Multimedia Tools and Applications (2023) 82:13617–13636
https://doi.org/10.1007/s11042-022-13762-7

Multimodal recognition of frustration during game-play with deep neural networks

Carlos de la Fuente[1] · Francisco J. Castellanos[1] · Jose J. Valero-Mas[1] · Jorge Calvo-Zaragoza[1]

Received: 1 February 2021 / Revised: 19 April 2022 / Accepted: 1 September 2022 /
Published online: 27 September 2022

Abstract
Frustration, which is one aspect of the field of emotional recognition, is of particular interest to the video game industry as it provides information concerning each individual player's level of engagement. The use of non-invasive strategies to estimate this emotion is, therefore, a relevant line of research with a direct application to real-world scenarios. While several proposals regarding the performance of non-invasive frustration recognition can be found in literature, they usually rely on hand-crafted features and rarely exploit the potential inherent to the combination of different sources of information. This work, therefore, presents a new approach that automatically extracts meaningful descriptors from individual audio and video sources of information using Deep Neural Networks (DNN) in order to then combine them, with the objective of detecting frustration in Game-Play scenarios. More precisely, two fusion modalities, namely *decision-level* and *feature-level*, are presented and compared with state-of-the-art methods, along with different DNN architectures optimized for each type of data. Experiments performed with a real-world audiovisual benchmarking corpus revealed that the multimodal proposals introduced herein are more suitable than those of a unimodal nature, and that their performance also surpasses that of other state-of-the–art approaches, with error rate improvements of between 40% and 90%.

Keywords Multimodal · Audiovisual · Neural network · Emotion · Frustration

1 Introduction

The field of affective computing, which is understood as the process of estimating human emotions by means of computational tools [35], is becoming progressively more relevant

The first author acknowledges the support from the Spanish "Ministerio de Educación y Formación Profesional" through grant 20CO1/000966. The second and third authors acknowledge support from the "Programa I+D+i de la Generalitat Valenciana" through grants ACIF/2019/042 and APOSTD/2020/256, respectively.

✉ Carlos de la Fuente
cdlf4@alu.ua.es

[1] Department of Software and Computing Systems, University of Alicante, Alicante, Spain

in the video game industry owing to the inherent relationship between the emotions evoked in the player and the overall user experience [30], with the objective being to create personalized game scenarios for each player [25]. Video games are currently considered to be possibly the main Human-Computer Interaction environment in which users are more open to an alteration in emotional state in order to enhance their own experience [49]. The actual estimation of the user's emotional state is, therefore, a key element in the process of providing proper feedback to the system [7].

In this regard, scientific literature reports several proposals with which to estimate such emotional states and engagement levels. It is generally possible to broadly divide these proposals into two main families: (i) the so-called *invasive* techniques, in which this process is carried out using sensing devices, most typically data from electroencephalography (EEG) headsets [6], although electrocardiogram, facial electromyography, electrodermal activity and respiration arterial data have also being considered [13], and (ii) *non-invasive* approaches, in which the premise is to gather the information required without introducing external elements suchas tackling the problem as a facial recognition task [3], as the face is considered the most expressive part of the body [32, 37], or considering the use of eye-tracking technologies [23].

Of all the different emotions a video game user may experience, frustration is one of the key elements to estimate because it is significantly correlated with engagement success [19]. Frustration appears when the user is not able to achieve a goal and, if not properly monitored, may lead the user to disregard not only the goal but also the actual game [12]. The study of this particular problem is, therefore, clearly beneficial for the video game industry.

Despite the aforementioned relevance of quantifying a user's frustration level, this particular task has been poorly addressed and the few works can be found in literature basically differ as regards the principle used to estimate this emotion. For instance, Miller and Mandryk [28] studied this problem by assuming that the players' affective state is related to the touch pressure they apply to game controllers. Other authors, however, address this task by relying on the use of audio recordings, video captures, and combinations of both sources of information obtained from the actual game users [12, 41, 42].

Regardless of the actual proposal used to estimate user emotions, most approaches analyze the data collected by considering Machine Learning (ML) within a general framework. Furthermore, the current Deep Learning paradigm, which is represented by Deep Neural Networks (DNNs), is the current trend in most recent emotion-estimating proposals owing to its demonstrated effectiveness and great capacity for generalization in highly disparate tasks [11, 21, 36, 45].

In this regard, it is necessary to point out that the use of DNNs to solve the frustration recognition problem is not new, since state-of-the-art approaches already consider them [41, 42]. Nevertheless, these proposals do not take advantage of the complete potential of these learning techniques, signifying that there is significant room of improvement that should be further explored and studied.

This work, therefore, proposes a *non-invasive* multimodal approach based on DNNs that detects frustration by making use of audiovisual data. More precisely, we consider the separate exploitation of the audio and video data in order to take advantage of the nuances of each modality so as to then explore different policies with which to synergistically combine both sources of information. We specifically propose two fusion modes with the aim of improving the results of the single-source methods, which deal with audio and video separately, along with the results provided by the existing approaches described in Section 2.

Our results for real-world benchmarking data for frustration detection show that this multimodal approach is beneficial as regards addressing the proposed task and that it outperforms existing state-of-the-art proposals.

While it could be argued that this proposal is possibly limited in terms of the number of data modalities considered and the sophistication of the neural architectures, it should be noted that, as mentioned previously, the results are already better than those of state-of-the-art strategies. In this regard, and as will be discussed below, additional sources of information, along with other learning schemes, may further improve these results, thus making them of great relevance for future work.

The present work is organized as follows: Section 2 provides a literature review on emotional recognition, focusing on the particular case of frustration, and also on multimodal learning systems. Section 3 then goes on to describe the multimodal approach proposed in this work, while Section 4 provides details on the corpus and metrics considered for the evaluation of the proposal. Section 5 presents the results of the experimentation carried out, and finally, Section 6 concludes the work and proposes future lines with which to further study the topic.

2 Background

This section provides the background required for the remainder of this work. It first explores the topic of emotion analysis so as to then concentrate on that of frustration recognition, after which the topic of multimodal exploitation of information is presented.

2.1 Emotion recognition

The emotion recognition task is highly complex, and scientific literature, therefore, provides a wide range of methods with which to tackle it [5]. Of these, speech analysis constitutes quite a common framework in which to perform this task. Kwon et al. [20] proposed a combination of specific features such as pitch, energy, and Mel Frequency Cepstral Coeffecients (MFCCs), among others, to be later processed by a Support Vector Machines (SVM) classifier in order to recognize the emotion presented in the audio recording. Another example is the work by Yang et al. [48], which detects emotions in songs by using regression algorithms, obtaining the best results with Support Vector Regression (SVR).

