rule, with each stream having its own field. These streams, however, organized using a particular Lite 3x3 layer (see Figure 2).

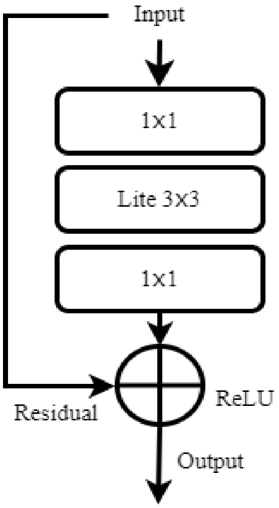

Figure 1. Border bottleneck

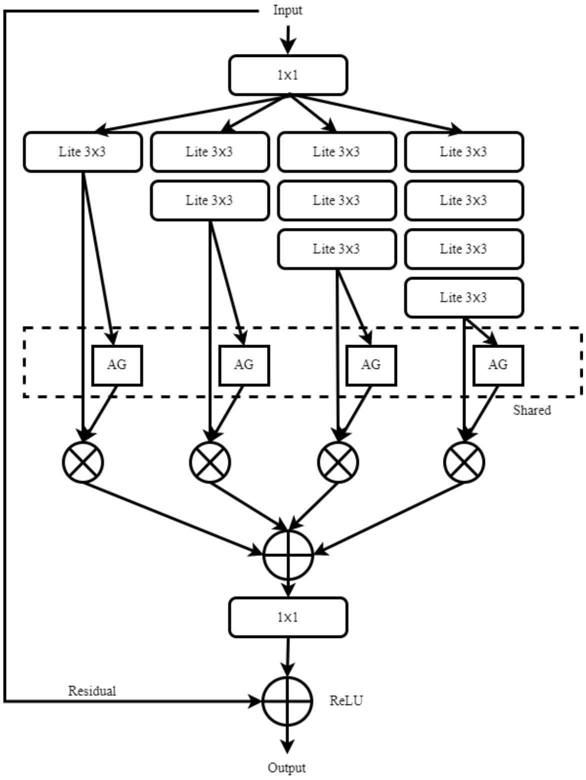

Figure 2. The light bottleneck of OSNet

That layout works over a wide range of scales. As a result, Inception designed to lower the cost of computation by dividing computations among many streams. The structure of Inception then created by hand using a combination of convolution and pooling procedures. Since ResNeXt contains many streams that are at the same stage, learning representations are accurate. 2) OSNet employs aggregation gates AG, which integrated, whereas ResNeXt/Inception collect characteristics through addition/concatenation. This makes combinations learning (combined forms of multiple-scale characteristics) easier, resulting in adaptive and dynamic fusing of each input image. OSNet is therefore architecturally distinct from ResNeXt/Inception. 3) Because factorized convolutions are used, both the OSNet network and the resulting building block are small and quick to run.

OSNet rained in this work using MARS dataset as a large-scale re-identification dataset. In Table 1 MARS and some video dataset with all their challenges, number of identities and cameras shown.

3.2 Yolov5

One of the most popular Convolutional Neural Networks (CNNs) for object identification is You Only Look Once (YOLO). The task of object detection comprises locating specific elements on a photograph and classifying them according to their location. Before YOLO created, CNNs like Region-Convolutional Neural Networks (R-CNN) used Regions Proposal Networks (RPNs), which produce proposal-bounding boxes on the input image first, then run a classifier on the bounding boxes, and finally apply post-processing to dispose of redundant detections and refine the bounding boxes. It was not appropriate to train each stage of the R-CNN network separately. YOLO is well-liked because it combines outstanding accuracy with real-time functionality.

