Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220606

A Deep Learning based Human Detection and Tracking for Security Surveillance Systems

Tahira Irshad[1], Muhammad Asif[1], Arfa Hassan[1], Umair Bin Ahmad[1], Toqeer Mahmood[2,3], Rehan Ashraf[2] and C.M. Nadeem Faisal[2]

[1] Department of Computer Science, Lahore Garrison University, Lahore, Pakistan
[2] Faculty of Computer Science, National Textile University, Faisalabad, Pakistan

[3]Corresponding author: toqeer.mahmood@yahoo.com

Abstract. All around the world, the crime rate has been increasing day by day, causing a rise in security issues. Closed-Circuit Television (CCTV) cameras have been installed throughout the world with the aim of decreasing crime and increasing public safety. The usage of CCTV cameras helps to increase crime detection accuracy significantly. Daily, a considerable amount of data has been recorded through CCTV cameras. Detection and recognition of culprits in the recorded data is a challenging task as it takes a lot of time, and human interaction is also involved. So, there is a need to develop a system that performs real-time detection and tracking of humans. This paper proposes a human detection and tracking system based on deep learning that assigns a unique ID to humans who enter the video scene. Multi-Task Cascaded Convolutional Neural Networks (MTCNN) and FaceNet models are used to achieve the desired target. The MTCNN model is trained on the WIDER SPACE dataset to perform human detection. FaceNet is used for human identification that is trained on the LFW dataset. The proposed system has been evaluated on 50 video sequences captured in different environments and achieved 97% average accuracy.

Keywords: human detection, tracking, security and surveillance, CCTV, FaceNet, MTCNN

1. Introduction

Public safety is one of the main focuses of any government. Street crime is meant by any criminal offense such as robbery, assault, homicide, rape, and arson in a public place [1]. Street crime is one the main challenge faced by law enforcement agencies around the globe. Street crimes create a significant impact on human lives as they did not feel safe to travel which significantly affects their routine life and pushes them into depression [2]. The rate of street crime increases day by day as the world's population grows. Table 1 lists the top 10 countries with the highest street crime rate index. It can be seen that Venezuela is at the top of the list with a crime rate index of 83.76% [3]. To calculate the crime rate index, the authorities find the total number of reported cases, divide it by the total number of populations, and then multiply the answer by 100000 [4, 5].

DOI: 10.54254/2755-2721/2/20220606

Table 1. Top ten countries' with high crime rate index [6].

Sr. No.	Country Name	Crime Rate Index
1	Venezuela	83.76
2	Papua New Guinea	80.79
3	south Africa	76.86
4	Afghanistan	76.31
5	Honduras	74.54
6	Trinidad and Tobago	71.63
7	Guyana	68.70
8	EI Salvador	67.79
9	Brazil	67.49
10	Jamaica	67.42

The security expert introduces different kinds of methods to control this highly alarming street crime in developed and underdeveloped countries. Installation of CCTV cameras on-street is one of them. It can be considered as a progression of inquiries involved in the '5WH' examination model. This model describes who was associated with an occurrence, where did it occur, what occurred, when did it occur, for what reason did it occur, and how any offenses were submitted [6]. While using CCTV cameras detection rate has been enhanced to the accuracy of 55.7%. Due to the absence of cameras, detection rates of just about 2% [7].

Daily, a vast amount of data has been recorded through CCTV cameras. Detection and recognition of culprits in the recorded data is a challenging task. It takes a lot of time, and human interaction is also involved. So, there is a need to develop a system that performs real-time detection and tracking of humans. Many researchers have been working on the implementation of detection and tracking. Several challenges are faced, such as segmentation errors, problems in tracking complex objects such as faces and shadows, and changes in lighting conditions. There has been a low accuracy rate of 69% [8]. A system with higher efficiency protocol has been presented, but there is a problem in storage capacity while numerous companies enhance their storage capability (up to 50%). Generally, memory has been removed weekly in a CCTV to give new footage storage because of low storage capacity [9]. To overcome these challenges, an approach of computer vision has been used, which might start with identifying and detecting individuals in the scene that can be classified as humans and performing tracking [10]. After the human bodies are detected, the counting procedure by facial identification is straightforward.

