Face detection from in-car video for continuous health monitoring

Selvaraju, Vinothini, Spicher, Nicolai, Swaminathan, Ramakrishnan, Deserno, Thomas

Vinothini Selvaraju, Nicolai Spicher, Ramakrishnan Swaminathan, Thomas M. Deserno, "Face detection from in-car video for continuous health monitoring," Proc. SPIE 12037, Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, 1203708 (4 April 2022); doi: 10.1117/12.2612911

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Face detection from in-car video for continuous health monitoring

Vinothini Selvaraju *[a,b], Nicolai Spicher[b], Ramakrishnan Swaminathan[a] , Thomas M. Deserno[b]

[a]Department of Applied Mechanics, Indian Institute of Technology Madras, India;

[b]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

Email: vinothini.selvaraju@plri.de

ABSTRACT

Face detection in videos from smart cars or homes is becoming increasingly important in human-computer interaction, emotion recognition, gender and age identification, driving assistance, and vital sign measurements, as heart rate and respiratory rate is derived from the video. However, face detection suffers from variations in illumination, subject motion, different skin colors, or camera distances. We compare three algorithms for in-car application: Haar cascade classifier (HCC), histogram of oriented gradients (HoG), and a deep neural network (DNN). For evaluation, we consider the freely available "driver monitoring dataset" multimodal database (DMD) and self-collected videos recorded in a research car. We analyze run-time, accuracy, and F1-score. HoG has highest computational time as compared to HCC and DNN with 2.99 frames per second(fps), 7.00 fps, and 18.25 fps, respectively. For DMD, the F1-scores are 91.75%, 95.91%, and 99.48% for HCC, HoG, and DNN respectively, and 88.05%, 83.68%, and 99.66% for our dataset, respectively. All in all, DNN is fastest and most accurate. Moreover, we observed that DNN handles occlusions and varying illumination better than the other approaches. In conclusion, DNN can be applied successfully for in-car face detection as a first step towards real-time continuous health monitoring.

Keywords: Face detection, Haar cascade classifier, histogram of oriented gradients, deep neural network, in-car environment

1. INTRODUCTION

Face detection is prominent in computer vision for digital images or videos [1]. While it is rather simple for humans, it is challenging for algorithms and much efforts have been dedicated to this study area [2]. It has a wide range of applications namely, face alignment, face modelling, face verification, head pose tracking, human-computer interaction, facial expression recognition, gender identification, age recognition, driving assistance, surveillance, and vital signs measurement [3].

There are numerous articles on face detection [4-6]. However, face detection is still challenging due to a greater diversity in scale, position and posture of the body, head orientation, facial expression, illumination, background and occlusion. In addition, the presence of eye glasses, headgear, facial hair or masks for COVID-19 prevention, or any other objects partially obscuring the face, negatively influence the recognition [7].

Face detection identifies the position of faces in photographs or frames of videos. The methods vary with conceptual levels [3][8].

- Template matching correlates the image with predefined face patterns that are either manually extracted from standard images or parameterized by a function [9-10]. As a drawback, these types of methods suffer from changes in scale and viewing position as well as pose and gesture.
- Knowledge-based methods employ rules that are generated from the researcher's understanding of human faces [11]. It is nevertheless, difficult to convert human knowledge into rules. However, it is difficult to model the expert's knowledge and adopt this for the different instances [8].

Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications,
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 12037,
1203708 · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2612911

- Feature-based approaches extract landmark coordinates, color, and texture from eyebrows, eyes, nose, mouth, forehead, and cheek [8]. Haar cascade classifier (HCC) and histogram-oriented gradients (HoG) are widely used [12-15]. HCC employs the cascade of classifiers by feeding the extracted features from the integral image and the Adaboost algorithm. [14]. HoG extracts the intensity of gradients or edge directions in the image [12]. HCC is efficiently calculated and yields good performance for frontal faces but fails in extreme illumination conditions [16].
- Data based methods use statistical analysis and machine learning to find significant aspects of images showing and not-showing faces. They divide reference datasets into train and test sets, and then evaluate the result. Appearance-based methods sub-divide to eigenface-based, distribution-based, neural networks, support vector machine, naive bases classifier, hidden Markov models, and information theoretical approach [8][17]. Recently, deep neural networks (DNN) have showed remarkable performance [18-20].

