Face Detection and Localization in Video Using HOG with CNN

Faqeda Hassen Kareem[1,*], Mohammed Abdullah Naser[1]

[1] College of Science for Women, University of Babylon, Iraq

Emails: faqeda.albermany.gsci141@student.uobabylon.edu.iq; wsci.mohammed.abud@uobabylon.edu.iq

Abstract

Face detection is important in computer vision and image processing, particularly in surveillance, security systems, video analytics, and facial recognition applications. However, face detection algorithms face challenges like position variations, lighting fluctuations, size and resolution differences, facial expressions, and background clutter. This research aims to develop a system that achieves high accuracy in detecting and localizing faces using local descriptors and spatial feature extraction techniques, specifically the Histogram of Oriented Gradients method (HOG). Using videos from the YouTube Face database, features were extracted from frames and trained using a convolutional neural network (CNN). The HOG technique achieved a 94% accuracy rate and good localization compared to CNN without feature extraction.

Keywords: Face detection; HOG feature extraction; CNN; Euclidean distance

1. Introduction

Computer vision is a subfield of computer science that enables machines to see and understand the content of images and videos. Videos are a sequence of images presented at a consistent frame rate. Object detection is a specialized task in artificial intelligence that focuses on accurately recognizing and locating objects within images or videos [1]. Face detection is a significant challenge in computer vision, given the growing importance of visuals in today's multimedia-driven society. The human face is a prominent and easily recognizable feature in image and videos [2]. Despite notable progress in facial detection methods, accurately and effectively detecting faces in real-world scenarios still requires improvement. Factors such as position, occlusion, scale, illumination, image quality, and facial emotions influence the performance of facial recognition technology [3]. Convolutional neural networks (CNNs) have achieved remarkable success in various computer vision applications, including picture classification, object identification, and semantic segmentation [4]. Deep learning algorithms offer an advantage over traditional computer vision systems by eliminating the need for manual design processes. These algorithms can be evaluated using established benchmarks like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [5]. In this research, the HOG method is employed to extract features from video frames, which are then fed into a CNN for improved face detection performance. The objective is to enhance accuracy by extracting significant characteristics from the image. The research is structured into sections covering previous research, the suggested methodology, experimental findings, analysis, and conclusions with an outlook on the proposed approach's prospects.

2. Related Works

This section provides a thorough examination of several methods employed for face detection, spanning from initial approaches to recent advancements. In recent times, multiple techniques have been developed to detect and classify facial characteristics.

Wankou Yang and et al (2019) [6]**.** This research introduces a new approach for cascaded convolutional neural networks with two primary stages. In the initial phase, a low-pixel candidate window is employed as an input, enabling the rapid extraction of the candidate window by the shallow convolutional neural network. During the

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

second stage, the window from the previous stage is resized and utilized as an input to the corresponding network layer. The detection accuracy of the discrete score on the FDDB dataset is 93.4%.

Zhishuai, and et al (2020) [7]**.** The research provides a novel approach for enhancing the performance of face detection algorithms. Resilience by collecting knowledge about minor facial features on complex images, they adopt VGG16 as the underlying convolutional neural network (CNN) architecture. This fusion process ultimately yields the desired detection feature map and achieves accuracy scores of 95.7%, 94.9%, and 89.7% on the easy, medium, and complex, respectively, on the WIDER FACE dataset.

Qi Guo, et al (2020)[8],They suggested By employing data augmentation, restructuring the detection network with deep separable convolution, and optimizing the NMS algorithm, the enhanced MTCNN algorithm outperforms the original MTCNN algorithm in processing multi-face detection tasks. The accuracy rate is 92%. The output of the MTCNN on face quality has been enhanced to improve the robustness of face identification and raise recognition accuracy for subsequent face recognition.

Shaoqi Hou and et al. (2021) [9]**.** They suggested use of multiscale Hybrid Pyramid Convolutional Network (HPCNet), a one-stage fully convolutional network with three modules: Hybrid Dilated Convolution, Hybrid Feature Pyramid, and Context Information Extractor. The HPCNet also introduces an enhanced Online Hard Example Mining technique for improved face detection accuracy. On the Easy subset of WIDER FACE, the approach achieved an accuracy of 93%; on the Medium subset, it achieved 92%; on the hard subset, it achieved 84%.

