A Machine Learning Approach to assess the interlocutor's attention and understanding during robotic interactions

Caterina Borzillo[1,†], Federica Cocci[1,†]

[1]*Student of Artificial Intelligence and Robotics, DIAG, Sapienza University of Rome*

Abstract

The ability of a system to detect the heedfulness of the interlocutor in human-robot interactions, as well as in general human interactions, is fundamental to establishing a stable and lasting communication. The main issue, however, is the small amount of data needed for training such systems; to this purpose in this work first of all we create a brand-new dataset containing about 120k images, extending an existing one with the aim of improving the overall performance accuracy with respect to the Face Orientation detection task and in a second place we aim at solving another newly task-specific problem: determining whether the interlocutor is heedful or not based on eye orientation. Several experiments are conducted and subsequent comparisons are made to compare the results obtained with respect to existing models.

HIGHLIGHTS

- Creation of a new labeled Dataset containing about 120k images
- Successfully addressed the Face orientation task by achieving higher accuracy compared to recent systems
- Creation of the Dataset for the Eye orientation task plus implementation of 2 baselines solution for it
- Implementation of an Eye detection module to improve baselines' performances
- Conceptualization of a Modular Pipeline Architecture for overall Attention Detection task

1. Introduction

Assessing interlocutor attention in social interactions is crucial for enhancing the quality of the communication, understanding human behavior and developing intelligent systems such as social robots. Attention detection, in fact, plays a vital role in various aspects of human life because it allows to establish strong, meaningful and effective communication. An attention detection system, in fact, finds applications in numerous real-life situations across different relevant domains such as in the Driver Monitoring in which the detection of distraction is fundamental to prevent accidents caused by driver inattention, in the Education domain, in which the Attention detection can be employed to assess students' engagement and focus during lessons but also in Healthcare domains where the examination of the patient attention and engagement during therapy sessions or cognitive assessments is essential to evaluate the effectiveness of the treatments.

To address this, in recent years, several machine learning systems have been developed with the purpose of assessing the heedfulness of the interlocutor. Nevertheless, the main issue for the evolution and expansion of such systems is the scarcity of available data that makes the training difficult and complicated as they do not acquire the proper knowledge and abilities to discriminate an heedful with respect to a not heedful interlocutor. To tackle this issue, in 2022 [1] attempts to partially solve the problem of data scarcity by creating a large dataset for the attention recognition task: in particular they address the problem of attention detection based on the face orientation of users. In this paper, we worked to improve and to explore the core foundation of this project by starting from its fundamental elements and expanding it. Firstly, we have doubled our dataset for the task of classifying facial orientations, creating approximately 120k additional images containing 5 different face orientations. Secondly, we have studied and analyzed another problem concerning attention detection, specifically the analysis of attention based on eye orientation. There is limited research in this area, which prompted us to implement a simple model that could learn whether an individual's eyes are looking at the camera (heedful) or elsewhere (not heedful). One possible application, as mentioned earlier, could be in an educational context, such as an elementary school, where the teacher can receive real-time feedback on the students' level of attention; this feedback can help make the lesson more interesting and engaging, if necessary.

Another situation in which an attention detection system can be extremely advantageous is during Human-Robot interactions, where such a system can bring significant improvements to the quality of the interaction. For

EAI - AI for Visual Perception in HCI & HRI, Final Project

[†] These authors contributed equally.

borzillo.1808187@studenti.uniroma1.it (C. Borzillo); cocci.1802435@studenti.uniroma1.it (F. Cocci)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

example, if a robot is able to understand and interpret the non-verbal cues such as gaze direction and facial expressions of its interlocutor, it's plausible that the robot itself can adapt in a better way its behavior better towards the user, leading to a more effective and natural communication.

Another scenario where attention detection can be highly beneficial is in the context of assistive robots. These robots are designed to assist and support humans in various tasks, working alongside humans to provide physical assistance, social interaction, or cognitive support. By detecting the user's attention, these robots can anticipate their needs and intentions, becoming more responsive and proactive in their assistance or interaction.

However, it's important to note that currently there are not many efficient attention systems that study nonverbal human behavior. Therefore, it is crucial to further explore this field and create new datasets to facilitate the development of such systems. These systems not only have practical utility but also contribute to improving the quality of life for individuals in their interactions with robots, considering the increasing frequency of human-robot interactions in the near future.

