Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 2556–2566

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

# Bimodal Emotion Recognition Based on Vocal and Facial Features

Mateusz Wozniak[a,*], Michal Sakowicz[a], Kacper Ledwosinski[a], Jakub Rzepkowski[a], Pawel Czapla[a] and Szymon Zaporowski[a,b]

[a]*Multimedia Systems Department, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland*
[b] *Audio Acoustic Laboratory, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland*

## Abstract

Emotion recognition is a crucial aspect of human communication, with applications in fields such as psychology, education, and healthcare. Identifying emotions accurately is challenging, as people use a variety of signals to express and perceive emotions. In this study, we address the problem of multimodal emotion recognition using both audio and video signals, to develop a robust and reliable system that can recognize emotions even when one modality is absent. To achieve this goal, we propose a novel architecture based on well-designed feature extractors for each modality and use model-level fusion based on a TFusion block to combine the information from both sources. To be more efficient in real-world scenarios, we trained our model on a compound dataset consisting of RAVDESS, RML, and eNTERFACE'05. It is then evaluated and compared to the state-of-the-art models. We find that our approach performs close to the modern solutions and can recognize emotions accurately when one of the modalities is missing. Additionally, we have developed a real-time emotion recognition application as a part of this work.

*Keywords:* Emotion recognition; TFusion; Deep Learning; Affective Computing;

## 1. Introduction

Human emotions play a vital role in our society and have a high impact on the way we communicate. Multimodal emotion recognition is the process of identifying and interpreting emotions from multiple sources, such as text, speech, facial expressions, and body language. The proposed technology could be used in a variety of applications,

[*] Corresponding author.
*E-mail address:* s175610@student.pg.edu.pl

10.1016/j.procs.2023.10.247

including human-computer interaction, virtual assistants, and video surveillance. The goal of multimodal emotion recognition is to provide a more accurate and comprehensive understanding of a person's emotional state, as emotions can be conveyed through multiple channels. Advances in machine learning and computer vision have made it possible to develop sophisticated algorithms that can analyze and interpret multiple modalities simultaneously [2] [3] [10]. This approach will be more extensively described in Section 1.

In [1] authors suggest that the brain combines information received from different sources, such as vision and sound, through a process called multimodal integration. This phenomenon is named the McGurk effect [9], where audio signals can be affected by conflicting visual information. This type of integration is important for perception in challenging conditions, such as a noisy environment. Moreover, studies have found that speech and facial expressions can work together in tasks related to recognizing emotions and identifying speech [1]. Additionally, previous studies have shown that using only one modality for emotion recognition is not effective [11] [4]. Using multiple modalities, such as audio, video, and text, presents better results in emotion recognition.

Referring to the aforementioned issues, an innovative network architecture was proposed that implements separate audio and video branches. The results from these independent paths are fused using the TFusion method [6]. The developed model is described in more detail in Section 3.

As a part of our study, 7 different emotions were implemented that are expressed most often: calm, happiness, sadness, anger, fear, disgust, and surprise. To train our algorithm, a diverse dataset was created that consists of three databases: RAVDESS, RML, and eNTERAFACE'05. Section 4 contains detailed descriptions of each subset of the data and the results of the experiment.

The final stage of the project was to create an application that allows for the detection of emotions. Using our model, a multi-threaded application was created that returns predictions for 3.6-second time windows. A detailed description of the structure and operation of our application can be found in Section 5.

## 2. Related works

The challenges of Emotion Recognition related tasks have gained popularity among Machine Learning Scientists, proposing new methods to tackle the problem [16] [15]. In our work we specifically attempt to address the issue of multi-modal emotion recognition, emphasizing implementing an efficient fusion method. Multimodal fusion for emotion recognition refers to the collection of techniques that utilize information from various sources to predict emotions [2] [6] [24]. Modalities, implemented in the mentioned methods, include images, speech, text, or biosignals.