The approach "looks at the picture just once" in the sense that it only does one forward propagation pass through the neural network while making predictions. It returns discovered items together with their bounding boxes after conducting non-max suppression, which guarantees that the object detection method only recognizes each item once. Alexey Bochkovsky's exclusive framework, Darknet [21], was used to create YOLO models. The business known as Ultralystic converts older iterations of YOLO to PyTorch, one of the most well-known deep learning systems created in the coding language Python. Five distinct sizes of Yolov5 were released, n stands for extra small (nano) model size, s stands for tiny model, the model abbreviation is m, huge size model l, and x for an extra-large model. Except for the number of layers and parameters, there is no difference in the operations employed by the five models. The same three parts used in all Yolov5 models: a backbone made of CSP-Darknet53, a neck, and head made of SPP and PANet, and the head from YOLOv4.

3.3 Yolov5 description

3.3.1 CSP-Darknet

The foundation of Yolov5 is CSP-Darknet53. The authors simply implemented the Cross Stage Partial (CSP) network [22] technique to the Darknet53 convolutional network, which served as the foundation for YOLOv3. The YOLO deep network employs residual and dense blocks to allow information to travel to the deepest levels and circumvent the vanishing gradient issue. The issue of repeated gradients is one benefit of having dense and residual blocks, though. By truncating the gradient flow, CSPNet aids in solving this issue. Applying this approach has significant benefits for Yolov5, as it reduces the number of parameters and requires less processing (few FLOPS), which increases the inference speed the key factor in real-time object detection models.

3.3.2 Neck of Yolov5

Two significant alterations to the model neck came with Yolov5. After using a variety of Spatial Pyramid Pooling (SPP), the Path Aggregation Network (PANet) [23] was altered by adding the BottleNeck CSP to its architecture. A feature pyramid network called PANet utilized in an earlier version of YOLO (YOLOv4) to enhance information flow and aid in accurate pixel localization for the task of mask prediction. This network altered in Yolov5 by the CSPNet technique applied to it. Without slowing down the network speed, Spatial Pyramid Pooling (SPP) offers the advantage of greatly expanding the receptive area and isolating the most important context elements. However, in Yolov5 (6.0/6.1), SPP has been employed to increase network speed.

3.3.3 Head of Yolov5

It made up of three convolution layers that forecast where the bounding boxes (x, y, height, and width), scores, and object classes would be.

In Figure 3, the architecture of Yolov5 shown and divided into three components: Head, Neck, and Backbone.

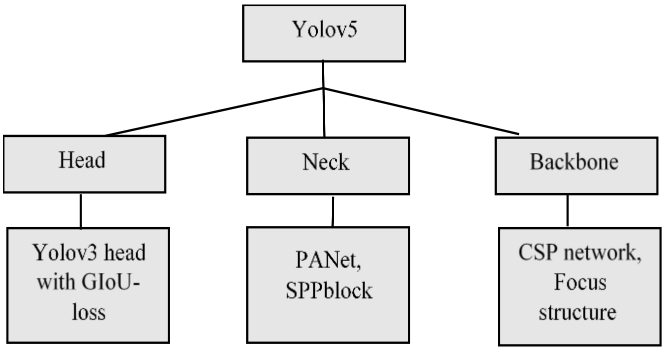

Figure 3. Yolov5 archeticture

Table 1. Video benchmark datasets

Dataset	No. of identities	Cameras	Year	Challenges	No. of images or videos
PRID	983	2	2011	Pose, viewpoint, background variation, and light	100-150 images for a person
iLIDs_VID	300	2	2014	Pose, viewpoint variation, and similar clothing-different person	600 videos
MARS	1261	6	2016	Pose, viewpoint variation, inter-class, and intra-class	20715 videos
DukeMTMC_ Video Re_Id	1812	-	2019	Noisy occlusion, pose, lighting, background, viewpoint variation	12 frames per second