Vital signs can be measured from face video due to variations in the reflected light [21]. For instance, cyclic body movement indicates the respiratory rate, and cyclic changes in skin color indicates the blood pulses and heart rate [22]. Unobtrusive assessment of vital signs is recommended for continuous health monitoring in hospitals, in smart homes monitoring [23], and smart cars [19],[24-25] as well as in pandemic situations (e.g., COVID-19). In a first step, the face must be accurately extracted. This is impacted by motion and illumination, which is least controlled in driving conditions. The above-mentioned approaches have several known pros and cons but it is still unclear which method is best suited for real time in-car videos. In particular, we pose the following research questions on HCC, HoG, and DNN:

- **RQ-A:** Which method achieves more accurate and fastest on face recognition?
- **RQ-B:** How does the indoor and outdoor environment affect the performance?
- **RQ-C:** How does the static and dynamic environment affect the performance?

2. MATERIAL AND METHODS

We consider two databases, namely the freely available driver monitoring dataset (DMD) multimodal database [26] and self-collected in-car videos recorded in a research car. We analyze the performance using accuracyF-score and run time.

2.1 Datasets

The DMD consists of two scenarios: outdoor environment in a real car and indoor environment in a driving simulator. In outdoor, the recordings were obtained under three different conditions i) sitting in a car, changing the steering wheel, and reaching behind, ii) various physical movement (e.g. hair make up, answering calls, steering wheel movement, messaging on the phone), iii) performing physical movement during driving (e.g. tuning the radio, drinking water, taling with a passenger, shifting gears). i) and ii) were performed while the car was stopped (static environment) and iii) was captured while during driving. The video lengths for i), ii), and iii) is on average one, seven, and four minutes respectively. For the indoor environment only one condition iv) was tested including diverse actions (e.g. simulated driving, hair make up, phone calling, and messaging, drinking water) for around nine minutes. In total, DMDcontains 20 videos from five subjects with light skin color. They attached an RGB camera (RealSense D415 camera, Intel Corp., Santa Clara, CA, USA) to the windshield, which delivers 1920 x 1080 pixels and 30 frames per second (fps)) [26].

We recorded additional data similar to DMD experiments i) and ii) with a windshield-mounted RGB camera (Realsense D435i, Intel Corp., Santa Clara, CA, USA). It also delivers 1920 x 1080 pixels at 30 fps. We recorded seven videos of two minutes' length from seven subjects in a real car but in a static environment with actions like fastening the seat belt, changing the steering wheel, adjusting radio, shifting gears, and changing facial expressions. Volunteers have light skin (three subjects), brown skin (two subjects) and dark skin (two subjects). All volunteers provided written consent prior to their participation in this study.

2.2 Face detection algorithms

We feed the videos to three different algorithms: HCC as provided by Dalal and Triggs [12-13], HoG as provided by Viola and Jones [14-15], and DNN as a Resnet-based architecture [18-20]. All algorithms are written in Python 3.8 using OpenCV v4.5.3. For HCC, we loaded the .xml file which is a pretrained model for frontal face. We use dlib library for HoG. We implement the DNN with caffe and TensorFlow. The caffe-based face detector requires the two files: the prototxt file, which defines the model architecture and the caffe model file which includes the weights [27]. We run the algorithms

on an off-the-shelf laptop (Latitude 5410, Dell Technologies Inc., Texas, USA) with an Intel core i7 10810U CPU and 16 GB RAM.

The pipeline for face detection loads the captured video and extracts the frame (Fig. 1). The video frames are looped over for detecting face only if frame number is less than or equal to the total number of frames in the video. Loop is terminated if the condition is not satisfied. Face detection algorithms namely, HCC, HoG, and DNN employed, which allows imported libraries to recognize the face. Thereby, each frame delivers with none, one, or multiple region-of-interests (ROI) containing a face for each algorithm.