Several papers that review previous research examine the various techniques used for recognizing emotions using multiple modalities. Besides choosing the modalities it is also crucial to determine which features are the most descriptive and best suited for the algorithm. The descriptors for face images are in most cases geometric face features, whereas there are multiple approaches implemented in literature such as spectral, cepstral, and prosodic features e.g., MFCC, Tonnetz, Mel spectrogram, pitch, energy, and many more [2] [3] [24]. These features may be implemented independently in the model or concatenated forming a single vector containing audio information.

In the field of machine learning using multiple modalities, typically three types of fusion approaches are identified: early fusion, where the data from different modalities are combined and processed together, with the features being integrated immediately after the extraction [24]; late fusion, which involves the fusion of modalities after a distinct model has made an independent prediction for each one; intermediate feature fusion [10], which concatenates the input feature representations and then passes them through a model that computes a learned, internal representation before making its prediction [12].

Most of the fusion approaches take advantage of the Transformer architecture [2] [6] or simply multi-head attention [17]. When combining information from two modalities, such as audio and visual, self-attention can be used as a fusion approach by calculating queries from one modality and keys and values from the other. This leads to a representation learned from one modality attending to the corresponding features from the other modality, and then applying the resulting attention matrix to the representation learned from the second modality [2].

Additionally, the aim of creating an optimal method of fusion is that it is able to work in situations where one modality is missing. This issue was specifically addressed in the articles [2] and [6]. They portray entirely different approaches to the problem, where [2] requires training the model to handle modalities that occasionally drop out, being replaced by zero-padding, whereas [6] does not need zero-padding by introducing a modal attention mechanism that builds a shared representation. The approach proposed by the authors of [24] is based on a completely different idea, where the correlation between modalities is learned in a self-supervised pre-training stage utilizing contrastive loss. However, the method was not evaluated in the scenario of a missing modality, additionally resulting in minor accuracy than other state-of-the-art methods in favour of less complicated fusion variants.

## 3. Proposed method

This section describes the proposed architecture for recognizing emotions using a multimodal approach. The architecture consists of two parts, each responsible for extracting features from a different modality (video and audio). The section also introduces the TFusion fusion module, which is implemented for the first time in the context of multimodal emotion recognition. The overall architecture is shown in Figure 1.

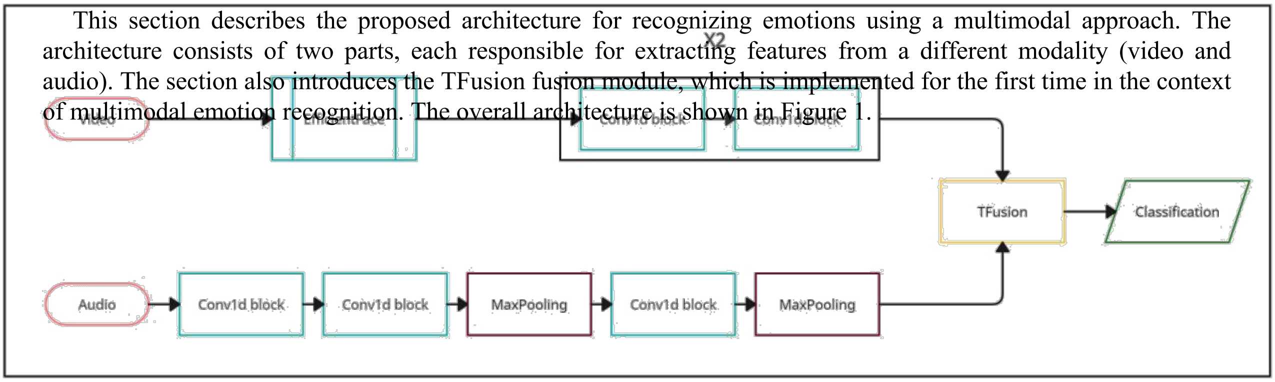


Fig. 1. Architecture of the proposed model

### *3.1 Data preprocessing*

Video: For video data, recordings were separated into 3.6-second samples, from which 15 frames were selected. These frames were then processed using the MTCNN model for face detection, and face images were cropped and resized to 224x224 pixels.