4. EXPERIMENTS

4.1 Experiment on Mars

(1) The Convolutional Neural Network (CNN) in a variety of visual tasks, including person re-identification, has attained modern-day accuracy. To train the re-id model in classification mode, OSNet [24] is used which produces decent re-id accuracy. During training, photos downsized to 256 by 128 pixels and fed in batches to CNN along with their IDs (label). We define two fully connected layers, each with 1,024 blobs, through five convolutional layers with the OSNet structure [24]. In the case of MARS as shown in Table 2 below, there are 625 training identities, 626 as query, and 621 identities in the gallery; hence, there are exactly as many blobs in the eighth layer as identities. There are 518k training boundary boxes in total on MARS. The learning rate started at 0.0003 for 60 epochs. Data expansion includes arbitrary crop, arbitrary flip, arbitrary patch (create a patch pool that saves replicated picture patches arbitrarily), and past arbitrary patch (choosing from patch pool onto input image at a random place). OSNet has received training from AMSGrad [25]. Data augmentation involves random flipping and erasing. The created software is based on Torchreid [26].

(2) CNN model used to retrieve probe and gallery information anterior to the metric learning steps in testing because re-id differs from image classification in that the identities used for training and testing do not overlap. To be more precise, the model extracts the features for every bbox in an input tracklet. Then, a 1,024-dim vector created for a tracklet of random length using max/average pooling.

Table 2. MARS statistics

subset	#identities	#traklets
train	625	8298
query	626	1980
Gallery	621	9330

4.2 Experiment using Yolov5

It is known that one of the most popular deep learning-based object detection algorithms called "You Only Look Once," or YOLO. This section demonstrates how Yolov5 has been trained using a unique dataset. More specifically, a person's dataset used to train the Yolov5 detector (a custom-created dataset collected from many identities of college students). After the training, a person-localization and classification-capable object detector has created. Because Yolov5 is the most actively updated Python port of YOLO, it chosen over other variations. Other variations, such as YOLO v4, coded in C, which also might be less user-friendly than Python for the average deep-learning practitioner.

At the beginning, a work environment is prepared and the requirements for the work of Yolov5 have installed within it. Then, the dataset has prepared for training. Each identity considered as a class in Yolov5. The training process is applied to Yolov5 with a batch size equal to 32 for 100 epochs. All images in the dataset have partitioned into validation and training groups each group has its labels and images. After training ended, a model of Yolov5 capable of recognizing people that are registered in the dataset with good precision.

The Custom-Created Dataset has collected from two cameras of 11 identities of students of the computer engineering department at Basrah University (see Figure 4). The dataset has 5910 images and four tracklets for each identity on average. It has annotated in Yolov5 Oriented Object Detection format figure. Roboflow [27], which is a computer vision end-to-end platform, has been used to organize, annotate and create the dataset.

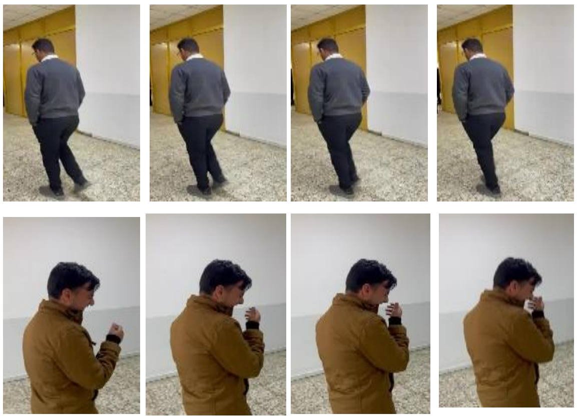

Figure 4. Example of the custom-created dataset

5. RESULTS

5.1 Results and visualizations of MARS

(1) After training the OSNet network on MARS dataset, the results come out as groups of folders each folder contains all the tracklets that have a feature similar to the probe tracklet. The tracklets that are exactly similar to the probe one are labeled as true tracklets and others are labeled as false tracklets. In Figure 5 an example of a person tracklet as a probe and the results of the true and false tracklets to the right (a: is the tracklet of the probe, b: is the true one, and c: is the false one. The mean average precision (mAp) and rank1 (R1) that resulted after testing the model on MARS are 80.7% and 82.4% respectively which are very good results for MARS performance in person re-id.

