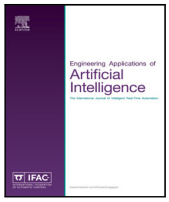
Engineering Applications of Artificial Intelligence 130 (2024) 107708

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets

U. Bilotti [a], C. Bisogni [a,*], M. De Marsico [b], S. Tramonte [b]

[a] University of Salerno, Via Giovanni Paolo II, 132, Fisciano, 84084, Italy
[b] Sapienza University of Rome, Via Salaria 113, Rome, 00198, Italy

ARTICLE INFO

Keywords:
Emotion recognition
Multimodal emotion recognition
Multi-input model
Biometrics
Deep learning

ABSTRACT

The aim of this paper is to investigate emotion recognition using a multimodal approach that exploits convolutional neural networks (CNNs) with multiple input. Multimodal approaches allow different modalities to cooperate in order to achieve generally better performances because different features are extracted from different pieces of information. In this work, the facial frames, the optical flow computed from consecutive facial frames, and the Mel Spectrograms (from the word melody) are extracted from videos and combined together in different ways to understand which modality combination works better. Several experiments are run on the models by first considering one modality at a time so that good accuracy results are found on each modality. Afterward, the models are concatenated to create a final model that allows multiple inputs. For the experiments the datasets used are BAUM-1 ((Bahçeşehir University Multimodal Affective Database - 1) and RAVDESS (Ryerson Audio–Visual Database of Emotional Speech and Song), which both collect two distinguished sets of videos based on the different intensity of the expression, that is acted/strong or spontaneous/normal, providing the representations of the following emotional states that will be taken into consideration: angry, disgust, fearful, happy and sad. The performances of the proposed models are shown through accuracy results and some confusion matrices, demonstrating better accuracy than the compared proposals in the literature. The best accuracy achieved on BAUM-1 dataset is about 95%, while on RAVDESS it is about 95.5%.

1. Introduction

Biometrics systems have been originally devised to recognize a person based on physiological characteristics, such as fingerprints, facial appearance, iris patterns, or behavioral characteristics including handwriting style, e.g., in signature, key typing pace, or walking pattern. Therefore, biometrics-related technologies have been traditionally used for forensic applications, e.g., fingerprint-based identity identification and verification, and user authentication. However, the scope has increasingly expanded to, e.g., smart ambient personalization and even healthcare. Automatic emotion recognition is one of the possible applications of the analysis of different biometric traits, especially facial expression, voice, gesticulation style, and body posture. Actually, these are the physical and behavioral cues that allow interpreting the emotions expressed by another person in order to communicate, understand and empathize with each other. Allowing computers to develop this ability could have a major impact on improving our daily lives by supporting, e.g., affective computing and a more humanized interaction with technology, and unintrusive diagnosis of particular diseases or mental health disorders. The wide range of applications includes augmented video gaming, intelligent interaction with smart devices, empathic virtual assistants, social avatars or robots, marketing, vehicle automation, and advanced personalized e-learning.

For several years now, Face Emotion Recognition (FER), like many other biometric applications, has benefited from the potential of Artificial Intelligence methods reaching ever higher levels of accuracy and speed (Canal et al., 2022). However, with the increased availability of computational resources and theoretical advancements, it makes more sense to classify facial expressions as a dynamic event, through all possible detectable information from a video sequence, rather than from static facial images (Bisogni et al., 2023). As a further extension, useful information can also be extracted from the voice characteristics extracted from the video or from other behavioral and vital signals (Ahmed et al., 2023). This work investigates multimodal emotion recognition from face and voice. In order to determine

* Corresponding author.
E-mail addresses: ubilotti@unisa.it (U. Bilotti), cbisogni@unisa.it (C. Bisogni), demarsico@di.uniroma1.it (M. De Marsico), tramonte.1464086@studenti.uniroma1.it (S. Tramonte).

https://doi.org/10.1016/j.engappai.2023.107708
Received 11 August 2023; Received in revised form 30 October 2023; Accepted 11 December 2023
Available online 14 December 2023

which modality combination performs better, facial frames, optical flow computed from subsequent facial frames, and Mel spectrograms are extracted from videos, and either used as input for a multimodal CNN-based architecture, or the obtained responses from single-mode CNNs are fused in various ways. The document continues as follows. Section 2 introduces the theoretical background of emotion recognition and states the problem tackled in this paper from the point of view of automatic multimodal emotion classification. Section 3 describes the two multimodal datasets used for the experiments, namely RAVDESS and BAUM-1. Section 4 briefly discusses some related work especially focused on the same datasets. Section 5 underlines some aspects of the following sections. Section 6 introduces the models used for the video channels and the tested fusion strategies for 2 inputs, and reports the experimental results. Section 7 does the same for 3 input-based emotion recognition after introducing the method to handle the audio channel. Section 8 reports a comparison with recent works in the literature using the same benchmarks. Finally, Section 9 draws some conclusions and sketches some future works.

2. Background and problem statement

2.1. Theoretical framework

There are many intrinsic variables that characterize an emotional phenomenon. Much before the rise of computer science-based automatic emotion recognition, other branches of science tried to investigate which variables are most related to our human emotional reactions, possibly stemming from the precise context in which we are immersed. Their findings have led to the construction of several theoretical models that can contemplate possible linguistic and sociocultural differences. Thus, without any loss of generality, the choice of a theoretical framework may follow from the particular problem proposed. For the automatic classification of emotions, the most suitable choice is the *discrete approach*. This approach involves the identification of different emotions that establish a discrete partition of classes into which the individual detected phenomena can be allocated. Given the complexity of sentiment-related events, the first problem is the choice of the categories/labels to take into account, which is a fundamental component of the domain (Clavel and Callejas, 2016).

The process of emotion discretization is resolved by Ekman et al. (1999) through the identification of six main emotion families, within which other sub-families can be detected. The study discusses evidence to support the universality of facial expressions, which are categorized as happiness, sadness, anger, fear, surprise, and disgust. Facial expressions present strong similarities in different cultures, even though possible differences are governed by "display rules" in different social contexts. For instance, Japanese subjects in formal contexts refrain from showing their true emotion. Since babies usually exhibit a wide range of facial expressions, it is possible to hypothesize that these expressions are innate (Izard, 1994). Subsequent studies have broadened this set to eight: anger, fear, sadness, disgust, surprise, happiness, calm, and neutral. However, as early as Tomkins (1962), who distinguishes emotions arising from negative and positive affect, a new classification strategy begins to emerge. Indeed, as Posner et al. (2005) point out, the collateral problems of Ekman's approach cannot be solved by simply increasing the number of classes. Actually, as testified by related studies, the same facial expression can be paired with different emotions, depending on different personality, temperament, and social traits not only of the observed subject but also of the observer (Stahelski et al., 2021). The proposed suggestion is instead to identify a discrete set of universal constituents of emotions.

2.2. Problem statement

Affective computing aims to enable intelligent and possibly empathic and proactive systems to interpret human emotions and trigger suitable actions that depend on the application context. Of course, this field entails cross-disciplinary work involving psychology, neurophysiology, and social and cognitive sciences as a source of authoritative theory and computer science for the design of reliable and accurate methodologies. The evolution of the technologies and their increased functionalities and availability has allowed us to pass from unimodal emotion recognition, based on either video or audio, to the present research works focusing on multimodal emotion recognition using both information channels or even more, e.g., adding text or body posture and gestures or signals from wearable sensors. However, most state-of-the-art frameworks still rely on a single modality, i.e., audio or video. These systems generally lack sufficient robustness, accuracy, and generalizability, which could be increased by multimodal fusion of data (Poria et al., 2017). In particular, this work deals with the fusion of video and audio signals.

There are two main different methodologies for dealing with emotion recognition, more specifically when regarding facial expressions:

- image-based methods (static);
- video sequence-based methods (dynamic).

The static approach is able to effectively derive spatial information, but because emotion is more of a dynamic event, images are not able to capture the entailed temporal variability, so it is more effective to extract frames from video sequences to get more information.

Contrarily to facial expression, emotion detection from audio mainly relies on processing a continuous signal based on two kinds of signal characteristics:

- audio features;
- prosodic features.

Several studies demonstrate that emotional characteristics are better highlighted by prosodic features, which are therefore widely used in the literature (Luengo et al., 2005).

The information underlying emotion classification can be obtained by applying two main strategies:

- by hand-crafted features engineering
- by using deep learning-trained models

Regarding the face, the first ones employ techniques like Local Binary Patterns from Three Orthogonal Planes (TOP) (Guo et al., 2014) or Local Phase Quantization from TOP (Jiang et al., 2014), which stem from already well-known image descriptors that have been extended to be applied on video sequences. Other solutions exploit the dynamics of face landmarks as in Bisogni et al. (2023). As for audio, the most used features are those usually extracted for speech analysis, e.g., Mel Frequency Cepstral Coefficients (MFCC) collectively forming a mel-frequency cepstrum (MFC), or features expressing prosodic aspects like pitch, intensity, and speaking rate, or measuring voice quality.

The second strategy involves deep neural networks which have been proven to be able to learn strong features. Their multiple convolution and pooling layers are well-suited for many computer vision tasks, such as image classification, object detection, and face recognition.