As mentioned previously, more recent works rely on DNN architectures. For example, the work by Wootaek et al. [24] presents an approach based on a Convolutional Recurrent Neural Network (CRNN) that automatically extracts potential features to be used in the classification of emotions obtained from recordings of speech. Another similar example is the work by Mirsamadi et al. [29], which biases the feature extraction process by using a weighted-pooling strategy to promote those features that best represent the emotions in question.

Although speech is commonly employed to study emotions, other sources of information have also been explored. For instance, the work by Ebrahimi et al. [8] presents an approach based on Recurrent Neural Networks (RNNs) in order to analyze facial expressions and classify them according to a set of predefined categories. In their work, Bahreini et al. [2] similarly recognize facial emotions but employ a fuzzy logic approach. Finally, as stated above, other *invasive* approaches base their performance on the acquisition of additional data such as brain activity by using EEG devices and processing them [16, 33, 39], but they have the constraint of specific hardware requirements.

2.1.1 Analysis of frustration

Since frustration is a particular type of emotion, the general frameworks presented above can be adapted to its sole detection and recognition. However, given the relevance of this topic, several strategies have been especially devised in order to tackle this problem. For example, Fernandez and Picard [10] proposed a method based on Hidden Markov Models in order to recognize frustration in speech signals. More recently, Malta et al. [26] presented a work in which the frustration of drivers was detected by using a Bayesian network, considering the correlation between frustration and several inputs, such as speech, video recordings and the driver's use of pedals.

In the particular context of this work, despite being scarce, there are, nevertheless, some works that analyze frustration in the context of video games. Song et al. [42] proposed a multimodal approach with which to estimate the frustration level by combining audio and video inputs through the use of neural networks. However, the authors relied on hand-crafted facial features extracted from the video and MFCCs, which may arguably not be the best descriptors for the task in hand. The approach uses a standard Long Short-Term Memory (LSTM) model to process both audio and video features, and does not, therefore, completely exploit the capabilities of both DNNs and the different sources of information available. This work was further improved by Meishu et al. [41], who employed more complex neural networks with residual connections but relied exclusively on speech data, thus ignoring the information provided by the video images.

2.2 Multimodal audiovisual analysis

Multimodality [44] is the trend in ML of exploiting different sources of data in order to then carry out a certain combination, which results in a more robust and proficient model. Of all the different combination possibilities, namely fusion policies, the following are highlighted:

- *Early fusion*: this combines the data sources before they are processed by the learning-based model. Its main advantage is that only one model has to be trained, but it requires a proper preprocessing stage for the data sources to enable them to be combined. The high degree of source variability, therefore, hinders the creation of a proper model that is able to correctly classify data.
- *Late or decision fusion*: this is based on the processing of each data source by an independent model and then combining their individual classification decisions. In contrast to early fusion, each model learns a specialized set of features, which is a much easier to achieve. This strategy is typically used when the sources are significantly different from each other.
- *Intermediate or feature fusion*: this is a feature-level fusion of the learning models, and is typically carried out by concatenating the features obtained before the final decision is made. This allegedly makes it possible to obtain more robust classifiers. However, this scheme increases the complexity of the model, since it consists of a single model with several inputs and one output.

Early and late fusion are fairly common in the literature. For example, Snoek et al. [38] compare both strategies in semantic video analysis, concluding that late fusion tends to provide a slightly better performance. Another comparison is presented in the work by Gunes and Piccardi [15], which combines facial and body-gesture features in order to recognize emotions by employing two traditional ML algorithms, namely Decision Trees and

Bayesian Networks. These authors conclude that fusion modalities are better than unimodal models, and that feature fusion performs particularly well. Another example is the work by Wimmer et al. [46], which proposes the use of the SVM classifier with a low-level combination of features extracted from audiovisual data for emotion recognition, thus improving performance with regard to unimodal scenarios. It also highlights the work by Pantic et al. [34], in which an adaptive neural network classifier is presented and assessed in different study cases, such as the combination of hand gestures and facial features or the combination of speech and video features. Güçlütürk et al. [14] studied the fusion of audiovisual and textual information for first impression analysis using Deep Residual Networks. All these works endorse the benefits of multimodal fusion approaches. More related works can be found in the work of Wu et al. [47], in which there are multiple strategies for multimodal methods for emotion recognition.

This work further explores the idea of frustration recognition by considering multimodal strategies in Game-Play scenarios using audiovisual data. More precisely, in contrast to previous works tackling this task, we propose the use of DNNs in order to automatically extract a set of meaningful descriptors from the audio and video sources of information so as to then assess the synergistic capabilities of different data fusion policies. This particular strategy is a new approach in this respect and, as stated in Section 5.2, outperforms the results achieved by state-of-the-art unimodal speech-based [41] and audio-and-video multimodal [42] approaches to a remarkable extent.

3 Methodology

In this section, we present our multimodal proposal for frustration detection during Game-Play, which considers the information from both audio and video recordings. We, therefore, first describe the approach considered for the audio source of information, and then do so for the video input. Finally, we introduce the proposal employed to combine both sources of information in order to determine the presence of frustration.

3.1 Audio classification

With regard to the audio data, previous work [42] proposes an approach based on LSTM to process a set of MFCCs previously extracted from the raw signal. While MFCCs have been considered to a great extent in audio speech analysis [31], it has been proved that a configuration based on Convolutional Neural Networks (CNNs) applied to an initial time-frequency representation of the signal is a more appropriate way in which to find suitable features with which to detect frustration in audio recordings [41]. Our audio analysis will, therefore, also consider this idea rather than that of using hand-crafted features.

Formally, let $\mathcal{X}_a$ be an audio recording in raw format and let $\mathcal{S}_a$ be its associated time-frequency representation. We consider a neural network architecture based on CNN layers in order to process $\mathcal{S}_a$ and automatically extract the most suitable features for the frustration classification. This scheme is shown in Fig. 1.