Regarding some implementation details, to solve the 5-class classification task, we intuitively used both our brand-new dataset and the dataset of [1], resulting in a total of 235k images. Additionally, for both classification tasks (the one involving 5 face orientations and the one involving 2 eye orientations), we examined and applied various versions of the entire dataset. Firstly, we utilized the entire dataset consisting of integral images of users in different face orientation positions (center, up, down, left, right) and performed classification on them. Then, in the second step, we applied the face detection module to the entire dataset, resulting in a dataset where each image is cropped around the face. Once again, classification was performed on this dataset. The reason for using the face detection module is that by training the model on the cropped face dataset, we anticipate it encountering fewer difficulties in classifying the images, thus achieving higher accuracy.

2. Related Works

Facial expressions of humans play a vital role in social interaction. Typically, communication encompasses both spoken and unspoken elements. Non-verbal cues have been recognized as significant channels of communication [2], enabling humans to connect through eye contact, gestures, facial expressions, body language, and paralanguage. Attention serves as a prominent indicator, revealing the level of observation and communication among individuals. The visual focus of attention signifies the specific object or subject that captures a person's interest in a given moment. This focus can be gauged through the examination of eye gaze, head pose, and body orientation as reported in [3, 4, 5] studies. Authors of [6] investigate the issue related to the examination of eyes' center in nonfrontal faces: the head pose estimation represents a solution to this problem indeed this paper introduce a hybrid scheme to combine head pose and eye location information to obtain the gaze estimation with the usage of a transformation matrix obtained from the head pose. Actually estimating the head pose of a person is a crucial problem. There are lots of works where head pose estimation is performed by keypoints from the target face. In [7] a robust way to determine pose is presented and it consists of training a multi-loss convolutional neural network to predict intrinsic Euler angles. Also in [8, 9] machine learning approaches are proposed in order to classify head poses and in particular in [8] deep convolution networks such as ResNet50 are used. However, accurately measuring gaze in everyday and unrestricted environments poses significant challenges. As an alternative, estimating the visual focus of attention can be achieved through the utilization of a visual saliency map model. The saliency map model, initially introduced by Koch and Ullman [10], draws inspiration from psychological research on visual attention [11] and has proven effective in approximating the distribution of human attention. Supporting evidence from studies utilizing gaze measurements, papers [12, 13, 14] further confirm the alignment between saliency maps and actual human attention patterns. Instead of relying solely on eye gaze, head orientation and body direction are key determinants for discerning an individual's focus of attention (FoA), as emphasized in [15]. The detection of human attention holds paramount importance across various scenarios. It serves as a valuable tool for warning drivers about drowsiness and lapses in attention [16, 17, 18], evaluating levels of engagement [19, 20], and acts as a foundational element for facial expression recognition. In particular, a novel approach towards real-time drowsiness detection is proposed in [16] which is based on a deep learning lightweight model. The authors of [17] introduce a face monitoring system whose compactness is achieved by a multi-scale pyramidal face representation to capture local and global information which is passed to deep Convolutional Neural Networks for the classification. Also in [18] a CNN is applied to identify the driver's tiredness in real-time along with Haar-Classifier that is used for eye feature extraction. About the evaluation of the level of participation, applications with students are understandable and in fact both [19, 20] present machine learning approaches for systems which detect scholars' attention during lessons. Moreover, attention detection finds application in fields where the accurate identification of pain levels is critical [21, 22], as well as in diagnosing syndromes in infants [23]. The study of human attention extends beyond its applications in machine learning and encompasses extensive research in the field of biology. Scientists have delved into numerous methodologies and approaches to comprehend and identify human attention from a biological standpoint. Within the realm of neuroscience, investigations have utilized functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) to explore the neural underpinnings of attention. These techniques enable researchers to examine and analyse the neural correlates associated with attentional processes. The degree of attentiveness exhibited by students throughout their learning process holds immense influence over their overall learning efficacy. By promptly discerning whether students are attentive or not, educators can offer timely reminders to help them sustain focus, ultimately augmenting the outcomes of the learning experience. In [24] authors aim to determine students' attentiveness during instruction by analysing their EEG signals, collected using mobile sensors, using a support vector machine (SVM). As already citeted, [1] is a remarkable work since the authors make a manually annotated dataset consisting of 120,000 images of people (important because of the scarcity of the data in the field of human attention detection) and, using as baseline VGG16, they get satisfactory results adding also the GAN-based data augmentation technique. In its study, [25] puts forth a mathematical model that leverages image processing techniques to detect head movement where they predominantly utilized the Lucas-Kanade (LK) algorithm for pattern recognition. On a similar note, [26] developed the OHMeGA analyzer, a system that detects head movement by employing an optical flow-based method and a gesture analyzer. Accurate estimation of head motion is crucial for their approach. Moving forward, [27] proposed a computer vision technology centered around eyeball tracking and detection. They developed a MATLAB-based system, and in their work, they focused on creating a pupil tracking-based application for controlling mouse movement using a webcam. In a study conducted by Frutos-Pascual et al. [28], attention skills were assessed in a group of children aged 8 to 12 years. The researchers analysed the children's gaze patterns and communication using eye-tracking devices. They achieved a classification accuracy of 88% by employing a random forest classifier. However, the study faced certain limitations. Firstly, the dataset was relatively small, consisting of only 32 children. Additionally, the processing speed of the system utilized in the study was relatively slow. In [29] authors introduced a system that utilized Raspberry Pi to estimate head direction and track humans. The system employed the Haar cascade classifier for detecting human faces. A triangle similarity-based approach for detecting eyeball movement was proposed, utilizing the Haar cascade classifier to detect the face and applying the Hough transform to identify the eyeball area [30]. This method enables the detection and tracking of eyeball movement using geometric principles. The Viola-Jones object detection algorithm offers an alternative approach for detecting facial features [31]. In the context of driver fatigue/drowsiness detection, a real-time monitoring system [32] was proposed, utilizing the Haar Cascade file for face detection. However, this study has a notable limitation: the system may fail to function properly under low-light conditions and when the driver is wearing sunglasses. The AdaBoost algorithm possesses the capability to adapt sample weights based on the classification error rate of the weak classifier. By iteratively constructing a classifier with the lowest error rate, AdaBoost creates a strong classification model. In the context of human eye detection, a different study [33] proposes an algorithm based on AdaBoost. Additionally, Dasgupta et al. [34] introduce the Percentage of Eye Closure (PERCLOS) system as a mean to quantify attentiveness levels. Another study [35] presents a range of machine learning techniques for predicting the degree of visual focus in human attention. Eight different classifiers, namely Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), AdaBoost, Multilayer Perceptron (MLP), Extra Tree classifier, and Voting Classifier, were trained to classify the attention level of each participant into one of three classes: High, Average, and Low.