3. Multimodal datasets

This section will focus on the datasets used for this work. The collection of datasets to test FER methods, besides benefiting from the improved technological context, occurred progressively as the theoretical frameworks developed. It is worth pointing out that, first of all, the construction and labeling of a good database of facial expressions requires expertise, time, and training of subjects. In addition, the use

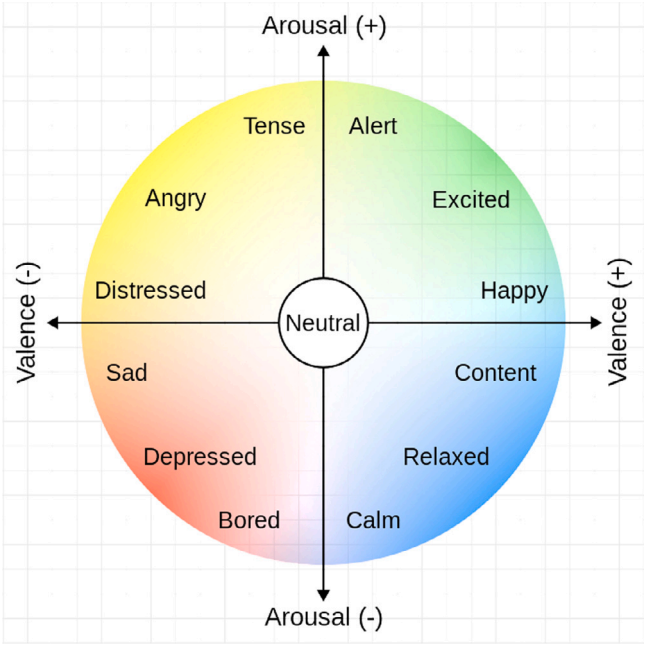


Fig. 1. The circumplex model with valence and arousal.

of posed versus spontaneous (authentic) expressions in selecting facial stimuli, is a debated issue. Actually, most psychologists deem that spontaneous expressions are the only ones truly reflecting an inner emotion, and therefore consider them as the only ones deserving investigation (Sebe et al., 2007). on the contrary, most early datasets for emotion recognition were posed ones. An overview of the main available datasets can be found in the work of Mollahosseini et al. (2019). The one they created contains more than 1 million images extracted from the Internet and classified according to the circumplex model of affect (Russell, 1980). The axes of this model are in relation to valence (from positive to negative) and arousal (intensity) of an emotion (Feldman Barrett and Russell, 1998) (Fig. 1). Of course, the position of the affect expression on the arousal axis can influence the accuracy in identifying the valence. This is why acted emotions are easier to classify. Once the problem of authentic expressions is solved, another one arises. Any dataset containing images of facial expressions cannot fully represent the dynamic and multisensory nature of the phenomenon of emotion.

In light of these findings, the Ryerson Audio–Visual Database of Emotional Speech and Song *(RAVDESS)* (Livingstone and Russo, 2018) was one of the datasets selected for the experiments included in this paper, which aim at investigating multimodal emotion recognition from face expression as well as from voice features. Unfortunately, at present, the number of multimodal datasets suited for this research is even more limited, not to mention the spontaneous ones. RAVDESS dataset collects 7356 recordings with acted-emotional content. It contains clips collected from 24 subjects, 12 male and 12 female, all professional actors. Each subject performs 60 trails, so the dataset contains a total of 1440 clips that include calm, happy, sad, angry, fearful, surprise, and disgust, each of which is produced at two levels of intensity: normal and strong, plus the addition of the neutral and calm baseline expressions. Only five emotions will be generally taken into consideration if not differently pointed out: angry, disgust, fear, happy and sad, for allowing both a performance comparison with the other used dataset BAUM-1 (see below) and some cross-dataset test. The files are separated into two vocal channels (speech and song) and three AV modalities (full AV, video-only, and audio-only). The eight emotional categories are represented by a single actor in each file and except for the neutral emotion, which only comprises regular intensity, the emotions are created at both regular and strong degrees of emotional intensity. Fig. 2 shows samples from the dataset. Notice the uniform background. While this simplifies face segmentation, on the other side, as it will be clear later, may represent a fixed feature in the images that may hinder the generalizability of classifiers trained on such a set.

While the choice of RAVDESS is dictated by the possibility of easily fulfilling the characteristics assumed by the theoretical model, the numerous practical applications suggest the analysis of emotional phenomena as natural and spontaneous as possible. Given that most datasets of this type, as opposed to the universal nature of the emotional phenomenon, consist of video–audio files in English, the identification of more general characteristics can be achieved by comparison with manifestations detected in languages other than English. In *BAUM-1* (Zhalehpour et al., 2017) there are 31 Turkish participants (17 men and 14 women). The collected audio–video clips are divided based not only on the kind of emotion but also on whether the emotion is acted or spontaneous. During the video collection, the participants view a series of still images and brief video clips that have been painstakingly planned and timed to elicit a variety of feelings and mental states. Without employing prewritten scripts, the respondents speak in their own words (in Turkish) about how they feel and what they think about the stimulus they have just observed on a screen. The target emotions include the six basic ones, namely happiness, anger, sadness, disgust, fear, and surprise, as well as boredom and contempt. Although there is no instruction given to the subjects regarding how to make facial expressions, using video snippets to generate emotions is a tried-and-true technique (Gross and Levenson, 1995). Fig. 3 shows some subjects from this dataset. It is possible to notice how that quality is lower than RAVDESS: a frame extracted from RAVDESS has a weight of around 45 KB and a dimension of $420x \times 20$ pixels, while a frame extracted from BAUM1 has a weight of 15 KB and a dimension of 250×250 pixels, so it has a lower quality. Most of all, the background is still uniform but of bright green.

The differences in the two datasets will allow testing non only the accuracy of the proposed approaches, but also their cross-dataset generalizability. It is worth noticing that equivalent differences also hold for the audio channel: the Mel spectrograms extracted from RAVDESS have a silence period both at the beginning and at the end of the recording, making them look very different with respect to the BAUM-1's ones.

It is well known in literature that emotion recognition performances may be influenced by the language of the speaker (Pell et al., 2009). A more recent survey (Rajoo and Aun, 2016) found that there are linguistic variations in emotion perception, with English having a better recognition rate than Malay and Mandarin. Since RAVDNESS uses English and BAUM-1 Turkish, even in this case it is reasonable to expect language-dependent results. In fact, the literature reports better performances on RAVDNESS than on BAUM-1. On the other hand, the results of experiments summarized in this paper present no substantial difference when the audio signal is combined with facial cues.

4. Related works

Given the strong interest of the scientific community as well as the industrial community in the problem dealt with herein, a variety of solutions have been proposed, which differ in terms of both solution strategy and type of data to be analyzed. In this chapter, we will focus on multimodal approaches using data extracted from video and audio files.

4.1. Multimodal emotion recognition

In addition to being a dynamic phenomenon, which therefore requires a study that considers its evolution over a period of time, human emotion is a manifestation resulting from various conscious and unconscious psycho-motor processes. Consequently, as for the human being and also for a machine, a better interpretation of emotion can

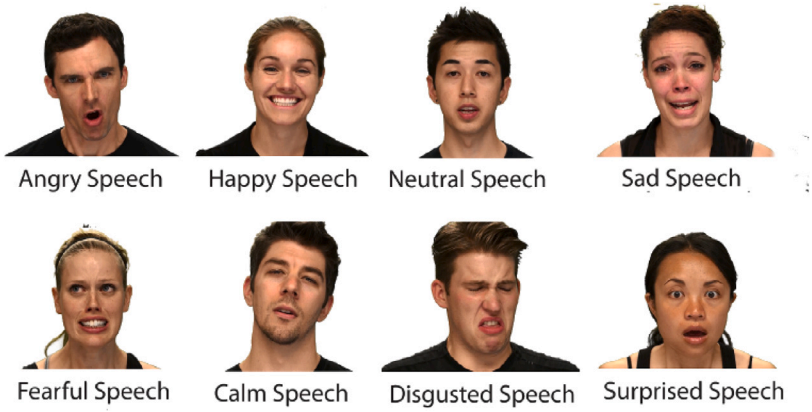


Fig. 2. Samples from RAVDESS dataset.

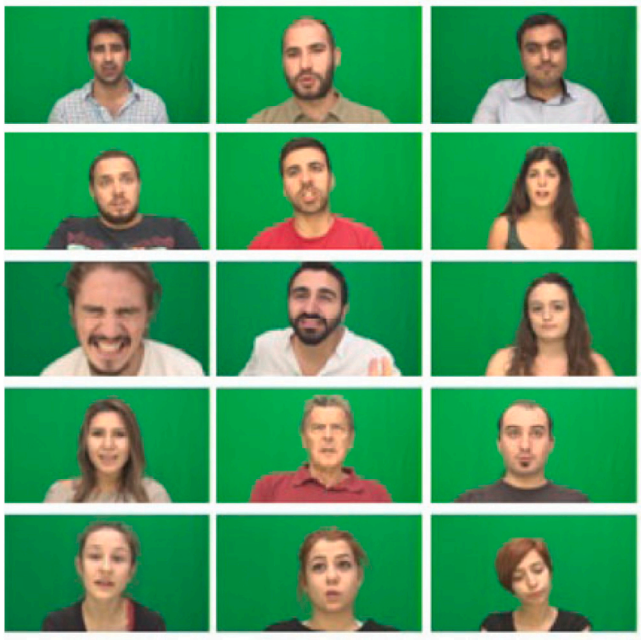


Fig. 3. Subjects participating in BAUM-1 dataset.

be achieved when it is possible to effectively interpret different types of information. These reflections have encouraged the creation of emotion recognition methods that exploit different types of data and are therefore multimodal. The literature classifies the various conventional multimodal methods into two major groups based on the different fusion strategies of the data collected.