The main difference between the audio processing performed by [42] and our method is that we propose the use of convolution layers to automatically extract those features that are most appropriate from the point of view of the neural network. Our premise is that this will support the performance of the classification task, since the neural network is responsible for it. In contrast, [42] directly uses those features provided by the MFCCs, which, although

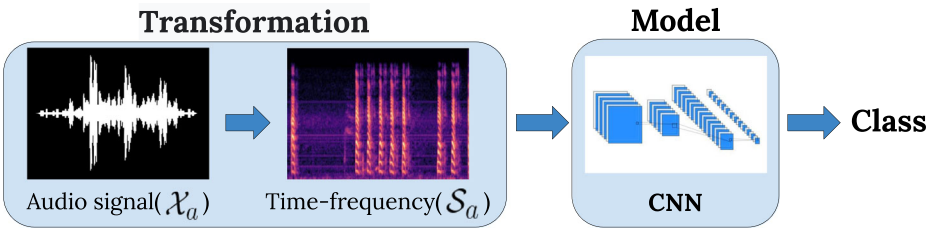


Fig. 1 Scheme of the unimodal frustration classifier approach based on speech data

they may appear adequate according to visual perception, may not be the most suitable for processing by the neural network.

Since the use of time-frequency representations is common in audio processing, note that different representations other than the MFCCs can also be considered. In Section 5.1.1, therefore, we shall study the input representation, along with other parameters, such as sample rate or the specific classifier model to be used, in order to discover the most suitable configuration for the task in hand.

3.2 Video images classification

Let $\mathcal{X}_v$ be a video composed by a sequence of N image frames such that $\mathcal{X}_v = \{x_{v,i} \mid 1 \leq i \leq N\}$.

As shown in Fig. 2, the objective of the proposed method is to detect frustration in single-frame video images $\mathcal{X}_v$. However, since only the facial expressions are useful as regards this detection, our system preprocesses each frame $x_{v,i}$ in order to obtain its trimmed and resized version $s_{v,i}$ and subsequently create a new video $\mathcal{S}_v = \{s_{v,i} \mid 1 \leq i \leq N\}$ in which only the face is present. This makes it possible to automatically extract features with CNN rather than having to employ hand-selected features. The details of this preprocessing are described in Section 5.1.2.[1]

An RNN model then provides the decision concerning the presence of frustration in the $\mathcal{S}_v$ trimmed version of the video. This kind of architectures receives a time-correlated sequence of data, in this case, each single $s_{v,i}$ image frame, and makes the classification decision after processing all the frames in the sequence.

Owing to the nature of the data in question and the relatively high sampling resolution, the particular facial expression is barely modified in consecutive frames, signifying that a model able to learn these long-term dependencies is required. We consequently considered the use of the well-known LSTM architecture [43], since it is capable of modeling such dependencies and is also considered in the work by Song et al. [42].

Please note that in this latter work, the set of hand-crafted features are a manual selection of the features from the Facial Action Coding System [9], which defines a series of facial features through the use of specific action units (AUs). The method selects 18 of these AUs features and performs the extraction for each frame $x_{v,i}$ of the initial video data $\mathcal{X}_v$, which are then handled by an LSTM in order to perform the frustration classification task. Nevertheless, as explained for the audio case in the previous section, since these features may not be the most representative for the task in hand, we propose to automatically learn

[1] Results without this trimming stage are not reported, since preliminary experimentation showed a remarkable drop in the overall performance of the system.

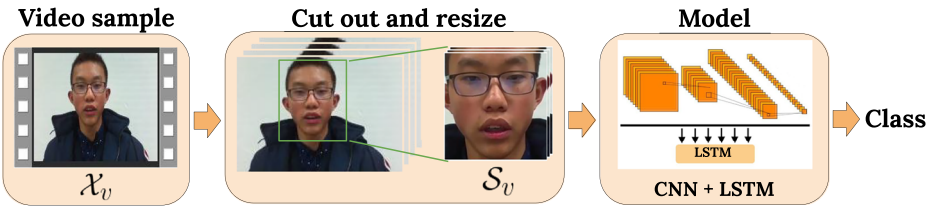


Fig. 2 Scheme of the unimodal frustration classifier approach for video images

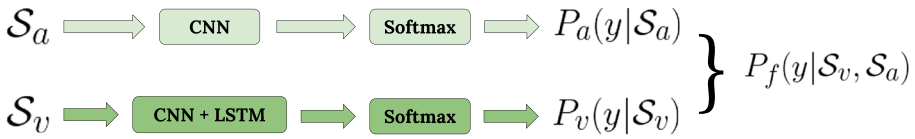
them by employing CNN layers from the trimmed version $\mathcal{S}_v$, since we assume that these features may provide a better overall performance given that they are estimated for the actual task in question.

3.3 Multimodal fusion

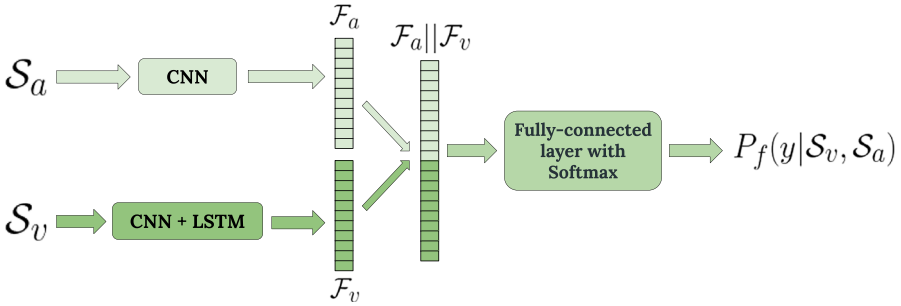
The key aspect of our proposal is the fusion of both audio and video information as described above. In this regard, we considered two particular types of combinations to be studied in the experiments: *decision fusion* and *feature fusion*.

This *decision fusion* case (see Fig. 3a) is based on the combination of the single decisions made by the audio and video models through the use of a weighting factor α. This fusion can be mathematically represented as

$$P_f(y|\mathcal{S}_v, \mathcal{S}_a) = \alpha\, P_a(y|\mathcal{S}_a) + (1-\alpha)\, P_v(y|\mathcal{S}_v), \tag{1}$$



(a) *Decision fusion* scheme. The independent decisions of each unimodal classifier are aggregated by a weighted (α) sum following Eq.1.



(b) *Feature fusion* scheme. $\mathcal{F}_a$ and $\mathcal{F}_v$ denote the features learned by the audio and video neural networks, respectively, which are concatenated $(\cdot||\cdot)$ before the final decision layer.

Fig. 3 Multimodal fusion schemes with which to compute $P(y \mid \mathcal{S}_v, \mathcal{S}_a)$

where $P_a(y|\mathcal{S}_a)$ $P_v(y|\mathcal{S}_v)$ represent the unimodal scores obtained by the DNNs over the time-frequency audio representation $\mathcal{S}_a$ and the face-based trimmed video data $\mathcal{S}_v$, respectively, while $\alpha \in [0, 1]$ represents a weighting factor.