Table 3. Results on MARS

Base model	MARS	
	mAP	Rank1
Resnet50	72.0%	83.0%
Inception-v2	74.6%	84.2%
Resnet50-X	72.0%	83.4%
OSNet	**80.7%**	**82.4%**

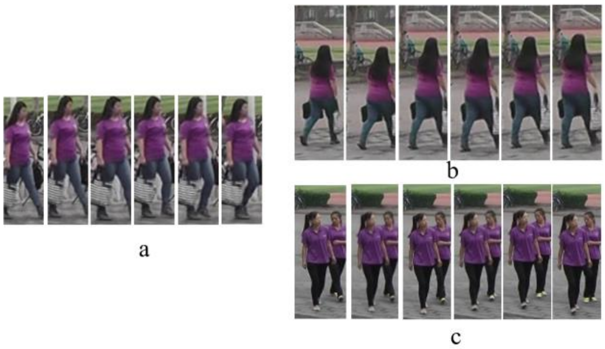

Figure 5. Visualization of MARS results. a: Probe tracklet, b: True tracklet, and c: False tracklet

(2) The same experiment on MARS has applied to some networks such as Resnet50 [28], Inception-v2 [29], and ResNeXt [22] on MARS, and the results are shown in Table 3 below which shows a comparison in performance on MARS among networks. Cumulative matching characteristics (CMC) (R1: rank 1 from CMC) and mean average precision (mAp) are measured for evaluation on MARS using OSNet and some of deep learning models ResNeXt, Inception, and Resnet50.