Audio: Many features were extracted from the audio recording corresponding to the video clip, including MFCC, Mel spectrogram, spectral contrast, and Tonnetz as suggested by [10]. The resulting feature tensor had a size of 156x181.

### *3.2 Architecture*

Video Branch: The first part is responsible for extracting features from a temporal series of facial images. The preprocessed footage is fed as an input to a pre-trained model called EfficientFace. This model is trained to recognize emotions from a single image and its facial features. Further details of the EfficientFace architecture can be found in [19]. In order to use it for video processing, we skipped the classification layer and used its ability to extract a feature vector describing a given face for each of the 15 frames. This approach is effective because the model was trained on a large dataset and is capable of generalizing new data. The extracted facial features are then concatenated and represented in the temporal dimension. This allows us to use 1D convolution operations, which have several advantages over other video processing techniques such as 3D convolution or ConvLSTM2D. The most important advantage is the reduction in computational complexity, which leads to shorter training times and allows the model to run in real-time. The determined descriptors are then fed as an input to a module called Conv1Dblock.

Each block consists of a 1D convolution layer with a kernel size of 3x3, a BatchNormalization layer, and a ReLU activation function. In the video path, there are four blocks that each have a different number of filters. Specifically, the first two blocks contain 64 filters, while the following two contain 128. This approach of using 1D convolution layers allows for capturing the temporal dynamics of the facial features while keeping the model computationally efficient.

Audio Branch: This section describes the proposed architecture for recognizing emotions using audio features. The designated audio features are processed by a branch consisting of blocks similar to those used in the video branch, with a few key differences. The blocks in the audio branch use a filter size of 5x5 and do not include padding, to effectively extract features from the audio data. Additionally, MaxPooling operations are utilized to create a downsampled feature map, which helps to reduce computational complexity and improve the model's performance. The first two blocks in the audio branch have 64 filters, followed by a layer that reduces the size of the feature vector. The next block uses 128 filters and is finished with a MaxPooling layer to further downsample the feature map. The final step in the audio branch is a Dropout layer with a parameter p=0.4, which helps to prevent overfitting and improve the model's generalization capabilities. To ensure that the feature vector has the same number of output channels as the video branch feature vector, AdaptiveAvgPool1d was used. Overall, this approach effectively extracts relevant features from the audio data and prepares them for fusion with the features extracted from the video data. Details are included in Table 1.

Table 1. Architectural details of the proposed model

| | Video Branch |
|---|---|
| 1. | EfficientFace |
| 2. | Conv1d(out_channels=64, kernel_size=3, padding=”same”) + BatchNorm + ReLU |
| 3. | Conv1d(out_channels=64, kernel_size=3, padding=”same”) + BatchNorm + ReLU |
| 4. | Conv1d(out_channels=128, kernel_size=3, padding=”same”) + BatchNorm + ReLU |
| 5. | Conv1d(out_channels=128, kernel_size=3, padding=”same”) + BatchNorm + ReLU |

| | Audio Branch |
|---|---|
| 1. | Conv1d(out_channels=64, kernel_size=5, padding=”valid”) + BatchNorm + ReLU |
| 2. | Conv1d(out_channels=64, kernel_size=5, padding=”valid”) + BatchNorm + ReLU |
| 3. | MaxPool1d(kernel_size=2, stride=1) |
| 4. | Conv1d(out_channels=128, kernel_size=5, padding=”valid”) + BatchNorm + ReLU |
| 5. | MaxPool1d(kernel_size=2, stride=1) |
| 6. | Dropout(0.4) |
| 7. | AdaptiveAvgPool1d(15) |

