Sanchez, P & Bennett CC. (2023). Facial Expression Recognition via Transfer Learning in Cooperative Game Paradigms for Enhanced Social AI. *Journal of Multimodal User Interfaces*. 17(3), pg. 1-15. https://doi.org/10.1007/s12193-023-00410-z

Facial Expression Recognition via Transfer Learning in Cooperative Game Paradigms for Enhanced Social AI

Paula Castro Sánchez[1,2], Casey C. Bennett[2,3*]

[1]Department of Computer Science, Universidad Carlos III de Madrid, Madrid, Spain
[2]Department of Intelligence Computing, Hanyang University, Seoul, Korea
[3]Department of Computing & Digital Media, DePaul University, Chicago, IL USA

*Corresponding Author: Casey C. Bennett, cabennet@hanyang.ac.kr
Corresponding Author ORC-ID: 0000-0003-2012-9250

Funding: This work was supported through funding by a grant from the National Research Foundation of Korea (NRF grant# 2021R1G1A1003801).

Abstract

Facial Expression Recognition (FER) is an effortless task for humans, and such non-verbal communication is intricately related to how we relate to others beyond the explicit content of our speech. Facial expressions can convey how we are feeling, as well as our intentions, and are thus a key point in multimodal social interactions. Recent computational advances, such as promising results from Convolutional Neural Networks (CNN), have drawn increasing attention to the potential of FER to enhance human-agent interaction (HAI) and human-robot interaction (HRI), but questions remain as to how "transferrable" the learned knowledge is from one task environment to another. In this paper, we explore how FER can be deployed in HAI cooperative game paradigms, where a human subject interacts with a virtual avatar in a goal-oriented environment where they must cooperate to survive. The primary question was whether transfer learning (TL) would offer an advantage for FER over pre-trained models based on similar (but the not exact same) task environment. The final results showed that TL was able to achieve significantly improved results (94.3% accuracy), without the need for an extensive task-specific corpus. We discuss how such approaches could be used to flexibly create more life-like robots and avatars, capable of fluid social interactions within cooperative multimodal environments.

Keywords: facial expression recognition, emotion detection, human-robot interaction, computer vision, transfer learning, social intelligence

1. Introduction

1.1 Background

Facial expressions play a vital role in human communication. It is through them that we can non-verbally communicate our emotions and intentions to others beyond the explicit content of our speech [1]. The multimodal combination of explicit and implicit communication channels is a key element in human interactions [2]. Indeed, previous studies have shown that up to 55% of human communication is conveyed through facial expressions [3]. As such, facial expression recognition (FER) is becoming an increasingly relevant topic in human-computer interaction (HCI), human-robot interaction (HRI), and human-agent interaction (HAI), where we aim to mimic this non-verbal mode of communication to create more life-like virtual avatars, capable of natural and fluid social interactions with humans [4]. The applications of this topic are vast, ranging from virtual reality to applications in health care [5,6].

In this research paper, we aim to study FER in the context of multimodal interaction during cooperative game paradigms. More specifically, we utilize a multiplayer online survival game (MOG), in which an autonomous virtual avatar and a human participant have to interact in order to survive in a hostile environment [7]. Such a goal-oriented environment requires communication and coordination between the avatar and human across a variety of tasks related to survival. FER takes on particular relevance in this scenario, as being able to infer the emotional state of the player will allow the virtual avatar to act accordingly based on non-verbal cues, potentially creating a more human-like social experience for the participant. Examples of this could be changing the avatar's expressions to match that of the participant, changing the avatar's tone of voice, or even triggering certain in-game actions in response to some FER emotional states.

This study provides a controlled setting without the necessity of defining a specific task scenario since there are multiple potential paths leading to survival, in which we can test how different interaction behaviors of the avatar affects the subject's perception of the interaction. The ultimate goal would be to determine which behavioral patterns create a more natural, human-like communication. Such findings may have a broader impact beyond just MOGs, which could lead to new insights that produce better interactive technology in general (e.g. chatbots, personal assistants, in-home devices) [8,9]. However, the first step towards this is creating a robust system for FER in such cooperative game paradigms that can address various challenges. Common FER challenges include limited task-specific corpuses, proper fusing of temporal information, and the requirement of algorithms being able to run in real-time or near real-time, among others. Two different primary approaches are tested here. First, we examine the performance of pre-trained FER models, deriving some post-processing criteria for better results. Second, we consider the application of transfer learning (TL) as a comparison to the previous pre-trained architectures.

1.2 Prior Research

1.2.1 Facial Expression Recognition

Emotion recognition in humans (including FER) is an extensively researched topic, which has been approached in many different ways. For instance, various sensing technology (e.g. cameras, eye tracking technology, electrocardiograms, electromyographs, electroencephalographs) have been used in emotion recognition systems [10,11]). However, some of those require obtrusive equipment that may bias the results and are not easily deployable in real-

world scenarios. Considering that "visual expressions are one of the main information channels in interpersonal communication" [11], cameras are a very popular sensor choice when it comes to recognizing emotion in human faces. They are widely available and typically have easy-to-use interfaces, while still yielding promising results.

Camera-based approaches for FER are typically composed of 3 steps: face detection, feature extraction, and expression classification. In the first step, the face is detected from the input images. A very common approach for such detection is the use of *Haar Cascades*, which were originally introduced by Viola and Jones in 2001 [12]. After detection, the important landmarks are extracted, so that they can be subsequently processed by classifiers into different expressions. The feature extraction process can be handled in multiple ways, but handcrafted algorithms to extract specific, targeted features are one popular approach. Those include texture-based features (e.g. Gabor filters, local binary patterns, histogram of oriented gradients) and appearance-based features (e.g. pixel intensities, landmark points, optical flow.) [13].

1.2.2 Deep Learning for FER

Due to recent advances in computational resources, and the increasing attention to deep learning, the traditional hand-crafted approach mentioned in the previous section is sometimes substituted with Convolutional Neural Networks (CNN) based-approaches [14]. This has the advantage of providing 'end to end' learning, in which feature extraction is handled by the network itself, rather than relying on face physics models. Such Deep learning FER approaches vary in terms of architecture and performance, as shown in Ko (2018) [11]. A 2020 review on the methods proposed for FER via deep learning provides a good summary of many recent and successful approaches [15]. For example, Mollahosseini et al. merged several of the existing available facial expression datasets, then used data pre-processing and augmentation techniques to develop a deep CNN (consisting of two convolution-pooling layers with two inception styles blocks), which was then trained on the aforementioned consolidated dataset [16]. They achieved increased performance, and a substantial reduction in overfitting versus the individual datasets. Elsewhere, Lopes et al. explored the impact of different types of image pre-processing on model performance [17]. Results showed that combining data augmentation, rotation correction, cropping, down sampling with 32x32 pixels and intensity normalization yielded better accuracies than applying each pre-processing type in isolation. Some other approaches combined CNNs with LSTMs, to

capture the temporal dimension of image sequences for various applications such as video processing [18].

However, all the approaches described above have similar limitations. It is still an open research question regarding how to best obtain accurate FER in-the-wild, where illumination changes, occlusion, and different backgrounds can greatly challenge our models [19]. Also, the amount of available data in this area is limited, and many times it requires human annotation before it is readily usable for modeling [20]. Transfer learning is a common approach to solving those problems, as it repurposes knowledge acquired by other existing model architectures from large datasets of similar tasks and adapts them to new tasks [21]. Strategic initialization of network parameters allows the model to learn how to solve the problem at hand more accurately and faster, while needing fewer task-specific training examples. Bin Li [22] reported improvements in accuracy of up to 13% against other popular methods in the field when using ResNet-101 for TL. Promising results for FER have also been obtained by M. A. H. Akhand et al. [23] when using TL methods based on deep CNNs.

