Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Video Based Crossmodal Representation Learner for Emotion Recognition

Shangrui Nie

Course of Study: Softwaretechnik

Examiner: Prof. Dr. Andreas Bulling

Supervisor: Ekta Sood, M.Sc.
Florian Strohm, M.Sc.

Commenced: November 04, 2022

Completed: May 04, 2023

CR-Classification: I.7.2

Abstract

Emotion recognition is a critical aspect of human-computer interaction, with numerous applications in fields such as psychology, social robotics, and affective computing. Recent advances in deep learning have significantly improved the performance of emotion recognition models, particularly when integrating multimodal data sources such as audio and video. Despite these advancements, the potential of gaze information as an auxiliary modality in emotion recognition has been relatively underexplored, leaving room for further innovation. Additionally, there is a growing interest in pre-training feature extractors to enhance the performance of emotion recognition models, including gaze encoders, which have also been understudied. This thesis presents a novel approach to emotion recognition by incorporating gaze as the additional modality into a multimodal architecture with pre-trained audio, video, and gaze feature extractors on Voxceleb1 [NCZ17], specifically focusing on the Gaze-enhanced L3-Net and AVE-Net architectures.

In addition to the exploration of gaze as an auxiliary modality and the pre-training of feature extractors, it is crucial to investigate the performance of emotion recognition models under single modality conditions. Real-world applications often face challenges in obtaining multimodal data due to factors such as limited resources, privacy concerns, and environmental constraints. By evaluating the effectiveness of our proposed models in single modality settings, we aim to provide a comprehensive understanding of their potential applicability and robustness in diverse scenarios. This aspect of our study highlights the importance of developing high-performing single modality models alongside multimodal approaches for emotion recognition.

In this study, we provide compelling evidence that leveraging pre-trained feature extractors with gaze-enhanced visual and audio embeddings leads to substantial performance gains in emotion recognition models on the OMG Emotion dataset [BCL+18]. Our results show that the pre-trained Gaze-enhanced L3-Net outperforms both the original L3-Net and the AVE-Net, achieving F1 micro scores of 49.28 and 45.94 for the video and audio channels, respectively. Furthermore, our model also surpasses the performance of Abdou et al. [ASMB22]'s model-level fusion and early fusion techniques and achieves state-of-the-art performance on the OMG Emotion dataset [BCL+18] in both video channel and audio channel. Notably, our model's architecture does not outperform Abdou et al. [ASMB22]'s architecture, and the improvements we observe can be largely attributed to the pre-training process.

Ultimately, this study sheds light on the significant potential of incorporating pre-training and gaze information in emotion recognition tasks, paving the way for more accurate and robust models in real-world applications. By thoroughly investigating these aspects, we aim to motivate further research in the field and encourage the development of

innovative approaches that capitalize on the unique advantages of gaze information and pre-training techniques for improved performance in emotion recognition tasks.

Contents

List of Figures

List of Tables

1 Introduction

Emotion recognition plays a vital role in human communication and interaction. Emotions are inherently multimodal human behaviors, expressed through a combination of facial expressions, body language, and speech [[MAV+15][PSCH05][ZPRH07]]. By recognizing these multiple channels of communication, we can gain a more comprehensive understanding of a person's emotional state. In the context of machine learning, leveraging this comprehensive understanding is crucial for developing more accurate and effective emotion recognition models. Consequently, it is natural to consider using multimodal data when designing emotion recognition models to enhance their accuracy. However, while the integration of multiple modalities can improve model performance, obtaining multimodal data in real-world applications can be challenging due to factors such as the need for specialized equipment, complexities in integrating and synchronizing data from different sources, and privacy or ethical concerns. This makes developing robust emotion recognition models for single modality data a valuable pursuit.

Various studies have investigated the relationship between gaze and emotion, which serves as the foundation for incorporating gaze information in our model [[IB09][AK03][MHLB11][BML08][Kel95]]. Numerous researchers have also focused on emotion recognition using both single and multimodal approaches [[SOA21][CCCF14][CJZW17][RSSL13][HZRS19][ASMB22][HZRS19]]. However, there has been relatively less development in incorporating gaze as one of the modalities for multimodal emotion recognition. Notably, based on the EmoBed architecture proposed by Han et al. [HZRS19], Abdou et al. [ASMB22] added gaze as an additional modality, resulting in a multi-modal emotion recognition framework that includes gaze information. Abdou et al. [ASMB22]'s work is one of the few examples of such architectures that consider gaze as an important aspect of emotion recognition, and they also achieved the previous state-of-the-art result on the OMG Emotional challenge dataset [BCL+18]. In this thesis, we build upon Abdou et al. [ASMB22]'s architecture as the foundation for our study, aiming to further explore the potential of gaze-based emotion recognition frameworks.

Recently, pre-trained feature extractors have emerged as a popular technique that has significantly boosted performance in many tasks, including emotion recognition. Pre-trained encoders are useful across tasks due to their ability to capture higher-level,

abstract representations, and generalize well to different domains. However, there has been limited research on pre-training architectures that involve gaze information. We believe that pre-training can not only provide substantial gains for common models, such as image-based or video-based networks but also yield considerable improvements in processing gaze information, thus extending the benefits of pre-training to this crucial modality. By incorporating pre-training in our multimodal emotion recognition model, we aim to leverage this powerful approach to enhance the model's accuracy under single modality conditions.

In this thesis, our objective is to investigate the impact of pre-training on the accuracy of a multimodal emotion recognition model that utilizes audio, video, and gaze as input. We chose Abdou et al. [ASMB22]'s gaze-enhanced EmoBed architecture as our base model and defined the pre-training tasks according to our objective [SRS22]: detecting whether video frames and audio information originate from the same video. Furthermore, we replaced the audio and video feature extractors in the gaze-enhanced EmoBed [ASMB22] architecture with those from Arandjelovic and Zisserman [AZ18]'s L3-Net and AVE-Net, which are video-based pre-training models, to ensure the effectiveness of pre-training. Considering the importance of data commonality between the pre-training and fine-tuning datasets, we chose VoxCeleb1 [NCZ17] as the pre-training dataset for this study. Both pre-trained architectures achieved over 80% accuracy in pre-training tasks.

Upon completion of the training, we performed single modality tests (i.e., using only audio or video embeddings for prediction) on both architectures on the OMG Emotion Challenge dataset [BCL+18], and the results showed significant improvements in both video and audio channels for the pre-trained models. As an additional test, we also used full modality for prediction, that is combining the information of audio, video, and gaze together to perform the prediction, and the results were considerably higher than the single modality results. In terms of comparison, the pre-trained models outperformed Abdou et al. [ASMB22]'s results in both video and audio channels, while the non-pre-trained models did not. This indicates that in terms of architecture selection, i.e. the selection of the feature extractors, the architecture used in this thesis may not have been superior to Abdou et al. [ASMB22]'s architecture, this suggests that the improvement in performance can be largely attributed to the pre-training process rather than the model architecture itself.

2 Related Work

Video-based emotion recognition is an active area of research in the domains of computer vision and machine learning. Various approaches have been proposed to address the challenges of recognizing human emotions from videos, including single-modal and multi-modal methods. Single-modal methods use only one type of feature, such as facial expression or voice, while multi-modal methods combine multiple types of features to improve recognition accuracy. These approaches have shown promising results in emotion recognition tasks.

In the single-modal emotion recognition category, researchers have explored different techniques such as facial expression analysis, speech analysis, and gaze analysis. These methods have achieved good performance in recognizing emotions from individual modalities. However, they may not capture the full spectrum of human emotions since emotions are often expressed through multiple modalities.

Although pre-training techniques have been widely used in natural language processing and computer vision, there have been few studies on pre-training in the context of video-based emotion recognition. Transfer learning and self-supervised learning have been used to pre-train video models in the related task of action recognition. However, the application of pre-training techniques in video-based emotion recognition is still relatively unexplored.

In this section, we provide a comprehensive review of the latest developments in video-based emotion recognition with a particular focus on four main areas: Gaze and Emotions, Gaze-based Emotion Recognition, Multimodal Emotion Recognition, and Video-based Pre-training. We aim to identify the research gaps and challenges in each of these areas and highlight potential research directions to improve the performance of video-based emotion recognition models. By providing an overview of the related work in these areas, this section seeks to establish a solid foundation for our research approach and contribute to the advancement of the field.

2.1 Gaze and Emotions

In this section, we review the literature focusing on the relationship between gaze and emotions, emphasizing the critical role gaze plays in recognizing and distinguishing various emotional states.

