Open Access

OBM Neurobiology

Original Research

Human Attention Assessment Using A Machine Learning Approach with GAN-based Data Augmentation Technique Trained Using a Custom Dataset

Sveva Pepe [1], Simone Tedeschi [1], Nicolo' Brandizzi [1], Samuele Russo [2, *], Luca Iocchi [1], Christian Napoli [1, *]

1. Department of Computer, Automation and Management Engineering, Sapienza University of Rome, Via Ariosto 25, Roma 00185, RM, Italy; E-Mails: sveva.pepe@gmail.com; tedeschi@diag.uniroma1.it; brandizzi@diag.uniroma1.it; iocchi@diag.uniroma1.it; cnapoli@diag.uniroma1.it
2. Department of Psychology, Sapienza University of Rome, Via dei Marsi 74, Roma 00185, RM, Italy; E-Mail: samuele.russo@uniroma1.it

* **Correspondence:** Samuele Russo and Christian Napoli; E-Mails: samuele.russo@uniroma1.it; cnapoli@diag.uniroma1.it.

Academic Editor: Raul Valverde

Special Issue: Neuroscience and Information Technology

OBM Neurobiology
2022, volume 6, issue 4
doi:10.21926/obm.neurobiol.2204139

Received: May 27, 2022
Accepted: September 20, 2022
Published: October 04, 2022

Abstract

Human–robot interactions require the ability of the system to determine if the user is paying attention. However, to train such systems, massive amounts of data are required. In this study, we addressed the issue of data scarcity by constructing a large dataset (containing ~120,000 photographs) for the attention detection task. Then, by using this dataset, we established a powerful baseline system. In addition, we extended the proposed system by adding an auxiliary face detection module and introducing a unique GAN-based data augmentation technique. Experimental results revealed that the proposed system yields superior performance compared to baseline models and achieves an accuracy of 88% on the test set. Finally, we created a web application for testing the proposed model in real time.

Keywords
Attention assessment; dataset; GAN

1. Introduction

Human–robot interaction (HRI) [1], or more generally, human–computer interaction (HCI) [2], has received increased research interest in the last two decades thanks to the recent technological advances in the hardware and systems fields [3].

Extensive research has been performed to improve the quality of HRI by exploiting modern machine learning (ML) techniques [4, 5] to allow voice and gesture commands. Such interactions are largely based on visual perception, and the system must be able to distinguish whether the user is attentive. However, few approaches relying upon eye gaze have been proposed to date [6, 7].

Detecting attention is an essential stage in various applications. For example, when dealing with hyperactive children, knowing if the child is following the recommended action (e.g., an educational session) is necessary to decide whether the same activity must be presented again. Furthermore, in conventional HRI, if the user does not pay attention when the robot asks a question (e.g., "What do you want to order?" in a restaurant scenario), the robot can increase the volume and ask the question again to get the user's attention. In a driver monitoring system, the robot can check if the driver is looking ahead and decide to notify him and refocus his attention.

Visual tracking plays an important role in the assessment of human attention. This task requires merging many computer vision applications, which include image and video processing, pattern recognition, information retrieval, automation, and control. While the correlation between eye movement and attention level is generally acknowledged, there are no standards to define a direct mapping between the two. However, eye tracking is an extensive field and is employed in many applications, such as mobile robotics, solar forecasting, particle tracking in microscopy images, biological applications, and surveillance [8-11]. Although there are numerous use cases of gaze tracking and attention detection, few datasets are freely available online. This shortage is due to the lack of a clear and established baseline for labeling image data. Moreover, constructing a dataset is time-consuming and expensive.

Massive amounts of data are required to train deep learning (DL) systems capable of solving such tasks in a supervised manner.

In this study, we focused on the attention detection task in its entirety, from addressing the issue of data scarcity, building a baseline system, and proposing unique ways to boost performance further to developing a web application. For the attention detection task, we constructed a new manually annotated dataset containing approximately 120,000 images from 18 users. Moreover, we developed a strong baseline system to achieve competitive detection results on the developed dataset. Furthermore, we enhanced the baseline system by adding an auxiliary face detection module, enhancing the performance. In addition, we proposed a novel GAN-based data augmentation technique to further enrich the collected data and achieve performance improvements. We extensively evaluated the benefits of using the constructed test set by performing a statistical analysis of the obtained results. Finally, we designed a web application to further analyze the system’s behavior and perform real-time testing. With this work, we aim to encourage further studies on the development of high-performance attention detection systems.

The paper is structured as follows. In Section 2, a brief overview of the related works for the topics covered in the introduction is presented, particularly attention assessment through different human signals and their use in real-world scenarios. In Section 3, the methodology employed for constructing the dataset and some statistical analysis are described. In Section 4, the proposed architecture for assessing human attention is presented, and the baseline system and the augmentation implementation by using GAN are discussed. The experimental setup and results are presented in Section V. The real-time application developed for evaluation is described in Section 6. Finally, the conclusion and ideas for future research are presented in Section 7.

2. Related Works

Extensive research has been conducted on human attention detection from many perspectives, and thanks to recent advances in sensor technologies and computer vision techniques, the research focus has shifted from manual to automated techniques [12-14].

Previously, studies mainly employed experiments on physiological factors such as fatigue [15], brain signal data [16], blood flow and heart rate [17], and galvanic skin conductance [18]. Recently, efforts have been made to accomplish driver attention prediction as a computer vision technique. In a previous study [19], semantics-induced scene variations were analyzed using a novel multi-branch deep architecture that integrates three sources of information: raw video, motion, and scene semantics. In another study [20], a semantic context-induced attentive fusion network was designed. Eye-tracking [21, 22], gaze-tracking [23], and face-tracking algorithms [24] have also been proposed.

