Pietro Bernardo Santos Masur

EmotiRam-FAU: Leveraging Facial Action Units Activation Knowledge for Emotion Recognition on Faces

Universidade Federal de Pernambuco
www.cin.ufpe.br/~graduacao

Recife
2022

Pietro Bernardo Santos Masur

EmotiRam-FAU: Leveraging Facial Action Units Activation Knowledge for Emotion Recognition on Faces

A B.Sc. Dissertation presented to the Centro de Informática of Universidade Federal de Pernambuco in partial fulfillment of the requirements for the degree of Bachelor in Computer Science.

Area: *Affective Computing, Machine Learning*
Advisor: *Prof. Veronica Teichrieb (UFPE)*
Co-Advisor: *Prof. Lucas S. Figueredo (UFRPE)*
Co-Advisor: *Willams Costa (UFPE)*

Recife
2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Masur, Pietro Bernardo Santos.
EmotiRam-FAU: Leveraging Facial Action Units Activation Knowledge for Emotion Recognition on Faces / Pietro Bernardo Santos Masur. - Recife, 2022.
37 : il., tab.

Orientador(a): Veronica Teichrieb
Cooorientador(a): Lucas Silva Figueredo
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Ciências da Computação - Bacharelado, 2022.

1. Computação Afetiva. 2. Aprendizagem de Máquina. 3. Visão Computacional. 4. Unidades de Ação Facial. 5. Reconhecimento de Emoções. I. Teichrieb, Veronica. (Orientação). II. Figueredo, Lucas Silva. (Coorientação). III. Título.

000 CDD (22.ed.)

Folha de Aprovação

EmotiRam-FAU: Leveraging Facial Action Units Activation Knowledge for Emotion Recognition on Faces

Trabalho apresentado à disciplina de Trabalho de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Prof. Veronica Teichrieb (UFPE)

Data de aprovação: 20/10/2022

Banca Examinadora:

Professora Veronica Teichrieb
Professor Cleber Zanchettin

Recife - 2022.1

Dedico este trabalho a todos aqueles que contribuíram em minha árdua e amarga jornada de estudos formais: meu avô (in memoriam), minha avó, meu pai e meu amor.

ACKNOWLEDGEMENTS

Agradeço, em primeiro lugar a Deus, por me conceder o fôlego de vida e sustentar a minha existência. À santíssima, puríssima, bem-aventurada Mãe de Deus e sempre Virgem Maria eu agradeço por sempre interceder em favor de mim à Santíssima Trindade. Essas intercessões, são com toda a certeza, o que me permitiram realizar a conclusão desse trabalho, e o que me permitiu estar com vida para tal, através da graça da Santíssima Trindade.

Agradeço também muitíssimo a Santo Agostinho, meu onomástico, que através de sua biografia e intercessão me conduziu à Santa Igreja e me deu razões para viver.

Esse trabalho, selo final de minha educação superior, pôde ser concluído graças a uma miríade de pessoas da minha vida. Em primeiro lugar, preciso agradecer a toda a minha família que sempre esteve a meu lado e me concedeu o suporte necessário para ir em frente. Especialmente, sou muitíssimo grato ao meu falecido avô que me ensinou o que é ser um homem e ter responsabilidade, *não com palavras, mas com poder*. Também agradeço a minha avó, por sempre estar ao meu lado e me ajudar em tudo. Agradeço também a meu pai, que me ensinou muitas coisas e em tantas outras me corrigiu.

Agradeço a Luana Espíndola, meu amor, por estar ao meu lado em tudo e sempre me levar à alegria.

Agradeço a meu irmão, Leão Masur, que compartilhou comigo muito conhecimento.

Agradeço também a todos os meus amigos, que sempre me acompanharam e tornaram mais leve a carga de cursar a universidade.

Agradeço a todos os membros do Voxar Labs que me acessoraram e ajudaram em muitas questões técnicas e acadêmicas. Em especial, agradeço ao Professor Willams Costa , meu co-orientador, por me auxiliar na pesquisa. Também ao Professor Lucas Figueredo, por me dar o suporte necessário para realizar a conclusão de tudo a tempo. Agradeço também a Daniel Perazzo, por sua prestatividade e por me ajudar com as ferramentas para escrita desse documento.

Agradeço também ao professor Antônio Nigro, do departamento de música. Este é um verdadeiro professor, que me mostrou o que é ser mestre no que faz e trabalhar com seriedade e paixão.

“ When a man is just born, he is weak and flexible. When he dies, he is hard and insensitive. When a tree is growing, it’s tender and pliant. But when it’s dry and hard, it dies. Hardness and strength are death’s companions. Pliancy and weakness are expressions of the freshness of being. Because what has hardened will never win “

–Andrei Tarkovsky

ABSTRACT

People naturally understand emotions, thus permitting a machine to do the same could open new paths for human-computer interaction. Facial expressions can be very useful for emotion recognition techniques, as these are the biggest transmitters of non-verbal cues capable of being correlated with emotions. Several techniques are based on Convolutional Neural Networks (CNNs) to extract information in a machine learning process. However, simple CNNs are not always sufficient to locate points of interest on the face that can be correlated with emotions. In this work, we intend to expand the capacity of emotion recognition techniques by proposing the usage of Facial Action Units (AUs) recognition techniques to recognize emotions. This recognition will be based on the Facial Action Coding System (FACS) and computed by a machine learning system. In particular, our method expands over EmotiRAM, an approach for multi-cue emotion recognition, in which we improve over their facial encoding module.

Keywords: human behavior recognition, emotion recognition, facial unit activation, deep learning.

RESUMO

As pessoas entendem as emoções naturalmente, portanto, permitir que uma máquina faça o mesmo pode abrir novos caminhos para a interação humano-computador. As expressões faciais podem ser muito úteis para técnicas de reconhecimento de emoções, pois são as maiores transmissoras de pistas não verbais capazes de serem correlacionadas com emoções. Diversas técnicas são baseadas em Redes Neurais Convolucionais (CNNs) para extrair informações em um processo de aprendizado de máquina. No entanto, CNNs simples nem sempre são suficientes para localizar pontos de interesse no rosto que possam ser correlacionados com emoções. Neste trabalho, pretendemos ampliar a capacidade das técnicas de reconhecimento de emoções propondo o uso de técnicas de reconhecimento de Unidades de Ação Facial (UAs) para reconhecer emoções. Esse reconhecimento será baseado no Facial Action Coding System (FACS) e computado por um sistema de aprendizado de máquina. Em particular, nosso método se expande sobre o EmotiRAM, uma abordagem para reconhecimento de emoção multi-cue, na qual melhoramos seu módulo de codificação facial.