Itier and Batty [IB09] laid the foundation by highlighting the central role of eyes and gaze in social cognition. Gaze can reveal information about a person's attention and intentions, which is crucial for successful social interaction. Building on this idea, several studies have identified specific gaze patterns corresponding to different emotions.

One of the key findings is that gaze direction can facilitate the processing of certain emotions. Adams Jr and Kleck [AK03] suggested that direct gaze facilitates the processing of approach-oriented emotions (e.g., anger and joy), while averted gaze facilitates the processing of avoidance-oriented emotions (e.g., fear and sadness). This relationship between gaze direction and emotional expression can help individuals recognize and differentiate emotions.

Consistent with this idea, Milders et al. [MHLB11] found that fearful faces with averted gaze were detected more frequently than those with direct gaze, while angry and happy faces were detected more frequently with direct gaze. These findings further support the notion that gaze direction is associated with specific emotional expressions.

Expanding on these findings, Bindemann, Mike Burton, and Langton [BML08] reexamined the interaction between eye gaze and selected facial emotional expressions, revealing that the perception of happy, sad, angry, and fearful expressions was impaired when eye gaze was averted compared to direct gaze conditions.

In addition to emotion recognition, gaze also plays a role in communicating emotions. Keltner [Kel95] indicated that gaze is the first aspect to change when individuals experience embarrassment. This finding highlights the importance of gaze in social communication.

Lastly, Emery [Eme00] discussed the evolutionary role of gaze in primates, emphasizing its importance in distinguishing emotions and its adaptive function in social interactions.

In conclusion, these studies underline the significant relationship between gaze and emotions. They demonstrate that gaze direction not only corresponds to specific emotional states but also assists in recognizing and distinguishing emotions, facilitating successful social interaction.

2.2 Gaze-based Emotion Recognition

In this section, we discuss the role of gaze-based features in emotion recognition and the various methodologies employed in recent studies to leverage these features for improved emotion recognition performance.

The shared signal hypothesis (SSH) serves as a foundation for understanding the relationship between gaze and emotions, Liang et al. [LZL+21] explored. According to SSH, direct gaze shares an approach-oriented signal with the emotions of anger and joy, whereas averted gaze shares an avoidance-oriented signal with fear and sadness. This hypothesis has been verified using different materials and participant populations, highlighting the importance of gaze direction in emotion perception.

O'Dwyer, Murray, and Flynn [OMF19] has made significant contributions to the field of gaze-based emotion recognition by proposing feature engineering methods that have been widely adopted in other studies. The authors' work focused on the development of gaze features using various statistics such as mean, interquartile range, and standard deviation based on the features generated from Openface [BZLM18]. Utilizing these features, the authors trained an LSTM network for the task of continuous affect prediction on the RECOLA [RSSL13] dataset. They found that their model performed better for arousal prediction when trained on gaze features. In another study, O'Dwyer, Murray, and Flynn [OMF18] assessed the effectiveness of eye gaze as a supportive modality in a bimodal continuous affect prediction system, demonstrating improvements in prediction performance when combining eye gaze with speech features. The addition of eye gaze features to speech yielded an improvement of 19.5% for valence prediction and 3.5% for arousal prediction.

Several studies have investigated the use of gaze-based features in emotion recognition systems, such as Aracena et al. [ABSV15] and Van Huynh et al. [VYL+19]. Aracena et al. [ABSV15] trained a shallow feed-forward neural network using pupil size and gaze position information to classify emotions into positive, negative, and neutral categories. The study also suggested that incorporating additional modalities and gaze features could enhance recognition performance.

Van Huynh et al. [VYL+19] presented an emotion recognition method that combines facial and eye movement information to improve system accuracy. the authors employed a deep learning model, specifically a CNN, to process facial information. Then, eye movement features were extracted from Openface [BZLM18]. A new set of 51 features were used to obtain related information about each emotion for the corresponding sample, and the emotion for a sample was recognized based on the combination of the knowledge from the two previous stages. The study demonstrated a 2.87% improvement in accuracy for the face model when incorporating eye movement features. As a strategy

for feature engineering, the authors also employed various statistical values calculated from the Openface [BZLM18] output but differed from the feature set of O'Dwyer, Murray, and Flynn [OMF19]. Van Huynh et al. [VYL+19] also used the statistics of facial landmark coordinates, while in the work of O'Dwyer, Murray, and Flynn [OMF19], coordinates were only used for pupil size calculation. Additionally, O'Dwyer, Murray, and Flynn [OMF19] calculated features such as eye blink intensity, gaze approach, pupil dilation, and pupil constriction, which were not included in Van Huynh et al. [VYL+19]'s method.

Overall, these studies demonstrate the potential of gaze-based features for improving emotion recognition performance. By leveraging the relationship between gaze and emotions, researchers have been able to develop more accurate and effective emotion recognition systems.

2.3 Multimodal Emotion Recognition

In this section, we delve into the domain of Multimodal Emotion Recognition, exploring the diverse approaches and techniques that researchers have adopted to harness the power of multiple modalities, such as facial expressions, speech, and gestures, in order to enhance emotion recognition performance and accuracy.

Emotion recognition is often associated with visual and auditory modalities, focusing on facial expressions and speech, respectively. Numerous studies have explored these modalities in developing effective emotion recognition models.

Schoneveld, Othmani, and Abdelkawy [SOA21] proposed a deep learning-based approach for audio-visual emotion recognition. For facial feature extraction, they employed a CNN, while for speech, they used a CNN to process raw Mel spectrograms. The embeddings were concatenated using a linear layer, followed by an LSTM and another linear layer to predict arousal and valence. Their results outperformed the state-of-the-art methods in predicting valence on the RECOLA [RSSL13] dataset.

Chen et al. [CCCF14] also focused on visual and audio modalities in their emotion recognition framework. However, their approach differed in the processing of visual features and the fusion of modalities. They introduced a feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG_TOP) for facial expression representation and utilized Multiple Kernel Learning (MKL) for optimal feature fusion.

Chen et al. [CJZW17] went beyond visual and audio modalities by also considering text generated from audio in their emotion recognition study. The authors primarily

discussed feature extraction and fusion across these modalities and investigated the effectiveness of non-temporal support vector regression (SVR) and temporal LSTM-RNN models. Their results demonstrated that LSTM-RNNs significantly improved recognition performance.

However, in real-world scenarios, it may not always be possible to obtain ideal data from multiple modalities. Thus, it becomes crucial to develop models that can still deliver high-accuracy results based on single modality inputs. This is where the work in [HZRS19] becomes particularly relevant. Han et al. [HZRS19] proposes an emotion recognition framework EmoBed, that, while trained using both audio and visual modalities, can perform emotion recognition using only one modality during inference. The framework leverages a VGG face CNN [PVZ15] as the feature extractor for video frames and an openSMILE [EWS10] toolkit for audio. The main focus of this work is the joint Triplet training loss, which encourages the generation of similar embeddings for different modalities when dealing with the same source information.

Building upon Han et al. [HZRS19], Abdou et al. [ASMB22] presents a novel approach that incorporates gaze information as an additional modality, enhancing facial feature information. In [ASMB22] the author utilizes the same joint Triplet training loss but with the addition of gaze-enhanced features and voice features. This updated architecture outperforms the original EmoBed [HZRS19] framework proposed by Han on the One-Minute Gradual Emotion Recognition dataset and reached the state-of-the-art on OMG [BCL+18] datasets on single modality test.

In summary, while substantial progress has been made in the field of multimodal emotion recognition, it is important to recognize the practical need for models that can perform well using single modality inputs. Han et al. [HZRS19] and Abdou et al. [ASMB22] provide valuable insights and advancements in this direction, demonstrating the potential for high-performing emotion recognition models even in situations where only limited data is available.

2.4 Video-based Pre-training

In this section, we discuss video-based pre-training architectures and their applications in the domain of emotion recognition. Pre-training, particularly self-supervised learning, has demonstrated significant performance improvements in numerous computer vision and natural language processing tasks. By conducting self-supervised training on large amounts of unlabeled data, pre-trained models can learn rich feature representations, providing a more robust starting point for downstream tasks. This strategy has achieved remarkable success in various tasks, including emotion recognition. Notably, in Han et al.

[HZRS19]'s architecture, which serves as the foundation for this research, the author also suggests that large-scale pre-training could potentially lead to further enhancements in emotion recognition performance.