The *feature fusion* alternative (see Fig. 3b) consists of designing an actual neural architecture with two inputs (audio and video) and one output with the classification result. The idea is based on two parallel streams that process $\mathcal{S}_a$ and $\mathcal{S}_v$ separately in order to obtain a particular representation of each modality by employing the neural network. These neural-based features, denoted as $\mathcal{F}_a$ and $\mathcal{F}_v$ for the audio and video streams, respectively, are then concatenated before making the final decision. Note that, in contrast to the previous case, this feature fusion is performed implicitly, given that the neural network is trained simultaneously with $\mathcal{S}_a$ and $\mathcal{S}_v$.

Both fusion modes considered will be evaluated and compared with their unimodal versions and other state-of-the-art results in Section 5.

4 Materials and metrics

This section presents the experimental setup considered in order to assess the frustration detection method proposed. More precisely, we shall describe the corpus used and the set of evaluation metrics.

In a technical sense, we considered the following libraries and toolkits for the proposed experimentation:

- **Python**.The research has been carried out in the Python programming language (v. 3.6.9).
- **Tensorflow** [1]: Framework for the implementation of the DNNs models (v.2.3.1).
- **Keras**: Collection of functions that make it possible to design architectures for neural networks. It also includes the tools employed to train the models and use them to evaluate performance (v. 2.4.3).
- **NumPy**. This is an open source library in the Python programming language used for the creation of vectors and multidimensional matrices. It provides powerful data structures that are capable of carrying out vector operations easily and efficiently. In this research, it will be used to deal with the structures of data used with the different neural architectures (v. 1.19.5).
- **librosa** library [27]: Audio analysis library used for the extraction of part of the time-frequency representations considered (v.0.8.0).
- **Cassani** toolkit [4]: Audio analysis toolkit used for the extraction of the Modulation-Spectral representation (v.0.1).
- **dlib** library [17]: Image analysis library used as a face detector when trimming the initial video data (v.19.18.0).

4.1 Corpus

For the evaluation of our approach, we have considered the Multimodal Game Frustration Database [42] of real-world recordings used by both the unimodal [41] and multimodal [42] state-of-the-art methods. The database comprises over 5 hours of 960 $\times$ 540-pixel video recordings at 30 frames per second split into 10-second excerpts and annotated as either representing or not representing frustration. This corpus was created thanks to the participation of 67 students from the Shanxi Province in China, with ages ranging from 12 to 16.

This dataset is already provided in three separated partitions for its correct benchmarking: a training set of 3,979 videos, a validation partition of 1,326 videos, and a test set of 1,328 elements. This configuration is maintained for comparison purposes in our experiment. Table 1 provides a summary of the details of the corpus.

4.2 Metrics

Since this work tackles an imbalanced scenario, as shown in Table 1, the evaluation requires a metric that is able to avoid any bias towards a particular class. We have, therefore, considered the use of the *F-measure* (F_1). In a two-class classification problem, as in our case, F_1 is described as

$$F_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

where TP represents the *True Positives* or correctly classified elements, FP represents the *False Positives* or type I errors and FN represents the *False Negatives* or type II errors.

Nevertheless, since the works compared consider the recall metric, we shall also introduce it for comparison purposes. In the same terms as the F_1, the *recall* $\mathcal{R}$ metric can be defined as

$$\mathcal{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

Moreover, since it may provide some additional insights into the performance of our proposal, we also consider the *precision* $\mathcal{P}$ metric, which is defined as:

$$\mathcal{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

Finally, in order to be consistent with the other works dealing with this problem, all these metrics will be computed by taking the minority class, i.e., the frustration class as the positive class.

5 Experimentation

This section presents the different results obtained after considering several DNNs models in several scenarios. We first consider an initial stage so as to correctly adjust the parameters of our proposed models in order to then assess their performance on the test partition and compare them with the results reported by the state-of-the-art works in this field.

Please note that the models were trained until 115 epochs, and that the weights of these epochs were maintained, which maximizes the results in the validation set. We used the well-known Adam optimizer [18] and the categorical cross-entropy loss function. We considered batch sizes, taking the maximum allowed by the memory restrictions up to 32.

Table 1 Amount of audiovisual recordings of the corpus considered with training, validating and testing partitions, according to the presence of frustration

Emotion	Train	Validation	Test
Neutral	3,564	1,188	1,189
Frustration	415	138	139

5.1 Model optimization

As mentioned previously, in this first part of this section our objective is to correctly optimize the different parameters of the model proposed. The stand-alone audio and video models are, therefore, first analyzed and optimized in order to subsequently study the multimodal cases proposed. Note that this optimization study considers only the training and validation partitions of the data in question.

5.1.1 Audio model

With regard to the implementation of the audio processing, as described in Section 3, our approach first processes the audio in order to extract the corresponding time-frequency representation. We considered the use of the well-known Mel spectrogram, as occurred in the work by Meishu et al. [41] in which its effectiveness for this particular task was proven. We additionally included two other representations widely applied in the field of audio-based emotion recognition: the Modulation-Spectral one [50] and MFCCs [22].

Both the Mel spectrogram and the MFCCs were obtained using the *librosa* library, whereas the Modulation-Spectral representation was attained by employing the *Cassani* toolkit. With regard to the neural architectures, we selected ResNet50 and Xception, since both have been used in the reference works considered [40, 41]. Note that, since we are modeling a two-class problem (frustration/non-frustration), the output of these models consists of two neurons representing each of these labels.

As mentioned above, we trained the models using the training and validation partitions of the corpus. The hyper-parameter tuning took place in a two-step fashion: we first studied the most suitable type of time-frequency representation for the task in hand and we then analyzed different values of hyper-parameters in order to eventually optimize the classification results.

For the first analysis, we fixed a sample rate (sr) of $22,050\ Hz$ and a hop length of 512 samples, which resulted in temporal and frequency resolutions of $23.2\ ms$ and $43\ Hz$, respectively. When the input corresponded to an MFCCs spectrogram, we fixed a total of 39 common coefficients: 13 MFCCs, 13 delta-MFCCs, and 13 second-order delta-MFCCs.

The results obtained with the validation partition using the aforementioned conditions are shown in Table 2.

