

Framework for multiple person identification using YOLOv8 detector: a transfer learning approach

Dileep Jayaram¹, Supriya Vedagiri², Manjunath Ramachandra³

¹Department of Electronics and Communication Engineering, Sir M. Visvesvaraya Institute of Technology, Affiliated to Visvesvaraya Technological University, Bengaluru, India

²Department of Electronics and Communication Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, India

³Department of Electronics and Communication Engineering, Atria Institute of Technology, Bengaluru, India

Article Info

Article history:

Received Jul 24, 2023

Revised Nov 14, 2023

Accepted Jan 12, 2024

Keywords:

Deep learning

Person identification

Precision

Transfer learning

YOLOv8 classifier

ABSTRACT

Video surveillance extensively uses person detection and tracking technology based on video. The majority of person detection and classification techniques currently in use encounter challenges in video sequences brought on by occlusion, ambient lighting, and variations in human facial position. This paper proposed an effective person identification and classification system based on deep learning, which comprises you only look once at version 8 (YOLOv8) detection and classification model, to classify human faces in video sequences accurately. This work proposes a new staff-detection and classification (S-DEC) dataset for comprehensive performance evaluation. visual tracker benchmark (VTB) standard database is used for performance comparison with the proposed S-DEC dataset. The proposed technique achieved 98.67% precision accuracy. For the S-DEC dataset, the system gave 94.67% accuracy in identifying facial images from a video sequence of 38 people addressing the pose variation occlusion challenge. Earlier methods used to provide approximately 85% to 90% results taking more execution time. Many existing techniques were successful in detecting people only-identification of the detected person has been done in limited papers. The proposed method uses the cross-stage partial connections (CSPDarknet53) model, integrated with YOLOv8, to achieve faster results. The proposed framework took 35 minutes to train a deep learning model. A testing time of 2 minutes ensured that the proposed framework outplayed other existing methodologies and successfully identified extra information about the detected person.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Dileep Jayaram

Department of Electronics and Communication Engineering, Sir M. Visvesvaraya Institute of Technology

International Airport Road, Hunasamaranahalli, Yelahanka, Bengaluru - 562157, India

Email: dileep1721991@gmail.com

1. INTRODUCTION

One of the critical study areas of target tracking involves facial identification and tracking technology. It is now widely employed in many aspects of everyday life, including video surveillance, human-computer interaction, intelligent identification of targets, secure transportation, and healthcare diagnosis. Face identification has the benefit of being challenging to decipher and extremely private when compared to conventional digitized passwords and handwriting impersonations [1]. The human face is a naturally occurring structural target that undergoes more intricate and nuanced alterations, making face detection a problematic research topic. Furthermore, it is more challenging to identify faces when auxiliary features like a person's hair fringe, eyes, and makeup are present. The facial expression is also influenced by

the clarity of pictures of faces and the trajectory and brightness of light sources [2]. Depending upon the person's face complexity, variations in the output of the facial tracking pro exist. The following elements, in broad terms, may impact facial identification and tracking efficiency. Firstly, targets that move. The tracking frame may lose the desired face if the target face travels instantly, even beyond the border-secondly, there are Environmental consequences. Facial features are frequently influenced by similar backgrounds and fluctuating light sources, which makes face tracking more challenging. Therefore, increasing a person's face-tracking precision in complicated real-world contexts is crucial. Various deep learning (DL) based face tracking techniques have also been put forth [3]. Person identification and tracking techniques in video clips have started to use DL architecture as a hotspot for target-tracking research. Before deep learning methodologies, conventional feature extraction techniques like principal component analysis (PCA), kernel-PCA (K-PCA), fisher linear discriminant analysis (FLDA), independent component analysis (ICA), and many more were tested by researchers [2], [3]. It gave good results, but execution time was more time-consuming and limited for images. Few researchers used the above feature extractors for video and got sufficient results at the cost of execution time (nearly 1 to 2 hours for a 30-second video). Along with feature extractors, single and multi-layer classifiers are available for predicting an individual. A few examples of single-layer classifiers are probabilistic neural network (PNN) and general regression neural network (GRNN). Learning vector quantizer (LVQ) can be one example of a multi-layer neural networks. The number of hidden layers is higher in multi-layer classifiers. Apart from this, all classifiers will have input and output layers at the initial and final stages. A multi-layer classifier takes a lot of time to execute. Most of the researchers have worked for person identification in images itself. Few of them worked on videos but detected single individuals comparing from the given template. This limitation is addressed in the following sections of this paper.

The study [1] addressed the feature scaling problem. The more accurate regression network-based face tracking (RNFT) approach increases the sense of feature scaling and retrieves the rectangle structure of the intended face in the preceding frame. This method can only detect and track a single face in a video frame. Wen *et al.* [3] proposed their video database, and a new framework to classify an individual faster regions with convolutional neural networks (R-CNN) model was used to obtain accurate results for the University of Albany detection and tracking (UA-DETRAC) video database. In study [4], the facial appearances of Indian freedom warriors are predicted. Cascaded Haar classifiers were used to categorize the page system. The system was used to produce 76% to 91% recognition accuracy. The study [5] consists of a HashNet deep neural network to extract deep hash appearance features of pedestrians. Object occlusion is eliminated due to the inclusion of feature fusion techniques. The drawback of this architecture is that it has a complex network structure. Deep vision neural network [6] is implemented to verify an anomaly based on the period, repeated number, and elapsed eye blink time. Paper [7] includes features such as grey, luminance, and luminance relative extracted from images. But the disadvantage is the execution time. Hammad *et al.* [8] proposed a method that includes two deep neural networks for identifying the individual and physical activity. This methodology requires 30% less area than an architecture involving two different deep neural networks. Other papers from study [9], [10] presented their methodologies on convolutional neural networks. It gives a significant idea of using a convolutional neural network to identify a person in a video frame or image.