3. Dataset creation

As mentioned earlier, the issue of limited data to train attention detection systems is quite pressing. For this reason, to assist future researchers in implementing new models and approaches to capture the heedfulness of human users, we have created a new dataset for the face orientation detection task. In particular, we added an additional collection of 120k images to the existing dataset of 120k images from [1]. This resulted in a large dataset of 240k images, divided into five different classes: center, up, down, right, and left. Each class corresponds to a specific face orientation, with the intuition that if interlocutors' face is oriented towards the camera (center class), they can be considered heedful, whereas if their face is oriented away from the camera (left, right, up, or down), they can be classified as not heedful. To obtain the images for the dataset, we recorded 125 videos for the training set in which a group of 7 users is involved and 50 videos for the validation and test sets in which 2 users are involved respectively. Each video lasts approximately 30 seconds, and users were asked to change their outfit or location every five videos to ensure a diverse and representative dataset. In Table 1, it's possible to see the number of samples extracted for each class in the training, validation, and test sets. Table 2, instead, provides an

Split	Class	#Samples
Train	CENTER	19.697
	DOWN	15.927
	LEFT	15.915
	RIGHT	16.012
	UP	15.909
–	–	–
Validation	CENTER	3.122
	DOWN	3.128
	LEFT	3.134
	RIGHT	3.133
	UP	3.230
–	–	–
Test	CENTER	3.867
	DOWN	3.035
	LEFT	3.112
	RIGHT	3.121
	UP	3.059

Table 1
Number of samples for each class in the dataset splits of BORZILLO-COCCI Dataset.

