International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 3, June 2022, pp. 3083~3091
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i3.pp3083-3091

Real-time face detection in digital video-based on Viola-Jones supported by convolutional neural networks

Tameem Hameed Obaida[1], Abeer Salim Jamil[2], Nidaa Flaih Hassan[3]
[1]Computer Systems Techniques, Al-Furat Al-Awsat Technical University, Najaf Technical Institute, AL Najaf, Iraq
[2]Department of Computer Technology Engineering, Al-Mansour University College, Baghdad, Iraq
[3]Department of Computer Science, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received Aug 3, 2021
Revised Jan 15, 2022
Accepted Feb 3, 2022

Keywords:

Convolutional neural network
Face detection
Kanade-Lucas-Tomasi
Track faces
Viola-Jones

ABSTRACT

Face detection is a critical function of security (secure witness face in the video) who appear in a scene and are frequently captured by the camera. Recognition of people from their faces in images has recently piqued the scientific community, partly due to application concerns, but also for the difficulty this characterizes for the algorithms of artificial vision. The idea for this research stems from a broad interest in courtroom witness face detection. The goal of this work is to detect and track the face of a witness in court. In this work, a Viola-Jones method is used to extract human faces and then a particular transformation is applied to crop the image. Witness and non-witness images are classified using convolutional neural networks (CNN). The Kanade-Lucas-Tomasi (KLT) algorithm was utilized to track the witness face using trained features. In this model, the two methods were combined in one model to take the advantage of each method in terms of speed and reduce the amount of space required to implement CNN and detection accuracy. After the test, the results of the proposed model showed that it was 99.5% percent accurate when executed in real-time and with adequate lighting.

Corresponding Author:

Tameem Hameed Obaida
Department of Computer Systems Techniques, Al-Furat Al-Awsat Technical University, Najaf Technical Institute
AL Najaf, Iraq
Email: tameem_hameed@yahoo.com

1. INTRODUCTION

Several studies have focused on the detection and analysis of human beings using facial features in recent decades [1]. Face detection (FD) is a form of social communication for humans; it involves both verbal and non-verbal. Face detection is a communication of non-verbal form, in which the important communication signs are expressed by face, such as eye contact. Gestures and body language are examples of non-verbal communication. Human faces are easy for people to detect and interpret [2]. However, developing an automated system that achieves the same comprehension remains tough. There are various issues in this area, including detecting an image part as an actual face depending on illumination or occlusions, fluctuation in head postures, extraction of face info, detection of facial landmark, or classification [3], [4]. Face detection is an active study subject in the artificial intelligence (AI) field, with applications in video surveillance systems, biometrics for access control, social humanoid robots, security applications, among others [5]. However, to do face analysis jobs, face detection must be done to recognize the human face automatically. For decades, several studies [6]–[8] have investigated the subject of face detection. Despite tremendous advances, in an uncontrolled environment, the robust face detection is remained rather murky area due to the emergence of substantial variations of faces impacted by occlusion, position fluctuations, poor resolutions, scale variations, light fluctuation, and other factors [9].

The proposed method employs the Viola-Jones algorithm in conjunction with convolutional neural networks (CNN). The Viola-Jones algorithm is used to detect a human face, and CNN is used to identify the face as a witness or non-witness, this work used CNN with various parameters to obtain high accuracy. The suggested model comprises of three steps: the first step collects frames from the camera and reads each frame as an image, the second step use Viola-Jones to image and detection human face object in images, and the final step use CNN to classify the objects and identify face or non-face. The rest of the paper is laid out as follows: section 2 is overviewed a brief of the related works. The proposed approach has been identified and described in depth in Section 3. The experimental results are described in Section 4. Section 5 concluded the work and recommends new viewpoints for future works.