In video, self-supervised learning can improve the performance of video understanding and analysis tasks by learning useful feature representations. In recent years, numerous researchers have proposed a variety of self-supervised learning methods, which can be categorized into four different types: Pretext tasks, Generative learning, Contrastive learning, and Cross-modal agreement [SRS22]. Pretext tasks aim to learn useful feature representations by solving a prediction problem unrelated to the actual task, such as predicting the order of frames in a video or colorizing grayscale images. Generative learning seeks to learn useful feature representations by generating data, for instance, using autoencoders or GANs to generate video frames or images. Contrastive learning strives to learn useful feature representations by comparing differences between similar and dissimilar samples, like comparing video clips with other clips from different timepoints or videos. Cross-modal agreement focuses on learning useful feature representations by matching information between multiple modalities, such as visual and audio data.

As the ultimate objective of this thesis is to perform emotion recognition based on multiple modalities, and the subsequent tests will mainly involve single modality tests, it is crucial for each modality in the architecture to have independent predictive capabilities. Therefore, Cross-modal agreement was chosen as the pre-training task category. In the following, architectures based on this training method will be introduced.

Rouditchenko et al. [RBH+20], proposed an audio-visual learning network called AVLnet that learns shared audio-visual embeddings directly from raw video input. The authors used a large-scale pre-training task, conducting self-supervised learning on the HowTo100M [MZA+19] dataset. Specifically, they trained AVLnet with randomly segmented video clips and their original audio waveforms to learn audio-visual representations. As an instantiation of Cross-modal agreement, the authors maximized the difference between audio and video embeddings with the different label while minimizing the difference for same labels as a pre-training task, allowing AVLnet to learn better audio-visual representations. Furthermore, they employed random data augmentation techniques during training to increase data diversity and reduce overfitting risk. This architecture achieved state-of-the-art performance in multiple benchmark tests and was applicable to various image retrieval and video retrieval tasks.

For Cross-modal agreement, Arandjelovic and Zisserman [AZ17] introduced a different implementation using a three-part network structure called L3-Net, composed of visual, audio, and fusion subnetworks. The visual subnetwork employed a convolutional neural network (CNN), the audio subnetwork used a combination of CNN and LSTM layers, and the fusion network utilized a multi-layer perceptron (MLP). Instead of calculating

differences between embeddings, the authors proposed a pre-training method for audio-visual correspondence learning (AVC) that aimed to train visual and audio networks by learning the correspondence between video frames and audio clips using unlabeled video data. Specifically, the task required the model to determine whether a given video frame and audio clip came from the same video (binary classification). This pre-training objective could be performed in a completely unsupervised manner, as it did not require any labeled data. Experimental results showed that this method effectively learned the correspondence between visual and audio features, achieving excellent performance in multiple audio-visual tasks.

Using a similar pre-training method, Arandjelovic and Zisserman [AZ18] improved upon the previous architecture, introducing a new one called AVE-Net. In contrast to the previous embedding concatenation-classification method, AVE-Net first calculated the MSE loss between the two embeddings and passed this loss value into the MLP for binary classification. Additionally, AVE-Net and L3-Net did not differ in their subnetworks, i.e., the audio and video subnets. As a result, AVE-Net outperformed other baseline methods in cross-modal retrieval tasks and achieved impressive results in localizing objects that emit sounds in images. Specifically, in cross-modal retrieval tasks, the performance of this architecture was more than 10% higher than other baseline methods.

In the experiment of this thesis, to ensure the effectiveness of pre-training, we adapted the audio and video subnets from AVE-Net and L3-Net and also employed the audio-video correspondence (AVC) as the pre-training objective.

3 Method

In this chapter, we will 1) introduce the datasets used for pre-training and fine-tuning, as well as the methods employed for gaze feature extraction; 2) describe the feature extractors for each of the three modalities: audio, video, and gaze; 3) introduce the two pre-training architectures, AVE-Net and L3-Net, which are based on these modalities; 4) delve into the fine-tuning architecture utilized in the later stages of the model development; 5) outline the training strategies implemented during the fine-tuning phase of our model.

3.1 Datasets

In this chapter, we introduce two datasets, VoxCeleb1, and OMG Emotion Challenge dataset. VoxCeleb1 is used as the pre-training dataset for training the feature extractor, while OMG is used as the fine-tuning dataset in the next stage. Furthermore, OMG dataset will be used for both single modality tests and joint tests.

3.1.1 Dataset for pre-training

In this thesis, the VoxCeleb1 [NCZ17] dataset serves as our pre-training dataset on gaze-enhanced AVE-Net and Gaze-enhanced L3-Net. VoxCeleb is a large-scale voice dataset that includes video clips of celebrities from various fields, designed for training and testing Automatic Speech Recognition (ASR) and speaker recognition systems. The dataset encompasses a wide range of natural and diverse speaking scenarios, covering various languages, accents, and environmental noises. In our thesis, this dataset is primarily used for training feature extractors for different modalities, allowing them to learn how to process video frames, audio, and gaze. After filtering, we have a total of 897,739 video data, which we split into training and testing sets in an 8:2 ratio. Upon examining the actual videos, we found that the face positions in VoxCeleb videos are relatively stable, and the size of the faces is also appropriate for gaze feature extraction (see Figure 3.1). Moreover, the video resolution is already 224×224, which is convenient

Figure 3.1: Sample frames from Voxceleb dataset. In this dataset, the position and size of the head are relatively stable, which also facilitates the extraction of gaze data. [NCZ17]

for the model to process. These factors make VoxCeleb an excellent choice for a pre-training dataset. The only known drawback is that the dataset is multilingual, while our final fine-tuning task is entirely based on English. The multilingual audio may have a slight negative impact on the final training results because different languages may express the same emotions in different ways. For example, the expression of happiness in English may differ from that in another language, and this could introduce noise into the training data, potentially reducing the model's accuracy.

3.1.2 Dataset for Finetuning

The fine-tuning dataset used in this thesis is the OMG Emotion Challenge dataset [BCL+18]. Sample frames can be seen in Figure 3.2. From the OMG dataset, we can observe that the proportion of the face area is relatively small. If we directly use the resized frames as input, the facial patterns learned by the visual modality may not be very effective. Furthermore, due to the smaller faces, gaze extraction could be less accurate. Based on these two reasons, we first performed face cropping on the OMG dataset and included a portion of the background surrounding the face to make the proportion of the face area as close as possible to that in the VoxCeleb dataset. We then performed gaze extraction based on these cropped faces, making the gaze features more accurate.

Consistent with the prior work of this thesis [ASMB22], we used the training set and validation set of the OMG dataset. In total, we utilized 3,055 data samples. For the distribution of classifications, please refer to Table 3.1.

Upon examining the actual videos, we found that the OMG dataset is not as clean as the VoxCeleb dataset. First, not every frame in the dataset contains a person; sometimes,

Figure 3.2: Sample frames from OMG emotion challenge dataset. In this dataset, the emotions displayed are acted by the performers, and due to the smaller and less stable head position, the accuracy of the gaze data is somewhat reduced compared to that in the Voxceleb1 dataset.[BCL+18]

there are frames with only the background. Additionally, the OMG dataset is not a true emotion dataset; The videos feature actors performing scripted lines, and their task is to express the emotions in the lines. This sometimes leads to overly dramatic performances and exaggerated body movements, causing challenges in face recognition and gaze extraction. We believe this is a reason for reducing the accuracy of the system. However, on the other hand, this also presents an opportunity to test the robustness of the system.

Table 3.1: The data distribution of OMG emotion challenge datasets.[BCL+18]

Neutral	Happy	Sad	Anger	Disgust	Fear	Surprise
1084	920	425	354	160	78	34

3.1.3 Gaze data

In this thesis, all gaze features were extracted using the OpenFace toolkit [BZLM18]. Given a video as input, OpenFace outputs eye gaze direction vectors in world coordinates for both eyes, as well as 2D and 3D eye region landmarks. Additionally, it provides Action Unit Presence and Action Unit Intensity to describe eye movement patterns. Building upon these features, Barreto, Zhai, and Adjouadi [BZA07] further refined the features by calculating various statistics to enhance their representational capabilities. The author also conducted a correlation analysis between the refined features and arousal and valence. Based on this analysis, Abdou et al. [ASMB22] selected a subset of these features as gaze features for the OMG Emotion Challenge dataset. In this thesis, we also adopted the same feature set, calculating the features for both the VoxCeleb1 and OMG datasets. The gaze feature set used in this thesis is presented in the table 3.2.