Qingqing Xu and et al. (2021) [10]. This study introduces an innovative two-tier face detection algorithm, SR-YOLOv5, designed to address the difficulties of detecting closely packed small faces in real-world situations. The first stage of the project focused on enhancing the structure and loss function used in YOLOv5. Later, attempts were made to improve the ability to detect faces in situations with blurry or low-resolution settings. The method attained accuracies of 96.3%, 94.9%, and 88.2% for the low, medium, and high difficulty levels, respectively.

3. Proposed System of Face Detection and Localization

The proposed model finds faces in the video by translating the video into individual frames, doing pre-processing on the frames, and directly feeding them into a (CNN) without extracting any additional features. In a later stage of the model, we utilize techniques to extract the features and feed them into the CNN to assess the detection accuracy of the method. As depicted in Figure (1):

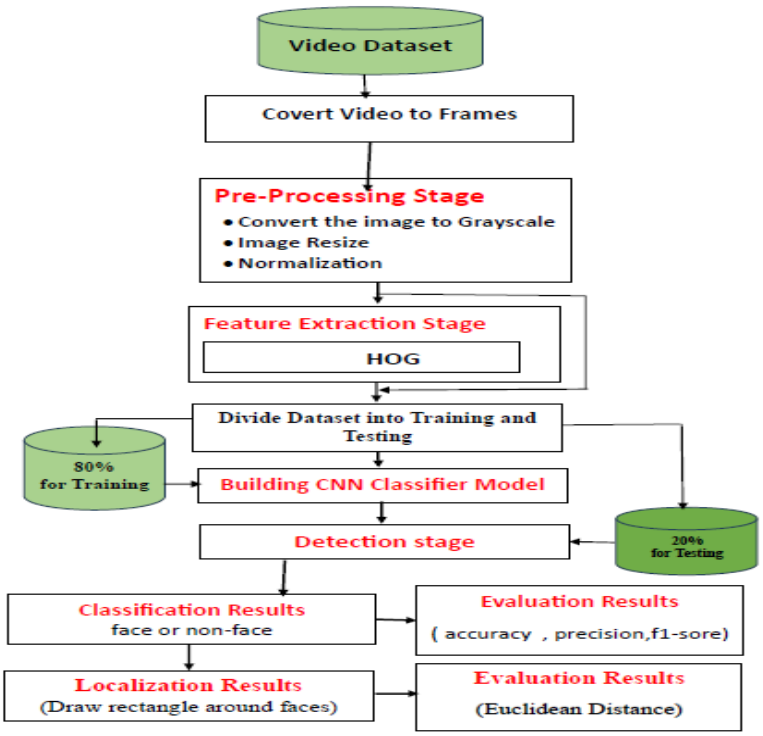

Figure 1. Displays the block diagram of the system that is being suggested

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

3.1. Dataset used

The dataset used for this study is the YouTube Faces dataset, often known as YTF. It contains over 3,000 videos of faces from more than 1,000 individuals, which were collected from YouTube. The average number of videos per person is two. The video segments vary in duration, with the shortest clip consisting of 46 frames and the longest 1 including roughly 6,000 frames. The collection comprises video clips with an average duration of around 180 frames[11] **.**

3.2. Converting a Video Stream into Consecutive Frames

The technique entails extracting individual frames from the video stream and saving them as separate image files. The process of extracting video frames is widely acknowledged as video frame extraction[12].

3.3. Data preprocessing: Image preprocessing is a methodical approach to manipulating an image or frame in order to prepare it for further processing. This involves organizing and filtering the data, as well as optimizing the dataset for extracting specific features. The preprocessing stage encompasses the execution of the subsequent procedures on the dataset:

3.3.1. Images Resizing: Resizing images to a fixed dimension of 224 x 224 pixels is a proactive measure to improve compatibility across different apps. Resizing images in deep learning systems reduces computational operations by reducing the computational cost of handling images with different sizes[3].