Cao *et al.* [11] proposed a paper on a 2D pose estimation-based person identification technique. This includes extracting several angles of a moving person. This technique provided an error rate concerning classification-papers [12]–[14] compared different neural networks concerning person detection. Vinay *et al.* [15] identified a method that classifies the person efficiently from a video database. He has used the Dominant feature of two self-contained convolutional neural networks. The classifier gave a maximum efficiency of 91% for 100 epochs. Yu *et al.* [16] addressed the complexity of identifying relevant movement caused by various objects of interest, such as persons, animals, or vehicles. Relevant motion event detection network (ReMotENet), a three-dimensional convolutional neural network, was implemented to increase the speed. This methodology is efficient and lightweight. Zhou *et al.* [17] introduced recurrent neural networks. This network makes use of metric and feature-based techniques to re-identify an individual. The study's [18] goal was to guess an accurate number of objects and group them in a low-resolution frame. The study [19] used deep neural network concepts to enhance the efficiency of universal object identifiers on video frames to discover a few characteristics. The study's [20] goal was to identifying faces in violent snapshots. The violent flow descriptor was used for violence identification. The study [21] used a sparse classification algorithm, which includes two minimization methods. Papers [22]–[25] had been developed on convolutional network algorithms to recognize a person/object. Papers [26] and [27] have been published on low short-term memory classifiers, which can be applied to a video frame or an image to get effective results.

Visual tracker benchmark (VTB) database has been used as a standard database [28] to compare the results with the newly created staff-detection and classification (S-DEC) database. Convolutional networks and random sample convolutions were employed in study [29] to detect the person. Although it provided a

maximum accuracy of 98%, the dataset's sample size was small. Residual deep neural networks were employed [30] to identify that person. The person's body was equipped with accelerometer, magnetometer, and gyroscope sensors in 5 separate places. Various bodily actions were used to identify individuals. The proposed method was applied to the residual network (ResNet) V1 and V2 samples. For a minimal activity, 94% success was achieved. A growth in the number of activities revealed a research deficit. The study [31] suggested multi-scale deep supervision using attention characteristic learning deep model. The MSMT17 sample and the DukeMTMC-ReID dataset were utilized to obtain a mean average precision of 89%. Most recent studies have focused on designing discriminative global features while ignoring local and salient variables related to this problem. For the Market-1501 dataset, Hu *et al.* [32] described individual reidentification-identification utilizing a deep batch active learning. The author got 86% accuracy but was restricted to mere pictures. This suggested approach cannot process videos. The labeled dataset

WL-DukeMTMC-ReID was used in [33]. The deep graph metric learning method was used for a maximum mean average precision of 90%. The limitation is that noisy clips cannot be processed properly. In [34], the author presented a review paper that consists of convolutional and deep neural networks for identifying an individual. The ResNet-50 model was used by [35] along with the multi-task part-aware network (MPN), designed to extract semantically aligned part-level features from pedestrian images. Market-1501, DukeMTMC-ReID, and CUHK03 databases achieved mean average precisions of 89%, 82%, and 81%, respectively. Research gaps were: i) the time and space complexities of MPN are slightly increased and ii) it was applied to images only.

Pedestrian feature identification was developed using the you only look once version 4 (YOLOv4) deep learning network and hybrid transformers [36]. Richly annotated pedestrian version 2 (RapV2), RapV1, pedestrian attribute (PETA), and PA100K databases were used to obtain mean accuracy of 79%, 81%, 86% and 84% respectively. It used to take 46 minutes to produce the results. This was the limitation in [36]. Wang *et al.* [37] proposed a person re-recognition method with a multi-grain size generative adversarial network considering pedestrian images. The author addressed the problem of occlusion. The Market-1501 and DukeMTMC-ReID datasets achieved mean accuracy of 91% and 88%, respectively. It was applied only to images. The author of paper [38] describes a multi-scale pyramid attention (MSPA) model for P-ReID that changes how well semantic attributes and visual appearance work together. The Market-1501 and DukeMTMC-ReID datasets achieved maximum accuracy of 96% and 97%, respectively. The proposed framework was designed for short-term reidentification-identification only and applied to images. The papers mentioned above were studied, and Table 1 shows the comparison and illustrates the research gap in the chosen area. It was observed that many researchers proposed their methodology on images by using different datasets. Few of them used video datasets, which had many limitations: i) more execution time, ii) a single face can only be detected, iii) some algorithms only detect different faces, no identification, iv) few algorithms concentrate only on moving objects neglecting static objects, and v) few papers do not work on challenges parts such as pose variation, illumination, and occlusion. The main objective of this work includes i) reducing execution time, ii) detecting and identifying multiple faces from a video frame, iii) both static and moving persons should be identified with additional information called metadata, and iv) challenges such as pose variation (>30 degrees) and illumination (>40%) should be addressed.

Table 1. Research gap analysis

Paper Reference	Methodology used	Pose variation limit	Illumination/Occlusion limit in %	Precision in %	Execution time	Limitations
[1]	Regression network-based face-tracking model	30 deg	Illumination: 40	89.2	50 minutes for 100 epochs	Single face in a video frame
[3]	Regions with convolutional neural networks	25 deg	---	91	40 minutes for 100 epochs	It does not address the illumination problem
[5]	HashNet deep neural network	15 deg	Illumination: 40	72.8	---	Complex structure, low precision
[15]	Convolutional neural network	20 deg	---	90	---	Illumination and pose problem
[16]	3D convolutional neural network	15 deg	---	87	56 minutes for 100 epochs	Mis-classification/error percentage was more
[17]	Joint spatial and temporal recurrent neural networks	---	Illumination: 20	90	42 minutes for 100 epochs	Concentrates only on moving objects
[18]	Expectation maximization	---	---	90	4 hours for 100 epochs	It detects people does not identify
[21]	Regularized sparse representation classification algorithm	25 deg	---	65.56	---	Less precision
[38]	MSPA model	---	Illumination: 30	97	---	Applied only for images

Figure 3 shows the YOLOv8 backbone unit. The foundation of YOLOv8 is a modified version of the CSPDarknet53 framework. Fifty-three convolutional layers make up this design, using cross-stage restricted correlations to enhance information transfer between the various layers. Convolutional layers consist of a set of Conv2d (2-dimensional) functions and a 2-dimensional batch normalization function for reducing the dimensionality of the considered frame and extracting facial features. Sigmoid linear unit (SiLU) is an activation function calculated by multiplying the sigmoid function with input.

