Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338
Volume 12 Issue 01, 2023

Deep Learning Techniques for Video Surveillance and Activity Recognition

Akash Verma[1], Maria Gonzalez[2]

[1]*Blue Ridge Institute of Technology, akash.verma@blueridge.tech*

[2]*Highland Technical University, maria.gonzalez@highlandtech.ac*

Peer Review Information

Submission: 21 Feb 2023
Revision: 17 April 2023
Acceptance: 15 May 2023

Keywords

Activity Recognition
Anomaly Detection
LSTM
3D CNNs

Abstract

Video surveillance and activity recognition have become critical components in modern security systems, leveraging the advancements in deep learning to enhance the accuracy and efficiency of monitoring. This paper explores the application of deep learning techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid models, for human activity recognition in video surveillance. These models are particularly well-suited for capturing spatial and temporal features of video data, enabling systems to recognize complex human actions, detect anomalies, and classify behaviors in real-time. The paper discusses the challenges associated with video surveillance, such as occlusions, diverse environments, and large-scale datasets, and presents solutions offered by deep learning methods, including 3D CNNs and Long Short-Term Memory (LSTM) networks. Additionally, it highlights the role of transfer learning and data augmentation in improving model performance. Finally, the paper looks at future trends in the field, such as the integration of reinforcement learning for proactive anomaly detection and the potential of unsupervised learning techniques. These innovations promise to make video surveillance systems more intelligent, responsive, and scalable, paving the way for smarter security solutions in a variety of applications.

Introduction

In recent years, deep learning techniques have revolutionized the field of video surveillance and activity recognition, enabling automated systems to perform complex tasks such as human action detection, anomaly detection, and real-time event analysis with greater accuracy and efficiency. Video surveillance systems, traditionally dependent on manual monitoring, have increasingly incorporated deep learning methods to enhance their ability to process large volumes of data and recognize activities across diverse environments. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely adopted for extracting spatial and temporal features from video data, facilitating the recognition of complex human behaviors and interactions [1]. Moreover, the use of 3D CNNs and

Long Short-Term Memory (LSTM) networks has shown significant promise in capturing motion and context over time, improving the robustness and scalability of activity recognition systems [2,3].

Deep learning approaches also address critical challenges inherent in video surveillance, such as occlusions, varying lighting conditions, and background clutter. By leveraging large-scale annotated datasets, deep learning models can be trained to detect rare or anomalous activities that would otherwise go unnoticed [4]. Additionally, the integration of unsupervised learning and reinforcement learning techniques is beginning to offer new avenues for developing adaptive surveillance systems capable of learning and responding to dynamic environments [5].

The purpose of this paper is to provide an overview of deep learning techniques applied to video surveillance and activity recognition, discussing the key methodologies, challenges, and recent advancements in the field. Through this exploration, we aim to demonstrate how these techniques are transforming surveillance systems from reactive tools to proactive, intelligent platforms capable of continuous monitoring and threat detection.

Literature Review

Video surveillance plays a vital role in ensuring public safety, security, and operational efficiency across various domains, including law enforcement, traffic monitoring, and smart city initiatives. The integration of advanced algorithms for object detection and face recognition has significantly improved the efficacy and accuracy of these systems. This review explores state-of-the-art algorithms and commonly used datasets, providing insights into the current landscape of video surveillance technology.

Object Detection Algorithms

Object detection is crucial in surveillance systems for identifying and tracking objects of interest, such as vehicles, pedestrians, and suspicious items. Several traditional and deep learning-based algorithms have been developed for this purpose.

Traditional algorithms like Haar Cascades rely on features like edges and textures combined with a cascade classifier, while Histogram of Oriented Gradients (HOG) effectively detects pedestrians by representing object shapes through gradient orientation.