Note that, although both ResNet50 and Xception attain high-performance figures, the Mel spectrogram is the scenario with the best recall results, with $\mathcal{R} = 93.7\%$, when

Table 2 Results on the validation partition in terms of $\mathcal{R}$, $\mathcal{P}$, and F_1 measured in % for audio classification according to the type of input data representation and the neural network model

Input	Model	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	F_1 (%)
MFCCs	Xception	91.4	80.2	85.4
	ResNet50	90.9	74.1	81.6
Modulation	Xception	82.2	70.8	75.7
	ResNet50	78.0	65.4	71.1
Mel	Xception	**93.7**	**87.8**	**90.6**
	ResNet50	85.3	69.9	76.8

The figures in bold type represent the best results obtained for each metric

compared to the MFCCs case which obtains a maximum recall of $\mathcal{R} = 91.4\%$ and the Modulation spectral which obtains $\mathcal{R} = 82.2\%$, all of which correspond to the Xception model. While ResNet50 also attains high-performance results, Xception outperforms all of them. For instance, upon considering the MFCCs input, the recall decreases from the $\mathcal{R} = 91.4\%$ obtained by the Xception model to the $\mathcal{R} = 90.9\%$ obtained with ResNet50; the Modulation spectral case also shows a reduction in recall from $\mathcal{R} = 82.2\%$ to $\mathcal{R} = 78.0\%$. With regard to the Mel spectrogram, which provided the best results of all the metrics considered, there was the same tendency as with the other results, decreasing the recall from the $\mathcal{R} = 93.7\%$ obtained by Xception to the $\mathcal{R} = 85.3\%$ obtained by ResNet50. Note also that the $\mathcal{P}$ and F_1 metrics are highly correlated with the previous recall analysis, obtaining the best result in all cases for the Mel spectrogram representation processed by the Xception model. All of the above eventually led us to select the Mel spectrogram as the input data for the audio classifier.

The second step was the optimization of the hyper-parameters of the input representation, i.e., the hop length and sample rate parameters. For this analysis, we considered three different sample rate values, along with two different hop lengths. Table 3 shows the results obtained from the previously selected input representation for the two neural models considered in the work.

One relevant difference between this and the previous experiment was that the best results for all the metrics considered were not consistently attained by one particular DNNs and hyper-parameter configuration. In general, the Xception model outperformed the ResNet50 for all metrics and hyper-parameter configurations considered. More specifically, according to the $\mathcal{P}$ and F_1 metrics, Xception achieved the best overall results for the task in hand with a hop length of 512 samples and 22,050 Hz of sample rate, attaining $\mathcal{P} = 87.8\%$ and $F_1 = 90.6\%$. With regard to ResNet50, the best results were $\mathcal{P} = 79.3\%$ and $F_1 = 81.0\%$. The first was obtained with 1,024 hop length samples and a sampling rate of 22,050 Hz, whereas the second was obtained with 512 hop length samples and 11,025 Hz. Since the

Table 3 Results obtained with the validation partition for the Mel spectrogram hyper-parameter tuning

Model	hop length (samples)	sr (Hz)	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	F_1 (%)
Xception	512	11,025	**94.3**	85.7	89.8
		22,050	93.7	**87.8**	**90.6**
		44,100	92.6	80.2	86.0
	1024	11,025	91.1	84.6	87.7
		22,050	93.1	86.5	89.7
		44,100	92.1	81.1	86.3
ResNet50	512	11,025	84.0	78.3	**81.0**
		22,050	85.3	69.9	76.8
		44,100	**85.5**	69.8	76.9
	1024	11,025	78.1	79.1	78.6
		22,050	81.0	**79.3**	80.1
		44,100	85.1	68.0	75.4

The figures in bold type highlight the best results for each model and metric considered

baseline state-of-the-art approaches [41, 42] reported the results in terms of $\mathcal{R}$, we consequently decided to focus on this particular metric in order to determine the configuration for the final experiments. The hyper-parameter optimization using the validation partition shows that Xception attained the best recall results ($\mathcal{R} = 94.3\%$) when considering a hop length of 512 samples and a sample rate of 11,025 Hz, which resulted in a temporal and frequency resolution of 46.4 ms and 21.5 Hz, respectively. We, therefore, considered this configuration for the final comparison with the state-of-the-art proposals.

5.1.2 Video model

As described in Section 3.2, our proposal first trims the frames in order to obtain smaller images that focus on the face of the individual using the face detector of the aforementioned *dlib* library. We have resized the resulting images to 64 $\times$ 64 pixels for reasons of simplification.[2]

The classifier used in this case is a combination of CNN and LSTM, whose details are shown in Table 4. We considered three variations of the model, henceforth denominated as $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, with an increasing number of layers, respectively.

One of the most important hyper-parameters to adjust in this section is the number of frames of the video data to be introduced into the network. As mentioned previously, the data collection contains 10-second video excerpts recorded at 30 frames per second, which results in videos of 300 frames. However, since the difference between consecutive frames in terms of expression is usually almost imperceptible, some of them could be ignored.

On that premise, in our experiments, we subsampled the number of frames so as to decrease the complexity of the learning task. More precisely, we experimented with two particular subsampling rates: taking one frame either every five or ten of the initial frames. This preprocessing results in excerpts of 30 and 60 frames, depending on the subsampling rate selected.

The results obtained with the different network models proposed for each of the subsampling policies considered are shown in Table 5.

We discovered that the most complex model—$\mathcal{M}_3$—provided the best results with the 60-frame policy for all the metrics computed. With regard to the other models, $\mathcal{M}_2$ attained better results than $\mathcal{M}_1$ for both subsampling policies, with a maximum recall of $\mathcal{R} = 90.1\%$ for $\mathcal{M}_1$ with 30 frames and $\mathcal{R} = 93.7\%$ for $\mathcal{M}_2$ with 30 frames. However, since $\mathcal{M}_3$ with 60 frames obtained the best overall recall results with $\mathcal{R} = 94.2\%$, this particular configuration was selected for the final experiments. The other metrics were found to follow a similar trend, and it consequently became clear that $\mathcal{M}_3$ was superior to the other alternatives for this task.

5.1.3 Fusion model

Having optimized the individual models for audio and video classification, we then combined them to build the multimodal audiovisual approach. As mentioned in Section 3.3, we considered two possible combinations: *decision fusion* and *feature fusion*.

With regard to the decision fusion, according to (1), it is necessary to study the optimal value of α for our multimodal method. Figure 4 shows the result of this combination with

[2]These particular values were selected on the basis of preliminary experiments.