Palavras-chave: reconhecimento de comportamento humano, reconhecimento de emoções, ativação de unidades faciais, aprendizagem profunda.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1

INTRODUCTION

Emotions are an inherent aspect of a human being's life. They are an indicator of one's inner state, which is constantly communicated through verbal and nonverbal cues in one's speech. The ability to recognize what is being expressed by an individual is vital for communication to happen. In particular, reading someone's emotional state is essential to correspond with them in the correct tone. For example, when talking to someone sad, people act differently than when talking to someone happy.

Knowing the user's emotions during or after a software interaction can be precious information. On its recommendation system, Netflix includes feedback given by the user on the movies displayed for him. Facebook, YouTube, and much other software have similar approaches to their recommendation algorithms [5]. In the examples cited, the emotion displayed by the user is collected through heuristics based on the user's input during his interaction. However, for some applications, gathering information on the user's feelings in a more direct way may be interesting. For instance, educational, safety aid, and entertainment softwares may benefit from having insight into the user's displayed emotions while watching a lesson [19]. Another approach to acquire feedback on the users' emotions is to ask them for it directly, however, this approach depends on the users' will to collaborate and cannot be utilized in all cases. Many components go into building a robust solution for emotion recognition; for instance, the metrics and outputs from such a model must be consistent with psychological principles. One of the main ways to classify emotions is by placing them in a valence and arousal space [8]. Valence is linked to the pleasure or displeasure a given feeling contains, whereas arousal is related to the intensity of the emotion. Figure 1 shows some emotions and how they are placed on the valence and arousal plane.

Many publicly available emotion recognition data sets, such as AffectNet [15], utilize valence and arousal as their annotations. However, others, such as CAER-S [10], rely on a more direct approach by simply labeling emotions according to their categories. The main difference between the two is that CAER-S-based models are categorical, whereas models based on valence and arousal are continuous. Both approaches have been widely used as the base for machine learning (ML) models.

When it comes to performing recognition of emotions on faces, another fundamental of

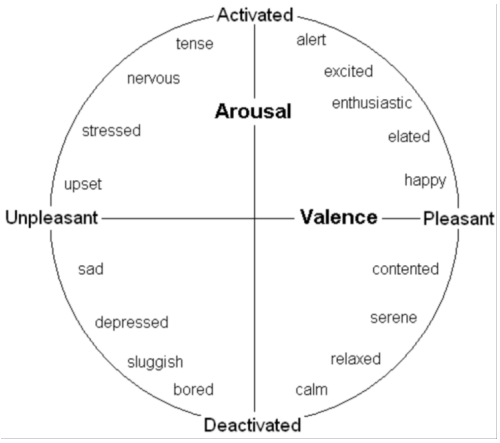

Figure 1: Valence and arousal graph showing different emotions and their positions [8].

psychology can be applied: the Facial action encoding system (FACS) [3]. FACS is a system capable of translating someone's non-verbal cues into an emotion. It was coded based on empirical observation of subjects' facial and corporal displays while feeling different emotions.

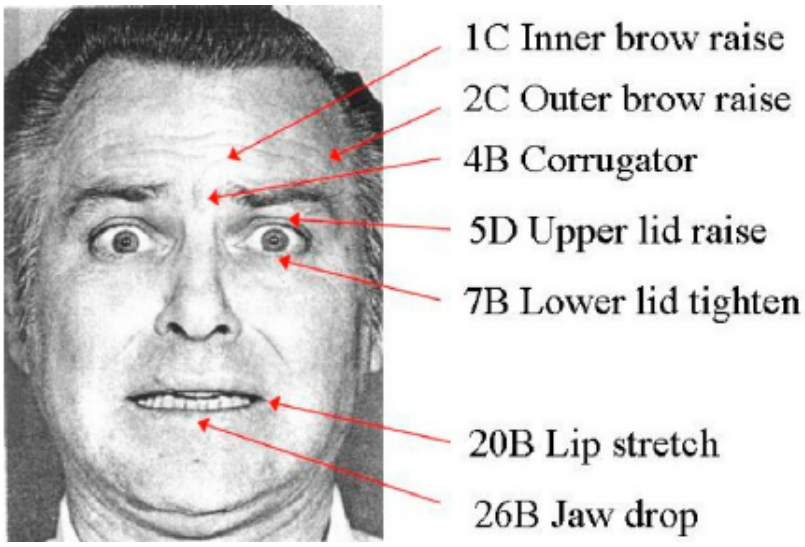

Figure 2: The FACS encoding of fear, from *Littlewort et al.* [11].

FACS encodes specifically the behaviors behind the activation of AUs; those associated with facial muscles are referred to as facial action units (FAUs). When someone feels an emotion, his body and face give a physiological response to that feeling in the form of action units. The response intensity varies according to each subject's physiology, but the correlation between emotions and action unit activation is universal. Besides, since this activation is a physiological response, it cannot be posed consistently. Thus, AUs analysis is the most reliable way to recognize someone's emotion through the facial display.

In the affective computing literature, many techniques to predict FAUs have arisen. Similar to the emotion recognition techniques, many of them rely on CNNs as their main resource. However, most recent approaches based on Graph Neural Networks (GNN) have shown some significant improvements over traditional techniques [6, 12]. The objective of such

techniques is to tell what FAUs are active and their intensity, given a face image. To this end, specific datasets have been built. Two of the most utilized academic works are DISFA [13] and BP4D [22].

Given the known correlation between FAUs and emotions, in this work, we propose that models trained to recognize FAUs can provide a better face encoding to perform emotion recognition than the naive application of encoding techniques. To the best of our knowledge, this is the first time such a hypothesis has been tested.

1.1 OBJECTIVES AND METHODOLOGY

In this work, we aim to evaluate the hypothesis that ML models trained to recognize FAUs provide a more efficient face encoding than naive emotion-recognition techniques. We also sketch a brief investigation on the possibility of introducing explicability into the EmotiRAM [2] method for emotion recognition, proposed by *Costa et al.*. This explicability involves identifying what FAUs are responsible for the recognized emotion outputs. We present these hypotheses backed by psychological fundamentals of emotion recognition and the FACS system. To check the validity of our thesis, we need to answer the following questions.

1. On performance: Initializing an emotion recognition model with a pre-trained FAU recognition model provides a better result for emotion recognition tasks? (performance)

2. On validation of FAU as useful for emotion recognition: After performing training to detect facial emotions as its labels, is the model still able to perform well on FAUs activation recognition tasks?

To perform our evaluation, we shall use the EmotiRAM's face encoding module, which we will refer to as EmotiRAM-f, and test it in combination with two different FAU techniques: ME-GraphAU [12] and Facial Action Unit Intensity Estimation via Semantic Correspondence Learning with Dynamic Graph Convolution [6]. Firstly we shall train EmotiRAM-f in its original implementation. Then, we will perform a new training for both of the analyzed AU techniques and see how the model results behave. After the retraining process, we will once again evaluate the model in its original FAUs recognition task.