Manual monitoring through CCTV cameras is a hectic and time taking job. It needs a lot of attention, and many incidents can be unaddressed during manual checking. To address this issue, in this work, deep learning-based CCTV cameras monitoring system is introduced that detects the human from video, identifies the human face, assigns a unique ID, and performs tracking. Multi-Task Cascaded Convolutional Neural Networks (MTCNN) and FaceNet models are used to achieve the desired target. The MTCNN model is trained on the WIDER SPACE dataset to perform human detection. FaceNet is used for human identification that is trained on the LFW dataset. The experimental results indicate that the proposed system performs the desired task with 99% accuracy. The security agencies can automatically identify unauthorized persons and activities through this system without delay and biasness.

The rest of the paper is organized as follows: Section 2 presents a literature review. The proposed human detection and tracking system are presented in section 3. Experimental analysis is made in section 4. Finally, the conclusion is drawn in section 5.

DOI: 10.54254/2755-2721/2/20220606

2. Literature review

In literature, several efforts have been made by proposing a variety of frameworks, algorithms, models, and different tools and techniques to detect and track humans from video. Visakha et al. [11] have developed a technique in which humans are detected from the video scene and tracked as long as they stay in the scene by recognizing individual persons. They used multiple cameras to capture the scenes and a Haar classifier for human detection. The experimental results are not mentioned and discussed are missing in this work. Jin et al. [12] have proposed a technique to identify humans in pre-recorded videos by using global and local structural information. Initially, they performed pedestrian detection after that tracking is done through facial recognition. To develop a system, they captured 93 video sequences in the corridor. They achieved 91 % accuracy.

Ojha et al. [13] have presented a survey of techniques that perform object tracking in video surveillance. They classified techniques into different types and discussed their positive and negative aspects. They also presented background subtraction, temporal differencing, and the optical flow tracking process. They categorized object tracking techniques as region-based, active contour-based, and feature-based. Prabhakar and Ramasubramanian [14] have proposed an algorithm to track unknown and abandoned objects with the help of background subtraction and morphological filtering. They tested the algorithm on a real-time video surveillance system and claimed very low false alarms and missing detection. Akama et al. [15] have used multiple omnidirectional cameras and a machine-learning algorithm to develop a method for human tracking and posture evaluation. By using two omnidirectional cameras, the trajectory of Head and foot position could be obtained continually to identify the behavior of residents. A successive human tracking system can only be possible in daily life by adding a machine learning method to identify the action between abnormal and normal situations. Chang and Liu [16] have proposed a technique to detect multiple heads that can be used in the smart human tracking system. They also presented a parallel design for optimizing the computational complexity of this approach to deploying it on embedded platforms. This technique performed head detection with 85% accuracy. Kang et al. [17] have developed a multiple moving body human detection and tracking technique with the help of CCD and thermal image sensor. They used a chess board to calibrate the CCD and thermal image that display the texture and thermal data for heterogeneous sensor fusion. To track the moving body, a segment blob tree is developed. They claimed that the technique performed well in dynamic environments. Dinama et al. [18] have proposed a system for human detection and tracking for surveillance of video footage. To track the detected persons, tracking algorithms are developed using a deep neural network in which channel and spatial filters are used.

Shehzed et al. [19] have presented a multi-person tracking technique for crowd counting and abnormal/normal events detection for both outdoor and indoor environments. The proposed technique comprises four main modules including human detection with the help of inverse transform and median filter, head-torso template extraction, human tracking using Kalman filter with Jaccard similarity and normalized cross-correlation, and crowed clustering analysis based on Gaussian mapping for normal and abnormal events detection. The experimental analysis is performed on PETS2009 and UMN crowd analysis datasets that showed that the system has achieved 88.7% counting accuracy and 95.5% detection rate. Jin et al. [20] have proposed a system to perform human identification based on global and local structural information. Initially, they performed pedestrian detection and tracking using facial recognition. After that, a pedestrian identification is made through a selective algorithm. They used their own dataset and claimed that their method had better results as compared to techniques that are using only facial information. Othman and Aydin [21] have introduced a human detection technique that is capable of catching images and sending them to android phones. In this work, people are detected with the help of an IoT-based system that also used computer vision techniques. To develop a system, a Raspberry PI 3 card along with a PIR sensor and camera is used. PIR sensor help to perform any movement detection and a camera is used to capture the movement. To send the captured images to a Smartphone telegram application is used. Experimental results showed that a 93.89% detection rate has been achieved.