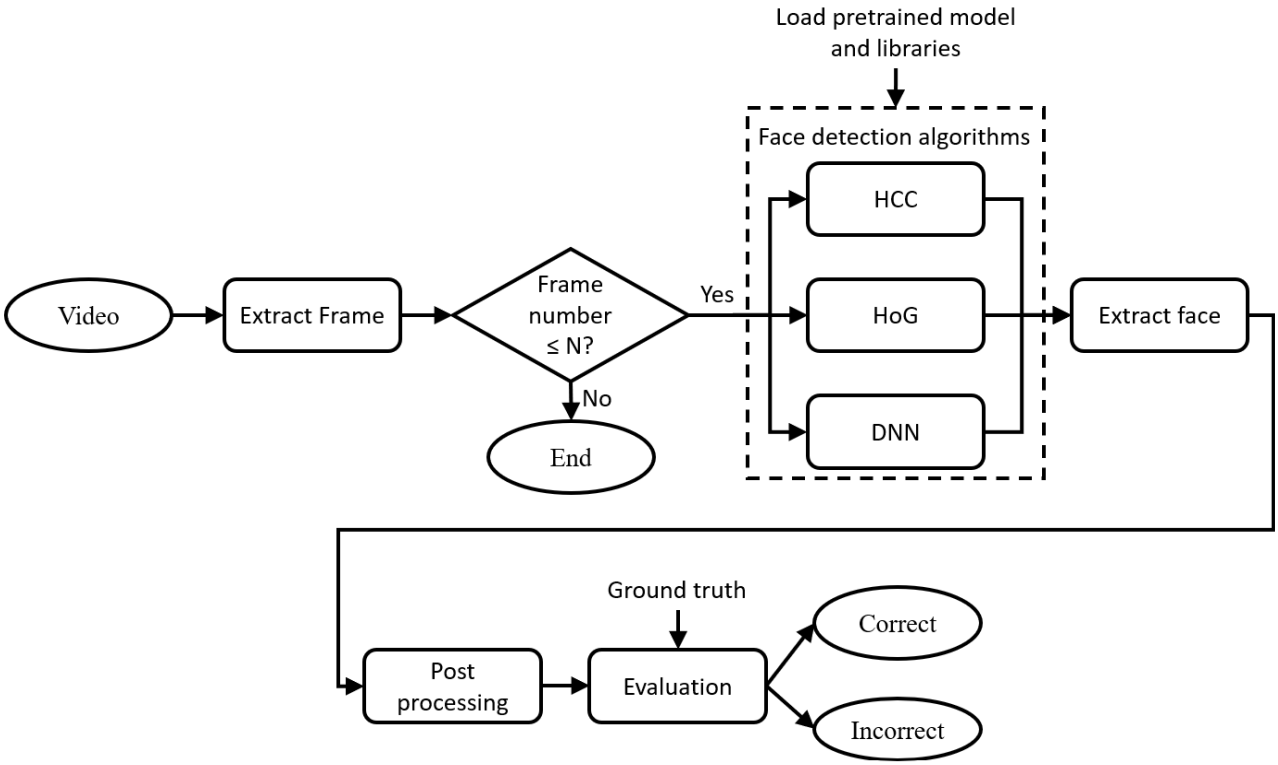


Figure 1. The pipeline of face detection algorithm (Note: N denotes the total number of frames in video)

2.3 Post Processing

To reduce false positive detections that are caused by changes in the illumination (e,g., reflections in the windshield), We calculated the height and width of the ROI and accept only ROI's area exceeding the area of 12% of the entire image. We determined this threshold empirically using the camera's aperture and its distance to the driver's seat.

2.4 Evaluation

Every frame of each input video was checked manually for the presence of the face before being fed in to the pipeline. Some frames in the self-recorded videos do not contain a face (e.g., when turning to fasten the safety belt) and were labeled as "no face". The remaining frames were found to show one face only. The output of the algorithms is also analyzed by manual inspection. If the detected ROI is fully covering the face, the frame is labeled "correct". If the face is missed, only partly covered, or surrounded by a substantial portion of background, the frame is labeled "incorrect". For quantitative assessment, we compare runtime, accuracy, and F1-score [28].