Methods that perform *Early Fusion (or Multisensor Fusion or Data-Level Fusion)* assemble a vector consisting of the data from the different sources, which then becomes the input of a machine-learning algorithm. The use of raw data, besides not ensuring good performance, can lead to problems of inconsistency and desynchronization of data. While there are several solutions to the synchronization problem, the non-homogeneity of the raw data is solved by replacing it with features extracted manually.

The other main fusion strategy is the *Late Fusion (or Decision-Level Fusion)*. The methods belonging to this class involve the training of a separate model for each different mode in the first phase, followed by a second phase for the creation of a final model that learns from the features detected in the first phase. The way the methods of this class are structured, they are immune to the synchronization problem; on the other hand, the risk of producing redundant information is greater (Luna-Jiménez et al., 2021).

In addition to the above, deep neural networks (DNNs) allow a third form of multimodal fusion, i.e., intermediate fusion of learned representations. Deep-learning architectures learn a hierarchical representation of the data across the hidden layers so that learned representations between different modalities can be fused at various levels of abstraction (Ramachandram and Taylor, 2017).

The work in literature that appears most similar to ours is Mamieva et al. (2023). The authors here also use Mel for speech, and face expression to perform the classification. However, they prefer MFCC (Mel-Frequency Cepstral Coefficients) over the Mel Spectrogram. While the Mel Spectrogram is a time–frequency representation, the MFCC represents a set of compact spectral features. The authors also use an attention mechanism at the last layers of the network, while for the architecture proposed here the fully connected layers are preferred. Finally, the three inputs get the same weight, while in Mamieva et al. (2023) the audio outputs are first concatenated and then combined with the facial output.

The choice of following one strategy over another is therefore to be judged on the basis of how well the different modes may be correlated. The experimentation reported in this paper is not intended to highlight the greater effectiveness of one class over another, but rather to provide reference points for new, more efficient solutions. In the following, we briefly analyze some literature proposals using the datasets described above.

4.2. Related works on RAVDESS

The work of Radoi et al. (2021) uses a simple Convolutional Neural Network (CNN) consisting of 3 main convolutional blocks and a fully connected layer for the extraction of the video features. Each RAVDESS video is divided into N time sequences, and for each sequence, a frame is chosen at random as a representative of the sequence. For each such frame, the relative time instant is saved, and this is used to construct an interval of the audio file to be associated with the frame. The emotion classification problem of an audio–video pair becomes a family of N emotion classification problems for the derived frame–interval audio pairs. Thus, after an evaluation of the N pairs, the starting pair is assigned the emotion that verifies the greater sum of the N contributions of the frame–audio pairs. To increase accuracy in the training phase, Data Augmentation is performed.

The method proposed in Siddiqui and Javaid (2020) exploits two channels, one for the images and the other for the audio file. The information from the images is derived in both visible (RGB) and infrared spectra, the extracted features of which are preliminarily merged and then merged again with the features from the audio file. The algorithm chosen to perform the fusion of the RGB and infrared images

is Canonical Correlation Analysis (CCA) and the precision achieved using the resulting feature vector is greater than using both vectors resulting from the individual inputs. The accuracy achieved is 86.36% but the dataset they created i.e. VIRI (which has no audio files) exploits chimeric subjects, taking the audio channel from RAVDESS audio files. The results in this case are less questionable than using video and speech, since there is, in principle, a lower correlation to maintain when creating the chimeric subjects.

In the work of Middya et al. (2022), 54 possible combinations of video and audio feature extractors are analyzed (9 extractors for video and 6 extractors for audio). The features obtained from the two modalities are concatenated (vectorized) and the resulting vector is given as input to a dense layer with a ReLU activation and 128 filters. The final layer is a fully connected layer with 8 units for each of the eight groups of emotions. For the ultimate emotion categorization, it makes use of the softmax function. The precision achieved among all 54 possible combinations is 86% with the combination V8+A4 (see the cited paper for details).

Speech and facial expressions are integrated to accomplish emotion recognition in the work on paper (Luna-Jiménez et al., 2021). These modalities are integrated utilizing two separate models joined by a late fusion method. The audio-related model will use speech signals as input and transform them into spectrograms, after which it will use a Support Vector Machine to recognize speech emotions.

The work of Alshamsi et al. (2018) proposes a lighter framework that can be used on mobile devices. They use a module that extracts facial features and classifies them with Support Vector Machines (SVM) and another module that does the same but on the speech signal. The results are combined by a fusion features module, which then proceeds with the classification of the emotion.

4.3. Related works on BAUM-1

One of the first research groups to conduct experiments on this dataset was that of Zhang et al. (2021). Being made up of videos of emotions expressed by common subjects person (not actors), this dataset is on the one hand more representative of a possible real-world application context and on the other hand more complicated as an object of analysis. More than the result in terms of accuracy achieved (amounting to 44.06% following 300 epochs), the observations made on the different CNN methods used and the two fusion strategies will be useful for the following work.

Four types of data are combined in Jiang et al.'s work (Jiang et al., 2020) using a DBN algorithm to extract EEG features, an AlexNet DCNN network to extract advanced features from speech's Mel spectrum, a VGGNet DCNN network to extract features from users' facial expressions, and a CNN network to extract features from social network text. The huge amount of data and the method workflow suggested by Zhang et al. allow 90% accuracy after 200 epochs.

An example of a different architecture is proposed in Hsu and Wu (2020). In the first step, a VGG model is used for the extraction of video features and a CNN model for audio features. In the second step, a Neural Tensor Network is used to determine the level of correlation between two modalities based on how representative they are of the same emotion. Finally, the composed feature vector becomes the input of a Coupled Long Short-Term Memory (C-LSTM) (Su et al., 2020b) that performs a feature refinement process through the alternating cooperation of another pair of video–audio models.

5. A study on different fusion approaches for multimodal emotion recognition: models, results, and comparisons

The following sections introduce the models and the different multimodal strategies that have been tested on both single datasets and in a cross-dataset setting. When not differently pointed out, the experiments aimed to classify five emotions: angry, disgust, fear, happy, and sad. The experiments in Section 6 were run over BAUM-1 acted subset, except for the comparison with the approach in Zhang et al. (2019) that is detailed below. They are maintained apart for easier comparison with Zhang et al. (2019) that uses the same visual channels without speech. Section 7 presents models and experiments to further add the speech channel to emotion recognition.

6. Two input CNN using frames and optical flow frames

6.1. Emotion recognition from facial expressions and optical flows: a relevant compared approach

The first step of our experiments focuses on the video channel only, though exploiting both RGB images and optical flow images extracted from the videos. It takes inspiration from the approach presented in Zhang et al. (2019) and tested on BAUM-1 dataset. We give a description of that approach which is later compared with our proposal. In Zhang et al. (2019), to take full advantage of a 2-stream CNN, a fusion network built with a Deep Belief Network, DBN, captures the spatiotemporal features. Each video is passed as input to the model by generating the facial frames (see Fig. 4 for a sequence of frames from a video labeled as *disgust*) and the optical flow facial frames. The facial frames are used to represent the spatial features, while the optical flow frames represent the temporal features. The video is segmented, with a segment length $L = 16$. If frames are shorter or longer, the frames are eliminated or duplicated if needed. From each segment spatial and temporal inputs are generated. For the temporal input, optical flow images are extracted from the segments. The optical flow images are first normalized into the interval $[0, 255]$, then the images are resized to $227 \times 227 \times 3$, where the last number refers to the number of channels used. Note that from a video segment $L = 16$, 15 optical flow images are generated as inputs of the temporal CNN. For the spatial input, each frame of the video segment undergoes real-time face detection to crop the face out of the image with a $150 \times 110 \times 3$ size, which is then resized to $227x \times 227 \times 3$. Note that the first frame is always discarded, so that out of $L = 16$, 15 frames are extracted and given as input to the spatial CNN.

The spatial and temporal CNNs used in Zhang et al. (2019) both have the same structure, which is a VGG16 model (Fig. 5), a variant of the VGG. They are both pre-trained on ImageNet, and then individually re-trained using the backpropagation techniques, so to update the network parameters. Instead of going in the softmax layer, the outputs of the second fully connected layer of the two CNNs, which contains 4096 units, get concatenated into an 8192 vector that is then passed as input to the deep fusion network, the DBN. The DBN is a multi-layered neural network formed by stacking a series of Restricted Boltzmann Machines, RBMs, each of which is a bipartite graph and because of the multiple RBMs, DBNs can effectively discover properties and hierarchical features of input data. Finally, a softmax layer is added to perform classification. The DBN is trained in two steps, with an unsupervised pre-training that aims to update the RBMs weights and supervised training that updates the network parameters with fine-tuning. After this training, the output of the last hidden layer of the DBN will represent the jointly learned discriminative spatiotemporal features of the video segments. Then an average-pooling approach on all divided segments is applied to produce a fixed-length global video feature representation, and a linear SVM classifier is adopted to perform the final classification of the video sequences. More details can be found in Zhang et al. (2019).