These methods are promising as they are inexpensive and unobtrusive [25-28]. In addition, the automatic recognition of facial expressions has resulted in applications that span several disciplines. For example, facial expressions are used to identify pain [29, 30], diagnose syndromes in infants [31, 32], detect driver’s drowsiness [33-35], recognize emotions [36-40], and detect engagement [41, 42].

These systems have been employed for important tasks such as the detection of driver attention and autism-related diseases. There has been an increased interest in gaze detection information to determine a driver’s focus of attention. In a previous study [43], a model based on Sanger’s neural network [44] was proposed to monitor real-time driver attention through binary classifiers and iconic data reduction. In another study [45], the driver’s attention behavior and the road scene were parsed to predict potentially risky maneuvers. A DL-based gaze detection approach that can detect the driver’s head and eye movement by using a near-infrared (NIR) camera sensor has been proposed [46]. The driver’s gaze in a pre-attention environment has been investigated using intention prediction based solely on fixation maps [47]. However, these systems suffer from the problem of grabbing the attention of a driver and rely on a restricted number of predefined safety-inspired rules. In a previous study [48], the driver’s attention toward pedestrians and motorbikes at junctions was inspected, and object saliency was employed to avoid the looked-but failed-to-see effect. On the contrary, in the absence of eye-tracking systems and reliable gaze data, several studies [49-53] have focused on the driver’s head and the detection of facial expressions to predict head orientation. Such latter techniques are more robust to varying lighting conditions and occlusions; however, there is no certainty about the adherence of predictions to the true gaze during the driving task.

Most studies have focused on a specific aspect, such as the driver or the environment. To solve this problem, in a study [54], a novel driver attention estimation model that considers the environment's saliency map and the driver's gaze was proposed; both the gaze and the scenario image were used to estimate the driver's attention area and establish a multiresolution network to model.

For children with autism spectrum disorders (ASDs), attention recognition plays a vital role in providing learning support. The unobtrusiveness of face-tracking techniques enables establishing automatic systems to detect and classify attentional behaviors. In a previous study [55], an attention detection model was established based on the kid's behavior during engagement with the robot in HRI systems; the model was used in an adaptive interaction system where the robot detects the action depending on the child's attention. However, constructing such systems is challenging because of the complexity of attentional behavior in ASDs. To overcome this problem, in a study [56], a face-based attention recognition model was presented based on geometric feature transformation and a support vector machine (SVM) classifier [57].

These semi-autonomous adaptive systems are complex and require high-performance hardware, such as GPU chips [58], to process real-time data and update interactions. Moreover, fully autonomous and complex robots and systems are not yet reliable outside controlled research setups. In a previous study [59], a simple autonomous assessment system based on attention cues was developed and deployed and then combined with an enhanced adaptive semi-autonomous interaction system. This technique can aid in ASD intervention to facilitate adaptive interactions with patients while involving minimal subjective biases.

In the present study, inspired by recent advancements, we developed a simple yet effective high-performance architecture to perform attention detection; the proposed system can be easily integrated into the aforementioned applications.

3. Data Collection Process

Thanks to the recent advances in neural networks [60], various tasks can be performed by training network models on large amounts of data. However, such data are scarcely available. Therefore, we created a new large dataset for the attention detection task.

The direction of the head is usually one of the determining aspects of whether the interlocutor is heedful, whether it's between two humans or between a human and a machine [60-64].

Thus, we determined attention based on the direction of the interlocutor's face.

In the proposed system, when an interlocutor faces their interaction partner, full attention is assumed. In contrast, if the interlocutor is looking in another direction (e.g., left, right, up, or down), the interlocutor is considered distracted. We used five classes—CENTER, LEFT, RIGHT, UP, and DOWN—rather than a binary label to better cluster different situations, letting the model distinguish between various kinds of inattention. The CENTER class is the only positive label, indicating that the interlocutor is heedful. We recorded 270 videos from 18 users, each video lasting approximately 20 s. Each user was asked to record 15 videos (each user recorded three series of five videos, where each of the five videos corresponded to one specific label), changing their location and/or their outfit—including glasses—every five videos to let the dataset be as general as possible. The videos were segmented manually and double-checked to ensure the validity of the annotations.

The created dataset was divided into the training, testing, and validation sets. The people involved in the training set were not included in the testing and validation sets, and vice versa. The average age of the people in the dataset is between 20 and 30 years. To test the proposed system's ability to generalize across different ages, we added a person who was approximately 60 years old. Details regarding the dataset are presented in Table 1.

Table 1 Dataset statistics describing the number of samples for each class in the training, testing, and validation sets. The last line shows the total number of produced samples independently from the class and the split.

Dataset Split	Class	# Samples
Train	CENTER	16.5K
	LEFT	17.8K
	RIGHT	18.0K
	UP	17.3K
	DOWN	17.0K
Validation	CENTER	2.5K
	LEFT	2.7K
	RIGHT	2.6K
	UP	2.4K
	DOWN	2.6K
Test	CENTER	3.3K
	LEFT	3.4K
	RIGHT	3.4K
	UP	3.7K
	DOWN	3.3K
∑	—	116K

4. Methodology

In this section, we first describe the proposed baseline system in Section 4.1. Next, we propose two extensions to improve such systems: face detection (Section 4.2) and GAN-based data augmentation (Section 4.3).

4.1 Baseline System

Training ML models is a challenging task. The training algorithms may not work as intended, the training times are too long, and the training data may be problematic. Transfer learning (TL) [65] is one of the most effective ML techniques to facilitate training. TL involves storing the knowledge gained in solving one problem and applying it to a different but related problem. TL can be applied to many ML models, including DL models such as artificial neural networks and reinforcement learning models. In addition, TL offers several advantages in the development of ML models, such as resource-saving and improved efficiency in training new models. Furthermore, TL can be used to train models when only unlabeled datasets are available, as most of the model is already trained.