| | Classification |
|---|---|
| 1. | Linear(128) + ReLU |
| 2. | Linear(7) |

Fusion: The feature representations extracted from the video and audio branches are fused using the TFusion [6] module. To our knowledge, we are the first to use it to solve the problem of bimodal emotion recognition. The method is based on the transformer architecture using the attention mechanism, allowing it to effectively combine the information from both modalities even in the absence of data from one of them. The primary advantage of TFusion is that it introduces an efficient way to learn the correlation between modalities and is easy to implement for different types of multimodal classification problems. In contrast to predetermined formulas [24] or methods based on convolution, TFusion handles missing modalities without the need to synthesize or pad them with zeros [2], which may introduce undesirable bias. This approach is illustrated in Figure 2. The feature representations are projected as tokens and fed into transformer layers to generate latent multimodal correlations. An attention mechanism is then applied to reduce the impact of a particular modality and create a shared representation for both modalities. The final step of the proposed architecture is the classification module, which consists of dense layers and a last layer with 7 neurons for recognizing the 7 emotions.

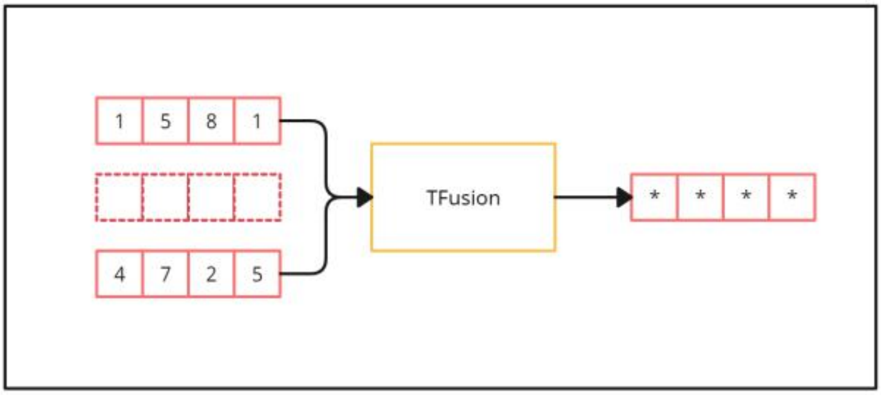


Fig. 2. TFusion strategy for handling modality missing data

## 4. Experiments and results

### *4.1 Datasets*

Three datasets were selected in order to create a diverse training set. The datasets consist of video and audio signals and include basic emotions.

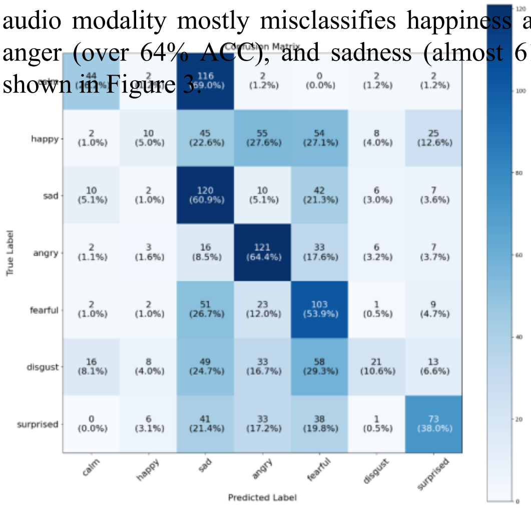
RAVDESS dataset: The RAVDESS database [7] is an extensive collection (7356 clips) of speech and songs, collected from 24 actors, with a North American accent. Emotions are classified into 8 categories - calm, happiness, sadness, anger, fear, surprise, disgust, and neutral. In this research, solely the speech part of the database was used. Neutral expression was excluded due to the fact that it is considered to be an absence of emotion rather than a particular one itself. Neither in Russel's circumplex model, [14] nor in Plutchnik's wheel of emotions [13] neutral category is included. Calmness has been populated with augmented records in order to compensate for the absence of emotion in other databases.

RML dataset: The RML database [5] contains 720 audio-visual expression samples, where 6 principal emotions are categorized: anger, disgust, fear, happiness, sadness, and surprise. The recordings' length varies from 3 to 6 seconds. However, as mentioned in Section III, they were shortened to 3.6 s while maintaining the main expression part.

eNTERAFACE'05 dataset: The eNTERFACE'05 database [8] consists of a total number of 1166 audio-video sequences, of which 23% are women and 73% men. The database considers emotions such as anger, disgust, fear, happiness, sadness, and surprise. What is crucial, the subjects were not actors - they do not have previous experience with professional acting. They were mostly the research engineers, which caused not an excessively exaggerated performance in expressing emotions.