$$S = \varphi(x) \quad (1)$$

where, x = Input, φ = Transfer function symbol, S = Activation value

$$S_y = \varphi_y \sum_{i=0}^n W_{i-y} S_i \quad (2)$$

here, S_y = Final activation signal, W_{i-y} = Weighted signal

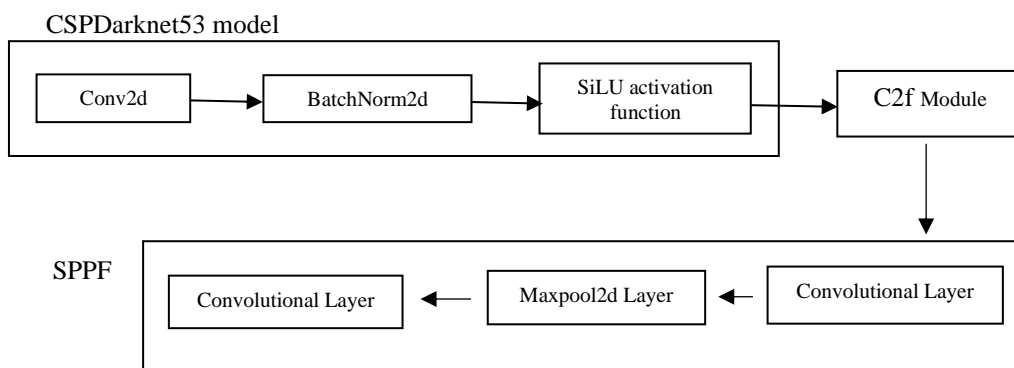


Figure 3. YOLOv8 Backbone unit

Two independent flows of gradient streams are integrated into the C2f module's design, enabling a more reliable trajectory data movement. Spatial pyramid pooling fast (SPPF) is used in the YOLOv8 structure. By optimizing the pooling procedure, SPPF prevented SPP from having to run repeatedly and considerably increased the module's operating speed. SPPF consists of a set of convolutional layers, 2-dimensional max-pooling layers concatenated with a final convolutional layer to increase the operating speed of the framework.

Figure 4 shows the YOLOv8 head unit. A number of convolutional layers and a string of fully connected layers make up the head of the YOLOv8 algorithm. For the elements discovered in a picture, these layers are in charge of estimating box boundaries, objectless ratings, and likelihoods of classes. Using a self-attention option in the network's brain is one of YOLOv8's distinguishing characteristics. Finally, the detect function consists of a set of convolutional layers (both 1 and 2-dimensional) to compute bounding box and class losses. Binary cross-entropy (BCE) loss calculates the classification work failure. In other words, the error in determining whether an object is present in a specific grid cell or not is calculated using a binary cross-entropy loss. Complete intersection over union (IoU) loss quantifies the mistake in object localization within the grid cell. The predicted output was stored in the runs folder after passing train and test data through the Backbone and Head unit.

2.3. Training phase flow chart

Figure 5 shows the flowchart for the training phase. Stored video of different frame rates can be applied as training input to the proposed system. Video frames can be extracted from 24fps, 30fps, and 60fps video. Labelled data is created for all the faces considering x, y, length, and width parameters available from the converted frame. The labeled data file is stored as a notepad file in the " data " folder. Class information is provided through the YAML file after installing all the packages required to run the program in the Python environment. Convolutional layers are essential to map required information from the labelled data and extract the required feature matrix the YOLOv8 classifier used to identify a given face with the mapped data. Classified weights had to be saved for comparison with testing data in the next phase.

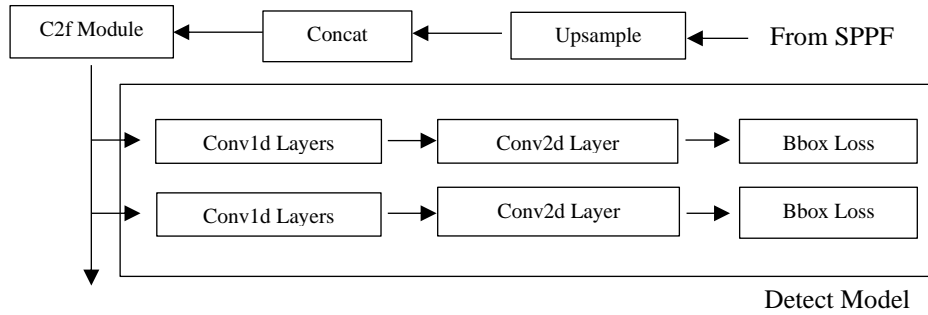


Figure 4. YOLOv8 head unit

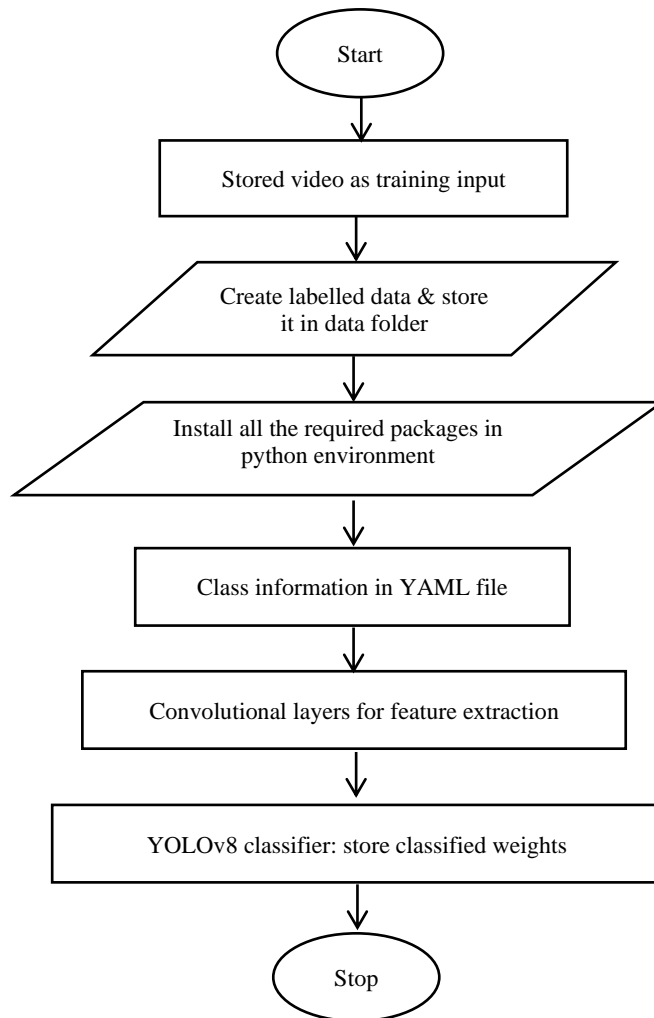


Figure 5. Training phase flow chart

2.4. Testing phase flow chart

Figure 6 shows the flowchart for the testing phase. Real-time or stored video of different frame rates can be applied as an input in the testing phase. Converted video frames and class data are applied to the feature extractor to detect faces. If the detected face is classified and authenticated as a particular individual from the training phase, the detected face is classified and authenticated as a particular individual. In the output window, the metadata file is given as input to provide additional information about the classified people. Tracked videos can be saved and used in the future. If the detected face vectors do not match stored weights, they exit the testing phase. Note that this process continues until all the frames complete its execution.