Table 3.2: Base features from OpenFace output and the corresponding statistical features. These 103 features are subset of the feature set in [OMF19]. LR refers to a linear regression fitted to the time series of feature values in the window. The time ratio is the proportion of time during which a binary feature is detected in the analysis window. IQR denotes the interquartile range, i.e. IQR 2-3 refers to the difference between third and second. quartile. [ASMB22]

Base feature	Statistical functionals	# Statistical functionals
gaze angle x, gaze angle y, Δ gaze angle x, Δ gaze angle y, pupil diameter mm	min, max, mean, median, quartile 1, quartile 3, std, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept, LR slope	60
Δ pupil diameter mm	min, max, mean, quartile 1, quartile 3, std, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept, LR slope	11
eye blink intensity	max, mean, median, quartile 3, std, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept, LR slope	10
pupil dilation, pupil constriction	time ratio, mean time, max time, total time	8
gaze approach	time ratio, mean time, max time, median time	4
eyes closed, gaze fixation	time ratio, min time, max time, mean time, median time	10

3.2 Multi-modal Subnetworks for Emotion Recognition

3.2.1 Audio Subnetwork

As illustrated in Figure 3.3, the audio subnetwork in our experiment takes log Mel spectrograms as inputs instead of raw audio sequences.

Log Mel Spectrogram is a widely used feature representation in audio processing tasks, which combines the principles of both the Mel scale and the Short-Time Fourier Transform (STFT). The STFT is applied to the audio signal to obtain the time-frequency representation, providing insight into the spectral content of the signal. However, the

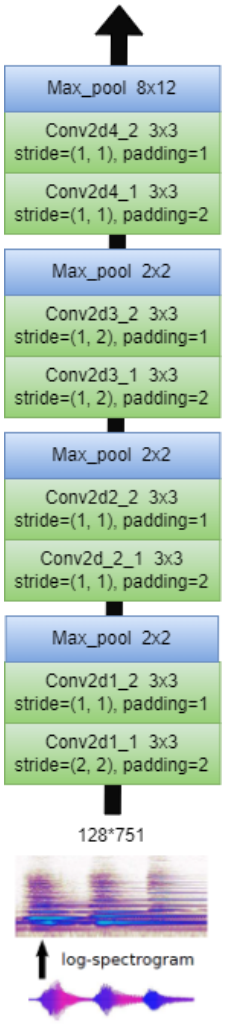

Figure 3.3: The feature extractor of the audio input. As with most audio pre-training architectures, we use the log mel spectrogram of the audio as input and employ 2D CNN layers to extract information. After every two 2D CNN layers, a pooling layer is utilized, resulting in a final 512-dimensional embedding vector.

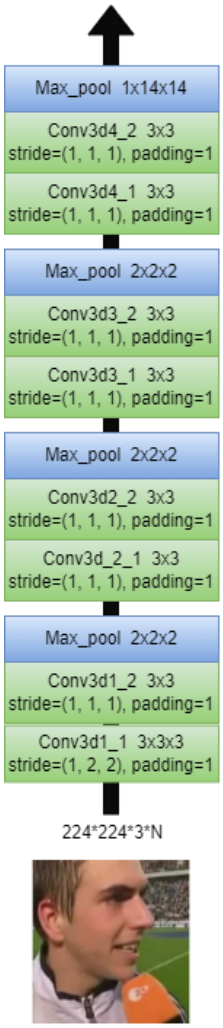

Figure 3.4: The feature extractor of the video input, which takes 224x224 three-channel frames as input. To capture the temporal information between frames, we employ 3D CNN layers, with a pooling layer after every two 3D CNN layers. The final output is a 512-dimensional embedding vector.

standard linearly spaced frequency bins of the STFT do not accurately reflect human perception of sound, which is more sensitive to lower frequencies than higher ones.

The Mel scale is a perceptual scale of pitches, designed to better align with human auditory perception. By converting the frequency axis of the STFT into the Mel scale, we obtain a more meaningful representation of the audio signal, known as the Mel Spectrogram. To further improve the representation, the logarithm of the Mel Spectrogram is computed, resulting in the Log Mel Spectrogram. This logarithmic transformation emphasizes the more relevant and perceptually important components of the audio signal.

In summary, the Log Mel Spectrogram is chosen as the input to our audio subnet due to its ability to capture the spectral content of audio signals in a perceptually meaningful way. The combination of the Mel scale and logarithmic transformation results in a feature representation that emphasizes the most relevant and perceptually important components of the audio signal, which is crucial for tasks such as emotion recognition. By employing Log Mel Spectrogram, we aim to extract meaningful information from audio signals, allowing our model to better understand the nuances of speech and ultimately enhancing the overall performance of our emotion recognition system. Due to the 2D nature of the Log Mel Spectrogram, we use CNN as the feature extractor for audio.

Although the original paper by Abdou et al. [ASMB22], which is the foundation of this thesis, uses OpenSMILE [BZLM18] as a feature extractor, we decided against it for two main reasons. First, OpenSMILE outputs an 88-dimensional vector containing domain-specific attributes, such as Critical Band spectra, Loudness, Zero and Mean Crossing rate, etc. [1] However, these outputs do not contain rich continuous information, making it less suitable as a feature extractor compared to pattern-capturing CNNs. Second, the CNN-based structure is more in line with the pre-training architectures found in the literature. In our audio subnetwork, each CNN layer is followed by a batch normalization layer to increase the model's generalization capabilities and avoid overfitting. In the end, the Audio Subnetwork outputs a 512-dimensional vector.

3.2.2 Video Subnetwork

The video subnetwork is as shown in Figure 3.4. As the input for the video subnetwork, we use 224x224 3-channel frames. In the work of Abdou et al. [ASMB22], the feature extractor for the video part is 2D VGG face CNN [PVZ15], which means that for each video, the feature extractor generates an embedding with an output shape of $(N, 4096)$,

[1] https://audeering.github.io/opensmile/about.html%audio-features-low-level

where N represents the number of frames, and 4096 is the dimension of the embedding. Based on this output, Ahmed feeds it into a single-layer GRU layer to extract continuous information between different frames. In contrast, Arandjelovic and Zisserman [AZ18] extracts features using 3D CNN, which establishes relationships between adjacent frames through its three-dimensional convolutional kernels. We believe that a multi-layer 3D CNN is more capable of recognizing human facial patterns than a single-layer GRU layer based on embeddings. Furthermore, to ensure the pre-training performance, we use the same video feature extraction network as in the AVE-net and L3-net mentioned by Arandjelovic and Zisserman [AZ18]. Besides, similar to the audio feature extraction network, a batch normalization layer follows each CNN layer to improve the model's generalization ability and prevent overfitting. In the end, the Video Subnetwork also outputs a 512-dimensional vector.

3.2.3 Gaze Subnetwork

The Gaze subnetwork in our model consists of two MLP (Multilayer Perceptron) layers. The base gaze features are extracted on a per-frame basis using OpenFace [BZLM18], and we further refine these features by calculating various statistics, the same as the methods used by Abdou et al. [ASMB22] and O'Dwyer, Murray, and Flynn [OMF19]. After refining, each video has a 103-dimensional vector as its gaze feature, containing statistical parameters such as mean, max, interquartile range (IQA), and standard deviation (STD). Since this vector is unordered and each video has only one gaze feature vector, there are no dynamic patterns to recognize. Therefore, the advantages of traditional CNN and LSTM modules may not be well-expressed on gaze features. Based on these reasons, we chose to use MLP layers as the feature extractor for the gaze modality. The defined MLP layers map the 103-dimensional gaze feature to a 128-dimensional representation, with ReLU activation functions between the layers and BatchNorm layers following each layer.