2

BACKGROUND CONCEPTS

In this chapter, we review some concepts that will be helpful during this document and some of the theoretical background essential to our research.

2.1 FACIAL ACTION CODING SYSTEM

As discussed in the introduction, FACS is a manual observation system to code facial actions. It was developed based on empirical observation, and further research has shown that FACS holds even on subjects from far-off and isolated civilizations [4]. However, different subjects express themselves with varying levels of expressiveness [6].

Action units are fundamental actions from individual muscles or groups of muscles. They can be activated at five different levels ranging from a slight trace of contraction (A) to maximum contraction (E). FACS also describes action descriptors; they describe unitary movements like *forward thrusting movement of the jaw*. In this work, we will be focusing our efforts only on those emotions that can be expressed exclusively through AUs. Crossing information from AUs activation and their respective intensities, it is possible to determine what sentiment is being expressed in a face image. In Table 1, we display a table that illustrates some AUs combinations and what emotion is associated with them.

Emotion	AU
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+17
Contempt	R12A+R14A

Table 1: Activated AUs and their associated emotions.

According to *Sanchez et al.* [17], since AUs are facial muscle actions, they're naturally correlated with other parts of the face. Action Units rarely occur alone, and most of the time

appear combined with others. Thus, when trying to extract knowledge from a facial display, it is important to consider this fact.

2.2 HEATMAP REGRESSION

A heatmap is a graphical data representation that relies on colors to simultaneously express data location and intensity. Heatmaps can be used on the AU recognition field to depict what areas on a face contain AU activations and their intensity, as shown in 3. Noticing the fitting of applying heatmaps to represent AUs, *Sanchez et al.* [17] proposed using heatmaps as labels to the AU location problem. This idea came from the fact that AUs rarely occur alone; thus, taking their correlation into account can be a rich feature for ML models.

In heatmap regression employed for AUs, different map sections are activated according to the AU's location and intensity value on the FACS scale. The inferring of the AUs consists of, after locating the AU landmarks, outputting of the intensity of the active AUs.

Figure 3: Heatmap for AU10, predicted with *Yingruo Fan et al.* [6] method.

2.3 GRAPH NEURAL NETWORKS

Graphs are a mathematical structure that is very useful for representing correlations. They are built from vertices and edges; edges are paths between vertices that can be weighted. In computing, this structure is commonly utilized to represent problems related to distances or costs between points. One of the most notable problems that showcase the use of graphs is the traveling salesman. It proposes an apparently simple question: Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city? The class of graphs that match this description is

called a Hamiltonian cycle. Oddly, finding Hamiltonian cycles on a graph is an open question and pertains to the NP-hard class.

Many times, data sample types are collected in the form of graphs. However, translating graphs into tensors is not trivial, and many of them can lead to information loss. To address that, recently, graphs neural networks (GNNs) have been proposed [18] as an alternative to CNNs. GNNs are a way of performing ML considering the topological relation between data encoded in the form of graphs.

For FAU-related ML problems, many approaches that utilize GNNs in their architecture were proposed [6, 12]. Those approaches aim to take advantage of graph encoding to extract knowledge from FAUs mutual correlations.

3

RELATED WORK

We dedicate this chapter to some works related to our research. We review two of the most utilized datasets in the non-posed facial expressions recognition field: DISFA [13], and BP4D [22]. A dataset for emotion recognition in a context, CAER-S [10], is also reviewed. Finally, we present techniques that are related to our work.

3.1 DATASETS

Well-labeled datasets on the facial expression field are essential for progress in this area. There are not many datasets built aiming to allow for exploration of non-posed expressions [13]. Labeling facial expression data requires the job of expert personnel, which can be costly. Therefore, there is a distinction between datasets built on non-controlled settings (in the wild), which usually make no distinction between posed and non-posed expressions, and datasets built on controlled settings, which have more precise labeling and a controlled generation process.

3.1.1 Datasets for action units activation

There are various datasets available in the literature for recognizing facial expressions. They are not the same as datasets for recognizing AUs. AUs datasets are usually built in controlled environments by inducing the participants to genuinely feel emotions and making the correct labeling of associated AUs. Here we disclose two of the most utilized datasets for AUs recognition tasks. We utilize these two datasets as a basis for fine-tuning in emotion recognition tasks.

3.1.1.1 DISFA

DISFA [13] is a dataset built from the collection of videos from twenty-seven young adults. The participants were recorded while videos aiming to induce spontaneous expressions were shown to them. The recordings were made with two cameras, one on the right and the other one on the left side. The data collected was labeled by FACS experts, frame by frame, to indicate the presence (or absence) and intensity of FAUs. DISFA contains labels for 12 AUs: 5 on the

upper face and 7 on the lower face. In joint with BP4D, DISFA is a default benchmark for AUs recognition techniques.

3.1.1.2 BP4D

BP4D [22] is a database of 2D and 3D videos. Similar to DISFA, BP4D contains labels for non-posed expressions. However, it is a much larger database. The dataset was motivated by the observation that since faces are 3D deformable objects, 2D video may be insufficient to encode all features needed for AUs activation research.

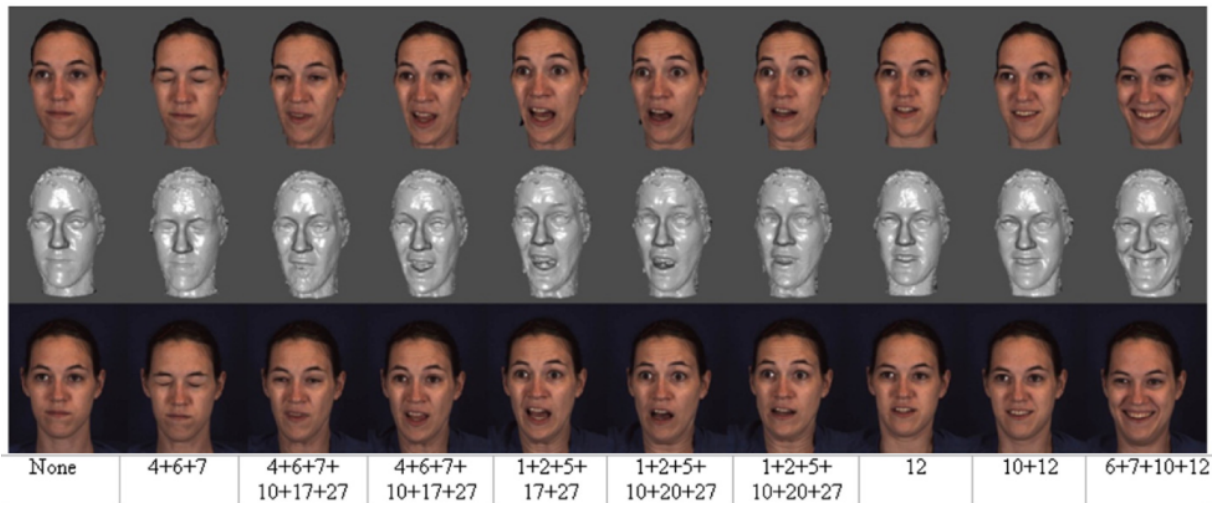

Figure 4: The figure, collected from the BP4D publication [22], displays set of entries on the BP4D database. The rows display the textured 3D models, shaded 3D models, original 2D videos, and annotated AUs.