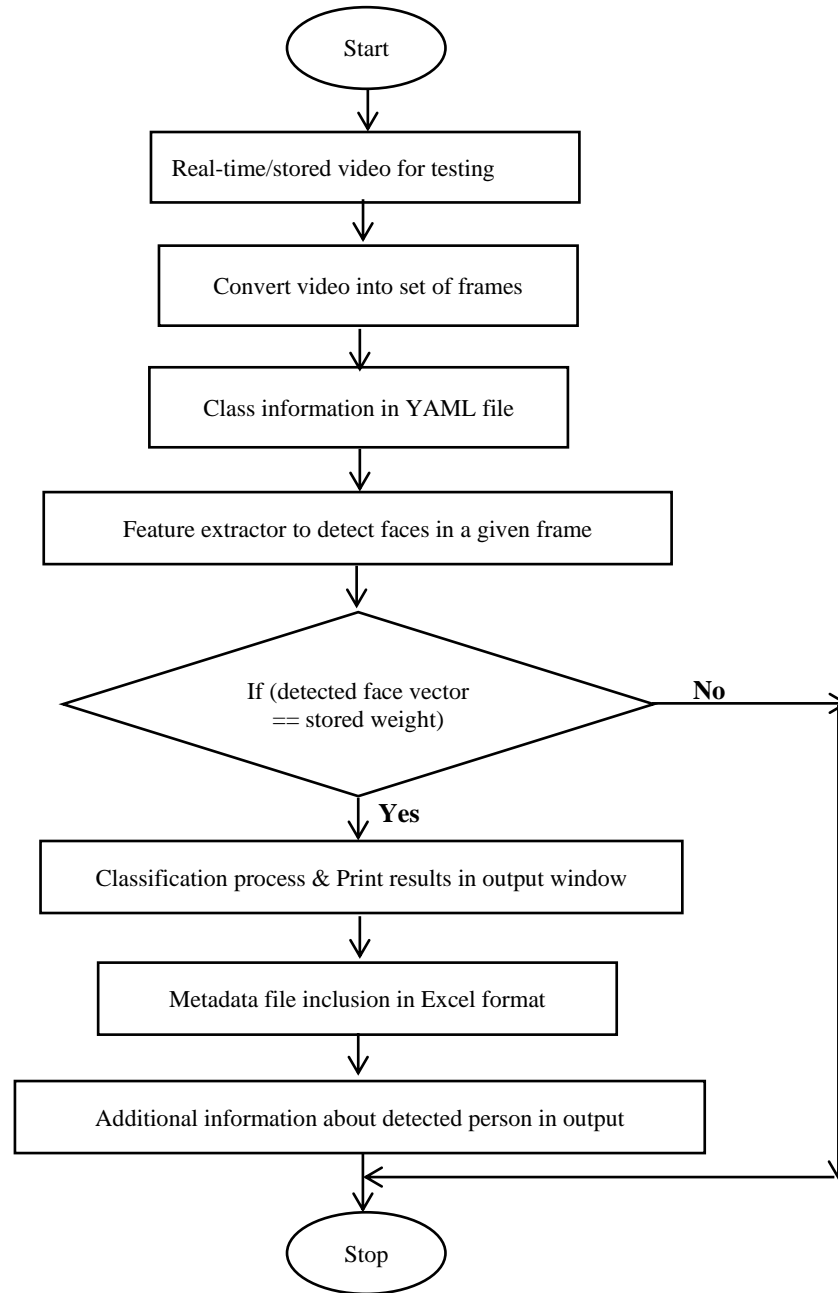


Figure 6. Testing phase flow chart

3. RESULTS AND DISCUSSION

The performance of any deep learning neural network can be measured by computing precision, Recall, and F1-Score-these performance metrics depend on true positive, false positive, and false negative values. S-DEC database consists of 38 people video datasets. Videos were taken in varying backgrounds, different poses, and various expressions. 50 video files were considered, which consist of multiple people. Different frame rate video files were considered (24fps, 30fps, and 60fps). The VTB database [28] consists of multiple-person videos taken in different backgrounds, with varying expressions and poses.

3.1. Evaluation metrics

True positive (TP) is where the algorithm correctly identifies the existing face. False negative (FN) is a state where the algorithm offers a zero-confidence estimate of an actual face. False positive (FP) occurs when a device inaccurately estimates the existence of a face at an extremely high trust level. Precision measures the percentage of cases among those the model identified as positive that are actual positive

examples. Equation (3) shows the computational metric for precision. Recall, commonly referred to as sensitivity, is the percentage of positive examples among all positive examples. Equation (4) shows the designated formula for the recall metric.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

Figures 7(a) and 7(b) shows the precision metric for VTB and S-DEC databases, respectively. The graph was plotted by considering the number of epochs as 100 on the x-axis and precision (%) on the y-axis. Different frame rates, such as 24fps, 30fps, and 60fps, were considered in the plot. 91.42%, 93.58%, and 94.67% precision accuracy were obtained for 24fps, 30fps, and 60fps of S-DEC database, respectively. For the VTB database, 92.28%, 95.43%, and 98.67% precision rates were achieved for 24fps, 30fps, and 60fps, respectively.

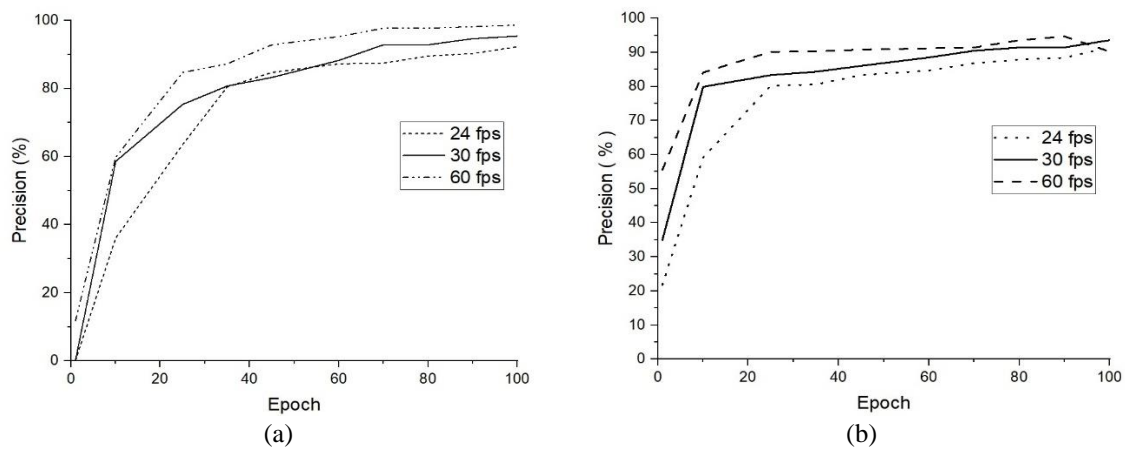


Figure 7. Precision metrics, (a) VTB database and (b) S-DEC database

Figures 8(a) and 8(b) shows the Recall metric for VTB and S-DEC databases, respectively. The graph was plotted by considering the number of epochs as 100 on the x-axis and recall on the y-axis. Different frame rates such as 24fps, 30fps and 60fps were considered in the plot. 82.36%, 84.02%, and 88.09% of recall accuracy were obtained for 24fps, 30fps, and 60fps of the S-DEC database, respectively. For the VTB database, 94.36%, 96.88%, and 100% recall rates were achieved for 24fps, 30fps, and 60fps, respectively.

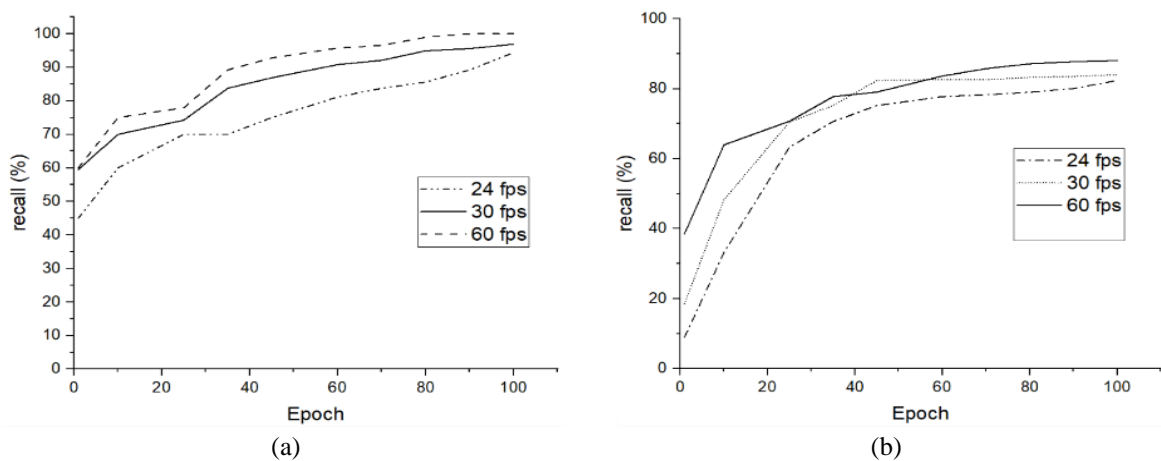


Figure 8. Recall metrics, (a) VTB database and (b) S-DEC database

The F1-Score can be described as the Harmonic median of the framework's precision and recall. It is a gauge of the model's consistency presented in (5). Figures 9(a) and 9(b) shows the F1-Score metric for VTB and S-DEC databases, respectively. This metric should be "As close to ONE." Practically, more than 0.9 is considered as the best output. The X-axis consists of the number of epochs ranging from 1 to 100. The Y-axis includes the des F1-Score metric ranges from 0 to 1. Different frame rate videos were considered to measure. The proposed framework produces 0.91 of the F1-Score for the S-DEC database and 0.99 for the F1-Score for the VTB database.

$$F1 - Score = \frac{2}{\left(\frac{1}{Precision}\right) + \left(\frac{1}{Recall}\right)} \quad (5)$$

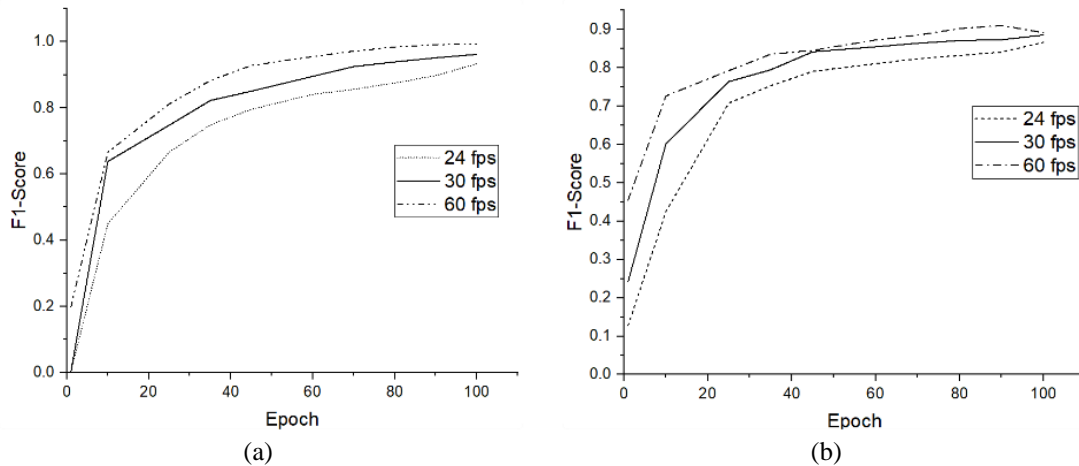


Figure 9. F1-Score metrics, (a) VTB database and (b) S-DEC database

Figure 10 shows the predicted output for the S-DEC database with 24fps in Figure 10(a), 30fps in Figure 10(b), and 60fps in Figure 10(c). Different frame rate videos were analyzed using the proposed technique, yielding the best result in 60 fps video. All 6 people in the S-DEC dataset of 60fps video frame were identified successfully and displayed additional information related to authenticated people on the output window. It was observed that in the 24fps video, one of the faces remained unidentified, yielding less class precision. Even though the noise was present in 24fps video, the proposed algorithm gave an acceptable result.

Figure 11 shows the predicted output for the VTB database with 24fps in Figure 11(a), 30fps in Figure 11(b), and 60fps in Figure 11(c). Regarding 30fps and 60fps videos, the proposed framework produced an excellent number of accurate detections by considering occlusion, various facial expressions, variation of light, and different poses. For the VTB database, considering color and black and white video as input in different frame rates, the system obtained the best results in detecting and identifying faces. In the 60fps video frame, a significant amount of noise was present. Even then, the proposed system had identified the person effectively, yielding 99% accuracy.

All the above said performance metrics were tabulated for different frame rates and different datasets. Table 2 shows the resultant table for the S-DEC database considering the number of epochs to be trained, precision, mean average precision, recall, and F1-Score of 3 different frame rates. Precision and recall metrics should be considered to draw the qualitative measure of the designed algorithm. A mean average precision of 93% was achieved from the proposed framework. Table 3 shows computations for the VTB database. It gave a mean average precision of 99.5% for 60fps video. In many papers, convolutional neural networks are used as feature extractors, and a separate classifier was used to get the required outcome. In my proposed work, a modified CSPDarknet53 framework with the YOLOv8 has been used to address occlusion and variation of light and pose variation challenges. Further, the illumination and background clutter problems have been addressed and tabulated in Table 4, and execution time has been given greater importance in addition to accuracy. There was a trade-off between accuracy and execution time in many papers, shown in Table 4. Pose variation challenge is addressed up to 45 degrees with respect to any person. 98.67% of best precision was obtained for a video of 50 seconds from this framework, which takes 35 minutes of training time and 2 minutes of testing time, having 38 total classes.



Figure 10. Predicted video output of S-DEC database: (a) 24fps, (b) 30fps, and (c) 60fps

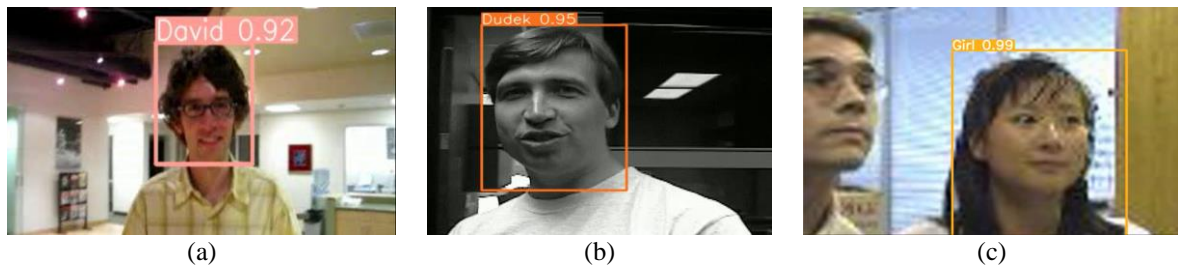


Figure 11. Predicted video output of VTB Database: (a) 24fps, (b) 30fps, and (c) 60fps

Table 2. Result in comparison of S-DEC database with different frame rates

Epoch	24fps				30fps				60fps			
	Precision (%)	mAP 50(%)	Recall (%)	F1-Score (0-1)	Precision (%)	mAP 50(%)	Recall (%)	F1-Score (0-1)	Precision (%)	mAP 50(%)	Recall (%)	F1-Score (0-1)
1	21.80	0.29	9.04	0.13	34.94	9.03	18.56	0.24	55.61	25.77	38.54	0.46
10	59.13	38.35	33.17	0.43	79.91	50.41	48.26	0.60	84.10	60.85	63.95	0.73
25	80.26	74.28	63.33	0.71	83.32	79.09	70.52	0.76	90.19	80.35	70.71	0.79
35	80.63	82.94	70.66	0.75	84.32	85.69	75.20	0.79	90.29	87.26	77.76	0.84
45	83.37	85.69	75.20	0.79	86.04	89.30	82.36	0.84	90.84	90.83	79.05	0.85
60	84.70	87.26	77.76	0.81	88.51	89.42	82.66	0.85	91.17	91.66	83.64	0.87
70	86.83	87.29	78.28	0.82	90.50	90.44	82.66	0.86	91.43	92.35	85.76	0.89
80	87.89	89.58	79.02	0.83	91.42	90.92	83.21	0.87	93.58	92.57	87.15	0.90
90	88.46	89.86	80.12	0.84	91.46	91.13	83.53	0.87	94.67	93.13	87.72	0.91
100	91.42	90.83	82.36	0.87	93.58	92.18	84.02	0.89	90.23	93.28	88.09	0.89

Table 3. Result in comparison of VTB database with different frame rates

Epoch	24fps				30fps				60fps			
	Precision (%)	mAP 50(%)	Recall (%)	F1-Score (0-1)	Precision (%)	mAP 50(%)	Recall (%)	F1-Score (0-1)	Precision (%)	mAP 50(%)	Recall (%)	F1-Score (0-1)
1	0.16	0.24	45.00	0.00	0.20	0.49	59.50	0.00	11.87	14.66	60.00	0.20
10	36.11	50.58	60.00	0.45	58.60	56.09	70.00	0.64	59.89	60.43	75.00	0.67
25	63.56	75.17	70.00	0.67	75.35	93.88	74.24	0.75	84.74	93.02	77.97	0.81
35	80.51	84.28	70.00	0.75	80.76	94.27	83.76	0.82	87.29	93.88	89.21	0.88
45	84.74	92.72	75.00	0.80	83.29	94.27	86.88	0.85	92.87	94.27	92.80	0.93
60	87.29	92.72	81.14	0.84	88.29	95.00	90.80	0.90	95.27	95.00	95.74	0.96
70	87.49	93.88	83.76	0.86	92.87	95.00	92.06	0.92	97.75	99.50	96.54	0.97
80	89.56	94.27	85.59	0.88	92.87	96.58	94.94	0.94	97.75	99.50	98.97	0.98
90	90.29	94.27	89.21	0.90	94.65	98.07	95.59	0.95	98.20	99.50	100.00	0.99
100	92.28	95.00	94.36	0.93	95.43	98.07	96.88	0.96	98.67	99.50	100.00	0.99