3.3.2. Grayscale Image Conversion: In order to decrease the number of channels in the input image graphs, it is essential to convert these images (frames) from three-dimensional Red, Green, and Blue (RGB) bands to grayscale. Using grayscale images improves the speed and efficiency of the verification process[2].

3.3.3 Normalization: Efficiently speeding up the training process and improving the generalization capabilities of deep neural networks (DNNs) is of utmost importance. It has demonstrated efficacy in a diverse array of applications. CNN gathers and examines a wide variety of images, with values ranging from 0 to 1. To get the desired outcome, each pixel with a value ranging from 0 to 255 is transformed by dividing it by 255, resulting in a new value between 0 and 1[13].

3.4. Feature Extraction Techniques:

The process of feature extraction is crucial in different applications, including diagnosis, classification, clustering, recognition, and detection. The technique involves transforming raw data into features that accurately capture essential elements of the data. [14].

The process of extracting features has a direct influence on the efficiency and accuracy of an application. Specialized expertise and understanding of the data are required for feature extraction. In order to optimize performance, researchers focus on refining and enhancing feature sets. The extraction approach depends on the data's features, the computational resources available, and the individual use cases. [15]. Applying methods to derive spatial characteristics from nearby data. Enhancing and converting the image to grayscale format will permit the application of local descriptors and provide more semantic information. The characteristics will be retrieved at the pixel level using **HOG** approaches, which will be explained in detail in the next section:

- **Histogram of Oriented Gradients (HOG)**

The feature descriptor has proven to be effective in identifying objects and pedestrians. The representation represents an object as a single value vector instead of a set of feature vectors where each vector refers to a particular area inside the image. The calculation of Hidden Optical Gradients (HOGs) entails the utilization of a sliding window detector that traverses the entire image. To acquire the corresponding HOG feature extraction, the HOG descriptor is computed for each point, while adjusting the image scale[16]. The first step in computing HOG characteristics is summed up as follows[17]:

Gradient calculation: During this step, the spatial gradients are calculated in both the horizontal and vertical directions. These two gradations are employed to compute the magnitudes and orientations of the gradient. as in mathematical equations(1), (2)[18]:

$$GX = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (1)$$

$$GY = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (2)$$

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

The gradients in the horizontal and vertical directions are denoted as G x and G y, respectively. It is worth mentioning that the Sobel mask can be employed to convolve images to do computations for equations (1) and (2). The algorithm computes the gradient of image intensity at each pixel inside the image in order to determine its operational properties. The identification of edges in images is a widely utilized method in the field of image analysis. The strategy may be influenced by the center difference, as it exhibits a preference for pixels located at the center. The process incorporates both differentiation and refinement. *Sobel edge detection* is a method that computes the gradient of every pixel within an image. Subsequently, the magnitude and direction of the gradient can be determined by employing equations (3) and (4)[18].

$$M=|GX|+|GY| \quad (3)$$

$$\theta(x , y)= |GX|+|GY| \quad (4)$$

Where θ (x, y) indicates the gradient's direction and m its magnitude.

- Orientation binning: During this stage, the image is divided into small, interconnected parts known as cells. The magnitude of the gradient for each pixel in a cell is partitioned into multiple orientation bins based on the gradient angle. In this stage, neighboring cells are arranged into groups, and each group goes through normalization. The process of combining normalized block histograms is used to generate a description inside a detection window[18].

3.5. The Convolutional Neural Network (CNN) Structure:

CNN is one of the most often used multi-layered deep neural network types in computer vision applications. It is a helpful tool for image recognition, object detection, and visualization. In AI, CNNs are the most popular neural networks for Deep Learning and image processing because they can handle large amounts of data [19].