Table 4 Architecture of the neural networks considered for the video classification task

# Layer	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
1	Conv(32 x 3 x 3) MaxPool(2 x 2)	Conv(32 x 3 x 3) MaxPool(2 x 2)	Conv(32 x 3 x 3) MaxPool(2 x 2)
2	Conv(64 x 3 x 3) MaxPool(2 x 2)	Conv(64 x 3 x 3) MaxPool(2 x 2)	Conv(64 x 3 x 3) MaxPool(2 x 2)
3	Conv(128 x 3 x 3) MaxPool(2 x 2)	Conv(128 x 3 x 3) MaxPool(2 x 2)	Conv(128 x 3 x 3) MaxPool(2 x 2)
4	Dropout(0.2) BatchNorm() Flatten()	Dropout(0.2) BatchNorm() Flatten()	Conv(256 x 3 x 3) MaxPool(2 x 2)
	LSTM(200) Dropout(0.2)	LSTM(200) Dropout(0.2)	
5	Dense(2, "softmax")	LSTM(100)	Dropout(0.2) BatchNorm() Flatten()
			LSTM(200) Dropout(0.2)
6		Dense(2, "softmax")	LSTM(100)
7			Dense(2,"softmax")

Conv($f \times w \times h$) denotes the convolution operation with f filters and a kernel with a size of $w \times h$ pixels and the Rectified Linear Unit (ReLU) as an activation function; MaxPool($w_p \times h_p$) represents a max-pooling operator with a size of $w_p \times h_p$ pixels; Dropout (d) applies this type of regularization with a ratio of d neurons; BatchNorm() performs the batch normalization operation; LSTM(n) includes an LSTM layer with n neurons and a hyperbolic tangent activation; Flatten() converts the input into a 1-D vector, and Dense(t, "af") is a fully-connected layer with t units and an "af" activation function

Table 5 Results obtained for the validation set of the video classification for the different models and subsampling policies considered

Model	# frames	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	F_1 (%)
$\mathcal{M}_1$	30	90.1	82.7	86.3
	60	88.9	81.1	84.8
$\mathcal{M}_2$	30	93.7	83.0	88.0
	60	92.8	80.6	86.3
$\mathcal{M}_3$	30	93.3	78.6	83.9
	60	**94.2**	**83.7**	**88.7**

The figures in bold type represent the best results attained for each metric

$\alpha \in [0, 1]$ and a granularity of 0.1. Note that $\alpha = 0$ corresponds to the unimodal audio model, while $\alpha = 1$ represents the unimodal video model.

In this graph, the red curve represents the performance of our multimodal proposal in terms of the recall metric. The maximum goodness of the method when the fusion is performed equally is evident, i.e., $\alpha = 0.5$, obtaining a result of $\mathcal{R} = 97.6\%$. If this value is compared with the second-best result—95.5% when $\alpha = 0.4$—, it will be noted that there is an important difference between them. Indeed, the error rate is reduced by over 46%, from 4.5% (when $\alpha = 0.4$) to 2.4% ($\alpha = 0.5$), thus proving the need to properly adjust this hyper-parameter. Furthermore, the frustration detection of our multimodal proposal outperforms both the unimodal models considered. After this analysis we, therefore, eventually selected $\alpha = 0.5$ for the final experiments. However, it is worth highlighting the wide range of values of α with which our multimodal approach based on decision fusion outperforms the performance of the unimodal methods, particularly when $\alpha \in [0.1, 0.8]$. Moreover, the experiment proves that this combination never provides worse figures than the unimodal

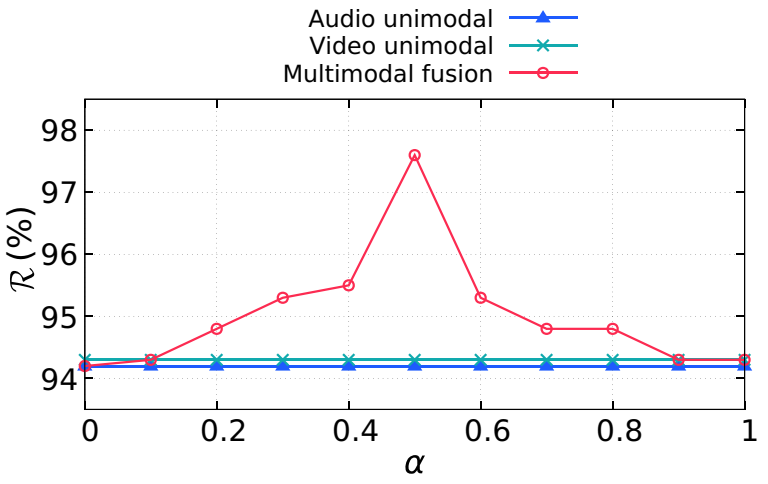


Fig. 4 Effect of the hyper-parameter α on the multimodal decision fusion when compared with the unimodal approaches in terms of $\mathcal{R}$ on the validation partition

approaches, the worst case being when $\alpha = 0.9$, in which the performance of the fusion model equals that of the video unimodal. These results reinforce the idea that our combination of audio and video data may be beneficial for the detection of frustration, and this premise, in fact, holds true for the majority of values of α.

The second fusion scenario is the feature fusion. This consists of combining the probability predictions of the two unimodal approaches described above in order to make a single decision about the presence of frustration aided by the individual descriptors extracted for each type of data. Table 6 shows a comparison between the best decision fusion model obtained ($\alpha = 0.5$) and this second type of multimodal fusion.

Focusing on the recall metric, the decision fusion provides a performance of $\mathcal{R} = 97.6\%$, whereas the feature fusion remains at $\mathcal{R} = 95.6\%$. The other metrics evaluated also attained a superior performance when comparing the decision fusion with the feature fusion.

5.2 Final results and comparison with the state of the art

In this section, we show the results obtained in the test partition of the corpus considered and comparatively assess them using those obtained by the state-of-the-art approaches mentioned above. As commented on previously, the reference works in literature that address the frustration detection task report only the recall metric, and we shall, therefore, consider only this particular measure for comparative purposes.

Table 7 shows the results obtained with the test partition of the corpus considered for all the different unimodal and multimodal proposals studied in this work, along with the results reported by the reference state-of-the-art methods.

Upon studying the results reported for the state-of-the-art methods, it will be observed that the multimodal approach by Song et al. [42] provides the worst recall results, with a value of $\mathcal{R} = 60.3\%$, while the unimodal method [41] outperforms those results with $\mathcal{R} = 93.1\%$. This remarkable difference between the scores obtained suggests that the aforementioned multimodal approach may not be properly exploiting its available sources of information, as a unimodal approach clearly outperforms it. Note that [41] performed this classification task through the use of more complex models, thus exploiting the capabilities of the neural networks, but that only audio was considered for the experiments, signifying that a valuable amount of information that could have been useful in the classification task was missed.