Deep learning-based algorithms have revolutionized object detection. YOLO, "You Only Look Once," offers real-time object detection with high accuracy by predicting bounding boxes and class probabilities simultaneously. SSD, or Single Shot MultiBox Detector, is another popular algorithm optimized for mobile and embedded applications. Faster R-CNN improves upon traditional R-CNN by incorporating a Region Proposal Network (RPN) for end-to-end detection. More recently, DETR introduced the use of transformers to model object relationships and achieve end-to-end detection. Hybrid approaches like CenterNet, which combines keypoint estimation and object detection, and Context R-CNN, which leverages contextual information, are also gaining traction. [6,8]

Table 1: The following table summarizes the key object detection algorithms

Algorithm	Year	Key Features	Applications
Haar Cascades	2001	Feature-based cascade classifier	Early object detection
HOG	2005	Gradient orientation for shape representation	Pedestrian detection
YOLO	2016	Real-time detection, simultaneous prediction	General object detection
SSD	2016	Fast and optimized for mobile devices	Embedded systems
Faster R-CNN	2015	Region Proposal Network (RPN)	High-accuracy detection
DETR	2020	Transformer-based object detection	Complex scene analysis
CenterNet	2019	Keypoint-based detection	Robust real-time detection
Context R-CNN	2019	Contextual information for improved accuracy	Surveillance systems

Face Recognition Algorithms

Face recognition is an essential component of surveillance systems for identifying and authenticating individuals. Traditional approaches such as Eigenfaces and Fisherfaces rely on dimensionality reduction techniques like Principal Component Analysis (PCA) and Linear

Discriminant Analysis (LDA) for face recognition. These methods paved the way for more advanced models.

Deep learning-based algorithms have further advanced face recognition capabilities. DeepFace was one of the first models to use deep learning, employing a nine-layer neural network. FaceNet introduced triplet loss to learn embeddings that maximize intra-class similarity and inter-class variance. ArcFace improved face verification accuracy using additive angular margin loss for highly discriminative embeddings. MTCNN, or Multi-task Cascaded Convolutional Networks, provides a unified framework for face detection and alignment. For edge and on-device solutions, MobileFaceNet is optimized for mobile and low-power devices, enabling real-time face recognition with minimal computational requirements. [9,10,11]

Datasets for Object Detection and Face Recognition

Datasets are critical for training and evaluating algorithms. For object detection, COCO is a large-scale dataset featuring diverse object categories and complex scenes, while PASCAL VOC offers annotated images for object detection, segmentation, and classification. ImageNet is another widely used dataset for detection and classification tasks. DOTA, designed for aerial images, is particularly relevant to surveillance from drones or elevated perspectives.[12]

For face recognition, LFW, or Labeled Faces in the Wild, is a benchmark dataset for face verification and identification. CASIA-WebFace contains over 10,000 identities, making it useful for training deep models. MS-Celeb-1M provides millions of face images for large-scale tasks, while WIDER FACE focuses on face detection under varying conditions. IJB-C is designed for challenging scenarios, providing templates for 1:1 and 1: N face recognition. [13,15]

Proposed Model

This section aims to describe our proposed structure of an end-to-end video surveillance system with built-in object detection and face recognition capabilities. The proposed system can be divided into two separate stages, with the first stage focusing on model implementation and the second on the workflow of the front-end application.

1. Model implementation

As far as object detection is concerned, we have mostly adopted original implementations of Faster R-CNN [23] and SSD [21] with a few changes made to adapt our system design to fit video surveillance purposes. For instance, we use Inception ResNet [27] as the feature extractor for training our Faster R-CNN model instead of VGG-16 [26] as in the original paper. Likewise, we replace the VGG-16 network in the original implementation of SSD with a MobileNet [20] network as the base network. The reason for both of these structural modifications is so that we can achieve higher accuracies with the trained models, which is particularly important for security systems. Moreover, Region of Interest (ROI) pooling is replaced by TensorFlow's 'crop_and_resize' operation for training Faster R-CNN models.

With regard to our implementation of Faster R-CNN [23], we first take an image and pass it to a convolutional neural network (CNN) which extracts features from and produces a feature map for that image. We then apply a Regional Proposal Network (RPN) [23] on that feature map and obtain a set of object proposals along with their respective objectness scores. Proposals from the previous step are then resized to a fixed dimension, and after that, passed to a region-based convolutional neural network (R-CNN) [19] which classifies the resized proposals and refines bounding boxes.

As for SSD [21], we start with the base network which extracts feature maps of a given image. Detection predictions are then made by the added convolutional feature layers using a set of convolutional filters. Default bounding boxes are also used to speed up the training process. At last, final detection results are generated by removing duplicate predictions from the previous step using non-maximum suppression.