Liu et al. [22] have proposed a technique for resolving multi-target visual tracking problems. They considered across the frame and across camera data correlation along with appearance and dynamic

DOI: 10.54254/2755-2721/2/20220606

similarities scores to address this problem. For appearance attributes, the Local Maximal Occurrence Representation feature extraction technique for ReID is used. For dynamic features, they developed a Hankel matrix against each target tracklet, and rank estimation with Iterative Hankel Total Least Squares algorithm is applied to it. They evaluated the technique on sequences taken from EPFL CVLAB and Duke MTMC dataset and achieved satisfactory results. Ko et al. [23] have derived a new technique that integrates the extended Kalman filter-based 2D image tracking and 3D depth tracking to improve the performance of fall detection with the help of a single camera. The experimental analysis showed good results.

It is observed that several researchers have proposed techniques to perform human detection and tracking in videos. Some techniques targeted single camera detection and tracking while others performed multi-cameras human detection and tracking. Only a few techniques assigned a unique ID to a person but the ID changed every time whenever the person re-enters an environment. There is a need to develop a comprehensive system that performs human detection and tracking by assigning a unique ID to any human being who enters this environment each time.

3. Proposed system

The complete design of the proposed system is discussed in this section. The proposed method takes the video as an input and applies a different operation to find the solution to the problem. The flow diagram of the proposed system is shown in Figure 2. The following sections describe the main components of the proposed system.

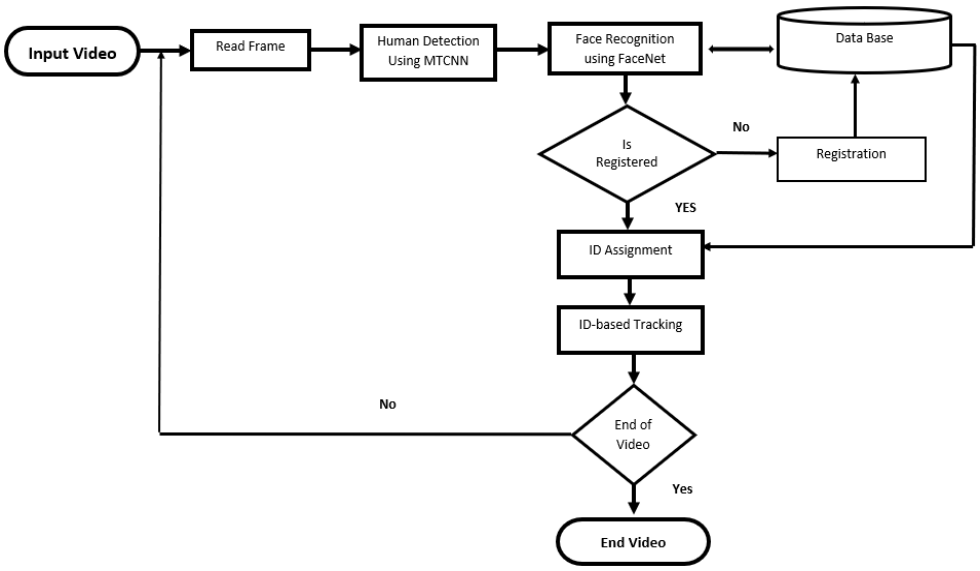

Figure 1. Flow diagram of the proposed system.

3.1. Input video

To develop the proposed system, 50 video sequences are captured in both indoor and outdoor environments through CCTV and mobile phone cameras. The details of a few captured video sequences are listed in Table 2.

DOI: 10.54254/2755-2721/2/20220606

Table 2. Sample test video sequences.

Video Source	Duration (Second)	Resolution	No. of Person	Environment	Total No. of Frames
CCTV	33	1280x720	3	Outdoor	792
Mobile	19	640×360	3	Outdoor	456
CCTV	45	1280×720	7	Indoor	1080
Mobile	55	1280×720	3	Indoor	1320
CCTV	60	1280×720	4	Outdoor	1440

3.2. Human detection using Multi-Task cascaded convolutional neural network (MTCNN)

MTCNN is a type of neural network algorithm that is a pre-trained model introduced in 2016 by Zhang et al. This algorithm is used for human face detection on human face image datasets [24]. It uses three different networks, P-Net (Proposal-Network), R-Net (Refine-Net), and O-Net (Output-Net), to find out the alignments of the face. The primary formation of the MTCNN is shown in Figure 4. Initially, images are resized on different scales; the next step is to stack them into an image pyramid. After that, the system can produce the same face on a different scale. In this way, the system enhances the network's capability. Finally, a pivot window will be put into the pyramid and split the image into its domain, which is input to the network.