The convolution layer is a crucial part of a convolutional neural network (CNN), which uses mathematical methods to extract features from an input image. It generates feature maps by applying input data arrays and collecting output for each location point. The max pooling layer reduces the parameter count and prevents overfitting by reducing the size of the feature map. There are two main types of pooling: average pooling and maximal pooling[19]. Flattening converts two-dimensional arrays into a single linear vector. The fully connected layer uses individual pixels from the flattened matrix to identify an image. The CNN architecture includes fully connected layers (FC) to classify incoming images based on training data. The final layer includes a loss function to optimize the model and improve prediction accuracy[20]. Below figure (2) is a simple depiction of the CNN model:

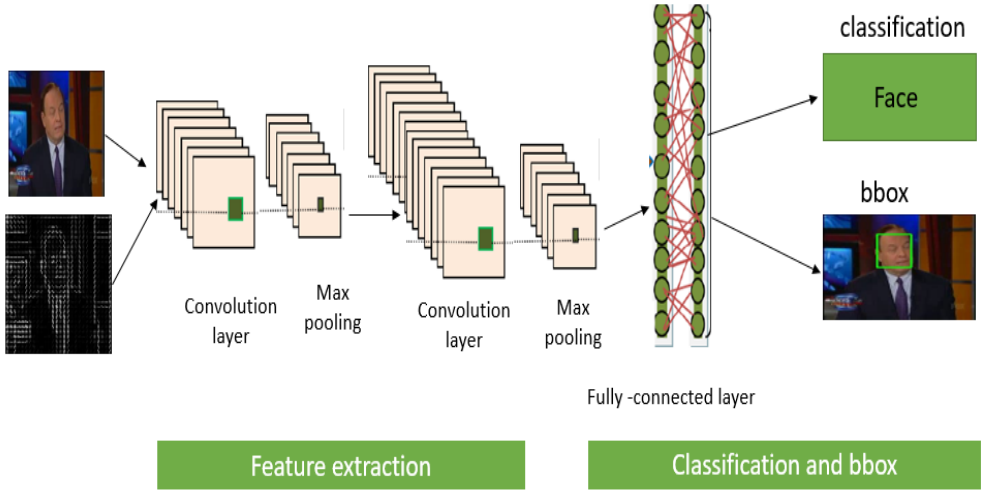

Figure 2. CNN model for face detection.

The (CNN) model is effectively classify facial characteristics in images and accurately determine their locations. This is a specialized (CNN) model designed exclusively for the purpose of detecting faces. It consists of two stages: the first stage classifies faces in images, while the second stage determines the precise location of the detected faces. The architecture incorporates several Conv2D layers for convolution and MaxPooling2D layers for pooling, which are then followed by Dense layers for data classification and then localization.

Figure (2) shows using the raw image as input to the CNN once, and again using images of spatial features and inputting them to the CNN.

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

3.5.1.CNN without feature extraction:

In this instance, the CNN receives the dataset directly, with pre-processing steps, and uses it as input. The dataset in this case does not have any extracted features. Next, classify the images into two categories: face and non-face. Then, identify the position of face.

3.5.2. CNN with HOG feature extraction method:

In this scenario, the CNN is provided with an image containing a specific feature, which undergoes pre-processing stages before being used as input. The dataset used in this case already contains extracted features obtained by the HOG approach. The next step is to categorize the images into two categories: face and non-face. Finally, CNN identifies the position of the face inside the image.as show in algorithm 1.

Algorithm 1: face detection model with **HOG** Method &CNN

Input: video face detection

Output : detection face and drew rectangle around them within video

Begin:
Step 1: import video
Step2 convert video to frames(images)and save them in (DS)
Step3: pre-process (frames)
- Resizing images to a fixed dimension of 224 x 224 pixels
- Grayscale Image Conversion
- Normalization

Step3: apply HOG method for feature extraction
- Repeat through each images in the dataset(DS).
- Calculate G x and G y by apply equation (1), (2)
- Calculate M and θ(x,y) defined by equation (3),(4)
- Return image of feature.

end repeat
Step4: Save Image features extracted as HOG _IMAGES
Step5: Input HOG _IMAGES to **CNN** for face detection by:
- Classify image to face or non –face
- Drew rectangle about location of face(**bbox**)for HOG_IMAGES

Step 6: Retrieve rectangle about location of face for original images
Step 7: Convert images with face(**bbox**)to video
End.

3.6. Evaluation result

In order to evaluate the model's performance, it will be evaluated once based on its ability to classify faces in the image. Once again, the evaluation is dependent on determining the exact location of the face within the image.