3.3 Pretrain Architecture

In this section, we will delve into the pre-training architectures utilized in our study, which serve as the foundation for extracting features from different modalities, as introduced in the previous chapter. It is essential to emphasize that the primary purpose of pre-training is to effectively train feature extractors for audio, video, and gaze, ensuring that they can capture rich representations from their respective data sources before being integrated into the emotion recognition framework. Specifically, we will discuss two pre-training architectures, Gaze-enhanced AVE-Net and L3-Net, both of

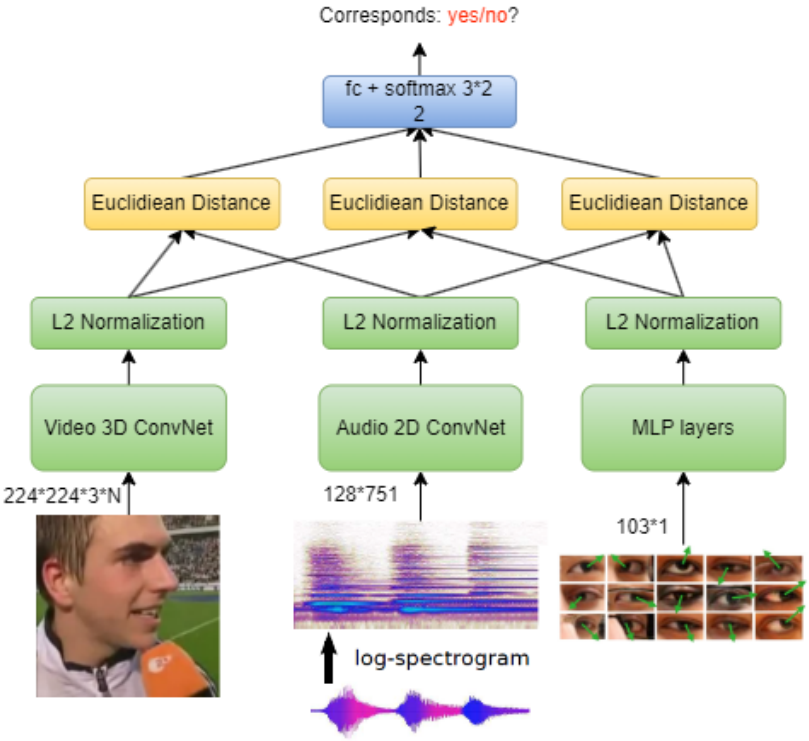

Figure 3.5: The structure of the Gaze-enhanced AVE-Net. The audio and video feature extractors are as depicted in Figure 3.3 and 3.4. Based on the output of feature extractors, the Euclidean distance is calculated to measure the differences between embeddings, and then the Euclidean distance is used for the binary classification.

which were previously employed by Arandjelovic and Zisserman [AZ18]. The original architecture of AVE-Net and L3-Net only used audio and video as input modalities, in this thesis, gaze was added as a new modality with its own feature extractor. Based on the audio, video, and gaze channels, we used the same pre-training task as in Arandjelovic and Zisserman [AZ18]'s work: detecting whether audio and video frames come from the same video. As for the gaze channel, because the gaze information is directly extracted from the video frames, we can ensure that the video frames and gaze information are always from the same source. With the pretrain method, we aim to enable the feature extractors of audio, video, and gaze to understand and utilize the information from videos. These architectures have been proven to effectively train feature extractors [AZ18], leading to improved performance during the pre-training process.

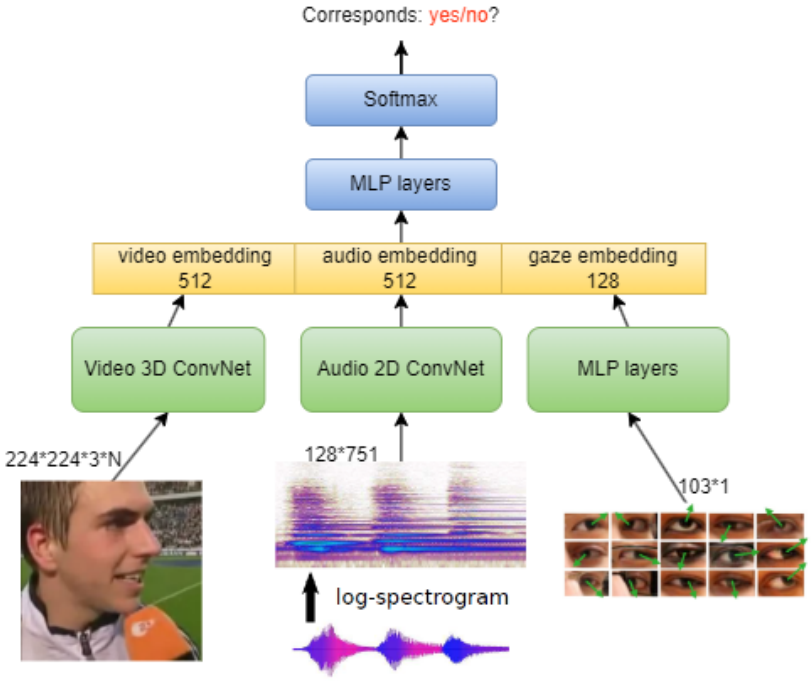

Figure 3.6: The structure of the Gaze-enhanced L3-Net. Compared to Gaze-enhanced AVE-Net the structure is simpler, with the output of the same feature extractors, Gaze-enhanced L3-Net concatenates all embeddings together and feeds them into MLP layers to perform classification.

3.3.1 Gaze-Enhanced AVE-Net

The architecture of the Gaze-Enhanced AVE-Net is as shown in Figure 3.5. In this thesis, the Gaze-Enhanced AVE-Net architecture is adapted from the paper Arandjelovic and Zisserman [AZ18]. The original AVE-Net in that paper only has two modalities as input: audio and video frames. The authors made the assumption that if audio and video frames are from the same source (i.e., the same video), different feature extractors should produce similar embeddings. Based on this assumption, the distance between the two embeddings should be smaller when the audio and video are from the same source, and larger when they are from different sources. To measure this distance, the authors calculated the mean square error between the outputs of the two feature extractors and used it for subsequent binary classification.

In the Gaze-Enhanced AVE-Net architecture employed in this thesis, the same concept is extended to the gaze modality as well, We believe that the commonality of embeddings for homologous information exists not only in audio and video but also in gaze. Hence, we decided to include gaze embeddings in the difference calculation. The mean square error between the embeddings of all three modalities (audio, video, and gaze) is

calculated and used as the basis for subsequent classification. Additionally, in the original AVE-Net, the audio and video feature extractors were followed by MLP layers that transformed the 512-dimensional embeddings into 128-dimensional ones. In this thesis, the gaze feature extractor is composed of MLP layers, which transform a 103-dimensional gaze statistical feature set into a 128-dimensional embedding vector. Therefore, we have audio embedding, video embedding, and gaze embedding with the same dimensions, ensuring compatibility with the mean square error calculation process in the AVE-Net architecture.

3.3.2 Gaze-Enhanced L3-Net

The L3-Net architecture, initially introduced by Arandjelovic and Zisserman [AZ17] and later used as a baseline by Arandjelovic and Zisserman [AZ18], is depicted in Figure 3.6. Compared to the AVE-Net architecture, L3-Net is relatively simple with respect to the complexity of the architecture. It concatenates the embeddings generated by the different feature extractors horizontally, resulting in a single vector. This concatenated vector is then passed through multiple MLP layers and a softmax layer for classification. It is worth mentioning that the performance of L3-Net and AVE-Net in pre-training tasks is not significantly different; however, AVE-Net shows better results in downstream tasks.

3.4 Fine-tuning Architecture

The fine-tuning architecture is as shown in 3.7. In this thesis, the fine-tuning architecture is largely consistent with the structure presented by Abdou et al. [ASMB22]. According to the classification in the paper, the implemented architecture belongs to the model-level fusion structure, which achieved the previous state-of-the-art results in the video channel of the single modality test. Building upon method Abdou et al. [ASMB22] proposed, we replaced the original audio feature extractor, OpenSMILE, with a 2D CNN structure based on log mel spectrograms 3.2.1 and replaced the original video feature extractor, 2D VGG Face CNN, with a 3D CNN architecture. The gaze feature extractor remained unchanged, utilizing MLP layers. The substituted feature extractors have been pre-trained on the VoxCeleb1 [NCZ17] dataset.

On top of the feature extractors for each modality, the authors also combined the video and gaze portions using MLP layers to form a more informative visual embedding, which has been preserved in the implemented architecture. Therefore, before the classifier, we still have two different embeddings: audio embedding and gaze-enhanced

visual embedding. It is worth noting that the classifier in the fine-tuning architecture is modality-agnostic, meaning that both audio embedding and gaze-enhanced visual embedding can serve as inputs to the classifier for single modality inference. As an additional experiment, we also attempted to concatenate the embeddings, allowing for joint prediction using all modalities as input. The results are presented in Chapter 4.