### *4.2 Results*

Results obtained during this study have shown that the multimodal variation of the system provides the best score. Table 2 shows metrics for the multimodal approach, as well as the circumstances when video or audio modality is excluded. As it was apparent beforehand, the audio results were far worse than the ones collected from the video modality. It is directly caused by difficulty in distinguishing between emotions based merely on audio signals. The audio modality mostly misclassifies happiness and disgust, while having the highest performance in categorizing anger (over 64% ACC), and sadness (almost 61% ACC). The confusion matrix (CM) for the audio modality is shown in Figure 3.



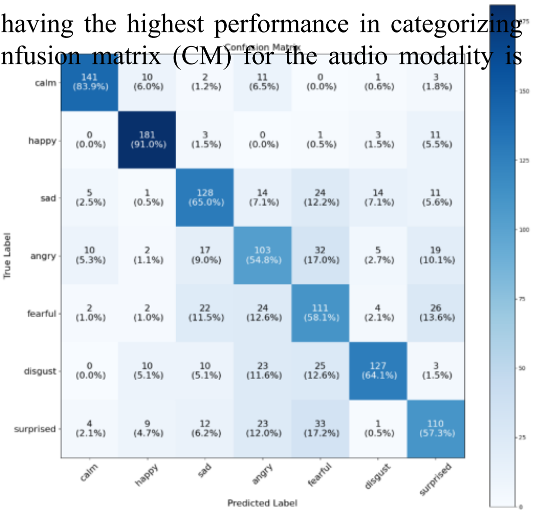


Fig. 3. Confusion matrix for the audio modality

Fig. 4. Confusion matrix for the video modality

On the other hand, the video modality achieves the best results when it classifies happiness (upwards of 90% ACC), and calm (83.9% ACC). CM for the video modality is shown in Figure 4. However, as can be observed in Figure 5, joining the video and audio modalities has a beneficial impact on the outcome, which results in as many as 77.56% ACC for the 30% of the joint database's records.

Naturally, results vary by database. For example, the RAVDESS database accuracy score, obtained during a multi-modal system test, equals a noticeable 82,99%, whereas RML accuracy is 82.47% and eNTERFACE'05 ACC is equal to 72.27%. The eNTERFACE'05 database results are the lowest, due to more ordinary subjects' acting.

Table 2. Performance (%) of different modalities on joint datasets

| | Accuracy | Top 3 Accuracy | Precision | Recall |
|---|---|---|---|---|
| **Bimodal** | **77.56** | **96.02** | **78.12** | **77.56** |
| Video | 67.74 | 92.27 | 68.63 | 67.74 |
| Audio | 37.01 | 77.64 | 41.57 | 37.01 |

Also, the comparison results of the proposed multimodal emotion recognition system with the previous methods are given in Table 3. None of the authors used the same combination of datasets as we did. However, in all of the articles, the proposed system was tested on at least one commonly used dataset. Therefore, the comparison of the 3 datasets is presented separately. The number of emotions recognized varies depending on the dataset and used architecture. In the third column, emotions are denoted by abbreviations, meaning: A-anger, C-calm, D-disgust, F-fear, H-happiness, N-neutral, Sa-sadness, Su-surprise. As can be seen, our model outperforms different models on the RML dataset and has a very promising performance on the others. Additionally, it is worth noting that the findings reported in [20] demonstrate high performance on the eNTERFACE’05 dataset. However, on the BAUM-1 database [21], the accuracy is recorded at 59.17%. Moreover, the comparison of results from the vast majority of given articles is not entirely straightforward when only a single accuracy metric is employed.