Our aim in this research is to apply such TL approaches to FER during HAI in cooperative gameplay environments, in order to evaluate its utility in such task-oriented environments. We further compare TL to using simply using pre-trained models from other tasks on the same dataset, in order to quantify whether it provides significant advantages or not. The paper is laid out as follows. First, in Section 2, we describe the methods used and experimental design. In Section 3, we provide an overview of the modelling concepts explored and their theoretical background as it pertains to our experimental gameplay paradigm. In Section 4, we present the results of those models applied to our dataset. Section 5 provides a discussion of those results and our main conclusions.

2. Methods

2.1 Cooperative Game Environment

To carry out experiments, a cooperative game environment provided by the video game '*Don't Starve Together'* (https://www.klei.com/games/dont-starve-together) was employed, which was then modified for purposes of the experiments (as described in [7,24]). The game is publicly downloadable from online sources such as Steam. It is a social survival game, in which players must cooperate with each other and perform a variety of tasks such as resource collection or

monster fighting to ensure survival. Similarities can be drawn to other popular videogames of this type, such as Minecraft.

As explained in [7], the game allows both single-player or multi-player mode, in which up to 6 players are allowed. In this experiment, multi-player mode was used, and limited to two-player games. One of these players was the human participant, and the other was an autonomous virtual avatar designed to test various HAR and related HRI concepts [7]. The experimental setup entailed different computers ("participant" computer and the "confederate" computer respectively), in separate locations, but connected to the same online server. During the experiments, the avatar was run from the confederate computer, while a human confederate recorded the gameplay using the same computer (see Section 2.2 for data description). These experiments had a duration of thirty minutes and were conducted in a private room on the server to avoid interruptions.

As the original experiment was aimed at studying the components of socially intelligent AI, many modifications were implemented to a "Game Mod" to explore a range of HAI hypotheses by allowing researchers to create customizable interaction scenarios. However, these modifications are not relevant to the current FER analysis as they focused on in-game behavior rather than social interaction outside the game, so details will be omitted here. Further information can be found in [7].

2.2 Experiments

The data used in this research was derived from a series of focused experiments. As explained above, each experiment was comprised of a human player and an autonomous AI virtual avatar, which engaged in a 30-minute gameplay session. A total of 40 sessions were conducted. Out of these, 20 were conducted in Korean, and the remaining 20 were conducted in English. This attracted both Asian and Western participants to the study, implying that our system for FER would need to be able to recognize expressions irrespective of differences in physical facial features or cultural forms of expression [25]. During each experiment, several types of data were collected, including audio-video recordings, written game data, and several common HRI scales such as the Godspeed instrument [26]. However, for our goals we will only focus on the experiment video recordings from OBS (https://obsproject.com/) studio. These were used to create annotations for facial recognition analysis, test our models, and later as the source of our task-specific image corpus.

The general setup of the experiments was kept consistent throughout. A Zoom meeting connected the participant with the virtual avatar. The avatar was created through the Loomie application (https://www.loomielive.com/), and was capable of moving its lips while speaking autonomously as well as making basic expressions and gestures accordingly, in order to create a more human-like interaction. This Zoom meeting allowed player and avatar to freely interact through audio and visual channels. OBS was used to record the whole screen during experiments (both the game window and the Zoom window). An example of the setup can be seen below, in Figure 1.

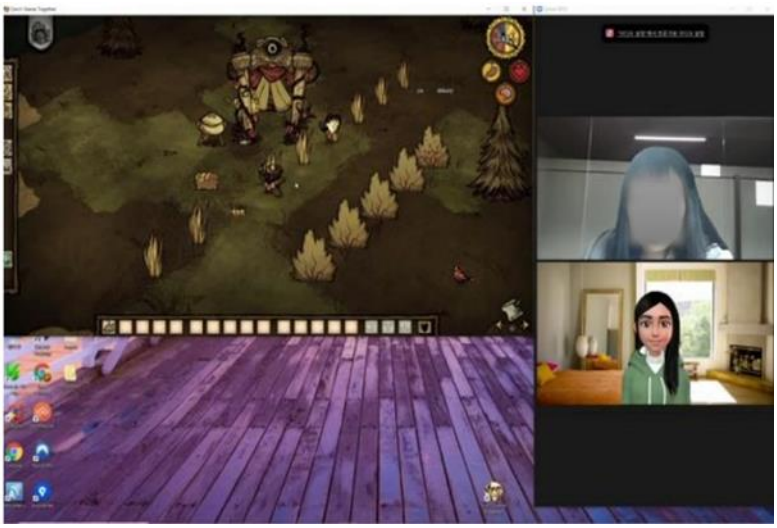

Fig. 1 Experimental Setup

2.3 Data Collection

Once all the experiments were completed, it was necessary to extract the relevant information from the recorded videos to detect emotional responses to gameplay events and/or avatar interactions. In this case, the relevant data was related to the facial expressions of participants during the gameplay. We aimed to detect the 6 basic Ekman facial expressions: Happy, Sad, Surprising, Angry, Fearful, and Disgusting (plus a 7th "Neutral" expression, i.e. no expression), as these expressions have been shown to be produced in similar ways by people across different cultures around the world. Though there is some more recent literature discussing potential cross-cultural differences in the depiction of these expressions, the majority of prevailing evidence suggests a great degree of similarity across cultures despite those nuances [27, 58]. We thus adopted the latter assumption. Such an approach was necessary here, as we were dealing with participants from a wide range of social and cultural backgrounds.

Initially, we had 40 experiment videos from which to extract the needed data. However, roughly half of these presented quality problems for our purposes. The experiment room was specifically setup ahead-of-time to address lighting issues, which including blocking windows and providing frontal lighting for the participant's face. Nevertheless, in some videos the participant's sitting posture and/or shifting in the seat during gameplay combined with variations in exterior light in the background made it difficult for current state-of-the art computer vision algorithms (see Section 3) to reliably detect the participants' faces and/or facial expressions, which is a known issue with FER in naturalistic settings [19]. For that reason, those videos we judged as not meeting minimum quality standards were excluded from further analysis. After that, we were left with 20 videos from the experiments (10 Korean and 10 English speaking).

The videos then needed to be manually annotated in order to provide ground-truth labels for later modeling. An initial analysis showed that during gameplay, 89% of the time participants displayed a Neutral expression. This is not surprising, as the majority of the time during gameplay participants were concentrating on playing the game rather than socially interacting [7]. Therefore, during manual annotation we adopted the strategy of assuming Neutral to be the base expression for any given moment in a video, and focused on annotating those times in the video in which the expression was different from Neutral. This meant a great reduction in the time needed to analyze each video. The distribution of annotated facial expressions from the sample is displayed in Table 1. The predominant expression was Neutral, due to the reasons mentioned above.

Neutral	Disgusting	Happy	Angry	Sad	Surprising	Fear
22698	1395	105	900	450	870	45

Table 1 Expression distribution in a sample of training videos

For each experiment we derived an excel file, in which each row contained the starting and ending time for each of the expressions detected. A basic Power Query M script (https://docs.microsoft.com/en-us/powerquery-m/) was then used to reformulate this into more properly structured format, with the exact video frames linked to each expression being shown.