6.2. 1-Input emotion recognizer from either video or optical flow with BAUM-1 dataset

For this set of experiments, a modified model structure was adopted as a preliminary step toward developing a streamlined framework for facial frame classification. Running a single epoch would consume over

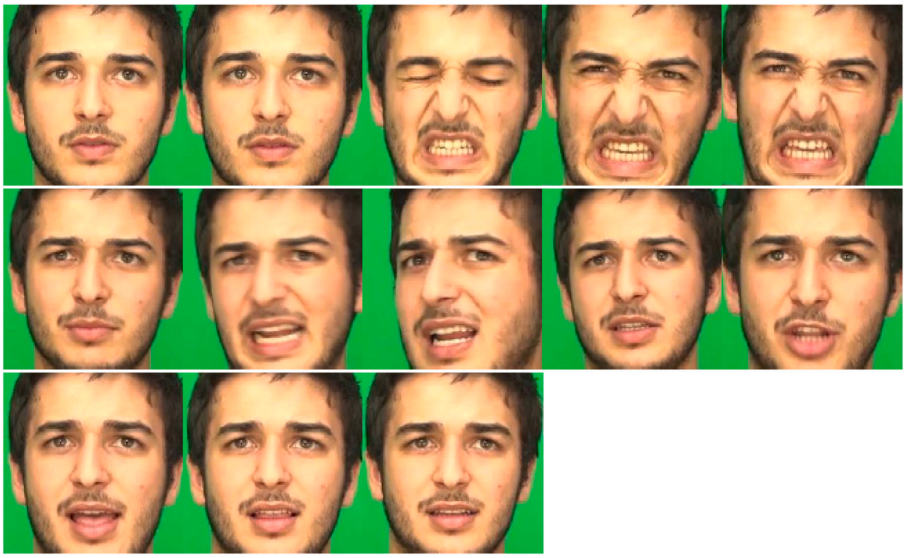


Fig. 4. A sequence of frames from a BAUM-1 video labeled as *disgust*.

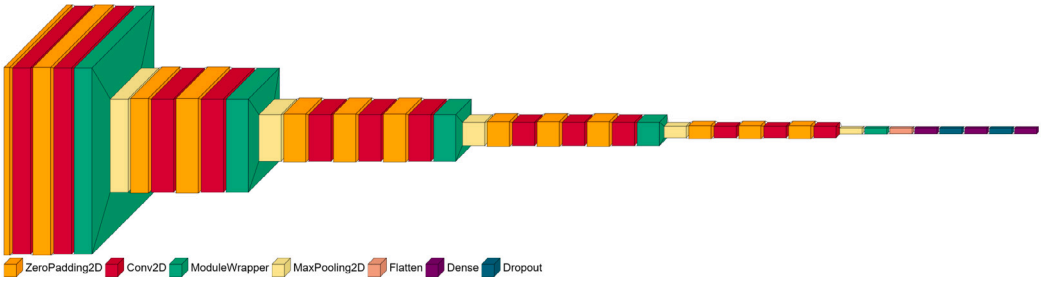


Fig. 5. The VGG16 model.

20 min, making it impractical for our purposes. To address memory limitations and lengthy execution times, the complete VGG16 architecture was reduced to the one in Fig. 6. The architecture was developed in Python using the Keras API.[1] The sequence of Conv2D layer is specified by a 3 × 3 kernel with output spaces of size 32, 64, 18, 56, and 51 respectively, while the MaxPooling2D interleaved layers downsample the input along by a × pooling window size. The Dropout layer randomly sets input units to 0 with a specified rate at each step during training, which helps prevent overfitting. The chosen 0.25 and 0.4 values were proven to be optimal after several experiments. The Flatten layer flattens the input. The three dense layers have output dimensions of 2048, 1024, and 5 respectively, with the last one activated by the softmax function, which actually performs the classification and takes as input the number of classes that the model is classifying. The average time to compute an epoch with the selected model is 60 s. For this reason, the time required for 40 epochs is 40 min. Considering that the training process is carried out only once and off-line, this time can be considered reasonable given the obtained performance.

In contrast to Zhang's approach in Zhang et al. (2019), our method does not involve passing videos as input in the form of stacked frames. Instead, each frame is individually processed by the model for classification. In other words, in our case, the model learns to classify each frame of the video independently. To generate Dense optical flow images are also generated from the acted clips in BAUM-1. This process involves computing optical flow vectors,[2] which are then mapped onto colors for better visualization (Fig. 7).

The optical flow works on the assumption that there is no change in the pixel intensity of an object when considering consecutive frames and that the motion of close pixels is similar. Assuming dt as the time in which the movement happens and (dx, dy) the distance covered by the pixel (x, y), according to the first assumption $I(x, y, t) = I(x + dx, y + dy, t + dt)$, where t is the initial time and I is the pixel intensity. By approximating the right side of the equation by the Taylor series, removing common terms, and dividing by dt, the following equation can be obtained:

$$\frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + f_t = 0 \tag{1}$$

This is called the optical flow equation. Then, the Lucas–Kanade method (Lucas and Kanade, 1981) assumes that the 3 by 3 neighbors of the pixel have the same motion. It follows that the problem consists of nine equations with two unknown variables. By the least squares fit method, this solution can be obtained, and from a practical point of view, it is possible to give some points to track and receive the optical flow that well represents a small motion of those points.

To evaluate the model, accuracy results are calculated on the test data. Although several tests were carried out, from the confusion matrix pairs for the two different modes, the highest accuracy values always resulted from the frames rather than the optical fluxes. Consequently, the model's capability to extract meaningful features from the optical flow frames is comparatively limited compared to the video frames. This is testified by the results in Table 1 in the next subsection. However, despite this limitation, the optical flow frames will still be included and examined in subsequent experiments.

[1] https://keras.io/
[2] https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html

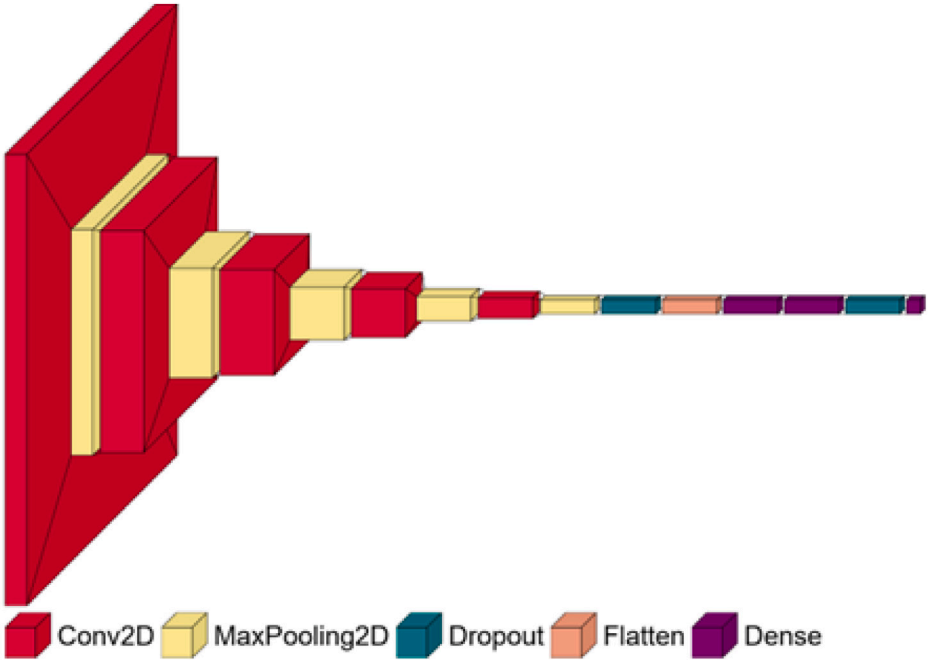


Fig. 6. Proposed model for emotion recognition by frames and/or MEL-spectrograms.

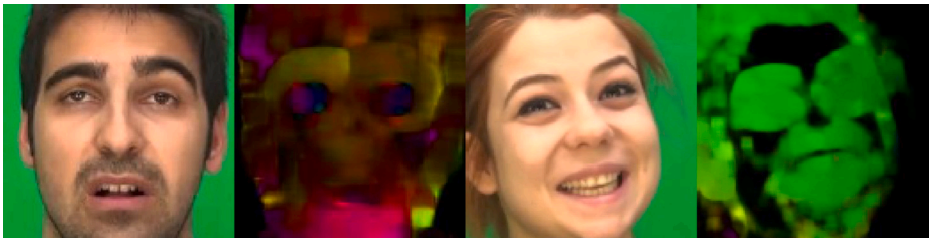


Fig. 7. Examples of frames and the corresponding optical flow frames computed considering the magnitude and direction of the displacement of points with respect to the previous frame.

Table 1
2-input model evaluation results compared with results of 1-input model evaluation results over frames and optical flow frames.