With regard to the unimodal audio and video strategies proposed in this work, note that they also attain competitive results, with recall values of $\mathcal{R} = 91.7\%$ and $\mathcal{R} = 89.6\%$, respectively. Note also that, when compared with [41], our audio-based model attains slightly worse results, since the recall decreases from $\mathcal{R} = 93.1\%$ to $\mathcal{R} = 91.7\%$. Although our unimodal approaches do not achieve the results obtained by the unimodal state-of-the-art method, they significantly outperform the existing multimodal method. The benefits of the influence of the automatic learning performed by the convolution layers when compared

Table 6 Comparison between the two fusion modalities considered on the validation partition of the corpus

Fusion type	$\mathcal{R}(\%)$	$\mathcal{P}(\%)$	$F_1(\%)$
Decision	**97.6**	**95.3**	**96.5**
Feature	95.6	86.9	91.0

In the case of the decision fusion, we consider the $\alpha = 0.5$ value, which optimizes the result.
The figures in bold type highlight the best results each figure of metric considered

Table 7 Results regarding the test partition of the corpus

Method		$\mathcal{R}$ (%)
State of the art		
	Multimodal [42]	60.3
	Unimodal [41]	93.1
Our approach		
	Audio	91.7
	Video	89.6
	Decision fusion	**95.9**
	Feature fusion	93.8

State-of-the-art methods are clearly differentiated from the results attained with the proposals described in this work. The figures in bold type represent the best results in terms of recall

with the hand-crafted feature extraction is, therefore, demonstrated. It is important to state that, while the actual audio architecture proposed in literature was reproduced here for the sake of comparability, our results are slightly lower than those reported in the reference work.[3]

However, when analyzing the two multimodal approaches proposed in this work, it is evident that they are consistently better than the figures attained by the unimodal architectures. With regard to feature fusion, it yields a recall score of $\mathcal{R} = 93.8\%$, which is thus better than the results of both $\mathcal{R} = 60.3\%$ by [42] and $\mathcal{R} = 93.1\%$ by [41]. The decision fusion attains the best overall recall value, with a score of $\mathcal{R} = 95.9\%$.

While the improvements made may appear to be relatively limited, it should be noted that decision fusion obtained only a 4.1% recall error which, when compared with the 39.7% error provided by the state-of-the-art multimodal approach, implies a relative improvement of 89.7% for this figure of merit. A similar analysis comparing the decision fusion with the audio-based state-of-the-art result also yields a relative improvement of slightly more than 40%. This shows that our multimodal method is much more reliable as regards providing feedback about users' game-play experience.

Although the feature fusion does not achieve the best performance, it is also worth highlighting its high recall, with 93.8%, which from the point of view of the error made, supposes 6.2% of the absolute recall error, or in other words, a relative reduction in the error of over 10% with respect to the best state-of-the-art method, i.e., the audio-based one. These results reinforce the premise on which we this research is based, i.e., that a multimodal approach may be a more appropriate means to carry out the classification task than unimodal models, since it is able to leverage the information provided by the two data sources involved—audio and video—in order to make remarkable improvements to the detection of the frustration when compared with single-source based models.

The results obtained, therefore, confirm that both multimodal fusion methods presented in this work are considerably better than unimodal approaches, which only tackle either audio or video information. Moreover, the results obtained also outperform those attained by the state-of-the-art works addressing this same task, including the multimodal and the unimodal proposals. Finally, note that the experimentation presented validates our proposed

[3]This is probably owing to certain implementation nuances that are difficult to replicate completely.

multimodal strategies as it attains the best recall scores of all the benchmarked methods, with the decision fusion scheme being that which obtains the best overall results.

6 Conclusions

Frustration detection constitutes the discovery of an emotion of particular interest in the video game industry, since it is directly correlated with the users' engagement. However, its estimation and tracking remains an open research question, especially when invasive tracking devices are not considered. In this context, this work introduces a new approach with which to detect frustration in non-invasive scenarios by considering multimodal strategies that fuse the information extracted with a feature-learning stage based on Deep Neural Networks (DNNs) obtained from the different individual data sources. More precisely, when considering audiovisual data, the idea is to extract meaningful descriptors from the audio and video sources of data and combine them in order to eventually perform frustration detection. Note that this fusion synergistically exploits the capabilities of DNNs to obtain a suitable set of features with which to detect the frustration emotion for each particular data source.

We specifically propose two multimodal approaches with which to merge the audio and video pieces of information: a *decision-level* approach, which combines the individual decisions made with each data source, and a *feature-level* policy, which combines the individual features extracted by the DNNs from each type of data in order to then make a single decision. The experiments reveal that the two proposed multimodal fusion methods outperform unimodal strategies, along with providing better results than the state-of-the-art schemes obtained from the related literature. The best results were specifically obtained with the *decision-level* fusion, with a recall score of 95.9%, thus improving the error rate by almost 90% in comparison to the multimodal state-of-the-art approach, and by over 40% when compared to that of the unimodal audio-based method.

The remarkable improvement obtained with our approach validates not only the use of multimodal approaches as regards merging different sources of information in a synergistic manner, but also the use of DNNs as feature extractors for emotion recognition tasks other than those related to frustration. However, this proposal still has considerable constraints, such as the limited amount of sources of information or the simple neural architectures considered. In this respect, future work should consider the inclusion of other complementary data i.e., eye gazing or information related to playing time. We also aim to further study other fusion modalities, such as *early fusion* or to explore *feature fusion* methods in greater depth. Finally, a further objective is that of exploring other neural architectures based on residual connections since, as proved by other works in literature, they may further improve the results obtained.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/. Software available from tensorflow.org
2. Bahreini K, van der Vegt W, Westera W (2019) A fuzzy logic approach to reliable real-time recognition of facial emotions. Multimed Tools Applic 78(14):18,943–18,966
3. Carvalhais T, Magalhães L (2018) Recognition and use of emotions in games. In: 2018 International conference on graphics and interaction (ICGI), pp 1–8. IEEE
4. Cassani R (2019) Amplitude-modulation-analysis-module, https://github.com/MuSAELab/amplitude-modulation-analysis-module Accessed April 2022
5. Chandrasekar P, Chapaneri S, Jayaswal D (2014) Automatic speech emotion recognition: a survey. In: 2014 International conference on circuits, systems, communication and information technology applications (CSCITA), pp 341–346. IEEE
6. Chen D, James J, Bao F, Ling C, Fan T (2016) Relationship between video game events and player emotion based on eeg, pp 377–384
7. Dworak W, Filgueiras E, Valente J (2020) Automatic emotional balancing in game design: use of emotional response to increase player immersion. In: Marcus A, Rosenzweig E (eds) Design, user experience, and usability. Design for contemporary interactive environments. Springer International Publishing, Cham, pp 426–438
8. Ebrahimi Kahou S, Michalski V, Konda K, Memisevic R, Pal C (2015) Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 467–474
9. Ekman R (1997) What the face reveals: basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA
10. Fernandez R, Picard RW (1998) Signal processing for recognition of human frustration. In: Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, ICASSP'98 (Cat. No. 98CH36181), vol 6, pp 3773–3776. IEEE
11. Gadekallu T, Rajput D, Reddy P, Lakshman K, Bhattacharya S, Singh S, Jolfaei A, Alazab M (2020) A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. J Real-Time Image Proc, 1–14
12. Gilleade KM, Dix A (2004) Using frustration in the design of adaptive videogames. In: Proceedings of the 2004 ACM SIGCHI international conference on advances in computer entertainment technology, pp 228–232
13. Granato M, Gadia D, Maggiorini D, Ripamonti LA (2020) An empirical study of players emotions in vr racing games based on a dataset of physiological data. Multimed Tools Applic 79(45):33,657–33,686
14. Güçlütürk Y, Güçlü U, Baro X, Escalante HJ, Guyon I, Escalera S, Van Gerven MA, Van Lier R (2017) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9(3):316–329
15. Gunes H, Piccardi M (2005) Affect recognition from face and body: early fusion vs. late fusion. In: 2005 IEEE international conference on systems, man and cybernetics, vol 4, pp 3437–3443. IEEE
16. Horlings R, Datcu D, Rothkrantz LJ (2008) Emotion recognition using brain activity. In: Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing, pp II–1
17. King DE (2009) Dlib-ml: A machine learning toolkit. J Mach Learn Res 10:1755–1758
18. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: 3rd International conference on learning representations. San Diego, USA
19. Kosa M, Uysal A (2021) Need frustration in online video games. Behav Inform Technol, 1–12