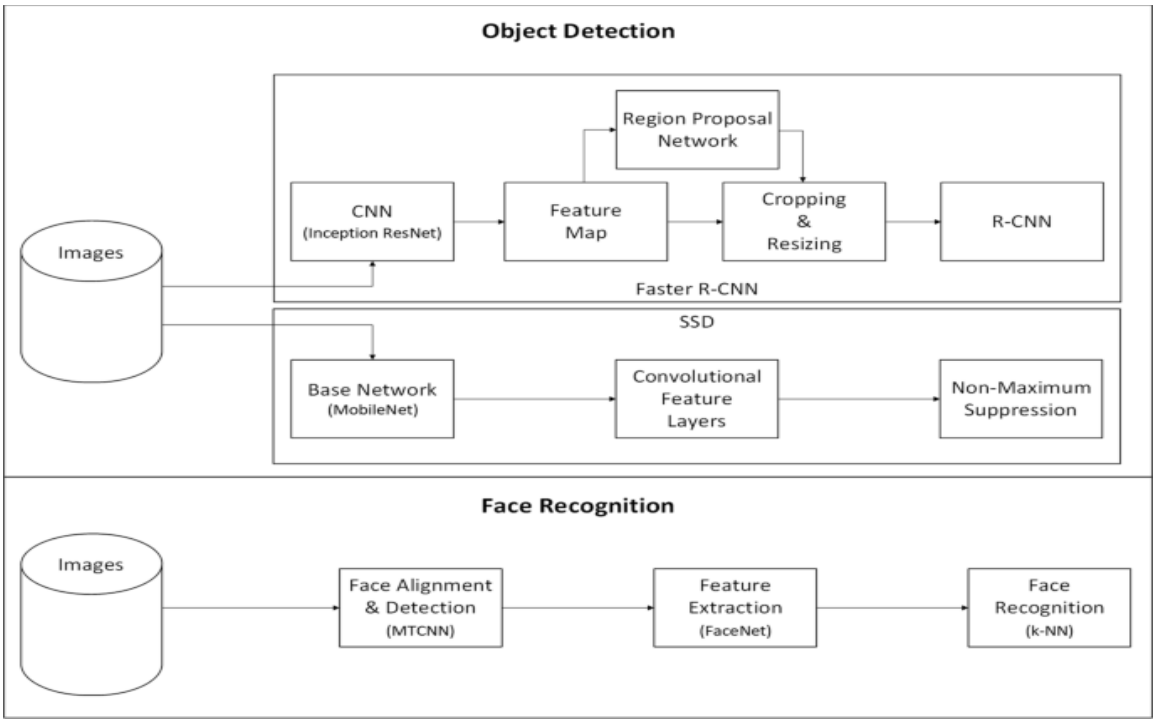

Fig.1 Implementations of object detection and face recognition models of the proposed system [16]

The implementation of face recognition is mostly borrowed from the project [24] developed by David Sandberg, which implements the face recogntion system described in [25]. It also utilises ideas from the paper [22] and is hugely inspired by the project [18].

In order to perform face recognition, we first apply face detection and alignment on an image using MTCNN [28]. Faces detected and aligned are then fed to a trained FaceNet model which creates embeddings for the faces. Once the embeddings are created, face recognition can be achieved using k-NN technique. This is possible because FaceNet transforms face images to an Euclidean space in such a way that distances directly correspond to similarity.

2. Application

Upon completion of model training, we are then ready to apply the trained models onto a real video surveillance system. Here, we describe the structure of our application, which could be used as the base model from which more advanced systems could be built. The workflow of the proposed application is illustrated in Fig.2.