Type of input \Epochs	10	20	30	40
Two-input (posed)	0.7822	0.7902	0.803	0.8223
Frames	0.7739	0.8229	0.8378	0.8589
Optical flow frames	0.556	0.6297	0.6088	0.63
Two-input (spontaneous)	0.83	0.8218	0.835	0.846

6.3. 2-Input multimodal emotion recognizer from video and optical flow with BAUM-1 dataset

In order to test the results achieved by the combined use of the two frame sequences, the merge modality proposed is an early fusion strategy. This means that the features extracted from the different modalities are joined before training a final model. An advantage of this strategy is that the model can discover the correlations between features to remove redundant information, and also learn the interactions between modalities. The main problem that could occur using this strategy regards synchronization, which is needed when aligning the data from different modalities because each modality might have a different sampling rate. This will happen for this bunch of experiments because only the video frames are considered, which have the same sampling rate. Differently from Zhang et al. (2019), we used the concatenate layer in Keras. It works by taking as inputs a list of tensors, all of the same shape except for the concatenation axis, and returns a single tensor that is the concatenation of all inputs. To this aim, we slightly modified the architecture used for the 1-modality experiments. As can be seen from Fig. 8, the architecture used for the individual modes is adapted for an early fusion strategy. The first part is used to determine a tensor for each mode, and the resulting tensors are then concatenated before becoming the input for the second part of the architecture. Conv2D and MaxPooling2D layers remain the same as in the previous experiments, i.e., five blocks of ConvD-MaxPooling2D, followed by a single DropOut Layer (parameter 0.5) and a Flatten layer to instantiate two intermediate Tensors, one for each frame sequence. Then they are passed over to perform the concatenation, after which the missing layers (with respect to Fig. 6) are added again. In more detail, after concatenation, there is a Flatten layer, two Dense Layers (size 2048 and 104 respectively), a Dropout (0.4), and a Dense layer activated by softmax for 5 classes (see Fig. 8) The findings in Table 1 demonstrate that this 2-input approach using both frames and optical flow frames does not actually improve upon the outcomes of employing frames alone. This may either depend on the limited discriminative power of optical flow frames alone or demonstrate that the concatenation layer approach could be potentially inadequate to capture distinguishing characteristics from these two types of inputs. The comparison with the results in Zhang et al. (2019) required making the experiments on the spontaneous videos of BAUM-1, and also to consider the 6 emotions taken into account there: anger, disgust, fear, happiness, sadness, and surprise. The results show that the proposed 2-input classifier achieves better accuracy than with acted videos. This is probably due to the larger amount of available training data. Moreover,

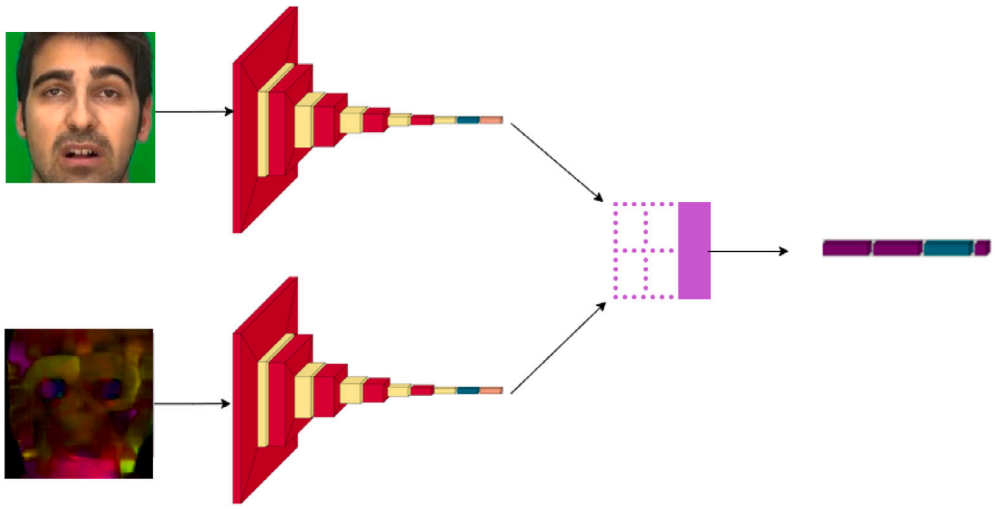


Fig. 8. Our proposed 2-input CNN model.

the best accuracy reported in Zhang et al. (2019) with the same kind of videos is 55.85%. Even with only 10 epochs, the results reported testify a relative improvement of about 48%.

7. Three input CNN

7.1. Emotion recognition from speech and video: a relevant compared approach

The work in Luna-Jiménez et al. (2021) is presented here in some detail because it represents an example of a different approach to multimodal emotion recognition with respect to Zhang et al. (2019) for two reasons: (1) speech and facial expressions are combined to perform emotions recognition; (2) these modalities are combined by employing two independent models connected by a late fusion strategy. The latter means that different models, as many as the modalities, are trained and then combined to create a final model that learns the posterior or the concatenated features.

The speech emotion recognizer implements two transfer learning techniques for avoiding training a CNN from scratch. The techniques used are either feature extraction or fine-tuning. The feature extraction uses two different frameworks, DeepSpectrum (Amiriparian et al., 2017) and PANNs (Kong et al., 2020). In the first case, the descriptors are denoted as deep-spectrum features. They are derived from forwarding spectrograms through very deep task-independent pre-trained CNNs: activations of the last two fully connected layers from two common image classification CNNs, AlexNet and VGG19, are used as feature vectors. The model performs the speech emotion recognition on them using a Support Vector Machine (SVM). In the second case, pretrained audio neural networks (PANNs) are trained on the large-scale AudioSet dataset. The best accuracy results are achieved when fine-tuning the CNN-14 of the PANNs framework. This seems to confirm that, when sufficient data is not available, the training can be more robust when it does not start from scratch, of course, given that the tasks are similar.

Also for the facial emotion recognizer, the cited work explores both feature extraction and fine-tuning. This time, as a pre-trained model, a Spatial Transformer Network, or STN (Jaderberg et al., 2015) is used. The STN was modified in order to use saliency maps on the frames, so as to capture the main regions of the face, because the more relevant the information to process is the better the STN's accuracy. The frames are extracted at 30 fps, the lateral pixels are removed, and then the frames are resized to 48 × 48, as STN expects at its input. In both feature extraction and fine-tuning, the authors consider a pooling solution, with majority voting, and a dynamic solution with a modified version of a bi-LSTM (Baziotis et al., 2018), with an attention layer. The final multimodal emotion recognizer is implemented with a late fusion strategy so that all the above structures can be exploited without synchronization problems. So, the best posteriors embeddings of both the speech emotion recognizer, from the fine-tuned CNN-14, and the facial emotion recognizer, from the bi-LSTM trained from the STN's outputs, are concatenated. Performance comparison of the final model exploits Logistic Regression, a k-NN with majority voting, and an SVM with 'linear' and 'rbf' kernels.

7.2. Speech based emotion recognizer

A mandatory step for being able to use CNNs with audio files is to represent audio as images because CNNs only accept arrays that represent images as input. The images extracted from an audio file are typically representative of some of the audio's features, depending on the kind of chosen representation. So, for being able to transform audios in images, the python library Librosa[3] is used here. This library provides a set of methods to extract audio features easily. For the experiments, the following audio representation are considered:

- Waveform: refers to the shape of the signal over time
- Spectral centroid: provides the center of gravity of the magnitude spectrum, so where most of the energy is concentrated
- Spectral bandwidth: is derived from the spectral centroid, as the range of interest around the spectral centroid
- Chromagram: analyzes meaningful pitches in the audio
- Spectrogram: represents the spectrum of frequencies of the signal over time
- Mel spectrogram: is a spectrogram with a Mel scale on the y-axis.

Fig. 9 shows an example of each of these features, extracted from the same audio file.

Depending on the type of representation used, the images generated from an audio file are often representatives of some of the audio's qualities. A slightly modified version of the model described in the previous sections (see Fig. 6) was used to evaluate which graphic representation of the audio files was the most successful. In more detail, it is composed of four Convolutional layers, each of which is followed by a MaxPooling layer, of a Flatten layer, two Dense layers, a Dropout layer, and at last a softmax layer. The model uses Adam optimizer and the *sparse_categorical_crossentropy* as a loss function. It has a lower

[3] https://librosa.org/doc/latest/feature.html

Table 2
Evaluation results over different audio representations.

Epochs \Audio features	Waveform	Spectral centroid	Spectral bandwidth	Chromagram	Spectrogram	Mel spectrogram
10	0.3041	0.2693	0.2951	0.2225	0.3011	**0.4**
20	0.4141	0.3083	0.3076	0.3386	0.3314	**0.5537**
30	0.42	0.45	0.3261	0.48	0.3371	**0.6096**
40	0.40	0.43	0.3427	0.3965	0.3371	**0.5721**

Table 3
Comparison of model changes.