The dataset comprises artifacts sampled from forty-one participants, all being young adults. They were shown videos aiming to elicit them into feeling a given emotion; while watching the video, they were recorded. FACS experts annotated the collected data. Figure 4 displays a sample from entries at the BP4D dataset.

3.1.2 Datasets for emotion recognition

Here we review some datasets made specifically for emotion recognition tasks. They differ from the AUs datasets in their purpose and labeling process.

3.1.2.1 EMOTIC

EMOTIC [9] is the first dataset to propose the use of context on emotion recognition tasks. It is an in-the-wild dataset and contains labels of 26 different emotion categories as well as labels with continuous valence, arousal, and dominance values. Valence and arousal were

already discussed in this work introduction. Dominance metric measures the level of control a person feels over the situation. It ranges from submissive to dominant.

3.1.2.2 CAER benchmark

CAER [10] stands for context-aware emotion recognition. In this work, the authors propose a network (CAER-Net) capable of inferring the emotion of people on a picture, taking the face and the image context into consideration. To benchmark their results, they have built a database of video clips.

The CAER benchmark comprises videos collected from 79 TV shows, totaling 20,484 clips. They were manually annotated with six different emotion labels. A static version of the CAER benchmark, CAER-S, was also proposed.

Something that must be taken into consideration is that CAER and CAER-S annotations are not very precise and can sometimes be inconsistent with what the image is presenting.

3.1.3 EmotioNet database

Unlike the other reviewed datasets, EmotioNet database [1] comprises both AUs and emotion labels. It is an in-the-wild dataset made through the collection of images from the internet. They've utilized WordNet [14] to search for images related to specific emotions and performed ML-based AU recognition to annotate AU labels. It is a very useful dataset for both AU and emotion recognition fields.

3.2 MULTI-CUE EMOTION RECOGNITION

Multi-cue emotion recognition refers to techniques that consider not only people's faces in a given image to extract their emotions but also other elements of the image, such as context and body pose. In this section, we will review the CAER-Net-S [10] technique and EmotiRAM (Multi-cue adaptive emotion recognition network) proposed *Willams L. Costa et al.* [2] .

3.2.1 Context-aware emotion recognition

Lee et al. [10] proposes that facial encoding applied on its own is limited due to not considering the whole scene where the emotion happens. The context of a given scene is understood as everything that is not the face of the person whose emotion is being analyzed.

To achieve their objective, the researchers have utilized two separate encoders, one for context and one for faces. The two encoders' outputs are then passed into an adaptive fusion network, which combines the encoders' features.

This article's advances are based on an insight very significant to the emotion recognition field: context may be essential to recognize emotions. This was originally proposed by EMOTIC

[9] However, as seen in the next subsection, the context can be better considered if a more sagacious encoding is performed.

3.2.2 Multi-cue adaptive emotion recognition network

The technique proposed in this article [2] extends the idea behind CAER. Its pipeline works by encoding an image in three different modules: face, context, and body encoding stream. Figure 5 displays the EmotiRAM architecture in its fullness.

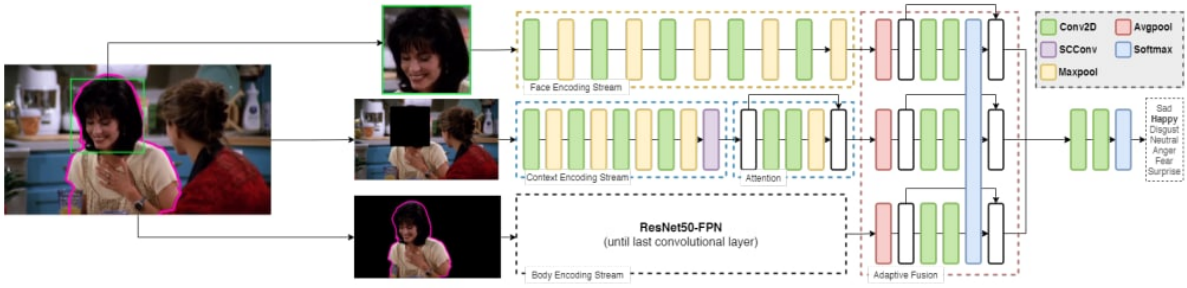

Figure 5: Architecture described in the EmotiRAM article. [2]

Like CAER-Net, all the encoding streams receive a different image section and perform learning separately. The encoding results are then input into an adaptive fusion network.

To evaluate the impact of each encoding stream on the final accuracy results, the article evaluates EmotiRAM in various module combination settings. One considers only the face, the other, the face and context, and finally, the full approach with face, context, and body is evaluated. With its approach, EmotiRAM improved significantly over CAER-Net-S, as shown in Table 2.

Method	Accuracy
CAER-Net-S	0.74
MCAER-Net (face+context)	0.87
MCAER-Net (face+context+body)	0.89

Table 2: Here are displayed the accuracy metrics achieved by EmotiRAM compared to the CAER-Net-S method on the CAER-benchmark dataset.

3.3 MACHINE LEARNING TECHNIQUES FOR ACTION UNITS

Despite the inherent correlation between emotions facial display and AUs activation, these two problems are usually treated with different techniques. This section aims to give a brief overview of three articles that we revised and utilized as references for developing our work.

3.3.1 Heatmap regression

In the publication entitled "Joint Action Unit localisation and intensity estimation through heatmap regression"[17] the authors proposed a supervised learning approach for learning AU heatmaps. To make the labels, first, an estimation of the facial landmarks was made, then 2D Gaussians were drawn around the points where AUs are known to cause changes. Their technique was validated on the BP4D dataset.

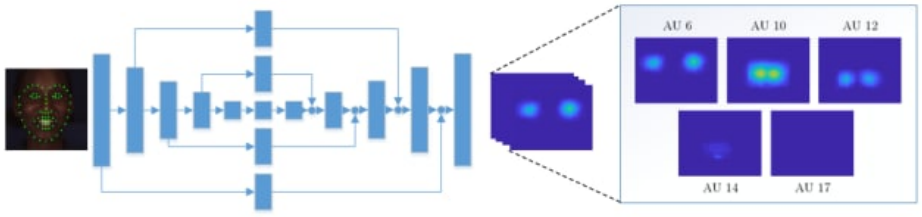

Figure 6: Hourglass network, generating the heatmaps corresponding to the Action Units. Figure taken from *Sánchez-Lozano et al.*[17].