- Accuracy: the ratio of correctly classified frames to total number of frames.
- F-score: the harmonic mean between precision and recall. Precision is defined as the ratio of correctly classified Face frames to total number of frames, containing Face, whereas recall is the ratio of correctly classified Face frames to total number of frames with ground truth label of Face

3. RESULTS

3.1 Face detection

Fig. 2 illustrates the detection of "faces" using various detection algorithms before post processing. DNN detected the face in all the frames successfully, while HCC misidentified the reflections on the glass and detects the small areas through the glass (Fig. 2-a, c-d). In addition, HoG also detected the small area due to the changes in illumination (Fig. 2-a), and incorrectly detected the face as "no face" (Fig. 2-c). Multiple false faces are eliminated after post processing (Fig. 3). DNN and HoG detected the faces correctly (Fig. 3). However, HCC misidentified the face as "no face" (Fig. 3-c).

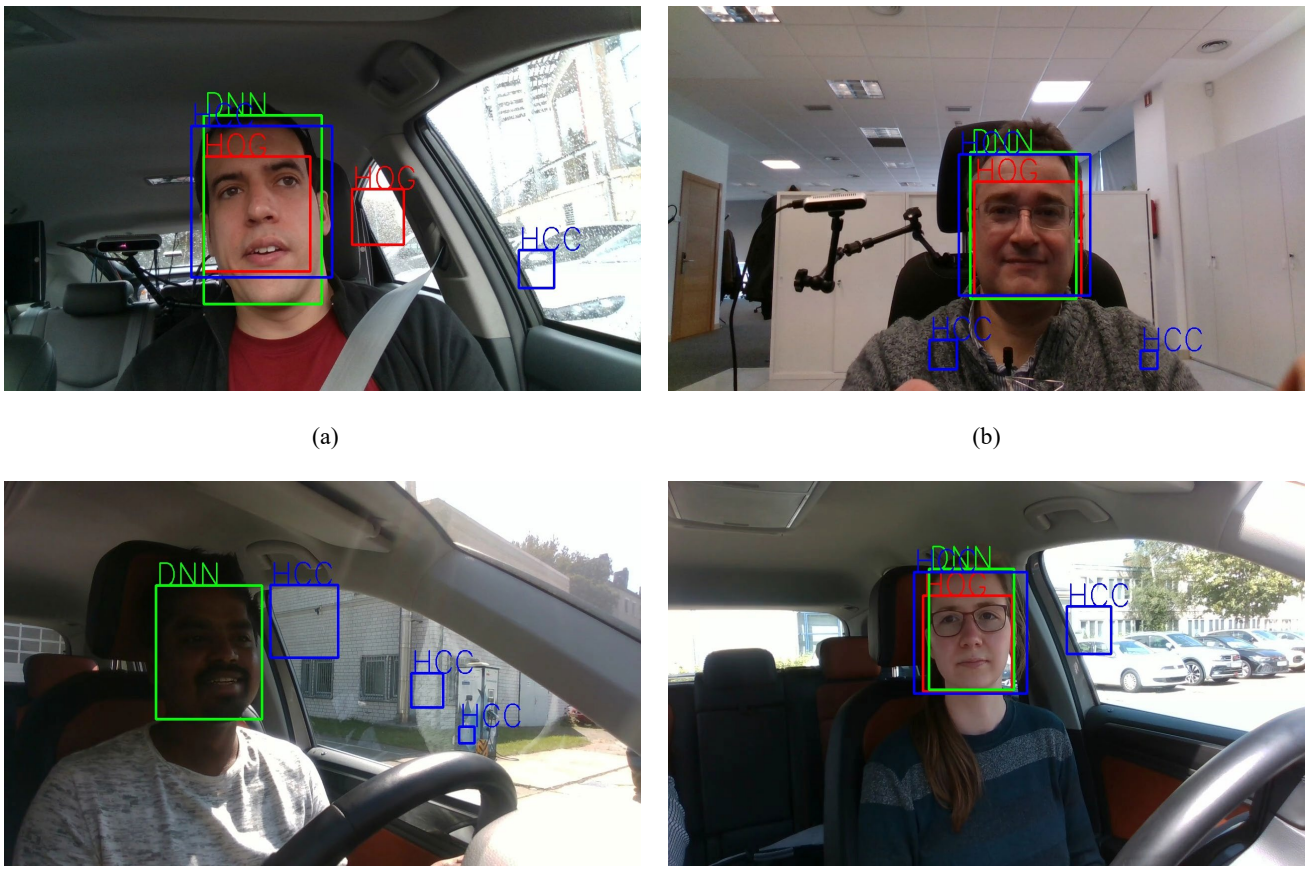
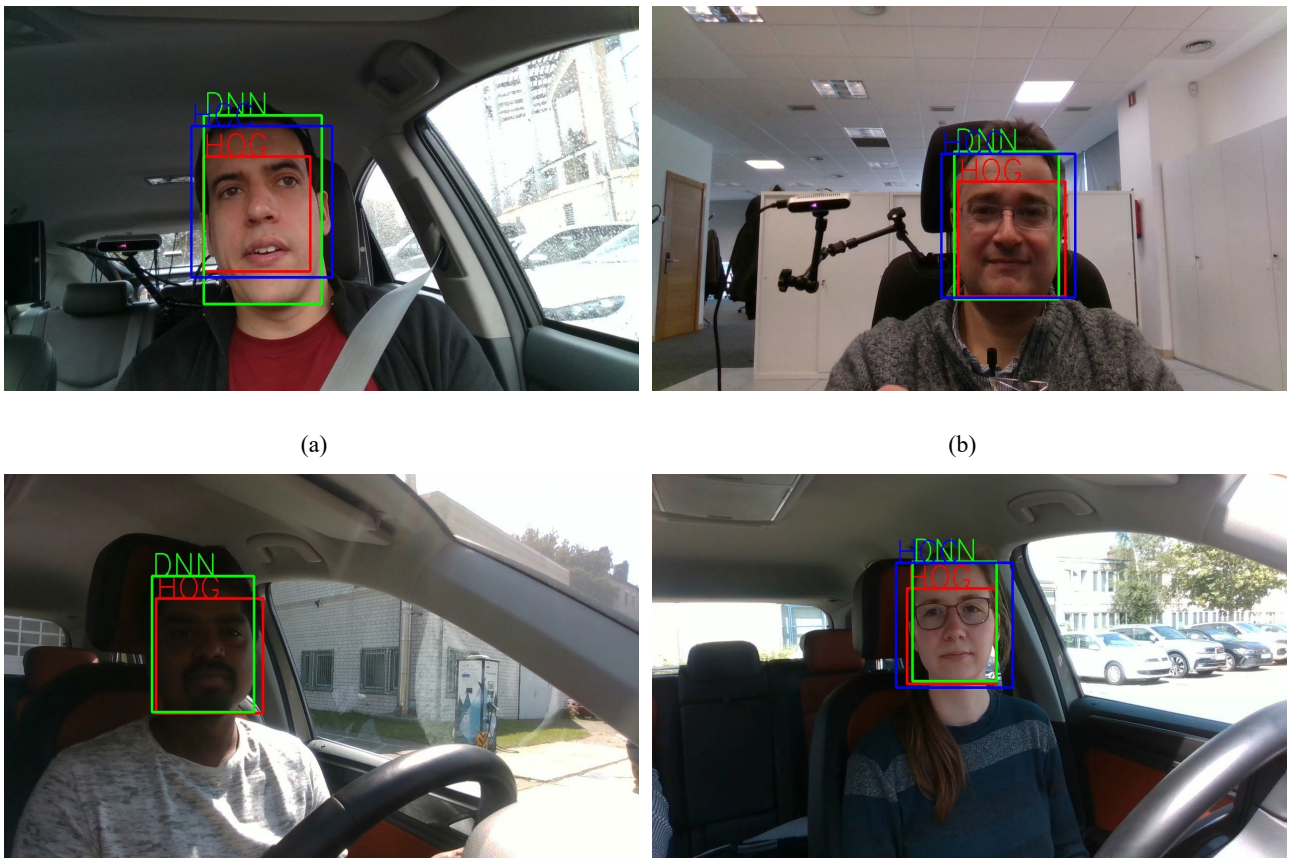


Figure 2. Example frames before post-processing a-b) stem from DMD, and c-d) stem from our own self recorded dataset. Red, blue, and green color boxes indicate HoG, HCC, and DNN respectively.