Epochs \Model change	Starting model	Change 1	Change 2	Change 3	Change 4	Change 5
10	0.40	0.35	0.3741	**0.4964**	0.2348	0.1761
20	0.5537	0.50	0.6287	**0.519**	0.4871	0.4021
30	0.6096	0.44	0.608	**0.6817**	0.3401	0.4524
40	0.5721	0.44	0.6833	**0.7159**	0.6950	0.5677

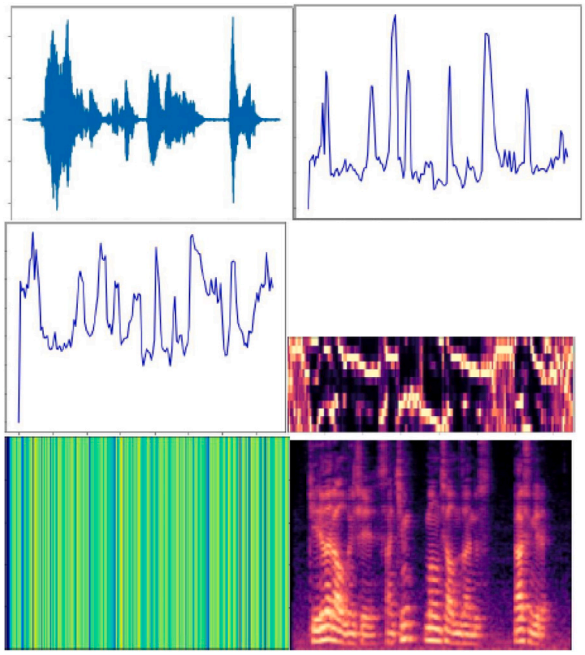


Fig. 9. From the top left: a waveform, a spectral centroid, a spectral bandwidth, a chromagram, a spectrogram, and a mel spectrogram extracted over the same audio file.

number of Conv2D layers, and the learning rate used is greater. Table 2 demonstrates that the MEL Spectrogram achieves better performance, and also improves in accuracy by increasing the number of epochs. Therefore, only Mel spectrograms are utilized in further studies.

After identifying the Mel spectrogram as the best representation, we aimed to improve the level of precision of the method. A series of experiments were therefore carried out by applying some changes to the architecture in a generally incremental way, meaning that for example, change number 3 is an experiment that runs using also changes 1 and 2. The tested changes are:

(1) add extra Conv2D and connected MaxPool2D layers;
(2) change the number of filters in the layers to a bigger number;
(3) change the batch size to 30;
(4) change the batch size to 50;
(5) change batch size to 30 and add an extra Conv2D and MaxPool2D.

Table 4
Comparison of model changes.

Epochs \Dataset	Original	Augmented x 2	Augmented x 4
10	0.4964	**0.5684**	0.84
20	0.519	**0.8536**	0.953
30	0.6817	**0.9007**	0.9154
40	0.7159	**0.8043**	0.9451

Table 3 shows that change (3), including (1) and (2) also, returned a better accuracy for any number of epochs. In practice, the structure of the resulting best architecture is the same as shown in Fig. 6.

To further increase performance, an additional crucial step is taken. All the aforementioned outcomes are derived from a dataset comprising 172 samples, which is relatively small for effective training. So additional experiments were conducted to implement data augmentation and effectively expand the dataset. Data augmentation involves generating new data by introducing minor yet significant modifications to the original samples. This approach not only increases the dataset size but also aids in constructing more generalized models and mitigating overfitting. Commonly adjusted sound characteristics for generating new samples include noise addition, which adds white noise randomly to the original sample; time shifting, which performs a shifting of the wave by a specified factor on the time axis; pitch shifting, which changes the pitch of a sound without altering the speed; time stretching, which changes the speed and duration of a sample without affecting the pitch. Fig. 10 shows an example of how these alterations change an original MEL Spectrogram. The example testifies how some modifications produce very similar images, that the CNN is very smart at comparing. In order to avoid redundancies, it was decided to use only two of the four possible alterations, resulting in a dataset of 516 samples. The adoption of this strategy immediately rewards: using only two modifications and training for 30 epochs produces only slightly lower performance than applying 4 modifications for the same number of epochs, reaching about 90% accuracy (see Table 4). Further insight is given by Fig. 11.

Although seemingly counterintuitive, we prefer to apply fewer variations since, in the case of the 30 epoch, the performances are not significantly improved by double variations. On the other hand, fewer variations mean a smaller dataset, or, in other words, that a lower computational training time is required.

7.3. Synchronizations of audio–visual data

Once the best representation of the audio signal has been established, and thanks to the flexibility of the realized structure, different mode combinations can be considered. However, before proceeding with the new experiments, a problem of synchronization must be solved. While a video has a series of frames, the audio file is unique and

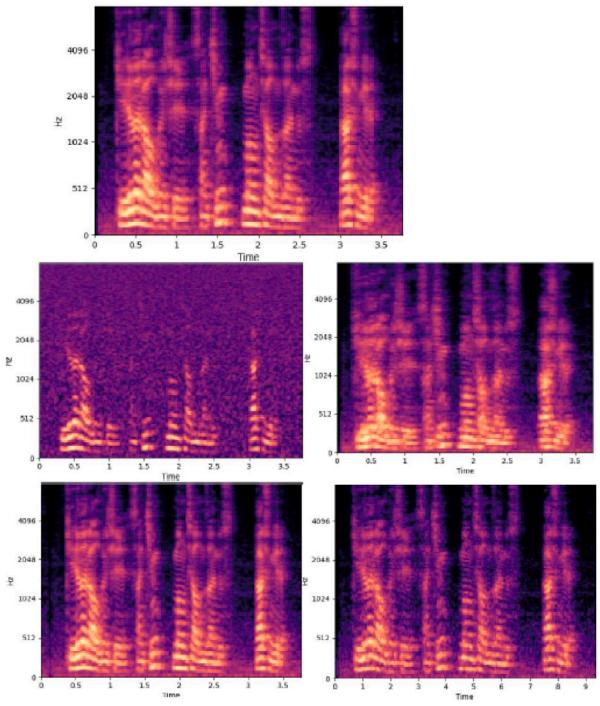


Fig. 10. From the top, an original MEL Spectrogram, noise addiction, pitch shifting, time shifting, and time stretch.

its segmentation would severely damage the information. This means that, while with different video channels like, e.g., RGB and optical flow, we can adopt a solution like the one described in Section 5, this is hardly possible with audio and video. Using the same MEL Spectrogram for each frame of a video is not an effective approach. This was demonstrated when discussing audio augmentation, where it was found that excessive audio transformations lead to redundant data. The solution was inspired by the audio augmentation strategy. Instead of extracting all video frames, only three of them will be taken from each video. In other words, three frames, three optical flow frames, and three MEL Spectrograms will be extracted from each video. Specifically, the spectrograms will include the original one, as well as the augmented versions with added noise and pitch shifting. Nevertheless, a smart and useful selection of three frames from videos poses a new challenge. The random extraction of a frame every ten ones proved to be ineffective. It was also observed that within a video sequence, several frames lack significance in representing the intended label. Looking at Fig. 4, which shows an example of a sequence of frames extracted from a video for disgust, it is easy to observe that not all frames are representative of the emotion that the video is labeled with, especially the first ones, which seems more a neutral expression, and some other frames, which seems more an angry expression. A specific model must be created to extract the most significant frames. Several studies have been done in the literature to correctly match a section of an audio file with the best of the possible frames representing the emotional state. In the work of Huang K. Huang et al. (2019) a classification of audio segments is made according to the different content of silence or emotionally significant verbal (tone of voice, frequencies) or non-verbal (laughter, shouting, sighing, etc.) expressions. Dominguez's work (Domínguez Bajo et al., 2016) analyzes an important feature of speech such as prosody as an effective tool for a weighted decomposition of audio files with the presence of verbal expressions into prosodic units.

In Table 5 we present the main differences in the proposed architectures. Activation functions, optimization strategies, and batch sizes are the same. Each single network has 5 convolutional layers and

Table 5
The main differences among the several tested architectures.

Input	Conv levels	Dense levels
F	5	3
Opt	5	3
Mel	5	2
F & Opt	10	3
Opt & Mel	10	3
F & Mel	10	3
3-input	15	3

Table 6
The comparison of results with different 2-input models and a 3-input model on BAUM-1. Opt stands for Optical flow frames, F for frames and Mel for Mel spectrograms.

Model type \Epochs	10	20	30	40
F & Opt	0.6187	0.8224	0.8412	0.8477
Opt & Mel	0.7558	0.8564	0.8803	0.8667
F & Mel	**0.7558**	**0.9416**	**0.9667**	**0.95**
3-input	0.8374	0.9212	0.9264	0.9392

3 dense layers, with the exception of the Mel spectrogram network, which uses only 2 dense levels. The combined networks parallelize the convolutional layers, so in total, the number of layers is the sum of the layers of each network. The dense levels are always 3 because they come after the fusion of the 2 or 3 convolutional parts of the networks.

7.4. Results on BAUM-1

To achieve the goal of audio–video synchronization, we propose a specific model using the same structure as depicted in Fig. 6. However, the model is trained using frames carefully selected "by hand" from all the acted data in BAUM-1. A total of 1254 thoughtfully chosen frames are used for training, and the trained model is saved. This specialized model is then employed to preprocess the BAUM-1's acted dataset again. It extracts precisely three frames from each video, and the extraction of optical flow frames is dependent on the frame chosen by the specialized model. So now, from BAUM1's acted dataset, 513 samples are ready to be used for training and testing the new 3-input model. But first, some small changes on the model with early fusion shown above are required in order to allow it to work with three inputs, instead of two, taking into account that the same architecture has proven to be effective for both video and audio. The tensors that are used for the concatenation are instantiated the same as before. Then, the model is created as above, but now instead of two tensors, three tensors are passed to the concatenate method. As we can see from Table 6, the best result is achieved by combining the video frames and the Mel spectrogram images, given more than 10 epochs of training. This underlines again that the optical flow frames are the less discriminative kind of input, which causes the performances to be lower. At last, something that it is important to point out is that having a set of data that was extracted by the special model improved the previously found results over the 2-input model previously shown (see Table 1).