The article's insight is that AUs are jointly activated. Thus, it is possible to model AUs more realistically by considering them all when performing regression. The network utilized in this work is a single Hourglass network [16], as shown in figure 6. In combination with their label design and pre-processing techniques, the hourglass network sufficed to surpass state-of-the-art results.

However, as shown in the next section, defining labels with a pre-made rule is not the better approach for jointly regressing AUs. Each face contains a unique setting; taking this into consideration provides a fuller approach to this problem.

3.3.2 Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution

Relying on pre-defined rules for modeling AU co-occurrence leads to limited generalization. To contour that, this article proposes a novel GNN-based technique [6]. Their framework is such that latent relationships of AUs are automatically learned via establishing semantic correspondences between feature maps. This is achieved by employing a combination of heatmap regression and Semantic Correspondence Convolution (SCC). SCC modules are based on the work by *Wang et al.* [20], which focuses on dynamic graph convolutions in geometry modeling. Intuitively, feature channels that are simultaneously activated are considered to have a latent co-relation; this corresponds to a co-occurrence pattern of AU intensities.

The method basic framework is built by adding several de-convolutional layers on a ResNet [7], which generates the AUs feature maps. Feature maps are then input into SCC modules 3 times. The network outputs one heatmap per AU. On figure 7 the full architecture is

displayed.

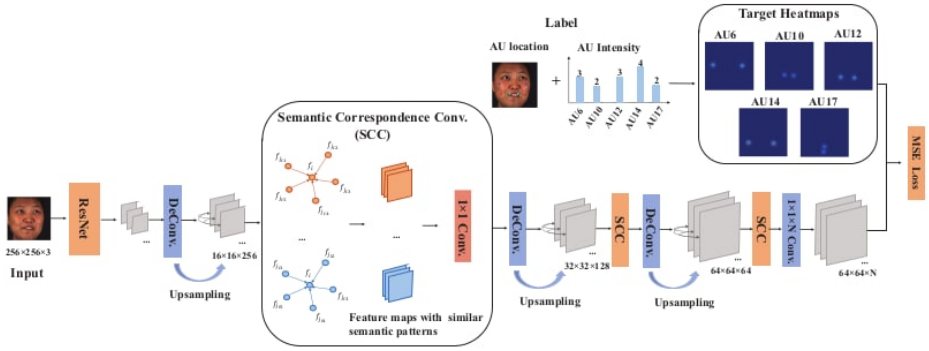

Figure 7: Display of the network architecture taken from the original article publication. [6]

3.3.3 Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition

Considering each facial topology when building AU graphs is essential for building a robust GNN-based technique. However, the number of edges between the AUs may also be a very important factor in the representation ability of this graph. *Luo et al.* [12] propose that a single edge between AUs in a graph representation is insufficient for dealing with their complex relations. To address this problem, they propose a strategy that, for each AU pair, encodes a pair (two edges) of multi-dimensional features, as shown in figure 8.

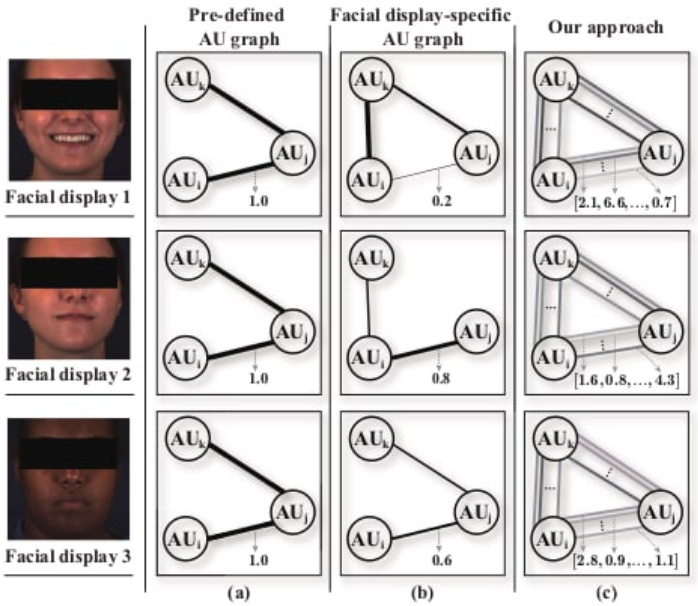

Figure 8: Graph representation of faces in three different strategies:(a) Pre-defined AU graphs; (b) Facial display-specific AU graph; (c) Facial display-specific multi-dimensional edge. [12]

To achieve their objective and deal with the problems they've encountered with previous approaches to model AUs relationships as graphs, the authors utilize three resources:

1. Modeling a full face representation into an explicitly AUs relationship describing graph through the Attention Node Feature Learning (ANFL) module

2. Modeling the relationship between AUs pair on a multi-dimensional edge feature graph with Multi-edge Feature Learning (MEFL) module

3. Considering the uniqueness of each facial display by utilizing full face representation as input to the two modules listed above

The ANFL module is composed of two blocks:

1. AU-specific Feature Generator (AFG): composed of Fully Connected (FC) and global average polling layers which jointly act as a feature extractor

2. Facial Graph Generator (FGG): computes node similarity via KNN and GCN. This block is only used as a reinforcement to the AFG block; thus it is not used on training

MEFL module also has two blocks:

1. Facial Display-specific AU Representation Modeling (FAM) Receives full face representation and facial features from the AFG FC layer and locates activation cues of each AU

2. AU Relationship Modeling (ARM)

The ARM outputs are utilized to extract features from the located cues which relate to both AUs activations. On Figure 9 we can see the full architecture.

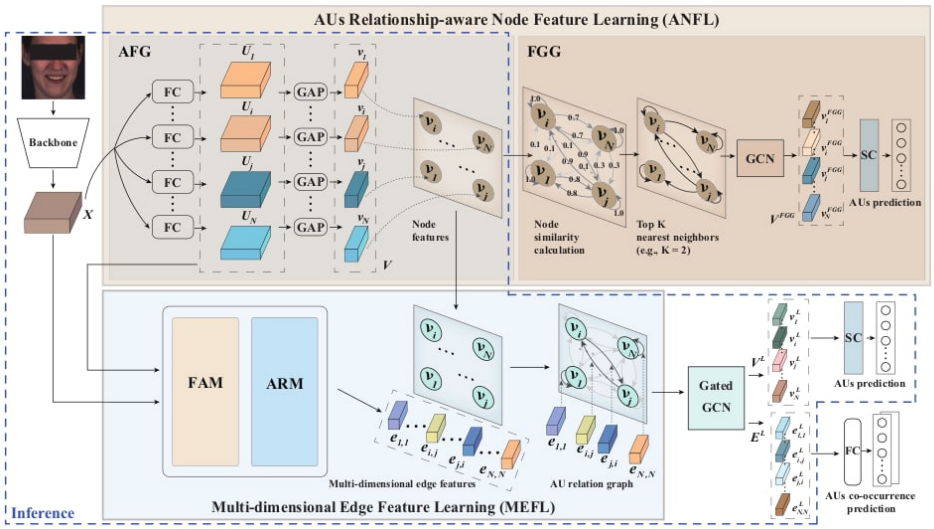

Figure 9: The pipeline of the AU relationship modeling approach proposed in the article. Only the modules and blocks within the blue dashed lines are used at the inference stage. [12]