3.6.1. For classification

The model classifies images as either faces or non-faces and consequently utilizes evaluation metrics.
Accuracy (ACC): refers to the degree to which predictions are correct. The determination is made by dividing the number of accurate predictions by the total number of forecasts. This will assign a singular value to the entire network, Display equation (4).

$$\text{Accuracy(acc)}=\frac{TP+TN}{TP+TN+FN+FP} \quad (4).$$

3.6.2. For localization

to calculate the distance between the centroid of face(nose) and centroid of bounding box of detected face by used Euclidean distance

Euclidean Distance Loss: The loss function appears in this location. The primary application of embedding loss is in situations where two inputs need to be compared rather than in cases involving classification considerations. It calculates the distance between two locations or vectors. The equation (8) represents the Euclidean Distance Loss[21].

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

$$Euclidean\ loss=\sqrt{\sum_{i=1}^{n}(yi-pi)^2} \quad (8)$$

Let yi represent the input vector and n represent the length of the vector. The projected vector is denoted as pi.

4. Result and Discussion

In this part, we will discuss how to detect face in video frames despite the low resolution. It will be proven through training and testing of the model that inputting an image with feature into the CNN is better than inputting the raw image, as will be explained in detail with the figures:

- Utilize CNN exclusively without doing any feature extraction, as depicted in Figure (3): The utilization of CNN only for face detection yielded an accuracy rate of 92%. Although the accuracy is acceptable, it still needs to be improved compared to the accuracy achieved by the HOG feature extraction approach in Figure (4).

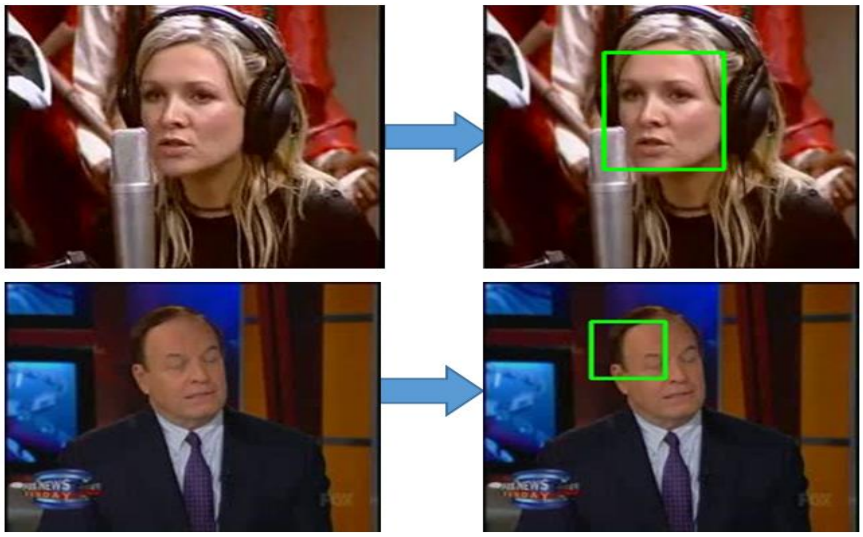

Figure 3. Show face detection using CNN without feature extraction

- The utilization of HOG in combination with CNN is illustrated in Figure (4). By applying the HOG feature extraction approach to preprocessing video frames prior to utilizing CNN, the detection accuracy was enhanced to 94%. The Histogram of Oriented Gradients (HOG) algorithm is adept at extracting features based on gradients and is highly proficient in representing shape and contour information. This approach likely facilitated effectively differentiating faces from the surrounding.