Table 3. Performance of the proposed models in recent years

| Author | Accuracy bimodal (%) | Emotions | Dataset |
|---|---|---|---|
| E. Ghaleb et al. [3] | 67.7 | A,C,D,F,H,N,Sa,Su | RAVDESS |
| K. Chumachenko et al. [2] | 81.58 | A,C,D,F,H,Sa,Su | RAVDESS |
| **Ours** | **82.99** | **A,C,D,F,H,Sa,Su** | **RAVDESS** |
| A.I. Middya et al. [10] | 86.0 | A,C,D,F,H,N,Sa,Su | RAVDESS |
| S. Zhang et al. [18] | 54.57 | A,D,F,H,Sa,Su | eNTERFACE05 |
| **Ours** | **72.27** | **A,D,F,H,Sa,Su** | **eNTERFACE05** |
| Y. Ma et al. [20] | 85.69 | A,D,F,H,Sa,Su | eNTERFACE05 |
| E. Avots et al. [22] | 69.3 | A,D,F,H,Sa,Su | RML |
| C. Fadil et al. [23] | 79.72 | A,D,F,H,Sa,Su | RML |
| **Ours** | **82.47** | **A,D,F,H,Sa,Su** | **RML** |

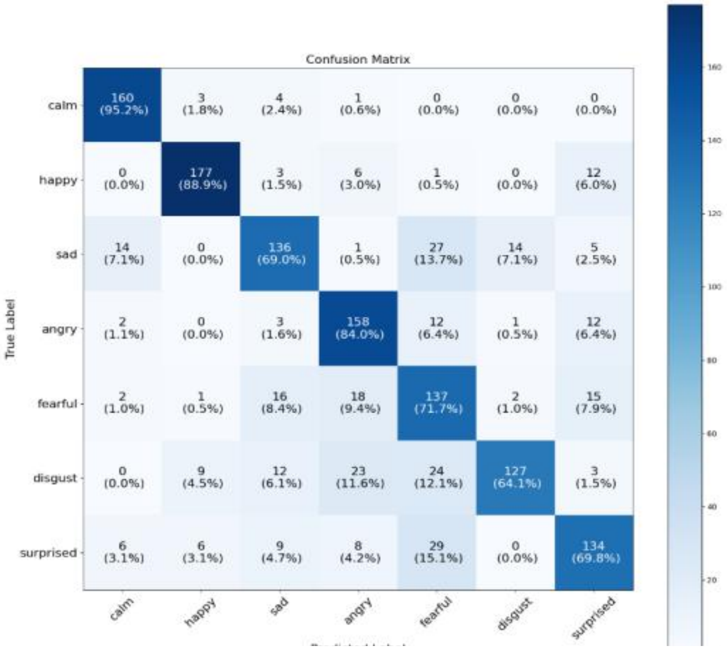


Fig. 5. Confusion matrix for the joint modalities

### *4.3 Experiments*

All calculations were carried out on an NVIDIA A40 graphics card, which is based on the NVIDIA Ampere architecture. The card is equipped with 48GB of GDDR6 memory and delivers 37.4 teraflops of computing power (FP32).

In the conducted experiments, aside from the best architectural design, we focused on optimizing the hyperparameters of our model. For this task, we used the weight and biases platform [25], which provided us with tools to automatically track and document the results of each iteration of the experiment. We used its visualization and analysis features to fully understand the impact of various hyperparameters on the results of our models.

Our attention was focused on selecting the optimal optimizer, finding the best batch size and the finest learning rate value. In the very first iterations, we found that the optimal optimizers for our model were Adam and SGD. The difference in the impact on model evaluations between the two optimizers was relevantly small. Higher learning rate values signaled an overfitting model. Choosing a small number of samples in the batch also influenced overfitting.

Subsequent iterations allowed us to find adequate values for the rest of the hyperparameters we were looking for. The final model was trained for the following hyperparameters: optimizer Adam, 32 samples in every batch and a learning rate value set to 0.0003.