Split	Class	#Samples
Train	CENTER	36.170
	DOWN	33.768
	LEFT	33.929
	RIGHT	33.263
	UP	32.868
–	–	–
Validation	CENTER	5.568
	DOWN	5.801
	LEFT	5.707
	RIGHT	5.513
	UP	5.844
–	–	–
Test	CENTER	7.161
	DOWN	6.480
	LEFT	6.519
	RIGHT	6.818
	UP	6.353

Table 2
Number of samples for each class in the dataset splits of the merged dataset (Borzillo, Cocci, Pepe, Tedeschi et al. dataset)

overview of the total number of samples available used in this work for the attention detection task, with also the inclusion of the dataset from [1]. It's important to note that the training, validation, and test sets are disjoint, meaning the samples used to train the model are different from those used for validation, and likewise, the validation set samples are different from those used for testing, as different users were involved. About the age of the users, it's worth mentioning that since attention detection system applications encompass various domains, including children's healthcare and education, we included in our dataset the collection of frames of a 16-year-old girl in our dataset. For the other users, their ages ranged from 20 to 30 years old.

4. Method

In this section, we provide a comprehensive explanation of the experiments conducted for the two different attention detection tasks: attention detection task based on face orientation and attention detection task based on eyes orientation. Furthermore, as we will explain later, each experiment is part of a larger structure, a sort of pipeline, where we start with a full image representing the interlocutor and we gradually study his heedfulness, considering the crucial aspects of face orientation and eye orientation. Finally, after carefully analyzing and comparing all the conducted experiments, we will also compare them with existing systems to demonstrate the improvement in performances. Moreover, to ensure a fair comparison of our test results with those of other papers, we decided to employ the technique of transfer learning, as also done by [1]. The use of transfer learning technique is crucial because, for many years now, it has significantly influenced the field of computer vision, particularly in image classification tasks. For both our 5-classes and 2-classes classification tasks, we utilized the pre-trained VGG (Visual Geometry Group) model for feature extraction. The VGG model is a deep convolutional neural network architecture that has undergone extensive training on large-scale image classification datasets such as ImageNet. As a result, the VGG model has developed exceptional capabilities to extract visual features from images, capturing hierarchical and abstract representations of various objects, textures, and patterns present in the images. In the following subsections we will analyze in detail the implementation of the two separated classification tasks based on the Face orientation (Task 1) and on the Eye orientation (Task 2). Then, we will explain how we combine togheter the Face orientation module and the Eye orientation module in a third task (Task 3) where we use the results obtained by Task 1 to execute the Task 2 in order to obtain a final pipeline result.

4.1. Task 1: Attention Detection task based on face orientation

With regard to the attention detection task based on face orientation, we address the 5-classes classification problem, in which each class corresponds to the 5 different face orientations. As CNN, we use the VGG model because it has already been proven in [1] to be the most efficient and successful one with respect to other CNNs like ResNet50 or Xception. Concerning the dataset used for training, validation, and testing we use the Borzillo, Cocci, Pepe, Tedeschi et al. one which contains a total of 240k samples (see Table 2). As the initial experiment, we employed the entire collection of images for the classification task; then we trained the model on these images and evaluated its performance on the test set to assess its generalization ability to new and unseen images. As second experiment, following the approach of [1], we incorporated a Face detection module wherein we applied a pre-trained OpenCV Caffe Model face detector to crop the images specifically around the face. The result of this cropping is illustrated in the Figure 1. Intuitively, what we expected before running the experiment turned out to be true: the results obtained with this face-cropped dataset are better with respect to the results obtained with the non-cropped dataset.

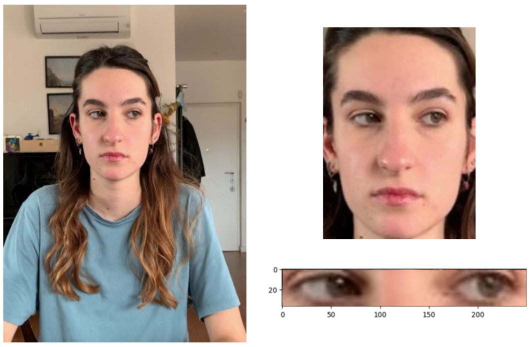

Figure 1: Example of face cropping and eyes cropping.

4.2. Task 2: Attention Detection based on eyes orientation

Regarding the second part of the work, we focus on a task that has received, in recent years, limited attention and investigation: the identification of the direction or orientation of the interlocutors' eyes in images. As a preliminary study that can be used for future research, this paper specifically addresses a binary classification problem. If the interlocutor's eyes are looking at the camera, they are classified as "heedful," whereas if the eyes are looking elsewhere away from the camera, they are classified as "Not Heedful." Although this two-classification task may seem straightforward, the problem of accurately determining the orientation of eyes can be challenging for a model due to various factors: these include the variability and complexity of eye appearance across individuals and images, as well as variations in shape, size, color, and texture. To tackle this task, we conducted three experiments, that differ based on the analyzed dataset. In the first experiment, we trained the model on the dataset with integral images for the 2-classification task. In the second experiment, we utilized a face-cropped dataset to classify eye orientation. Finally, we implemented an eye detection module, utilizing a machine learning toolkit called Dlib, to detect eyes in all images of the dataset and subsequently crop them for final classification. An example of this cropping carried out by the Eye Detector model is shown in Figure 1.

4.3. Task 3: Face Orientation + Eyes Orientation task (modular pipeline architecture)

So far, analyze in detail the implementations of Task 1 (Face Detection) and Task 2 (Eye Detection) separately. In Task 3, we aim at building a modular pipeline architecture where each part or component represents a distinct module that is sequentially connected to form a unified and integrated system. This "virtual" architecture, in fact, is composed of the 2 modules mentioned in the last two sections (4.1, 4.2): a Face Detection module and an Eye Detection module which are interconnected in a sequential manner. The pipeline architecture for task 3 is shown in Figure 2: the first module of Face Detection, which performs a 5-classes classification, is followed by a second module (Eye Detection) that, based on the results obtained from the previous step, performs the final 2-classes classification to determine the overall attentional state of a user. In particular, the second Eye Detection module takes in input only the samples that are classified as "Center" by the face detection module because all the others samples are automatically identified as "Not Heedful". In this way, we can interpret this architecture as a integrated system that, starting from a dataset of images classifies, firstly observing the face and secondly inspecting the eyes, whether a user is concentrated and attentive or not.

A potential future work could involve the practical development of this highly unified system, in which all components of the architecture are incorporated in a unique training model.

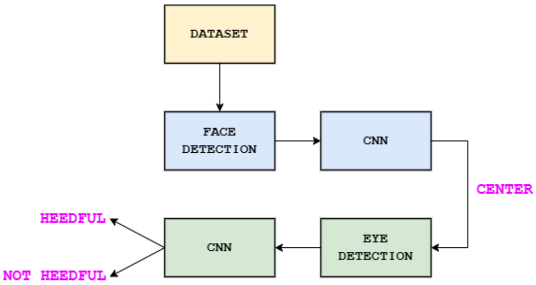

Figure 2: This is the pipeline introduced: a Face Detection module followed by an Eye Detection module.

5. Results

In this section all the results achieved with our models are described and commented. We highlight the fact that the reported accuracies (Tables 3, 4, and 5) are the ones evaluated at test time. More specifically, in Section 5.1, we discuss the results for the 5-class Face Orientation task (Task 1) using both the dataset of complete images and the dataset of cropped facial images. In Section 5.2, we examine the results of the 2-class Eye Orientation task (Task 2) using two different datasets: the dataset of cropped facial images and the dataset of cropped eye images. Finally, in Section 5.3, we present the outcomes related to Task 3 (Face Orientation + Eye Orientation). In this last subsection, the obtained results are the ones related to the continuous structure of the pipeline, where the samples are processed by both the Face detection and the Eye detection modules.

5.1. Task 1: Face orientation task

The obtained results for the Face orientation task are presented in Table 3. As we can notice, it is evident that the accuracy achieved by the model on the face-cropped dataset surpasses the accuracy on the whole image dataset: the reason is that on face-cropped images the model acquires, during training, a greater ability to detect the most significant features (such as the face orientation of the user) within images. Another point to mention is that there is an improvement also compared to the accuracies reported in the [1] paper; this means that as the dataset increases, the performance of the model improves correspondingly. In the Figure 3 is possible to analyse the results obtained with the face-cropped images.

5.2. Task 2: Eye orientation task

The results of two experiments conducted for the second task are presented in the Table 4. As we can observe, when analyzing images that exclusively depict the eyes, the model can focus more effectively on extracting relevant features and specific patterns related to eye orientation. This representation of the input allows the model to pay closer attention to distinctive characteristics such as the shape, position, and relative arrangement of pupils and eyelids. Consequently, the accuracy related to the face-cropped dataset is lower than the one related to the eye-cropped dataset.

TASK 1: FACE ORIENTATION DETECTION

Images type	Dataset	Accuracy
Entire face images	Pepe, Tedeschi et al.	73.31
Entire face images	Borzillo, Cocci, Pepe, Tedeschi et al.	**86.23**
Cropped face images	Pepe, Tedeschi et al.	86.17
Cropped face images	Borzillo, Cocci, Pepe, Tedeschi et al.	**88.48**

Table 3
Here's the accuracy values for the TASK 1.

TASK 2: EYE ORIENTATION DETECTION

Images type	Accuracy
Cropped face images	78.53
Cropped eyes images	**88.00**

Table 4
Here's the accuracy values for the TASK 2 (computed on Borzillo, Cocci, Pepe, Tedeschi et al. dataset).

TASK 3: FACE ORIENTATION + EYE ORIENTATION DETECTION

Images type	Accuracy
Cropped face images	78.52
Cropped eyes images	**87.02**

Table 5
Here's the accuracy values for the TASK 3 (computed on Borzillo, Cocci, Pepe, Tedeschi et al. dataset) where the dataset images undergo first the Face Orientation module and then the Eye Orientation module.

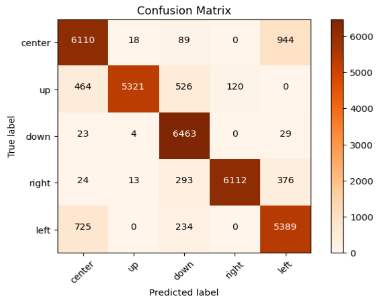

Figure 3: Confusion matrix of TASK 1, taken from the model which works with cropped face images.

5.3. Task 3: Face Orientation + Eye Orientation results (pipeline)

In Table 5 the results obtained for Task 3 (Face Orientation + Eye Orientation Detection) are presented and compared. In this experiment, the images from each dataset (face-cropped images and eye-cropped images) are processed within a modular pipeline. As we can realize, the results of Task 3 using cropped face images is lower with respect to the results of Task 3 using cropped eye images: the system, in fact, is more effective in discerning user attention (based on the eyes) by looking at images cropped around the eyes, with respect to those cropped around the face.

6. Conclusions

Finally, we can conclude that the amplification of the existing dataset in the attention detection task leads us to achieve better performances and results with respect to other recent systems. This increase in performance is achieved despite the small challenges that we encountered on Google Colaboratory (we trained each model for few epochs), therefore it means that by training these fine-tuned models for more epochs and with more computational resources, it is possible to achieve much higher accuracies.

Moreover, the introduction of a new dataset for the attention detection task based on eye orientation can be considered as a starting point for future studies concerning this topic that can be subject of further explorations. Furthermore, this dataset unlocks possibilities for studying the generalization capabilities of attention detection models across individuals with different age groups, cultural backgrounds, and varying eye characteristics.

Our hope, however, is that this work will be studied and analyzed in the future. A first improvement could be the accuracy and efficiency of the implemented mode, or also the construction of the already mentioned "virtual" pipeline architecture representing a unified real-time attention detection system, for example a device placed in a classroom of students that is capable of intercepting the students' attentiveness during the lesson, enabling the teacher to receive real-time feedbacks. Concluding, in our opinion, it's fundamental that experts and researchers carry on into the study of attention detection systems. In this way, firstly, the society could be able to benefit from these systems to enhance the quality of educational institutions and in general the quality of life and secondly, we would be able soon to open the doors to Human-Robot interactions in different situations and settings.

References

[1] S. Pepe, S. Tedeschi, N. Brandizzi, S. Russo, L. Iocchi, C. Napoli, Human attention assessment using a machine learning approach with gan-based data augmentation technique trained using a custom dataset, OBM Neurobiology 6 (2022). doi:10.21926/obm.neurobiol.2204139.

[2] A. Kendon, Gesticulation and speech: Two aspects of the process of utterance in m, The Relationship of Verbal and Nonverbal Communication 25 (1980).

[3] T. Afroze, Exploring the similarities of influencers in online brand communities, in: Proceedings of the 7th 2016 International Conference on Social Media Society, SMSociety '16, Association for Computing Machinery, New York, NY, USA, 2016. URL: https://doi.org/10.1145/2930971.2930981. doi:10.1145/2930971.2930981.

[4] J. M. Henderson, Visual Attention and Eye Movement Control During Reading and Picture Viewing, Springer New York, New York, NY, 1992, pp. 260–283. URL: https://doi.org/10.1007/978-1-4612-2852-3_15. doi:10.1007/978-1-4612-2852-3_15.

[5] D. W. Hansen, Q. Ji, In the eye of the beholder: A survey of models for eyes and gaze, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 478–500. doi:10.1109/TPAMI.2009.30.

[6] R. Valenti, N. Sebe, T. Gevers, Combining head pose and eye location information for gaze estimation, IEEE Transactions on Image Processing 21 (2012) 802–815. doi:10.1109/TIP.2011.2162740.

[7] N. Ruiz, E. Chong, J. M. Rehg, Fine-grained head pose estimation without keypoints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.

[8] S. Afroze, M. M. Hoque, Head pose classification based on deep convolution networks, in: R. Misra, N. Kesswani, M. Rajarajan, V. Bharadwaj, A. Patel (Eds.), Internet of Things and Connected Technologies, Springer International Publishing, 2021, pp. 458–469.

[9] S. Afroze, M. M. Hoque, Classification of attentional focus based on head pose in multi-object scenario, in: P. Vasant, I. Zelinka, G.-W. Weber (Eds.), Intelligent Computing and Optimization, Springer International Publishing, Cham, 2020, pp. 349–360.

[10] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human neurobiology 4 (1985) 219–227.

[11] A. M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychology 12 (1980) 97–136. URL: https://www.sciencedirect.com/science/article/pii/0010028580900055. doi:https://doi.org/10.1016/0010-0285(80)90005-5.

[12] T. Foulsham, G. Underwood, What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition, Journal of Vision 8 (2008) 6–6. URL: https://doi.org/10.1167/8.2.6. doi:10.1167/8.2.6.

[13] L. Itti, Quantitative modelling of perceptual salience at human eye position, Visual Cognition 14 (2006) 959–984. URL: https://doi.org/10.1080/13506280500195672. doi:10.1080/13506280500195672.

[14] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vision research 42 (2002) 107–123. URL: https://doi.org/10.1016/s0042-6989(01)00250-4.

[15] S. R. Langton, R. J. Watt, V. Bruce, Do the eyes

have it? cues to the direction of social attention, Trends in Cognitive Sciences 4 (2000) 50–59. URL: https://www.sciencedirect.com/science/article/pii/S1364661399014369. doi:https://doi.org/10.1016/S1364-6613(99)01436-9.

[16] R. Jabbar, K. Al-Khalifa, M. Kharbeche, W. Alhajyaseen, M. Jafari, S. Jiang, Real-time driver drowsiness detection for android application using deep neural networks techniques, Procedia Computer Science 130 (2018) 400–407. URL: https://www.sciencedirect.com/science/article/pii/S1877050918304137. doi:https://doi.org/10.1016/j.procs.2018.04.060.

[17] A. Moujahid, F. Dornaika, I. Arganda-Carreras, J. Reta, Efficient and compact face descriptor for driver drowsiness detection, Expert Systems with Applications 168 (2021) 114334. URL: https://www.sciencedirect.com/science/article/pii/S0957417420310241. doi:https://doi.org/10.1016/j.eswa.2020.114334.

[18] A.-U.-I. Rafid, A. I. Chowdhury, A. R. Niloy, N. Sharmin, A deep learning based approach for real-time driver drowsiness detection, in: 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2021, pp. 1–5. doi:10.1109/ICEEICT53905.2021.9667944.

[19] H. Monkaresi, N. Bosch, R. A. Calvo, S. K. D'Mello, Automated detection of engagement using video-based estimation of facial expressions and heart rate, IEEE Transactions on Affective Computing 8 (2017) 15–28. doi:10.1109/TAFFC.2016.2515084.