3.3.4 Comparison between machine learning techniques for AUs

We have chosen to evaluate three methods that introduced significant landmarks in the AU recognition field from our review. From all the reviewed techniques, *Fan et al.* [6] method provides the best approach regarding accuracy to the FAU recognition problems. Despite the evaluation metrics utilized on *Luo et al.* [12] differing from the other two articles, we make that remark based on the experiments performed during this work (which shall be detailed in the next section). Table 3 describes the results achieved by each method on the BP4D and DISFA datasets.

	DISFA	BP4D
Sánchez-Lozano et al.[17]	N/A	0.68
Fan et al.[6]	0.47	0.72

Table 3: Score reported on DISFA and BP4D datasets. *Sánchez-Lozano et al.* and *Fan et al.* metrics are the ICC score.

4

METHODS

Our work's main goal is to achieve better accuracy on the face encoding module over *Costa et al.* [2] technique (EmotiRAM). To achieve that, we combine FAU ML techniques with raw emotion recognition and train a model to perform emotion recognition tasks. Our experiments are performed on the CAER-S [10] dataset.

Furthermore, to analyze how much the final model relies on knowledge of action units to determine the emotion output, we perform some qualitative experiments on both DISFA [13] and CAER-S datasets.

4.1 INTUITION

As shown by *Ekman et al.* [3], emotions can be recognized based on the AUs activated. Furthermore, the emotion displayed by the AUs are non-posed, thus, AUs consists a better source for emotion inference than naive facial observation.

Transporting this scientific observation into the ML field, we propose that encoding action units consist of a better source of information than naive facial encoding techniques. This is the postulate we depart from to develop our EmotiRAM-FAU technique.

4.2 DATA PRE-PROCESSING

It is very important to pre-process the images to feed the models with the correct data. In this section, we describe data pre-processing performed on CAER-S and DISFA in our work.

4.2.1 CAER-S

In the CAER-S dataset, we utilize the same pre-processing approach made by *Costa et al.* in their work on the face-encoding stream. We had access to their original code and implemented our technique on its top. The pre-processing consists of making facial crops and resizing them to 96x96 resolution.

4.2.2 DISFA

In the experiments performed in the DISFA dataset, we followed the pre-processing approach utilized by *Luo et al.* [12]. Only the left-side videos from the dataset are used. Firstly, each video frame is converted into an image. Then, these images are fed into MTCNN [21] for face detection and alignment. Then, the images are cropped to 256x256 resolution. Figure 10 showcases this pre-processing pipeline.

Figure 10: Pre-processing pipeline on DISFA samples.

4.3 EMOTIRAM-FAU

Our best results were achieved by utilizing *Fan et al.* [6] technique in combination with a bottleneck as our pipeline. As described in the literature review section, *Fan et al.* relies on adding de-convolutional layers on a ResNet backbone to generate heatmaps. Our approach extends this architecture by taking the generated heatmap as input to a bottleneck of fully connected (FC) layers.

A bottleneck of FC layers is a crucial component to achieving our objective of utilizing information encoded by FAU knowledge to classify emotion. Data from the heatmap is not in the expected shape to be related to one emotion, so our bottleneck works as a translator of FAU encoding information into emotions. To achieve this, our model as a whole (including the pre-trained model) is fine-tuned to recognize emotions.

As described in the experiments section, we have experimented with several different bottlenecks. We will now describe the architecture of our best approach.

The heatmap generated by the pre-trained model has a shape of 10x24x24. This is a fixed shape that is restrained by the output of *Fan et al.* architecture for the BP4D dataset. The shape is flattened to a resulting 5760x1 shape and passed through the following arbitrated FC layers.

1. In 5760, Out 2048
2. In 2048, Out 102
3. In 1024, Out 512
4. In 512, Out 256
5. In 256, Out 7

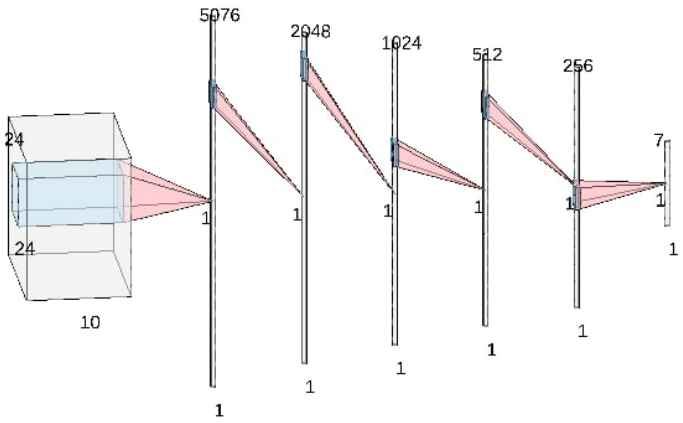

Figure 11: Graphical representation of our architecture. The first box displays the output received by *Fan et al.* technique. The other layers are our described bottleneck.

The output shape is seven because of the number of classes in the CAER-S dataset. Image 11 displays our implemented architecture. From now on, we will refer to EmotiRAM-FAU with *Fan et al.* technique simply as EmotiRAM-FAU.

5

EXPERIMENTS

5.1 DATASETS

In our work, we had access to DISFA [13] and CAER-S [10] database. CAER-S was split into three partitions for performing the experiments: train, validation, and test. Reported statistics were mainly achieved on the test partition. We utilized models pre-trained on FAUs datasets with ML AU recognition techniques to implement our idea. Our architecture consists of inserting data into the pre-trained model and passing them through a bottleneck of FC layers. The bottleneck outputs represent the detected emotions.

We also wished to utilize BP4D [22] during our work; however, this dataset is not easily accessible. Due to its large volume of data and the potentially sensible data of participants in the dataset, the author limits the distribution only to University professors or their students through a very bureaucratic process. It is required that people who wish to gain access to the dataset send a physical hard drive for them to record data on. The hard drive is then returned to the solicitors.

5.2 FRAMEWORK

The base for the code we utilized is the same of *Costa et al.* [2]. Voxar Labs provided us with a machine to run the experiments. We have utilized the PyTorch python library as our ML framework and WandB to track our results. Experiments were performed on a Ubuntu 20.04 machine with an RTX 2080 Ti GPU, 24GB of RAM, and a 4-core processor with 8 threads.