7.5. Results on RAVDESS

As previously indicated, there are a number of distinctions between the BAUM-1 and the RAVDESS dataset, with the most significant ones relating to the caliber of the video, the quantity of data, and the fact that the RAVDESS subjects are actors. However, we can hypothesize that these differences do not necessitate a different choice of either architectures or modality combinations that are the same used for BAUM-1. In this case, frames are sampled randomly from videos. In

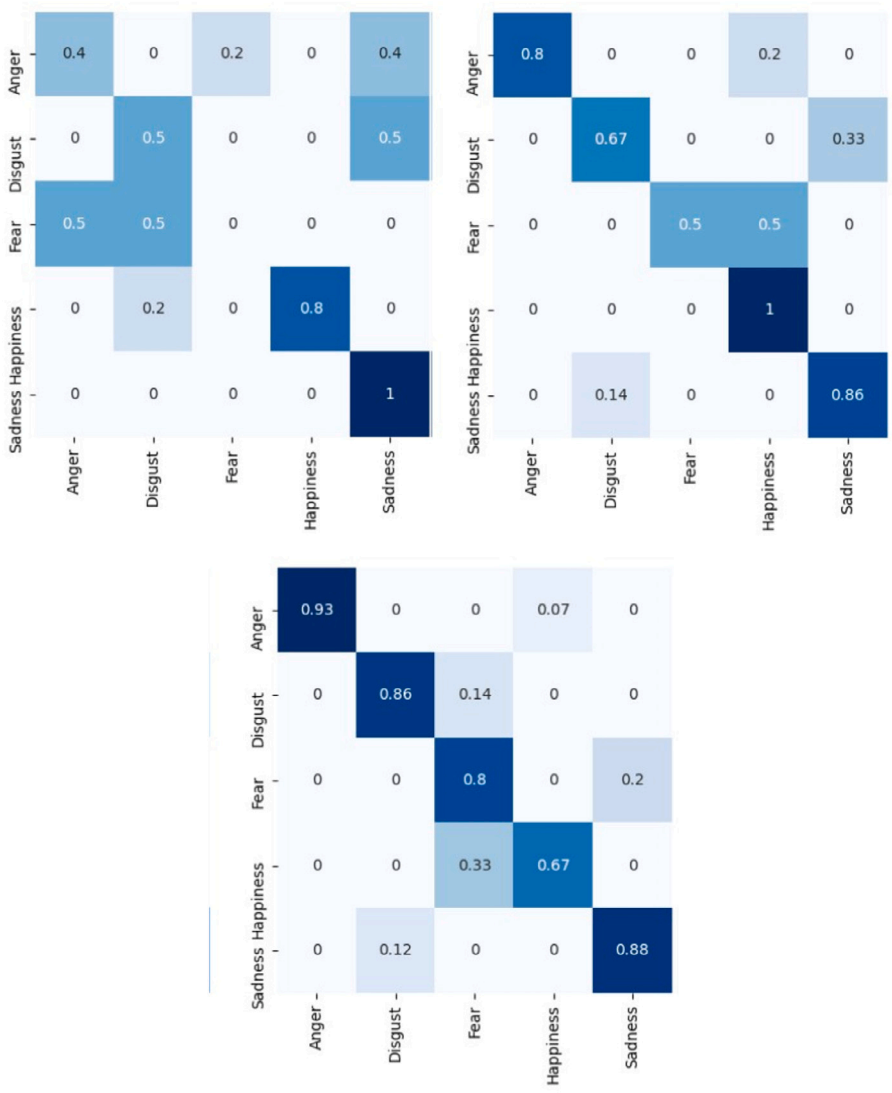


Fig. 11. From left to right and from up to down: the confusion matrices of the initial model, the model after change three, and the model with the augmented dataset with two alterations, respectively.

contrast to BAUM-1, the RAVDESS dataset is structured differently; each video is labeled with a set of codes that identify the individual, the emotion being expressed, the strength of the emotion, and the number of repetitions. It was essential to modify the pre-processing method in order to process the movies and extract the frames and Mel spectrograms as above. The retrieved pieces will be split into two categories according to the emotion label in the film and the strength of the emotions: normal or strong. Table 7 displays the findings for movies with regular intensity, whereas Table 8 displays the results for videos with strong intensity. What is evident and could be expected is that, in comparison to the strong intensity, which is comparable to acted emotion intensity, the outcomes over the normal intensity, which is comparable to spontaneous expressions, are consistently lower. As with BAUM-1 the 2-input model seems to work better, with respect to the 3-input model, proving again that the optical flow frames have a lower discriminative power with respect to the Mel spectrograms and the normal facial frames. Something to notice, that was not expected, is the fact that the results obtained using the video frames only as input have very good performances as the 2-input and the 3-input models. This could be explained by the fact that RAVDESS outperforms BAUM-1 both in terms of a higher number of samples and single sample quality.

Table 7
The comparison of results with different 2-input models and a 3-input model on RAVDESS — normal data. Opt stands for Optical flow frames, F stands for frames and Mel stand for Mel spectrograms.

Type of input \Epochs	10	20	30	40
Frames	**0.8922**	**0.91**	**0.926**	**9377**
Opt	0.6642	0.6739	0.6851	0.6953
Mel	0.6642	0.8375	0.833	0.8623
2-input : F & Mel	0.7914	0.8378	0.8468	0.9014
3-input	0.7955	0.8934	0.9146	0,9127

8. Comparison with the state-of-the-art

For comparing these results with those reported in Luna-Jiménez et al. (2021), we consider our results with the 2-input model using video frames and Mel spectrograms, because the same type of inputs are used there. There are many differences between this work and the work on paper Luna-Jiménez et al. (2021) to keep in consideration. First of all in the amount of labels considered, five in our case against the eight in the paper. Also, the fusion approaches are different, namely late fusion versus early fusion in our proposal. Another difference regards

Table 8
The comparison of results with different 2-input models and a 3-input model on RAVDESS — strong data. Opt stands for Optical flow frames, F for frames and Mel for Mel spectrograms.

Type of input \Epochs	10	20	30	40
F	0.933	0.9285	0.9316	0.9529
Opt	0.7038	0.7064	0.7021	0.7011
Mel	0.7423	0.9086	0.9306	0.9439
2-input : F & Mel	**0.9205**	**0.9344**	**0.952**	**0.9595**
3-input	0.9	0.9289	09 366	0.9359

Table 9
Comparison on BAUM-1.

Methods	Year	Accuracy
Zhalehpour et al. (2017)	2017	51.29
Zhang et al. (2019)	2018	54.57
Cornejo and Pedrini (2019)	2019	56.01
Ma et al. (2020)	2020	67.59
Guanghui and Xiaoping (2021)	2021	71.26
Kansizoglou et al. (2022)	2022	56.01
Hina et al. (2022)	2022	61.68
Our work	**2023**	**95.00**

Table 10
Comparison on RAVDESS.

Methods	Year	Accuracy
Ghaleb et al. (2019)	2019	79.00
Su et al. (2020a)	2020	74.86
Luna-Jiménez et al. (2021)	2021	80.08
Radoi et al. (2021)	2021	78.70
Mocanu and Tapu (2022)	2022	87.89
Middya et al. (2022)	2022	86.00
Our work	**2023**	**95.95**

the way the tests are performed, because the paper used a subject-wise 5-cross validation, while in this work the test data are always sampled randomly, meaning that with a high probability the model sees at least one sample for each subject in the training phase. The last difference worth mentioning is that typically the architecture of Luna-Jiménez et al. (2021) is tested with a model that trains up to 500 epochs, while in this work the highest number of epochs considered is 40. The experiments in Luna-Jiménez et al. (2021) produce several results, one for each different tested approach, and the best accuracy result obtained by them is $80,\%$ on the RAVDESS dataset, while in this work the best accuracy result is $95,95\%$ on the same dataset. To better compare the two approaches, a new experiment was done by re-training a 2-input model, with frames and Mel spectrograms, but this time, by using all the eight labels provided by RAVDESS's normal intensity videos. The proposed approach is still performs better, even if the results are a bit lower than before. In more detail, we achieve 0.8424, 0.8787, 0.8898, and 0.9162 accuracy with 10, 0, 30, and 40 epochs respectively

We further report the comparison with other works based on the same benchmarks. As highlighted in the previous sections, multi-modal fusion methods differ in both structural components and fusion strategy. Another distinction is the manner in which the tests are conducted; in contrast to previous efforts, our proposal always uses a random sampling of the test data, ensuring that the model in the training phase sees at least one sample for each individual. The last distinction worth noticing is that this work only takes into account 40 epochs, whereas the other designs are normally examined with a model that trains up to 100 epochs. Given the many variables, it is therefore not easy to compare the proposed solution with others in the literature on these characteristics. As already done in the study for the optimal multimodal approach, the main measure of the quality of the solutions identified on the two datasets will be the achieved accuracy (see Tables 9 and 10).