[20] A. I. KABIR, S. AKTER, S. MITRA, Students Engagement Detection in Online Learning During Covid-19 Pandemic Using R Programming Language, Informatica Economica 25 (2021) 26–37. URL: https://ideas.repec.org/a/aes/infoec/v25y2021i3p26-37.html.

[21] S. D. Roy, M. K. Bhowmik, P. Saha, A. K. Ghosh, An approach for automatic pain detection through facial expression, Procedia Computer Science 84 (2016) 99–106. URL: https://www.sciencedirect.com/science/article/pii/S1877050916300874. doi:https://doi.org/10.1016/j.procs.2016.04.072, proceeding of the Seventh International Conference on Intelligent Human Computer Interaction (IHCI 2015).

[22] S. El Morabit, A. Rivenq, M.-E.-n. Zighem, A. Hadid, A. Ouahabi, A. Taleb-Ahmed, Automatic pain estimation from facial expressions: A comparative analysis using off-the-shelf cnn architectures, Electronics 10 (2021). URL: https://www.mdpi.com/2079-9292/10/16/1926. doi:10.3390/electronics10161926.

[23] E. Vezzetti, D. Speranza, F. Marcolin, G. Fracastoro, G. Buscicchio, Exploiting 3d ultrasound for fetal diagnostic purpose through facial landmarking, Image Analysis Stereology 33 (2014) 167–188. URL: https://www.ias-iss.org/ojs/IAS/article/view/1100. doi:10.5566/ias.1100.

[24] L. Ning-Han, C.-Y. Chiang, , H. C. Chu, Recognizing the degree of human attention using eeg signals from mobile sensors, Sensors 13 (2013). doi:10.3390/s130810273.

[25] T. Siriteerakul, Y. Sato, V. Boonjing, Estimating change in head pose from low resolution video using lbp-based tracking, in: 2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), 2011, pp. 1–6. doi:10.1109/ISPACS.2011.6146173.

[26] S. Martin, C. Tran, A. Tawari, J. Kwan, M. Trivedi, Optical flow based head movement and gesture analysis in automotive environment, in: 2012 15th International IEEE Conference on Intelligent Transportation Systems, 2012, pp. 882–887. doi:10.1109/ITSC.2012.6338909.

[27] R. Ramesh, M. G. R. Rishikesh, Eye ball movement to control computer screen, Journal of Biosensors and Bioelectronics 6 (2015) 1–3.

[28] M. Frutos-Pascual, B. Garcia-Zapirain, Assessing visual attention using eye tracking sensors in intelligent cognitive therapies based on serious games, Sensors 15 (2015) 11092–11117. URL: https://www.mdpi.com/1424-8220/15/5/11092. doi:10.3390/s150511092.

[29] S. S. A. Abbas, M. Anitha, X. V. Jaini, Realization of multiple human head detection and direction movement using raspberry pi, in: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 1160–1164. doi:10.1109/WiSPNET.2017.8299946.

[30] R. P. Prasetya, F. Utaminingrum, Triangle similarity approach for detecting eyeball movement, in: 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), 2017, pp. 37–40. doi:10.1109/ISCBI.2017.8053540.

[31] J. Coster, M. Ohlsson, Human attention: The possibility of measuring human attention using opencv and the viola-jones face detection algorithm, Dissertation (2015).

[32] J. Suryaprasad, D. Sandesh, V. Saraswathi, D. Swathi, S. Manjunath, Real time drowsy driver detection using haarcascade samples, volume 3, 2013, pp. 45–54. doi:10.5121/csit.2013.3805.

[33] L. Shang, C. Zhang, G. Gao, Eye detection and attention recognition based on opencv, DEStech Transactions on Computer Science and Engineering (2018). doi:10.12783/dtcse/icmsie2017/18694.

[34] A. Dasgupta, A. George, S. L. Happy, A. Routray, A vision-based system for monitoring the loss of attention in automotive drivers, IEEE Transactions on Intelligent Transportation Systems 14 (2013) 1825–1838. doi:10.1109/TITS.2013.2271052.

[35] P. Chakraborty, M. A. Yousuf, S. Rahman, Predicting level of visual focus of human's attention using machine learning approaches, in: M. S. Kaiser, A. Bandyopadhyay, M. Mahmud, K. Ray (Eds.), Proceedings of International Conference on Trends in Computational and Cognitive Engineering, Springer Singapore, Singapore, 2021, pp. 683–694.