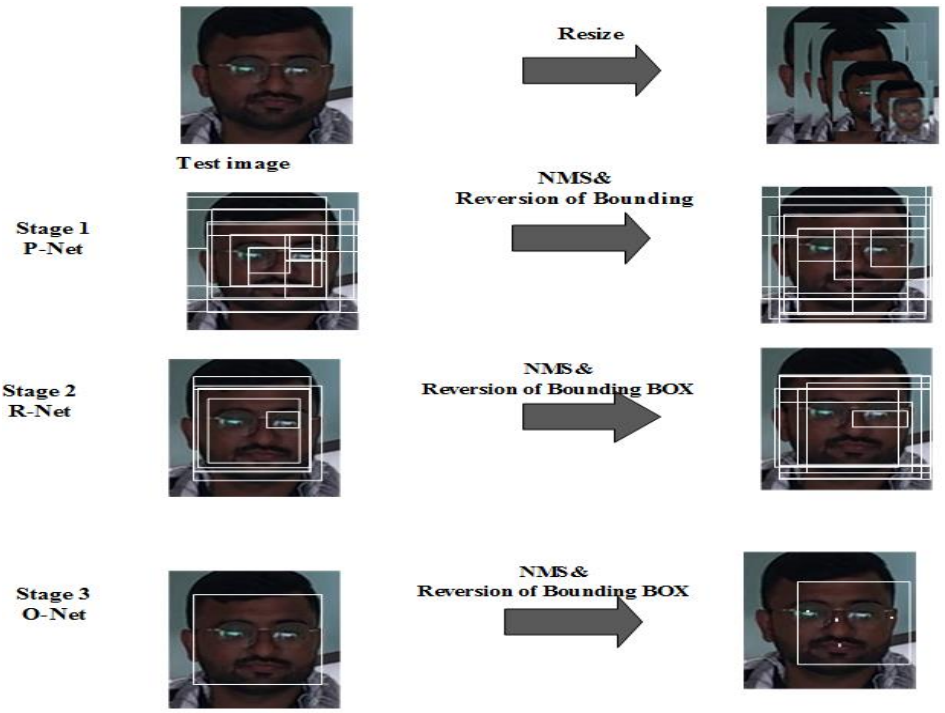

Figure 2. The pipeline of the cascaded framework of MTCNN.

3.3. Face recognition

To perform face recognition, FaceNet is used in the system. FaceNet was launched in 2015 by Google examiner Schroff et al. [25]. The Architecture of FaceNet consists of a deep convolutional neural network. Batch is used for the input layer, and that is further followed by L2 normalization, which provides face embedding. In his process, the triplet loss function is used for face validation. This function tells us about f(x) creating embedding in d-dimensional space for an image x. In the training phase, FaceNet

DOI: 10.54254/2755-2721/2/20220606

learns and extracts different facial features. After that, these features are directly converted to 128D embedding in which the same faces shall be close to each other and various faces shall be huge apart in the embedding space. When FaceNet starts work, it takes multiple images of a particular person, saves them, and learns and matches images. To compare two or multiple images, it can create the embedding for multiple images by saving images throughout the model separately. FaceNet takes a picture of the individual's face as a participation and production vector of 128 numbers that address a face's main elements.

3.4. Registration and tracking

For a new face, the registration process is performed by assigning a unique ID and storing the facial image in the database. After registration, every time a human will enter the environment, the system detects and recognizes his/her face and will assign the same ID by marking the bounding box. The system works efficiently and repeats the same procedure automatically.