2. RELATED WORK

Detection face is the first step before the tracking process, which depends on identifying facial features. There are different methods for detecting the face, which are divided into two main categories: local feature-based and global methods [10], [11]. Guo *et al.* [12] introduced a deep CNN with numerous inputs, including visible light images and near-infrared imagery. The authors combined the information loss approach and the nearest neighbor technique to obtain predictions. The experimental results demonstrated that it was extremely resistant to illumination and outperformed other state-of-the-art approaches. Kamencay *et al*. [13] used a deep neural network (DNN) to develop face detection while analyzing face detection approaches. The performance improvement was superior to several previously published studies. A comparison of CNN's performance against three well-known image detection algorithms, principal component analysis (PCA), local binary pattern histogram (LBPH), and K-nearest-neighbors (KNN), revealed that CNN outperforms them all. The ORL dataset's experimental findings detected the efficacy of CNN-based face detection. Face detection accuracy of 98.3% was achieved by the proposed CNN.

Arya and Agrawal [14] published a review of various face recognition algorithms and methodologies that promised to produce efficient and optimal face detection. Li *et al.* [15] introduced a new identification algorithm based on the decision level function of the C2D-CNN model. When there were significant variations between the test and the training sets, the proposed approach was put to the test. A new CNN model was presented with a faster convergence process and shorter training time than existing approaches, resulting in superior performance.

Taherkani *et al*. [16] proposed a deep learning approach for improving face detection by predicting facial features. A CNN with two outputs was used to create the proposed model. The results of experiments showed that this approach outperformed existing face detection and feature prediction approaches. Said *et al.* [17] suggested a way to modify the biometric system for recognition of face using convolutional neural networks by structured a model for deep learning, that improved accuracy and processing time. The proposed method achieved an accuracy of 98.75%.

3. THE PROPOSED MODEL

The suggested model is based on wittiness detection of the face. We capture video with the camera, then read the frames of video, treating each frame as an image of the witness, then detecting the face in the image, recognizing the witness, and finally creating a dataset with the witness face. Face detection performance is improved by using an efficient face detection method. The face region of the witness is extracted and a dataset is created for further processing. The extracted face image's Viola-Jones algorithm is used to create the building dataset, which is then downsize using crop operations. As for face detection, we must capture an image of witnesses from various angles and build a training dataset that will be utilized by the CNN classifier to detect the witness's face, after which the Kanade-Lucas-Tomasi (KLT) algorithm will track it. The proposed model is depicted in Figure 1.

3.1. Dataset description

The major purpose of this work is to detect the faces of witnesses. The key issue is how to train CNN; thus, we'll need to create a dataset to do this. We collect and construct our collection because finding public datasets that suit our requirements is tough. We need to collect multiple angles of the witness's face because we need to capture the witness's facial images, as illustrated in Figure 2. Each face is classified as a witness or non-witness in the dataset, which has 300 different faces from various angles and with different size of pixels. The images of witnesses that were extracted from the webcam make up our dataset. The extraction procedure begins with converting a video stream from a camera to frames, then using the algorithm of Viola-Jones to extract the face of a human, cropping each image extracted from the Viola-Jones method, and finally building our dataset.