20. Kwon OW, Chan K, Hao J, Lee T (2003) Emotion recognition by speech signals. In: Eighth European conference on speech communication and technology
21. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
22. Likitha M, Gupta SRR, Hasitha K, Raju AU (2017) Speech based human emotion recognition using mfcc. In: 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pp 2257–2260. IEEE
23. Lim JZ, Mountstephens J, Teo J (2020) Emotion recognition using eye-tracking: taxonomy, review and current challenges. Sensors 20(8):2384
24. Lim W, Jang D, Lee T (2016) Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), pp 1–4. IEEE
25. López C, Tucker C (2018) Toward personalized adaptive gamification: A machine learning model for predicting performance. IEEE Trans Games 12(2):155–168
26. Malta L, Miyajima C, Kitaoka N, Takeda K (2010) Analysis of real-world driver's frustration. IEEE Trans Intell Transp Syst 12(1):109–118
27. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, vol 8, pp 18–25
28. Miller MK, Mandryk RL (2016) Differentiating in-game frustration from at-game frustration using touch pressure. In: Proceedings of the 2016 ACM international conference on interactive surfaces and spaces, pp 225–234
29. Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2227–2231. IEEE
30. Ng Y, Khong C, Thwaites H (2012) A review of affective design towards video games. Procedia - Social and Behavioral Sciences 51, 687–691 (2012). The World Conference on Design, Arts and Education (DAE-2012), May 1-3. Antalya
31. Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015) Audio-visual speech recognition using deep learning. Appl Intell 42(4):722–737
32. Noroozi F, Kaminska D, Corneanu C, Sapinski T, Escalera S, Anbarjafari G (2018) Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing
33. Oh S, Lee JY, Kim DK (2020) The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals. Sensors 20(3):866
34. Pantic M, Caridakis G, André E, Kim J, Karpouzis K, Kollias S (2011) Multimodal emotion recognition from low-level cues. In: Emotion-oriented systems, pp 115–132. Springer
35. Picard RW (2000) Affective computing
36. RM SP, Maddikunta PKR, M P, Koppu S, Gadekallu TR, Chowdhary CL, Alazab M (2020) An effective feature engineering for dnn using hybrid pca-gwo for intrusion detection in iomt architecture. Comput Commun 160:139–149
37. Sharma G, Dhall A (2021) A survey on automatic multimodal emotion recognition in the wild. In: Advances in data science: methodologies and applications, pp 35–64. Springer
38. Snoek CG, Worring M, Smeulders AW (2005) Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on multimedia, pp 399–402
39. Soleymani M, Pantic M, Pun T (2011) Multimodal emotion recognition in response to videos. IEEE Trans Affect Comput 3(2):211–223
40. Solovyev RA, Vakhrushev M, Radionov A, Romanova II, Amerikanov AA, Aliev V, Shvets AA (2020) Deep learning approaches for understanding simple speech commands. In: 2020 IEEE 40th international conference on electronics and nanotechnology (ELNANO), pp 688–693. IEEE
41. Song M, Mallol-Ragolta A, Parada-Cabaleiro E, Yang Z, Liu S, Ren Z, Zhao Z, Schuller B (2021) Frustration recognition from speech during game interaction using wide residual networks. Virt Real Intell Hardware 3(1):76–86
42. Song M, Yang Z, Baird A, Parada-Cabaleiro E, Zhang Z, Zhao Z, Schuller B (2019) Audiovisual analysis for recognising frustration during game-play: introducing the multimodal game frustration database. In: 2019 8th International conference on affective computing and intelligent interaction (ACII), pp 517–523. IEEE
43. Staudemeyer RC, Morris ER (2019) Understanding lstm–a tutorial into long short-term memory recurrent neural networks, arXiv:1909.09586
44. Toselli AH, Vidal E, Casacuberta F (eds.) (2011) Multimodal interactive pattern recognition and applications, 1st edn. Springer

45. Vasan D, Alazab M, Wassan S, Naeem H, Safaei B, Zheng Q (2020) Imcfn: image-based malware classification using fine-tuned convolutional neural network architecture. Comput Netw 171(107):138
46. Wimmer M, Schuller B, Arsic D, Radig B, Rigoll G (2008) Low-level fusion of audio and video feature for multi-modal emotion recognition. In: Proc. 3rd Int. conf. on computer vision theory and applications VISAPP, Funchal, Madeira, Portugal, pp 145–151
47. Wu CH, Lin JC, Wei WL (2014) Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing, 3
48. Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. IEEE Trans Audio Speech Lang Process 16(2):448–457
49. Yannakakis GN, Isbister K, Paiva A, Karpouzis K (2014) Guest editorial: emotion in games. Institute of Electrical and Electronics Engineers
50. Zhu Z, Miyauchi R, Araki Y, Unoki M (2016) Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants. In: INTERSPEECH, pp 262–266