TL is being used in several areas of ML, such as strengthening natural language processing [66], machine vision [67], and other real-world applications [68].

We used the Visual Geometry Group 16 (VGG16) [69] pretrained model for the attention task. The proposed model's architecture is illustrated in Figure 1 and consists of 13 convolutional layers (separated by five pooling layers) and three dense layers. We augmented the architecture with two additional dense layers and an output layer with the softmax activation function.

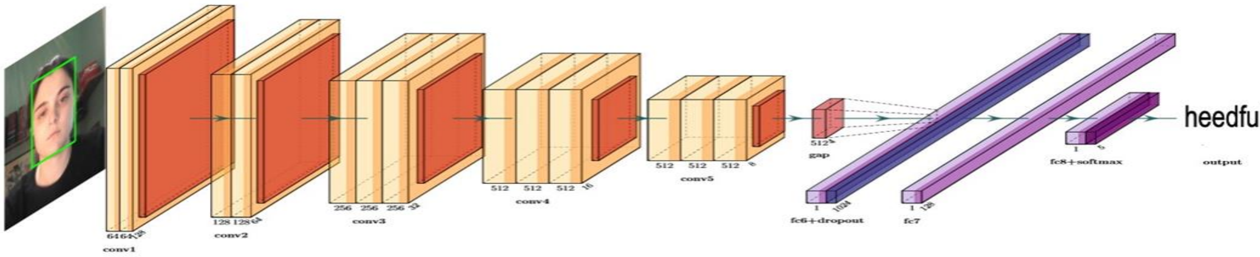

Figure 1 Architecture of the proposed model. *Conv* denotes convolutional layers, *fc* denotes fully connected layers, and *gap* is the global average pool.

The proposed system is designed to generalize as much as possible to achieve robustness to noise and unseen samples. Generally, the training data is augmented for this purpose [70]; however, not all data augmentation techniques are applicable in our case. For instance, *horizontal flipping* cannot be used because left-labeled images would become right-labeled and vice-versa, creating confusion in the training process. However, *brightness* and *shifting* can be used instead; *brightness* is beneficial if the test set contains samples whose brightness levels differ from those in the training data, whereas *shifting* is useful when the position of the interlocutor varies greatly between training and testing. In our case, *shifting* does not have any effect on the model's performance, as we already have this feature in our training set. Further details are provided in Section 5.

4.2 Face Detection

Another fundamental task of computer vision is face detection [71], wherein the human face is identified by relying on the key points that characterize it (e.g., eyes, mouth, and nose). Face detection can be used as an additional step for a variety of tasks, such as understanding facial expressions [72], lip reading [73], user identification [74], and marketing applications [75].

In the proposed system, face detection is used to focus on the features that are relevant to our task. For instance, our attention detection system finds it challenging to classify low-contrast images where the background color is similar to the face color or in cases of overexposed images. To solve this problem, we employed an auxiliary face detection system (a pretrained OpenCV Caffe model) that outputs the four vertices of the bounding box surrounding the face and used them to crop the original image.

4.3 GAN-based Data Augmentation

As already mentioned in Section 4.1, data augmentation techniques are beneficial for several computer vision tasks and enable the development of highly robust systems. However, such techniques involve elementary mathematical operations such as *shifting*, *rotating*, and *flipping* images, which may not be sufficient in some cases.

The GAN framework estimates generative models by using an adversarial process in which two models are trained simultaneously: a generative model G captures the data distribution p_g over the data x, and a discriminative model D estimates the probability that a sample came from the training data rather than G. The generator constructs a mapping from a prior noise distribution p_z to a data space as $G(z;\theta_G)$, where θ_G denotes generator parameters. The training procedure for G is to maximize the probability of D making a mistake.

The original GAN framework poses this problem as a min-max game in which two players (G and D) compete against each other, playing the following zero-sum min-max game:

$$\underset{G\quad D}{minmax} V(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z(x)}\{log[1 - D(G(z))]\} \tag{1}$$

Training GANs consists in finding a a point of equilibrium between the two competing concerns. Unfortunately, finding an equilibria is a very difficult problem. Both the generator model and the discriminator model are trained simultaneously in a game. This means that improvements to one model come at the expense of the other model. The proposed GAN architecture is illustrated in Figure 2.

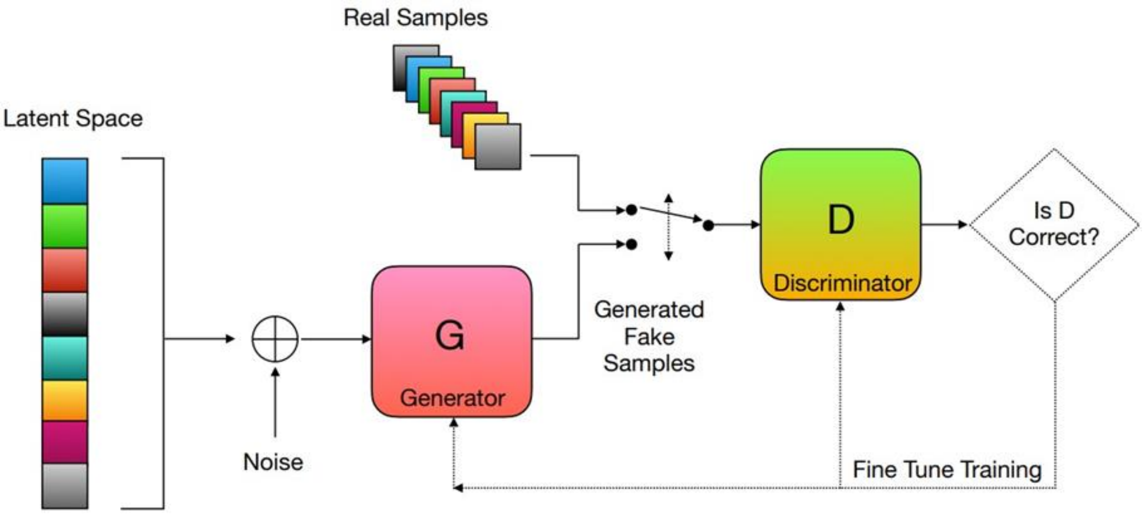

Figure 2 GAN architecture.

We developed a novel GAN-based data augmentation technique to further improve the system's performance. We used this technique to generate new samples, starting from the images of the users in the training set and making them a different age. For this purpose, we adopted SAM [76], a C-GAN for face aging, which takes as input an image of a person P and a desired age a_d and outputs the transformed image of P with age a_d.

This strategy is fundamental in any computer vision application that deals with users of all ages, but the corresponding training set does not cover all the required ages. Only users between 20 and

30 years were included in our case; however, the proposed system is intended for use by both young and old people. To overcome this problem, in addition to the data augmentation techniques already described in Section 4.1, we applied this face-aging solution to make our system insensitive to age changes, as illustrated in Figure 3.

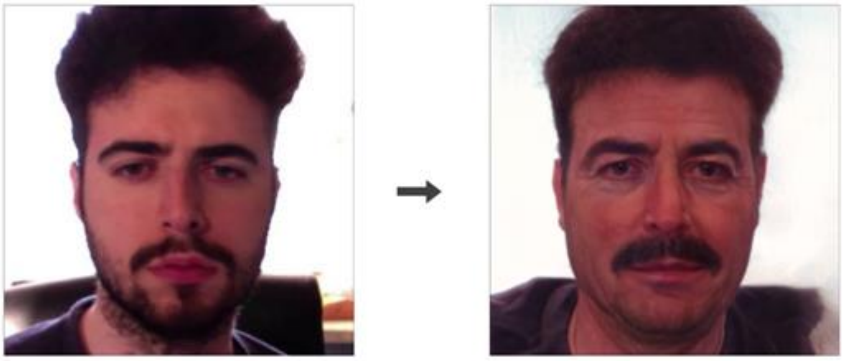

Figure 3 Example of face aging with target age $a_d = 60$.

5. Experiments

In this section, first, the experimental setup is described (Section 5.1). Next, the results of the baseline system are presented (Section 5.2.1). Finally, the results of the baseline system + face detection system (Section 5.2.2) and the proposed baseline system + face detection system + GAN system are presented (Section 5.2.3).

5.1 Experimental Setup

We implemented the proposed baseline system and its extensions by using TensorFlow and the Keras framework to load and fine-tune the weights of the pretrained models. We trained every model by using the Adam optimizer [77] for a maximum of 30 epochs with a learning rate of 10^{-5} and a cross-entropy loss criterion. The best models were determined according to their accuracy on the validation set at the end of each training epoch. The hyperparameter values are listed in Table 2. Furthermore, for obtaining the results shown in Sections 5.2.2 and 5.2.3, we repeated each training on five different seeds, fixed across experiments, and reported the mean and standard deviation of their accuracy scores. We compared the experimental results by using the Student's *t*-test. All the experiments were performed on NVIDIA Tesla K80 with 12 GB of RAM.

Table 2 Hyperparameter values of the models used for the experiments.

Hyperparameter	Value
batch size	64
learning rate	1e-5
dropout	0.5
adam β_1	0.9
adam β_2	0.999
adam ϵ	1e-8

5.2 Results

5.2.1 Baseline System

We aimed to develop a robust baseline system to perform attention detection. For this purpose, we performed model selection on the validation set by comparing different architectures. The results are presented in Table 3.

Table 3 Accuracy of the different models on the test set.

Model	Accuracy
ResNet152V2	53.37
ResNet50	54.74
Xception	49.95
VGG16	74.21
VGG19	74.12

As can be seen from Table 3, the ResNet152V2, ResNet50 [78], and Xception [79] models exhibited low accuracy due to their high architectural complexity; they have 152, 50, and 36 layers, respectively, thus indicating that extremely deep architectures are not suitable. On the contrary, simpler architectures such as VGG16 and VGG19 [69] performed better. We selected VGG16 as the temporary baseline, on which we applied data augmentation techniques because it attains comparable performances with respect to the 19-layer version while being less complex.

To further improve the proposed system, we applied the data augmentation techniques discussed in Section 4.1. The results are presented in Table 4.

Table 4 Performances of the different data augmentation techniques when applied to the VGG16 model.

Model	Accuracy
VGG16	74.21
w/*shift*	74.32
w/*brightness*	75.11
w/*brightness + shift*	74.95

As can be seen from the results presented in Table 4, applying the shifting operation did not improve the system performance as some shifted data are already partially included in the dataset, as already mentioned in Section 4.1. In contrast, the brightness operation improved the model performance because the test set contained samples with highly variable levels of brightness. In the subsequent discussion, we refer to this final system, VGG16, with the brightness operation, as the baseline system.

5.2.2 Face Detection

As mentioned in Section 4.2, we started from the intuition that the background of the image, or the user’s outfit, does not provide any helpful information but can only introduce noise in the classification process. Therefore, we performed face detection as an auxiliary task to let the system focus only on the relevant features, yielding three advantages: it greatly improved the model performance, dramatically reduced the training times, and considerably decreased the amount of required disk space.

In terms of system performance, as can be seen from the second line of Table 5, the baseline system + face detection model yielded an average improvement of 12.85 accuracy points over the previously described baseline system. Moreover, an extreme statistical significance was observed in the results (*p-value* = 3.7e-07), thus, demonstrating how this technique is fundamental for our task. As can be seen in Figure 4, the validation accuracy of the two systems varied during epochs, and a consistent gap was observed between the two curves (pink and gray curves).

Table 5 Comparison of the performance of the improved models based on face detection and GAN-based data augmentation with the baseline system. ** denotes $p - value < 0.001$, and * denotes $p - value < 0.05$. Both statistical significance and differences (Δs) are always expressed concerning the immediately above row, therefore the Δ of the third row refers to the difference of the results obtained between the model of the third and second row.

Model	Accuracy	Δ
Baseline System	73.31 ±1.40	–
w/*Face Detection*	86.17 ±1.66**	12.85
w/*Face Detection + GAN*	88.45 ±1.11*	2.28

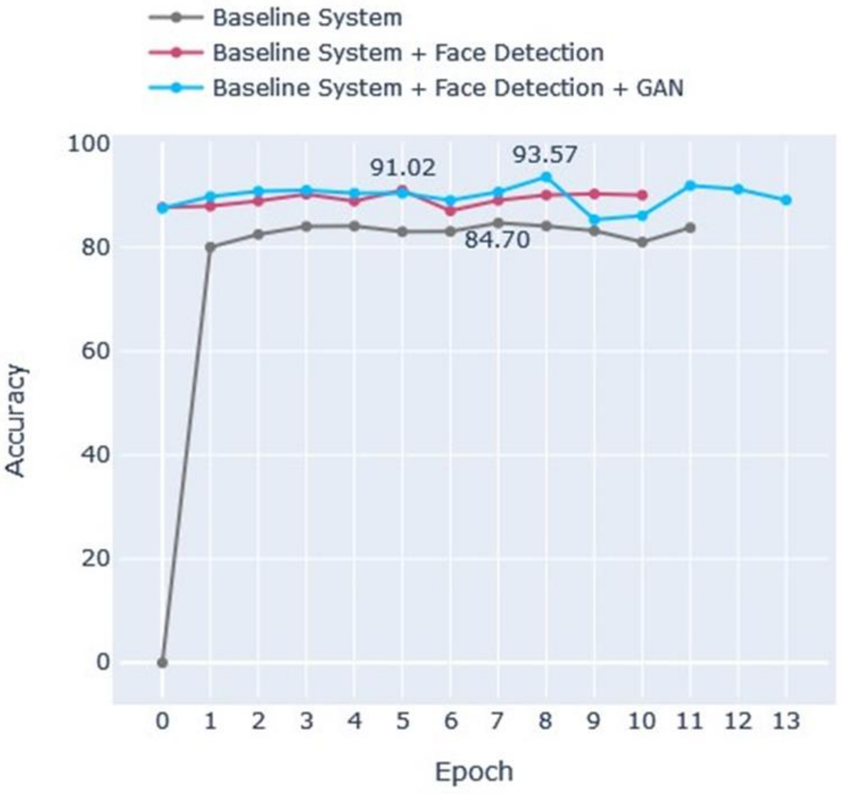

Figure 4 Validation performances of the baseline system, baseline system + face detection, and baseline system + face detection + GAN.

Considering training times instead, the baseline system required approximately 35 min for each epoch, whereas preprocessing the images by applying face detection and training the system directly on the cropped images increased the speed by up to 6×.

The final advantage is the reduced amount of required storage space. The standard dataset occupies approximately 25 GB of memory, whereas the processed dataset requires only approximately 4 GB of memory.

5.2.3 GAN-based Data Augmentation

As discussed in Section 3, our training set contains only images of users belonging to the age range of 20–30 years. Nevertheless, approximately 30% of our test set was constructed by including a 60-year-old user. While standard data augmentation techniques fail to let the system generalize well with users of different ages, the proposed GAN-based data augmentation technique (Section 4.3) helps overcome this problem. From Table 5, it can be seen that enhancing the baseline system by using GANs provides further improvements over the baseline for face detection system. The results were statistically significant.

6. Real-Time Application

After creating the dataset (Section 3), we developed the baseline system (Section 4.1) and applied the extensions—face detection (Section 4.2) and GAN-based data augmentation (Section 4.3)—to realize real-time applicability to further analyze and test the proposed system.

We developed an application consisting of a web interface that allows testing the proposed model by using the user's webcam. The web application allows to test the proposed system and observe the predictions in real time, although the prediction also depends on the frame rate (fps) of the user's computer. An example of positive and negative predictions is presented in Figure 5. We implemented the web application by using HTML, CSS, JavaScript, JQuery, and TensorFlowJS. The web application works on any browser and device, but it is not optimized for smartphones and tablets.

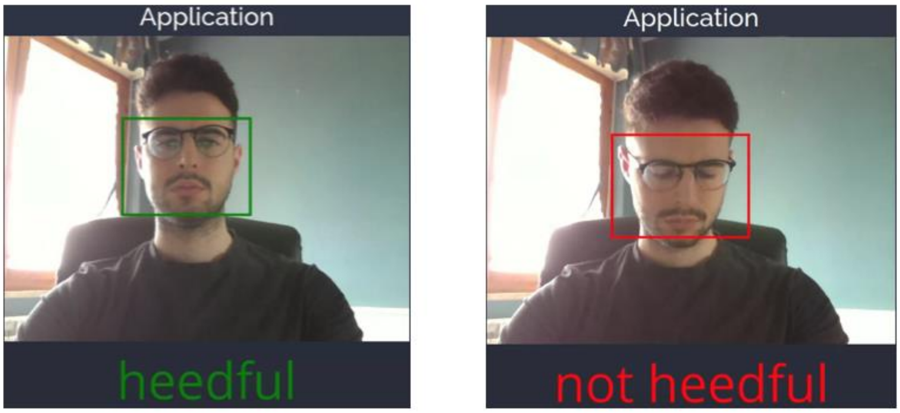

Figure 5 Example of positive and negative predictions in our web application.

7. Conclusions and Future Work

Although attention detection is a crucial step in various computer vision tasks and applications, particularly in the HRI field, few studies have been performed on this topic. In this study, we addressed the attention detection task in its entirety.

We created a manually annotated dataset consisting of approximately 120,000 images from 18 users with five labels (Section 3). Then, we used this dataset to train a baseline system and measure its performance, obtaining satisfactory results on the test set (Section 4.1).

Furthermore, as the first extension, we proposed an improvement over the baseline system by using an auxiliary face detection module to remove useless information (e.g., the background), obtaining consistent performance improvements over the baseline system (Section 4.2).

As the second extension, as our system is intended for users of all ages, we proposed a novel GAN-based data augmentation technique (Section 4.3) for face aging to make the proposed system robust to age changes. This extension yielded further performance improvements over the powerful baseline system. Moreover, we extensively evaluated the advantages of the proposed test set by performing statistical analysis and obtained statistically significant results (Section 5.2). Finally, to further analyze the behavior of the developed attention detection system and enable real-time testing, we designed a web application (Section 6).

Acknowledgments

This work has been supported by the project “HERMES WIRED” funded by Sapienza University of Rome within the Sapienza Research founding scheme 2020.

Author Contributions

All authors have equally contributed to this work.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Goodrich MA, Schultz AC. Human-robot interaction: A survey. Found Trends Hum Comput Interact. 2007; 1: 203-275.
2. Karray F, Alemzadeh M, Abou Saleh J, Arab MN. Human-computer interaction: Overview on state of the art. Int J Smart Sens Intell. 2017; 1: 137-159.
3. Balakrishnan N, Bytheway T, Carata L, Chick OR, Snee J, Akoush S, et al. Recent advances in computer architecture: The opportunities and challenges for provenance. Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance; 2015 July 8-9; Edinburgh, Scotland. Berkeley, CA: USENIX.
4. Trigueiros P, Ribeiro F, Reis LP. A comparison of machine learning algorithms applied to hand gesture recognition. Proceedings of the 7th Iberian Conference on Information Systems and Technologies (CISTI 2012); 2012 June 20-23; Madrid, Spain. New York: IEEE.

5. Tandel NH, Prajapati HB, Dabhi VK. Voice recognition and voice comparison using machine learning techniques: A survey. Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS); 2020 March 6-7; Coimbatore, India. New York: IEEE.
6. Macrae CN, Hood BM, Milne AB, Rowe AC, Mason MF. Are you looking at me? Eye gaze and person perception. Psychol Sci. 2002; 13: 460-464.
7. Yu C, Smith LB. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. PLoS One. 2013; 8: e79659.
8. Berg J, Lottermoser A, Richter C, Reinhart G. Human-Robot-Interaction for mobile industrial robot teams. Procedia CIRP. 2019; 79: 614-619.
9. Argyle EM, Gourley JJ, Kang Z, Shehab RL. Investigating the relationship between eye movements and situation awareness in weather forecasting. Appl Ergon. 2020; 85: 103071.
10. Günther U, Harrington KI, Dachselt R, Sbalzarini IF. Bionic tracking: Using eye tracking to track biological cells in virtual reality. In: Computer Vision – ECCV 2020 Workshops. Springer; 2020. pp. 280-297.
11. Wee HJ, Lye SW, Pinheiro JP. An integrated highly synchronous, high resolution, real time eye tracking system for dynamic flight movement. Adv Eng Inform. 2019; 41: 100919.
12. Dinesh D, Athi Narayanan S, Bijlani K. Student analytics for productive teaching/learning. Proceedings of the 2016 International Conference on Information Science (ICIS); 2016 August 12-13; Kochi, India; New York: IEEE.
13. Napoli C, Pappalardo G, Tramontana E. An agent-driven semantical identifier using radial basis neural networks and reinforcement learning. Proceedings of the XV Workshop "Dagli Oggetti agli Agenti"; 2014 September 25-26; Catania, Italy. Aachen: Sun SITE Central Europe.
14. Zaletelj J, Košir A. Predicting students' attention in the classroom from kinect facial and body features. EURASIP J Image Video Process. 2017; 2017: 80.
15. Wan Z, He J, Voisine A. An attention level monitoring and alarming system for the driver fatigue in the pervasive environment. Proceedings of the 2013 International Conference on Brain and Health Informatics; 2013 October 29-31; Maebashi, Japan. Cham: Springer.
16. Chen CM, Wang JY, Yu CM. Assessing the attention levels of students by using a novel attention aware system based on brainwave signals. Br J Educat Tech. 2017; 48: 348-369.
17. Di Palma S, Tonacci A, Narzisi A, Domenici C, Pioggia G, Muratori F, et al. Monitoring of autonomic response to sociocognitive tasks during treatment in children with autism spectrum disorders by wearable technologies: A feasibility study. Comput Biol Med. 2017; 85: 143-152.
18. Dehzangi O, Williams C. Towards multi-modal wearable driver monitoring: Impact of road condition on driver distraction. Proceedings of the 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN); 2015 June 9-12; Cambridge, MA, USA; New York: IEEE.
19. Palazzi A, Abati D, Calderara S, Solera F, Cucchiara R. Predicting the driver's focus of attention: The DR(eye)VE project. IEEE Trans Pattern Anal Mach Intell. 2019; 41: 1720-1733.
20. Fang J, Yan D, Qiao J, Xue J, Yu H. Dada: Driver attention prediction in driving accident scenarios. IEEE Trans Intell Transp Syst. 2022; 6: 4959-4971.
21. Billeci L, Narzisi A, Tonacci A, Sbriscia-Fioretti B, Serasini L, Fulceri F, et al. An integrated EEG and eye-tracking approach for the study of responding and initiating joint attention in autism spectrum disorders. Sci Rep. 2017; 7: 1-13.

22. Kollias KF, Syriopoulou-Delli CK, Sarigiannidis P, Fragulis GF. The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A review study. Proceedings of the 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST); 2021 July 5-7; Thessaloniki, Greece; New York: IEEE.
23. Ledezma A, Zamora V, Sipele Ó, Sesmero MP, Sanchis A. Implementing a gaze tracking algorithm for improving advanced driver assistance systems. Electronics. 2021; 10: 1480.
24. Geetha M, Latha RS, Nivetha SK, Hariprasath S, Gowtham S, Deepak CS. Design of face detection and recognition system to monitor students during online examinations using machine learning algorithms. Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI); 2021 January 27-29; Coimbatore, India. New York: IEEE.
25. Davis J, McKone E, Zirnsak M, Moore T, O'Kearney R, Apthorp D, et al. Social and attention-to-detail subclusters of autistic traits differentially predict looking at eyes and face identity recognition ability. Br J Psychol. 2017; 108: 191-219.
26. Mythili M, Mohamed Shanavas A. Early prediction of cognitive disorders among children using bee hive optimization approach.(CODEO). Biomed Pharmacol J. 2016; 9: 615-621.
27. Ponzi V, Russo S, Bianco V, Napoli C, Wajda A. Psychoeducative social robots for an healthier lifestyle using artificial intelligence: A case-study. Proceedings of the 2021 International Conference of Yearly Reports on Informatics Mathematics and Engineering. 2021 July 9. Aachen: Sun SITE Central Europe.
28. Rinehart NJ, Bradshaw JL, Moss SA, Brereton AV, Tonge BJ. Brief report: Inhibition of return in young people with autism and Asperger's disorder. Autism. 2008; 12: 249-260.
29. Roy SD, Bhowmik MK, Saha P, Ghosh AK. An approach for automatic pain detection through facial expression. Procedia Comput Sci. 2016; 84: 99-106.
30. El Morabit S, Rivenq A, Zighem MEn, Hadid A, Ouahabi A, Taleb-Ahmed A. Automatic pain estimation from facial expressions: A comparative analysis using off-the-shelf CNN architectures. Electronics. 2021; 10: 1926.
31. Vezzetti E, Speranza D, Marcolin F, Fracastoro G, Buscicchio G. Exploiting 3D ultrasound for fetal diagnostic purpose through facial landmarking. Image Anal Stereol. 2014; 33: 167-188.
32. Liu H, Mo ZH, Yang H, Zhang ZF, Hong D, Wen L, et al. Automatic facial recognition of Williams-Beuren syndrome based on deep convolutional neural networks. Front Pediatr. 2021; 9: 648255.
33. Jabbar R, Al-Khalifa K, Kharbeche M, Alhajyaseen W, Jafari M, Jiang S. Real-time driver drowsiness detection for android application using deep neural networks techniques. Procedia Comput Sci. 2018; 130: 400-407.
34. Moujahid A, Dornaika F, Arganda-Carreras I, Reta J. Efficient and compact face descriptor for driver drowsiness detection. Expert Syst Appl. 2021; 168: 114334.
35. Rafid AUI, Chowdhury AI, Niloy AR, Sharmin N. A deep learning based approach for real-time driver drowsiness detection. Proceedings of the 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT); 2021 November 18-20; Dhaka, Bangladesh. New York: IEEE.
36. Chu HC, Tsai WWJ, Liao MJ, Chen YM. Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning. Soft Comput. 2018; 22: 2973-2999.
37. Kotsia I, Pitas I. Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Trans Image Process. 2007; 16: 172-187.

38. Brandizzi N, Bianco V, Castro G, Russo S, Wajda A. Automatic RGB inference based on facial emotion recognition. Proceedings of the 2021 Scholar's Yearly Symposium of Technology, Engineering and Mathematics; 2021 July 27-29; Catania, Italy. Aachen: Sun SITE Central Europe.
39. Hua W, Dai F, Huang L, Xiong J, Gui G. Hero: Human emotions recognition for realizing intelligent internet of things. IEEE Access. 2019; 7: 24321-24332.
40. Lu CT, Su CW, Jiang HL, Lu YY. An interactive greeting system using convolutional neural networks for emotion recognition. Entertain Comput. 2022; 40: 100452.
41. Monkaresi H, Bosch N, Calvo RA, D'Mello SK. Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affect Comput. 2017; 8: 15-28.
42. Kabir AI, Akter S, Mitra S. Students engagement detection in online learning during COVID-19 pandemic using r programming language. Inf Econ. 2021; 25: 26-37.
43. Masala GL, Grosso E. Real time detection of driver attention: Emerging solutions based on robust iconic classifiers and dictionary of poses. Transp Res Part C Emerg Technol. 2014; 49: 32-42.
44. Sanger TD. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Netw. 1989; 2: 459-473.
45. Jain A, Koppula HS, Raghavan B, Soh S, Saxena A. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. Proceedings of the 2015 IEEE International Conference on Computer Vision; 2015 December 7-13; Santiago, Chile. New York: IEEE.
46. Naqvi RA, Arsalan M, Batchuluun G, Yoon HS, Park KR. Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. Sensors. 2018; 18: 456.
47. Pugeault N, Bowden R. How much of driving is preattentive? IEEE Trans Veh Technol. 2015; 64: 5424-5438.
48. Underwood G, Humphrey K, Van Loon E. Decisions about objects in real-world scenes are influenced by visual saliency before and during their inspection. Vision Res. 2011; 51: 2031-2038.
49. Tawari A, Trivedi MM. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. Proceedings of the 2014 IEEE Intelligent Vehicles Symposium Proceedings; 2014 June 8-11; Dearborn, MI, USA. New York: IEEE.
50. Vicente F, Huang Z, Xiong X, De la Torre, Zhang W, Levi D. Driver gaze tracking and eyes off the road detection system. IEEE Trans Intell Transp Syst. 2015; 16: 2014-2027.
51. Dat NN, Ponzi V, Russo S, Vincelli F. Supporting impaired people with a following robotic assistant by means of end-to-end visual target navigation and reinforcement learning approaches. Proceedings of the 2021 International Conference of Yearly Reports on Informatics Mathematics and Engineering; 2021 July 9. Aachen: Sun SITE Central Europe.
52. Fridman L, Langhans P, Lee J, Reimer B. Driver gaze region estimation without use of eye movement. IEEE Intell Syst. 2016; 31: 49-56.
53. Borghi G, Venturelli M, Vezzani R, Cucchiara R. Poseidon: Face-from-depth for driver pose estimation. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017 July 21-26; Honolulu, HI, USA. New York: IEEE.
54. Hu Z, Lv C, Hang P, Huang C, Xing Y. Data-driven estimation of driver attention using calibration-free eye gaze and scene features. IEEE Trans Industr Electron. 2022; 69: 1800-1808.

55. Wanglavan P, Jutharee W, Maneewarn T, Kaewkamnerdpong B. The development of attention detection model from child behavior for robot-assisted autism therapy. Proceedings of the 19th International Conference on Control, Automation and Systems (ICCAS); 2019 October 15-18; Jeju, Korea (South). New York: IEEE.
56. Banire B, Al Thani D, Qaraqe M, Mansoor B. Face-based attention recognition model for children with autism spectrum disorder. J Healthc Inform Res. 2021; 5: 420-445.
57. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl. 1998; 13: 18-28.
58. Ferrer EC, Rudovic O, Hardjono T, Pentland A. Robochain: A secure data-sharing framework for human-robot interaction. arXiv:1802.04480. 2018. doi:10.48550/arXiv.1802.04480.
59. Alnajjar F, Cappuccio M, Renawi A, Mubin O, Loo CK. Personalized robot interventions for autistic children: An automated methodology for attention assessment. Int J Soc Robot. 2021; 13: 67-82.
60. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit. 2018; 77: 354-377.
61. Stiefelhagen R, Fugen C, Gieselmann R, Holzapfel H, Nickel K, Waibel A. Natural human-robot interaction using speech, head pose and gestures. Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat No04CH37566); 2004 September 28-October 2; Sendai, Japan. New York: IEEE.
62. Yamazaki A, Yamazaki K, Kuno Y, Burdelski M, Kawashima M, Kuzuoka H. Precision timing in human-robot interaction: Coordination of head movement and utterance. Proceedings of the SIGCHI conference on human factors in computing systems; 2008 April 6. New York: Association for Computing Machinery.
63. Gaschler A, Huth K, Giuliani M, Kessler I, de Ruiter J, Knoll A. Modelling state of interaction from head poses for social human-robot interaction. Proceedings of the Gaze in Human-Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012); 2012 March 5-8; Boston, MA, USA.
64. Chakraborty P, Ahmed S, Yousuf MA, Azad A, Alyami SA, Moni MA. A human-robot interaction system calculating visual focus of human’s attention level. IEEE Access. 2021; 9: 93409-93421.
65. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010; 22: 1345-1359.
66. Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer learning in natural language processing. Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 June 2-7; Minneapolis, MN, USA. Stroudsburg, PA: Association for Computational Linguistics.
67. Li X, Grandvalet Y, Davoine F, Cheng J, Cui Y, Zhang H, et al. Transfer learning in computer vision tasks: Remember where you come from. Image Vision Comput. 2020; 93: 103853.
68. Ranaweera M, Mahmoud QH. Virtual to real-world transfer learning: A systematic review. Electronics. 2021; 10: 1491.
69. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6. 2015. doi:10.48550/arXiv.1409.1556.
70. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019; 6: 60.
71. 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.

72. Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw. 2003; 16: 555-559.
73. Lucey P, Potamianos G. Lipreading using profile versus frontal views. Proceedings of the 2006 IEEE Workshop on Multimedia Signal Processing; 2006 October 3-6; Victoria, BC, Canada; New York: IEEE.
74. Wozniak M, Polap D, Borowik G, Napoli C. A first attempt to cloud-based user verification in distributed system. Proceedings of the 2015 Asia-Pacific Conference on Computer Aided System Engineering; 2015 July 14-16; Quito, Ecuador. New York: IEEE.
75. Ishii Y, Hongo H, Kanagawa M, Niwa Y, Yamamoto K. Detection of attention behavior for marketing information system. Proceedings of the 7th International Conference on Control, Automation, Robotics and Vision; 2002 December 2-5; Singapore. New York: IEEE.
76. Alaluf Y, Patashnik O, Cohen-Or D. Only a matter of style: Age transformation using a style-based regression model. ACM Trans Graph. 2021; 40: 1-12.
77. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980. 2014. doi:10.48550/arXiv.1412.6980.
78. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the 2016 IEEE conference on computer vision and pattern recognition; 2016 June 27-30; Las Vegas, NV, USA. New York: IEEE.
79. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the 2017 IEEE conference on computer vision and pattern recognition; 2017 July 21-26; Honolulu, HI, USA. New York: IEEE.