### *4.4 Discussion*

In comparison to the given articles, we achieved the highest accuracy on the RML dataset and high accuracies on the remaining databases. Further, it should be emphasized that utilization of the TFusion module does not force an additional data modification - there is no need to upsample or replace signal from one modality with zeros or a particular scaling factor. The module itself completes the task. The implementation of the TFusion method allowed us to reach a prediction accuracy that can compete with other state-of-the-art methods, with the advantage of smoothly handling missing modalities.

## 5. Application

Beyond artificial intelligence, one of the key aspects of this project was the application use. The model alone is merely a scientific tool for processing video data. Our research has shown that scientific papers regarding emotion recognition, presented solely the results of evaluating gathered data. None of the papers implemented the main functionality of recognizing emotion with an audio-video signal in the form of an application. Thus, a model for fast evaluation was prepared to process emotions in real-time with the use of a typical PC. A schematic map of the application is shown in Figure 6.

The application fulfils the multi-platform requirement, being an application suitable for most operating systems. The Python programming language is perfect for that solution, especially with the TKinter library. TK provides us with a fast and reliable graphical interface, which can be developed with ease.

Reliability was one of the key values of our work, so the application was designed to be multi-threaded. This allowed providing live audio and video signals without experiencing any delay or latency. The main thread consists of a TK graphic user interface operation, which provides the main window with a camera preview and bar plot illustrating the detected emotion. Intuitively camera preview gives the user which face is now processed. This information is retrieved from the VideoSource thread. Emotion bar plot data is displayed as soon as the model finishes the classification.

The Model thread is constantly retrieving data captured from both Video and Audio Source threads which are consistently capturing data with a microphone and camera sensor connected to the PC. The audio signal is gathered with the PyAudio library in a specific size queue.

The application implements a functionality, that in case of the absence of video or audio signal, the model does not process the missing modality data. For instance, if a face was not detected, only the audio signal is fed into the network and the TFusion module handles the prediction using only one modality. Figure 7 depicts an example of emotion detection using video modality.

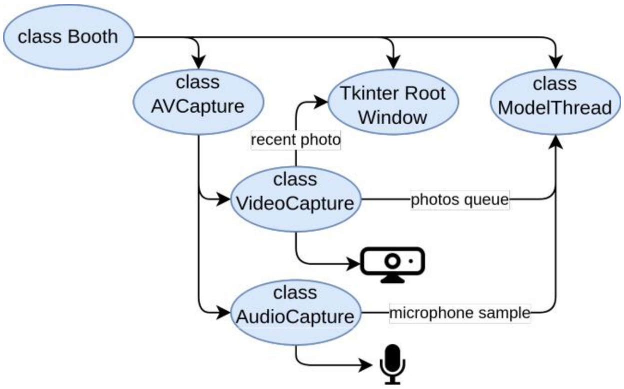
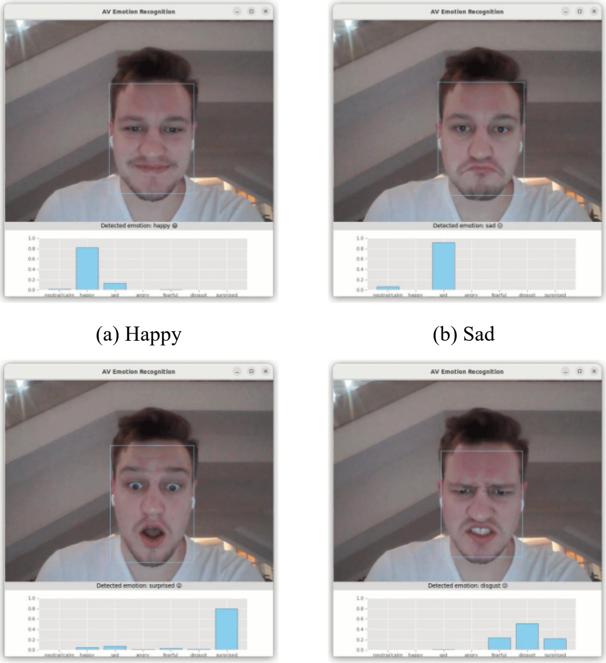


Fig. 6. Application threads schematic map



(a) Happy

(b) Sad

(c) Surprised

(d) Disgusted

Fig. 7. Application demonstration of person detection

## 6. Conclusion

In this paper, we presented a novel approach to bimodal emotion recognition using both audio and video signals. Our proposed model, based on the TFusion fusion module, achieved high performance in emotion recognition tasks. Furthermore, we demonstrated the potential of the model for real-world applications by developing a corresponding application using the trained model. Future research could focus on further developing the audio processing branch and exploring the integration of additional modalities, such as text or EEG signals, into the model's architecture using the capabilities of the TFusion module.

## Acknowledgements

This work is supported by the AI Tech program.

## References

[1] Pashiera Barkhuysen, Emiel Krahmer, and Marc Swerts. Crossmodal and incremental perception of audiovisual cues to emotional speech. Language and Speech, 53:3–30, 2010.

[2] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incom-plete data, 2022.

[3] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. Multimodal and temporal perception of audio-visual cues for emotion recognition. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 552–558, 2019.

[4] Xiaohua Huang, Jukka Kortelainen, Xiaobai Li Guoying Zhao, Antti Moilanen, Tapio Seppanen,¨ and Matti Pietikainen¨. Multi-modal emotion analysis from facial expressions and electroencephalogram. Computer Vision and Image Understanding, 147:14–124, 2016.

[5] Ryerson Multimedia Research Lab. Rml emotion database. http://shachi. org/resources/4965, 2017 (accessed November 02, 2022).

[6] Zecheng Liu, Jia Wei, and Rui Li. Tfusion: Transformer based n-to-one multimodal fusion block, 2022.

[7] Steven R. Livingstone and Frank A. Russo. 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, 13:1–35, 2018.

[8] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannnis Pitas. The enterface' 05 audio-visual emotion database. In 22nd International Conference on Data Engineering Workshops (ICDEW'06), pages 8–8, 2006.

[9] Harry McGurk and John MacDonald. Hearing lips and seeing voices. Nature, 264:746–748, 1976.

[10] Asif Iqbal Middya, Baibhav Nag, and Sarbani Roy. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-Based Systems, 244:108580, 2022.

[11] Michele Mukeshimana, Xiao juan Ban, Nelson Karani, and Ruoyi Liu. Multimodal emotion recognition for human-computer interaction: A survey, 2017.

[12] Bei Pan, Kaoru Hirota, Zhiyang Jia, Linhui Zhao, Xiaoming Jin, and Yaping Dai. Multimodal emotion recognition based on feature selection and extreme learning machine in video clips. Journal of Ambient Intelligence and Humanized Computing, 2021.

[13] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4):344–350, 2001.

[14] James Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 39:1161–1178, 1980.

[15] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognition Letters, 146:1–7, 2021.

[16] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. Recent advances and trends in multimodal deep learning: A review, 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[18] Zhang Shiqing, Zhang Shiliang, Huang Tiejun, Gao Wen, Tian Qi. Learning afective features with a hybrid deep model for audio-visual emotion recognition, 2017.

[19] Zhao, Zengqun, Qingshan Liu and Feng Zhou. Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. AAAI Conference on Artificial Intelligence, 2021.

[20] Yaxiong Ma, Yixue Hao, Min Chen, Jincai Chen, Ping Lum Andrej Košir. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach, 2019.

[21] Zhalehpour Sara, Onder Onur, Akhtar Zahid, Erdem, Cigdem. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States, 2016.

[22] Avots Egils, Sapiński Tomasz, Bachmann Maie, Kamińska Dorota. Audio-Visual Emotion Recognition in Wild, 2019.

[23] Fadil Carim, Alvarez Ramiroa, Martínez Cesar, Goddard John, Rufiner Hugo. Multimodal Emotion Recognition Using Deep Networks, 2014.

[24] Mandeep Singh, Yuan Fang. Emotion Recognition in Audio and Video Using Deep Neural Networks, 2020.

[25] Biewald Lukas. Experiment Tracking with Weights and Biases, 2020.