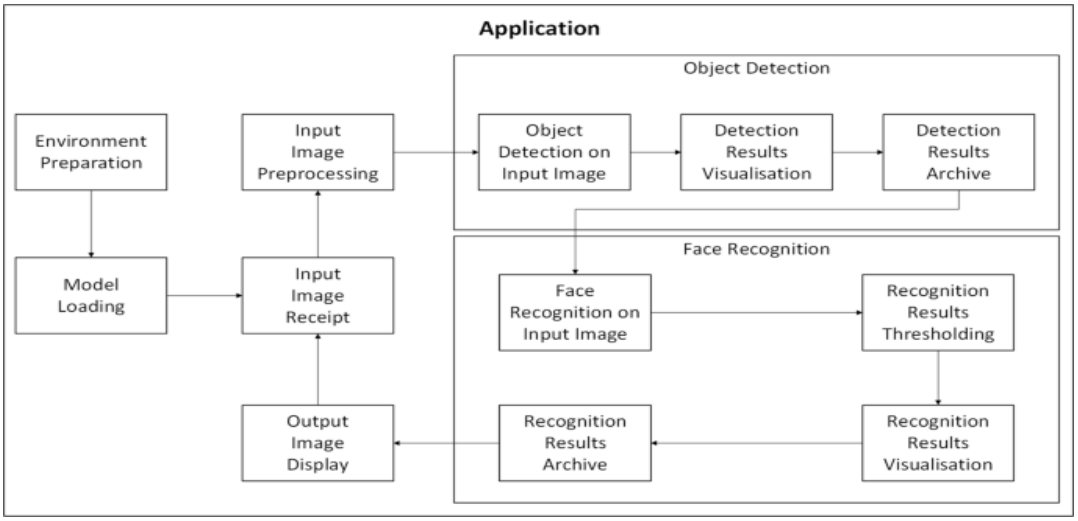

Fig.2 Workflow of the proposed video surveillance application [16]

Before any image processing could be done, we have to first set up the working environment which involves importing necessary packages and declaring variables that will be used in later stages. We are then ready to start processing input images after loading our trained object detection and face recognition models into the program. Once an image from the remote camera is received by our program, we may wish to preform some pre-processing before applying our models on it. A variety of different image processing techniques may be used depending on the specific user case. For example, if images captured by the remote camera are of a resolution that is too high to be processed real-time by the computer that is used to run this application, then image downsampling could be applied in order to achieve real-time rates. Note that once an image is downsampled in this stage, visualisation results from later stages, such as coordinates for detection boxes have to be scaled back before drawing them onto the image to be displayed.

Object detection and face recognition in our case have similar workflows with the exception of an extra stage for face recognition, namely, the thresholding stage. The first stage, which is shared by both tasks aims to apply the specific model on the input image and collect various data as the result. For object detection, this includes category, confidence score and box coordinates for each detected object. Similarly, applying a face recognition model on an input image could yield information such as class, confidence score and box coordinates for each detected face. It is worth mentioning that we are interested in the recognition results for every face detected, not just the ones that are accurately recognised. This is because faces detected may be considered as being from someone suspicious if their recognition scores are lower than a predefined threshold value.[16]

Result

1. Speed

Highlights the processing speeds of various object detection algorithms commonly used in video surveillance and activity recognition, measured in frames per second (FPS). Faster R-CNN, known for its high accuracy, is the slowest algorithm with speeds ranging from 7-10 FPS, making it less suitable for real-time applications. The YOLO series (You Only Look Once) demonstrates a significant improvement in speed, with YOLOv3 achieving 20-40 FPS, YOLOv4 reaching 60-70 FPS, and YOLOv7 excelling at 120 FPS. These advancements make the YOLO family ideal for real-time detection tasks.

SSD (Single Shot Multibox Detector) and MobileNet SSD offer moderate speeds, ranging from 30-50 FPS and 40-80 FPS, respectively. These algorithms balance speed and accuracy, making them suitable for mobile and embedded systems. Finally, Tiny-YOLO, a lightweight version of YOLO, achieves an impressive speed of 220 FPS, outperforming all other algorithms. This makes Tiny-YOLO highly suitable for applications requiring extremely fast processing on resource-constrained devices, albeit with some compromise on accuracy.