5.3 EXPERIMENTAL SETTINGS

We have experimented with three different approaches based on ResNet-50 for FAU recognition. First, the work by *Fan et al.* [6], in which we used a pre-trained model in the BP4D dataset, and also the work by *Luo et al.* [12] with a pre-trained model in the BP4D dataset and another pre-trained model in the DISFA dataset. *Luo et al.* model will be further referred to as ME-graph. We could not gain access to *Fan et al.* model pre-trained on DISFA, and thus it was not evaluated in this work.

1. *Fan et al.* [6] model pre-trained on BP4D
2. *Luo et al.* [12] model pre-trained on DISFA
3. *Luo et al.* [12] model pre-trained on BP4D

Each of these models is separately inserted on EmotiRAM's architecture in joint with a bottleneck of FC layers, it functions as a replacement for the face-encoding module. The fine-tuned models are then evaluated in the CAER-S dataset for emotion recognition. *Luo et al.* technique output is a vector of probabilities for each AU on the dataset. Thus, in our experiments with it, only one FC layer could be utilized.

5.4 EXPERIMENTS OVERVIEW

In this section, we will disclose the training process of our models. Our first step was to replicate *Costa et al.* results on the EmotiRAM-f. After that, we started to test our novel technique.

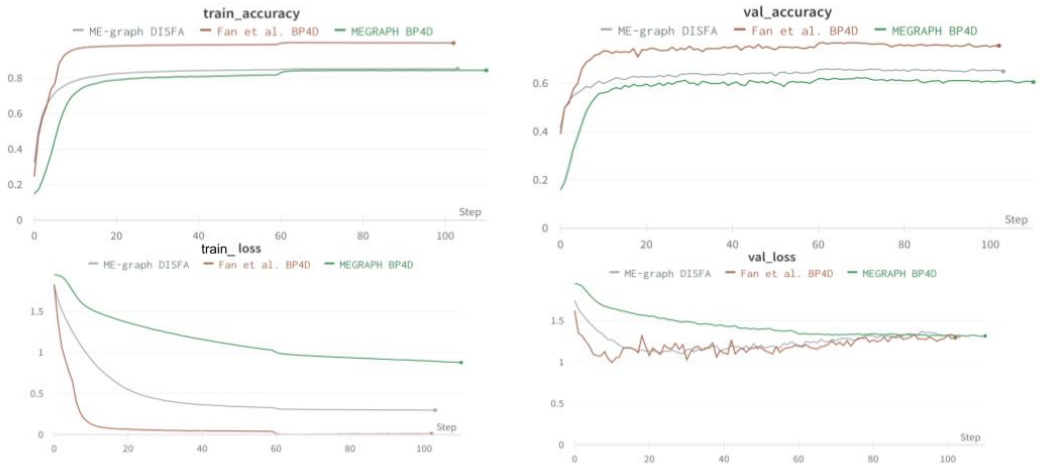

Figure 12: Graphics with metrics achieved during training epochs. Several experiments were performed, and this graph displays our run with the best results in each approach.

In Figure 12, we display some charts with metrics achieved during our experiments: train accuracy, validation accuracy, train loss, and validation loss. In our experiments, we have utilized cross entropy as our loss function, the batch size is 32, and the RMSProp optimizer, the same ones utilized by *Costa et al.*. We have varied values for learning rate and weight decay during our experiments.

Notice that the architecture relying on *Fan et al.* pre-trained on BP4D has, by far, outperformed the ME-graph approach. However, also notice that around epoch 58, *Fan et al.* model has started to suffer from overfitting. Thus, to evaluate our results, we utilized the model from the epoch immediately before.

5.5 EXPERIMENTS ON EXPLAINABILITY

Since our work relies on AUs, we've resolved to see how our system co-relates AUs and emotions, despite no mechanism to control that being utilized. The pre-trained models we use give output information about AUs. In particular, *Luo et al.* model outputs a vector of probabilities of the activations of each AU. For this experiment, we only utilize ME-graph models pre-trained on DISFA since it is the only AU dataset we have access to.

5.5.1 AU experiments on DISFA

To evaluate how much the fine-tuned model still relies on the AUs original representation, we perform experiments on the DISFA dataset. For this end, a publicly available implementation of *Luo et al.* code was utilized. First, we evaluate metrics for the original pre-trained ME-graph model. Then, we evaluate our fine-tuned model, truncating its bottleneck to fit with the network architecture required by *Luo et al.* code.

5.5.2 AU experiments on CAER-S

In this experiment, we set EmotiRam-FAU to provide two outputs: recognized AUs and recognized emotions. We work with data we have access to for building an explainable program based on EmotiRam-FAU. DISFA contains labels for a limited range of AUs; on *Luo et al.* work, the training process is limited to 8 AUs. Matching the AUs covered by ME-Graph with the emotions provided by the CAER-S dataset, we observe that the only full match between activated AUs and corresponded emotion is Happiness, as shown in Table 1. We label CAER-S happiness data as being (AU6+AU12) and evaluate the correctness of the AU output made by the FAU encoding module after our fine-tuning process. We also perform the same experiments with the original ME-graph pre-trained model and use it as a baseline. In the next section we will display our achieved results and perform some qualitative evaluations over selected images of the CAER-S dataset.

6

RESULTS

In this chapter, we discuss the results achieved in the experimentation of our approach using an FAU method, comparing it with the original baseline work, EmotiRAM-f, and other baseline approaches for FAU. By making those comparisons, we will see what impacts FAU encoding has on emotion recognition tasks and how FAU models are affected by fine-tuning for recognizing emotions.

6.1 RESULTS ON CAER-S

In this section, we detail the results achieved on CAER-S in terms of accuracy. Table 4 shows the results we acquired with each technique.

	CAER-S
EmotiRAM-f	0.70
EmotiRAM-FAU [BP4D]	**0.77**
EmotiRAM-FAU-ME-graph [BP4D]	0.62
EmotiRAM-FAU-ME-graph [DISFA]	0.66

Table 4: Results achieved on the CAER-S test partition, in terms of accuracy.

As shown, EmotiRAM-FAU improved 7 percent over EmotiRAM-f from *Costa et al.* [2] in terms of accuracy. This is a significant result and points out that models pre-trained on AUs can provide a better encoding than naive techniques.

The results achieved with *Luo et al.* technique were not as good as the others. We believe this occurred because their network outputs are shaped as AUs probabilities, while *Fan et al.* works with a heatmap richer in features as their output. Thus, when coupled with a bottleneck, *Luo et al.* output features have little margin to find new vector spaces representing emotions. From now on, we will refer to EmotiRAM-FAU with *Luo et al.* technique as EmotiRAM-FAU-ME-Graph.