3. Modeling

3.1 Modeling Overview

The aim of this study was to develop an algorithm that, given real-time input (as a stream of visual images) during cooperative gameplay with a virtual avatar agent, is able to determine the probabilities of any given point in the input stream displaying each of the 6 universal Ekman facial expressions (plus "Neutral"). However, a challenge exists in this scenario as the human interactor's attention is divided between the virtual avatar agent and the game itself, i.e. a *shared attention* context where social interaction is not necessarily the focal point [28,29]. The implication of that challenge is that the Ekman facial expressions may only occur sparsely, and further be inter-mixed with expressions of concentration or frustration related to the task rather than affective communication [30,31].

To address this, we developed a number of strategic approaches and weighting schemes for such a goal-oriented cooperative gameplay environment. At a high level, there is a series of basic steps we can follow. First, we need to detect whether there is an actual face in the image (Face Detection), and only after this can we apply some FER algorithm to predict the most likely expression (Facial Expression Recognition). To establish a baseline, we tested how well pre-existing pre-trained models would work in our experimental setup with our cooperative gameplay data. After that, we explored the use of novel post-processing "criteria", weighting schemes, and transfer learning in order to improve performance beyond the baseline.

3.2 Face Detection

Face Detection is a problem that has many real-world applications in recent times. We see it used in our smartphone unlocking system, the trendy 'filters' developed for social media, surveillance applications, and even in booths at the entrance of buildings where our temperature is measured for CoVID-19 purposes. However, technical methods to solve this problem have their origin further back, around 2001, with the developments made by Viola and Jones, popularly known as Haar Cascades [12]. Originally proposed as a "framework for robust and extremely rapid object detection" such as human faces, Haar Cascades are based on a sequential or cascade-like application of filters for edge/line detection where the information compounds from one filter to the next. It includes 3 main components: integral image representation, AdaBoost for feature

selection, and attentional cascades. To understand this method, it is useful to emphasize that the process is based on the detection of areas with sudden changes in pixel intensities. Viola and Jones used three types of shape features in their implementation, referred to as two-rectangle features, three-rectangle features or four-rectangle features. They are shown in Figure 2. These shape features evaluate the difference between the average of the pixels in each region, where an edge is then detected if this difference value is close to 1.

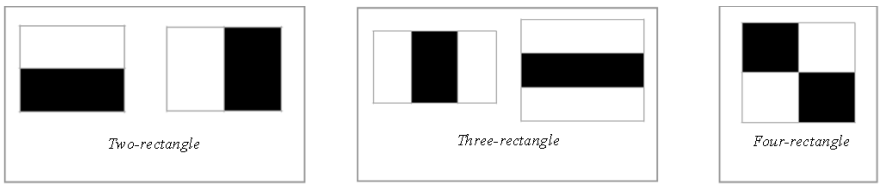

Fig. 2 Types of features evaluated by the Viola and Jones algorithm

This system has been widely adopted because of its fast computation, through the use of the *integral image.* Each pixel is set to be equal to the sum of all the pixels above and to the left of itself (inclusive). We can express the value of each pixel in the integral image as:

$$\boldsymbol{ii}(x,y) = \sum_{x' \leq x\,, y' \leq y} \boldsymbol{i}(x', y')$$

where $\boldsymbol{ii}(x,y)$ is a pixel in the integral image, and $\boldsymbol{i}(x,y)$ is a pixel in the original image. This can be computed in only one iteration through the original image. Once computed, only 4 operations are needed to determine the sum of the pixels in a rectangle, regardless of their size. This increase in speed is also supported by the use of AdaBoost to reduce the number of features considered, as well as attentional cascades that greatly limit the amount of cases (i.e. parts of the image) to be considered.

An example of what the first layer in an attention cascade during face detection looks like is shown in Figure 3. It is made up of two features; a two-rectangle and a three-rectangle feature. The first feature looks for any detectable edges in the eye and upper cheeks region. The second feature is based on the fact that eyes are normally darker than the upper-nose area, which it then tries to distinguish. Combining these two features thus produces very reliable face detection through a relatively simple approach.

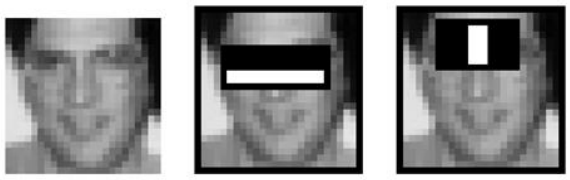

Fig. 3 First Layer of Haar Cascades obtained from [32]

Haar cascades are a very powerful algorithm capable of accurately detecting faces in input images. Thanks to its implementation based on integral images and attention cascades, as well as the feature selection performed using AdaBoost, it still remains a very popular method in the field of Computer Vision (CV). It is also very easily accessible, as the model is available to download in GitHub (https://github.com/topics/haar-cascade-classifier), and only takes a few lines of code to implement. Likewise, we adopt that approach here.

3.3 Facial Expression Recognition

Once a face has been successfully identified in an input image, the task remains to determine which of the 6 universal Ekman expressions (plus Neutral) it is most likely to be displaying (if any). To do so here, we applied a variety of different modeling classification approaches (see next section). The output was a 7-dimensional vector, in which each vector entry expressed the probability of its associated emotion being displayed in the image. This approach allows one to predict the detected expression as the one with the highest score in the output vector, as well as implement various weighting schemes or other criteria based on the vector. There are several alternative methods to perform FER. For example, some methods rely on face-physics-based models, or hand-coded facial features or landmarks into the model as a separate pre-processing step [11]. However, we opted here for a deep learning approach where the preprocessing is directly integrated in the model pipeline.

3.4 Modeling Approaches

3.4.1 Pre-trained Model

As mentioned above, the pre-trained models here are CNNs based on previous research [33]. They pose an advantage with image data over basic feed-forward Neural Networks (NN) because of the computational benefits of performing convolutions to simplify the task of matrix

multiplications. Their application entails the use of convolution filters (i.e. kernels) that efficiently down sample the input data in a structured way to reduce the number of parameters needed to be learned, and as such CNNs are a very common tool when dealing with CV tasks. The first convolutional layers of the CNN pick up on simpler patterns such as edges, lines, curves, etc. Then the following convolutional layers build on top of that information, creating successively more complex representations of the input image. For example, considering the problem of FER, one of the intermediate filters might detect shapes resembling particular facial features (eyes, nose, mouth, etc.).

Beyond the convolutional layers, CNN architectures typically include two other key types of layers, referred to as pooling layers and fully connected layers. Pooling layers are used to reduce the information passed on from the previous layer. To do this, it takes successive groups of pixels and applies a certain aggregation function to them. Fully connected layers, on the other hand, are those in which every node from the previous layer is connected to each of the nodes in the current layer. They are not technically convolutional layers in that sense because the weight of each of these connections can be different. Rather, they can be thought of as classification layers within the convolutional scheme, as they take the representation of the input data (after having gone through the previous pooling and convolutional layers) in order to calculate the probability of belonging to some category or class. The last final fully connected layer must have as many nodes as target classes in the given problem.