Accumulating all data, we obtained the best results with the DNN (accuracy: 98.96% and 99.34%; F-score: 99.48% and 99.66%) compared to HoG (92.15% and 72.34%; 95.91% and 83.68%) and HCC (84.76% and 67.21%; 91.75% and 80.05%) for DMD database and self-recorded database, respectively.

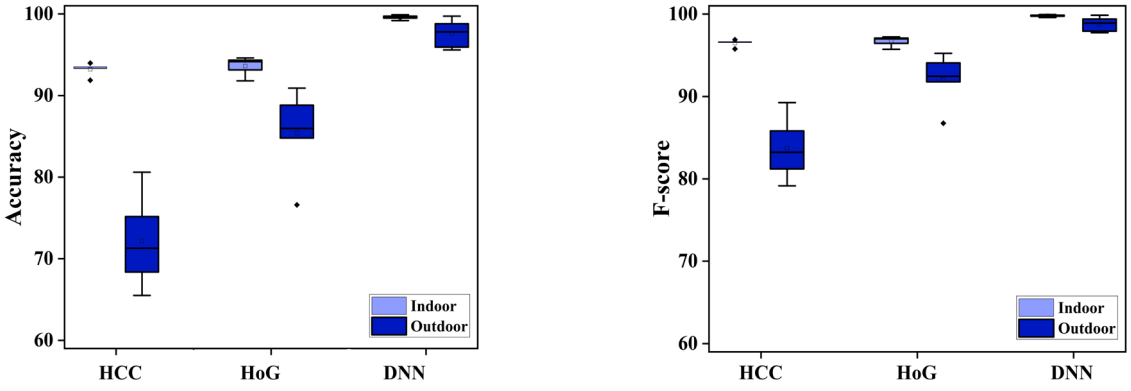
Moreover, we compared the indoor (iv) to the outdoor environments (iii) and the static (ii) to the dynamic environments (iii) in the DMD dataset (Fig. 4). We achieved improved performance in indoor environments (accuracy: 93.20%, 93.60%, and 99.58%; F-score: 96.48%, 96.69% and 99.79%) compared to outdoor environments (accuracy: 72.18%, 85.41%, and 97.57%; F-score: 83.73%, 92.05%, and 98.76%) for HCC, HoG, and DNN respectively (Fig. 4-a-b). We also observed the more accurate results in static environments (accuracy: 80.47%, 93.56%, and 98.98% and F-score: 88.93%, 96.64%, and 99.48%) compared to the dynamic environments (72.18%, 85.41%, and 97.57%; and 83.73%, 92.05%, and 98.76%) (Fig. 4-c-d). The best performance was obtained using the DNN which handles partial occlusions, motion, illumination variances, and environment (Fig. 4).



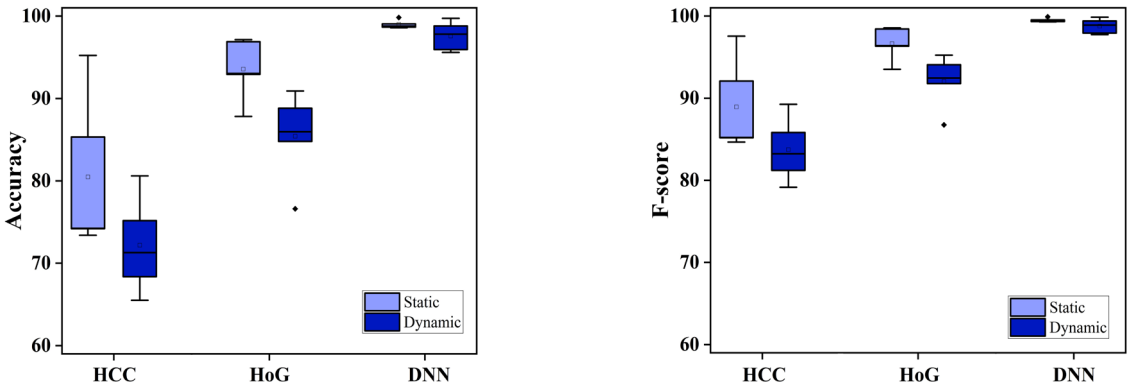
(a) (b)