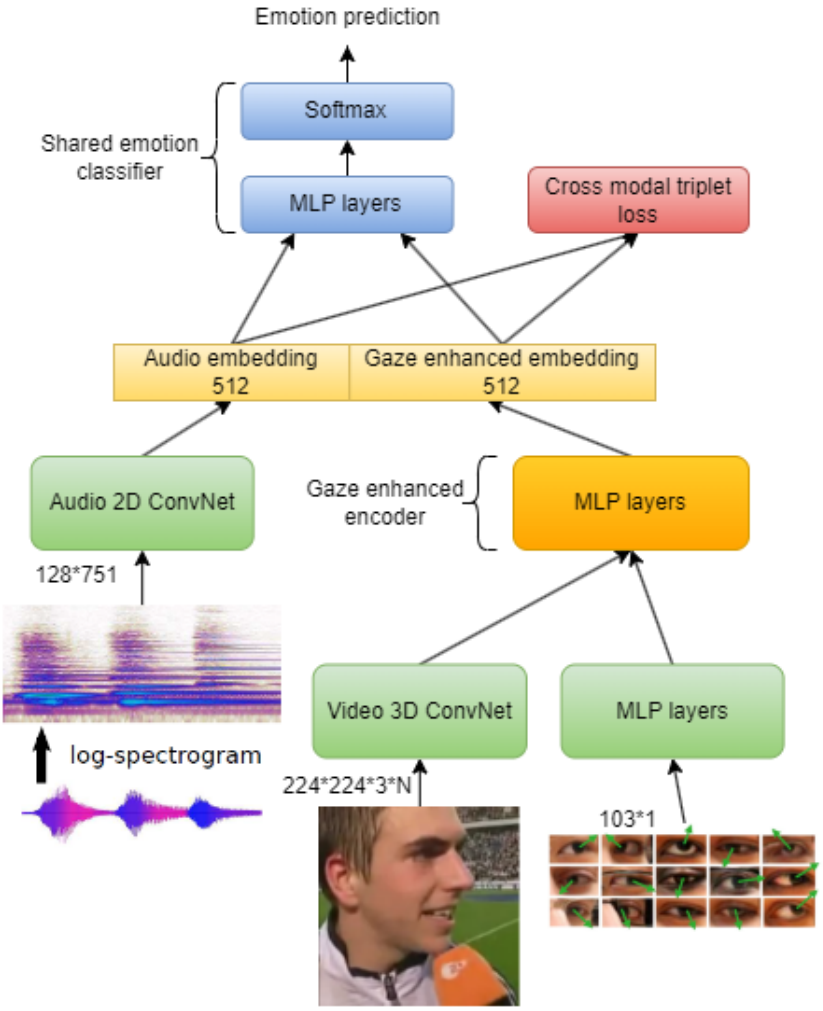

Figure 3.7: The fine-tuning structure used in this thesis. The audio, video, and gaze modalities are each fed into their own pre-trained feature extractors. On the basis of the feature embeddings, the video channel and gaze channel are fused by an MLP block to form the gaze-enhanced visual embedding. The audio embedding and gaze-enhanced visual embedding are then separately fed into a classifier for single-modality testing. In addition, we perform cross-modal triplet loss calculations based on these two embeddings to aid in training.

3.5 Triplet-training

In this section, we will focus on the loss function used during the fine-tuning stage of our model, which plays a crucial role in the training process and the model's overall performance.

During the fine-tuning stage, the objective is single modality inference, meaning that only audio embedding or gaze-enhanced visual embedding can be used. Consequently, treating the problem as a standard classification task and merely using cross-entropy loss could lead to confusion in the model, as gradients beneficial for the audio channel might not be beneficial for the visual channel. To address this issue, Han et al. [HZRS19] proposed the cross-model triplet loss, which allows the two channels to learn from each other, and this approach was also implemented in Abdou et al. [ASMB22]'s work.

The core concept of this loss function involves the following: within a single batch of instances, if the labels are the same, the audio embeddings generated during the forward pass should be close in distance. Conversely, if the labels are different, the distance should be greater. The same principle applies to gaze embeddings, which constitutes the intra-modality loss. Additionally, within the same batch of instances, if the labels are the same, the audio and visual embeddings of different instances should also be close in distance, while they should be far apart if the labels are different. This establishes a connection between the audio and visual channels, referred to as the inter-modality loss. As the distance between two embeddings e_i and e_j is defined as $d(e_i, e_j) = ||e_i - e_j||_2$. Detailed information on the loss terms is introduced in this section.

3.5.1 Intra-modality loss

In a batch with n instances, when considering one of the embeddings A, the intra-modality loss calculation begins with computing an $n \times n$ pairwise distance matrix, where the diagonal elements represent the distance between an embedding and itself, which is 0. The intra-modality loss is defined as the difference between the hardest positive and hardest negative. The hardest positive refers to the maximum distance between instances with the same label, while the hardest negative denotes the minimum distance between instances with different labels. The formula of Intra-modality loss is as shown in 3.1.

$$L_{Intra}(A) = \sum^{n}(d(e_a, e_a^+) - d(e_a, e_a^-)) \tag{3.1}$$

3.5.2 Inter-modality loss

Given a batch of n examples with both embeddings e_a of modality A and corresponding embeddings e_b of modality B. First, we still need to compute an $n \times n$ pairwise distance matrix. This matrix contains the distances $d(e_i, e_j)$ for all $e_a \in A$ and $e_b \in B$, Then, with the same strategy for searching for the hardest positives and negatives, the inter-modality loss is calculated by

$$L_{Inter}(A, B) = \sum^{n}(d(e_a, e_b^+) - d(e_a, e_b^-)) \tag{3.2}$$

3.5.3 Full loss term

After the explanations provided above, the complete triplet loss can be defined as the formula 3.4. In addition to the triplet loss, the conventional Cross entropy loss is also indispensable. It is defined as the sum of the Cross entropy of the classification results from the two modalities.

$$L_{Triplet} = L_{Inter}(A, V) + L_{Intra}(A) + L_{Intra}(V) \tag{3.3}$$

$$L_{full} = \alpha * L_{Triplet} + Crossentropy(A, label) + Crossentropy(V, label) \tag{3.4}$$

A refers to the audio embedding, V denotes the gaze-enhanced visual embedding, and α is a parameter that controls the influence of the triplet term. In the actual training process, the value of triplet loss usually ranges from several tens to 220, while the maximum value of cross entropy is only 1.95 (when the answers follow a uniform distribution). The disparity in loss values causes the impact of cross entropy to be diminished. Hence, we added the alpha parameter to control the influence of triplet loss. Through experimentation, it has been found that the best training performance is achieved when α is set to 0.1.

4 Results

In this section, we present and discuss the results obtained from our experiments, highlighting the effectiveness of the proposed methodologies and architectures. We will begin by presenting the pre-training results for both AVE-Net and L3-Net architectures, followed by a comparison of single modality test results for models with and without pre-training. Then, we will compare our findings with the results from the previous research, showcasing the improvements achieved through our approach in gaze-enhanced visual and audio embeddings. As a performance metric, we utilized F1 micro score, which is also the evaluation metric used in the upstream literature [ASMB22][HZRS19]. The objective of this section is to demonstrate the impact of incorporating gaze information and pre-training strategies on emotion recognition tasks and provide insights into the performance enhancements achieved with our methods.

4.1 Pre-training Results

In this section, we will present the results obtained during the pre-training phase of our multimodal emotion recognition model. The pre-training process plays a crucial role in preparing the model to effectively learn from the fine-tuning dataset. Here, we will discuss the performance of the two pre-training architectures, AVE-Net and L3-Net, in terms of their accuracy on the VoxCeleb1 dataset. This will provide a solid foundation for understanding the model's subsequent performance in single modality tests.

Table 4.1: The pre-training result of Gaze-enhanced L3-Net and AVE-Net. As the baseline performance for the pre-training task, we use random probability, that is, 50%.

Model	F1 micro score
Baseline (random)	50.00
L3-Net	84.62
AVE-Net	85.11

We would like to first reiterate that the pre-training task involves determining whether audio and video frames come from the same video, which is a binary classification problem. The baseline F1 micro score for this task is 50, achieved by random classification. Additionally, the pre-training dataset contains an equal number of negative and positive samples.

As shown in Table 4.1, both AVE-Net and L3-Net exhibit comparable performance on the pre-training task, with only a 0.49 difference in their respective results. This indicates that both models are capable of effectively processing different modalities of input and achieving satisfactory outcomes. However, it is crucial to emphasize that the pre-training results do not definitively establish the superiority of either model, as the primary focus of pre-training is binary classification. The subsequent fine-tuning stage, which involves a multi-classification task, will provide further insight into the performance of the two models in the context of emotion recognition.