The model architecture used for this paper is shown in Figure 4 [33]. We note that *Dense* layers in the figure are the *fully connected (*FC*)* layers mentioned above. There are also some additional layers included for reshaping data tensors when necessary (Flattening) and to reduce overfitting (Dropout). Based on the model architecture of the pre-trained model, the input image data must be grayscale with a size of (48×48) pixels, with 4 dimensions and normalized values in the range $[0,1]$. As such, we implemented those requirements as pre-processing steps on our gameplay video recording data prior to model usage.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 44, 44, 64)	1664
max_pooling2d_1 (MaxPooling2	(None, 20, 20, 64)	0
conv2d_2 (Conv2D)	(None, 18, 18, 64)	36928
conv2d_3 (Conv2D)	(None, 16, 16, 64)	36928
average_pooling2d_1 (Average	(None, 7, 7, 64)	0
conv2d_4 (Conv2D)	(None, 5, 5, 128)	73856
conv2d_5 (Conv2D)	(None, 3, 3, 128)	147584
average_pooling2d_2 (Average	(None, 1, 1, 128)	0
flatten_1 (Flatten)	(None, 128)	0
dense_1 (Dense)	(None, 1024)	132096
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 1024)	1049600
dropout_2 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 7)	7175

Fig. 4 Pre-trained CNN architecture

3.4.2 Transfer Learning Model

Along with testing the pre-trained model pipeline shown above, we wanted to explore alternative approaches to dealing with the challenge of FER during cooperative gameplay with virtual avatar agents. Our hypothesis was that the pre-trained models, although having been trained based on a similar task, may not capture the nuances present in a task-oriented HAI cooperative gameplay paradigm problem. As such, potential performance benefits might be obtained from creating a new approach, more specific to our problem. However, doing that from scratch is a formidable effort, which would furthermore also be limited in its broader applicability to other contexts beyond our specific task environment [7]. In short, the time needed to collect and label a big enough corpus of task-specific data was not feasible.

Transfer Learning is one proposed alternative to dealing with the above problem, which as mentioned in Section 1.2.2 has shown promising results when applied to FER tasks in other domains [34,35]. Transfer learning uses pre-trained neural networks as a base from which to build a more specific network for a particular problem in a given scenario. It takes advantage of the features already learned from the previous model, but fine-tunes the upper layers to interpret them within a new context. This means the needed resources for model development are not as high, as the bulk of learning is already done.

The first step in transfer learning implementation is to create a training dataset that can fit the pre-learned model features. To do so here, we took our previously manual annotations of video gameplay recordings, determined which emotion was displayed in each frame, extracted the participants faces from the frames indicated, and then tagged the emotion labels onto each frame. There was also some re-structuring of the dataset required based on our data analysis, which is described in the Results section below. After engineering the dataset for transfer learning, we used VGG16 as our convolutional base [36]. It has proven to work well with other FER tasks in previous research [35]. VGG16 model was trained on ImageNet, a dataset consisting of thousand different objects (including human faces) with millions of examples, on which it achieved an accuracy of 92.7%. It has also been verified on CK+, JAFFE benchmark datasets, with accuracies of 94.8% and 93.7% respectively.

For transfer learning, we removed the final layers of VGG16 aimed at classification. We then replaced the final layers with a flattening layer and fully connected Dense layer specific to our cooperative gameplay paradigm. The final model´s architecture can be seen in Figure 5. We note the parameters from the VGG16 convolutional base were initially set to be non-trainable (i.e. “frozen”), so the model focused on training the final layers. After having trained our final layers for our specific problem with the base frozen, the base layers were “unfrozen” in order to optimize them, i.e. what is known as “fine-tuning” in transfer learning.

```
Layer (type)                 Output Shape              Param #
=================================================================
vgg16 (Functional)           (None, 4, 4, 512)         14714688
_________________________________________________________________
flatten (Flatten)            (None, 8192)              0
_________________________________________________________________
dense (Dense)                (None, 256)               2097408
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 771
=================================================================
Total params: 16,812,867
Trainable params: 2,098,179
Non-trainable params: 14,714,688
_________________________________________________________________
```

Fig. 5 Transfer learning model architecture

4. Results

4.1 Pre-Trained Models

For the first part of the research, we used the pre-trained models explained in Section 3.4.1. It is important to note that the model used was not further trained on our research data, but rather used as a baseline. It was used here to test how well "out of the box" models could operate on unseen data in an HAI/HRI scenario. These models operated on single images (i.e. single frame), outputting for each one a 7-dimensional vector where each of the entries corresponded to the probability of each facial expression being: Happy, Sad, Surprising, Angry, Fearful, Disgusting, and Neutral. One issue of course is that facial expressions are not instantaneous, but occur over time. There are 2 main possible approaches to deal with that: 1) analyze frames separately and then merge the result, or 2) extend the architectures used in 2D image modeling to 3D in order to capture the temporal dimension of data as part of the learning process, for example by 3D Convolutional Neural Networks [37]. The former approach is limited in scope for real-time problems such as ours where humans engage in cooperative games with interactive virtual avatar agents, so we focus on the latter approach here in this paper, based on pre-trained models from previous research [38]. Those models utilized different approaches for temporal fusion, which can be seen in Figure 6.

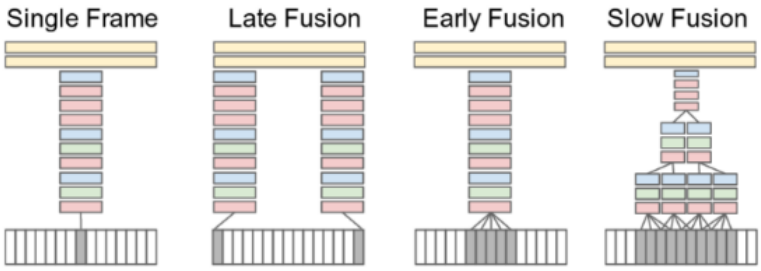

Fig. 6 Types of frame fusion proposed by [38], Copyright © 2014, IEEE.

In the previous research, four different approaches are taken to tackle the problem of fusing temporal information into the CNN architecture [38]. In Figure 6, the blue, red and green boxes symbolize different types of layers in the CNN. The *Single Frame* only considers one frame to derive the output (no temporality in data). It is used as a benchmark to test the other approaches. In the rest of the approaches, the general idea is the same. We connect the CNN output of different frame sets together, and then use a fully connected Dense layer to merge those outputs into predictions. In *Late Fusion*, we merge the results of two or more parallel CNNs with 50% overlap (i.e. if the total input size was 30, then they would be 15 frames between each input set). This allows for motion detection between frame sets that represents the delineation between different facial expressions. In *Early Fusion*, we combine information on the pixel level of T frames, by modifying the filters in the first convolution layer to a size of $11 \times 11 \times 3 \times T$. This proves to be more useful in detecting local motion and speed within the frame set that may represent the muscle-movement production of individual expressions. Lastly, *Slow Fusion* combines both of these approaches. It applies various "pipelines" of *Early Fusion* on different temporal scales, and then slowly merges those together in progressive layers such that the final output has access to global information over a wider range rather than a fixed range.

Although those fusion approaches yield promising results, they also add an extra layer of significant complexity and computational overhead to the modeling pipeline that is challenging in real-time environments [39]. Therefore, rather than fusion, here we adopted an approach that utilized post-processing "criteria" based on existing knowledge of human behavior during human-agent interaction in cooperative game paradigms [7]. These were designed based on a preliminary analysis that showed that the vast majority of time, participants had a Neutral expression during gameplay, as their focus was concentrated on the game. Looking back at the average distribution

of emotions shown in Table 1 (see Section 2), it was clear that the simplest strategy was to always assume a given expression swas Neutral, and only predict other expressions when sufficient evidence was available to do so (i.e. beyond some threshold). This entailed establishing a number of "criteria" that combined temporal information across a set of frames ("window"), based on various weighting and smoothing schemes. Those criteria were then used by the models to evaluate whether sufficient evidence existed for predictions. Results are shown below, with "Criteria" 0 being the raw predictions with no criteria applied.