4. Experimental analysis

To develop and evaluate the performance of the proposed system, the work station with following specifications is used:

- Intel core i7-1065G7
- CPU 1.30GHZ, Quad-core processor 16 GB DDR4 Ram,
- Integrated Intel iris plus graphics
- 10^{th} generation processor

To perform the experimental analysis, 50 test video sequences are used. Test sequences are recorded in both indoor and outdoor environments through CCTV and mobile phone cameras. Table 3 shows a few input frames and outputs of the system with ID assignment to humans appearing in the scene. It also lists the correct and wrong predictions made by the system for a particular test sequence. The model works equally well for both indoor and outdoor test sequences. It is observed that if a person appears multiple times in a test video the system identifies and gives the same ID every time. The performance of the proposed system is evaluated using accuracy, precision, recall, and F1-Score performance measure metrics. Table 4 lists the experimental results of the proposed system on a five selected test sequences. It indicates that the proposed system has achieved 99% accuracy, precision, recall, and F1 score [26-30]. For a complete dataset, the overall accuracy of the proposed system is 97%.

Table 3. Experimental results (continue).

Sr. No.	Input	Output	Results
1			Correct predictions =791 Wrong prediction=1
2			Correct predictions =453 Wrong prediction=3

DOI: 10.54254/2755-2721/2/20220606

Table 5. (continued).

3			Correct predictions =1074 Wrong prediction=6
4			Correct predictions =1314 Wrong prediction=4
5			Correct predictions =1439 Wrong prediction=1

Table 4. Experimental results of the proposed system.

Sr. No.	Human	Detected by ID	Non detected by ID	Accuracy	Precision	Recall	F1 Score
1	3	3	0	99%	99%	100%	99.4%
2	3	3	0	99%	99%	100%	99.4%
3	7	7	0	99%	99%	99%	98.0%
4	3	3	0	99%	99%	100%	99.4%
5	4	4	0	99%	99%	100%	99.4%
Total/ Average	**20**	**20**	**0**	**99%**	**99%**	**99.8%**	**99.1%**

5. Conclusion

Detection and recognition of culprits in the CCTV recorded data is challenging because it takes a lot of time, and human interaction is also involved. To automate this process, an automatic human detection and tracking system based on deep learning is presented in this work. The presented system assigns a unique ID to humans who enter the video scene. The system used MTCNN and FaceNet deep learning models to perform the desired task. The MTCNN model is trained on the WIDER SPACE dataset to perform human detection. FaceNet is used for human identification that is trained on the LFW dataset. Experimental results indicate that the proposed system has achieved 97% average accuracy. In the future, this system can be deployed in a real-time environment to help the law enforcement agencies that detect and track criminals.

DOI: 10.54254/2755-2721/2/20220606

References

[1] Jonathan, O. E., Olusola, A. J., Bernadin, T. C. A., & Inoussa, T. M. (2021). Impacts of Crime on Socio-Economic Development. *Mediterranean Journal of Social Sciences*, 12(5), 71-71.

[2] Ul Haq, H. B., Asif, M., Ahmad, M. B., Ashraf, R., & Mahmood, T. (2022). An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning. Mathematical Problems in Engineering, 2022.

[3] Taylor, R. B. (1995). The impact of crime on communities. The Annals of the American Academy of Political and Social Science, 539(1), 28-45.

[4] Abbas, S., Shouping, L., Sidra, F., & Sharif, A. (2018). Impact of Crime on Socio-Economic Development: A Study of Karachi. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *3*(3), 148-159.

[5] Gajjar, V., Gurnani, A., & Khandhediya, Y. (2017). Human detection and tracking for video surveillance: A cognitive science approach. In Proceedings of the IEEE international conference on computer vision workshops (pp. 2805-2809).

[6] Harrendorf, S., & Heiskanen, M. (2010). International statistics on crime and justice. S. Malby (Ed.). Helsinki: European Institute for Crime Prevention and Control, affiliated with the United Nations (HEUNI).

[7] La Vigne, N. G., Lowry, S. S., Dwyer, A. M., & Markman, J. A. (2011). Using public surveillance systems for crime control and prevention: A practical guide for law enforcement and their municipal partners.

[8] Space and Naval Warfare Systems Center Atlantic, (2013). System Assessment and Validation for Emergency Responders (SAVER) Handbook: CCTV Technology.

[9] Stelfox, P. (2013). Criminal investigation: An introduction to principles and practice. Willan., DOI: 10.4324/9781315880730.

