UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

Multimodal emotion recognition via face and voice

Master Thesis
submitted to the Faculty of the
Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona
Universitat Politècnica de Catalunya
by

Oriol Griera i Jiménez

In partial fulfillment
of the requirements for the master in
TELECOMMUNICATIONS ENGINEERING

Advisor: Maria de Marsico, Francisco Javier Hernando Pericas
Rome and Barcelona, June 2022

Abstract

Recent advances in technology have allowed humans to interact with computers in ways previously unimaginable. Despite significant progress, a necessary element for natural interaction is still lacking: emotions. Emotions play an important role in human communication and interaction, allowing people to express themselves beyond the language domain.

The purpose of this project is to develop a multimodal system to classify emotions using facial expressions and the voice taken from videos. For face emotion recognition, face images and optical flow frames are used to exploit spatial and temporal information of the videos. Regarding the voice, the model uses speech features extracted from the chunked audio signals to predict the emotion.

The combination of the two biometrics with a score-level fusion achieves excellent performance on the RAVDESS and the BAUM-1 datasets. However, the results remark the importance of further investigating the preprocessing techniques applied in this work to “normalize” the datasets to a unified format to improve the cross-dataset performance.

Acknowledgements

First of all, I would like to express my gratitude to Professor Maria de Marsico for providing me with the opportunity to work on this project and for her support and implication during its development. Also, my advisor at UPC, Professor Javier Hernando Pericas, for his eagerness to help.

Moreover, I would like to thank all the people I have met in Rome who have done of this Erasmus an incredible experience. Finally, I would not have been able to get here without the constant support of my family during all my studies. Your support means the world.

Contents

List of Figures

List of Tables

Revision history and approval record

Revision	Date	Purpose
0	09/05/2022	Document creation
1	07/06/2022	Document revision
2	23/06/2022	Document revision
3	02/07/2022	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Oriol Griera Jiménez	oriol.griera@estudiantat.upc.edu
Maria de Marsico	demarsico@di.uniroma1.it
Francisco Javier Hernando Pericas	javier.hernando@upc.edu

Written by:		Reviewed and approved by:	
Date	09/05/2022	Date	02/07/2022
Name	Oriol Griera Jiménez	Name	Maria de Marsico, Francisco Javier Hernando Pericas
Position	Project Author	Position	Project Supervisor

1 Introduction

In the last few decades, interest in investigating the significance of emotional intelligence and its influence on perception and behaviour has increased. It is known that the ability to express and control emotions is crucial, but so is the ability to understand, interpret, and respond to the emotions of others. Researchers suggest that there are four different levels of emotional intelligence including emotional perception, the ability to reason using emotions, the ability to understand emotions, and the ability to manage emotions [1]. This project focuses on the first level of understanding emotions: perceiving them accurately. In most cases, this involves interpreting nonverbal signals such as facial expressions.

Emotion influences practically all modes of human communication, and consequently, it can significantly change the message transmitted. Humans' interaction is mainly based on speech, but also through body gestures to stress a certain part of the speech and display emotions. According to R.W. Picard, affect recognition is most likely to be accurate when multiple modalities are combined [2]. Consequently, the new technologies are driving toward accommodating information exchanges via the combination of natural sensory modes of sound, sight, and touch, to use one to enhance and complement the others. Considering this aspect, B. Reeves and C. Nass believe that the AI research topic most likely to become widespread is multimodal context-sensitive human-computer interaction [3].

The rise in the importance of emotions has not gone unnoticed in the Computer Science field. Even though machines may never need all the emotional skills of humans, it is when interacting with people that they need this ability to appear intelligent. Interacting naturally with the user is required to achieve a truly effective human-computer intelligent interaction (HCII). Computers' functionality could be enhanced by being able to recognize the users' emotions, especially in those applications where they take a social role. For example, in the clinical sector, recognizing people's inability to express certain facial expressions may help to diagnose early psychological disorders. Furthermore, synthetic speech with emotions can sound more pleasing than a monotonous voice. It is also possible to mention an empathic behaviour by a digital agent, that improves the user experience especially in emotionally critical situations [4][5].

The purpose of this project is to develop a multimodal system to classify emotions using facial expressions and voice. The final model infers the emotion expressed by people speaking in different videos. Cropped face images, dense optical flows, and speech features are used to feed three Convolutional Neural Networks (one per feature) whose scores are fused to obtain the predicted emotion label.

This work is the continuation of two previous projects developed by students from La Sapienza university. The theses written by Sara Tramonte and Alessandro Linciano have been used as the starting point for this project. Regarding the software, one script implemented by Sara has been used to filter the peak high-intensity frames to train the classifiers.

The main objective of this project is to improve the cross-dataset results previously obtained. Consequently, the solution adopted is to pre-process the frames and the audios to make them as similar as possible between the multiple datasets used. In addition, the number of samples is increased by applying data augmentation.

To compare the results with those of the other works, the datasets used are RAVDESS and BAUM-1. The final classifier should be able to recognize the emotions that these datasets have in common (happiness, sadness, anger, fearfulness, and disgust) from the video files.

The development of the project (i.e., video preprocessing, data augmentation, design and adjustment of the different models used, as well as the training and testing phase) has been carried out using high-level APIs from the TensorFlow framework. The MTCNN framework [6] has been used to extract the cropped face image from the video frames, whereas the Librosa framework has been employed to obtain the audio features from the speech signal.

1.1 Work Plan

This section contains the four Work Packages in which this project is divided, the milestones, and the Gantt diagram.

1.1.1 Work Packages

Project: Literature review	WP ref: (WP1)	
Major constituent: Research	Sheet 1 of 1	
Short description: Analysis of existing research which is relevant to the topic.	Planned start date: 21/02/2022 Planned end date: 15/03/2022 Start event: 21/02/2022 End event: 15/03/2022	
Internal task T1: Study the project documentation previously done by the other students. Internal task T2: Review the state of the art regarding emotion recognition, focusing on facial emotion recognition and speech emotion recognition. Internal task T3: Download and analysis of the datasets.	Deliverables: None	Dates: None

Table 1: WP1 Literature review

Project: System development	WP ref: (WP2)	
Major constituent: Software	Sheet 1 of 1	
Short description: Processing of the multimedia files to extract features, and design and train the models.	Planned start date: 15/03/2022 Planned end date: 15/04/2022 Start event: 15/03/2022 End event: 15/04/2022	
Internal task T1: Process the dataset videos to extract the frames and the audio files. Internal task T2: Design the Deep Learning models. Internal task T3: Training of the models.	Deliverables: None	Dates: None

Table 2: WP2 System development

Project: Results validation	WP ref: (WP3)	
Major constituent: Simulation and Software	Sheet 1 of 1	
Short description: Evaluation of the performance of the models and application of improvements to enhance the results.	Planned start date: 15/04/2022 Planned end date: 06/06/2022	
	Start event: 15/04/2022 End event: 13/06/2022	
Internal task T1: Analysis of the results. Internal task T2: Retrain the models with the proposed improvements.	Deliverables: Result analysis report	Dates: 24/05/2022

Table 3: WP3 Results validation

Project: Documentation	WP ref: (WP4)	
Major constituent: Documentation	Sheet 1 of 1	
Short description: Prepare the thesis documentation.	Planned start date: 06/06/2022 Planned end date: 02/07/2022	
	Start event: 15/04/2022 End event: 13/06/2022	
Internal task T1: Write the thesis documentation. Internal task T2: Prepare the oral defence of the thesis.	Deliverables: Project documentation	Dates: 02/07/2022

Table 4: WP4 Documentation

1.1.2 Milestones

WP#	Task#	Short title	Milestone / deliverable
1	2	Review the state of the art regarding emotion recognition, focusing on facial emotion recognition and speech emotion recognition	Summary of the information
2	1	Process the dataset videos to extract the frames and the audio files	Save the cropped face images and audio files
2	3	Training of the models	Trained models
3	1	Analysis of the results	Result analysis report
4	1	Write the thesis documentation	Thesis document
4	2	Prepare the oral defence of the thesis	Oral presentation

Table 5: Milestones

1.1.3 Gantt Diagram

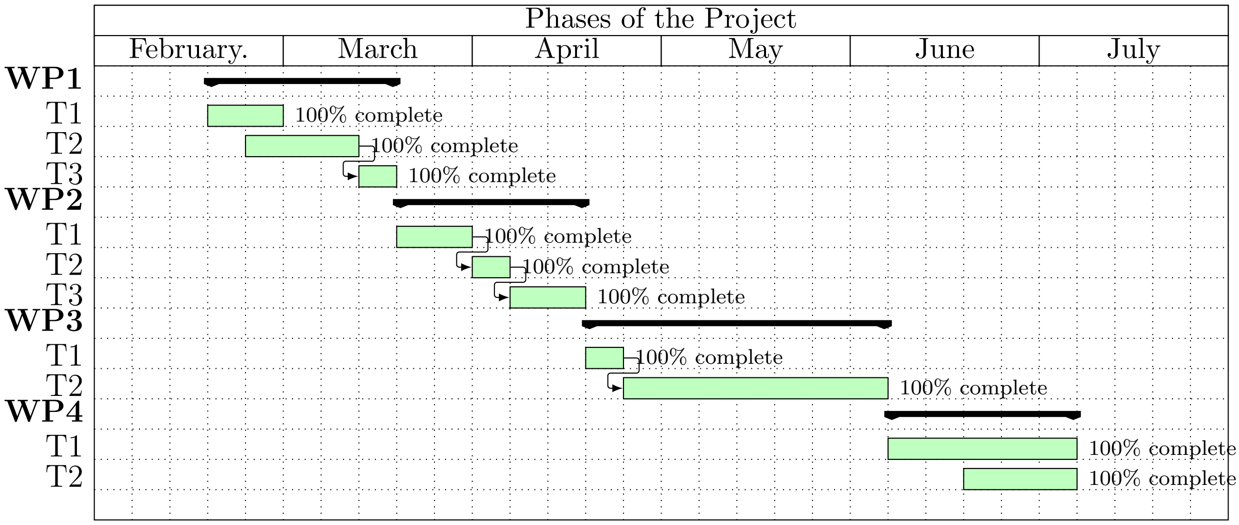


Figure 1: Gantt diagram of the project

1.2 Deviations and incidences

During the development of the project, there has not been any incidence that may have caused a delay in the different work packages. However, there has been an underestimation of both the time required to train the models and the number of retraining attempts. The main reason is that the computation capacity of the computer used was insufficient to deal with data augmentation. Therefore, the task regarding the improvement of the models and the retraining phase took more days than expected.

2 State of the art of the technology used or applied in this thesis:

The rising evidence of the importance of emotions in the interaction between humans has led to an increase in interest in researchers to develop automatic ways for computers to recognize emotional expressions as a goal toward achieving human-computer intelligent interaction. This chapter summarizes the research done on emotion recognition during the last few years that has mostly inspired this work. In addition, the model used to extract the faces from the video frames is described.

2.1 Face Emotion Recognition (FER)

Paul Ekman and his colleagues have been studying human face expressions extensively since the early 1970s [7]. They discovered evidence that facial expressions are universal. They looked at facial expressions in various cultures, including preliterate cultures, and came up with similarities in how people express and recognize emotions on their faces. Despite these similarities, it cannot be ignored the fact that each culture has its own nuances and singularities.

The extraction of the face region and the following landmark detection is necessary for many facial emotion recognition approaches. Some of them apply the Facial Action Coding System (FACS), which uses Units of Action (AU) derived from facial landmarks to represent emotions. An Action Unit could be to ”raise the upper lid” or to “raise de chicks”. An expression is defined by the activation of many AUs. Being able to appropriately detect AUs is a relevant step since it allows you to assess the level of emotion activation. See for example Yu and Zhang [8], who unveiled its Facial Expression Recognition technology that uses Convolutional Neural Networks to recognize facial expressions.

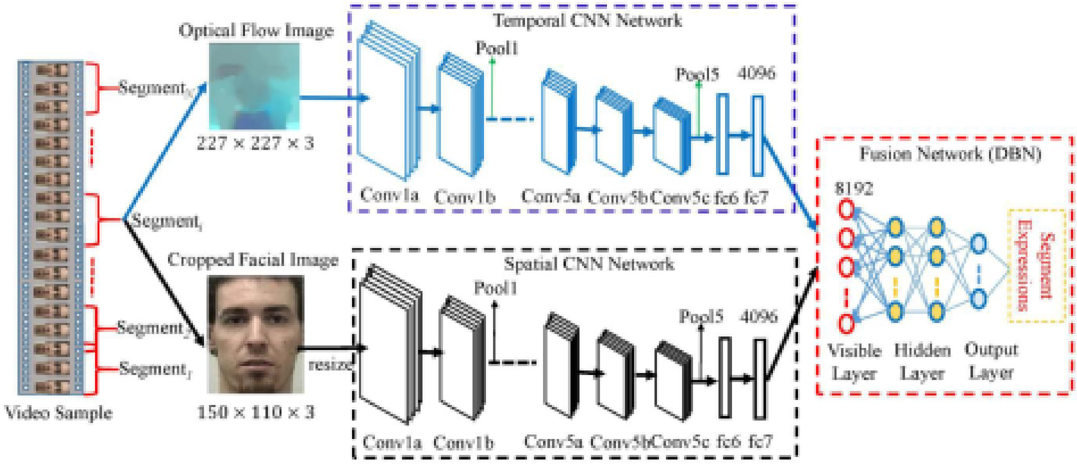


Figure 2: Deep Belief Network (DBN) architecture [9]

Apart from taking advantage of the spatial information provided by static features (images), Mase introduced optical flow to use the temporal information [9]. S. Zhang, X. Pan, Y. Cui, X. Zhao and L. Liu constructed a fusion network built with a Deep Belief Network (DBN) to take full advantage of a 2-stream CNN, demonstrating its effectiveness in capturing spatiotemporal information [10]. Figure 2 shows the model architecture presented in the study, which serves as the basis for the model proposed in this work.

2.1.1 MTCNN for face detection

MTCNN stands for Multi-task Cascaded Convolutional Networks, which was created as a solution for both face detection and face alignment. The method entails three stages of convolutional networks that can detect faces and landmarks. Landmarks consist of the coordinates of some key points allocated around the eyebrows, eyes, nose, mouth, and jaw.

MTCNN is proposed in the paper [6] to use multi-task learning to integrate both tasks (recognition and alignment). It uses a shallow CNN in the first stage to quickly generate candidate windows. It refines the recommended candidate windows in the second stage using a more complex CNN. Finally, in the third step, it employs a third CNN to refine the result and output face landmark positions.

Figure 3 shows the pipeline of the cascaded framework that includes the three-staged multi-task deep convolutional networks.

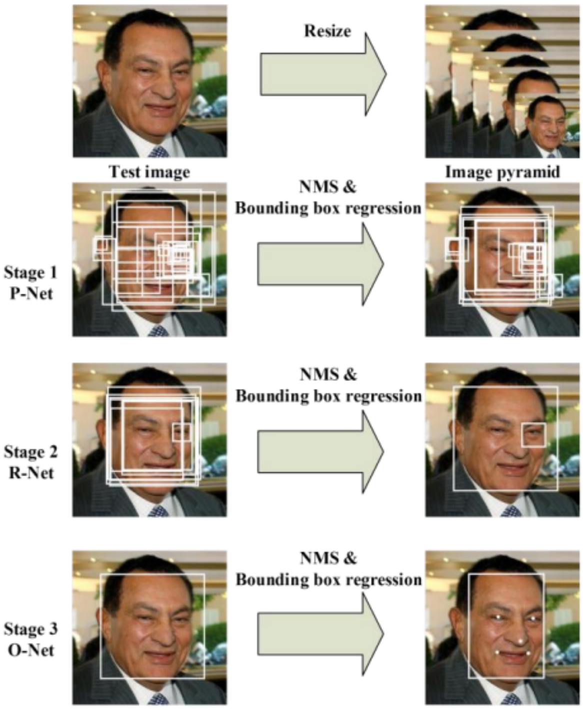


Figure 3: Three-staged multi-task deep convolutional networks [6]

The process of extracting the face starts by resizing the frame to different scales to build an image pyramid, which is the input of the following three-staged cascaded network.

Stage 1: The Proposal Network (P-Net) consists of a fully convolutional network (FCN). The difference between a CNN and a FCN is that the second network does not use a dense layer as part of the architecture. Candidate windows and their bounding box regression vectors are obtained using this Proposal Network.

Stage 2: The Refine Network receives all candidates from the P-Net. The R-Net decreases the number of candidates, performs calibration with bounding box regression, and merges overlapping candidates using non-maximum suppression (NMS). The R-Net outputs whether the input is a face or not, a 4-element vector which is the bounding box for the face, and a 10-element vector for facial landmark localization.

Stage 3: This step is similar to the second stage, but the Output Network's goal is to describe the face in greater detail and output the positions of the five facial landmarks for the eyes, nose, and mouth.

2.2 Speech Emotion Recognition (SER)

In speech emotion recognition challenges, traditional machine learning algorithms such as hidden Markov models (HMMs), support vector machines (SVMs), and decision tree-based methods have been used [11]. Researchers have recently developed several neural network-based architectures to increase speech emotion recognition performance. Deep neural networks (DNNs) were first used by Han et al. to extract high-level characteristics from raw audio data in a study that proved their effectiveness in voice emotion identification [12].

More complex neural-based architectures have been proposed as deep learning approaches have progressed. The information collected from raw audio signals utilizing spectrograms or audio features such as Mel-frequency cepstral coefficients (MFCCs) and low-level descriptors (LLDs) is used to train CNN-based models [13].

2.3 Multimodal Emotion Recognition

Two approaches have been considered for multimodal emotion analysis: feature level fusion and score level fusion. The first refers to the feature union from different recognition systems, and it allows to take advantage of the biometrics synchronism. The other approach combines the scores obtained from each of the classifiers.

For the first approach, Ronghe proposes a hybrid CNN-RNN architecture [14] (see Figure 4). First, the model is trained to categorize images into one of seven emotions. The videos are preprocessed and turned into a sequence of feature vectors, which are then used to train an SVM model. Then, a Multilayer Perceptron models the correlation between features of the emotions from the images and speech. The RNN is fed with the extra parameters to classify the emotional reaction to each video frame.

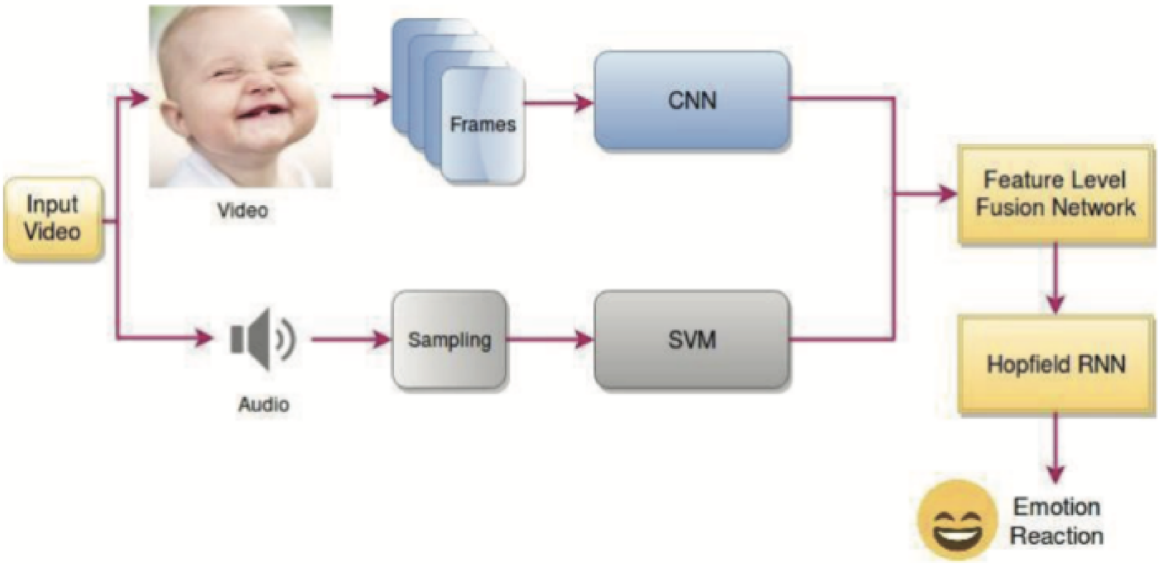


Figure 4: Scheme of the hybrid CNN-RNN model proposed by Ronghe et al. [14]

Regarding the second approach, it has studied the possibility of combining face and voice by employing two independent models connected by a late fusion strategy [15]. As Figure 5 shows below, the architecture proposed consists of two systems: the speech emotion recognizer and the facial emotion recognizer. The speech emotion recognizer uses two transfer learning approaches to avoid having to train a CNN from scratch, which would require a large quantity of data. Feature extraction and fine-tuning are the methods used.

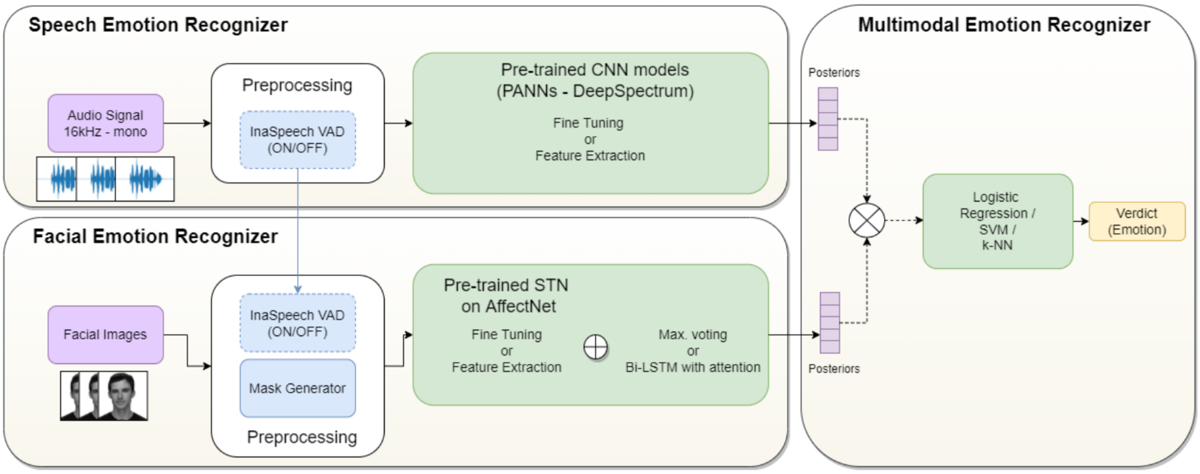

Speech Emotion Recognizer
Audio Signal 16kHz - mono
Preprocessing
InaSpeech VAD (ON/OFF)
Pre-trained CNN models (PANNs - DeepSpectrum)
Fine Tuning or Feature Extraction
Facial Emotion Recognizer
Facial Images
InaSpeech VAD (ON/OFF)
Mask Generator
Preprocessing
Pre-trained STN on AffectNet
Fine Tuning or Feature Extraction
Max. voting or Bi-LSTM with attention
Multimodal Emotion Recognizer
Posteriors
Logistic Regression / SVM / k-NN
Verdict (Emotion)
Posteriors

Figure 5: Multimodal approach with score fusion [15]

2.4 Model performance metrics

Several metrics and graphical techniques are used to assess the performance of a model in a classification problem.

A confusion matrix represents the performance of a supervised learning method. A problem with n classes requires a n x n confusion matrix with the rows representing the actual class and the columns representing the class predicted by the model. Figure 6 is an example of a confusion matrix for multi-class classification [16].

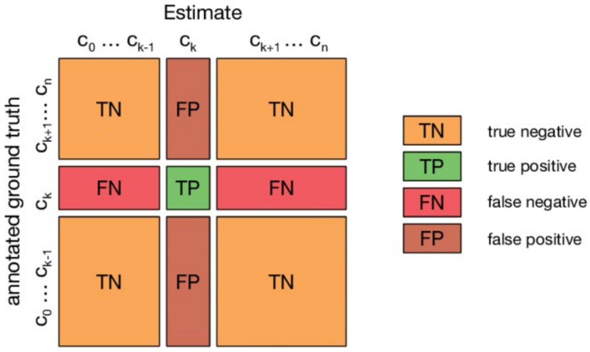


Figure 6: Confusion matrix for multi-class classification [16]

In general, the confusion matrix provides four types of classification results (each of them coded by a different colour in Figure 6) for a particular class k:

- True Positives (TP): Both predicted and actual classes are class k.
- True Negatives (TN): Both predicted and actual classes are not class k.
- False Positives (FP): Predicted class is class k, but the actual class is not.
- False Negatives (FN): Predicted class is not k, but the actual class is.

From the confusion matrix, it is possible to measure several metrics. The ones used in this work are precision, recall, and F1-score.

Precision is the percentage that the model correctly predicts positive when making a decision (see Equation 1).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall, also known as sensitivity, is the percentage of positives correctly identified out of all the existing positives (see Equation 2).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Accuracy, as defined in Equation 3, is the total number of correct predictions made over the total number of predictions. This metric is useful only when classes are equally distributed on the classification. An alternative metric used in imbalanced datasets is F1-score. The F1-score can be interpreted as a harmonic mean of the precision and recall (see Equation 4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

A different approach to analysing the results of a model is using the Receiver Operating Characteristic (ROC) curve. The ROC is a graphical representation of the true positive rate (TPR) (Equation 5) against the false positive rate (FPR) (Equation 6).

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{TN + FP} \tag{6}$$

Figure 7 shows ROC curves for different classifiers. A curve in the top left corner indicates that the classifier is successful, whereas a curve along the diagonal indicates that the classification is random.

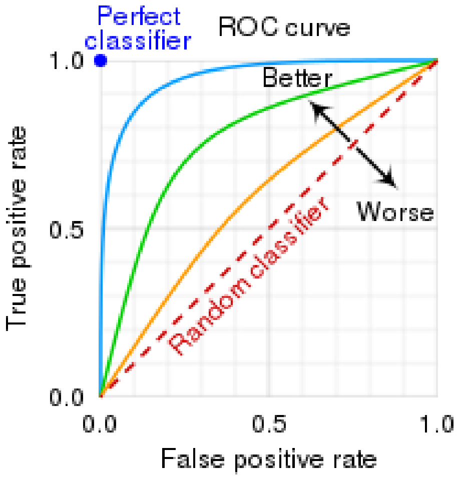


Figure 7: Example of ROC curves [17]

The quantitative value that provides a numerical measure of the ROC is the Area Under the Curve (AUC). The AUC value is between 0.5 (classifier that chooses randomly) and 1. The best value (1) is achieved when the model has an ideal measure of separability, and it is perfectly able to distinguish between the positive class and the negative classes. In this case, FPR is 0 and TPR is 1.

3 Methodology / Project development

This chapter presents the methodology that has been followed in this thesis. It starts by providing a general overview of the datasets used. It is followed by an explanation of the data preprocessing, the data augmentation, the features extracted, and the model developed for each mode (face expression and voice). Finally, it includes how the two biometrics are combined to obtain the final multimodal model.

3.1 Datasets

The datasets used to develop this project are RAVDESS and BAUM-1. While there are numerous datasets of emotions in use currently, the choice of the datasets is based on the aim to compare the results with previous works. Even though the datasets contain videos of various emotions, only those that both datasets have in common have been used: happiness, sadness, anger, fearfulness, and disgust. These emotions are labelled with numbers from 0 to 4.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18] contains videos collected from 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.

The Bahçeşehir University Multimodal Emotional Database (BAUM-1) [19] consists of multimodal facial video clips collected from 31 subjects (14 female, 17 male) who express their ideas and feelings about the images and video clips they have previously watched in an unscripted and unguided way in Turkish.

Both datasets contain expressions produced at two levels of emotional intensity: spontaneous/normal or acted/strong. Only the acted ones are considered since there is a major appreciation and comprehension of the emotions intended to express.

Table 6 shows the number of clips for each of the five emotions for the two datasets.

	Happinesss	Sadness	Anger	Fearfulness	Disgust	Total
RAVDESS	96	96	96	96	96	480
BAUM-1	26	37	42	35	34	174

Table 6: Number of videos per emotion of the datasets

The objective is to develop an emotion recognition model able to generalize and it is essential that the training data comes from videos of the maximum different people possible, having a database including several videos for each emotion and subject. The videos are divided into three sets: training, validation, and test.

The training set contains 60% of the videos and is used to fit the parameters to the model. The validation set consists of 20% of the videos and is used to obtain an unbiased evaluation of the model fitted on the training set while tunning the hyperparameters. The 20% missing is used to test the final model.

As seen in Table 6, the RAVDESS dataset contains enough videos to achieve the stated goal. However, in the case of the BAUM-1 dataset, for each emotion there is barely one video per person. The solution adopted is to split the BAUM-1 videos to double the number of videos per subject per emotion.

Table 7 presents the number of videos for each set once the BAUM-1 videos are split.

		Happy	Sad	Angry	Fearful	Disgust	Total
RAVDESS	Train	57	58	57	58	58	288
	Validation	19	19	20	19	19	96
	Test	20	19	19	19	19	96
BAUM-1	Train	31	44	42	38	40	195
	Validation	10	15	14	12	14	65
	Test	11	15	13	13	13	65

Table 7: Train, validation, and test number of clips per emotion

3.2 Face Emotion Recognition

Facial Emotion Recognition (FER) is the technology that analyses facial expressions from static photos and videos to disclose information about a person's emotional state.

3.2.1 Frames filtering

The framework used to extract the frames from the videos is OpenCV. The total number of frames generated is 64114 for the RAVDESS dataset and 21380 and the BAUM-1 dataset.

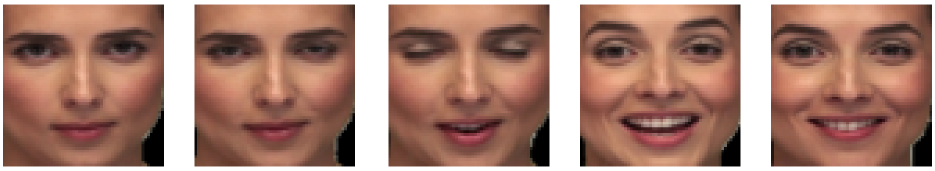


Figure 8: Sequence of frames labelled as "Happy"

One of the challenges of FER is dealing with lower-intensity expressions. Some frames in a video sequence cannot be considered meaningful to the emotion that is supposed to be represented. The figure above shows an example of a sequence of frames extracted from a video labelled as "happy". It can be observed that some frames express more a neutral emotion rather than happiness.

Consequently, the objective is to extract only the high-intensity expression or, in other words, remove the less significant frames. This process is carried out using the CNN model used in Sara Tramonte's work [20], which is fitted with approximately 600 relevant frames selected "by hand". Then, the trained model predicts the label of the dataset frames and removes the misleadingly predicted (Appendix A contains examples of removed lower-intensity frames).

After applying the filtering process to both datasets, Table 8 summarizes the number of frames per emotion used for training, validation, and testing.

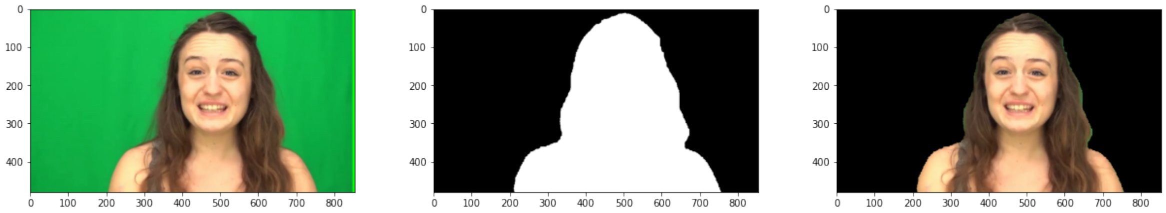
		Happy	**Sad**	**Angry**	**Fearful**	**Disgust**	**Total**
RAVDESS	Train	6371	6587	6799	6271	7030	33058
	Validation	2139	2150	2390	2072	2411	11162
	Test	2318	2131	2219	2105	2334	11107
BAUM-1	Train	1161	2804	1518	1037	1376	7896
	Validation	470	1111	588	479	503	3160
	Test	417	848	171	410	469	2315

Table 8: Train, validation, and test number of frames per emotion

3.2.2 Frame preprocessing

By taking a close look at the frames, it is possible to see differences between the datasets. As these discrepancies result in poor cross-dataset performance, it is imperative to preprocess the frames to "normalize" the images to a common format. This includes removing the background colour, extracting the face region, and resizing the images.

First of all, the background is removed using the MediaPipe framework. MediaPipe [21] is a framework for building machine learning pipelines for processing time-series data like video, audio, etc. Selfie Segmentation module segments the prominent humans in the scene. It processes the image and creates a mask with the segmented background. The result is obtained by applying the bit-wise AND operation of the original image and the mask. Figure 9 displays an example of the stated process.



(a) Original (b) Mask (c) Result

Figure 9: Background removal process

Then, faces are detected and extracted from the frames using MTCNN (see a detailed explanation of the model in Section 2.1.1). The model takes the frame as the input and returns the bounding box coordinates and the position of the keypoints (eyes, nose, and mouth).

The proposed face region contains irrelevant information that may lead to errors when recognizing emotions, such as parts of the background, ears, neck, or hair. Therefore, this work proposes using the positions of the five facial landmarks for the left eye, right eye, nose, left mouth, and right mouth to crop the region of interest more accurately. The algorithm below details de process.

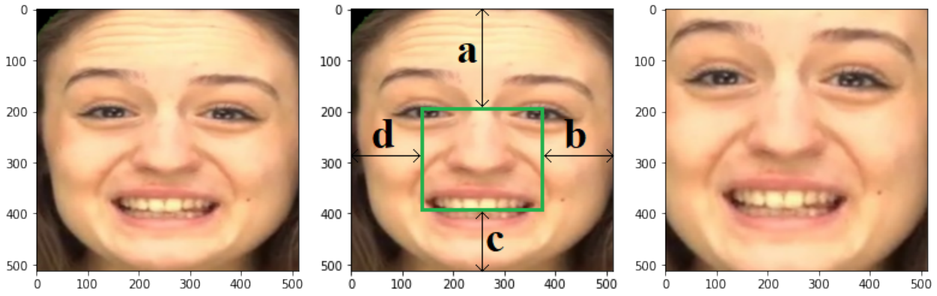
Algorithm 1 Face region extraction using landmarks algorithm

```
procedure EXTRACT-FACE-REGION($box, keypoints$)
    $startX, startY \leftarrow box[0], box[1]$        ▷ Upper-left (X,Y) coordinates of the initial box
    $width, height \leftarrow box[2], box[3]$         ▷ Width and height of the initial box
    $endX \leftarrow startX + width$                  ▷ Lower-right X coordinate of the initial box
    $endY \leftarrow startY + height$                 ▷ Lower-right Y coordinate of the initial box
    $startY_r \leftarrow 0$
    $startY_l \leftarrow 0$
    $endY_r \leftarrow startY + height$
    $endY_l \leftarrow startY + height$
    if $'mouth\_left'$ in $keypoints.keys()$ then
        $mouth\_left_Y \leftarrow keypoints['mouth\_left'][1]$
        $endY_l \leftarrow round(\frac{1}{6} * mouth\_left_Y + \frac{5}{6} * endY)$
    if $'mouth\_right'$ in $keypoints.keys()$ then
        $mouth\_right_Y \leftarrow keypoints['mouth\_right'][1]$
        $endY_r \leftarrow round(\frac{1}{6} * mouth\_right_Y + \frac{5}{6} * endY)$
    if $'left\_eye'$ in $keypoints.keys()$ then
        $eye\_left_X \leftarrow keypoints['left\_eye'][0]$
        $eye\_left_Y \leftarrow keypoints['left\_eye'][1]$
        $startX \leftarrow round(\frac{1}{3} * eye\_left_X + \frac{2}{3} * startX)$
        $startY_l \leftarrow round(\frac{1}{2} * (eye\_left_Y + startY))$
    if $'right\_eye'$ in $keypoints.keys()$ then
        $eye\_right_X \leftarrow keypoints['right\_eye'][0]$
        $eye\_right_Y \leftarrow keypoints['right\_eye'][1]$
        $endX \leftarrow round(\frac{2}{3} * endX + \frac{1}{3} * eye\_right_X)$
        $startY_r \leftarrow round(\frac{1}{2} * (eye\_right_Y + startY))$
    $startY \leftarrow max(startY, min(startY_l, startY_r))$
    $endY \leftarrow min(endY, max(endY_l, endY_r))$
    $return\ startX, startY, endX, endY$              ▷ Updated box
```

Given the MTCNN bounding box, and the keypoints (identified by a pixel position (x, y)), the algorithm returns the upper-left and the lower-right pixel coordinates that define the updated face region. First, the coordinates are initialized with those proposed by the MTCNN model. For each keypoint detected, the region is modified as follows:

- *Mouth_left* and *Mouth_right* refer to the landmarks that correspond to the extremes of the mouth. They are used to reduce the distance between the lower edge of the box and the mouth.
- *Left_eye* and *Right_eye* correspond to the centre of the left and right eye. They are used to reduce the lateral distance between the left and right edges of the box and the eyes. Moreover, they are used to reduce the distance between the upper edge of the box and the eyes.

Figure 10 shows the face detection process and the differences between the suggested regions. The final face region lies between the box that perfectly fits the landmarks and the proposed box by the face detector model. Figure 10b displays the distances between the edges of the two regions, which are used to draw the final region. The top edge of the final region is placed in the mediatrix of distance **a**. The lateral edges are located at 1/3 and 2/3 of distances **d** and **b**, respectively. Finally, the bottom edge is positioned at 5/6 of distance **c**. These proportions are computed to ensure no loss of relevant information about the face. Finally, the frame is resized to 512x512 pixels..



(a) Region proposed by MTCNN

(b) Region that perfectly fits the landmarks

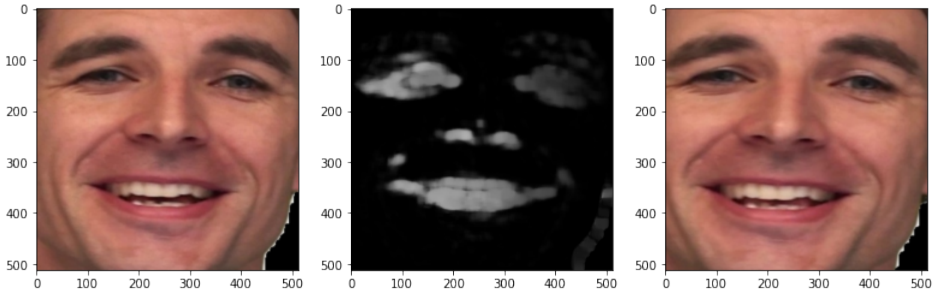
(c) Final face region

Figure 10: Face detection process

3.2.3 Optical flow frames

Apart from taking advantage of the spatial information provided by static features (face images), this work uses optical flow to gather temporal information. Two different models have been designed and trained for each type of information: In the first case, the CNN input is the face images and, in the other case, the optical flow frames.

Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of an object, in the case of this project, face muscles, lips, eyebrows, etc. Optical flow images are computed from optical flow vectors, which are 2D vectors where each vector represents a displacement showing the movement of points from one frame to the following one, with its magnitude and direction. Vectors are then mapped onto colour for better visualization. The optical flow frames are generated from the normalized face images using dense optical flow as implemented in [22] (see example in Figure 11).



(a) Previous frame

(b) Optical flow

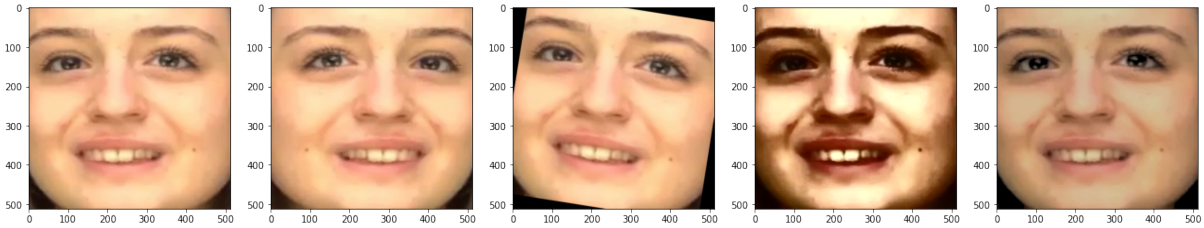
(c) Next frame

Figure 11: Optical flow frame

3.2.4 Data augmentation

The quality, quantity, and contextual significance of training data all play a role in the performance of deep learning models. However, one of the most common obstacles in developing deep learning models is the lack of data. The use of transfer learning with VGG-16 or ResNet50 models pre-trained on ImageNet has been considered, but the results were disappointing. Therefore, the models have been trained from scratch also applying data augmentation techniques.

Data augmentation is a process of artificially increasing the amount of data by generating new data points from existing data. Its main advantages are contributing to avoiding overfitting and balancing the dataset. The augmented data derive from original images with minor geometric transformations. In this work, the techniques used are mirroring, rotation, histogram equalization, and modifying the brightness(see Figure 12).



(a) Original (b) Mirroring (c) Rotation (d) Histogram Equalization (e) Brightness

Figure 12: Image data augmentation techniques

Mirroring consists of flipping horizontally the given image. Considering that the videos are filmed by cameras and can be placed in different places, it is interesting to have distinct points of view for recognition. Taking into account the fact that, in some videos, the subject moves its head while speaking, rotating the images at a random angle might be helpful. Histogram Equalization increases contrast in images by detecting the distribution of pixel densities in an image and plotting these pixel densities on a histogram. The distribution of this histogram is then analysed and if there are ranges of pixel brightness's that aren't currently being utilized, the histogram is then "stretched" to cover those ranges, and then is "back projected" onto the image to increase the overall contrast of the image [23]. As dataset recording conditions differ among them, it is preferred to train the model with images whose brightness has been modified, simulating different light conditions. After balancing the dataset and augmenting the training dataset, the number of frames per class for training is the following:

		Happy	**Sad**	**Angry**	**Fearful**	**Disgust**	**Total**
RAVDESS	Face images	13223	13223	13223	13223	13223	66115
	Optical flow	10072	10072	10072	10072	10072	50360
BAUM-1	Face images	8412	8412	8412	8412	8412	42060
	Optical flow	3964	3964	3964	3964	3964	19820

Table 9: Train set number of frames per emotion after data augmentation

3.2.5 Model

This section describes the two classifiers used to predict face emotions: the face image model and the optical flow model. The structure of the models is the same in both cases, and it is a simplified version of the VGG16 network [24]. Figure 13 displays the architecture of the face image model. The only difference between the classifiers is that the input layer is adapted to the number of image channels (3 for RGB face images and 1 for black and white optical flow frames). The five stages of convolutional layers and pooling layers extract the features from the images, and the dense layers classify the features to predict the label.

Face image model

Layer	Output shape	Param #
Input (InputLayer)	(None, 64, 64, 3)	0
Conv2d (1) (Conv2D)	(None, 64, 64, 32)	896
Batchnormalization (1) (BatchNormalization)	(None, 64, 64, 32)	128
Pool2d (1) (MaxPooling2D)	(None, 32, 32, 32)	0
Conv2d (2) (Conv2D)	(None, 32, 32, 64)	18496
Batchnormalization (2) (BatchNormalization)	(None, 32, 32, 64)	256
Pool2d (2) (MaxPooling2D)	(None, 16, 16, 64)	0
Conv2d (3) (Conv2D)	(None, 16, 16, 128)	73856
Batchnormalization (3) (BatchNormalization)	(None, 16, 16, 128)	512
Pool2d (3) (MaxPooling2D)	(None, 8, 8, 128)	0
Conv2d (4) (Conv2D)	(None, 8, 8, 256)	295168
Batchnormalization (4) (BatchNormalization)	(None, 8, 8, 256)	1024
Pool2d (4) (MaxPooling2D)	(None, 4, 4, 256)	0
Conv2d (5) (Conv2D)	(None, 4, 4, 512)	1180160
Batchnormalization (5) (BatchNormalization)	(None, 4, 4, 512)	2048
Pool2d (5) (MaxPooling2D)	(None, 2, 2, 512)	0
Dropout (1) (Dropout)	(None, 2, 2, 512)	0
Flatten (Flatten)	(None, 2048)	0
Dense (1) (Dense)	(None, 1024)	2098176
Dropout (2) (Dropout)	(None, 1024)	0
Dense (2) (Dense)	(None, 512)	524800
Dropout (3) (Dropout)	(None, 512)	0
Output (Dense)	(None, 5)	2565

Figure 13: Face image model architecture

The frames are resized to 64x64 pixels and normalized so the pixel values are between 0 and 1. The Conv2D layer creates a convolution kernel, where the first parameter is the filter that indicates the dimensionality of the output space, the second is the kernel size which specifies the dimensions of the 2D convolution window, and last is the padding used to add zeros to the output so it will have the same height/width dimension as the input. The MaxPool2D layer is used to down sample the input along its spatial dimension, and its parameter is the pooling window size. The Dropout layer randomly sets input units to 0 with a specified rate at each step during training time, which helps prevent overfitting. The Flatten layer flattens the input, and the Dense layer is a densely connected layer, where the first parameter, called units, is an integer indicating the dimensionality of the output. The last dense layer of the model is activated by the Softmax function, whose output is an array with the probabilities that the frame belongs to each class.

The model is compiled using Adam as an optimizer, which is a stochastic gradient descent method well suited for most problems and using as a loss function the *sparse_categorical_crossentropy*, which calculates the cross-entropy loss between the predictions and the labels.

Finally, the model is trained using *ReduceLROnPlateau* and *EarlyStopping*. The first callback reduces the learning rate when the validation accuracy is not improving after two epochs. The second one stops the training and restores the weights from the epoch with the best validation accuracy when the monitored metric does not improve.

3.3 Speech Emotion Recognition

Speech Emotion Recognition is the act of attempting to recognize human emotion and affective states from speech. Voice often reflects underlying emotion through tone and pitch.

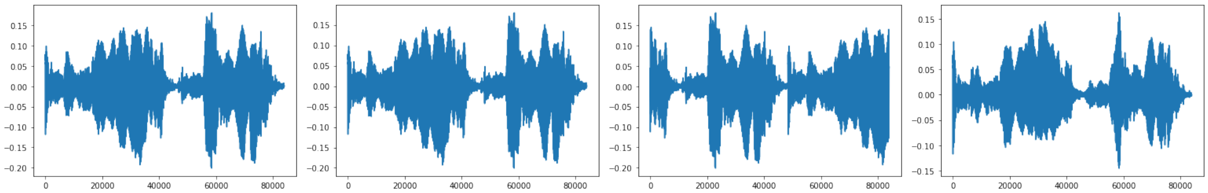
3.3.1 Audio extraction

Librosa is a Python library for analysing audio and music [25], and it is used to obtain the audio files from the video. The load function takes the video path as the input and returns the audio signal and the sample rate. Both datasets have in common the sample rate, but in the case of the RAVDESS dataset, the speaker waits one second before starts speaking. The silent parts of the signal at the beginning and the end are removed to homogenize the datasets. Mention that silence is considered when the peak amplitude is 30 dB lower than the maximum peak amplitude.

3.3.2 Data augmentation

Data augmentation techniques for speech are used to generate new audio files from the original data. This work applies the techniques used in Sara Tramonte's work that are noise injection and changing pitch. In addition, this work proposes also using time shifting [26].

Noise injection consists in adding random values to data. The result is the original signal plus Gaussian white noise (GWN). The other techniques include shifting the audio to the right or the left with a random second and randomly modifying the pitch. Figure 14 represents the mentioned speech augmentation approaches.

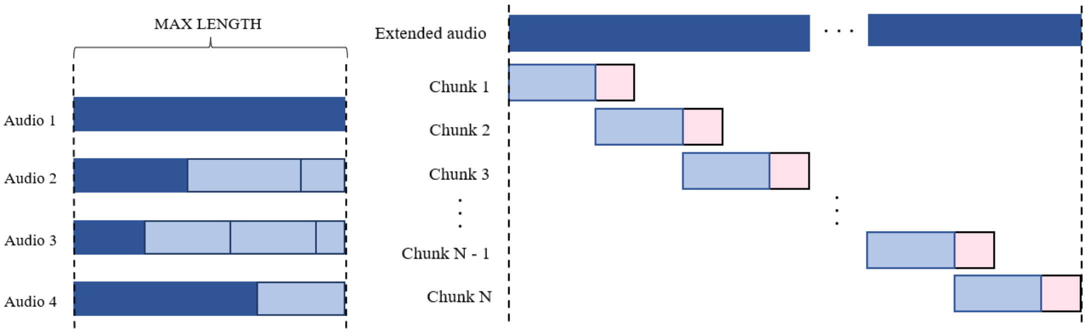


(a) Original (b) Noise injection (c) Time shifting (d) Pitch

Figure 14: Speech data augmentation techniques

3.3.3 Audio chunking

This section introduces a novel approach for the data preparation process implemented before the feature extraction. As videos have different durations, the extracted audio signal arrays vary in their lengths, and so does the feature vectors' length. The CNN model requires a fixed input dimension, so it is necessary to chunk the audios at a specific duration.



(a) Speech concatenation (b) Speech chunking

Figure 15: Speech chunking process

Once the dataset is balanced using speech data augmentation, it is ensured that all the audios have the same length to guarantee that the number of chunks per class is equal. The proposed algorithm calculates the maximum audio duration of the dataset and repeatedly concatenates the original signal to itself until it matches this length (see Figure 15a).

After the speech signals have been lengthened, they are segmented using the Librosa framework. According to [27], the optimal fixed length of the chunk is 1 s. After several experiments, the overlap between chunks was fixed to 250 ms (see Figure 15b).

Table 10 shows the total number of chunks per class for both datasets.

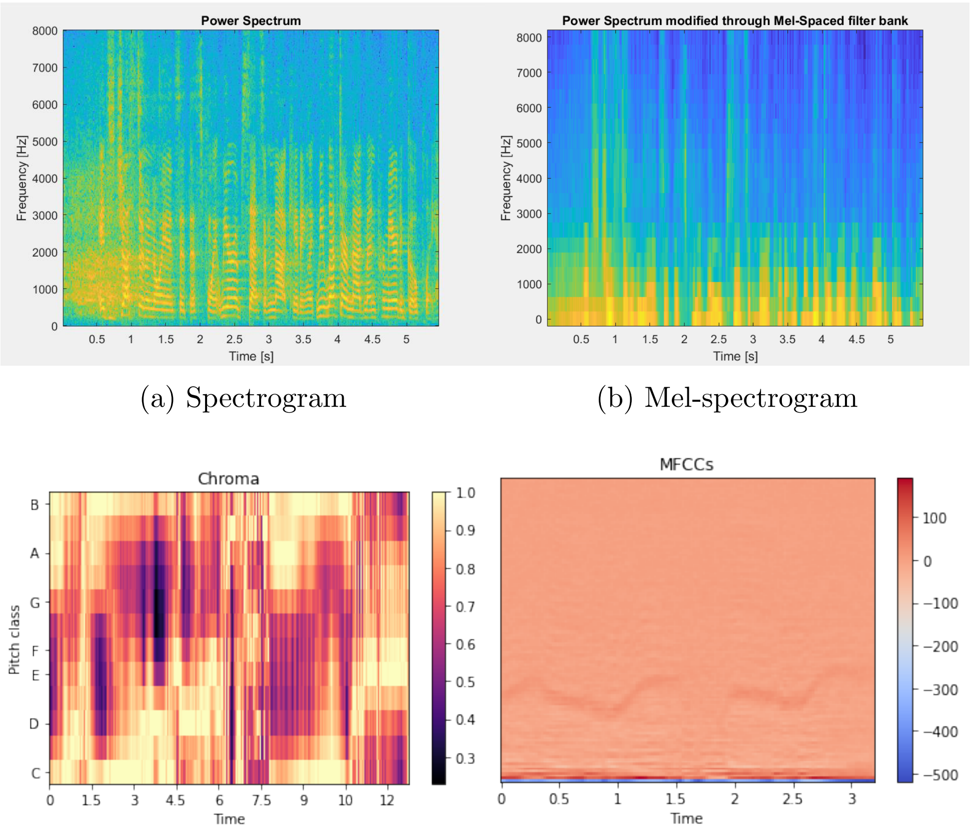
		Happy	**Sad**	**Angry**	**Fearful**	**Disgust**	**Total**
RAVDESS	Train	1160	1160	1160	1160	1160	5800
	Validation	100	100	100	100	100	500
	Test	100	100	100	100	100	500
BAUM-1	Train	1760	1760	1760	1760	1760	8800
	Validation	150	150	150	150	150	750
	Test	150	150	150	150	150	750

Table 10: Train, validation, and test number of audio chunks per emotion

3.3.4 Features extraction

There are two types of audio features: physical features and perceptual features [28]. Physical features refer to mathematical measurements computed directly from the sound wave, such as the energy function, the spectrum, the cepstral coefficients, the fundamental frequency, etc. Perceptual features are subjective terms related to the perception of sounds by human beings, including loudness, brightness, pitch, timbre, rhythm, etc. This work uses the Zero-Crossing rate, the Mel spectrogram, the chroma, the Mel Frequency Cepstral Coefficients, the delta MFCCs, and the delta-delta MFCCs, to classify the emotions from the audio files. For each of the speech chunks, a 525-length feature vector is computed using Librosa.

The zero-crossing rate (ZCR) is the rate at which a signal transitions from positive to zero to negative or negative to zero to positive. It can be utilized as a basic pitch detection algorithm for monophonic tonal signals [29].



(a) Spectrogram

(b) Mel-spectrogram

(c) Chroma

(d) MFCCs

Figure 16: Speech features

Spectrograms represent the signal power over time at different frequencies. Mel-Spaced filter banks are applied to the power spectrum to extract the Mel-spectrogram. The Mel-scale aims to mimic the non-linear human ear perception of sound by being more discriminative at lower frequencies and less discriminative at higher frequencies. Figure 16a and Figure 16b display the spectrogram and the Mel-spectrogram of a speech signal.

Chroma is a representation of music audio in which the entire spectrum is projected onto 12 bins representing the 12 different semitones of the musical octave (see Figure 16c). Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (i.e., the original octave) can give useful musical information about the audio and may even reveal perceived musical similarity that is not apparent in the original spectra [30].

The Mel-frequency cepstrum is highly effective in audio recognition and modelling the audio signal subjective pitch and frequency content. The Mel-frequency Cepstral Coefficients (MFCCs) are computed from the FFT power coefficients, which are filtered by a triangular band pass filter bank (Mel-Spaced filter bank) [31]. Figure 16d shows an example of the MFCCs.

Delta and Delta-Delta MFCCs are also known as differential and acceleration coefficients. The idea behind using differential and acceleration coefficients is to understand the dynamics of the power spectrum, i.e., the trajectories of MFCCs over time.

3.3.5 Model

The model used to classify the audios is a CNN with two stages of convolutional layers appended to a fully-connected neural network. The voice model takes as input the feature vectors explained in the previous section and returns, for each class, the probability that the chunk belongs to it. Figure 17 describes the architecture of the classifier.

Voice model

Layer	Output Shape	Param #
Input (InputLayer)	(None, 525, 1)	0
Conv1d (1) (Conv1D)	(None, 525, 128)	768
Conv1d (1) (Conv1D)	(None, 525, 128)	82048
Dropout (1) (Dropout)	(None, 525, 128)	0
Flatten (Flatten)	(None, 67200)	0
Dense (1) (Dense)	(None, 256)	17203456
Dropout (2) (Dropout)	(None, 256)	0
Dense (2) (Dense)	(None, 128)	32896
Dropout (3) (Dropout)	(None, 128)	0
Dense (3) (Dense)	(None, 128)	16512
Dropout (4) (Dropout)	(None, 128)	0
Output (Dense)	(None, 5)	645

Figure 17: Voice model architecture

The model uses Adam as an optimizer and the *sparse_categorical_crossentropy* as a loss function. It also uses the callbacks explained in Section 3.2.5: *ReduceLROnPlateau* and *EarlyStopping*.

3.4 Multimodal Emotion Recognition

Single-mode biometric solutions have limitations in terms of accuracy and susceptibility to spoofing. Several small-scale experimental experiments have shown that combining evidence from several sources enhances performance [32].

The fusion (i.e., combination) of the various biometric mode data is the key to multimodal biometrics. Fusion can occur at the feature extraction, match-score, or decision level [33]. Feature level fusion combines feature vectors at the representation level that essentially provides higher dimensional data points when comparing the matching score. Match-score level fusion combines the disjoint confidence scores. Decision level fusion combines the final decisions of the different systems.

The strategy adopted in this work is the score-level fusion, and it is implemented in two stages: an intra-model score fusion and an inter-model score fusion. Figure 18 describes the steps for one video, but the process is repeated for all the videos of the datasets.

The intra-model stage consists in averaging the frames scores or the audio chunk scores to obtain the video score. A video is preprocessed to obtain the cropped face images, the optical flow frames, and the voice chunks features, that are the input of their corresponding model. Then, each model makes its prediction and returns the scores, resulting in three matrices with K, M, and N rows and five columns (being K the number of image faces, M the number of optical flow frames, N the number of audio feature vectors, and five the number of emotions). For each matrix, each column is averaged to obtain the video scores. Once the first stage is completed, the results are three video score proposals, one per model, represented on three vectors. Each vector has five components ($\mathrm{P}_{\mathrm{happy}}$, $\mathrm{P}_{\mathrm{sad}}$, $\mathrm{P}_{\mathrm{angry}}$, $\mathrm{P}_{\mathrm{fearful}}$, $\mathrm{P}_{\mathrm{disgust}}$) representing the probabilities that the video belongs to each emotion.

The inter-model fusion refers to averaging the video scores of each model to generate the final video scores. The three vectors are averaged, and the result is a single vector with the final scores. At this point, the way is given to the argmax function that returns the final label by selecting the emotion with the highest score.

Figure 18 shows the schematic of the multimodal emotion recognition system.

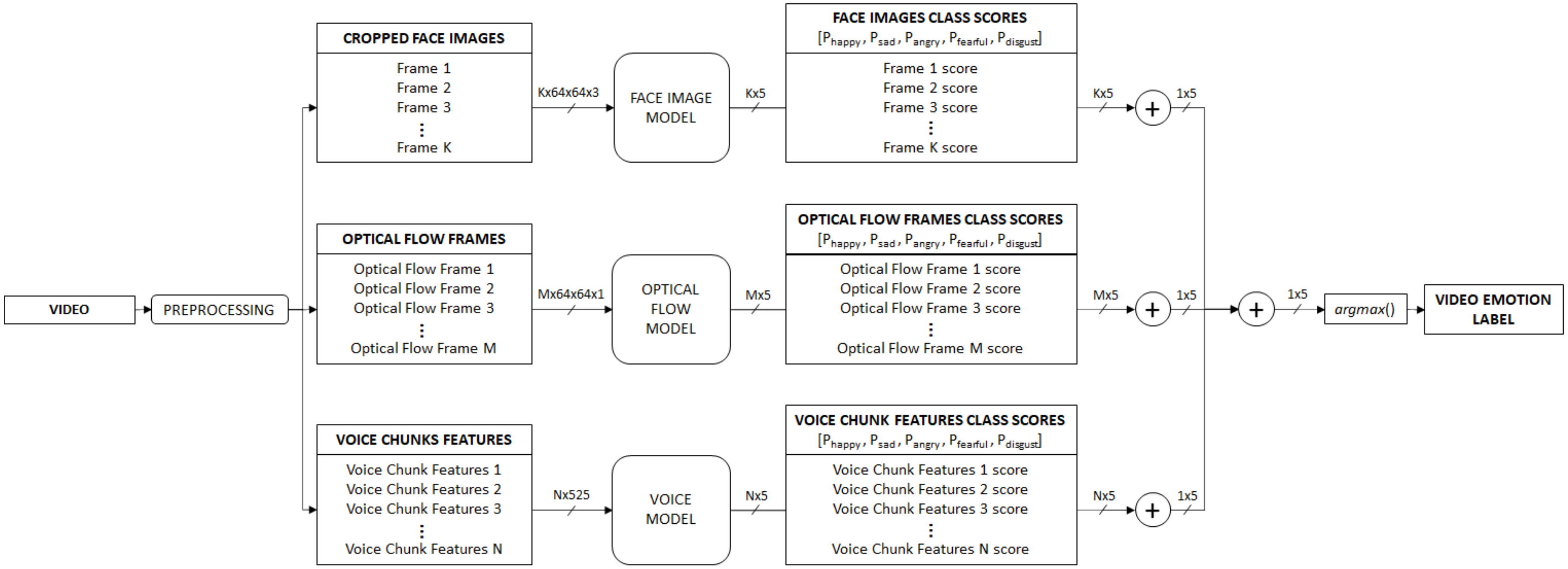

VIDEO
PREPROCESSING
CROPPED FACE IMAGES
Frame 1
Frame 2
Frame 3
⋮
Frame K
OPTICAL FLOW FRAMES
Optical Flow Frame 1
Optical Flow Frame 2
Optical Flow Frame 3
⋮
Optical Flow Frame M
VOICE CHUNKS FEATURES
Voice Chunk Features 1
Voice Chunk Features 2
Voice Chunk Features 3
⋮
Voice Chunk Features N
Kx64x64x3
Mx64x64x1
Nx525
FACE IMAGE MODEL
OPTICAL FLOW MODEL
VOICE MODEL
Kx5
Mx5
Nx5
FACE IMAGES CLASS SCORES
[P_{happy}, P_{sad}, P_{angry}, $P_{fearful}$, $P_{disgust}$]
Frame 1 score
Frame 2 score
Frame 3 score
⋮
Frame K score
OPTICAL FLOW FRAMES CLASS SCORES
[P_{happy}, P_{sad}, P_{angry}, $P_{fearful}$, $P_{disgust}$]
Optical Flow Frame 1 score
Optical Flow Frame 2 score
Optical Flow Frame 3 score
⋮
Optical Flow Frame M score
VOICE CHUNK FEATURES CLASS SCORES
[P_{happy}, P_{sad}, P_{angry}, $P_{fearful}$, $P_{disgust}$]
Voice Chunk Features 1 score
Voice Chunk Features 2 score
Voice Chunk Features 3 score
⋮
Voice Chunk Features N score
Kx5
Mx5
Nx5
+
1x5
1x5
1x5
+
1x5
argmax()
VIDEO EMOTION LABEL

Figure 18: Schematic of the multimodal emotion recognition system

4 Results

This chapter presents the main results obtained in this project for facial, speech, and multimodal emotion recognition. It also includes a comparison of this work with Sara Tramonte's proposal evaluated on the same datasets, giving an overview of the possible advantages of the techniques applied in this thesis. It details the individual results of the three models and the performance results of the face model (fusion of the face image model and optical flow model) and the multimodal system (three models fused), evaluated for each dataset and cross-dataset.

The following sections summarize the most relevant results for each model. They are calculated based on the classification of the individual frames or audio chunks (Ind. column of the tables) or based on the classification of the videos (Vid. column of the tables).

The tables below contain the weighted average results for each of the metrics explained in Section 2.4. Weighted average considers how many frames, audio chunks, or videos of each emotion are in its calculation. It means that the impact of a class on the weighted average of each of the metrics is directly proportional to the amount of data it has.

While this chapter presents the averaged results, Appendix B contains an extended analysis by emotion. It shows, for each of the models and emotions, the performance metrics, and the ROC curves. In addition, it includes the train, validation, and test confusion matrices of all the classifiers.

4.1 RAVDESS

This section aims to show the results obtained when the models are trained and tested with the RAVDESS dataset.

		Precision		Recall		F1-score		AUC	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
	Val	0.88	0.96	0.87	0.96	0.87	0.96	0.98	1.00
	Test	0.90	0.97	0.89	0.97	0.89	0.97	0.99	1.00
Optical flow	Train	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
	Val	0.78	0.90	0.77	0.90	0.77	0.90	0.95	0.99
	Test	0.78	0.94	0.78	0.94	0.78	0.94	0.95	1.00
Voice	Train	0.88	0.90	0.88	0.90	0.87	0.90	0.97	0.98
	Val	0.84	0.85	0.82	0.83	0.83	0.83	0.94	0.94
	Test	0.86	0.87	0.86	0.88	0.85	0.87	0.97	0.98
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.96	-	0.96	-	0.96	-	1.00
	Test	-	0.98	-	0.98	-	0.98	-	1.00
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.98	-	0.98	-	0.98	-	1.00
	Test	-	0.99	-	0.99	-	0.99	-	1.00

Table 11: RAVDESS averaged results

Table 11 outlines the performance of the models tested. The results indicate that the multimodal system can correctly classify the emotion in most videos (99%). Its performance is better than the individual performance of the three models, which means that the three modes can be used to complement each other to generate a more robust model.

The results in terms of video classification outperform the ones considering the individual classification of frames or audio chunks. The main reason is that the video score is obtained by averaging the individual scores of the frames, so if the number of misclassified frames is lower than the number of correctly classified frames, the predicted emotion of the video will be correct.

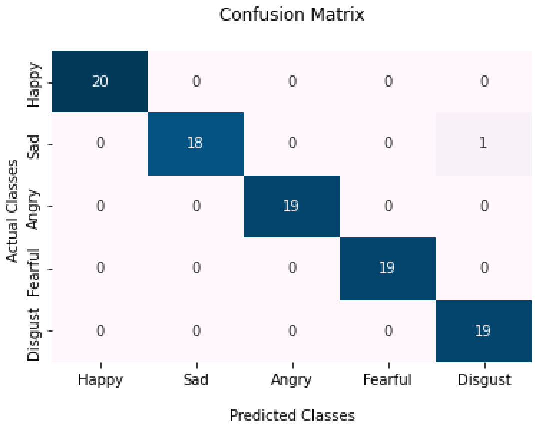
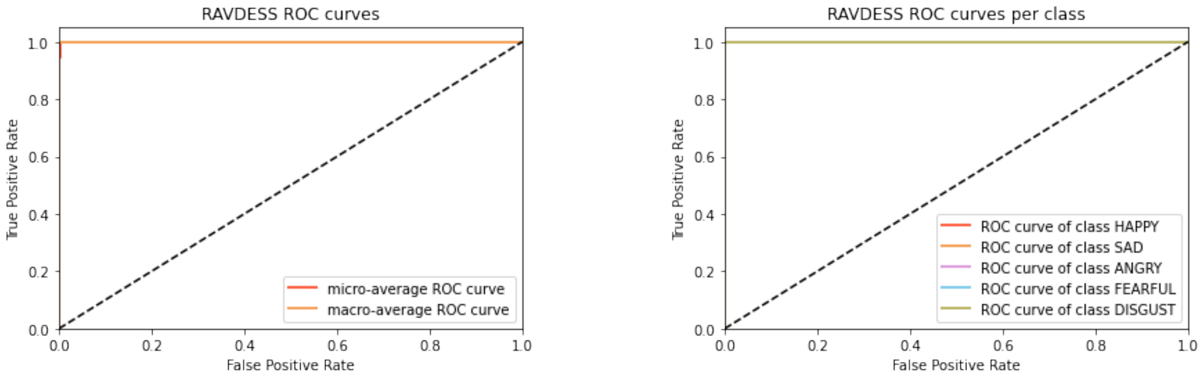


Figure 19: RAVDESS test confusion matrix

The confusion matrix in Figure 19 shows that just one video of the test dataset has been misleadingly classified by the final model. This emphasizes the outstanding results with the RAVDESS dataset.

Figures 20a and 20b display the ROC curves of the system. The average results and the analysis per class reveal an almost perfect classifier, as the curves are plotted in the top-left corner.



(a) Averaged

(b) Per class

Figure 20: RAVDESS ROC curves

As stated in the introduction, the objective is to improve the results of previous works. Table 12 compares the accuracies of the models developed in Sara Tramonte's work with the ones obtained in this project. It can be said that the accuracies of the individual model and the multimodal system have increased. The most relevant improvement is in the optical flow model, which may be caused by computing the optical flow frames once the frames have been precisely cropped to eliminate unnecessary information.

	Previous work	**This work**	**Improvement (%)**
Face Image	0.94	0.97	3.19
Optical flow	0.69	0.94	36.23
Voice	0.86	0.87	1.16
Face + Voice	0.91	0.99	8.79

Table 12: RAVDESS accuracy results comparison

4.2 BAUM-1

This section presents the results obtained when the models' training and the testing are done on the BAUM-1 dataset.

		Precision		**Recall**		**F1-score**		**AUC**	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.84	0.89	0.84	0.88	0.82	0.88	0.97	0.97
	Test	0.81	0.83	0.74	0.82	0.75	0.83	0.95	0.96
Optical flow	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.67	0.71	0.65	0.69	0.65	0.70	0.88	0.88
	Test	0.63	0.68	0.54	0.61	0.55	0.60	0.85	0.85
Voice	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.61	0.67	0.60	0.66	0.60	0.66	0.84	0.88
	Test	0.73	0.83	0.73	0.81	0.73	0.80	0.90	0.93
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.81	-	0.80	-	0.80	-	0.96
	Test	-	0.78	-	0.77	-	0.77	-	0.94
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.87	-	0.86	-	0.86	-	0.99
	Test	-	0.85	-	0.84	-	0.84	-	0.98

Table 13: BAUM-1 averaged results

The table above shows that the results are worse than when using the RAVDESS dataset. One possible explanation is that the quality of the images and the sound of this dataset is lower than the previous one. In addition, comparing the training, validation, and test results it is seen that the models are overfitted. Overfitting occurs when the model fits exactly against the training data and cannot perform accurately against unseen data. However, the models still can distinguish clearly between the five emotions.

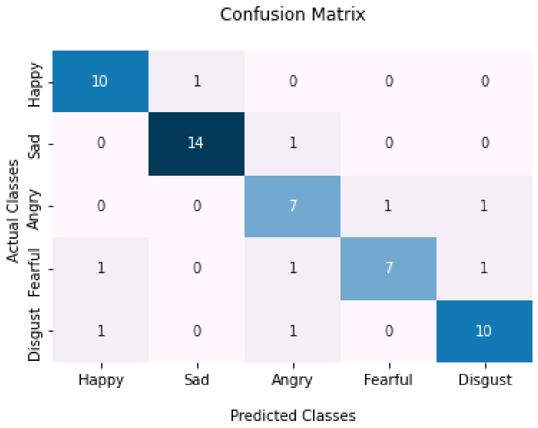
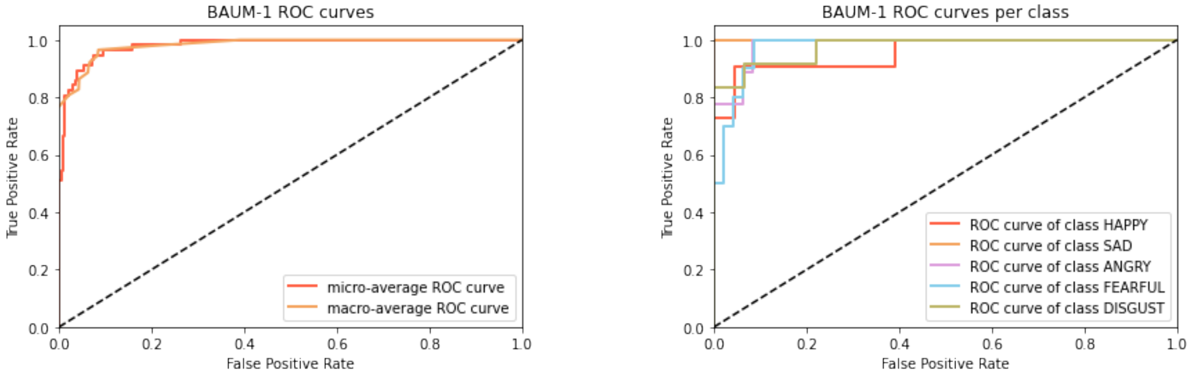


Figure 21: BAUM-1 confusion matrix

Figure 21 details which classes are the most difficult to identify for the multimodal model. In some cases, anger, fear, and disgust are misleadingly confused between them. Moreover, sometimes disgust and fear emotions are predicted as happiness. This fact is corroborated by the ROC curves below, where the red curve representing happiness is the lowest. Likewise, the decrease in performance when using the BAUM-1 dataset is seen by comparing the plots with the ones obtained with the RAVDESS dataset. The curves are no longer in the upper-left corner (ideal case).



(a) Averaged

(b) Per class

Figure 22: BAUM-1 ROC curves

The results obtained for the BAUM-1 results are worse than the ones in the previous work. The most significant decline is that of the voice model. Even though adding new speech features have resulted in a better performance for the RAVDESS dataset, in the case of the BAUM-1 dataset, it has not provided discriminant information to the model.

Another possible reason for this opposite effect on the BAUM-1 dataset is data augmentation. To compensate for the fact that the dataset contains fewer frames than the RAVDESS dataset, the number of generated images is higher. This fact may have caused the model to learn from redundant features, making it very difficult to generalize to other data.

	Previous work	**This work**	**Improvement (%)**
Face Image	0.86	0.82	-4.65
Optical flow	0.63	0.62	-1.59
Voice	0.94	0.81	-13.83
Face + Voice	0.94	0.84	-10.64

Table 14: BAUM-1 accuracy results comparison

4.3 Cross-dataset

It has been stated in the above sections the importance of having a model capable of generalizing and predicting emotions in different people. This section provides the results obtained when the models are trained and tested with different datasets.

First, Table 15 and Table 16 show that the cross-dataset results are much worse than the results previously explained. One thing to remark is that the face image model and the optical flow model work better when combined alone than when combined with the voice model.

	Precision		**Recall**		**F1-score**		**AUC**	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.49	0.55	0.48	0.53	0.45	0.53	0.73	0.78
Optical flow	0.47	0.44	0.38	0.40	0.36	0.40	0.67	0.76
Voice	0.28	0.24	0.23	0.21	0.23	0.21	0.58	0.55
Face	-	0.55	-	0.54	-	0.53	-	0.83
Face + Voice	-	0.49	-	0.51	-	0.49	-	0.81

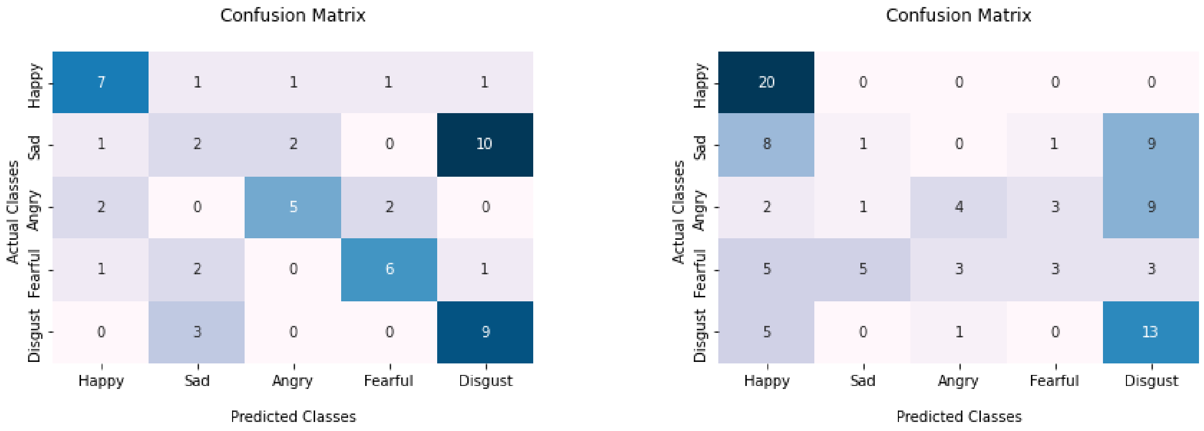
Table 15: RAVDESS - BAUM-1 averaged results

	Precision		**Recall**		**F1-score**		**AUC**	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.37	0.18	0.36	0.34	0.26	0.22	0.72	0.72
Optical flow	0.41	0.60	0.40	0.51	0.38	0.47	0.70	0.76
Voice	0.25	0.24	0.21	0.20	0.21	0.20	0.52	0.52
Face	-	0.57	-	0.40	-	0.32	-	0.74
Face + Voice	-	0.39	-	0.43	-	0.36	-	0.71

Table 16: BAUM-1 - RAVDESS averaged results

The voice model is not able to distinguish among the classes and acts nearly like a random classifier, especially when trained with the BAUM-1 dataset and tested with the RAVDESS dataset (Table 16). One possible cause is that this work tackles voice emotion recognition with a 1-D approach: the mean of the 2D features across the frequency dimension is computed and concatenated to one-dimensional features such as the Zero-Crossing Rate to form the feature vector. This process reduces the complexity and the computational effort with the cost of losing information. In addition, there is the fact that the language spoken in the two datasets is different (American English and Turkish), which may also be an extra source of errors.

Regarding the face image model, another explanation for the drop in performance is the difference in resolution between the datasets. Even though the cropped face images are resized, the CNN learns information about the resolution, which differs in each dataset.

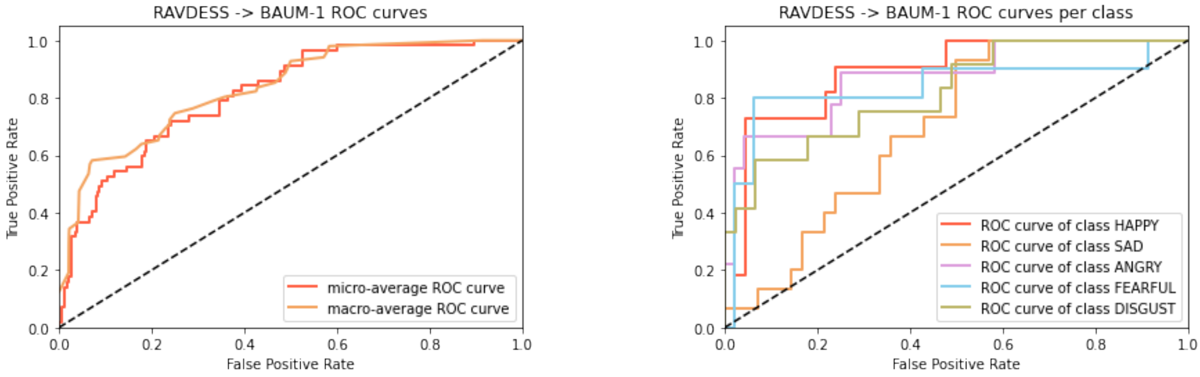


(a) RAVDESS - BAUM-1 confusion matrix

(b) BAUM-1 - RAVDESS confusion matrix

Figure 23: Cross-dataset confusion matrices

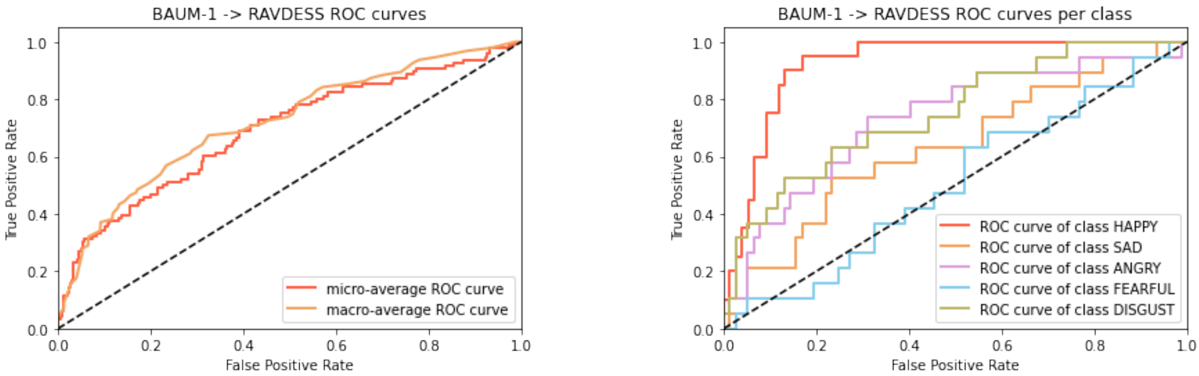
The confusion matrices above show the cross-dataset classification results. As previously seen in the tables, the final model has difficulties to classify the emotions when the dataset used for training and for testing is not the same.



(a) Averaged

(b) Per class

Figure 24: RAVDESS - BAUM-1 ROC curves



(a) Averaged

(b) Per class

Figure 25: BAUM-1 - RAVDESS ROC curves

On the one hand, when the model trains with the RAVDESS dataset (Figure 23a), sadness and disgust emotions are often confused between them. That is the main reason why their ROC curves are the lower ones. On the other hand, when the model is trained with BAUM-1, it tends to predict either happiness or disgust for all of the emotions (Figure 23b). In addition, for this second case, the classifier predicts the fear emotion wrongly as its ROC curve lies in the line with a unitary slope that divides the plot (random classifier).

	Previous work	**This work**	**Improvement (%)**
Face Image	0.39	0.53	35.90
Optical flow	0.22	0.44	100
Voice	0.21	0.21	0
Face + Voice	0.30	0.51	70

Table 17: RAVDESS - BAUM-1 accuracy results comparison

	Previous work	**This work**	**Improvement (%)**
Face Image	0.27	0.34	25.93
Optical flow	0.25	0.51	104
Voice	0.18	0.20	11.11
Face + Voice	0.28	0.43	53.57

Table 18: BAUM-1 - RAVDESS accuracy results comparison

Finally, the tables above show the comparison between the cross-dataset results of the previous work and the ones obtained during this project. Although there is an overall improvement in results, it varies depending on the model: the classifier with the best percentage of improvement is the optical flow model (with more than 100%), followed by the face image model and the voice model. The latter shows small improvement due to its poor performance on the BAUM-1 dataset.

The key to good cross-dataset performance lies primarily in minimizing the dissimilarities between datasets. The normalization applied to frames and speech signals has had a negative impact on the BAUM-1 dataset results, but these techniques have enhanced both the RAVDESS and the cross-dataset results. Nonetheless, it is worth mentioning that there is still much room for improvement.

5 Budget

The costs of this project are divided into human resources and the hardware needed. As all the programs used are open source, no fees are related to software. To calculate the budget, it is considered a project duration of 900 hours (equivalent to 30 ECTS).

On the one hand, the human resources cost includes the gross salary of a junior employee and social charges. A junior employee is paid 10€/h worked, which means a total wage of 9000€. The standard fees are 30% of the gross salary, which represents a cost of 2700€.

On the other hand, the hardware needed to develop the project is a computer with enough capacity to perform the training of the models. To calculate the amortisation, it is considered a computer price of 600€, a residual value of 60€, and a life span of 5 years. As a result, the first year's amortisation cost is 108€.

Concept	**Cost(€)**
Gross salary	9000
Social charges (30%)	2700
Computer amortization	108
TOTAL	**11808**

Table 19: Costs of the project

As can be seen in Table 19, the total cost of the project for one year is 11808€.

6 Conclusions and future development

This project has been developed with the intention of improving the cross-dataset results with respect to previous works when classifying emotions with a multimodal system based on facial expressions and voice. To this end, three models (face image, optical flow, and voice) have been trained and fused using score-fusion techniques to complement each other and create a robust final model.

After the testing phase, the developed system outperforms the previous related works. It can be concluded that the preprocessing techniques are the main reason of this improvement. Normalizing the frames reduces the differences between the datasets enhancing the cross-dataset performance. Moreover, chunking the audio files and using more voice features has resulted in an increment of discriminative information and, therefore, better classification results.

Despite the results being better, there is still a considerable difference between the performance of the models when they are trained and tested with the same dataset and when using cross-datasets. One of the possible causes is that, although there seem to exist similar patterns, there are still variations between individuals, especially when people are from different nationalities. However, the most challenging task is to unify the datasets into a common format. Further investigation into normalizing techniques must be done to deal with this issue. A future line of development can be to decrease the difference in the quality of both datasets, either by reducing the resolution of the dataset with better quality or by applying super-resolution techniques to enhance the worse quality dataset.

This project has focused on extracting information directly from the frames of the videos using Convolutional Neural Networks. An alternative that may also help with the normalization challenge is using a geometric approach. Most geometric feature-based approaches use the active appearance model (AAM) or its variations, to track a dense set of facial points. The locations of these facial landmarks are then used in different ways to extract the shape and the movement of facial features, as the expression evolves [34].

Finally, it is worth mentioning that the technologies used in this project do not only have a successful present but a promising future.

References

[1] P. Salovey and J. Mayer. Emotional Intelligence. Imagination, Cognition, and Personality. *Baywood Pub1ishlnl Co., Inc.*, 9(3):185–211, 1990.

[2] R. W. Picard. *Affective Computing.* MIT Press, Cambridge, Massachusetts, 1997.

[3] B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places.* Cambridge Univ. Press, Cambridge, Massachusetts, 1996.

[4] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, and L. Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. *Autonomous Agents and Multiagent Systems, International Joint Conference*, 2:194–201, 2004.

[5] K. Hone. Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with computers*, 18(2):227–245, 2006.

[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[7] P. Ekman. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994.

[8] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning.

[9] K. Mase. Recognition of facial expression from optical flow. *IEICE Trans.*, E74(10):3474–3483, 1991.

[10] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu. Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning. *IEEE Access*, 7:32297–32304, 2019.

[11] T. Seehapoch and S. Wongthanavasu. Speech emotion recognition using support vector machines. *5th International Conference on Knowledge and Smart Technology (KST)*, pages 86–91, 2013.

[12] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech*, 2014.

[13] A. Badshah, J. Ahmad, N. Rahim, and S. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. *International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017.

[14] N. Ronghe, S. Nakashe, A. Pawar, and S. Bobde. Emotion recognition and reaction prediction in videos. *Institute of Electrical and Electronics Engineers Inc.*, 2017.

[15] C. Luna-Jiménez, Z. Callejas, R. Kleinlein, F Fernández-Martínez, D. Griol, and Juan M. Montero. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. *Sensors*, 21(22), 2021.

[16] F. Krüger. Activity, context, and plan recognition with computational causal behaviour models., 2016.

[17] Receiver operating characteristic. [online]. `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`.

[18] SR Livingstone and FA 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(5), 2018.

[19] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Trans. on Affective Computing*, 2016.

[20] S. Tramonte. Multimodal emotion recognition with a 3-input CNN., 2021.

[21] Mediapipe. [online]. `https://mediapipe.dev/`.

[22] Opencv optical flow. [online]. `https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html`.

[23] Image augmentation for deep learning using keras and histogram equalization. [online]. `https://towardsdatascience.com/`.

[24] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015.

[25] Librosa. [online]. `https://librosa.org/doc/latest/index.html`.

[26] E. Ma. Data augmentation for audio. [online]. `https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6`.

[27] W. Lin and C. Busso. Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling. *IEEE Transactions on Affective Computing*, 2021.

[28] T. Zhang and C.-C. Jay Kuo. Audio Feature Analysis. In: Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing. *The Springer International Series in Engineering and Computer Science*, 606, 2001.

[29] D. Sharma. Analysis of Zero Crossing Rates of Different Music Genre Tracks. *Data Science Blogathon*, 2022.

[30] D. Ellis. Chroma feature analysis and synthesis. [online]. `https://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/`.

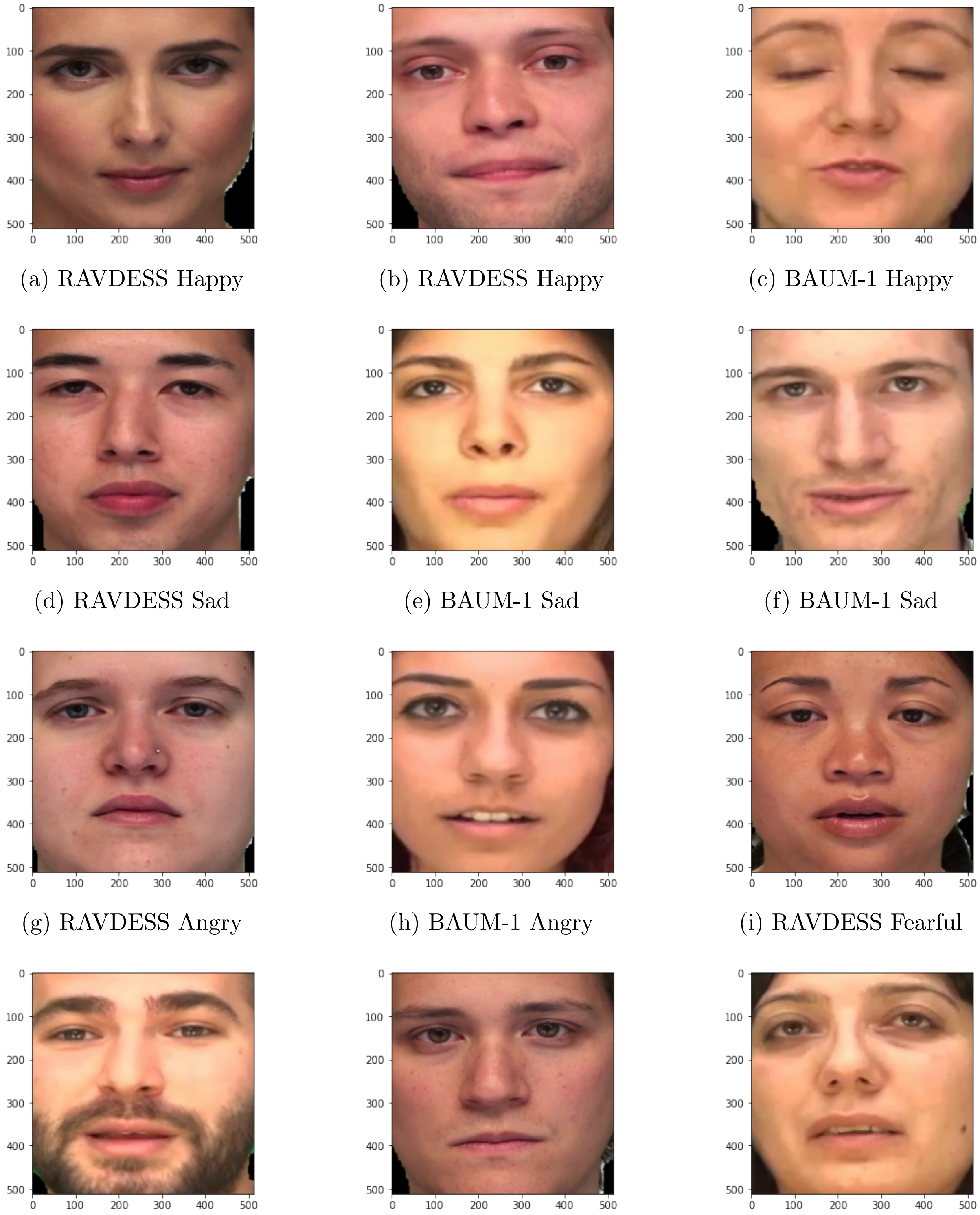
[31] M. Xu, L. Duan, J. Cai, L. Chia, C. Xu, and Q. Tian. HMM-Based Audio Keyword Generation. *Advances in Multimedia Information Processing*, 2004.

[32] A. K. Jain, R. Bolle, and S. Pankanti. Biometrics: Personal Identification in Networked Society. *Kluwer Academic Publishers*, 1999.

[33] R. Snelick, M. Indovina, J. Yen, and A. Mink. Multimodal Biometrics: Issues in Design and Testing. *Proceedings of the 5th International Conference on Multimodal Interfaces*, 2003.

[34] D. Ghimire and J. Lee. Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines. *Sensors*, 2013.

A Filtered frames examples



(a) RAVDESS Happy (b) RAVDESS Happy (c) BAUM-1 Happy

(d) RAVDESS Sad (e) BAUM-1 Sad (f) BAUM-1 Sad

(g) RAVDESS Angry (h) BAUM-1 Angry (i) RAVDESS Fearful

(j) BAUM-1 Fearful (k) RAVDESS Disgust (l) BAUM-1 Disgust

Figure 26: Filtered images using the CNN model

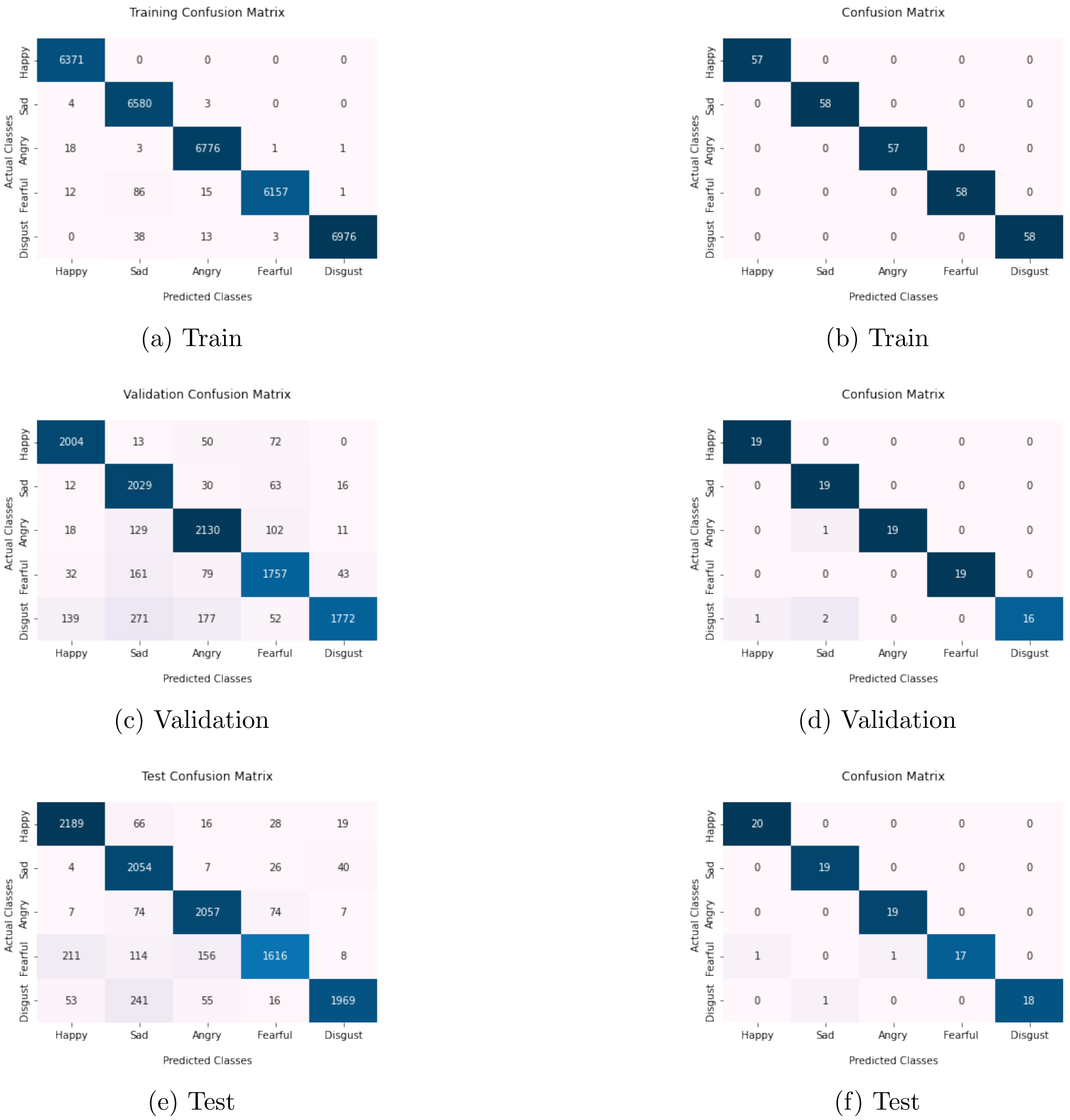
B Extended results

B.1 RAVDESS

Confusion matrices

Next, it is shown the train, validation, and test confusion matrices for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:



(a) Train

(b) Train

(c) Validation

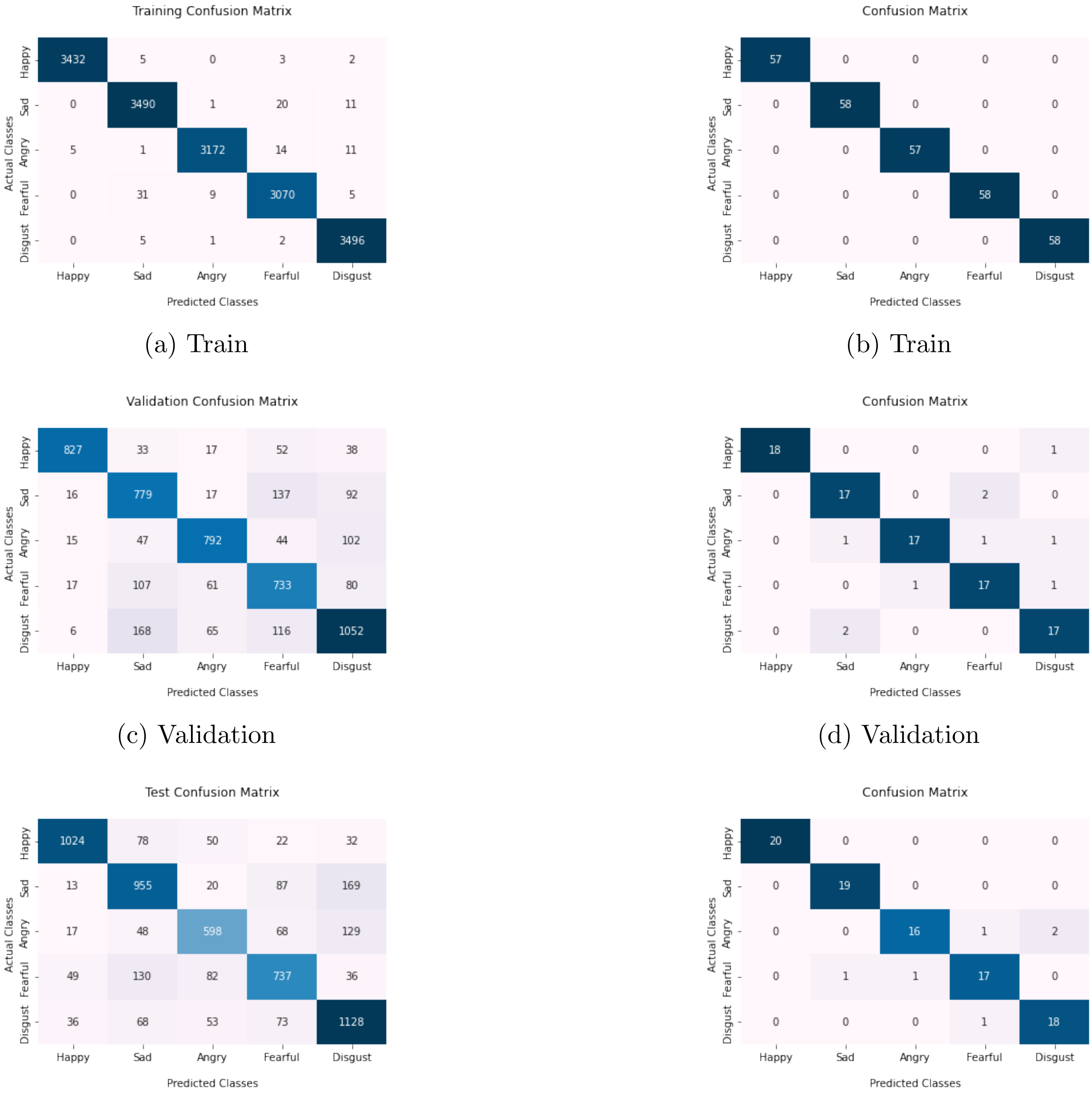
(d) Validation

(e) Test

(f) Test

Figure 27: RAVDESS face image model confusion matrices based on frames (a)(c)(e) and based on videos (b)(d)(f)

Optical flow model:



Training Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	3432	5	0	3	2
Sad	0	3490	1	20	11
Angry	5	1	3172	14	11
Fearful	0	31	9	3070	5
Disgust	0	5	1	2	3496

(a) Train

Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	57	0	0	0	0
Sad	0	58	0	0	0
Angry	0	0	57	0	0
Fearful	0	0	0	58	0
Disgust	0	0	0	0	58

(b) Train

Validation Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	827	33	17	52	38
Sad	16	779	17	137	92
Angry	15	47	792	44	102
Fearful	17	107	61	733	80
Disgust	6	168	65	116	1052

(c) Validation

Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	18	0	0	0	1
Sad	0	17	0	2	0
Angry	0	1	17	1	1
Fearful	0	0	1	17	1
Disgust	0	2	0	0	17

(d) Validation

Test Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	1024	78	50	22	32
Sad	13	955	20	87	169
Angry	17	48	598	68	129
Fearful	49	130	82	737	36
Disgust	36	68	53	73	1128

(e) Test

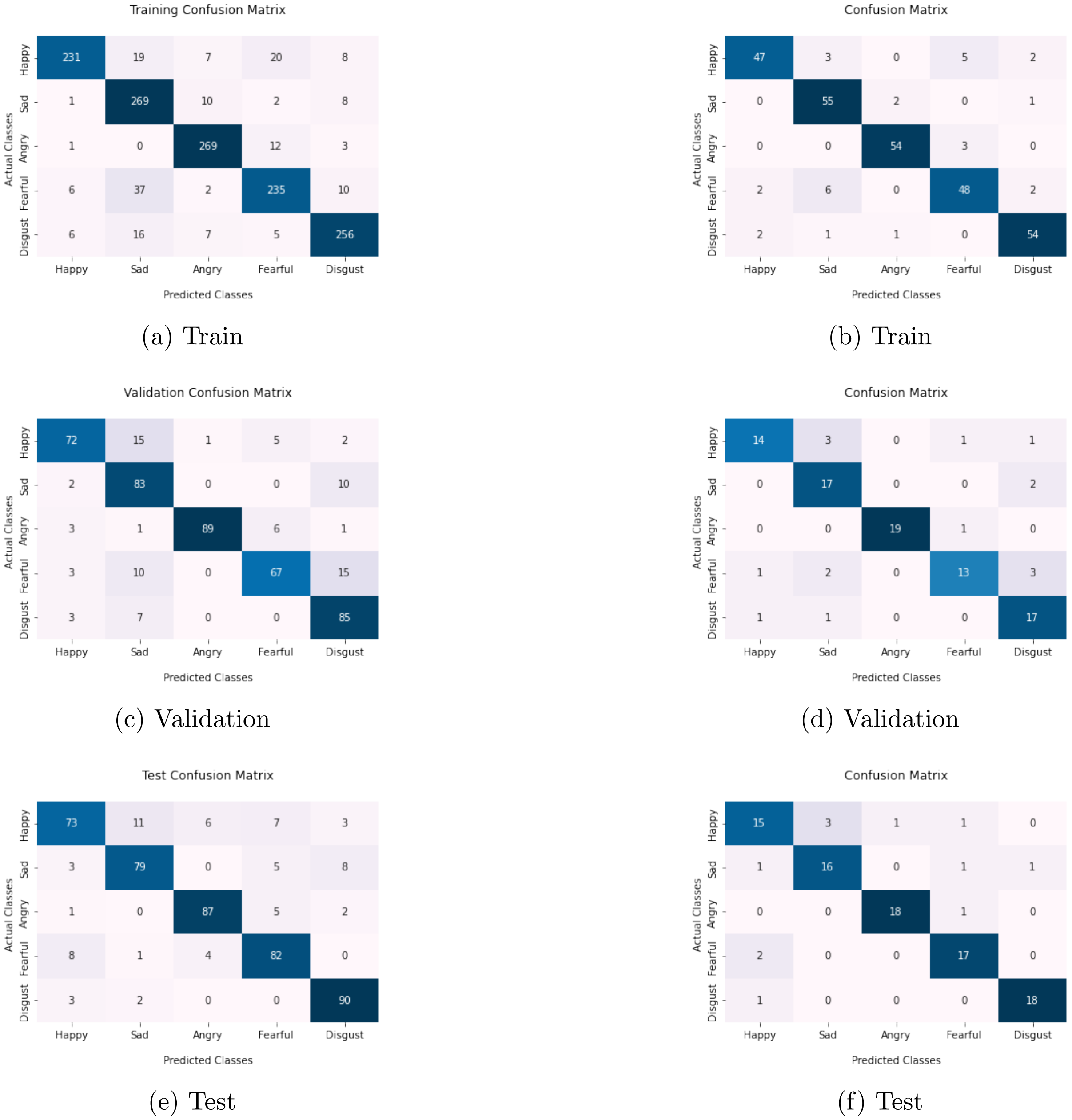
Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	20	0	0	0	0
Sad	0	19	0	0	0
Angry	0	0	16	1	2
Fearful	0	1	1	17	0
Disgust	0	0	0	1	18

(f) Test

Figure 28: RAVDESS optical flow model confusion matrices based on frames (a)(c)(e) and based on videos (b)(d)(f)

Voice model:



(a) Train

(b) Train

(c) Validation

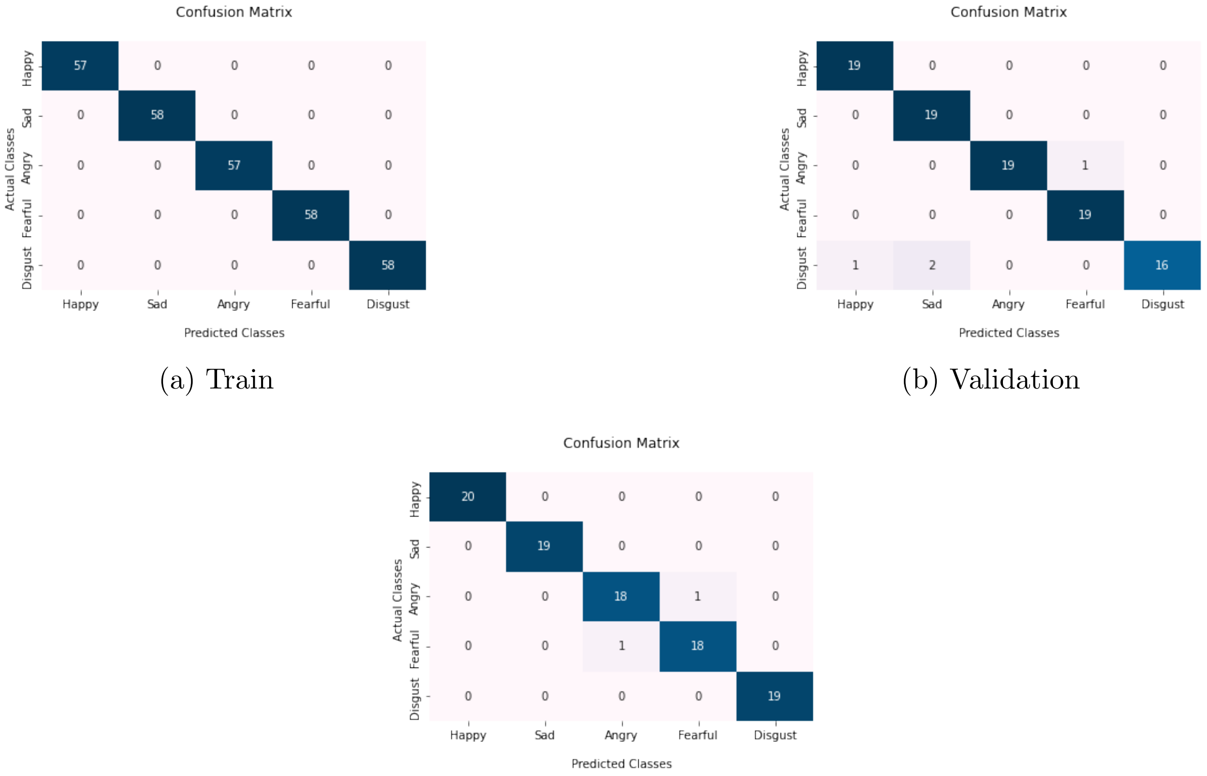
(d) Validation

(e) Test

(f) Test

Figure 29: RAVDESS voice model confusion matrices based on speech chunks (a)(c)(e) and based on videos (b)(d)(f)

Face model (Face image + optical flow):



(a) Train

(b) Validation

(c) Test

Figure 30: RAVDESS face multimodal model confusion matrices

Performance results

Next, it is shown the train, validation, and test performance results (precision, recall, F1-score, AUC) for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks (Ind.) and videos (Vid.).

Precision:

		Happy		Sad		Angry		Fearful		Disgust	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.91	0.95	0.78	0.86	0.86	1.00	0.86	1.00	0.91	1.00
	Test	0.89	0.95	0.81	0.95	0.90	0.95	0.92	1.00	0.96	1.00
Optical flow	Train	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	0.99	1.00
	Val	0.94	1.00	0.69	0.85	0.83	0.94	0.68	0.85	0.77	0.85
	Test	0.90	1.00	0.75	0.95	0.74	0.94	0.75	0.89	0.76	0.90
Voice	Train	0.94	0.92	0.79	0.85	0.91	0.95	0.86	0.86	0.90	0.92
	Val	0.87	0.88	0.72	0.74	0.99	1.00	0.86	0.87	0.75	0.74
	Test	0.83	0.79	0.85	0.84	0.90	0.95	0.83	0.85	0.87	0.95
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.95	-	0.90	-	1.00	-	0.95	-	1.00
	Test	-	1.00	-	1.00	-	0.95	-	0.95	-	1.00
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	0.90	-	1.00	-	1.00	-	1.00
	Test	-	1.00	-	1.00	-	1.00	-	1.00	-	0.95

Table 20: RAVDESS precision results

Recall:

		Happy		Sad		Angry		Fearful		Disgust	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00
	Val	0.94	1.00	0.94	1.00	0.89	0.95	0.85	1.00	0.73	0.84
	Test	0.94	1.00	0.96	1.00	0.93	1.00	0.77	0.89	0.84	0.95
Optical flow	Train	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
	Val	0.86	0.95	0.75	0.89	0.79	0.85	0.73	0.89	0.75	0.89
	Test	0.85	1.00	0.77	1.00	0.70	0.84	0.71	0.89	0.83	0.95
Voice	Train	0.81	0.92	0.93	0.85	0.94	0.95	0.81	0.86	0.88	0.92
	Val	0.76	0.88	0.87	0.74	0.89	1.00	0.71	0.87	0.89	0.74
	Test	0.73	0.79	0.83	0.84	0.92	0.95	0.86	0.85	0.95	0.95
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	1.00	-	0.95	-	1.00	-	0.84
	Test	-	1.00	-	1.00	-	0.95	-	0.95	-	1.00
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	1.00	-	1.00	-	1.00	-	0.89
	Test	-	1.00	-	0.95	-	1.00	-	1.00	-	1.00

Table 21: RAVDESS recall results

F1-score:

		Happy		Sad		Angry		Fearful		Disgust	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	Val	0.92	0.97	0.85	0.93	0.88	0.97	0.85	1.00	0.83	0.91
	Test	0.92	0.98	0.88	0.97	0.91	0.97	0.84	0.94	0.90	0.97
Optical flow	Train	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
	Val	0.90	0.97	0.72	0.87	0.81	0.89	0.70	0.87	0.76	0.87
	Test	0.87	1.00	0.76	0.97	0.72	0.89	0.73	0.89	0.79	0.92
Voice	Train	0.87	0.87	0.85	0.89	0.93	0.95	0.83	0.84	0.89	0.92
	Val	0.81	0.80	0.79	0.81	0.94	0.97	0.77	0.76	0.82	0.81
	Test	0.78	0.77	0.84	0.84	0.91	0.95	0.85	0.87	0.91	0.95
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.97	-	0.95	-	0.97	-	0.97	-	0.91
	Test	-	1.00	-	1.00	-	0.95	-	0.95	-	1.00
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	0.95	-	1.00	-	1.00	-	0.94
	Test	-	1.00	-	0.97	-	1.00	-	1.00	-	0.97

Table 22: RAVDESS F1-score results

AUC:

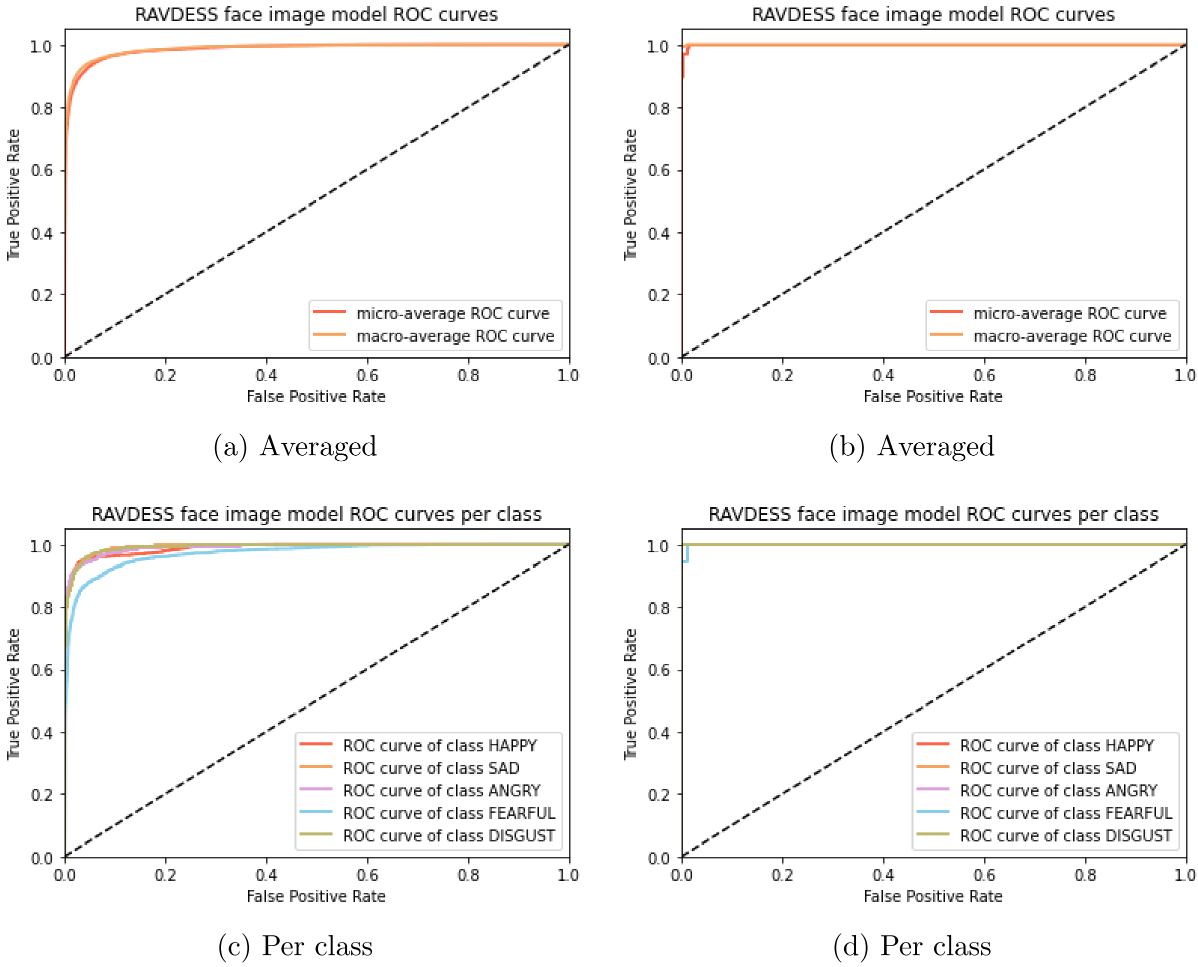
		Happy		Sad		Angry		Fearful		Disgust	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	0.95	0.99
	Test	0.99	1.00	0.99	1.00	0.99	1.00	0.97	1.00	0.99	1.00
Optical flow	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.98	1.00	0.94	0.99	0.97	1.00	0.93	0.98	0.93	0.99
	Test	0.98	1.00	0.94	1.00	0.94	0.99	0.94	0.99	0.95	0.99
Voice	Train	0.97	0.98	0.98	0.98	0.99	0.99	0.94	0.95	0.98	0.99
	Val	0.94	0.95	0.96	0.96	0.99	1.00	0.85	0.84	0.95	0.95
	Test	0.95	0.97	0.95	0.96	0.98	0.98	0.97	0.98	0.99	1.00
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	1.00	-	0.99	-	0.99	-	0.99
	Test	-	1.00	-	1.00	-	0.99	-	0.99	-	1.00
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	1.00	-	1.00	-	1.00	-	0.99
	Test	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00

Table 23: RAVDESS AUC results

ROC curves

Next, it is shown the weighted-average ROC curves and the ROC curves per class for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:



(a) Averaged

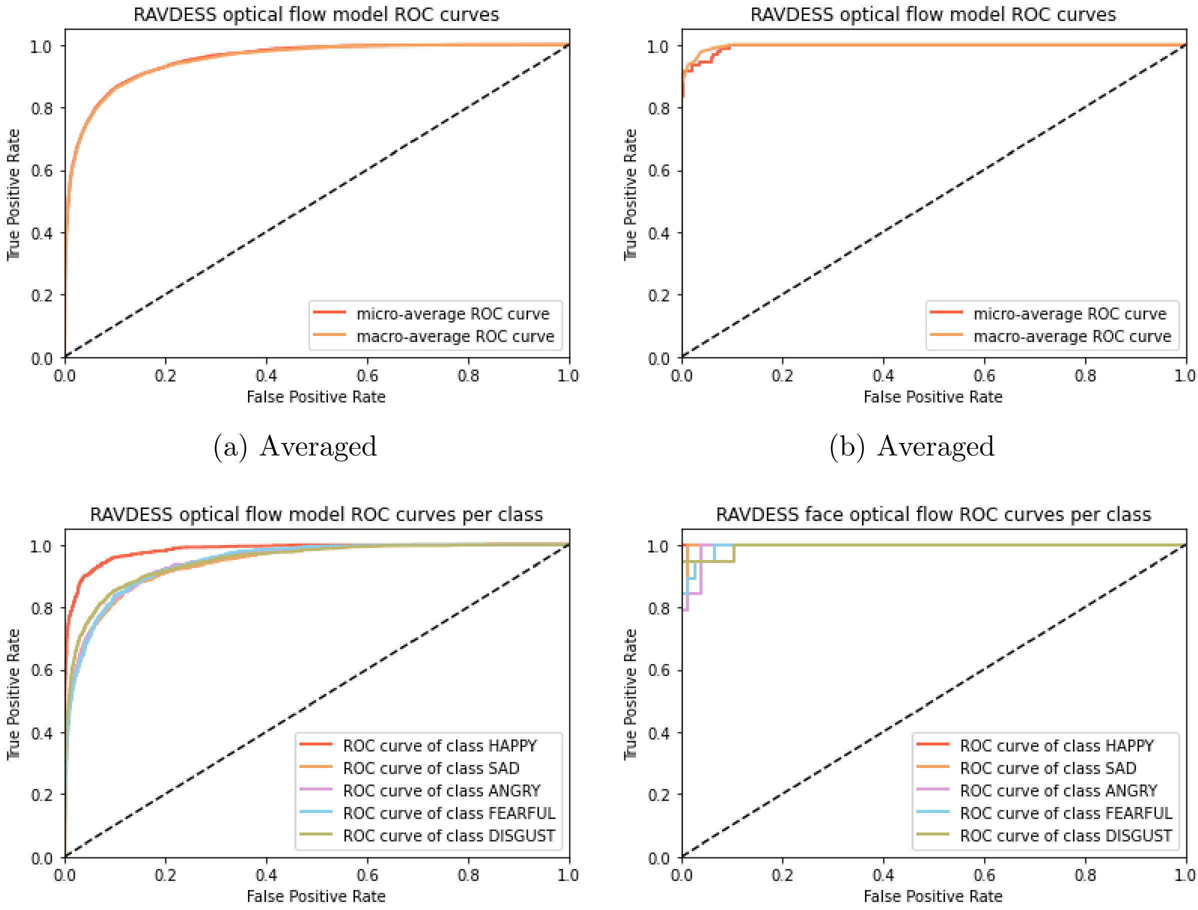
(b) Averaged

(c) Per class

(d) Per class

Figure 31: RAVDESS face image model ROC curves based on frames (a)(c) and based on videos (b)(d)

Optical flow model:



(a) Averaged

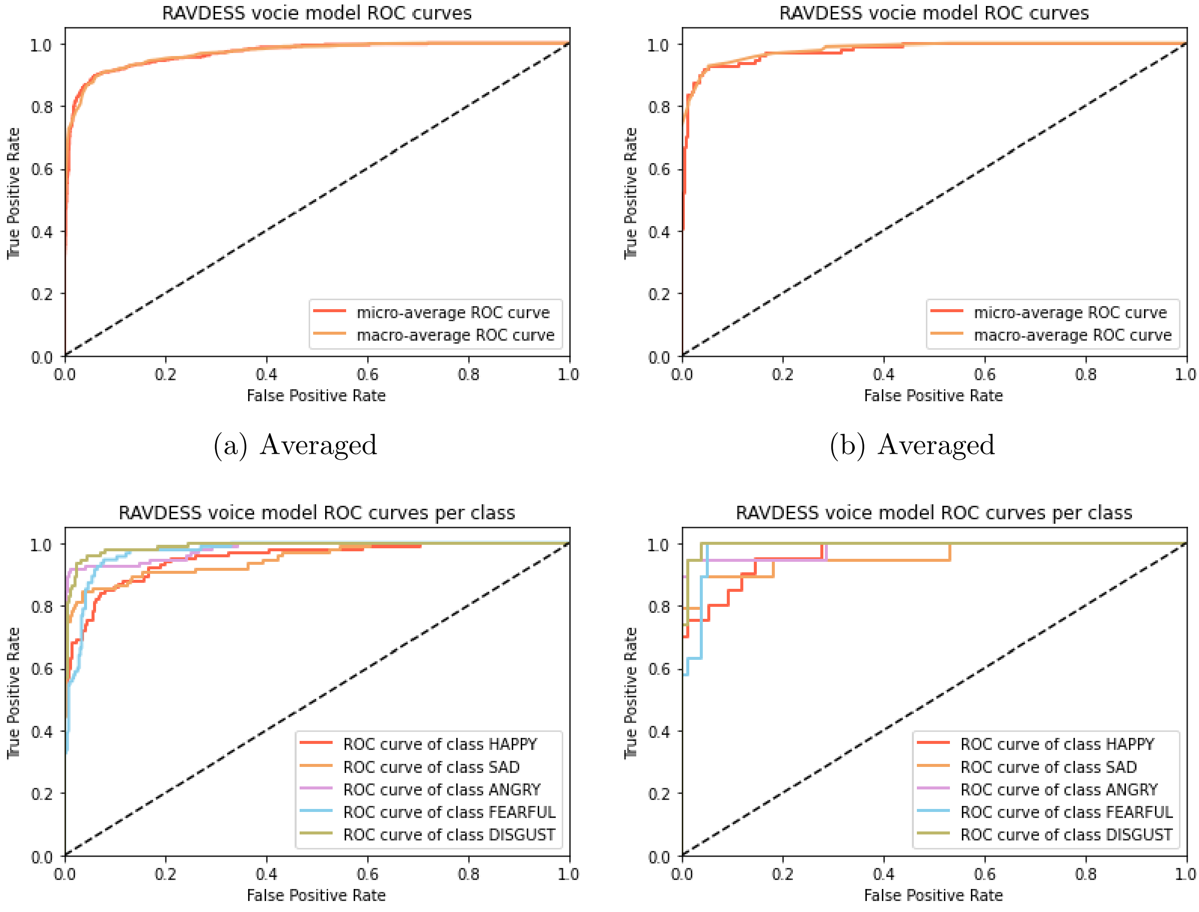
(b) Averaged

(c) Per class

(d) Per class

Figure 32: RAVDESS face optical flow ROC curves based on frames (a)(c) and based on videos (b)(d)

Voice model:



(a) Averaged

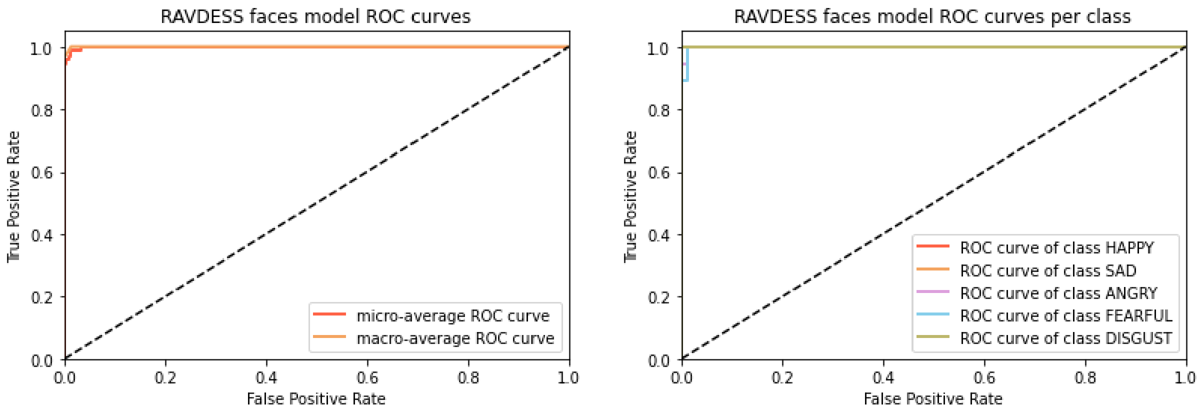
(b) Averaged

(c) Per class

(d) Per class

Figure 33: RAVDESS voice model ROC curves based on speech chunks (a)(c) and based on videos (b)(d)

Face model (Face image + optical flow):



(a) Averaged

(b) Per class

Figure 34: RAVDESS face multimodal model ROC curves based on frames (a) and based on videos (b)

Observations

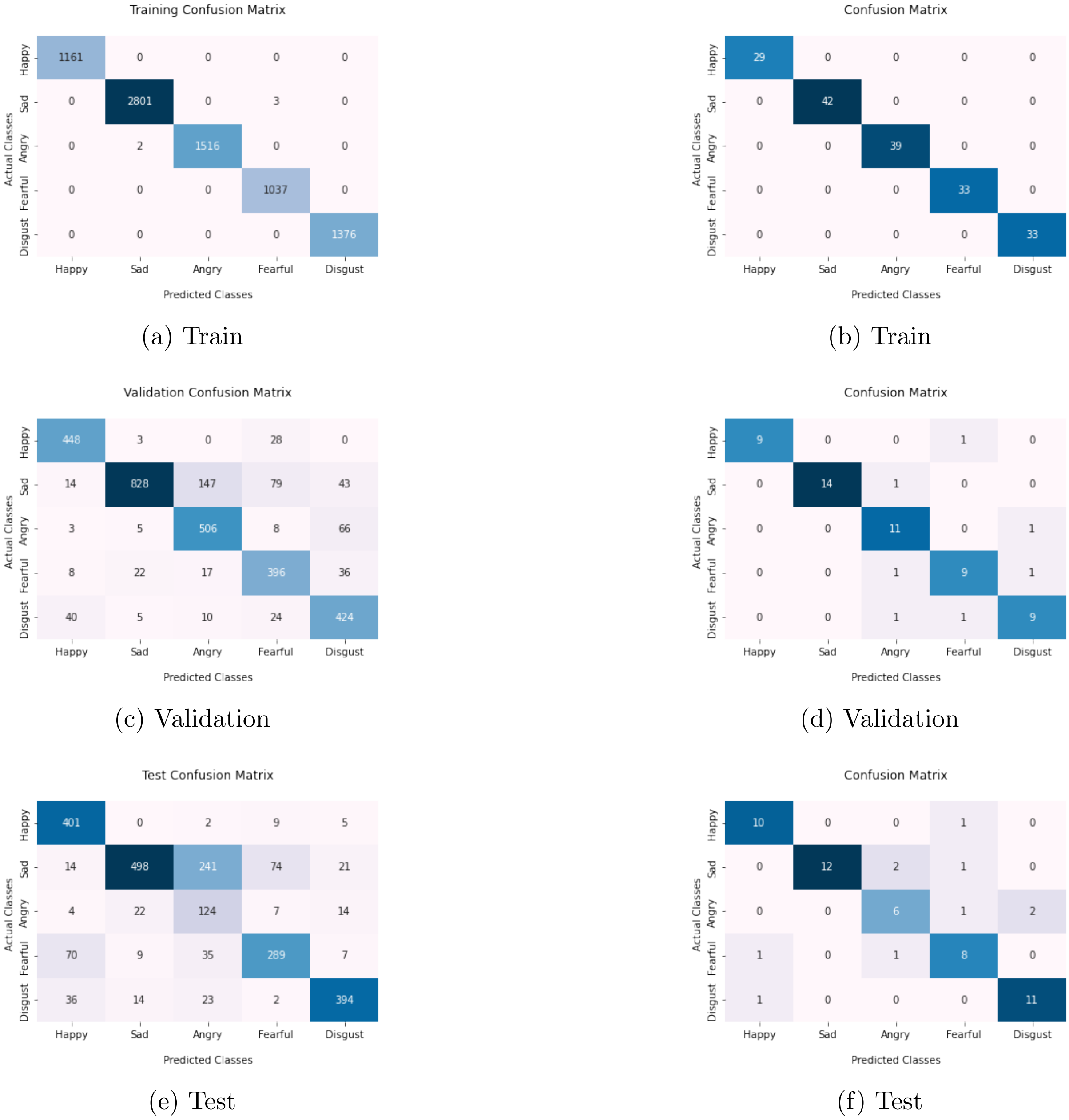
The results for the RAVDESS dataset are excellent. The face image and the optical flow model have shown outstanding performances, and when combined, the results are even better. The figures above show that using optical flow separately results in some confusion between sadness, anger, fear, and disgust. Moreover, it can be seen that in some cases, the voice model has misleadingly predicted sadness for videos representing happiness.

B.2 BAUM-1

Confusion matrices

Next, it is shown the train, validation, and test confusion matrices for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:



	Happy	Sad	Angry	Fearful	Disgust
Happy	1161	0	0	0	0
Sad	0	2801	0	3	0
Angry	0	2	1516	0	0
Fearful	0	0	0	1037	0
Disgust	0	0	0	0	1376

(a) Train

	Happy	Sad	Angry	Fearful	Disgust
Happy	29	0	0	0	0
Sad	0	42	0	0	0
Angry	0	0	39	0	0
Fearful	0	0	0	33	0
Disgust	0	0	0	0	33

(b) Train

	Happy	Sad	Angry	Fearful	Disgust
Happy	448	3	0	28	0
Sad	14	828	147	79	43
Angry	3	5	506	8	66
Fearful	8	22	17	396	36
Disgust	40	5	10	24	424

(c) Validation

	Happy	Sad	Angry	Fearful	Disgust
Happy	9	0	0	1	0
Sad	0	14	1	0	0
Angry	0	0	11	0	1
Fearful	0	0	1	9	1
Disgust	0	0	1	1	9

(d) Validation

	Happy	Sad	Angry	Fearful	Disgust
Happy	401	0	2	9	5
Sad	14	498	241	74	21
Angry	4	22	124	7	14
Fearful	70	9	35	289	7
Disgust	36	14	23	2	394

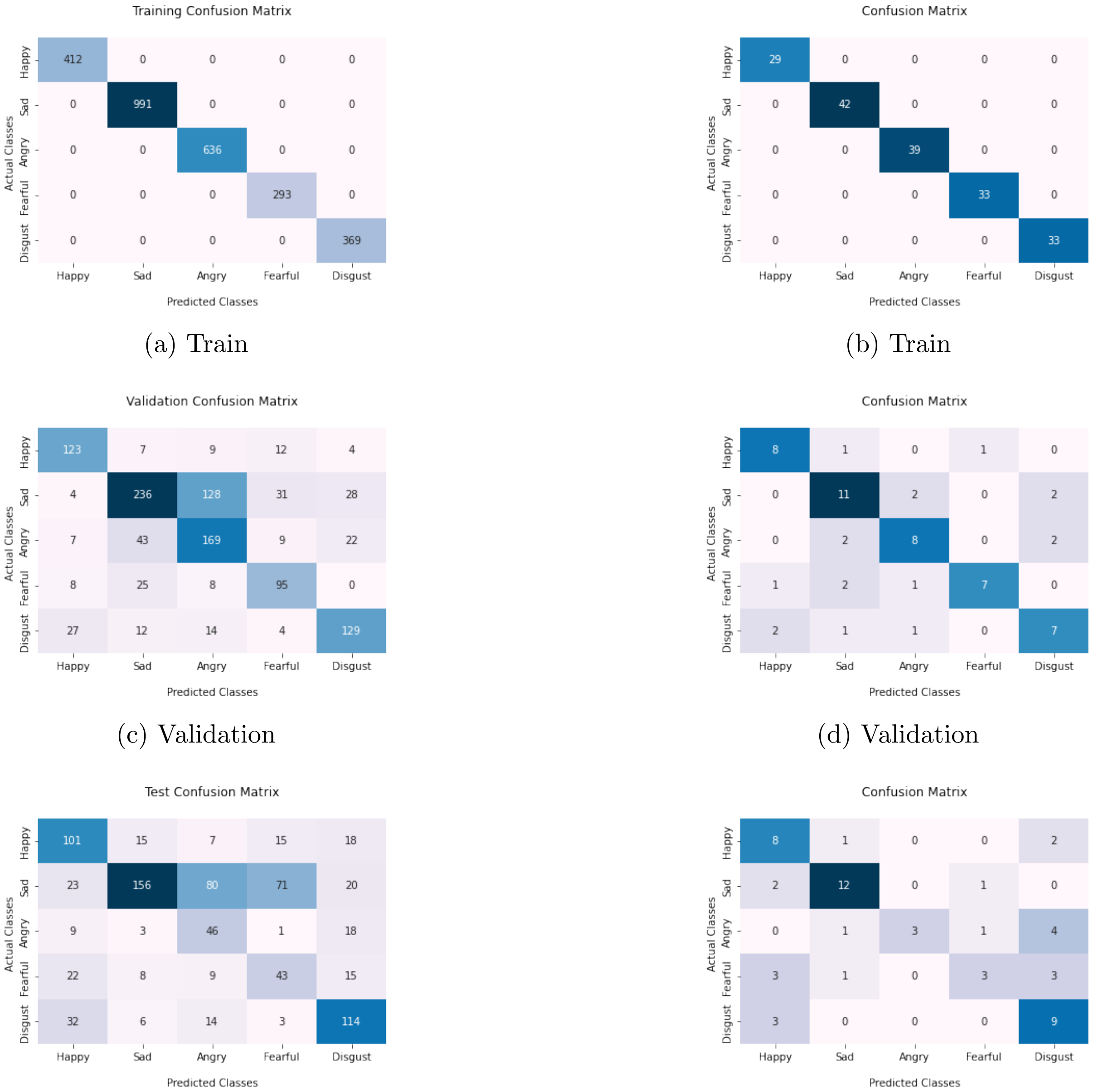
(e) Test

	Happy	Sad	Angry	Fearful	Disgust
Happy	10	0	0	1	0
Sad	0	12	2	1	0
Angry	0	0	6	1	2
Fearful	1	0	1	8	0
Disgust	1	0	0	0	11

(f) Test

Figure 35: BAUM-1 face image model confusion matrices based on frames (a)(c)(e) and based on videos (b)(d)(f)

Optical flow model:



(a) Train

(b) Train

(c) Validation

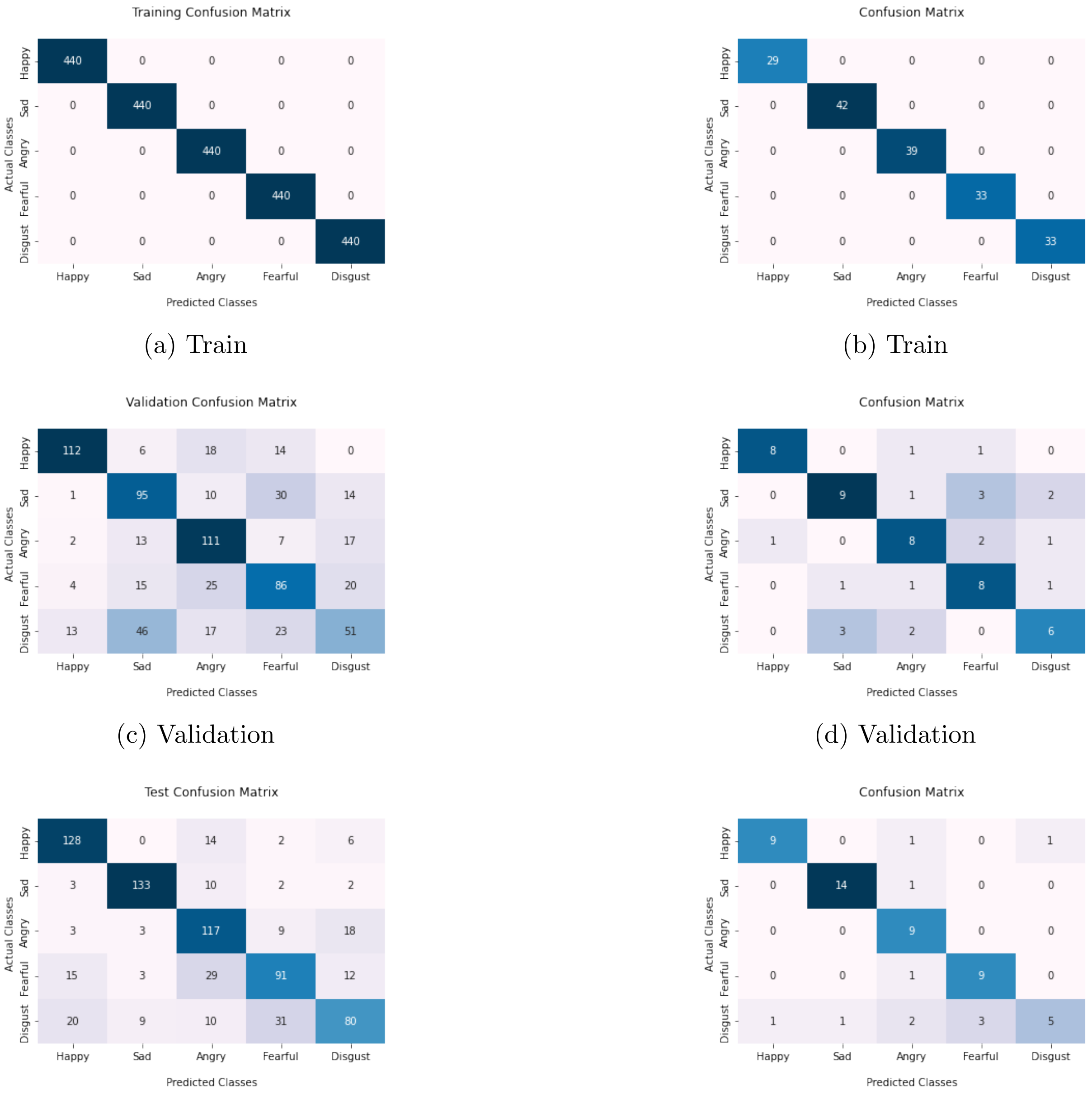
(d) Validation

(e) Test

(f) Test

Figure 36: BAUM-1 optical flow model confusion matrices based on frames (a)(c)(e) and based on videos (b)(d)(f)

Voice model:



Training Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	440	0	0	0	0
Sad	0	440	0	0	0
Angry	0	0	440	0	0
Fearful	0	0	0	440	0
Disgust	0	0	0	0	440

(a) Train

Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	29	0	0	0	0
Sad	0	42	0	0	0
Angry	0	0	39	0	0
Fearful	0	0	0	33	0
Disgust	0	0	0	0	33

(b) Train

Validation Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	112	6	18	14	0
Sad	1	95	10	30	14
Angry	2	13	111	7	17
Fearful	4	15	25	86	20
Disgust	13	46	17	23	51

(c) Validation

Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	8	0	1	1	0
Sad	0	9	1	3	2
Angry	1	0	8	2	1
Fearful	0	1	1	8	1
Disgust	0	3	2	0	6

(d) Validation

Test Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	128	0	14	2	6
Sad	3	133	10	2	2
Angry	3	3	117	9	18
Fearful	15	3	29	91	12
Disgust	20	9	10	31	80

(e) Test

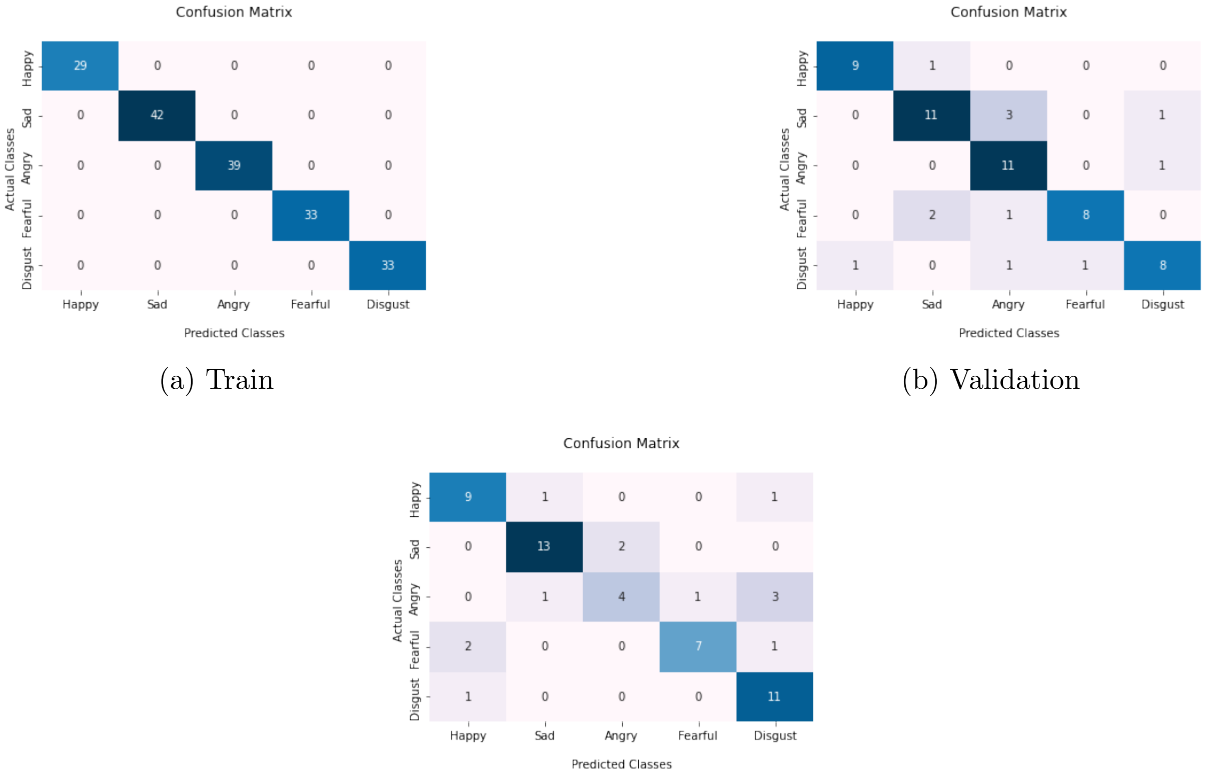
Confusion Matrix

	Happy	Sad	Angry	Fearful	Disgust
Happy	9	0	1	0	1
Sad	0	14	1	0	0
Angry	0	0	9	0	0
Fearful	0	0	1	9	0
Disgust	1	1	2	3	5

(f) Test

Figure 37: BAUM-1 voice model confusion matrices based on speech chunks (a)(c)(e) and based on videos (b)(d)(f)

Face model (Face image + optical flow):



(a) Train

(b) Validation

(c) Test

Figure 38: BAUM-1 face multimodal model confusion matrices

Performance results

Next, it is shown the train, validation, and test performance results (precision, recall, F1-score, AUC) for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks (Ind.) and videos (Vid.).

Precision:

		Happy		Sad		Angry		Fearful		Disgust	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.87	1.00	0.96	1.00	0.74	0.79	0.74	0.82	0.75	0.82
	Test	0.76	0.83	0.92	1.00	0.29	0.67	0.76	0.73	0.89	0.85
Optical flow	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.73	0.73	0.73	0.65	0.52	0.67	0.63	0.88	0.70	0.64
	Test	0.54	0.50	0.83	0.80	0.29	1.00	0.32	0.60	0.62	0.50
Voice	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.85	0.89	0.54	0.69	0.61	0.62	0.54	0.57	0.50	0.60
	Test	0.76	0.90	0.90	0.93	0.65	0.64	0.67	0.75	0.68	0.83
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.90	-	0.79	-	0.69	-	0.89	-	0.80
	Test	-	0.75	-	0.87	-	0.67	-	0.88	-	0.69
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.90	-	0.82	-	0.79	-	1.00	-	0.89
	Test	-	0.83	-	0.93	-	0.70	-	0.88	-	0.83

Table 24: BAUM-1 precision results

Recall:

		Happy		**Sad**		**Angry**		**Fearful**		**Disgust**	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.94	0.90	0.75	0.93	0.86	0.92	0.83	0.82	0.84	0.82
	Test	0.96	0.91	0.59	0.80	0.73	0.67	0.70	0.80	0.84	0.92
Optical flow	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.79	0.80	0.55	0.73	0.68	0.67	0.70	0.64	0.69	0.64
	Test	0.65	0.73	0.45	0.80	0.60	0.33	0.44	0.30	0.67	0.75
Voice	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.75	0.80	0.63	0.60	0.74	0.67	0.57	0.73	0.34	0.55
	Test	0.85	0.82	0.89	0.93	0.78	1.00	0.61	0.90	0.53	0.42
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.90	-	0.73	-	0.92	-	0.73	-	0.73
	Test	-	0.82	-	0.87	-	0.44	-	0.70	-	0.92
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.90	-	0.93	-	0.92	-	0.82	-	0.73
	Test	-	0.91	-	0.93	-	0.78	-	0.70	-	0.83

Table 25: BAUM-1 recall results

F1-score:

		Happy		**Sad**		**Angry**		**Fearful**		**Disgust**	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.90	0.95	0.84	0.97	0.80	0.85	0.78	0.82	0.79	0.82
	Test	0.85	0.87	0.72	0.89	0.42	0.67	0.73	0.76	0.87	0.88
Optical flow	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.76	0.76	0.63	0.69	0.58	0.67	0.66	0.74	0.70	0.64
	Test	0.59	0.59	0.58	0.80	0.39	0.50	0.37	0.40	0.64	0.60
Voice	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.79	0.84	0.58	0.64	0.67	0.64	0.55	0.64	0.40	0.57
	Test	0.80	0.86	0.89	0.93	0.71	0.78	0.64	0.82	0.60	0.58
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.90	-	0.76	-	0.79	-	0.80	-	0.76
	Test	-	0.78	-	0.87	-	0.53	-	0.78	-	0.79
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.90	-	0.87	-	0.85	-	0.90	-	0.80
	Test	-	0.87	-	0.93	-	0.74	-	0.78	-	0.83

Table 26: BAUM-1 F1-score results

AUC:

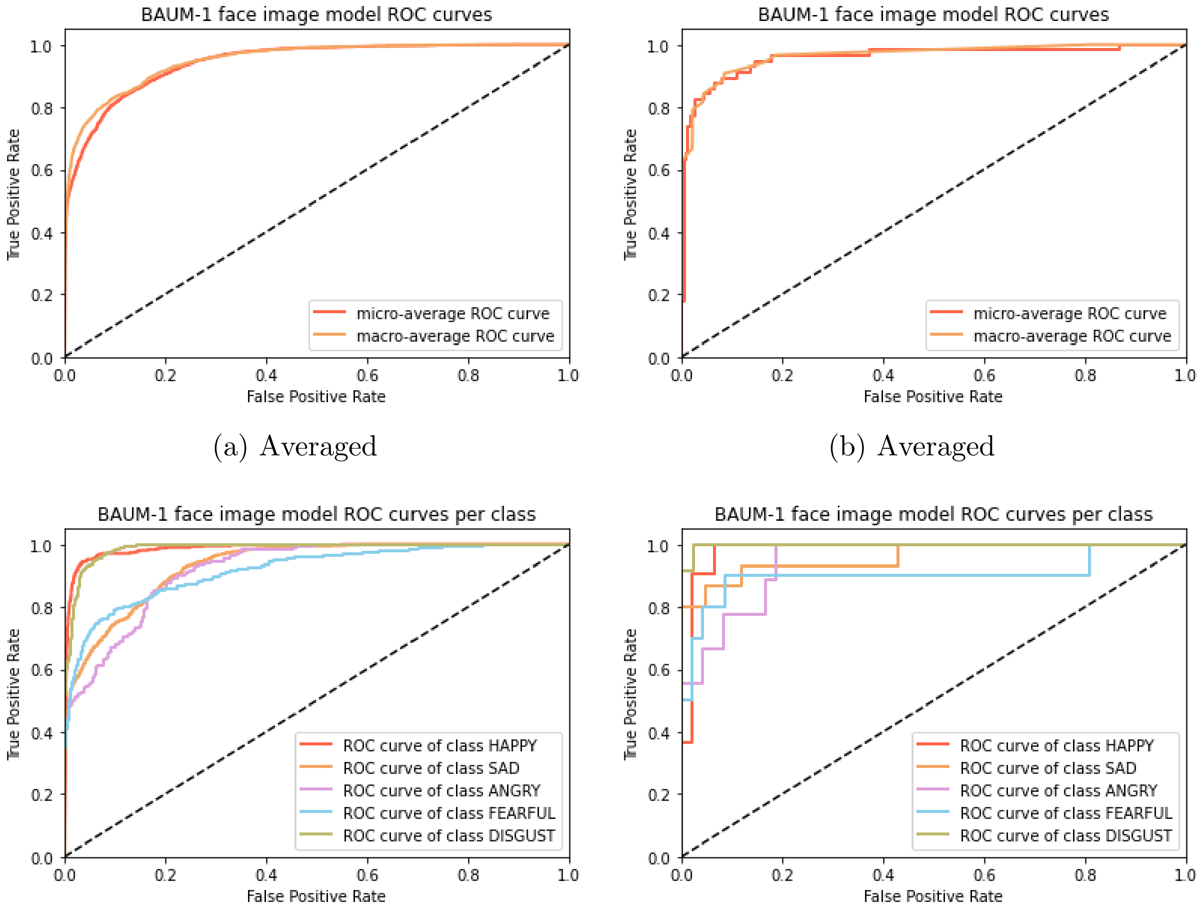
		Happy		**Sad**		**Angry**		**Fearful**		**Disgust**	
		Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	1.00	1.00	0.98	1.00	0.97	0.97	0.93	0.93	0.96	0.98
	Test	0.99	0.98	0.93	0.96	0.92	0.95	0.92	0.90	0.99	1.00
Optical flow	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.95	0.90	0.86	0.84	0.85	0.92	0.89	0.85	0.92	0.91
	Test	0.85	0.82	0.86	0.95	0.80	0.74	0.77	0.76	0.88	0.91
Voice	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Val	0.93	0.91	0.87	0.85	0.88	0.86	0.79	0.92	0.75	0.85
	Test	0.95	0.94	0.99	0.98	0.85	0.97	0.88	0.96	0.82	0.80
Face	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	0.99	-	0.97	-	0.96	-	0.92	-	0.96
	Test	-	0.95	-	0.98	-	0.91	-	0.88	-	0.97
Face + Voice	Train	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
	Val	-	1.00	-	0.98	-	0.98	-	1.00	-	0.97
	Test	-	0.96	-	1.00	-	0.98	-	0.98	-	0.98

Table 27: BAUM-1 AUC results

ROC curves

Next, it is shown the weighted-average ROC curves and the ROC curves per class for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:



(a) Averaged

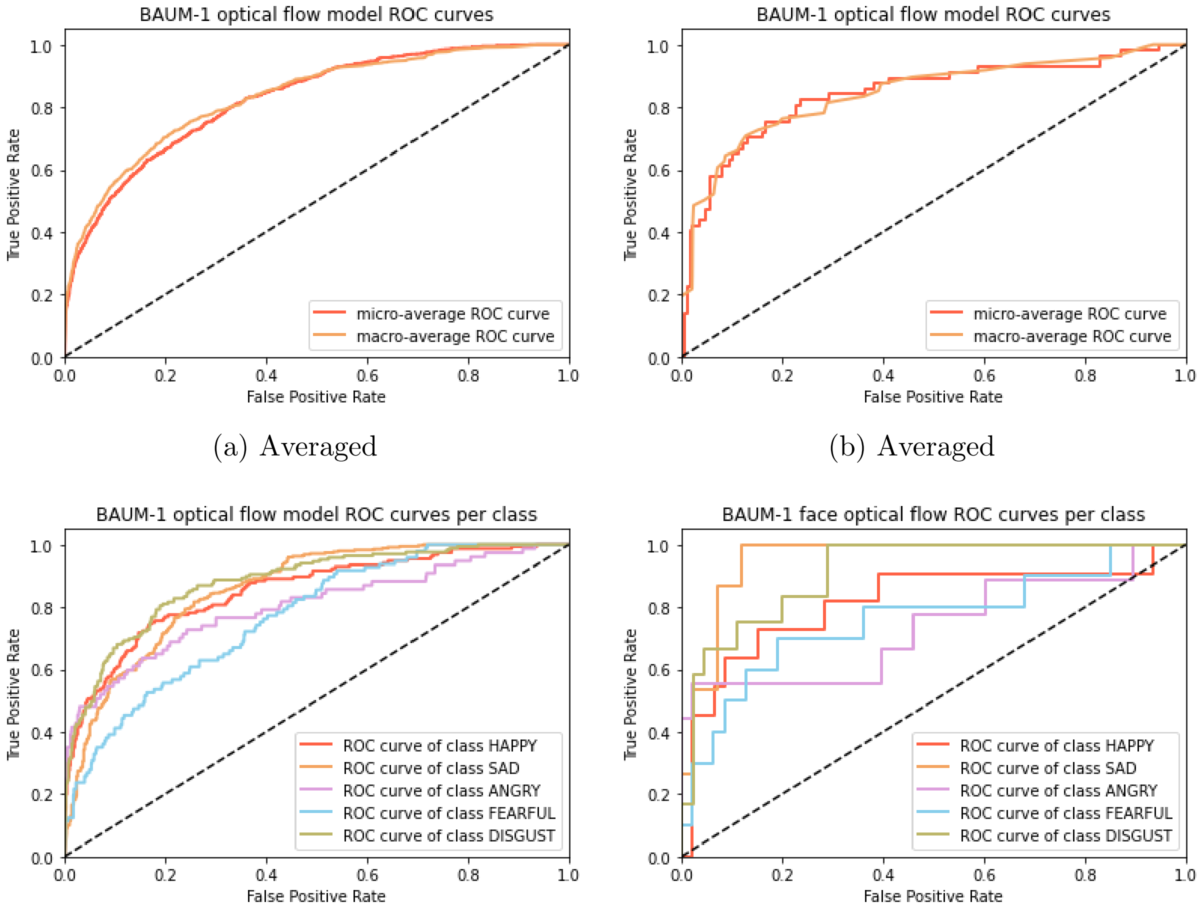
(b) Averaged

(c) Per class

(d) Per class

Figure 39: BAUM-1 face image model ROC curves based on frames (a)(c) and based on videos (b)(d)

Optical flow model:



(a) Averaged

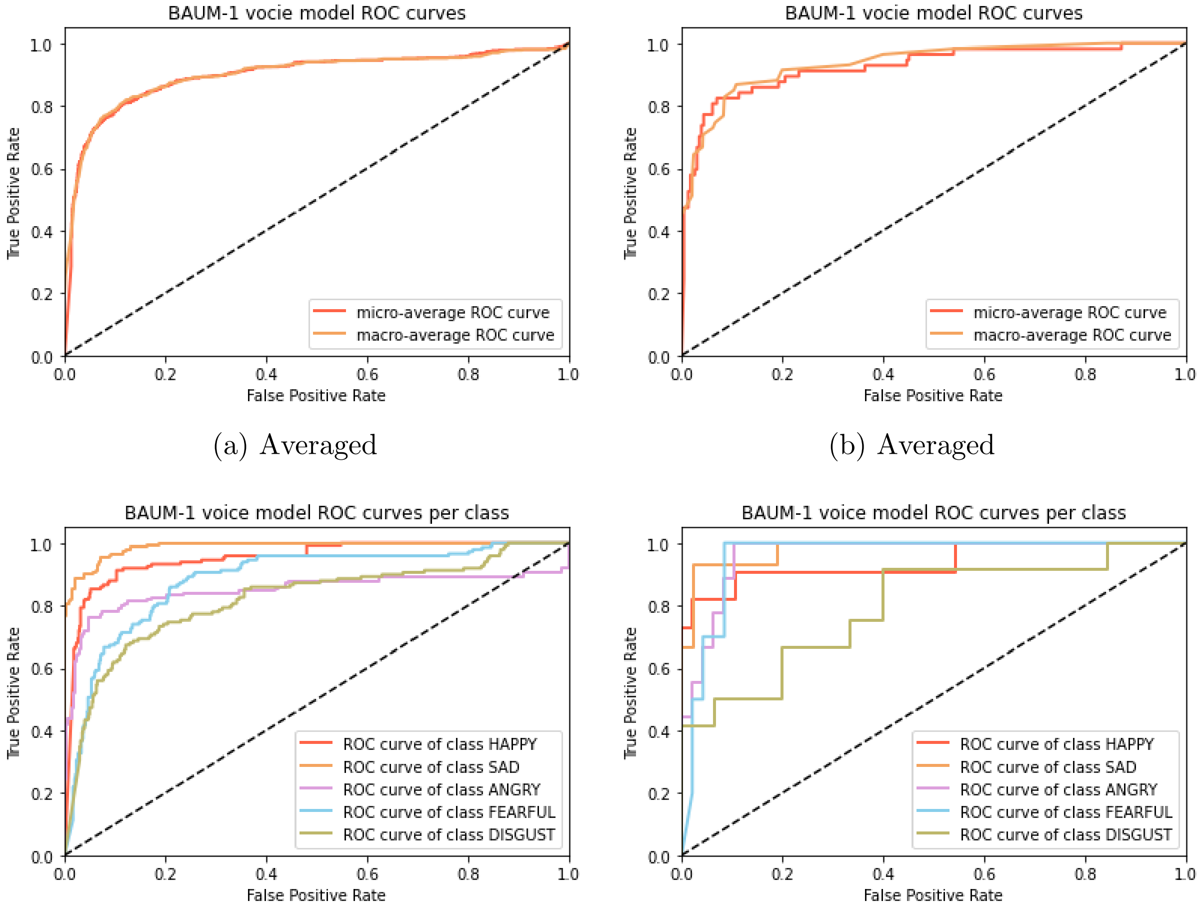
(b) Averaged

(c) Per class

(d) Per class

Figure 40: BAUM-1 face optical flow ROC curves based on frames (a)(c) and based on videos (b)(d)

Voice model:



(a) Averaged

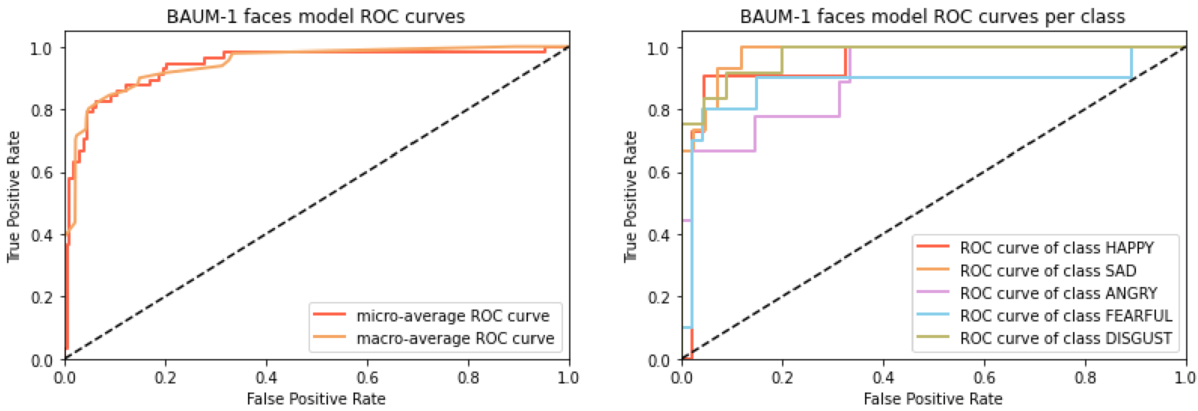
(b) Averaged

(c) Per class

(d) Per class

Figure 41: BAUM-1 voice model ROC curves based on speech chunks (a)(c) and based on videos (b)(d)

Face model (Face image + optical flow):



(a) Averaged

(b) Per class

Figure 42: BAUM-1 face multimodal model ROC curves based on frames (a) and based on videos (b)

Observations

Despite the results being worse than those obtained with the RAVDESS dataset, there is the same tendency to confuse sadness, anger, and fear. In addition, the voice model struggles when recognizing the disgust emotion.

B.3 Cross-datset

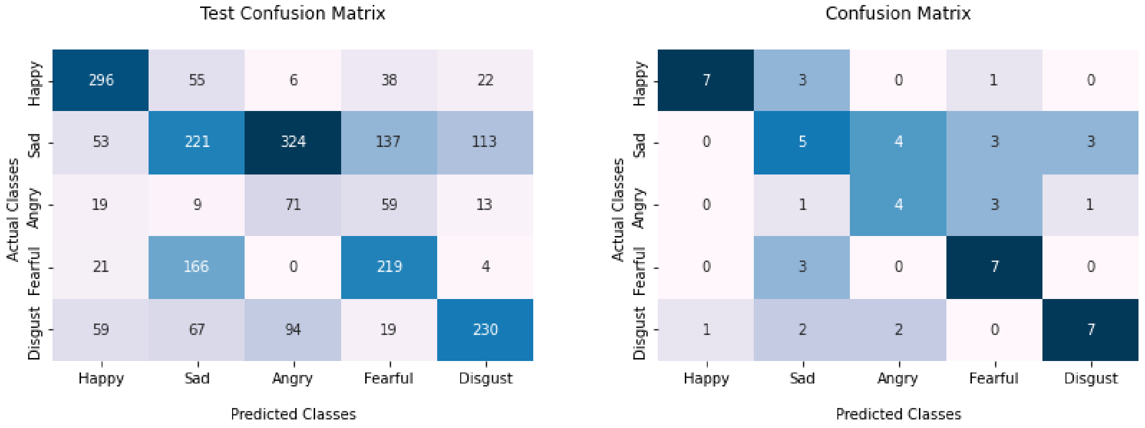
B.3.1 RAVDESS → BAUM-1

The results of this section are obtained by training the models on the RAVDESS dataset and testing on the BAUM-1 dataset.

Confusion matrices

Next, it is shown the test confusion matrices for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:

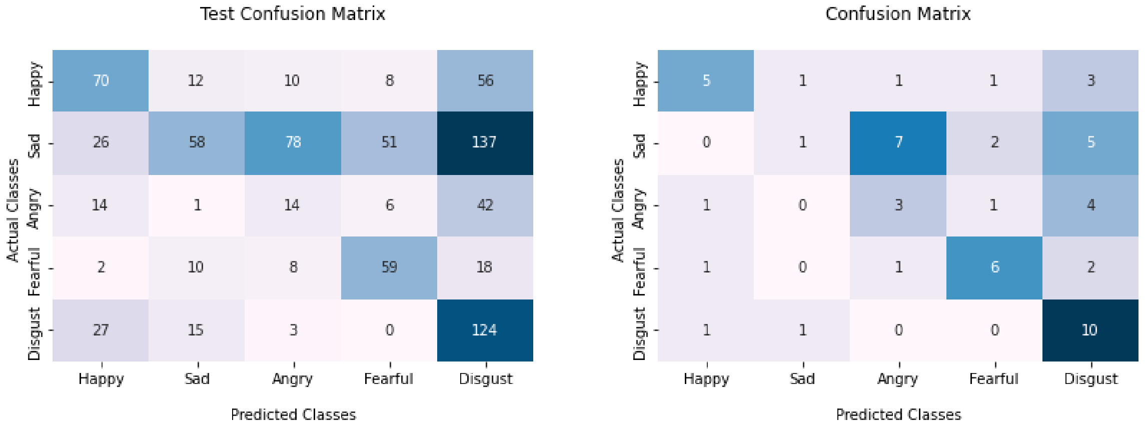


(a) Test

(b) Test

Figure 43: RAVDESS - BAUM-1 face image model confusion matrices based on frames (a) and based on videos (b)

Optical flow model:

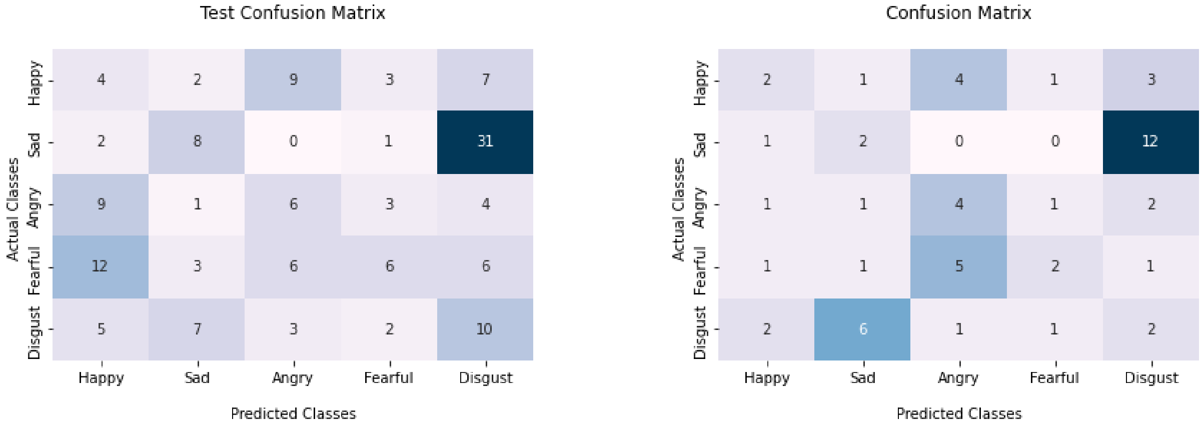


(a) Test

(b) Test

Figure 44: RAVDESS - BAUM-1 optical flow model confusion matrices based on frames (a) and based on videos (b)

Voice model:



(a) Test

(b) Test

Figure 45: RAVDESS - BAUM-1 voice model confusion matrices based on speech chunks (a) and based on videos (b)

Face model (Face image + optical flow):

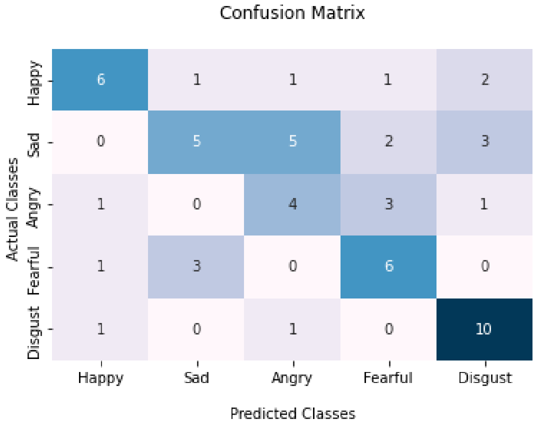


Figure 46: RAVDESS - BAUM-1 face multimodal model confusion matrices

Performance results

Next, it is shown the train, validation, and test performance results (precision, recall, F1-score, AUC) for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks (Ind.) and videos (Vid.).

Precision:

	Happy		Sad		Angry		Fearful		Disgust	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.66	0.88	0.43	0.36	0.14	0.40	0.46	0.50	0.60	0.64
Optical flow	0.50	0.62	0.60	0.33	0.12	0.25	0.48	0.60	0.33	0.42
Voice	0.12	0.29	0.38	0.18	0.25	0.29	0.40	0.40	0.17	0.10
Face	-	0.67	-	0.56	-	0.36	-	0.50	-	0.62
Face + Voice	-	0.64	-	0.25	-	0.62	-	0.67	-	0.43

Table 28: RAVDESS - BAUM-1 precision results

Recall:

	Happy		Sad		Angry		Fearful		Disgust	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.71	0.64	0.26	0.33	0.42	0.44	0.53	0.70	0.49	0.58
Optical flow	0.45	0.45	0.17	0.07	0.18	0.33	0.61	0.60	0.73	0.83
Voice	0.16	0.18	0.19	0.13	0.26	0.44	0.18	0.20	0.37	0.17
Face	-	0.55	-	0.33	-	0.44	-	0.60	-	0.83
Face + Voice	-	0.64	-	0.13	-	0.56	-	0.60	-	0.75

Table 29: RAVDESS - BAUM-1 recall results

F1-score:

	Happy		Sad		Angry		Fearful		Disgust	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.68	0.74	0.32	0.34	0.21	0.42	0.50	0.58	0.54	0.61
Optical flow	0.47	0.53	0.26	0.11	0.15	0.29	0.53	0.60	0.45	0.56
Voice	0.14	0.22	0.25	0.15	0.26	0.35	0.25	0.27	0.24	0.12
Face	-	0.60	-	0.42	-	0.40	-	0.55	-	0.71
Face + Voice	-	0.64	-	0.17	-	0.59	-	0.63	-	0.55

Table 30: RAVDESS - BAUM-1 F1-score results

AUC:

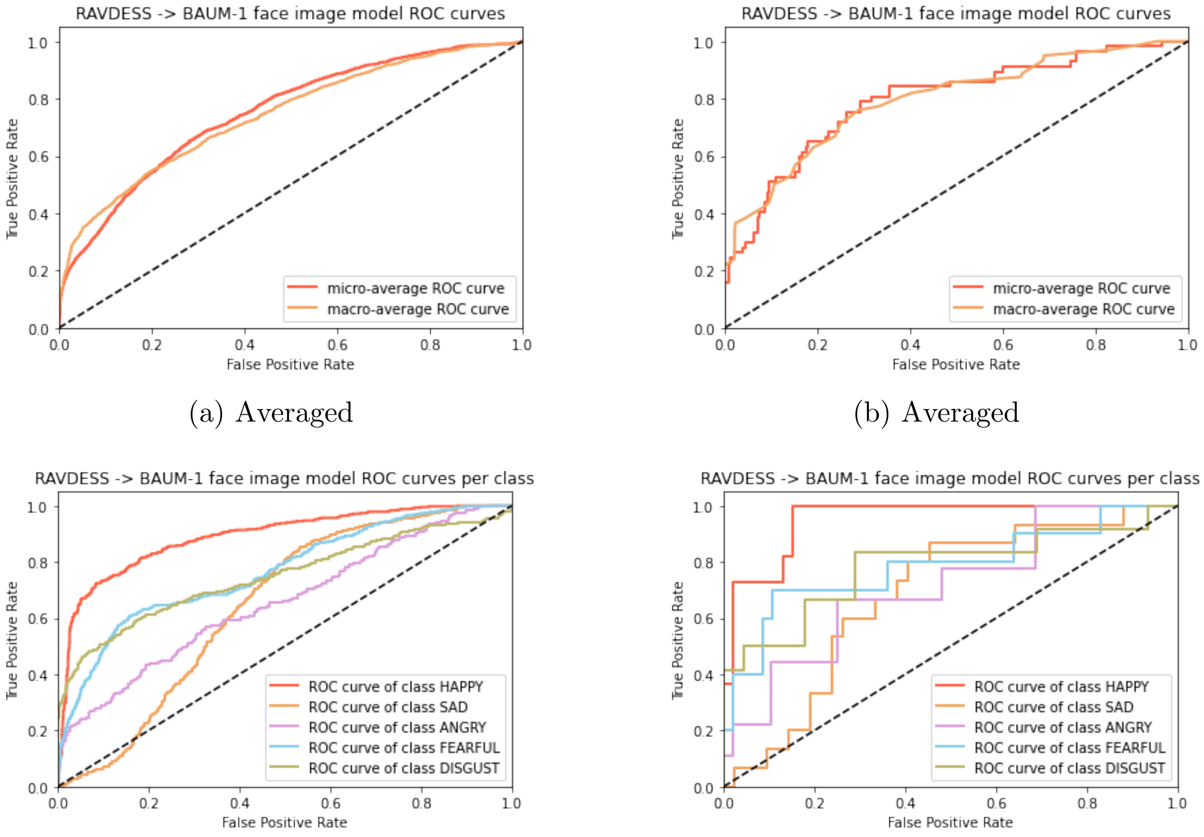
	Happy		Sad		Angry		Fearful		Disgust	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.89	0.95	0.65	0.69	0.66	0.71	0.77	0.79	0.75	0.78
Optical flow	0.66	0.66	0.61	0.68	0.54	0.72	0.88	0.86	0.74	0.91
Voice	0.62	0.59	0.52	0.52	0.68	0.75	0.63	0.45	0.48	0.51
Face	-	0.89	-	0.70	-	0.83	-	0.86	-	0.93
Face + Voice	-	0.89	-	0.70	-	0.87	-	0.84	-	0.82

Table 31: RAVDESS - BAUM-1 AUC results

ROC curves

Next, it is shown the weighted-average ROC curves and the ROC curves per class for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:



(a) Averaged

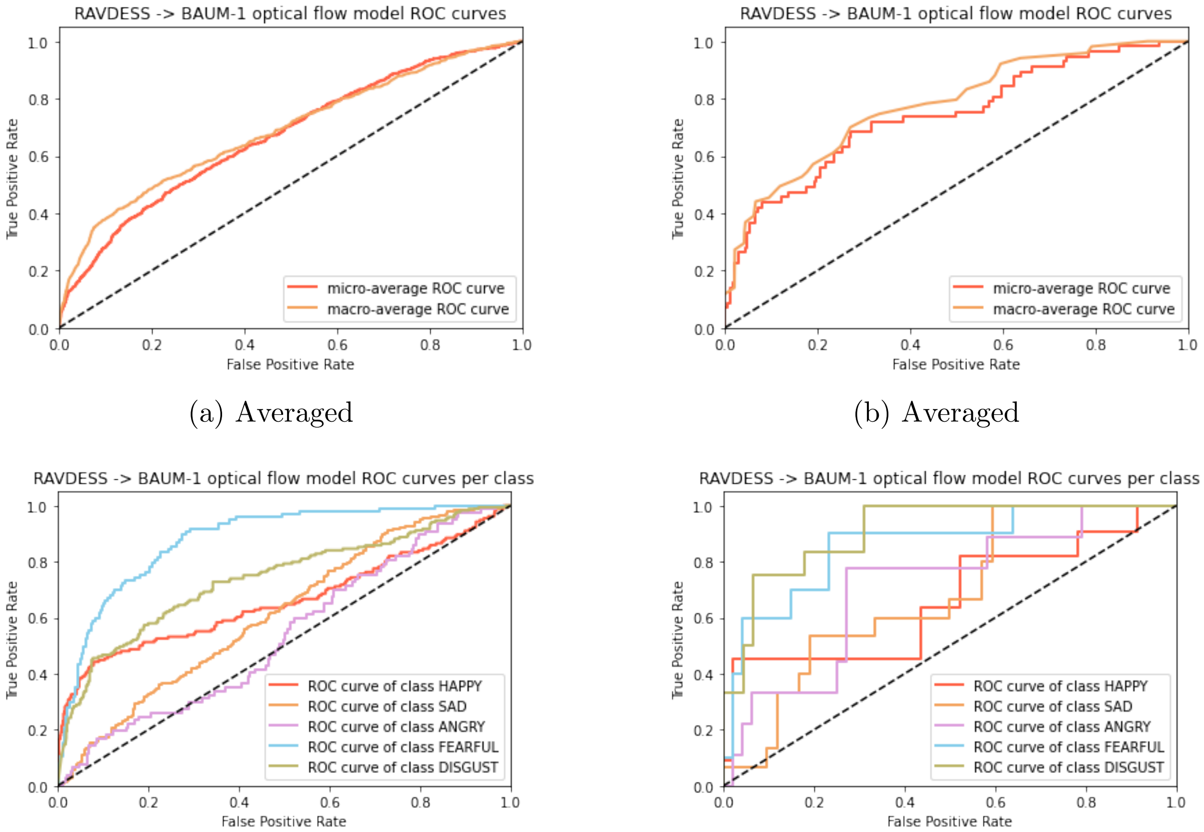
(b) Averaged

(c) Per class

(d) Per class

Figure 47: RAVDESS - BAUM-1 face image model ROC curves based on frames (a)(c) and based on videos (b)(d)

Optical flow model:



(a) Averaged

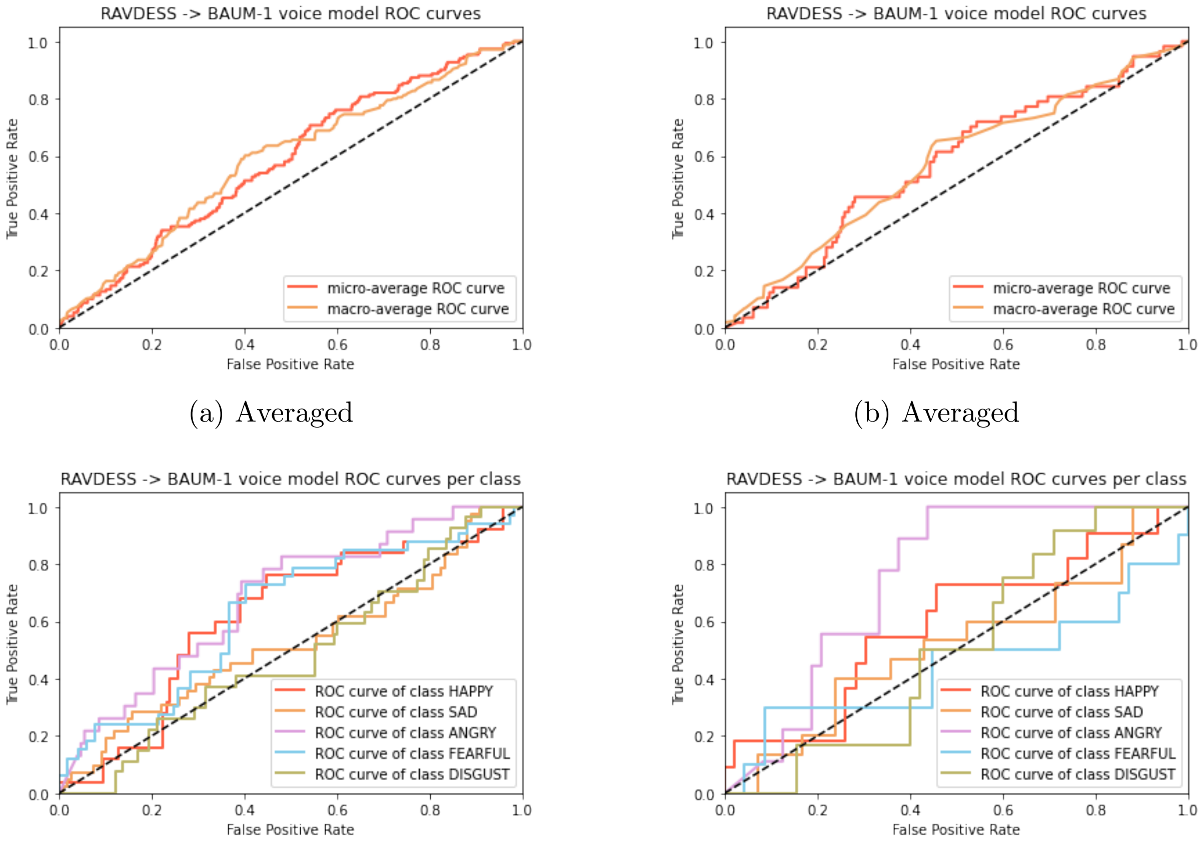
(b) Averaged

(c) Per class

(d) Per class

Figure 48: RAVDESS - BAUM-1 face optical flow ROC curves based on frames (a)(c) and based on videos (b)(d)

Voice model:



(a) Averaged

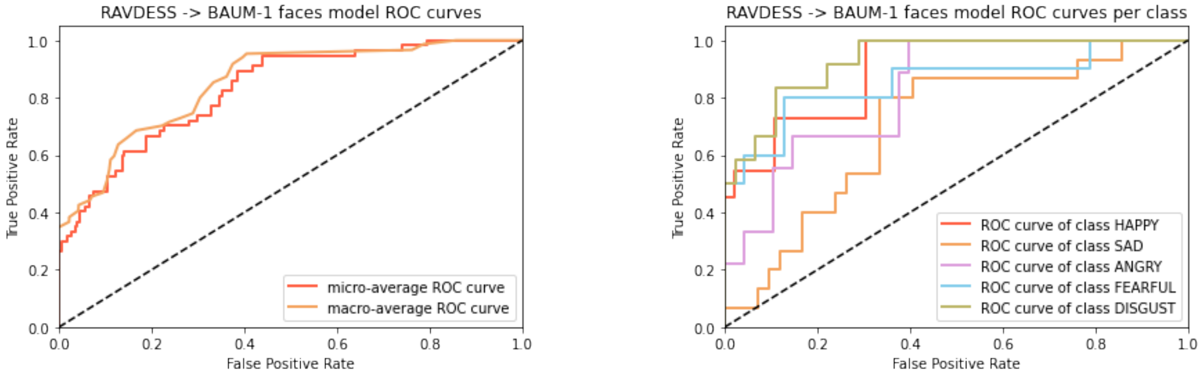
(b) Averaged

(c) Per class

(d) Per class

Figure 49: RAVDESS - BAUM-1 voice model ROC curves based on speech chunks (a)(c) and based on videos (b)(d)

Face model (Face image + Optical flow):



(a) Averaged

(b) Per class

Figure 50: RAVDESS - BAUM-1 face multimodal model ROC curves based on frames (a) and based on videos (b)

Observations

The figures above show that the cross-dataset results are much worse than the individual results of each dataset. The results highlight the effects of combining the Spatio-temporal information of the videos, as the face model outperforms both the face image model and the optical flow model. The voice model acts nearly as a random classifier in all the emotions except for the anger.

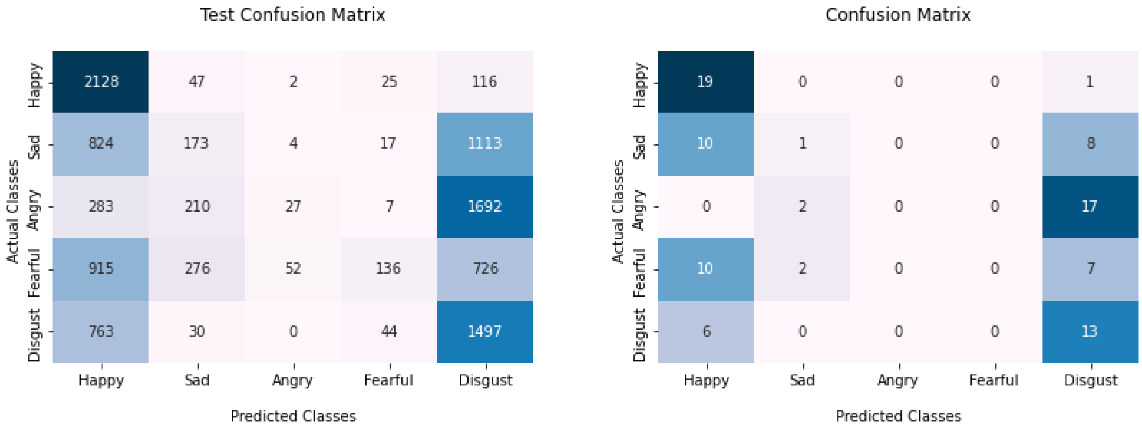
B.3.2 BAUM-1 → RAVDESS

The results of this section are obtained by training the models on the BAUM-1 dataset and testing on the RAVDESS dataset.

Confusion matrices

Next, it is shown the test confusion matrices for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

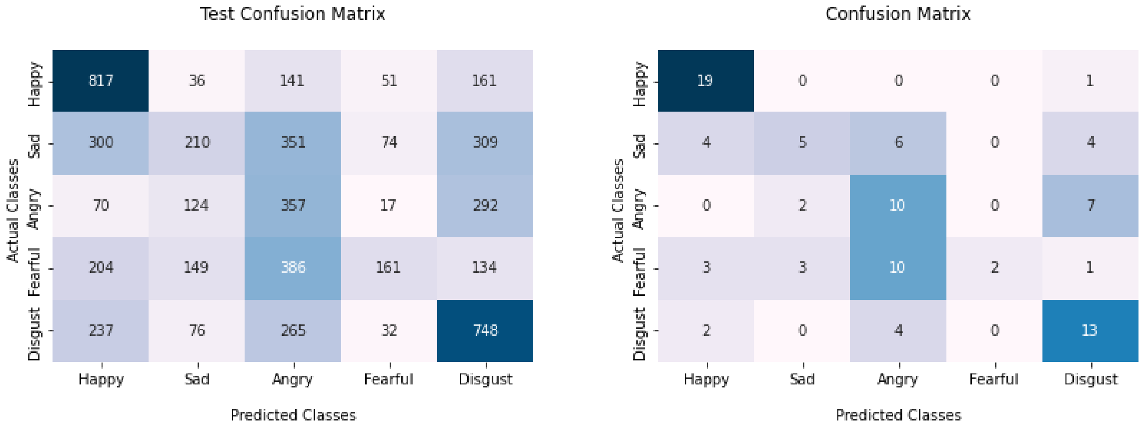
Face image model:



(a) Test (b) Test

Figure 51: BAUM-1 - RAVDESS face image model confusion matrices based on frames (a) and based on videos (b)

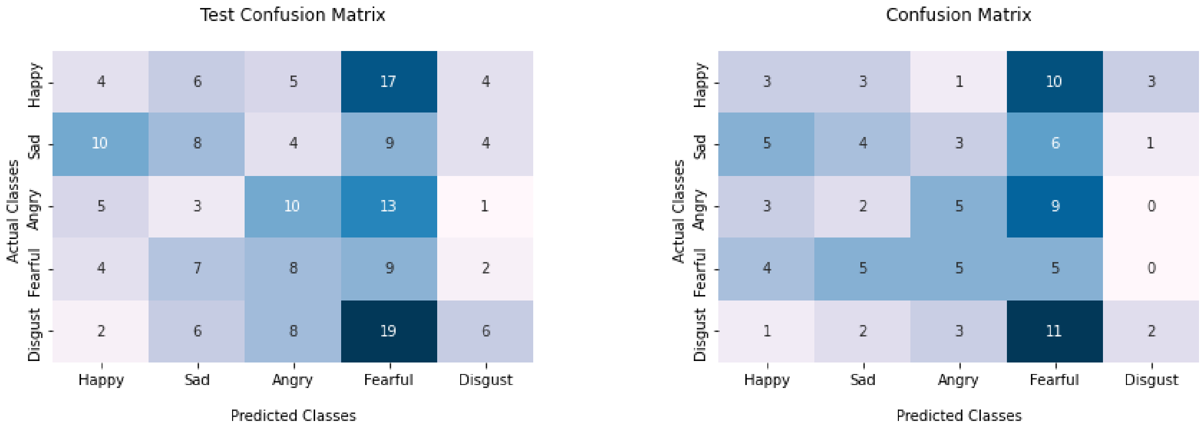
Optical flow model:



(a) Test (b) Test

Figure 52: BAUM-1 - RAVDESS optical flow model confusion matrices based on frames (a) and based on videos (b)

Voice model:



(a) Test (b) Test

Figure 53: BAUM-1 - RAVDESS voice model confusion matrices based on speech chunks (a) and based on videos (b)

Face model (Face image + Optical flow):

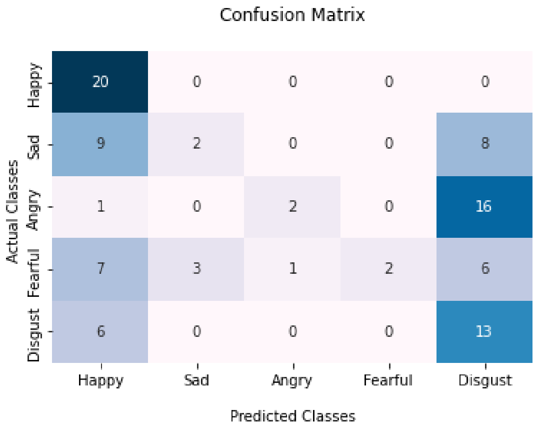


Figure 54: BAUM-1 - RAVDESS face multimodal model confusion matrices

Performance results

Next, it is shown the train, validation, and test performance results (precision, recall, F1-score, AUC) for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks (Ind.) and videos (Vid.).

Precision:

	Happy		Sad		Angry		Fearful		Disgust	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.43	0.42	0.24	0.20	0.32	0.00	0.59	0.00	0.29	0.28
Optical flow	0.50	0.68	0.35	0.50	0.24	0.33	0.48	1.00	0.45	0.50
Voice	0.16	0.19	0.27	0.25	0.29	0.29	0.13	0.12	0.35	0.33
Face	-	0.47	-	0.40	-	0.67	-	1.00	-	0.30
Face + Voice	-	0.50	-	0.14	-	0.50	-	0.43	-	0.38

Table 32: BAUM-1 - RAVDESS precision results

Recall:

	Happy		Sad		Angry		Fearful		Disgust	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.92	0.95	0.08	0.05	0.01	0.00	0.06	0.00	0.64	0.68
Optical flow	0.68	0.95	0.17	0.26	0.42	0.53	0.16	0.11	0.55	0.68
Voice	0.11	0.15	0.23	0.21	0.31	0.26	0.30	0.26	0.15	0.11
Face	-	1.00	-	0.11	-	0.11	-	0.11	-	0.68
Face + Voice	-	1.00	-	0.05	-	0.21	-	0.16	-	0.68

Table 33: BAUM-1 - RAVDESS recall results

F1-score:

	Happy		**Sad**		**Angry**		**Fearful**		**Disgust**	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.59	0.58	0.12	0.08	0.02	0.00	0.12	0.00	0.40	0.40
Optical flow	0.58	0.79	0.23	0.34	0.30	0.41	0.24	0.19	0.50	0.58
Voice	0.13	0.17	0.25	0.23	0.30	0.28	0.19	0.17	0.21	0.16
Face	-	0.63	-	0.17	-	0.18	-	0.19	-	0.42
Face + Voice	-	0.67	-	0.08	-	0.30	-	0.23	-	0.49

Table 34: BAUM-1 - RAVDESS F1-score results

AUC:

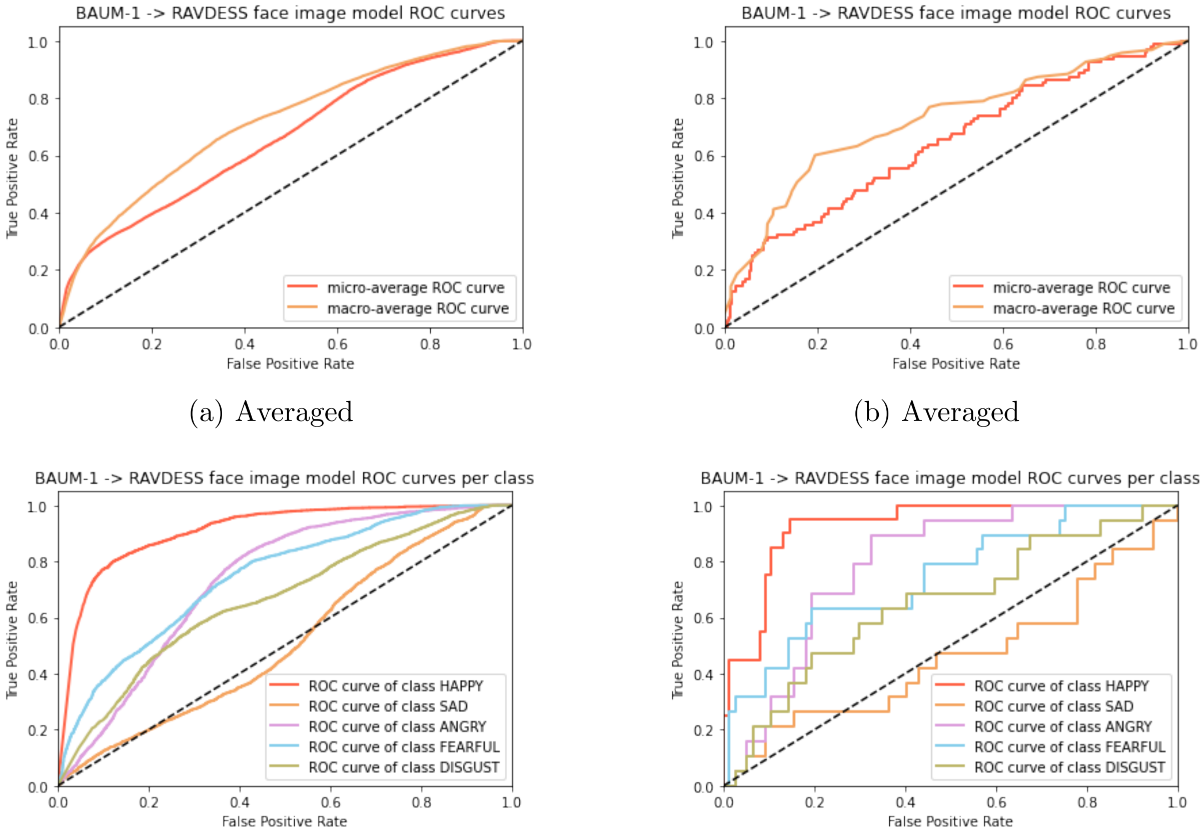
	Happy		**Sad**		**Angry**		**Fearful**		**Disgust**	
	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.	Ind.	Vid.
Face Image	0.91	0.93	0.51	0.46	0.73	0.79	0.75	0.74	0.66	0.65
Optical flow	0.85	0.98	0.62	0.65	0.64	0.68	0.64	0.62	0.74	0.83
Voice	0.37	0.39	0.59	0.57	0.51	0.54	0.40	0.40	0.68	0.73
Face	-	0.98	-	0.61	-	0.68	-	0.66	-	0.75
Face + Voice	-	0.92	-	0.63	-	0.73	-	0.51	-	0.75

Table 35: BAUM-1 - RAVDESS AUC results

ROC curves

Next, it is shown the weighted-average ROC curves and the ROC curves per class for the face image model, the optical flow model, the voice model, and the face model (face image + optical flow). The results are presented based on frames or audio chunks and videos.

Face image model:



(a) Averaged

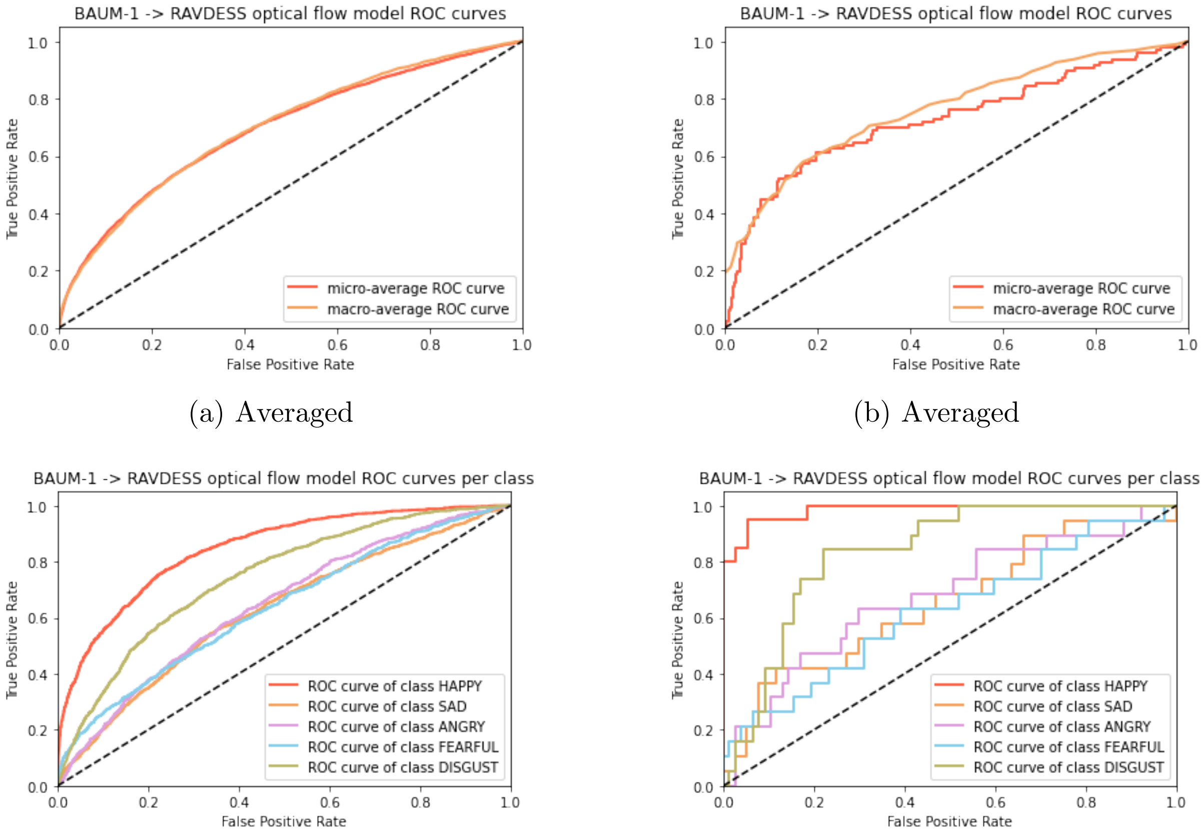
(b) Averaged

(c) Per class

(d) Per class

Figure 55: BAUM-1 - RAVDESS face image model ROC curves based on frames (a)(c) and based on videos (b)(d)

Optical flow model:



(a) Averaged

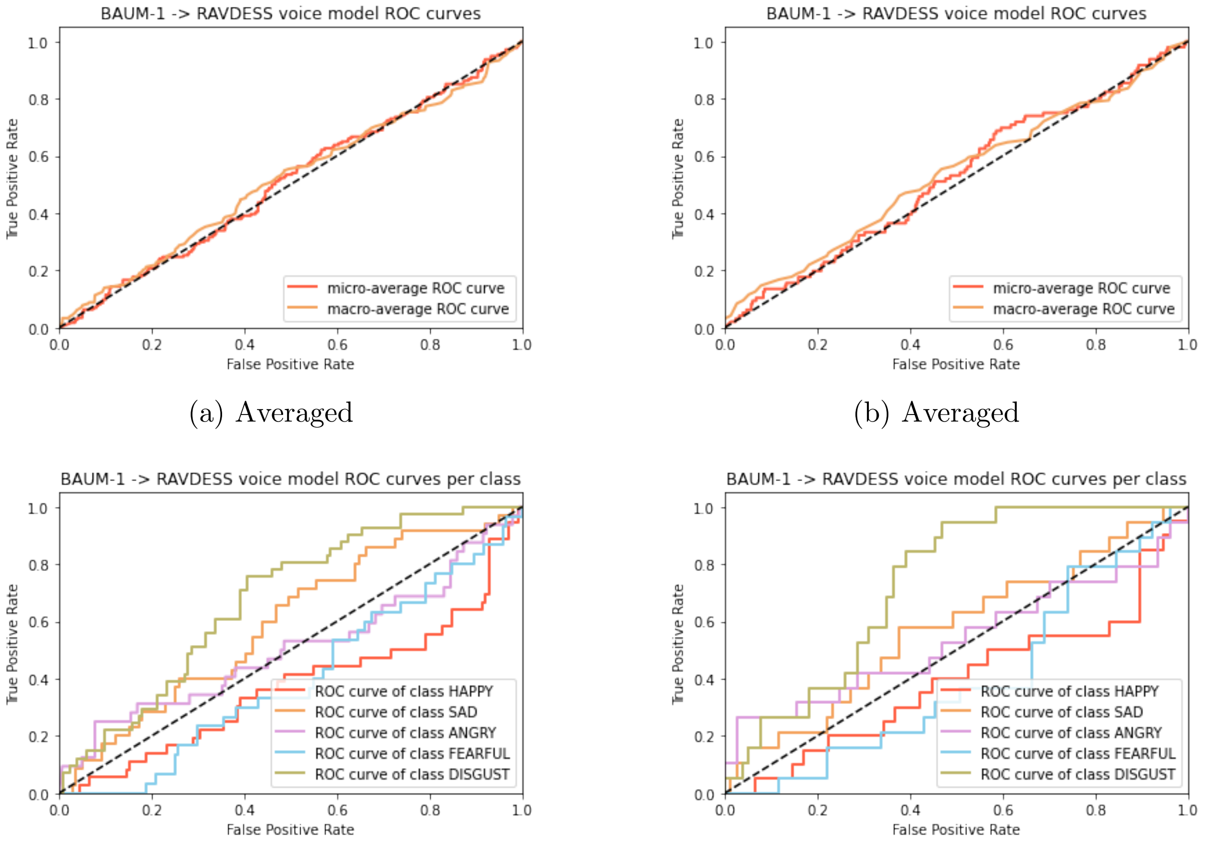
(b) Averaged

(c) Per class

(d) Per class

Figure 56: BAUM-1 - RAVDESS face optical flow ROC curves based on frames (a)(c) and based on videos (b)(d)

Voice model:



(a) Averaged

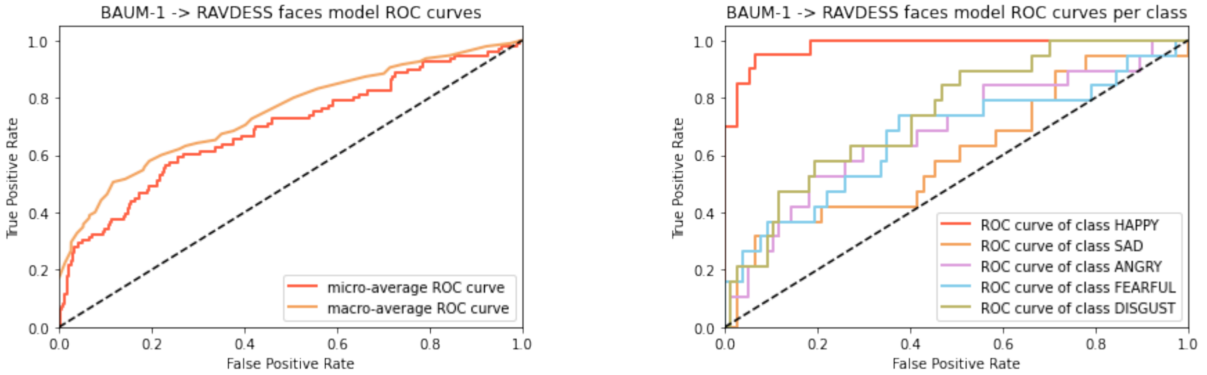
(b) Averaged

(c) Per class

(d) Per class

Figure 57: BAUM-1 - RAVDESS voice model ROC curves based on speech chunks (a)(c) and based on videos (b)(d)

Face model (Face image + Optical flow):



(a) Averaged

(b) Per class

Figure 58: BAUM-1 - RAVDESS face multimodal model ROC curves based on frames (a) and based on videos (b)

Observations

The results show that the worst performance is achieved when the model is trained with the BAUM-1 dataset and tested with the RAVDESS dataset. The face image model predicts happiness or disgust for almost all the videos. In the case of the optical flow model, also identifies anger emotions. Finally, the voice model predicts almost randomly sadness, anger, and fear emotions.

C Glossary

UPC Universitat Politècnica de Catalunya

AI Artificial Intelligence

HCII Human-Computer Intelligent Interaction

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song

BAUM-1 Bahcesehir University Multimodal Emotional Database

MTCNN Multi-task Cascaded Convolutional Networks

FER Face Emotion Recognition

FACS Facial Action Coding System

AU Action Unit

CNN Convolutional Neural Network

DBN Deep Belief Network

FCN Fully Convolutional Network

SER Speech Emotion Recognition

NMS Non-Maximum Supression

HMM Hidden Markov Model

DNN Deep Neural Network

RNN Recurrent Neural Network

SVM Support Vector Machine

TP True Positive

TN True Negative

FP False Positive

FN False Negative

TPR True Positive Rate

FPR False Positive Rate

ROC Receiver Operating Characteristic

AUC Area Under the Curve

GWN Gaussian White Noise

ZCR Zero-Crossing Rate

MFCC Mel Frequency Cepstral Coefficient

FFT Fast Fourier Transform