4.1.1 Criterion 1 - Weighted predictions

Emotion	Neutral	Disgusting	Happy	Angry	Sad	Surprising	Fear
Average Occurrence	89.72%	3.57%	0.99%	1.97%	1.33%	2.32%	0.1%

Table 2 Average occurrence per expression

In this criterion, we weighted the predicted probabilities by the overall average frequency of each expression, so that expressions that occurred more frequently in general (e.g., happy, surprising) were given more weight when determining the highest probability for a particular window (as shown in Table 2). For each frame window, this produced a weighted vector of length 7, from which was computed its dot product. The predicted label for that window was then the facial expression with the highest resulting value.

4.1.2 Criterion 2 - Average expression over the last N frames

This criterion aims at enforcing smoothness in our predictions. The gameplay videos are analyzed on a 15 fps basis. Therefore, in theory, 15 emotion changes are possible per second. However, we expect emotions to remain constant along a certain number of frames. This idea is reinforced by [40]. In the paper, it is described how emotions go through 3 main phases, from onset, to apex, to offset. This is represented in Figure 7.

Fig. 7 Stages of facial expressions. Image taken from [40], under creative commons license ***[CC-BY-3.0]***

For this reason, this criterion looks at the last N frames analyzed as the window, and then computes the average probability of each of the expressions for that window. Then, it selects the one with highest probability as the predicted label. By doing this, the predictions are smoothed out, and outliers would be counteracted by the rest of frames being analyzed in that specific window.

4.1.3 Criterion 3 - EMA weighted predictions

This criterion builds on the intuition that expressions will gradually build up to an apex, before starting decaying again. This time, exponentially moving averages (EMA) are used to materialize this idea. This implies that the probability of each emotion in the current frame will be weighted by the results yielded in previous frames within the window. Concretely, we are interested in EMA because it ensures that if probability of an emotion in previous frames was very low, it is not likely that it will suddenly jump up aggressively. Conversely, if the previous probabilities where high, it is more likely that the current frame also displays high probability of this emotion. By using EMA, the aim is to smooth out the predictions over an individual's distinct idiosyncratic micro-expressions during the production of a facial expression, in order to more generally capture how facial expressions change over time in reality. This is illustrated in Figure 8. It clearly shows how after applying EMA the peaks that we saw in the raw predictions are smoothed out, and a clearer pattern arises.

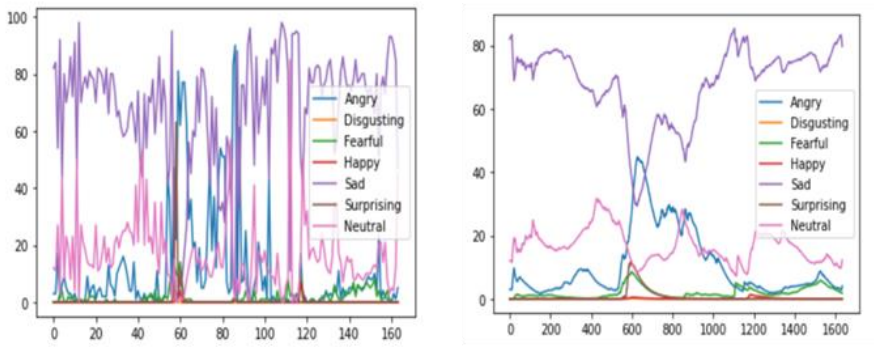

Fig. 8 Raw predictions (left panel) vs after applying EMA (right panel)

4.1.4 Criterion 4 - Minimum probability separation

In this criterion, we return to the idea of always predicting a Neutral expression unless sufficient evidence is available to predict some non-neutral expression. However, in this criteria we consider this to be the case only if the difference between the probability of highest-ranking expression and the probability of the second-ranking expression is higher than a certain threshold. In this case, that threshold was set to 15%, based on experimentation (results not shown for brevity). For instance, if our algorithm predicts Happy with a 53% confidence, and Surprise with a 46% confidence, we would conclude that there is no strong evidence to conclude that the actual expression is any of these two, and defer to a Neutral prediction. However, if the predicted emotion is Happy with a 55% confidence, and the expression with the highest largest probability is Surprise with a 35% confidence, we can say there exists enough evidence to suspect the real label is Happy.

4.1.5 Criterion 5 - Compact frame representation

Lastly, criterion 5 is similar to criterion 2. However, instead of averaging over frames by calculating the average probability of each emotion, it creates a compact frame representation of the last N frames (a new image computed as the average of N other face images) [41]. Then, the algorithm is applied to this new image, yielding a single prediction. Once again, we then predict the expression with the maximum probability in our output vector.

4.1.6 Criteria Results

The results for each of the criterions proposed is shown in Table 3. We note that for criterion 2 to criterion 5, we have also computed a weighted version for comparison, using the technique

explained in criterion 1. Note that this weighting did not affect the sample size considered. Rather, each original model output (a vector with one entry per emotion, depicting the probability of the corresponding facial expression being displayed) was scaled by the weighting vector, so that the total number of samples remained constant during analysis.

Criteria	Description	Weighted	Accuracy
0	Raw predictions	x	30.99%
1	Weighted with average expression occurrence	o	74.99%
2	Average expression occurrence over the last N	x	31.02%
	frames (N=15)	o	79.61%
3	EMA weighted predictions	x	30.83%
	(N=15)	o	81.72%
4	Minimum probability separation	x	47.23%
	(tsh = 15)	o	85.92%
5	Compact frame representation	x	20.40%
	(N=60)	o	66.33%

Table 1 Criteria Results of best-performing mode. (weighted= 'o', un-weighted = 'x'). 'N' refers to the frame window size, if applicable. 'tsh' refers to the threshold size, if applicable.

The best result is obtained when applying a weighted version of criteria 4 (in essence combining criteria 4 and criteria 1 together), with a performance of approximately 86%. This is similar to current state-of-the-art performance in the mid-80s percent range for FER during video game play reported by other researchers [54, 55]. However, a closer inspection of the various confusion matrices for all criteria revealed that there were issues. For interested readers, those full confusions matrices can be found in the appendix in the online Supplementary Material. In short, our higher-performing models here tended to predict Neutral for most expressions, and thus the performance was mainly due to high specificity but limited sensitivity. This supports that our original strategy that we should predict Neutral unless otherwise indicated, but it also highlights that the criteria we designed were not sensitive enough to clearly delineate neutral from non-neutral successfully when using the pre-trained model.

4.2 Transfer Learning

4.2.1 Data Restructuring

It was hypothesized that the poor performance of the previously explored pre-trained models (Section 4.1) may be due to the fact that the algorithms used for FER were derived from pre-built model architectures, which lacked the sensitivity needed to clearly distinguish between neutral and non-neutral expressions in these cooperative task-oriented gameplay environments where human facial expression distribution is highly imbalanced and users are often focused on the game rather than social interaction itself [7]. Indeed, those users often have a look of "concentration" that resembles a neutral expression. Although having been trained for with similar human facial expressions, the nuances present in this sort of HAI gameplay paradigm may not have been captured by that previous pre-trained architecture where the focus was more on the social interaction. Therefore, we set out to build our own model based on transfer learning, which was then fine-tuned with our cooperative gameplay experiment dataset.