Table 4. Result in comparison with proposed framework

Paper reference	Methodology used	Pose variation limit	Illumination/Occlusion limit in %	Precision in %	Execution time
[1]	Regression network-based face-tracking model	30 deg	Illumination: 40	89.2	50 minutes for 100 epochs
[3]	Regions with convolutional neural networks	25 deg	-----	91	40 minutes for 100 epochs
[5]	HashNet deep neural network	15 deg	Illumination: 40	72.8	---
[15]	Convolutional neural network	20 deg	---	90	---
[16]	3D convolutional neural network	15 deg	---	87	56 minutes for 100 epochs
[17]	Joint spatial and temporal recurrent neural networks	---	Illumination: 20	90	42 minutes for 100 epochs
[18]	Expectation maximization	---	---	90	4 hours for 100 epochs
[21]	Regularized sparse representation classification algorithm	25 deg	---	65.56	---
[38]	MSPA model	---	Illumination: 30	97	---
Proposed framework	Modified CSPDarknet53 method with YOLOv8	45 deg	Illumination: 40	98.67	37 minutes for 100 epochs

4. CONCLUSION

This study recommended the use of the modified CSPDarknet53 person identification and tracking as part of an effective video person identification and tracking strategy. Initially, issues such as ambient light variations, alterations in pose, and facial expressions related to person identification and tracking were discussed. CSPDarknet53 was used to determine the desired face's location in each frame during face tracking, and the YOLOv8 framework was utilized to determine the tracking pattern between subsequent frames. Considering the most challenging parameters like pose fluctuation, clutters, and changes in the background, this model guarantees consistency in delivering high-quality outcomes. To evaluate the suggested CSPDarknet53 frameworks with the related approaches in terms of precision and execution time, the VTB database was used as a standard dataset. The s-DEC database had been created and applied to the proposed model. After comparing both the database results, it was observed that 95% of similar results were achieved by addressing various challenges discussed above. Even though the pose variation is greater than 45 degrees and dissimilar lighting conditions, the system achieved 98.67% of best precision for a video of 50 seconds, which takes 35 minutes of training and 2 minutes of testing time, with 38 total classes for 100 epochs. In the future, the same system shall be modified to classify all the people present in 24fps video while addressing the false negative problem from Figure 10(a). It was observed that the system produced better results as the number of training epochs and frame rate increased. Even precision deteriorates if the frame rate is less, and false negative problems increase in-person identification. Some false negatives may be corrected by considering information from adjacent frames. Implementing tracking algorithms or using advanced, recurrent neural networks (RNNs) can help improve detection.

REFERENCES




- [1] G. Zheng and Y. Xu, "Efficient face detection and tracking in video sequences based on deep learning," *Information Sciences*, vol. 568, pp. 265–285, Aug. 2021, doi: 10.1016/j.ins.2021.03.027.
- [2] V. Pandimurugan, A. Jain, and Y. Sinha, "IoT based face recognition for smart applications using machine learning," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020, pp. 1263–1266, doi: 10.1109/ICISS49785.2020.9316089.

- [3] L. Wen *et al.*, "UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, Apr. 2020, doi: 10.1016/j.cviu.2020.102907.
- [4] V. Perumal, "Face recognition in video streams and its application in freedom fighters discovery - a machine learning approach," in *2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, Dec. 2020, pp. 1–5, doi: 10.1109/ICMLANT50963.2020.9355979.
- [5] Z. Ding, S. Liu, M. Li, Z. Lian, and H. Xu, "A Blockchain-enabled multiple object tracking for unmanned system with deep Hash appearance feature," *IEEE Access*, vol. 9, pp. 1116–1123, 2021, doi: 10.1109/ACCESS.2020.3046243.
- [6] T. Jung, S. Kim, and K. Kim, "DeepVision: deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020, doi: 10.1109/ACCESS.2020.2988660.
- [7] S. Thepade, P. Jagdale, A. Bhingurde, and S. Erandole, "Novel face liveness detection using fusion of features and machine learning classifiers," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, Feb. 2020, pp. 141–145, doi: 10.1109/ICIoT48696.2020.9089525.
- [8] I. Hammad and K. El-Sankary, "Using machine learning for person identification through physical activities," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Oct. 2020, pp. 1–5, doi: 10.1109/ISCAS45731.2020.9181231.
- [9] W. Rao, M. Xu, and J. Zhou, "Improved metric learning algorithm for person re-identification based on asymmetric metric," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Jun. 2020, pp. 212–216, doi: 10.1109/ICAICA50127.2020.9181918.
- [10] A. Rasool, S. Roohi, and A. Majumder, "Automated attendance system using few-shot learning approach," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Jun. 2020, pp. 43–46, doi: 10.1109/ICRITO48877.2020.9197814.
- [11] G. Cao, Y. Pu, Y. Li, and Z. Zhao, "Human motion capture using a multi-2D pose estimation model," in *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Aug. 2019, pp. 64–67, doi: 10.1109/IHMSC.2019.00023.
- [12] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "A survey on deep learning-based person re-identification systems," *IEEE Access*, vol. 7, pp. 175228–175247, 2019, doi: 10.1109/ACCESS.2019.2957336.
- [13] N. Narayan, N. Sankaran, S. Setlur, and V. Govindaraju, "Re-identification for online person tracking by modeling space-time continuum," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 1519–151909, doi: 10.1109/CVPRW.2018.00193.
- [14] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in *2017 36th Chinese Control Conference (CCC)*, Jul. 2017, pp. 11104–11109, doi: 10.23919/ChiCC.2017.8029130.
- [15] A. Vinay, D. A. Mundry, G. Kathiresan, U. Sridhar, K. N. B. Murthy, and S. Natarajan, "Dominant feature based convolutional neural network for faces in videos," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, Mar. 2017, pp. 17–22, doi: 10.1109/ICBDACI.2017.8070802.
- [16] R. Yu, H. Wang, and L. S. Davis, "ReMotENet: efficient relevant motion event detection for large-scale home surveillance videos," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 1642–1651, doi: 10.1109/WACV.2018.00183.
- [17] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6776–6785, doi: 10.1109/CVPR.2017.717.
- [18] Y.-L. Hou and G. K. H. Pang, "People counting and human detection in a challenging situation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 1, pp. 24–33, Jan. 2011, doi: 10.1109/TSMCA.2010.2064299.
- [19] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 1650–1657, doi: 10.1109/CVPR.2013.216.
- [20] V. E. Machaca Arceda, K. M. Fernández Fabián, P. C. Laguna Laura, J. J. Rivera Tito, and J. C. Gutiérrez Cáceres, "Fast face detection in violent video scenes," *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 5–26, Dec. 2016, doi: 10.1016/j.entcs.2016.12.002.
- [21] S. Nagendra, R. Baskaran, and S. Abirami, "Video-based face recognition and face-tracking using sparse representation based categorization," *Procedia Computer Science*, vol. 54, pp. 746–755, 2015, doi: 10.1016/j.procs.2015.06.088.
- [22] D. Chahyati, M. I. Fanany, and A. M. Arymurthy, "Tracking people by detection using CNN features," *Procedia Computer Science*, vol. 124, pp. 167–172, 2017, doi: 10.1016/j.procs.2017.12.143.
- [23] Y. Li, R. Ge, Y. Ji, S. Gong, and C. Liu, "Trajectory-pooled spatial-temporal architecture of deep convolutional neural networks for video event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2683–2692, Sep. 2019, doi: 10.1109/TCSVT.2017.2759299.
- [24] Bong-Nam Kang, Yonghyun Kim, and D. Kim, "Deep convolution neural network with stacks of multi-scale convolutional layer block using triplet of faces for face recognition in the wild," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2016, pp. 004460–004465, doi: 10.1109/SMC.2016.7844934.
- [25] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Jun. 2016, pp. 1–6, doi: 10.1109/ICIS.2016.7550888.
- [26] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018, doi: 10.1109/TGRS.2018.2848473.
- [27] L. Zhang and J. Zhang, "Synchronous prediction of arousal and valence using LSTM network for affective video content analysis," in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Jul. 2017, pp. 727–732, doi: 10.1109/FSKD.2017.8393364.
- [28] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: a Benchmark," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 2411–2418, doi: 10.1109/CVPR.2013.312.
- [29] M. R. Desai, S. A. Patel, M. Peerzade, and G. Chawhan, "Person re-identification via deep metric learning," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAIECC)*, Dec. 2020, pp. 1–9, doi: 10.1109/ICAIECC50550.2020.9339491.
- [30] S. Kilic, I. Askerzade, and Y. Kaya, "Using ResNet transfer deep learning methods in person identification according to physical actions," *IEEE Access*, vol. 8, pp. 220364–220373, 2020, doi: 10.1109/ACCESS.2020.3040649.
- [31] D. Wu, C. Wang, Y. Wu, Q.-C. Wang, and D.-S. Huang, "Attention deep model with multi-scale deep supervision for person re-identification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 1, pp. 70–78, Feb. 2021, doi: 10.1109/TETCI.2020.3034606.