9. Conclusions and future work

This study introduces a Convolution Neural Network (CNN) for emotion recognition by leveraging multimodal inputs extracted from video sequences. The BAUM-1 and RAVDESS datasets are employed, both containing videos where subjects react to stimuli while facing the camera. A less complex but more versatile version of the VGG-16 method was used as a basis for evaluating different combinations of data of different types through an early-fusion strategy. Interestingly enough, the same architecture was effective for both video and audio modeling. The experiments developed agree with many of the works in the literature that propose weighted video frame segmentation and MEL Spectrogram as the best representation of audio files for this class of problems. However, unlike the experiments on optical flow, the considerations made on the dynamic nature of emotional phenomena are still not effectively and completely reflected in the designed architectures. To summarize the results presented in this paper, the experiments on the three inputs achieved 0.9359 of accuracy on RAVDNESS strong data and 0.9127 on RAVDNESS normal data. On the other hand, the best result achieved in general is 0.9377 on RAVDNESS normal data by frames and 0.9595 on RAVDNESS strong data with the 2-input model (frames and Mel spectrograms). These considerations lead us to think future developments will attempt to find a more effective method of analyzing the optical flow, one that is physically more representative, and a structure capable of channeling a greater number of data modalities. Another problem to tackle in future work is the cross-dataset performance of the trained models. Some preliminary experiments demonstrated a dramatic drop in performance when training on one dataset and testing on the other one. This may be caused by different signal quality and also by different capture protocols. However, this is the most realistic setup to evaluate systems aiming at real-world applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Ahmed, N., Al Aghbari, Z., Girija, S., 2023. A systematic survey on multimodal emotion recognition using learning algorithms. Intell. Syst. Appl. 17, 200171.

Alshamsi, H., Kepuska, V., Alshamsi, H., Meng, H., 2018. Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference. IEMCON, pp. 730–738. http://dx.doi.org/10.1109/IEMCON.2018.8614831.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., Schuller, B., 2017. Snore sound classification using image-based deep spectrum features. In: INTERSPEECH 2017. pp. 3512–3516.

Baziotis, C., Athanasiou, N., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S., Potamianos, A., 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. arXiv preprint arXiv:1804.06658.

Bisogni, C., Cimmino, L., De Marsico, M., Hao, F., Narducci, F., 2023. Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models. Image Vis. Comput. 104724.

Canal, F.Z., Müller, T.R., Matias, J.C., Scotton, G.G., de Sa Junior, A.R., Pozzebon, E., Sobieranski, A.C., 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. Inform. Sci. 582, 593–617.

Clavel, C., Callejas, Z., 2016. Sentiment analysis: From opinion mining to human-agent interaction. IEEE Trans. Affect. Comput. 7 (1), 74–93. http://dx.doi.org/10.1109/TAFFC.2015.2444846.

Cornejo, J.Y.R., Pedrini, H., 2019. Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform. In: 2019 IEEE International Conference on Systems, Man and Cybernetics. SMC, IEEE, pp. 3396–3402.

Domínguez Bajo, M., Farrús, M., Wanner, L., 2016. An automatic prosody tagger for spontaneous speech. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016 Dec 11–17, Osaka, Japan.[Unknonw Place].

Ekman, P., et al., 1999. Basic emotions. Handb. Cogn. Emot. 98 (45–60), 16.

Feldman Barrett, L., Russell, J.A., 1998. Independence and bipolarity in the structure of current affect. J. Pers. Soc. Psychol. 74 (4), 967.

Ghaleb, E., Popa, M., Asteriadis, S., 2019. Multimodal and temporal perception of audio-visual cues for emotion recognition. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction. ACII, IEEE, pp. 552–558.

Gross, J.J., Levenson, R.W., 1995. Emotion elicitation using films. Cogn. Emot. 9 (1), 87–108.

Guanghui, C., Xiaoping, Z., 2021. Multi-modal emotion recognition by fusing correlation features of speech-visual. IEEE Signal Process. Lett. 28, 533–537.

Guo, Y., Tian, Y., Gao, X., Zhang, X., 2014. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In: 2014 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 3473–3479.

Hina, I., Shaukat, A., Akram, M.U., 2022. Multimodal emotion recognition using deep learning architectures. In: 2022 2nd International Conference on Digital Futures and Transformative Technologies. ICoDT2, IEEE, pp. 1–6.

Hsu, J.-H., Wu, C.-H., 2020. Attentively-coupled long short-term memory for audio-visual emotion recognition. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. APSIPA ASC, IEEE, pp. 1048–1053.

Huang, K.-Y., Wu, C.-H., Hong, Q.-B., Su, M.-H., Chen, Y.-H., 2019. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 5866–5870. http://dx.doi.org/10.1109/ICASSP.2019.8682283.

Izard, C.E., 1994. Innate and Universal Facial Expressions: Evidence from Developmental and Cross-Cultural Research. American Psychological Association.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28.

Jiang, Y., Li, W., Hossain, M.S., Chen, M., Alelaiwi, A., Al-Hammadi, M., 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Inf. Fusion 53, 209–221.

Jiang, B., Valstar, M., Martinez, B., Pantic, M., 2014. A dynamic appearance descriptor approach to facial actions temporal modeling. IEEE Trans. Cybern. 44 (2), 161–174. http://dx.doi.org/10.1109/TCYB.2013.2249063.

Kansizoglou, I., Bampis, L., Gasteratos, A., 2022. An active learning paradigm for online audio-visual emotion recognition. IEEE Trans. Affect. Comput. 13 (2), 756–768. http://dx.doi.org/10.1109/TAFFC.2019.2961089.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 2880–2894.

Livingstone, S.R., Russo, F.A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One 13 (5), e0196391.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence - Vol. 2. IJCAI '81, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 674–679.

Luengo, I., Navas, E., Hernáez, I., Sánchez, J., 2005. Automatic emotion recognition using prosodic parameters. In: Proc. Interspeech 2005. pp. 493–496. http://dx.doi.org/10.21437/Interspeech.2005-324.

Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F., 2021. Multimodal emotion recognition on RAVDESS dataset using transfer learning. Sensors 21 (22), 7665.

Ma, F., Zhang, W., Li, Y., Huang, S.-L., Zhang, L., 2020. Learning better representations for audio-visual emotion recognition with common information. Appl. Sci. 10 (20), 7239.

Mamieva, D., Bobomirzaevich, A., Kutlimuratov, A., Muminov, B., Whangbo, T.K., 2023. Multimodal emotion detection via attention-based fusion of extracted facial and speech features. Sensors 23, 5475. http://dx.doi.org/10.3390/s23125475.

Middya, A.I., Nag, B., Roy, S., 2022. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowl.-Based Syst. 244, 108580.

Mocanu, B., Tapu, R., 2022. Audio-video fusion with double attention for multimodal emotion recognition. In: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop. IVMSP, pp. 1–5. http://dx.doi.org/10.1109/IVMSP54334.2022.9816349.

Mollahosseini, A., Hasani, B., Mahoor, M.H., 2019. AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. 10 (1), 18–31. http://dx.doi.org/10.1109/TAFFC.2017.2740923.

Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009. Factors in the recognition of vocally expressed emotions: A comparison of four languages. J. Phonetics 37 (4), 417–435. http://dx.doi.org/10.1016/j.wocn.2009.07.005, URL https://www.sciencedirect.com/science/article/pii/S0095447009000448.

Poria, S., Cambria, E., Bajpai, R., Hussain, A., 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Inf. Fusion 37, 98–125.

Posner, J., Russell, J.A., Peterson, B.S., 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. 17 (3), 715–734. http://dx.doi.org/10.1017/S0954579405050340.

Radoi, A., Birhala, A., Ristea, N.-C., Dutu, L.-C., 2021. An end-to-end emotion recognition framework based on temporal aggregation of multimodal information. IEEE Access 9, 135559–135570.

Rajoo, R., Aun, C.C., 2016. Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages. In: 2016 IEEE Symposium on Computer Applications & Industrial Electronics. ISCAIE, pp. 35–39. http://dx.doi.org/10.1109/ISCAIE.2016.7575033.

Ramachandram, D., Taylor, G.W., 2017. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Process. Mag. 34 (6), 96–108.

Russell, J.A., 1980. A circumplex model of affect. J. Pers. Soc. Psychol. 39 (6), 1161.

Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S., 2007. Authentic facial expression analysis. Image Vis. Comput. 25 (12), 1856–1863.

Siddiqui, M.F.H., Javaid, A.Y., 2020. A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. Multimod. Technol. Interact. 4 (3), 46.

Stahelski, A., Anderson, A., Browitt, N., Radeke, M., 2021. Facial expressions and emotion labels are separate initiators of trait inferences from the face. Front. Psychol. 12, 749933.

Su, L., Hu, C., Li, G., Cao, D., 2020a. Msaf: Multimodal split attention fusion. arXiv preprint arXiv:2012.07175.

Su, M.-H., Wu, C.-H., Huang, K.-Y., Yang, T.-H., 2020b. Cell-coupled long short-term memory with L -skip fusion mechanism for mood disorder detection through elicited audiovisual features. IEEE Trans. Neural Netw. Learn. Syst. 31 (1), 124–135. http://dx.doi.org/10.1109/TNNLS.2019.2899884.

Tomkins, S., 1962. Affect Imagery Consciousness: Volume I: The Positive Affects. Springer publishing company.

Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E., 2017. BAUM-1: A spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. 8 (3), 300–313. http://dx.doi.org/10.1109/TAFFC.2016.2553038.

Zhang, S., Pan, X., Cui, Y., Zhao, X., Liu, L., 2019. Learning affective video features for facial expression recognition via hybrid deep learning. IEEE Access 7, 32297–32304.

Zhang, S., Tao, X., Chuang, Y., Zhao, X., 2021. Learning deep multimodal affective features for spontaneous speech emotion recognition. Speech Commun. 127, 73–81.