(c) (d)

Figure 3. Example frames of a video from the dataset a-b) DMD and c-d) self-recorded after post processing



(a) (b)



(c) (d)

Figure 4. Accuracy and F-score for a-b) indoor vs. outdoor environments and c-d) static vs dynamic environments

3.2 Processing speed

The average processing speed is calculated to 7.00 fps, 2.99 fps, and 18.25 fps for HCC, HoG, and DNN, respectively (Fig. 5).

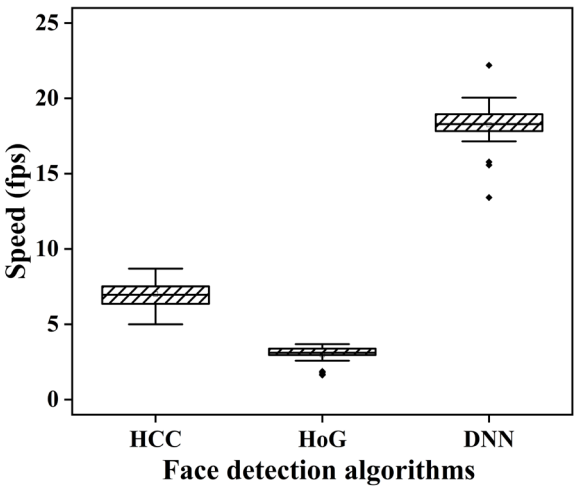


Figure 5. Processing speed for various face detection algorithms utilizing i7 processors

4. DISCUSSION

Our results show that an improved performance is achieved in indoor environments using a driving compared to outdoor environments in a real car. Most probably, this is due to the controlled illumination. Furthermore, we also attained the higher performance within static environments compared to dynamic environments which we attribute to the static background.

Furthermore, the lowest computation time is obtained for DNN, whereas HoG yields highest. From this, we answer the research question as follows:

- RQ-A: DNN obtains the highest accuracy and real-time speed in facial recognition (Fig. 5).
- RQ-B: Indoor environments provide high performance in all the considered algorithms (Fig. 4-a-b)

- RQ-C: Dynamic environments such as, movement, varying illumination and surroundings have a significant influence on the algorithm's performance when compared to static conditions (Fig. 4-c-d).

5. CONCLUSION

In this work, we analyze the suitability of three face detection algorithms for in-car face detection as a first step towards vital sign measurement. This method could also be used in numerous other fields of clinical (e.g., vital sign measurement in hospitals, pain detection) and nonclinical applications (e.g., smart home [23], smart car [25], surveillance systems). However, face detection is a complex task in real time environments due to the various challenges namely, dynamic environments, movements, varying skin color and illumination. Using the DMD dataset and a self-recorded dataset, we observe that DNN performs best in real-world situations. In addition, a higher processing speed was observed and compared to other algorithms. Our results might be useful for other researchers in the selection of the appropriate algorithm for in-car face detection.