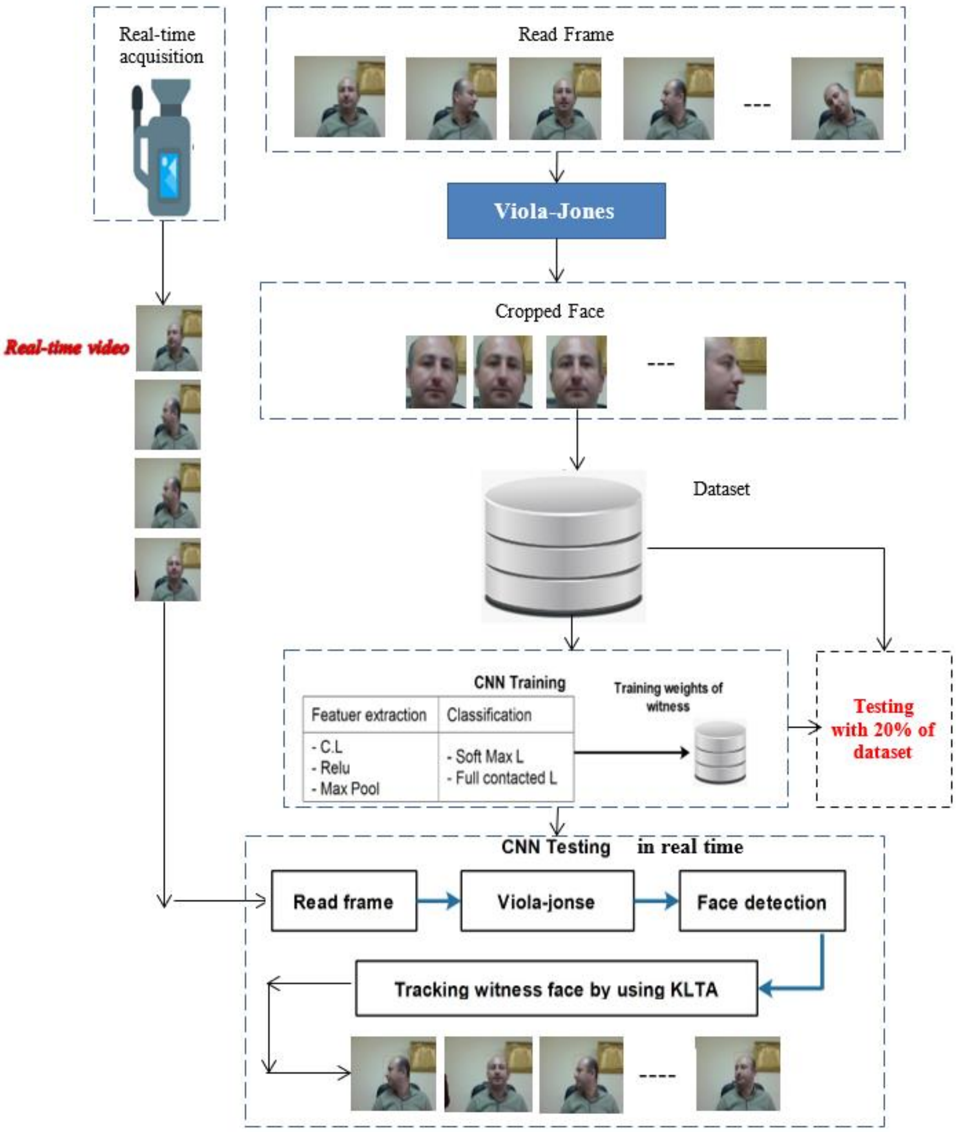


Figure 1. The proposed model

Figure 2. Sample of our dataset

3.1.1. Frame extraction from camera

Frame extraction (FE) can be used with both a camera and video input. When the FE receives video as an input, each frame is calculated. This project employed a frame-rate of 30 frames (frames per second). The frames occur, and an algorithm of the Viola-Jones detector was used to extract the human object in each of these frames. Following the computation of each frame, the frame with the human face is chosen since it contains face details (witness or non-witness). After that, this frame is chosen as an input for the cropped face.

3.1.2. Use the Viola-Jones algorithm

Viola-Jones algorithm (V-J) is used to detect human faces in real-time and has a high detection rate [18]. The haar-like features selection, AdaBoost training, and cascade classifiers are the three processes in this technique [19]. The first stage, Haar-like features selection, is a set of rectangular digital image features that divide an image into various sets of two adjacent rectangles at any scale and position within the image, and then these rectangles are applied to the picture [20]. AdaBoost, a machine learning approach for recognizing human faces, was employed in the second (1). The cascade classifier is used in the final phase to efficiently merge many of the features.

$$F(x_i) = sign(\sum_{t=1}^{L}(a_{ti} * (f_{ti}(x_i) + s))) \quad (1)$$

Where, F is the classification function, f_{ti} is the corresponding classifier, a_t is the AdaBoost coefficient $(a_t > 0)$, *x* is the input sample, and *s* is the sum of each rectangle determined by using (2) to find pixels A, B, C, and D in the image [21].

$$S = D - B - C + A \quad (2)$$

3.1.3. Cropped image

In this phase, the image generated by the Viola-Jones algorithm will cover the entire human body, and we will only clip the face. Cropping is required to resize the image using *mxn*, where *m*=182 and *n*=182. The cropped face in this work is based on spatial transformation [22]. The original image and the cropped image are shown in Figure 3.

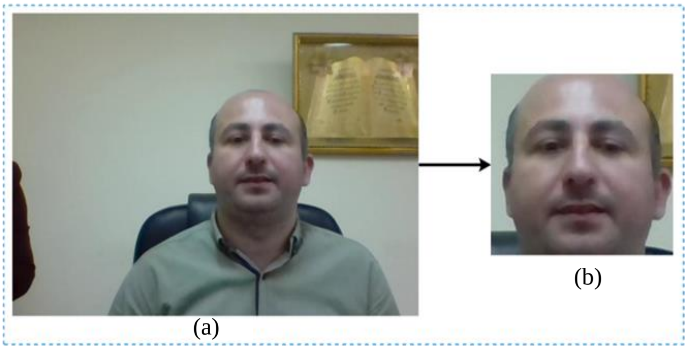


Figure 3. Cropped image with (a) the original image extracted using the Viola-Jones algorithm and (b) the cropped image

3.2. Apply the convolution neural network

To classify images into witness or non-witness, CNN was employed, since CNN is an up-to-date field of machine learning which is inspired by the brain of humans [23]. CNN is supposed to behave as the visual system of humans and is based on the concept that raw data is made up of 2-D images, allowing for the encoding of particular attributes. As a result, CNN was utilized, which generates feature maps by convolving images with kernels. Kernel weights connect units to preceding layers in a feature map, these weights are modified through training via a backpropagation procedure. Since all units used the same kernels, the convolutional layer only had to train a small number of weights. The sections listed below were used in conjunction with CNN to attain the classification of witnesses.

a) Activation function: the data was transformed into a non-linear form using this function rectified linear unit (ReLU) is the activation function employed, and it is expressed by (3).

$$F(x) = max(0, x) \tag{3}$$

b) Pooling: this layer was created to merge spatially adjacent features maps. To link features, either average pooling or max pooling is utilized; however, max-pooling was utilized in this work.

c) Architecture: because the visual cues of witnesses come from different angles, classification becomes a difficult process in this scenario. This complexity was decreased by fine-tuning the suggested CNN model to each image's intensity normalization transformation. Pooling, on the other hand, has removed certain critical characteristics; as a result, overlapping pooling with 3×3 receptive fields and 2×2 stride has been used to maintain position information. Feature maps have been padded before convolution in convolutional layers, as shown in (4). Padding guaranteed that feature maps were of the same size.

$$S(I * K)_{i,j} = \sum_m \sum_n I(m, n)K(i - m, j - n) \tag{4}$$

Where, I is the 2D array of witness faces segmentation, and K represents the kernel convolution function. Figure 4 is a representation of the proposed CNN architecture. The proposed architecture for witness and non-witness classification has been illustrated in Table 1. Table 2 lists the proposed architecture's hyper-parameters and their values, which were tweaked empirically. The CNNs were created using MATLAB 2018b and there are two classes to classify (witness and non-witness).

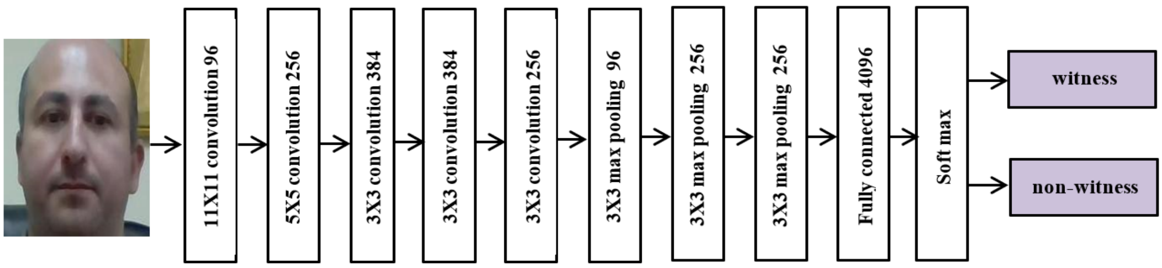


Figure 4. Proposed CNN architecture

Table 1. CNN's architecture was created to distinguish between witnesses and non-witnesses

Layer	Type of layer	Filter size	Stride	No of filters	Fully connected units	Input
1	Convolution	11x11	4x4	96	-	227x227x3
2	Convolution	5x5	1x1	256	-	27x27x96
3	Convolution	3x3	1x1	384	-	13x13x256
4	Convolution	3x3	1x1	384	-	13x13x384
5	Convolution	3x3	1x1	256		13x13x384
6	Max pooling	3x3	2x2	96	-	55x55x96
7	Max pooling	3x3	2x2	256		27x27x256
8	Max pooling	3x3	2x2	256		13x13x256
9	Fully connected	-	-	-	4096	1x1x4096

Table 2. The suggested CNN architecture's hyper-parameters

Stage	Hyper-parameter	Value
Initialization	Bias	0.1
	Weights	Random
Dropout	P	0.3
	Maximum epochs	25
	V	0.9
Training	Initial ε	0.0002
	Final ε	0.0002
	Batch	128

3.3. Apply the Kanade-Lucas-Tomasi (KLT) algorithm

The feature tracker is based on KLT [24]. This approach is used to find dispersed feature points with sufficient texture to track the needed points in a reasonable amount of time [22], [25]. The KLT algorithm was employed in this study to follow a witness's face in a video frame constantly. Calculate the displacement of the tracked points from one frame to the next with this procedure. It is simple to calculate the head movement using this displacement computation. An optical flow tracker [26] is used to track points of feature for a human face. The tracking algorithm KLT tracks the face in two easy steps: first, it traces feature points in the first frame, and then it uses the calculated displacement to track the identified features in subsequent frames. Let's pretend that the first of the corner points are (x, y). The displaced corner point of the frame will be the sum of the original point and the displaced vector in the next frame if it is displaced by some variable vector $(b_1, b_2, \ldots b_n)$. The new point's coordinates will be $x' = x + b_1$ and $y' = y + b_1$. As a result, the displacement should now be calculated for each coordinate. The warp function, which is a function with coordinates and a parameter, is used for this. It's written as (5), and it uses the warp function to estimate the formation.

$$W(x, p) = (x + b_1, x + b_2) \tag{5}$$

4. RESULT AND DISCUSSION

From image extraction from the camera to generating our dataset with witness faces, classifying photos comprising witness and non-witness needed multiple stages. The data was divided into three sections: 60% for training, 20% for cross-validation, and 20% for testing. Also, testing in real-time video with frames in different brightness backgrounds. The proposed method's results were also compared to those produced using state-of-the-art approaches. For two separate epochs and iterations, Table 3 shows the accuracy reached by CNN of 99.5%. Figure 5 depicts a plot representing accuracy and loss values. Figure 6 depicts the proposed system in real-time action from a 180-degree angle (far right or left). Figure 6(a) and (b) shows two images on the left and right with a 180-degree angle perspective. Our proposed approach can be used to detect the face of a witness.

Table 4 shows the results of a comparison acquired using the proposed approach with the results acquired using state-of-the-art approaches. Face detection for biometrics-based on CNN [17] was created as part of the research activity. Their proposed model has a precession of 90%, which is a low value [27]. A deep C2D-CNN approach based on decision level function was employed in another study for face detection [15]. To improve face detection, Taherkhani *et al.* [16] presented a new model based on CNN.

Table 3. The proposed CNN method yields accuracy

Epoch	Iteration	Accuracy	Loss	Learning rate
1	1	99.5	0	1.0000e-05
5	10	99.5	0	1.0000e-05

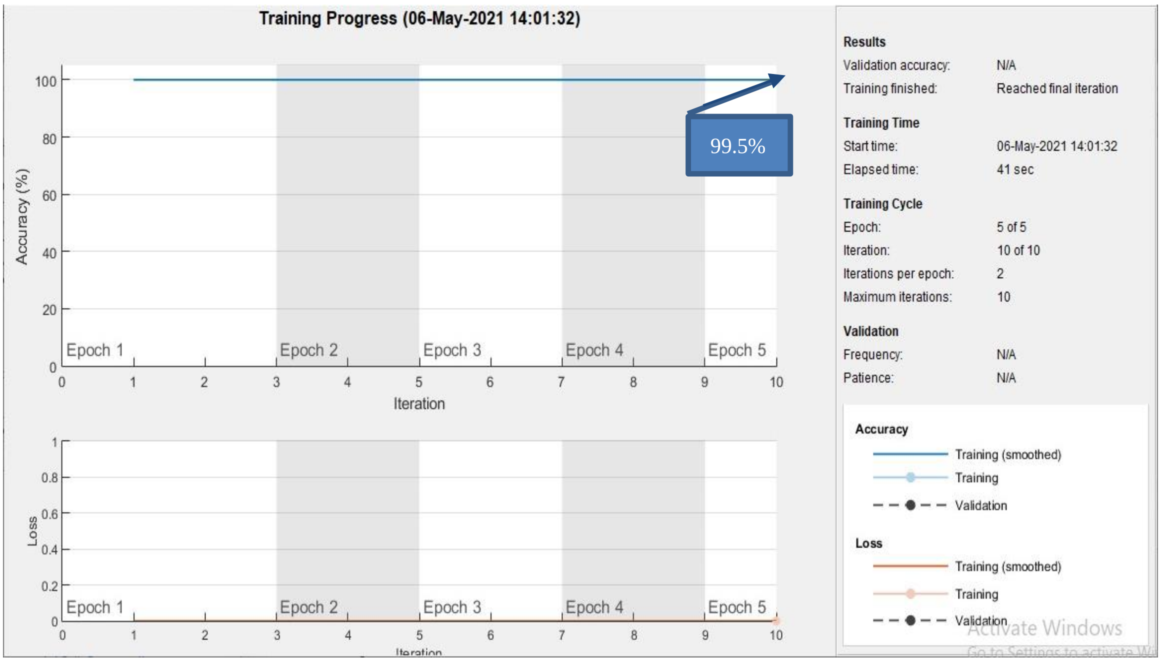
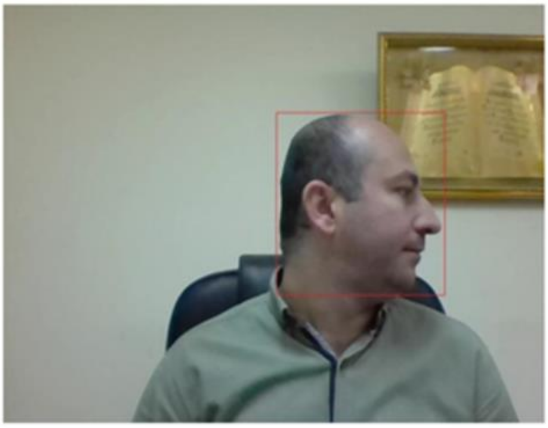
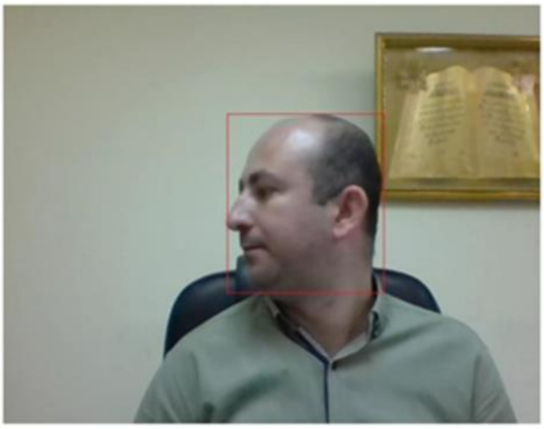


Figure 5. Performance plot showing accuracy and loos of the CNN method



(a)



(b)

Figure 6. Angle view of 180 degrees (a) far-right and (b) far left

Table 4. Methods that are currently in use are compared

No.	Author	Method	Dataset	Evaluation
1	Said *et al.* [17]	Develop face detection for a biometric method based on CNN	ORL dataset	Accuracy 98.5%
2	Hui Wang [27]	used ConvNets for face detection	Collected by the same author and has five classes	Precession 90%
3	Li *et al.* [15]	Used C2D-CNN for face detection based on decision level function	LFW and FRGC v2	Accuracy 91.98%
4	Taherkhani *et al.* [16]	This model is based on CNN. The output is separated into two branches: i) predicts facial attributes and ii) identifies face images.	CASIA-Web	Accuracy 78.82%
5	The proposed method	Viola-Jones algorithm for human face detection and CNN for training	Our dataset	Accuracy 99.5%

Table 4 demonstrates various state-of-the-art studies, with our proposed system achieving more accuracy than others in each research, with the proposed system achieving 99.5% accuracy for training and testing, respectively. In comparison to existing state-of-the-art methods, the suggested model was trained and tested on our dataset photos for each category. Our proposed model’s average accuracy is 99.5%.

5. CONCLUSION

Recently, many modern applications of face detection have appeared due to their great importance in security applications. In this paper, a new model of the face detection system is proposed based on the Merging of the Viola-Jones method and convolutional neural networks, which took place in two stages. The first stage is object detection, this is done by using the Viola-Jones algorithm as a pre-trained model. This algorithm is used for face detection to recognize each human face in an image coming from a video. In the second stage, CNN is applied with different parameters on an object that coming from the Viola-Jones algorithm to recognize which one is a witness. Because the proposed CNN outperforms the Viola-Jones algorithm in terms of accuracy because the Viola-Jones algorithm has specific viewing angles. Our dataset, which includes witness and non-witness images obtained as a result of real-time imaging, was used to train the proposed model. Finally, it is possible to improve the proposed model in the future to reduce the amount of space needed to implement CNN.

REFERENCES

[1] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, Jun. 2015, pp. 643–650, doi: 10.1145/2671188.2749408.
[2] A. D. Salman, H. B. Abdulwahab, "Proposed copyright protection systems for 3D video based on key frames," *Iraqi Journal of Computer, Communication, Control and System Engineering*, pp. 1–15, Jul. 2020, doi: 10.33103/uot.ijccce.20.3.1.
[3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, May 2009, doi: 10.1016/j.imavis.2008.08.005.
[4] A. R. Alawi and N. F. Hassan, "A proposal video encryption using light stream algorithm," *Engineering and Technology Journal*, vol. 39, no. 1B, pp. 184–196, Mar. 2021, doi: 10.30684/etj.v39i1B.1689.
[5] S. Singh and F. Nasoz, "Facial expression recognition with convolutional neural networks," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2020, pp. 0324–0328, doi: 10.1109/CCWC47524.2020.9031283.
[6] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2874–2883, doi: 10.1109/CVPR.2016.314.
[7] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 5325–5334, doi: 10.1109/CVPR.2015.7299170.
[8] H. Qin, J. Yan, X. Li, and X. Hu, "Joint training of cascaded CNN for face detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 3456–3465, doi: 10.1109/CVPR.2016.376.
[9] H. Mliki, S. Dammak, and E. Fendri, "An improved multi-scale face detection using convolutional neural network," *Signal, Image and Video Processing*, vol. 14, no. 7, pp. 1345–1353, Oct. 2020, doi: 10.1007/s11760-020-01680-w.
[10] F. Jalled and I. Voronkov, "Object detection using image processing," *ArXiv*, Nov. 2016, arxiv.org/abs/1611.07791.
[11] Z. Sufyanu, A. N. Musa, and K. A. Imam, "Approaches to face detection system-a survey," *International Review of Applied Engineering Research (IRAER)*, vol. 5, no. 3, pp. 5652–5658, 2016.
[12] K. Guo, S. Wu, and Y. Xu, "Face recognition using both visible light image and near-infrared image and a deep network," *CAAI Transactions on Intelligence Technology*, vol. 2, no. 1, pp. 39–47, Mar. 2017, doi: 10.1016/j.trit.2017.03.001.
[13] P. Kamencay, M. Benco, T. Mizdos, and R. Radil, "A new method for face recognition using convolutional neural network," *Advances in Electrical and Electronic Engineering*, vol. 15, no. 4, Nov. 2017, doi: 10.15598/aeee.v15i4.2389.
[14] S. Arya and A. Agrawal, "Face recognition with partial face recognition and convolutional neural network," *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, vol. 7, no. 1, pp. 91–94, 2018.
[15] J. Li, T. Qiu, C. Wen, K. Xie, and F.-Q. Wen, "Robust face recognition using the deep C2D-CNN model based on decision-level fusion," *Sensors*, vol. 18, no. 7, Art. no. 2080, Jun. 2018, doi: 10.3390/s18072080.
[16] F. Taherkhani, N. M. Nasrabadi, and J. Dawson, "A deep face identification network enhanced by facial attributes prediction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 666–6667, doi: 10.1109/CVPRW.2018.00097.
[17] Y. Said, M. Barr, and H. E. Ahmed, "Design of a face recognition system based on convolutional neural network (CNN)," *Engineering, Technology and Applied Science Research*, vol. 10, no. 3, pp. 5608–5612, Jun. 2020, doi: 10.48084/etasr.3490.
[18] K. Vikram and S. Padmavathi, "Facial parts detection using viola jones algorithm," in *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Jan. 2017, pp. 1–4, doi: 10.1109/ICACCS.2017.8014636.
[19] P. Viola and M. Jones, "Robust real-time face detection," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 747–747, doi: 10.1109/ICCV.2001.937709.
[20] T. Ephraim, T. Himmelman, and K. Siddiqi, "Real-time Viola-Jones face detection in a web browser," in *2009 Canadian Conference on Computer and Robot Vision*, May 2009, pp. 321–328, doi: 10.1109/CRV.2009.48.
[21] P. Tavallali, M. Yazdi, and M. R. Khosravi, "A systematic training procedure for Viola-Jones face detector in heterogeneous computing architecture," *Journal of Grid Computing*, vol. 18, no. 4, pp. 847–862, Dec. 2020, doi: 10.1007/s10723-020-09517-z.
[22] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *ArXiv*, Jun. 2015, arxiv.org/abs/1506.02025.
[23] G. Gilanie *et al.*, "Coronavirus (COVID-19) detection from chest radiology images using convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 66, Art. no. 102490, Apr. 2021, doi: 10.1016/j.bspc.2021.102490.
[24] J. K. Suhr, "Kanade-Lucas-Tomasi (KLT) feature tracker," *Computer Vision (EEE6503)*, 2009.
[25] R. Boda and M. J. P. Priyadarsini, "Face detection and tracking using KLT and viola jones," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 23, pp. 13472–13476, 2016.
[26] Z. Liu and Y. Wang, "Face detection and tracking in video using dynamic programming," in *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, vol. 1, pp. 53–56, doi: 10.1109/ICIP.2000.900890.
[27] S. Wang and K. Wang, "Real-time and accurate face detection networks based on deep learning," in *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, Oct. 2019, pp. 1541–1545, doi: 10.1109/EITCE47263.2019.9094843.

BIOGRAPHIES OF AUTHORS

Tameem Hameed Obaida received the B. Science degree in Computer Science from Babylon University, Iraq, in 2006 and the M.S.C degree in Computer Science/Information systems from Osmania University, India, in 2014. Currently, PhD student at the University of Technology in Iraq. Also she is a lecturer at the Najaf Technical Institute in the Department of Computer Systems Techniques, Al-Furat Al-Awsat Technical University. Her research interests include digital image processing, video processing, security and artificial intelligence applications. She can be contacted at email: tameem_hameed@yahoo.com.

Abeer Salim Jamil received the MSc. and PhD. in Computer Science from University of Technology, Iraq, 2004 and 2015 respectively. She has around 24 years of teaching experience and 11 years teaching in Cisco Network Academic (CISCO). Her areas of interests are Digital Image Processing, Video Processing, Security software Engineering, Networking and artificial intelligence applications. She can be contacted at email: abeer.salim@muc.edu.iq.

Nidaa Flaih Hassan received the MSc. and PhD. in Computer Science from University of Technology, Iraq, 1996 and 2005 respectively. She has around 25 years of teaching experience. Her areas of interests are Cyber Security, Multimedia Compression, and Image Processing. She can be contacted at email: 110020@uotechnolog.edu.iq.