4.2 Fine-tuning Results

In this section, we aim to assess the effectiveness of our pre-trained models in the fine-tuning phase on the OMG Emotion Challenge dataset. We will first examine the performance of single modality tests, where the gaze-enhanced video embedding and audio embedding are evaluated separately, following the testing approach adopted by the previous study that serves as the foundation of this work. Subsequently, we will explore the results of a more comprehensive approach, in which both modalities are combined to leverage all available information for emotion recognition. This analysis will provide valuable insights into the overall performance and potential improvements offered by our pre-training strategy.

As shown in Table 4.2, we can observe the performance of the different models on the OMG Emotion Challenge dataset. The results are presented in terms of F1 micro scores for each modality (video and audio) and their combination.

For the pre-trained Gaze-enhanced L3-Net, the F1 micro scores for the video and audio channels are 49.28 and 45.94, respectively. In comparison, the pre-trained gaze-enhanced AVE-Net yields F1 micro scores of 42.94 and 45.42 for the video and audio channels, respectively. These results indicate that the pre-trained Gaze-enhanced L3-Net outperforms the pre-trained gaze-enhanced AVE-Net in both modalities, particularly in the video channel.

When considering the models without pre-training, the Gaze-enhanced L3-Net achieves F1 micro scores of 42.37 and 38.65 for the video and audio channels, respectively, while the gaze-enhanced AVE-Net attains scores of 40.27 and 41.79 for the video and audio

Table 4.2: The single modality test result of Gaze-enhanced L3-Net and gaze-enhanced AVE-Net. The metric of evaluation is F1 micro score.

Model	Audio	Video
pre-trained Gaze-enhanced L3-Net	45.94	49.28
pre-trained gaze-enhanced AVE-Net	45.42	42.94
Gaze-enhanced L3-Net	38.65	42.37
gaze-enhanced AVE-Net	41.79	40.27
pre-trained Gaze-enhanced L3-Net	57.82	
pre-trained gaze-enhanced AVE-Net	52.35	

channels, respectively. These results demonstrate the importance of pre-training, as the pre-trained models consistently outperform their non-pre-trained counterparts in both modalities.

When combining all modalities for prediction, the pre-trained Gaze-enhanced L3-Net reaches an F1 micro score of 57.82, significantly outperforming the pre-trained gaze-enhanced AVE-Net, which achieves a score of 52.35. This indicates that the pre-trained Gaze-enhanced L3-Net is more effective at leveraging multimodal information to make predictions.

It is important to note that the results presented above are based on the average of three runs, due to computational resource constraints. However, the trends observed in the results provide valuable insights into the performance of the different models and the impact of pre-training on their effectiveness.

4.3 Benchmarking Against Prior Work

In this final section of the results chapter, we will benchmark the performance of our best-performing model, the pre-trained Gaze-enhanced L3-Net, against prior work conducted on the OMG Emotion Challenge dataset [BCL+18] as well as the provided baseline. This comparison will help us understand the effectiveness of our proposed model in relation to existing approaches.

As shown in Table 4.3, our model, the pre-trained Gaze-enhanced L3-Net, achieves F1 micro scores of 49.28 and 45.94 for the video and audio channels, respectively, outperforming other existing models on the OMG dataset.

Table 4.3: Comparison of pre-trained Gaze-enhanced L3-Net and results published in other works. The metric of evaluation is F1 micro score.

Model	Audio	Video
pre-trained Gaze-enhanced L3-Net	45.94	49.28
Abdou et al. [ASMB22], model-level fusion	42.6	45.0
Abdou et al. [ASMB22], early fusion	43.4	43.7
Han et al. [HZRS19]	41.7	43.9
OMG baseline [BCL+18]	33.0	37.0

For Abdou et al. [ASMB22]'s model-level fusion, our model outperforms their results by 4.28 points on the video channel (49.28 vs. 45.0) and by 3.34 points on the audio channel (45.94 vs. 42.60). In the case of Abdou et al. [ASMB22]'s early fusion, our model shows an improvement of 5.58 points for the video channel (49.28 vs. 43.7) and 2.54 points for the audio channel (45.94 vs. 43.40).

Comparing our model to EmoBed [HZRS19] architecture, our pre-trained Gaze-enhanced L3-Net outperforms their model by 5.38 points on the video channel (49.28 vs. 43.9) and 4.24 points on the audio channel (45.94 vs. 41.70).

Lastly, when considering the OMG baseline [BCL+18] results, our model demonstrates a significant improvement, outperforming the baseline by 12.28 points for the video channel (49.28 vs. 37.0) and 12.94 points for the audio channel (45.94 vs. 33.0). This clearly indicates that our pre-trained Gaze-enhanced L3-Net model outperforms the baseline and other existing models in emotion recognition on the OMG dataset.

An essential aspect to highlight is the significance of pre-training in our model's performance. By referring back to the previous Table 4.2, we can see that the models without pre-training did not perform as well as the models presented in Table 4.3. For instance, the non-pre-trained Gaze-enhanced L3-Net and AVE-Net lagged behind Abdou et al. [ASMB22]'s model-level fusion by 2.63 and 4.73 points in the video channel, and by 3.95 and 1.61 points in the audio channel, respectively.

This observation suggests that the improvement in our model's performance can be largely attributed to the pre-training process rather than the model architecture itself. Despite our model's architecture not being inherently superior to the other models, the substantial performance gains achieved by our pre-trained Gaze-enhanced L3-Net (with an increase of 6.91 in video and 7.29 in audio compared to non-pre-trained Gaze-enhanced L3-Net) demonstrate the effectiveness of pre-training in emotion recognition tasks, specifically on the OMG dataset.

4.4 Discussion

As shown in previous sections, the pre-trained Gaze-enhanced L3-net and gaze-enhanced AVE-Net both show significant improvements in the video and audio channels compared to their non-pre-trained counterparts. The pre-trained Gaze-enhanced L3-net achieves state-of-the-art performance, surpassing the results of previous research in both channels. This showcases the benefits of pre-training and supports our hypothesis that our architecture and objective can benefit from external information, specifically the VoxCeleb1 [NCZ17] dataset in this case.

It is important to note that the choice of external information is crucial for the success of our architecture. As our final objective is video-based emotion recognition, we selected the video-based interview dataset VoxCeleb1 to ensure that our architecture can effectively understand and utilize video frame information, audio information, and gaze information.

Although Abdou et al. [ASMB22]'s work also utilizes a pre-trained video frame feature extractor, VGG Face CNN [PVZ15], it is based on single-image pre-training and their gaze feature extractor is not pre-trained. Consequently, their results are not as strong as those presented in our study.

However, it is worth mentioning that the untrained Gaze-enhanced L3-net and gaze-enhanced AVE-Net do not outperform Abdou et al. [ASMB22]'s structure in both channels. This suggests that our choice of model architecture may not be more advantageous. It is possible that the use of 2D CNNs combined with GRU layers for video processing could be more effective than 3D CNNs, and the global domain-specific attributes output by OpenSMILE [EWS10] might have more robust performance than 2D CNNs on log mel spectrograms 3.2.1.

Lastly, it is also possible that we have not fully understood the details of Abdou et al. [ASMB22]'s architecture and may have missed some structural components when building the baseline architecture. To further investigate this, we plan to obtain the relevant code permissions and repeat the experiments for validation purposes.

5 Conclusion and Future Work

In conclusion, this study has presented a comprehensive examination of the benefits of pre-training on emotion recognition performance using gaze-enhanced visual and audio embeddings in the context of the OMG dataset. Our findings indicate that pre-training is a crucial factor in improving the performance of models for emotion recognition tasks, regardless of their architecture.

We have introduced the Gaze-enhanced L3-Net and AVE-Net models and demonstrated their effectiveness in pre-training tasks, achieving an accuracy of over 80%. Moreover, we have shown that the pre-trained Gaze-enhanced L3-Net outperforms the existing models on the OMG dataset, as well as the non-pre-trained versions of the same architectures, illustrating the potential of pre-training in enhancing emotion recognition capabilities.

Furthermore, our study highlights the importance of considering both single modality and multimodal approaches when evaluating the performance of emotion recognition models. The results indicate that combining information from both video and audio channels can lead to improved performance compared to using single modalities independently.

Despite the limitations of the study, such as the limited number of runs due to computational resource constraints, our findings underscore the potential of pre-training as a valuable tool for improving emotion recognition performance.

Overall, this research contributes to a deeper understanding of the role of pre-training in emotion recognition and provides a foundation for future work aimed at developing more accurate and efficient models in this domain.