REFERENCES

[1] Kaur P, Krishan K, Sharma SK, Kanchan T. Facial-recognition algorithms: a literature review. Med Sci Law. 2020;60(2):131-9.
[2] Zhang C, Zhang Z. A Survey of Recent Advances in Face Detection, Microsoft Res., Tech. Rep. MSR-TR-2010-66, 2010.
[3] Zafeiriou S, Zhang C, Zhang Z. A survey on face detection in the wild: past, present and future. Comput Vis Image Underst. 2015;138:1-24.
[4] Han CC, Liao HY, Yu GJ, Chen LH. Fast face detection via morphology-based pre-processing. Pattern Recognit. 2000;33(10):1701-12.
[5] Bhele SG, Mankar VH. A review paper on face recognition techniques. IJARCET. 2012;1(8):339-46.
[6] Albiol A, Monzo D, Martin A, Sastre J, Albiol A. Face recognition using HOG–EBGM. Pattern Recognit Lett. 2008;29(10):1537-43.
[7] Kumar A, Kaur A, Kumar M. Face detection techniques: a review. Artif Intell Rev. 2019; 52(2):927-948.
[8] Yang MH, Kriegman DJ, Ahuja N. Detecting faces in images: A survey. IEEE Trans PAMI. 2002;24(1):34-58.
[9] Miao J, Yin B, Wang K, Shen L, Chen X. A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template. Pattern Recognit. 1999;32(7):1237-48.
[10] Chai TY, Rizon M, Juhari M, Woo SS, Tan CS. Facial features for template matching based face recognition. Am. J. Appl. Sci. 2009;6(11):1897-1901.
[11] Yang G, Huang TS. Human face detection in a complex background. Pattern Recognit. 1994;27(1):53-63.
[12] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proc CVPR. 2005;886-93.
[13] Ryu J, Hong S, Liang S, Pak S, Chen Q, Yan S. Research on the combination of color channels in heart rate measurement based on photoplethysmography imaging. J Biomed Opt. 2021;26:025003.
[14] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proc CVPR. 2001; 511-18.
[15] Qi L, Yu H, Xu L, Mpanda RS, Greenwald SE. Robust heart-rate estimation from facial videos using Project_ICA. Physiol Meas. 2019;40(8):085007.
[16] Yadav S, Nain N. A novel approach for face detection using hybrid skin color model. J Reliab Intell Environ. 2016;2(3):145–58.
[17] Rowley HA, Baluja S, Kanade T. Neural network-based face detection. IEEE Trans PAMI. 1998;20(1):23-38.
[18] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot multibox detector. Proc ECCV. 2016;21-37.
[19] Hsu GS, Xie RC, Ambikapathi A, Chou KJ. A deep learning framework for heart rate estimation from facial videos. Neurocomputing. 2020;417:155-66.
[20] Susillo Ridao A, Zheng Y. A Comparison of Tools and Libraries for In-Class Face Detection and Emotion Recognition. Proc SIGITE. 2020;295-295.
[21] Hassan MA, Malik AS, Fofi D, Saad N, Karasfi B, Ali YS, Meriaudeau F. Heart rate estimation using facial video: A review. Biomedical Signal Processing and Control. 2017;38:346-60.
[22] Zaunseder S, Trumpp A, Wedekind D, Malberg H. Cardiovascular assessment by imaging photoplethysmography–a review. Biomed Eng. 2018;63(5):617-34.

[23] Wang J, Spicher N, Warnecke JM, Haghi M, Schwartze J, Deserno TM. Unobtrusive health monitoring in private spaces: The smart home. Sensors. 2021;21(3):864.
[24] Warnecke JM, Boeker N, Spicher N, Wang J, Flormann M, Deserno TM. Sensor fusion for robust heartbeat detection during driving. Proc. EMBC. 2021; 447-450.
[25] Wang J, Warnecke JM, Haghi M, Deserno TM. Unobtrusive health monitoring in private spaces: the smart vehicle. Sensors. 2020;20(9):2442.
[26] Ortega JD, Kose N, Cañas P, Chao MA, Unnervik A, Nieto M, Otaegui O, Salgado L. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. Proc ECCV. 2020;387-405.
[27] Sati V, Sánchez SM, Shoeibi N, Arora A, Corchado JM. Face Detection and Recognition, Face emotion recognition through NVIDIA Jetson Nano. Proc. ISAml. 2020;177-85.
[28] Taufeeque M, Koita S, Spicher N, Deserno TM. Multi-camera, multi-person, and real-time fall detection using long short term memory. Proc. SPIE. 2021;1160109.