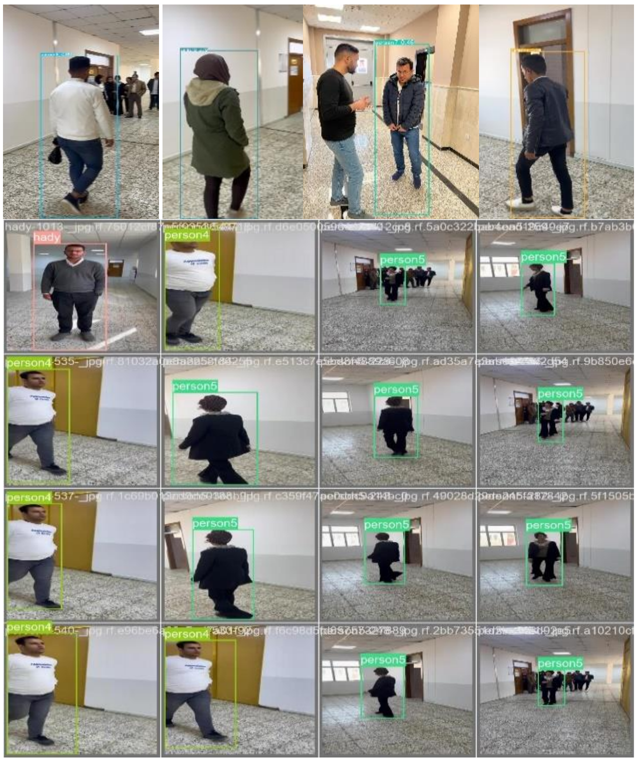

Figure 6. Visualization for images test

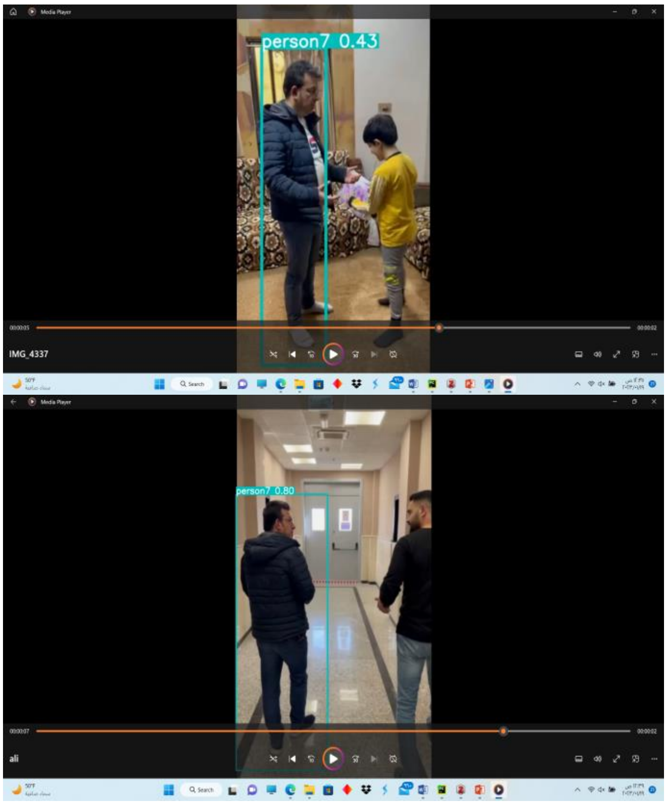

Figure 7. Visualization on video using Yolov5 framework

5.2 Results and visualizations using Yolov5

After training Yolov5 for 100 epochs to enhance the performance of re-identification people, a model of good results in re-id has resulted. In general, by using the model obtained from training for testing on the custom-created dataset and applying the testing for image, video, and real-time stream, the model can detect and re-id people that registered in the proposed dataset even without frontal view. Results and visualization are shown in Figure 6.

As shown in Figure 7 the results for the video and real-time re-id on person7 is satisfied (can recognized from the imposter) and the mean precision records very good scores (80%). The performance and results of the recall and precision are shown in Figure 8.

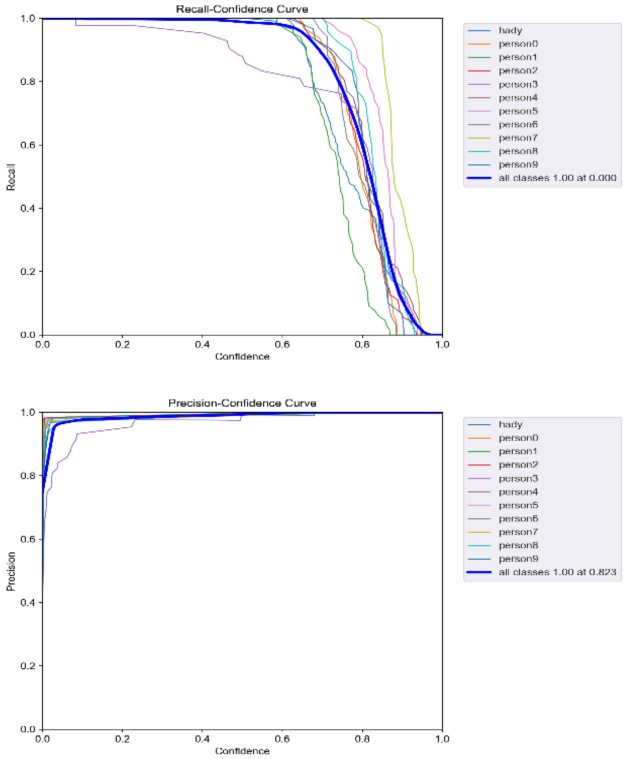

Figure 8. Precision and recall curves

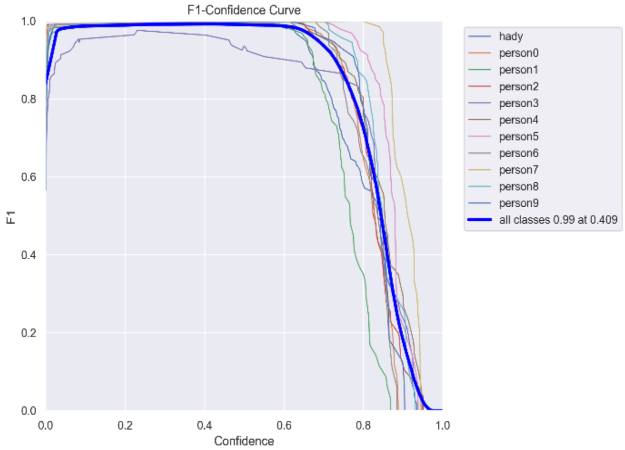

Figure 9. F1 score curve

In Figure 9, the score F1 is shown and it can be calculated from the following Eq. (1).

$$F = 2 * \left(\frac{p*R}{p+R}\right) \qquad (1)$$

where, F is the F1 score, P is the precision and R is the recall value. The mean average precision (mAp) curve is shown in Figure 10 (mAp= 0.99), as well as the complete results of Yolov5 shown in Table 4.

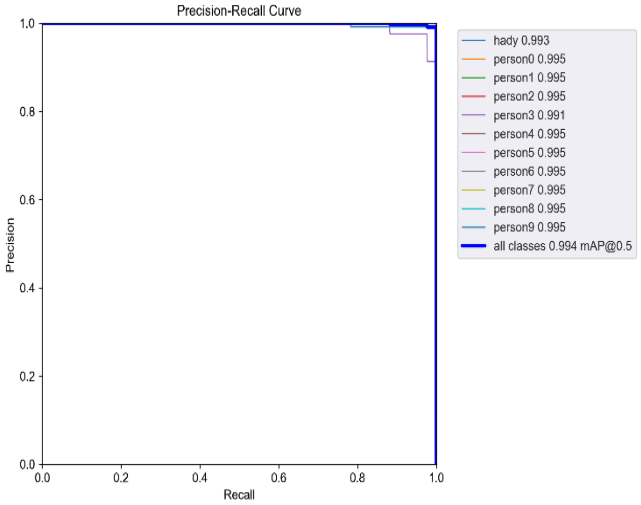

Figure 10. The curve of mAp50

Table 4. Evaluation of Yolov5 model

Classes	Precision	Recall	mAp50
Hady	0.99	1	0.99
Person0	0.99	1	0.99
Person1	0.99	1	0.99
Person2	0.99	1	0.99
Person3	0.97	0.94	0.99
Person4	0.99	1	0.99
Person5	0.99	1	0.99
Person6	0.98	1	0.99
Person7	0.98	1	0.99
Person8	0.99	1	0.99
Person9	0.98	1	0.99
All	0.99	0.99	0.99

6. CONCLUSIONS

In this paper, we present experiments of two deep learning methods to re-identify people that registered in the dataset. Two datasets are used MARS and custom-created dataset of 5910 images. The results on MARS using OSNet and different networks are compared. OSNet shows very good performance in re-id, which has reached 82.4% for rank1. OSNet produces decent re-id accuracy on MARS. It reaches 80.7% of mAp. The proposed method shows that Yolov5 can used to identify people. Yolov5 has reached 0.99 of mAp, which is an excellent result in re-id. Some of the challenges faced in the work while testing such as lightning, and bad resolution, which effect some video and real-time results. As future work, the method of using neural networks is open for enhancement such as using large data for training. The method of Yolov5 is a tough challenge, and there is room for more research. The use of the latest versions of Yolo (yolov6, yolov7, and yolov8) may lead to better results in performing the the same task.

REFERENCES

[1] Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds) Image Analysis. SCIA 2011. Lecture Notes in Computer Science, vol 6688. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21227-7_9

[2] Wang, T., Gong, S., Zhu, X., Wang, S. (2014). Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8692. Springer, Cham. https://doi.org/10.1007/978-3-319-10593-2_45

[3] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9910. Springer, Cham. https://doi.org/10.1007/978-3-319-46466-4_52

[4] Diwan, T., Anirudh, G., Tembhurne, J.V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. Multimedia Tools and Applications, 82(6): 9243-9275. https://doi.org/10.1007/s11042-022-13644-y