Given the challenges of distinguishing between neutral and non-neutral expressions using pre-trained FER models, the aim here was to classify each frame into *Negative, Neutral,* or *Positive* valence rather than individual facial expressions. Due to the heavy imbalance in the original dataset (roughly 84.9% of samples were Neutral), we resampled the data via under-sampling to produce a more balanced dataset of 250 cases. In the literature [42], emotional valence and arousal are the two primary axes used to categorize different expressions in a 2D space. Emotional valence describes the extent to which an emotion is positive or negative, whereas arousal refers to its intensity, i.e., the strength of the associated emotional state [43-45]. The visualized space of emotions is depicted in Figure 9, as per Russell's circumplex model. Our strategy here was to reduce the space of our transfer classes to make the feature adaption step of transfer learning more tractable, and therefore we limited the problem to the valence axis. As such, all emotions were collapsed to a negative, neutral, or positive class as shown in Table 4.

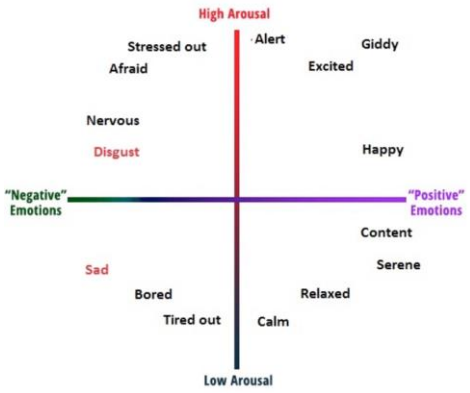

Fig. 9 Emotional dimension. Image obtained from [46], under license [CC BY-NC-SA 4.0], via ARROW@TU Dublin.

Upon further analysis, there were some challenges with some particular facial expressions, such as Surprise, which are more dependent on the arousal axis rather than valence. For instance, in our cooperative game scenario there were situations where the human player was surprised by monsters (negative valence), but also other situations where surprise occurred in response to the virtual avatar sharing resources with players (positive valence). An example can be seen in Figure 10. Thus to determine the appropriate valence category we used a *data-driven approach*, based on an analysis of game scenarios where those facial expressions occurred. Since Surprise occurred 77% of the time in what were judged to be positive valence situations, it was categorized as positive, as shown in Table 4. The final dataset consisted in a total of 2816 images. The split between training, testing and validation data per class can be seen in Table 5.

Original	Angry	Sad	Disgusting	Fear	Neutral	Surprising	Happy
New	Negative				Neutral	Positive	

Table 4 Target Class Transformation

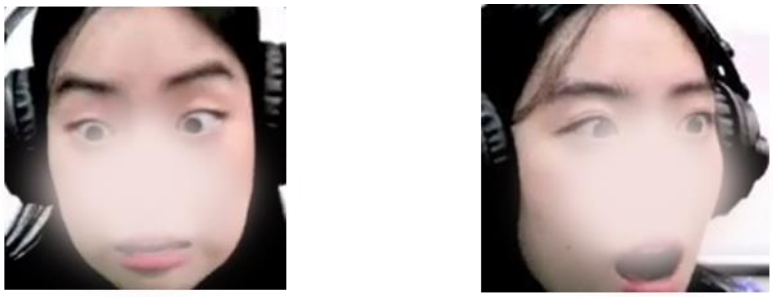

Fig. 10 Surprise expressions showing different emotional valences (images slightly blurred for privacy reasons)

	Train	Test	Validation	Total
Negative	778	105	155	1038
Neutral	786	106	157	1049
Positive	398	54	79	531
Total	1962	265	391	2618

Table 5 Final dataset composition

4.2.2 Transfer Learning vs. Pre-Trained Model Results

The results for the transfer learning model on the restructured data can be seen in Figure 11, visualized as a confusion matrix (with accuracy and F1 scores printed below). Categorical cross entropy was the loss function used. RMSprop was used as an optimizer, with a learning rate of $1e^{-4}$. Lastly, we optimized results with respect to the accuracy of the model. The model was trained using a batch size of 20, for 1x5 epochs, and that with the best results was chosen. For fair comparison, we also tested our pre-trained model from Section 4.1 on the same set of restructured data and using the same 3 valence targets. Those results can be seen in Figure 12.

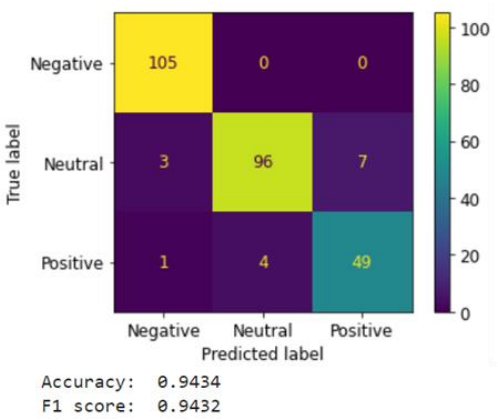

Fig. 5 Visualization of TL model performance on the test set

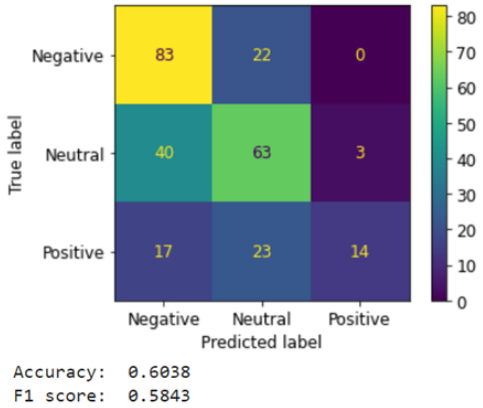

Fig. 6 Visualization of pre-trained model performance on the test set

As one can see, there is a noticeable difference between the TL model and pre-trained model. The TL model produced an accuracy and F1 score of around 94.3%, while the pre-trained model could only achieve around 60%. Additionally, given the reduction in performance of the pre-trained model from the mid-80% range in Table 4 to the results in Figure 12, the results further underscore the weaknesses of applying pre-trained models of human facial expressions to specific task-oriented environments. Indeed, those environments often necessitate certain forms of communication, which is where transfer learning may hold significant advantages for FER. Critically, we also note that for both pre-trained and TL models the inference speed (i.e. when making predictions on unseen data) was less than 1 second per image in all cases, indicating that the TL models were still capable of being used in real-time like the pre-trained models.

5. Discussion

5.1 Results Summary

Our aim in this research was to apply FER during multimodal HAI in cooperative gameplay environments, and assess the utility of pre-trained models versus transfer learning models in those kinds of task-oriented environments that require cooperation and communication between a human and AI agent in order to accomplish some goal. The assessment comprised iterations of analyses of different computational approaches, which are summarized as follows. The first approach (Section 3.4.1) attempted to use pre-trained models but proved to have unstable performance, which thus necessitated some post-processing of the results to ensure smoothness of the predictions. We further discovered it was necessary to engineer the post-processing criteria in a way so that Neutral facial expressions would be always assumed (i.e. “default expression"), unless sufficient evidence was given for any other non-neutral expression. This improved the raw predictions, though there was still a room for improvement, as the predictive power proved to be limited (maximal accuracy in the mid-80s) relative to human-to-human recognition accuracy of facial expressions, which has also been reported in the mid-80s by Ekman and other research groups [47,48]. Moreover, ideally as a tool for cognitive support, computer vision-based FER systems would enhance human performance, not simply replicate it.