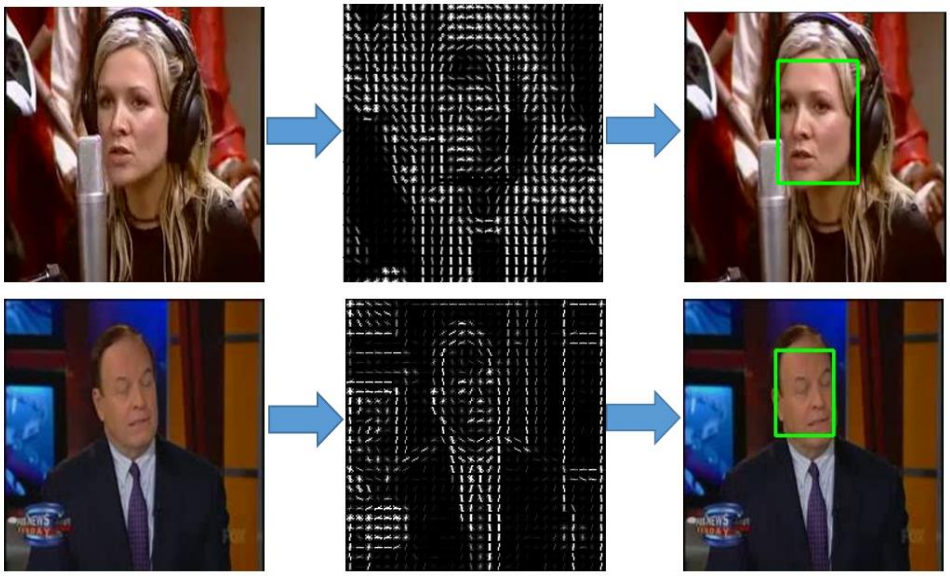

Figure 4. Show face detection using CNN with HOG feature extraction

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

After calculating the Euclidean distance to the square of the identified face, it is clear that relying alone on CNN for face identification produces inferior results compared to using HOG in combination with CNN. Upon computing the Euclidean distance to the detected face square, it is determined that the utilization of a combination of HOG and CNN for face recognition is superior to employing CNN alone. Figure (5) illustrates a comparison between the use of CNN and a combination of HOG and CNN.

(a) Evalut face localization with CNN

(b)Evalut face localization with CNN& HOG

Figure 5. (a), (b) show calculate Euclidean distance for location of face detected

Table (1) provides a comparison of various machine learning and deep face detection approaches. The new strategy outperformed the previous researchers' work in terms of classification measures, particularly accuracy, by using a distinct dataset.

Table 1: Comparison with Related Works

Researcher(s) Name	The Year	Method (s)	Dataset used	Accuracy
Wankou Yang and et al.[6]	2019	Cascade convolution structure	FDDB datasets	**93.4%**
Zhishuai , and et al.[7]	2020	CNN	WIDER FACE dataset.	**92%***
Qi Guo, et al [8]	2020	enhanced MTCNN	WIDER Face and FDDB datasets	**92%**
Shaoqi Hou and et al.[9]	2021	Hybrid Dilated Convolution, Hybrid Feature Pyramid, and Context Information Extractor	WIDER FACE	**89%***
Qingqing Xu and et al[10]	2021	SR-YOLOv5	WIDER FACE	**92%***
Our proposed	2024	ONLY CNN	YouTube Faces dataset	92%
		CNN & HOG		94%

*The meaning of accuracy for images categorized as easy, medium, and hard.

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

5. Conclusions

Face detection plays an important role in computer vision and processing images as it enables machine learning to detect and recognize human faces in videos. It is of greatest significance in the domains of surveillance, security systems, video analytics, and facial recognition applications. Detection algorithms have numerous challenges, including variations in position, fluctuations in lighting, disparities in size and resolution, facial expressions, and background clutter. Effective face detection requires the utilization of robust algorithms. This field utilizes feature extraction approaches, such as Histogram of Oriented Gradients (HOG). We performed a thorough examination of video frames obtained from the YouTube Face database. From these frames, we derived the features and employed a Convolutional Neural Network (CNN) to train a model capable of identifying faces and outlining them with bounding boxes. The empirical findings indicate that the CNN model, without employing feature extraction, acquired an accuracy rate of 92%, but the HOG technique attained a higher accuracy rate of 94%. They also used Ecclesiastes Distance to calculate the distance between the center of the face and the center of the rectangle of the detected face, and we showed that the results of the distance for the square detected by the HOG with CNN were better than the results of the CNN alone.

Reference

[1] Y. Xiao *et al.*, "A review of object detection based on deep learning," *Multimed. Tools Appl.*, vol. 79, pp. 23729–23791, 2020.