[5] Yap, M.H., Hachiuma, R., Alavi, A., et al. (2021). Deep learning in diabetic foot ulcers detection: A comprehensive evaluation. Computers in Biology and Medicine, 135: 104596. https://doi.org/10.1016/j.compbiomed.2021.104596

[6] McLaughlin, N., Del Rincon, J.M., Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, pp. 1325-1334. https://doi.org/10.1109/CVPR.2016.148

[7] Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X. (2017). Person Re-Identification via Recurrent Feature Aggregation. arXiv preprint arXiv:1701.06351. https://doi.org/10.48550/arXiv.1701.06351

[8] Liu, Y., Yan, J., Ouyang, W. (2017). Quality aware network for set to set recognition. arXiv preprint arXiv:1704.03373. https://doi.org/10.48550/arXiv.1704.03373

[9] Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y. (2019). Spatially and temporally efficient non-local attention network for video-based person re-identification. arXiv preprint arXiv:1908.01683. https://doi.org/10.48550/arXiv.1908.01683

[10] Li, J., Zhang, S., Huang, T. (2019). Multi-scale 3D convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 8618-8625. https://doi.org/10.1609/aaai.v33i01.33018618

[11] Subramaniam, A., Nambiar, A., Mittal, A. (2019). Co-segmentation inspired attention networks for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), pp. 562-572. https://doi.org/10.1109/ICCV.2019.00065

[12] Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., Shao, L. (2020). Learning multi-granular hypergraphs for

video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 2899-2908. https://doi.org/10.1109/CVPR42600.2020.00297

[13] Carreira, J., Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the Kinetics Dataset. arXiv preprint arXiv:1705.07750. https://doi.org/10.48550/arXiv.1705.07750

[14] Ji, S., Xu, W., Yang, M., Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231. https://doi.org/10.1109/TPAMI.2012.59

[15] Liao, X., He, L., Yang, Z., Zhang, C. (2019). Video-based person re-identification via 3D convolutional networks and non-local attention. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds) Computer Vision – ACCV 2018. ACCV 2018. Lecture Notes in Computer Science, vol. 11366. Springer, Cham. https://doi.org/10.1007/978-3-030-20876-9_39

[16] Abed, A.A., Jallod, A.A. (2022). GPU-based multiple face recognition using Yolov5x. In 2022 Iraqi International Conference on Communication and Information Technologies (IICCIT), Basrah, Iraq, pp. 298-302. https://doi.org/10.1109/IICCIT55816.2022.10010637

[17] Abed, A.A., Al-Ibadi, A., Abed, I.A. (2023). Real-time multiple face mask and fever detection using YOLOv3 and TensorFlow lite platforms. Bulletin of Electrical Engineering and Informatics, 12(2): 922-929. https://doi.org/10.11591/eei.v12i2.4227

[18] Zhao, C., Zhang, Z., Yan, J., Yan, Y. (2020). Local-global feature for video-based one-shot person re-identification. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 3662-3666. https://doi.org/10.1109/ICASSP40776.2020.9053134

[19] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. (2017). Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 5987-5995. https://doi.org/10.1109/CVPR.2017.634

[20] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[21] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[22] Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1571-1580. https://doi.org/10.1109/CVPRW50498.2020.00203

[23] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018). Path aggregation network for instance segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[24] Zhou, K., Yang, Y., Cavallaro, A., Xiang, T. (2019). Omni-scale feature learning for person re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 3701-3711. https://doi.org/10.1109/ICCV.2019.00380

[25] Reddi, S. J., Kale, S., Kumar, S. (2019). On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237. https://doi.org/10.48550/arXiv.1904.09237

[26] Zhou, K., Xiang, T. (2019). Torchreid: A library for deep learning person re-identification in pytorch. arXiv preprint arXiv:1910.10093. https://doi.org/10.48550/arXiv.1910.10093

[27] Roboflow: Give your software the power to see objects in images and video. https://roboflow.com/, accessed on Jan. 17, 2023.

[28] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[29] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308