- [32] Z. Hu, W. Hou, and X. Liu, "Deep batch active learning and knowledge distillation for person re-identification," *IEEE Sensors Journal*, vol. 22, no. 14, pp. 14347–14355, Jul. 2022, doi: 10.1109/JSEN.2022.3181238.
- [33] J. Meng, W.-S. Zheng, J.-H. Lai, and L. Wang, "Deep graph metric learning for weakly supervised person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6074–6093, Oct. 2022, doi: 10.1109/TPAMI.2021.3084613.
- [34] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022, doi: 10.1109/TPAMI.2021.3054775.
- [35] C. Ding, K. Wang, P. Wang, and D. Tao, "Multi-task learning with coarse priors for robust part-aware person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1474–1488, Mar. 2022, doi: 10.1109/TPAMI.2020.3024900.
- [36] S. Raghavendra, Ramyashree, S. K. Abhilash, V. M. Nookala, and S. Kaliraj, "Efficient deep learning approach to recognize person attributes by using hybrid transformers for surveillance scenarios," *IEEE Access*, vol. 11, pp. 10881–10893, 2023, doi: 10.1109/ACCESS.2023.3241334.
- [37] Y. Wang, Y. Sun, Z. Lan, F. Sun, N. Zhang, and Y. Wang, "Occluded person re-identification by multi-granularity generation adversarial network," *IEEE Access*, vol. 11, pp. 59612–59620, 2023, doi: 10.1109/ACCESS.2023.3285798.
- [38] S. U. Khan *et al.*, "Visual appearance and soft biometrics fusion for person re-identification using deep learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 575–586, May 2023, doi: 10.1109/JSTSP.2023.3260627.

BIOGRAPHIES OF AUTHORS






Dileep Jayaram    received M.Tech degree in 2015 from PESIT, Bengaluru, India. He is pursuing a Ph.D. degree at the Department of Electronics and Communication Engineering, Sir. M. Visvesvaraya Institute of Technology, Bengaluru, Affiliated to VTU, Belagavi. Currently, he works as an assistant professor in the Department of Electronics and Communication Engineering at Kammavari Sangham School of Engineering and Management, Bengaluru. His research areas include image processing, machine learning, deep learning, and computer vision. He can be contacted at email: dileep1721991@gmail.com.



Supriya Vedagiri    received Ph.D. degree in 2016 from Jain University, Bengaluru, India. Currently, she works as a Professor and Head of the Department of ECE, Sir. M. Visvesvaraya Institute of Technology, Bengaluru. She has about 34 years of work experience in the field of communication and image processing-including medical image processing and wireless/mobile networking. She has 20 journal and conference publications. She worked as Principal of GTTC Women Polytechnic for 10 years and started Mechatronics Training Centre. Her areas of interest include signal and image processing, cryptography, and renewable energy. She can be contacted at email: hod_ece@sirmvit.edu.



Manjunath Ramachandra    received Ph.D. degree in 2007 from Bangalore University. He has 24 years of work experience in the overlapping verticals of signal processing. He published 199 journal and conference publications, patent disclosures, and a book. He represented Philips in international standardization bodies such as the Wi-Fi Alliance, served as the editor for the regional profiles standard in the digital living network alliance (DLNA), and as the industrial liaison officer for the CE-Linux Forum. He has chaired about 30 conferences. His areas of interest include signal and image processing database architecture. He can be contacted at email: drmanjunathramachandra@gmail.com.