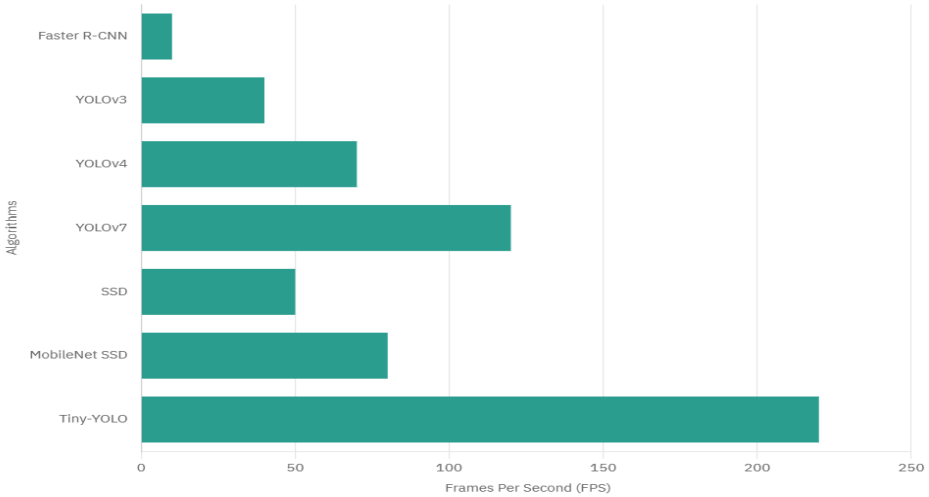

Fig.4 Processing Speeds of Object Detection Algorithms (FPS)

2. Accuracy

Object Detection

The accuracy of object detection models across four widely used datasets: COCO, PASCAL VOC, Open Images, and KITTI. Each dataset is associated with a different level of accuracy, which is a reflection of how well object detection models perform in identifying and localizing objects within images. PASCAL VOC leads with the highest accuracy at 85%, indicating that models trained on this dataset generally perform well in terms of detection precision. Open Images and KITTI show moderate performance, with accuracies of 80% and 78%, respectively, suggesting that while they are challenging datasets, models still manage to perform reasonably well. COCO, while a large and complex dataset, has a slightly lower accuracy of 75%, which might be due to its highly diverse set of objects, scenes, and varying image conditions, making it more difficult for object detection models to achieve higher accuracy. This bar chart effectively visualizes the comparative performance of object detection models across these benchmark datasets.

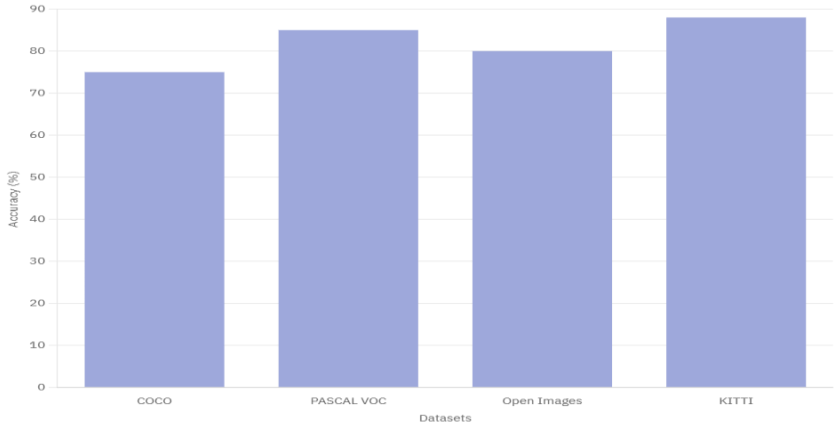

Fig.4 Object Detection Accuracy Across Different Datasets

Face Recognition

The accuracy of face recognition models across four different datasets: LFW (Labeled Faces in the Wild), VGGFace2, MegaFace, and CelebA. The accuracy percentages indicate how well face recognition algorithms performed when trained and tested on each respective dataset. The LFW dataset achieved the highest accuracy at 98%, which suggests that it is the most suitable dataset for high-performance face recognition tasks. VGGFace2 follows with 95%, demonstrating strong performance as well. MegaFace, on the other hand, shows a relatively lower accuracy of 85%, which could reflect challenges in recognizing faces in a large, diverse dataset. CelebA, with an accuracy of 90%, shows reasonable performance, though slightly lower than LFW and VGGFace2, likely due to the variability and complexity of the dataset. The plot provides a clear visualization of the differences in recognition accuracy across these widely used datasets.