To address the above limitations, transfer learning based on VGG16 was evaluated (Section 3.4.2). In order to accomplish this, the final layers of the VGG16 model were removed and replaced with custom layers specific to our HAI cooperative gameplay paradigm. After fine-tuning training of the TL model based on experiment data, this approach proved to have a great utility for FER recognition in that sort of task-oriented environment without the need for an extensive task-specific corpus, producing an accuracy and F1 score of 94.3%. That was a significant improvement over the pre-trained model performance, and shows the potential of TL to repurpose and adapt knowledge acquired from similar tasks to new tasks [21].

In short, HAI and HRI in cooperative environments tends to necessitate certain forms of multimodal communication to facilitate successful cooperation that goes beyond direct verbal communication, and transfer learning seems to be advantageous for detecting those kinds of communicative strategies (in this case non-verbal facial expressions) over generic pre-trained deep learning models for computer vision. This has significant implications for the design and development of AI agents interacting multimodally with humans on cooperative tasks in the future

[49]. Given that up to 55% of human communication is conveyed through facial expressions [3], robust FER systems become a pivotal point in creating more life-like virtual avatars, capable of natural and fluid social interactions with humans. The development of these systems will not only allow virtual agents to better understand the context of communication, but also to create deeper interactions by mimicking these non-verbal cues. More broadly, these findings may also contribute to our understanding of the computational nature of coordination between human and robot interactors during HAI/HRI in real-world environments [50].

5.2 Limitations and Future Work

Although the results were satisfactory, there still remains many issues to deal with. In the field of FER, there are few datasets available, and many of these concentrate in lab-scenarios, where the conditions are very controlled. This greatly varies from what we find in real-life. Changes in illumination, occlusion and non-plain backgrounds are very common, but no dataset manages to capture these. This makes FER in-the-wild a very challenging task. The time complexity of algorithms used is also an issue. Although the systems are designed with the time-constraints in mind, the processing and analysis of each frame is not as fast as it should, creating a certain 'lag' in the system. This may be solved by using a device with more resources, or down sampling the frames analyzed per second. Lastly, during multiplayer online games players are normally concentrated on the game, and very rarely show expressions different that neutral. However, expressions of boredom, frustration, and concentration are common during gameplay, which would be interesting to include in future analysis (they currently fall in the *Neutral* class).

For future studies, vocal analysis and gesture recognition could be incorporated to create a multimodal system that more accurately interprets the emotional state of players. Indeed, much social interaction is non-verbal and encompasses cues beyond facial expressions themselves [51]. Extending the multimodal input further would allow us to create a higher-level view of the interaction, and thus make more informed design choices for how the avatar behaves. Likewise, this would also allow the virtual avatar to autonomously engage in more fluid social interactions by picking up on subtle social cues during the interaction that belie any single mode of communication [52].

STATEMENTS & DECLARATIONS

Funding: This work was supported through funding by a grant from the National Research Foundation of Korea (NRF grant# 2021R1G1A1003801).

Compliance with Ethical Standards: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Hanyang University (protocol #HYU2021-138) for studies involving humans. Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are not publicly available due to the fact the data comprises video and audio recordings of identifiable human subjects during gameplay. However, extracted de-identified data may be made available from the corresponding author upon reasonable request.

Competing Interests: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Camerlink I, Coulange E, Farish M, Baxter EM, Turner SP (2018) Facial expression as a potential measure of both intent and emotion. Sci Rep 8: 17602. https://doiorg/101038/s41598-018-35905-3

2. Key, MR (2011) The Relationship of Verbal and Nonverbal Communication Berlin : New York: De Gruyter Mouton. https://doiorg/101515/9783110813098

3. Mehrabian A (2008) Communication Without Words.In Communication Theory (pp 193–200), Routledge. https://doiorg/104324/9781315080918-15

4. Jyoti J, Jesse H (2017) Continuous Facial Expression Recognition for Affective Interaction with Virtual Avatar, IEEE Signal Processing Society, SigPort.

5. Houshmand B, Khan N (2020) Facial Expression Recognition Under Partial Occlusion from Virtual Reality Headsets based on Transfer Learning. https://doiorg/1048550/arXiv200805563

6. Onyema EM, Shukla PK, Dalal S, Mathur MN, Zakariah M, Tiwari B (2021) Enhancement of Patient Facial Recognition through Deep Learning Algorithm: ConvNet. Journal of Healthcare Engineering. https://doiorg/101155/2021/5196000

7. Bennett CC, Weiss B, Suh J, Yoon E, Jeong J, Chae Y (2022) Exploring Data-Driven Components of Socially Intelligent AI through Cooperative Game Paradigms. Multimodal Technol Interact 6(2): 16. https://doiorg/103390/mti6020016

8. Carranza KR, Manalili J, Bugtai NT, Baldovino RG (2019) Expression tracking with OpenCV deep learning for a development of emotionally aware chatbots. 7th IEEE International Conference on Robot Intelligence Technology and Applications (RiTA), pp. 160-163. https://doiorg/101109/RITAPP20198932852

9. Castillo JC, González ÁC, Alonso-Martín F, Fernández-Caballero A, Salichs MA (2018) Emotion Detection and Regulation from Personal Assistant Robot in Smart Environment Personal Assistants. In: Personal Assistants: Emerging computational technologies (pp. 179-195), Springer Cham. https://doi.org/10.1007/978-3-319-62530-0_10

10. Samadiani N, Huang G, Cai B, Luo W, Chi CH, Xiang Y, He J (2019) A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. Sensors 9(8): 1863. https://doiorg/103390/s19081863

11. Ko BC (2018) A Brief Review of Facial Emotion Recognition Based on Visual Information Sensors (Basel, Switzerland). https://doiorg/103390/s18020401

12. Viola P, Jones M (2001) Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition (CVPR)1: I-I. https://doiorg/10.1109/CVPR.2001.990517.

13. Karnati M, Ayan S, Ondrej K, Anis,Y (2021) FER-net: facial expression recognition using deep neural net .Neural Comput Appl 33: 9125-9136.. https://doiorg/101007/s00521-020-05676-y

14. Sarker IH (2021) Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT SCI 2: 420. https://doiorg/101007/s42979-021-00815-1

15. Wafa M, Wahida H (2020) Facial emotion recognition using deep learning: review and insights. Procedia Computer Science 175: 689-694. https://doiorg/101016/jprocs202007101

16. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. IEEE Winter Conference on Applications of Computer Vision (WACV), 1-10 (2016)

17. Lopes AT, Aguiar ED, Souza AF, Oliveira-Santos T (2017) Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognit, 61: 610-628. https://doi.org/10.1016/j.patcog.2016.07.026

18. Kim DH, Baddar WJ, Jang J, Ro Y (2019) Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition. IEEE Transactions on Affective Computing, 10: 223-236. https://doi.org/10.1109/TAFFC.2017.2695999

19. Singh S, Prasad SVAV (2018) Techniques and Challenges of Face Recognition: A Critical Review. Procedia Computer Science 143: 536-543. https://doiorg/101016/jprocs201810427

20. Mohamad NO, Dras M, Hamey L, Richards D, Wan S, Paris C (2020) Automatic recognition of student engagement using deep learning and facial expression. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 273-289. https://doi.org/10.1007/978-3-030-46133-1_17

21. Rawat W, Wang Z (2017) Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput 29(9): 2352-2449. https://doiorg/101162/NECO_a_00990

22. Li B (2021) Facial expression recognition via transfer learning. EAI Endorsed Transactions on e-Learning 7(21): e4-e4. https://doiorg/104108/eai8-4-2021169180