[2] K. A. Majeed, Z. Abbas, M. Bakhtyar, J. Baber, I. Ullah, and A. Ahmed, "Face Detectors Evaluation to Select the Fastest among DLIB, HAAR Cascade, and MTCNN," *Pakistan J. Emerg. Sci. Technol.*, vol. 2, no. 1, pp. 50–62, 2021.

[3] N. Zhang, J. Luo, and W. Gao, "Research on face detection technology based on MTCNN," *Proc. - 2020 Int. Conf. Comput. Network, Electron. Autom. ICCNEA 2020*, pp. 154–158, 2020, doi: 10.1109/ICCNEA50255.2020.00040.

[4] L. Alzubaidi *et al.*, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. big Data*, vol. 8, pp. 1–74, 2021.

[5] J. Dai, "NIPS-2016-r-fcn-object-detection-via-region-based-fully-convolutional-networks-Paper.pdf," no. Nips, 2016.

[6] W. Yang, L. Zhou, T. Li, and H. Wang, "A Face Detection Method Based on Cascade Convolutional Neural Network," *Multimed. Tools Appl.*, vol. 78, no. 17, pp. 24373–24390, 2019, doi: 10.1007/s11042-018-6995-0.

[7] Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, and A. Yuille, "Robust face detection via learning small faces on hard images," *Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020*, pp. 1350–1359, 2020, doi: 10.1109/WACV45572.2020.9093445.

[8] Q. Guo, Z. Wang, C. Wang, and D. Cui, "Multi-face detection algorithm suitable for video surveillance," *Proc. - 2020 Int. Conf. Comput. Vision, Image Deep Learn. CVIDL 2020*, no. Cvidl, pp. 27–33, 2020, doi: 10.1109/CVIDL51233.2020.00013.

[9] S. Hou, D. Fang, Y. Pan, Y. Li, and G. Yin, "Hybrid Pyramid Convolutional Network for Multiscale Face Detection," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/9963322.

[10] Q. Xu, Z. Zhu, H. Ge, Z. Zhang, and X. Zang, "Effective face detector based on YOLOv5 and superresolution reconstruction," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–9, 2021.

[11] E. Solomon, A. Woubie, and E. S. Emiru, "Autoencoder Based Face Verification System," 2023, [Online]. Available: http://arxiv.org/abs/2312.14301

[12] H. Yeo, C. J. Chong, Y. Jung, J. Ye, and D. Han, "Nemo: enabling neural-enhanced video streaming on commodity mobile devices," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[13] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10173–10196, 2023.

[14] W. K. Mutlag, S. K. Ali, Z. M. Aydam, and B. H. Taher, "Feature Extraction Methods: A Review," *J. Phys. Conf. Ser.*, vol. 1591, no. 1, 2020, doi: 10.1088/1742-6596/1591/1/012028.

[15] H. Fei, B. Tu, Q. Chen, D. He, C. Zhou, and Y. Peng, "An overview of face-related technologies," *J. Vis. Commun. Image Represent.*, vol. 56, no. September, pp. 139–143, 2018, doi: 10.1016/j.jvcir.2018.09.012.

[16] J. Kaur and W. Singh, "Tools, techniques, datasets and application areas for object detection in an image: a review," *Multimed. Tools Appl.*, vol. 81, no. 27, pp. 38297–38351, 2022, doi: 10.1007/s11042-022-13153-y.

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024

[17] W. Zhou, S. Gao, L. Zhang, and X. Lou, "Histogram of Oriented Gradients Feature Extraction from Raw Bayer Pattern Images," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 67, no. 5, pp. 946–950, 2020, doi: 10.1109/TCSII.2020.2980557.
[18] M. G. Mohammed and A. I. Melhum, "Implementation of HOG Feature Extraction with Tuned Parameters for Human Face Detection," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 5, pp. 654–661, 2020, doi: 10.18178/ijmlc.2020.10.5.987.
[19] A. W. Salehi *et al.*, "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023.
[20] V. H. Phung and E. J. Rhee, "A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets," *Appl. Sci.*, vol. 9, no. 21, p. 4500, 2019.
[21] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, IEEE, 2017, pp. 17–24.

DOI: https://doi.org/10.54216/FPA.170117
Received: November 28, 2023 Revised: March 19, 2024 Accepted: July 24, 2024