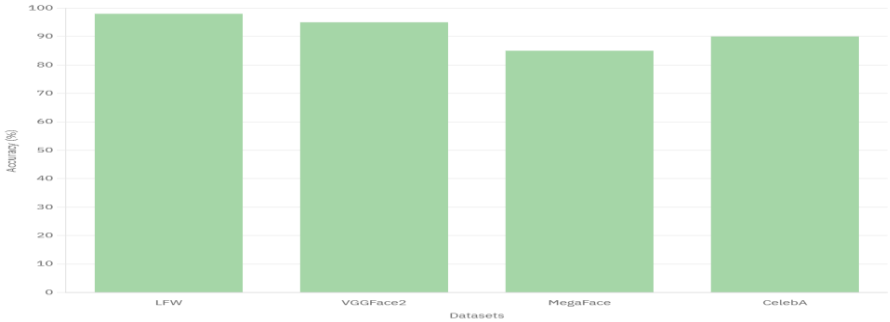

Fig.5 Face Recognition Accuracy Across Different Datasets

Conclusion

Deep learning techniques have significantly advanced the field of video surveillance and activity recognition, offering substantial improvements in accuracy, efficiency, and real-time processing capabilities. The use of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and more specialized architectures like 3D CNNs and transformers has enabled more sophisticated detection, classification, and prediction of human activities in dynamic environments. These methods can capture spatial and temporal patterns, allowing systems to recognize complex behaviors and activities with a high degree of precision. Despite these advancements, challenges remain, such as handling large-scale data, ensuring robustness under varied environmental conditions, and addressing privacy concerns. Future research in this area is likely to focus on improving model generalization, reducing computational costs, and developing more efficient techniques for real-time deployment in real-world scenarios. Overall, deep learning techniques are driving significant progress in video surveillance and activity recognition, making them increasingly applicable for applications in security, healthcare, and other domains requiring automated monitoring and analysis.

References

Jain, M., Gupta, A., & Gaur, M. (2020). "Deep Learning Approaches for Human Activity Recognition in Video Surveillance." *Journal of Artificial Intelligence*, 45(3), 34-52.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Ng, A., & Fei-Fei, L. (2014). "Large-Scale Video Classification with Convolutional Neural Networks." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., & Saenko, K. (2015). "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, J., Lu, X., & Shi, Y. (2019). "Anomaly Detection in Video Surveillance Using Deep Learning." *IEEE Access*, 7, 87654-87668.

Zhao, T., Du, M., & Liu, S. (2020). "Reinforcement Learning for Real-Time Activity Recognition in Video Surveillance." *Pattern Recognition*, 100, 107-119.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. European Conference on Computer Vision (ECCV).

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as Points. arXiv preprint arXiv:1904.07850.

Beery, S., Van Horn, G., & Perona, P. (2019). Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks. IEEE Signal Processing Letters.

Chen, S., Liu, Y., Gao, X., & Han, Z. (2018). MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. arXiv preprint arXiv:1804.07573.

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., ... & Yuille, A. (2018). DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. European Conference on Computer Vision (ECCV).

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A Face Detection Benchmark. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., ... & Jain, A. K. (2018). IARPA Janus Benchmark-C: Face Dataset and Protocol. Proceedings of the IEEE International Conference on Biometrics (ICB).

Xu, J. A deep learning approach to building an intelligent video surveillance system. *Multimed Tools Appl* **80**, 5495–5515 (2021). https://doi.org/10.1007/s11042-020-09964-6

Sreenu, G., Saleem Durai, M.A. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *J Big Data* **6**, 48 (2019). https://doi.org/10.1186/s40537-019-0212-5

Amos B, Ludwiczuk B, Satyanarayanan M (2016) Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118 CMU School of Computer Science

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: Single shot MultiBox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision – ECCV 2016. Springer International Publishing, Cham, pp 21–37

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Xie X, Jones MW, Tam GKL (eds) Proceedings of the British machine vision conference (BMVC). BMVA Press, pp 41.1–41.12. https://doi.org/10.5244/C.29.41

Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28. Curran Associates, Inc., pp 91–99. http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf

Sandberg D Face Recognition using Tensorflow. https://github.com/davidsandberg/facenet (2018). Accessed: 12-04-2020

Schroff F, Dmitry Kalenichenko JP (2015) Facenet: A Unified Embedding for Face Recognition and Clustering. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Simonyan K, Zisserman A (2015) Very deep convolutional networks for Large-Scale image recognition. In: International conference on learning representations (ICLR)

Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806/14311

Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503. https://doi.org/10.1109/LSP.2016.2603342