23. Akhand, MAH, Shuvendu R, Nazmul S, Kamal MAS, Shimamura T (2021) Facial Emotion Recognition Using Transfer Learning in the Deep CNN. Electronics 10(9): 1036. https://doiorg/103390/electronics10091036

24. Bennett CC, Weiss B (2022) Purposeful Failures as a Form of Culturally-Appropriate Intelligent Disobedience During Human-Robot Social Interaction. In: Autonomous Agents and Multiagent Systems Best and Visionary Papers (AAMAS 2022), Revised Selected Papers Springer-Verlag, Berlin, Heidelberg, 84–90. https://doiorg/101007/978-3-031-20179-0_5

25. Marsh AA, Elfenbein HA, Ambady N (2003) Nonverbal "Accents": Cultural Differences in Facial Expressions of Emotion. Psychological science 14(4): 373-376. https://doiorg/101111/1467-928024461

26. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robot. International journal of social robotics, 1: 71-81. https://doi.org/10.1007/s12369-008-0001-3

27. Ekman P, Friesen WV (2003) Unmasking the Face A Guide to Recognizing Emotions from Facial Clues Los Altos, CA: Malor Books.

28. Soussignan R, Schaal B, Boulanger V, Garcia S, Jiang T (2015) Emotional communication in the context of joint attention for food stimuli: effects on attentional and affective processing. Biological Psychology, 104: 173-183. https://doi.org/10.1016/j.biopsycho.2014.12.006

29. Mojzisch A, Schilbach L, Helmert JR, Pannasch S, Velichkovsky BM, Vogeley K (2006) The effects of self-involvement on attention, arousal, and facial expression during social interaction with virtual others: A psychophysiological study. Social neuroscience, 1(3-4): 184-195. https://doi.org/10.1080/17470910600985621

30. Blom PM, Methors, S, Bakkes S, Spronck P (2019) Modeling and adjusting in-game difficulty based on facial expression analysis. Entertainment Computing 31: 100307. https://doiorg/101016/jentcom2019100307

31. Mistry K, Jasekar J, Issac B, Zhang L (2018) Extended LBP based Facial Expression Recognition System for Adaptive AI Agent Behaviour. International Joint Conference on Neural Networks (IJCNN), pp. 1-7.

32. Viola P, Jones MJ (2004) Robust Real-Time Face Detection. International Journal of Computer Vision 57: 137-154. https://doiorg/101023/B:VISI000001308749260fb

33. Serengil SI (2022) TensorFlow 101: Introduction to Deep Learning for Python Within TensorFlow. https://githubcom/serengil/tensorflow-101 Accessed 12 December 2022

34. Yan H (2016) Transfer subspace learning for cross-dataset facial expression recognition. Neurocomputing 208: 165-173. https://doiorg/101016/jneucom201511113

35. Dubey AK, Jain V (2020) Automatic facial recognition using VGG16 based transfer learning model. Journal of Information and Optimization Sciences, 41: 1589-1596. https://doi.org/10.1080/02522667.2020.1809126

36. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doiorg/1048550/arXiv14091556

37. Ji S, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1): 221-231. https://doiorg/101109/TPAMI201259

38. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-Scale Video Classification with Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725-1732. https://doi.org/10.1109/CVPR.2014.223

39. Jeong M, Ko BC (2018) Driver's Facial Expression Recognition in Real-Time for Safe Driving. Sensors 18(12): 4270. https://doiorg/103390/s18124270

40. Wu C-H, Lin J-C, Wei W-L (2014) Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing, 3: e12. http://doiorg/101017/ATSIP201411

41. Yang J, Ren P, Zhang D, Chen D, Wen F, Li H, Hua G (2017) Neural Aggregation Network for Video Face Recognition. EEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5216-5225. https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.554

42. Citron FM, Gray MA, Critchley HD, Weekes BS, Ferstl EC (2014) Emotional valence and arousal affect reading in an interactive way: neuroimaging evidence for an approach-withdrawal framework. Neuropsychologia 56:79-89. https://doiorg/101016/jneuropsychologia201401002

43. Barrett LF, Russell JA (1999) The Structure of Current Affect: Controversies and Emerging Consensus. Current Directions in Psychological Science 8(1): 10-14. https://doiorg/101111/1467-872100003

44. Lang PJ, Bradley MM, Cuthbert BN (1997) Motivated Attention: Affect, Activation, and Action. Attention and Orienting: Sensory and Motivational Processes 97: 135.

45. Russell JA (2003) Core affect and the psychological construction of emotion. Psychological Review 110(1): 145. https://doiorg/101037/0033-295x1101145

46. Munoz-De-Escalona E, Cañas J (2017) Online Measuring of Available Resources. First International Symposium on Human Mental Workload: Models and Applications. https://doiorg/1021427/D7DK96

47. Tottenham N, Tanaka JW, Leon AC, et al. (2009) The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Research 68(3): 242-249. https://doiorg/101016/jpsychres200805006

48. Biehl MC, Matsumoto D, Ekman P, Hearn V, Heider KG, Kudoh T, Ton V (1997) Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability Data and Cross-National Differences. Journal of Nonverbal Behavior 21: 3-21. https://doi.org/10.1023/A:1024902500935

49. Holzinger AT, Müller H (2021) Toward human–AI interfaces to support explainability and causability in medical AI. Computer 54(10) 78–86. https://doi.org/10.1109/MC.2021.3092610

50. Thomaz A, Hoffman G, Cakmak M (2016) Computational Human-Robot Interaction. Foundations and Trends in Robotics 4: 104-223. http://dx.doi.org/10.1561/2300000049

51. Celiktutan O, Skordos S, Gunes H (2019) Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. IEEE Transactions on Affective Computing 10(4):484-497. https://doiorg/101109/TAFFC20172737019

52. Oh CS, Bailenson JN, Welch GF (2018) A Systematic Review of Social Presence: Definition, Antecedents, and Implications. Front Robot AI 5:114. https://doi.org/10.3389/frobt.2018.00114

53. Pantic M, Rothkrantz LJM (2000) Automatic Analysis of Facial Expressions: The State of the Art. IEEE Trans Pattern Anal Mach Intell 22(12): 1424-1445. https://doiorg/101109/34895976

54. Xu Q, Yang Y, Tan Q, Zhang L (2017) Facial Expressions in Context: Electrophysiological Correlates of the Emotional Congruency of Facial Expressions and Background Scenes. Frontiers in Psychology 8:2175. https://doiorg/103389/fpsyg201702175

55. Cha HS, Im CH (2022) Performance enhancement of facial electromyogram-based facial-expression recognition for social virtual reality applications using linear discriminant analysis adaptation. Virtual Reality 26(1):385-398. https://doiorg/101007/s10055-021-00575-6

56. Hasnul MA, Aziz NAA, Alelyani S, Mohana M, Aziz AA (2021) Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. Sensors 21(15): 5015. https://doiorg/103390/s21155015

57. Black MH, Almabruk T, Albrecht MA, Chen NT, Lipp O V, Tan T, Bolte S, Girdler S (2018) Altered Connectivity in Autistic Adults during Complex Facial Emotion Recognition: A Study of EEG Imaginary Coherence. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2752-2755. https://doiorg/101109/EMBC20188512802

58. Fang X, Rychlowska M, Lange J (2022) Cross-cultural and inter-group research on emotion perception. J Cult Cogn Sci 6, 1–7. https://doi.org/10.1007/s41809-022-00102-2