[10] Ashiq, F., Asif, M., Ahmad, M. B., Zafar, S., Masood, K., Mahmood, T., ... & Lee, I. H. (2022). CNN-based object recognition and tracking system to assist visually impaired people. IEEE Access, 10, 14819-14834.

[11] Visakha, K., & Prakash, S. S. (2018, July). Detection and tracking of human beings in a video using haar classifier. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1-4). IEEE.

[12] Jin, K., Xie, X., Wang, F., Han, X., & Shi, G. (2019, July). Human Identification Recognition in Surveillance Videos. In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 162-167). IEEE.

[13] Ojha, S., & Sakhare, S. (2015, January). Image processing techniques for object tracking in video surveillance-A survey. In 2015 International Conference on Pervasive Computing (ICPC) (pp. 1-6). IEEE.

[14] Prabhakar, G., & Ramasubramanian, B. (2012). An efficient approach for real time tracking of intruder and abandoned object in video surveillance system. International Journal of Computer Applications, 54(17), 22-27.

[15] Akama, S., Matsufuji, A., Sato-Shimokawara, E., Yamamoto, S., & Yamaguchi, T. (2018, November). Successive Human Tracking and Posture Estimation with Multiple Omnidirectional Cameras. In 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 46-49). IEEE.

[16] Chang, K. C., & Liu, P. K. (2017, October). Design and optimization of multiple heads detection for embedded system. In 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE) (pp. 1-2). IEEE.

[17] Kang, B. D., Jeon, K. H., Kyoung, D., Kim, S. H., & Hwang, J. H. (2011, October). Multiple human body tracking using the fusion of CCD and thermal image sensor. In 2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1-4). IEEE.

DOI: 10.54254/2755-2721/2/20220606

[18] D Dinama, D. M., A'yun, Q., Syahroni, A. D., Sulistijono, I. A., & Risnumawan, A. (2019, September). Human detection and tracking on surveillance video footage using convolutional neural networks. In 2019 International Electronics Symposium (IES) (pp. 534-538). IEEE.

[19] Shehzed, A., Jalal, A., & Kim, K. (2019, August). Multi-person tracking in smart surveillance system for crowd counting and normal/abnormal events detection. In 2019 international conference on applied and engineering mathematics (ICAEM) (pp. 163-168). IEEE.

[20] Jin, K., Xie, X., Wang, F., Han, X., & Shi, G. (2019, July). Human Identification Recognition in Surveillance Videos. In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 162-167). IEEE.

[21] Othman, N. A., & Aydin, I. (2017, September). A new IoT combined body detection of people by using computer vision for security application. In 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 108-112). IEEE.

[22] Liu, W., Camps, O., & Sznaier, M. (2017). Multi-camera multi-object tracking. arXiv preprint arXiv:1709.07065, doi: 10.48550/arxiv.1709.07065.

[23] Ko, M., Kim, S., Lee, K., Kim, M., & Kim, K. (2017, August). Single camera based 3D tracking for outdoor fall detection toward smart healthcare. In 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART) (pp. 1-4). IEEE.

[24] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.

[25] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[26] Ali, M., Asghar, M. Z., Shah, M., & Mahmood, T. (2022). A simple and effective sub-image separation method. Multimedia Tools and Applications, 81(11), 14893-14910.

[27] Mahmood, T., Shah, M., Rashid, J., Saba, T., Nisar, M. W., & Asif, M. (2020). A passive technique for detecting copy-move forgeries by image feature matching. Multimedia Tools and Applications, 79(43), 31759-31782.

[28] Asif, M., Bin Ahmad, M., Mushtaq, S., Masood, K., Mahmood, T., & Ali Nagra, A. (2021). Long multi-digit number recognition from images empowered by deep convolutional neural networks. The Computer Journal.

[29] Hussain, A., Asif, M., Ahmad, M. B., Mahmood, T., & Raza, M. A. (2022). Malware Detection Using Machine Learning Algorithms for Windows Platform. In Proceedings of International Conference on Information Technology and Applications (pp. 619-632). Springer, Singapore.

[30] Javed, S. H., Ahmad, M. B., Asif, M., Almotiri, S. H., Masood, K., & Ghamdi, M. A. A. (2022). An Intelligent System to Detect Advanced Persistent Threats in Industrial Internet of Things (I-IoT). Electronics, 11(5), 742.