6.2 RESULTS ON AUS

To evaluate how our model behaved on the AUs recognition task, we performed a test on the DISFA dataset. As previously explained could not work with the BP4D dataset due to our lack of access to the original database.

6.2.1 Results on DISFA

Table 5 display the results achieved by EmotiRAM-FAU-ME-Graph and compares them with the original *Luo et al.* model (ME-Graph).

	F1 Avg	F1 AU6	F1 AU12	Acc Avg	Acc AU6	Acc AU12
EmotiRAM-FAU-ME-Graph	0.18	0.0067	0.56	0.62	0.91	0.87
ME-Graph	0.58	0.64	0.82	0.94	0.94	0.94

Table 5: Results achieved on the CAER-S test partition, in terms of accuracy and F1 score. AUs 6 and 12 are responsible for the happiness facial expression, wich will be evaluated on the next section.

As we can see, the F1 score and accuracy metrics have greatly decayed compared to the original approach. There are some reasons we believe may have caused this. As previously observed, CAER-S is an on-the-wild dataset ensembled mainly from TV shows, thus, it is fair to expect some samples to contain posed emotions. Furthermore, since the annotation of CAER-S was not made by FACS experts, we believe that they are far from ideal in terms of correspondence between the real displayed emotion (based on the FACS system) and the labeled emotion. It is also reasonable to suppose that, if our hypothesis is true, there is no reliable way to determine whether CAER-S data provides a balanced distribution of FAUs activations examples. Thus, it is natural to suppose those factors lead our model to suffer some degree of forgetfulness at its original task. We believe that having more reliable training data could greatly benefit our model approach.

6.2.2 Qualitative Results on CAER-S

We have utilized data labeled as happiness on the CAER-S dataset and assumed that their corresponding activated AUs were AU6 and AU12. We then made a script to perform an accuracy test based on the emotion and AUs outputs produced by EmotiRAM-FAU. Table 6 shows our achieved results. Figure 13 shows a piece of data from the CAER-s dataset, which was correctly labeled for both AUs and emotions by our approach. On the other hand, figure 14 displays an image our model failed to recognize as having the correct AUs.

Notice that EmotiRAM-FAU-ME-Graph results at recognizing both AUs were near zero. This comes from the previously discussed forgetfulness suffered by the model. In particular,

Figure 13: A happiness-labeled image from the CAER-S dataset, which our system correctly labeled. Notice that AU12 is the main responsible for the happiness on this photo, and the AU6 region (eyes), in which our system has low performance, is covered by sunglasses.

Figure 14: A happiness labeled image from the CAER-S dataset, which our system mislabeled. Notice that in this sample, the main sign of happiness is given by AU6, the same which our system practically lost the ability to recognize.

	AU6 and AU12	AU6 or AU12
ME-Graph [DISFA]	0.52	0.98
EmotiRAM-FAU-ME-Graph [DISFA]	0.1	0.92
ME-Graph [BP4D]	0.49	0.81

Table 6: For the AU6 and AU12 metrics, we consider the model is right if it can recognize the activations of both AUs, the other metric considers recognizing at least one AU as a correct answer.

AU6 was greatly affected by this forgetfulness, as shown in table 5. This is the cause of such a low score in recognizing both AUs.

Unfortunately, due to time and accessible data limitations, it was impossible to conduct more extensive and quantitative research on the effectiveness of our system. In addition, the lack of annotated data and emotions compatible with the AUs learned by the ME-Graph technique was a significant drawback to our research.

7

CONCLUSIONS

7.1 CONSIDERATIONS

In this work, we have investigated the use of AUs to recognize emotions in ML models. We have shown how models pre-trained to recognize AUs fine-tuned for emotion recognition tasks provide a better basis for emotion recognition than training architectures from scratch.

Our approach is explicitly aimed at emotion recognition. However, we have also investigated how much the final model intermediary representations are still faithful to the original AUs representation. Our results showed that, when fine-tuned for emotions, the model still retains some ability to recognize AUs. This is an important step towards explicable systems for emotion recognition.

7.2 FUTURE WORKS

7.2.1 AU based emotion recognition

A significant drawback of our research was the lack of AU emotions in the CAER-S dataset. This has limited the scope of our AU-emotion correlation only to the happiness emotion. Further research could deepen the investigation of how important AUs are for emotion recognition via ML. For that, it is essential to utilize datasets with direct correspondence between emotions and activated AUs, such as EmotioNet [1]. Other ML architectures can also be investigated.

7.2.2 Explainable emotion recognition

To build explainable emotion recognition models, it is essential to investigate how they can achieve emotions and AUs consistency. Simultaneously considering AUs and emotions on the training loss can be a potential path towards that.

Another potential work in this field is utilizing two separate modules for recognizing AUs and emotions. Despite not correlating AU-emotion information, this could be a practical approach for retrieving both data.

REFERENCES

[1] Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5562–5570.

[2] Costa, W., Macêdo, D., Zanchettin, C., Figueiredo, L. S., & Teichrieb, V. (2021). Multi-cue adaptive emotion recognition network.

[3] Ekman, P., . F. W. V. (1978). Facial action coding system (facs).

[4] Ekman, P. (1972). Universals and cultural differences in facial expressions of emotions. *Nebraska Symposium on Motivation*, 207–282.

[5] FadhelAljunid, M. & Manjaiah, D. (2017). A survey on recommendation systems for social media using big data analytics. *International Journal of Latest Trends in Engineering and Technology*, 48–58.

[6] Fan, Y., Lam, J. C. K., & Li, V. O. K. (2020). Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition.

[8] Kadar, M., Gutierrez y Restrepo, E., Luis-Ferreira, F., Calado, J., Artifice, A., Sarraipa, J., & Jardim-Goncalves, R. (2016). Affective computing to enhance emotional sustainability of students in dropout prevention. 85–91.

[9] Kosti, R., Alvarez, J., Recasens, A., & Lapedriza, A. (2019). Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[10] Lee, J., Kim, S., Kim, S., Park, J., & Sohn, K. (2019). Context-aware emotion recognition networks.

[11] Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1.

[12] Luo, C., Song, S., Xie, W., Shen, L., & Gunes, H. (2022). Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*.

[13] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160.

[14] Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

[15] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

[16] Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation.

[17] Sánchez-Lozano, E., Tzimiropoulos, G., & Valstar, M. F. (2018). Joint action unit localisation and intensity estimation through heatmap regression. *CoRR*, abs/1805.03487.

[18] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

[19] Vinola, C. & Vimaladevi, K. (2015). A survey on human emotion recognition approaches, databases and applications. *ELCVIA : Electronic Letters on Computer Vision and Image Analysis*, 24–44.

[20] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds.

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

[22] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., & Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706. Best of Automatic Face and Gesture Recognition 2013.