In this study, as presented in the results section, the chosen model architecture did not outperform that of Abdou et al. [ASMB22]. The primary contribution of this research is to demonstrate the positive impact of pre-training on emotion recognition tasks. In the early stages of the project, Abdou et al. [ASMB22]'s code was not available, which leads us to the following future work directions.

Firstly, we plan to adapt the original architecture of Abdou et al. [ASMB22] by incorporating pre-training and conducting single modality tests under the same training settings. This will further verify that pre-training can indeed improve the performance

of the current state-of-the-art architectures, even though, as discussed in the results section, the Gaze-enhanced L3-Net and AVE-Net do not show inherent superiority in their structures. We also expect that the outcomes of this future work will surpass the results obtained in this study.

Furthermore, in both Abdou et al. [ASMB22]'s architecture and our own, the gaze data is processed using statistics as input for MLP layers, which prevents the model from capturing patterns representing eye movements. As an optimization, frame-based transformer architectures [VSP+17] have been proven effective in various tasks. Therefore, replacing the MLP layers with transformers is a feasible future direction.

Additionally, we will continue to explore replacing the feature extractors in Abdou et al. [ASMB22]'s original architecture with those used in this study to further validate the superiority of their approach. All these future experiments will be strictly based on Abdou et al. [ASMB22]'s architecture to ensure the comparability of the results.

6 Acknowledgement

I would like to extend my heartfelt gratitude to Ekta and Florian for their invaluable guidance and support throughout my research journey. Their insightful suggestions and direction not only helped me navigate the complexities of my research but also greatly improved my presentation skills. Additionally, I am grateful to the HCI-CS team for providing me with the computational resources that played a crucial role in expediting my experiments.

Bibliography

[ABSV15] C. Aracena, S. Basterrech, V. Snáel, J. Velásquez. “Neural networks for emotion recognition based on eye tracking data.” In: *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE. 2015, pp. 2632–2637 (cit. on p. 15).

[AK03] R. B. Adams Jr, R. E. Kleck. “Perceived gaze direction and the processing of facial displays of emotion.” In: *Psychological science* 14.6 (2003), pp. 644–647 (cit. on pp. 11, 14).

[ASMB22] A. Abdou, E. Sood, P. Müller, A. Bulling. “Gaze-enhanced Crossmodal Embeddings for Emotion Recognition.” In: *Proceedings of the ACM on Human-Computer Interaction* 6.ETRA (2022), pp. 1–18 (cit. on pp. 3, 11, 12, 17, 22–24, 26, 27, 30, 32, 35, 38, 39, 41, 42).

[AZ17] R. Arandjelovic, A. Zisserman. “Look, listen and learn.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 609–617 (cit. on pp. 18, 30).

[AZ18] R. Arandjelovic, A. Zisserman. “Objects that sound.” In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 435–451 (cit. on pp. 12, 19, 27–30).

[BCL+18] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, S. Wermter. “The OMG-emotion behavior dataset.” In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–7 (cit. on pp. 3, 11, 12, 17, 22, 23, 37, 38).

[BML08] M. Bindemann, A. Mike Burton, S. R. Langton. “How do eye gaze and facial expression interact?” In: *Visual Cognition* 16.6 (2008), pp. 708–733 (cit. on pp. 11, 14).

[BZA07] A. Barreto, J. Zhai, M. Adjouadi. “Non-intrusive physiological monitoring for automated stress detection in human-computer interaction.” In: *Human–Computer Interaction: IEEE International Workshop, HCI 2007 Rio de Janeiro, Brazil, October 20, 2007 Proceedings 4*. Springer. 2007, pp. 29–38 (cit. on p. 23).

[BZLM18] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency. “Openface 2.0: Facial behavior analysis toolkit.” In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE. 2018, pp. 59–66 (cit. on pp. 15, 16, 23, 26, 27).

[CCCF14] J. Chen, Z. Chen, Z. Chi, H. Fu. “Emotion recognition in the wild with feature fusion and multiple kernel learning.” In: *Proceedings of the 16th International Conference on Multimodal Interaction*. 2014, pp. 508–513 (cit. on pp. 11, 16).

[CJZW17] S. Chen, Q. Jin, J. Zhao, S. Wang. “Multimodal multi-task learning for dimensional and continuous emotion recognition.” In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 19–26 (cit. on pp. 11, 16).

[Eme00] N. J. Emery. “The eyes have it: the neuroethology, function and evolution of social gaze.” In: *Neuroscience & biobehavioral reviews* 24.6 (2000), pp. 581–604 (cit. on p. 14).

[EWS10] F. Eyben, M. Wöllmer, B. Schuller. “Opensmile: the munich versatile and fast open-source audio feature extractor.” In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462 (cit. on pp. 17, 39).

[HZRS19] J. Han, Z. Zhang, Z. Ren, B. Schuller. “EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings.” In: *IEEE Transactions on Affective Computing* 12.3 (2019), pp. 553–564 (cit. on pp. 11, 17, 32, 35, 38).

[IB09] R. J. Itier, M. Batty. “Neural bases of eye and gaze processing: the core of social cognition.” In: *Neuroscience & Biobehavioral Reviews* 33.6 (2009), pp. 843–863 (cit. on pp. 11, 14).

[Kel95] D. Keltner. “Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame.” In: *Journal of personality and social psychology* 68.3 (1995), p. 441 (cit. on pp. 11, 14).

[LZL+21] J. Liang, Y.-Q. Zou, S.-Y. Liang, Y.-W. Wu, W.-J. Yan. “Emotional gaze: The effects of gaze direction on the perception of facial emotions.” In: *Frontiers in psychology* 12 (2021), p. 684357 (cit. on p. 15).

[MAV+15] P. M. Müller, S. Amin, P. Verma, M. Andriluka, A. Bulling. “Emotion recognition from embedded bodily expressions and speech during dyadic interactions.” In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 663–669 (cit. on p. 11).

[MHLB11] M. Milders, J. K. Hietanen, J. M. Leppänen, M. Braun. “Detection of emotional faces is modulated by the direction of eye gaze.” In: *Emotion* 11.6 (2011), p. 1456 (cit. on pp. 11, 14).

[MZA+19] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic. “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2630–2640 (cit. on p. 18).

[NCZ17] A. Nagrani, J. S. Chung, A. Zisserman. “Voxceleb: a large-scale speaker identification dataset.” In: *arXiv preprint arXiv:1706.08612* (2017) (cit. on pp. 3, 12, 21, 22, 30, 39).

[OMF18] J. O’Dwyer, N. Murray, R. Flynn. “Affective computing using speech and eye gaze: a review and bimodal system proposal for continuous affect prediction.” In: *arXiv preprint arXiv:1805.06652* (2018) (cit. on p. 15).

[OMF19] J. O’Dwyer, N. Murray, R. Flynn. “Eye-based Continuous Affect Prediction.” In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 137–143 (cit. on pp. 15, 16, 24, 27).

[PSCH05] M. Pantic, N. Sebe, J. F. Cohn, T. Huang. “Affective multimodal human-computer interaction.” In: *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 669–676 (cit. on p. 11).

[PVZ15] O. M. Parkhi, A. Vedaldi, A. Zisserman. “Deep face recognition.” In: (2015) (cit. on pp. 17, 26, 39).

[RBH+20] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris, et al. “Avlnet: Learning audio-visual language representations from instructional videos.” In: *arXiv preprint arXiv:2006.09199* (2020) (cit. on p. 18).

[RSSL13] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions.” In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, pp. 1–8 (cit. on pp. 11, 15, 16).

[SOA21] L. Schoneveld, A. Othmani, H. Abdelkawy. “Leveraging recent advances in deep learning for audio-visual emotion recognition.” In: *Pattern Recognition Letters* 146 (2021), pp. 1–7 (cit. on pp. 11, 16).

[SRS22] M. C. Schiappa, Y. S. Rawat, M. Shah. “Self-supervised learning for videos: A survey.” In: *ACM Computing Surveys* (2022) (cit. on pp. 12, 18).

[VSP+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 42).

[VYL+19] T. Van Huynh, H.-J. Yang, G.-S. Lee, S.-H. Kim, I.-S. Na. “Emotion recognition by integrating eye movement analysis and facial expression model.” In: *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*. 2019, pp. 166–169 (cit. on pp. 15, 16).

[ZPRH07] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang. “A survey of affect recognition methods: audio, visual and spontaneous expressions.” In: *Proceedings of the 9th international conference on Multimodal interfaces*. 2007, pp. 126–133 (cit. on p. 11).

All links were last followed on October 14, 2022.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature