A study on Human Emotion Recognition in Video Images using Deep Learning

by

JARGALSAIKHAN ORGIL

Doctoral Thesis

Graduate School of Advanced Technology and Science
Tokushima University

Doctor of Engineering

in

Systems Innovation Engineering

March, 2022
Tokushima University
Japan

Abstract

From the beginning of this century, Artificial Intelligence (AI) has evolved to handle problems in image recognition, classification, segmentation, etc. AI learning is categorized by supervised, semi-supervised, unsupervised or reinforcement learning. Some researchers have said that the future of AI is self-awareness, which is based on reinforcement learning by rewards based on task success. Moreover, it is said that the reward would be harvested from human reactions, specially emotion recognition. On the other hand, emotion recognition is a new inspiring field, but the lack of enough amount of data for training an AI system is the major problem. Fortunately, in the near future, it will be necessary to correctly recognize human emotions because image and video dataset availability is rapidly increasing.

Emotions are mental reactions (such as anger, fear, etc.) marked by relatively strong feelings and usually causing physical reactions to previous actions in a short time duration focused on specific objects. In this Work, we are focusing on emotion recognition using face, body part, and intonation.

As stated earlier, automatic understanding of human emotion in a wild setting using audiovisual signals is extremely challenging. Latent continuous dimensions can be used to accomplish the analysis of human emotional states, behaviors, and reactions displayed in real-world settings. Moreover, Valence and Arousal combinations constitute well-known and effective representations of emotions. In this thesis, a new Non-inertial loss function is proposed to train emotion recognition deep learning models. It is evaluated in wild settings using four types of candidate networks with different pipelines and sequence lengths. It is then compared to the Concordance Correlation Coefficient (CCC) and Mean Squared Error (MSE) losses commonly used for training. To prove its effectiveness on efficiency and stability in continuous or non-continuous input data, experiments were performed using the Aff-Wild dataset. Encouraging results were obtained.

The contributions of the proposed method Non-Inertial loss function are as follows:

1. The new loss function allows for Valence and Arousal to be viewed together.

2. Ability to train on less data.
3. Better results.
4. Faster training times.

The rest of this thesis explains our motivation, the proposed methods and finally presents our results.

Acknowledgement

This research work could not have been possible without the help of some special and important people. At first, I would like to express the deepest appreciation to my family and friends.

Thanks to Professor Dr. Stephen Karungaru, my supervisor. Thank you for your encouragement and provision of a conducive research environment. Without your support and advice, this dissertation would not have been possible. Thank you for providing me with all the necessary facilities.

To Professor Dr. Kenji Terada, thank you for your professional advice, and to Dr. Akinori Tsuji, for your help, worries, and good comments. In addition, a thank you to Professor Shagdar Ganbold, who introduced me to B1 lab, and whose help had a lasting effect.

I would like to express my gratitude to the Mongolia-Japan higher Engineering Education Development project (MJEED), the Mongolian and the Japanese governments, who gave me chance to study in Japan.

I would also like to thank all B1 lab members, who studied with me, and thanks for a happy three years with you.

Contents

List of Figures

List of Tables

Publications

- Main

  1. “Multimodal Emotion Recognition Using Non-Inertial Loss Function”, Jargalsaikhan Orgil, Stephen Karungaru, Kenji Terada and Ganbold Shagdar, Journal of Signal Processing, Vol.25, Page 73-85, March, 2021, Published.

- Others

  1. “Facial emotion recognition accuracy improvement using deep learning”, Jargalsaikhan Orgil, Stephen Karungaru, ACEAT, Dec 2019, Page 88-93, Kyoto, Japan. Presented.
  2. “Continual learning: Linear layer classifier concatenation using image processing transform functions”, Jargalsaikhan Orgil, Stephen Karungaru, Kenji Terada, QCAV2021, June 2021. Presented.

Chapter 1

Introduction, Background and Related Works

The use of smart technology in society is growing rapidly, and as the industry develops, so does the need for technology that is able to assess the needs of potential customers and select the most suitable solutions for them. In human-machine interaction, to understand each other, emotions have played an important predictive role. Emotions are an extremely complex brain function that reacts to previous actions based on the human brain functional system, called the "Limbic System". Recently, there has been an enormous interest in this emotion recognition field.

As shown in Table 1.1[1], emotion recognition model input parameters can be linked to whether they provide the Valence or Arousal aspect quantification information. Emotional valence describes the extent to which an emotion is positive or negative, whereas arousal refers to its intensity or the strength of the associated emotional state.

Table 1.1: Emotion related parameters.

Emotion-related effect	Arousal	Valence
Emotion induced sweating	+	
Breathing rhythm variations	+	+
Heart rate variability	+	+
Blood pressure	+	
Core body temperature	+	
Heart rate	+	
Facial expression		+
Facial muscle activity	+	
Voice intonation	+	+
Questionnaire	+	+

The following relations are used in this study: facial expressions, facial muscle motion (around the lips, eyes and nose), voice intonation and body posture.

1.1 Emotion recognition from images

1.1.1 Limbic system

The limbic system [2] takes a leading role in emotional regulation. It is also responsible for the regulation of reflexive and endocrine function in response to emotional stimuli. It consists four main parts: the amygdala, the hippocampus the thalamus, and the hypothalamus.

Thalamus - collects sensory inputs like visual, audio and modality and its neurons can deliver project impulses to amygdala and better cortical regions for processing.

Amygdala - collects sensory inputs from the thalamus and coordinates responses to the environment, largely contributing to the processing of fear and anger. When the amygdala switches to the hypothalamus state, hormones e.g. adrenaline are released. In addition, the amygdala has a primary role in fear learning by the association between particular situations when fear is developed.

Hippocampus - Its functions are maintaining memories and transforming immediate memory to episodic memory. Essentially, memory (episodic memory, in particular) contributes abundantly to decision-making as emotions instance by memories assist upcoming behavior, immediate or episodic. Even if the hippocampus is broken, the brain loses the ability to make new memories but is still able to retain its episodic term memory.

Hypothalamus (less than 1% of brain weight), plays a huge role in adjusting various functions in the body muscle system. Dependent on emotions, it adjusts the Autonomic Nervous System (ANS) by adjusting the endocrine system which is engaged in the ease of different hormones into the bloodstream. Totally, it could be said that the hypothalamus is engaged in the expression of emotions instead of their generation. The lateral portions of the hypothalamus is associated closely feelings of pleasure whilst the median part is connected with more negative emotions.

The ANS commands the reflexive physiological changes that happen in response to emotional stimuli. It has two arms – the sympathetic arm and the parasympathetic arm. The sympathetic nervous system is associated with a fear response such as 'pass or reject' while the parasympathetic nervous system applies to the calm state and may be mentioned to as 'rest and digest'.

The limbic system is shown in Fig. 1.1.

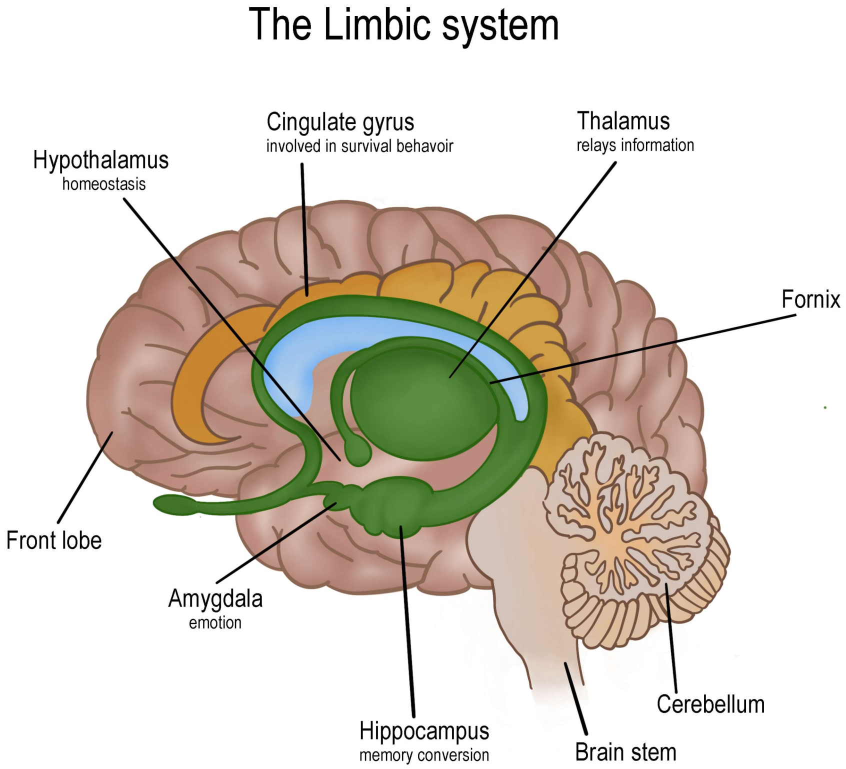


Figure 1.1: Limbic system.

1.1.2 Action units, Micro-expressions and Coding systems

The Facial Action Coding System (FACS) is a complex, anatomical system that describes all visible facial movements [3], [4]. It divides facial expressions into independent components of muscle movement, called Action Units (AUs). FACS refers to the movement of the facial muscles in response to emotions. Originally created by Carl-Herman Hjortsjö with 23 facial motion units in 1970, it was subsequently developed further by Paul Ekman, and Wallace Friesen.

FACS is a model that compares live emotions to facial expressions. It records how the facial muscles work and how their patterns evolve (the contraction or relaxation of facial muscles). FACS experts who have been trained will be able to tell the difference between a fake and voluntary smile (Pan American World Airways (Pan-AM Smile)) and a genuine and involuntary grin (Duchenne Smile). FACS classifies all visibly discernible facial activity into forty-four separate action units (AUs), as well as numerous different types of head and eye postures and movements. Every AU has a unique numerical code (the designation of that is fairly arbitrary).

Table 1.2 list the AUs coded in FACS, as well as the muscle groups involved in each action. It is crucial to point out that while FACS is anatomically based, there is not a 1:1 correspondence between muscle groups and AUs. This is because of the fact that a given muscle could act totally different in numerous ways for different regions—to manufacture visibly different actions. An example of this case is the frontalis muscle. Contraction of the medial portion of the frontalis muscle raises the inner corners of the supercilium solely (producing AU 1), whereas contraction of the lateral portion of the frontalis raises the outer brow (producing AU 2). FACS committal to writing procedures conjointly provide the intensity of every facial action on a 5-purpose intensity scale, for the temporal arrangement of facial actions, and for the writing of facial expressions in terms of "events". An occurrence is the AU-based description of every countenance, which can comprises one AU or several AUs contracted as one expression. FACS can verify the displayed feeling of a participant. This analysis of facial expressions is one among only a few techniques accessible for assessing emotions in a time period facial diagnostic technique (fEMG is another option). Alternative measures, like interviews and psychology tests, should be completed once a stimulant has been conferred. This delay creates an additional barrier to determining how a person feels in direct reaction to a stimulant. The benefits and downsides of measuring facial behavior with the assistance of FACS include different empirical cryptography schemes [5], etc.

FACS versus MAX

MAX refers to the maximally discriminative facial movement cryptography system. FACS and MAX are both observational coding schemes that describe expression in terms of constituent components. Componential coding systems can be classified on the basis of the manner in which they were derived: theoretically or anatomically, and according to whether they code facial behavior selectively or comprehensively. Theoretically derived procedures are those that classify facial components based on assumptions about which parts of the face should be engaged in certain stages (e.g., emotion). Since it codes only those face configurations that Izard [5] theorized, it corresponds to universally recognized facial expressions of emotion. Izard's MAX is theoretically derived. The problem with theoretically generated systems is that they can't detect behaviors that haven't been pre-determined. That is, they were created to explain whether the face accomplishes things it should do based on a certain theory, rather than to record all the face can do. In this sense, MAX and other theoretically based coding schemes like [6] are selective by definition. Moreover, because they do not comprehensively code all face movement, selective systems are time and effort efficient. However, if unfavorable results are obtained, they are as costly. For instance, if a researcher uses a selective coding system and fails to find "fear" expressions in subjects who were exposed to a supposedly frightening stimulus, the researcher will be unable to determine whether the subjects did experience fear or if the coding system simply failed to capture the

Table 1.2: Main facial action codes with muscular parts.

AU number	FACS name	Muscular baisis
0	Neutral face	
1	Inner brow raiser	frontalis (pars medialis)
2	Outer brow raiser	frontalis (pars lateralis)
4	Brow lowerer	depressor glabellae, depressor supercilii, corrugator supercilii
5	Upper lid raiser	levator palpebrae superioris, superior tarsal muscle
6	Cheek raiser	orbicularis oculi (pars orbitalis)
7	Lid tightener	orbicularis oculi (pars palpebralis)
8	Lips toward each other	orbicularis oris
9	Nose wrinkler	levator labii superioris alaeque nasi
10	Upper lip raiser	levator labii superioris, caput infraorbitalis
11	Nasolabial deepener	zygomaticus minor
12	Lip corner puller	zygomaticus major
13	Sharp lip puller	levator anguli oris (also known as caninus)
14	Dimpler	buccinator
15	Lip corner depressor	depressor anguli oris (also known as triangularis)
16	Lower lip depressor	depressor labii inferioris
17	Chin raiser	mentalis
18	Lip pucker	incisivii labii superioris and incisivii labii inferioris
19	Tongue show	
20	Lip stretcher	risorius w/ platysma
21	Neck tightener	platysma
22	Lip funneler	orbicularis oris
23	Lip tightener	orbicularis oris
24	Lip pressor	orbicularis oris
25	Lips part	depressor labii inferioris, or relaxation of mentalis or orbicularis oris
26	Jaw drop	masseter; relaxed temporalis and internal pterygoid
27	Mouth stretch	pterygoids, digastric
28	Lip suck	orbicularis oris

relevant facial behavior.

On the other hand, FACS is a complete measuring system. It measures all facial muscle movements that may be seen. Ekman and Friesen developed FACS by training themselves to isolate their face muscles and using needle EMG to ensure that their method contained every potential facial movement. As a comprehensive rather than selective system, FACS is not limited to measuring only those behaviors that are theoretically related to emotion MAX[5]. FACS allows for the discovery of new configurations of movements that might be relevant to extrinsic variables of interest. Many of the publications in this collection are excellent illustrations of FACS's use in this regard.

FACS versus EMG

In order to infer muscular activity, face electromyography (EMG) entails detecting electrical potentials from facial muscles. Surface electrodes are often put on the individual's facial skin, and variations in potential are monitored as the subject goes through various treatments (e.g., viewing emotional slides or films). The benefits of EMG are mostly based on its capacity to identify muscle activity that is not visible to the human eye and hence cannot be detected using observation-based coding methods like FACS or MAX. As a result, facial EMG appears to be a suitable measuring technique. However, there are issues with EMG, some of which may be more or less important depending on one's research questions. For starters, EMG is obtrusive—it draws attention to the fact that individuals' faces are being examined.

This might enhance the occurrence of self-conscious conduct while also potentially interfering with the behavior being investigated in other ways. Second, despite significant progress in recent years in refining EMG signals [7], there is still an issue with cross-talk, or neighboring muscular contraction (potentials) interfering with or muddying the signal of a specific muscle group. Cross-talk may misrepresent the "picture" of the contractions on the face, potentially leading to very different emotional interpretations. When it comes to emotion expressions, where slight changes in muscular activity of adjacent muscle groups can have very different emotional meanings, such cross-talk may misrepresent the "picture" of the contractions on the face, potentially leading to very different emotional interpretations.

1.2 Emotion recognition from videos

1.2.1 OpenFace

Over the past few years, there has been a growing interest in automatically analyzing and understanding facial behaviors. OpenFace [8] is presented for computer vision, mechanical engineering researchers, the Society of Computational Computation, and those interested in creating interactive applications based on facial behavior analysis. OpenFace is the first tool that can be used to identify facial landmarks, determine the position of the head, identify facial function units, and visualize using source code available for both model evaluating and training. The computer vision algorithms which represent the core of OpenFace demonstrate state-of-the-art results in all of the above mentioned tasks. Furthermore, their tool is capable of working in real time and can work from a simple webcam without any special equipment. Furthermore, their tool is capable of period performance and is in a position to run from a straightforward digital camera without any special hardware. An OpenFace extraction example is shown in Fig. 1.2.

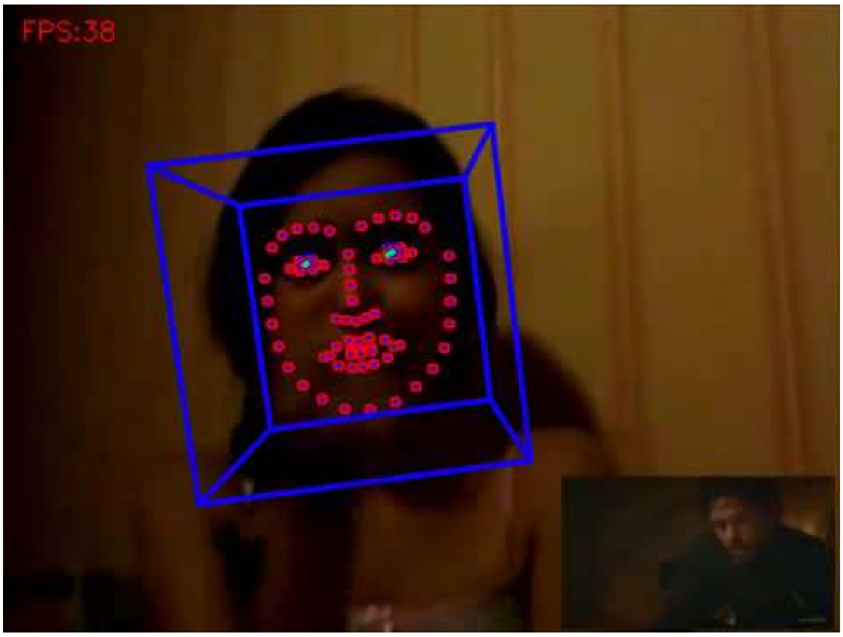


Figure 1.2: OpenFace example on Aff-Wild2 106.mp4.

1.2.2 OpenPose

Computing 2D images of real-time crowds is a key component in making the machine more aware of the people in the images and videos. OpenPose [9], concentrated on a real-time method of detecting 2D images of many people in an image. The proposed method uses a non-parametric representation called Part Affinity Fields (PAFs) to learn how to connect body parts with individuals in an image. This bottom-up system ensures high accuracy and real-time performance regardless of the number of people in the image. From their previous work, the calculation of the location of the PAF and the body part was improved simultaneously during the training phase. Instead of improving the location of PAF and body parts, authors proved that simply improving PAF significantly increases both performance and accuracy in runtime.

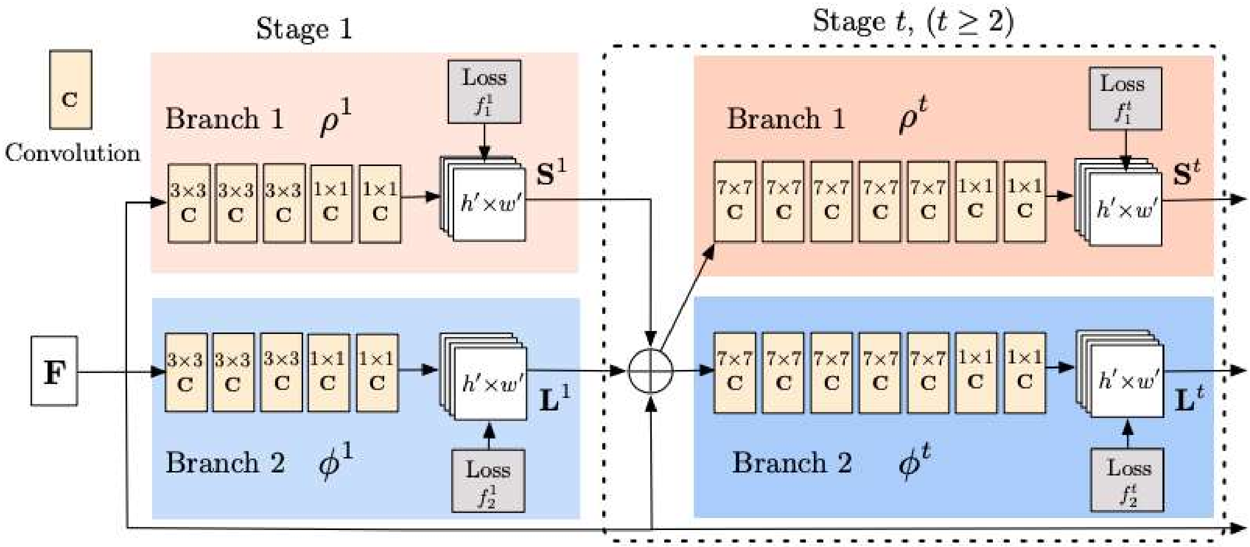


Figure 1.3: OpenPose model architecture https://bit.ly/3Er7Nz2.

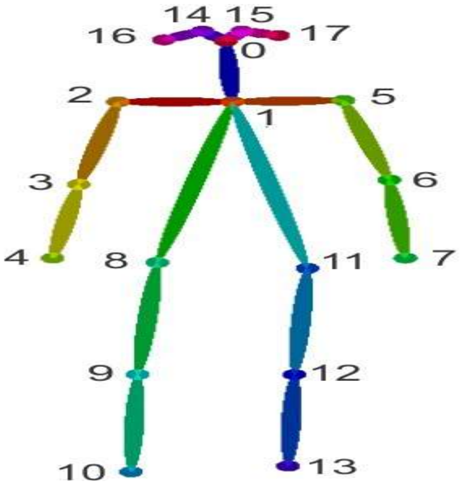


Figure 1.4: Body keypoints.

The first body and foot key sensor based on a database with internal foot explanations presented to the public are also introduced. The combined detector not only shortens the evaluation time compared to sequential operation, but also maintains the accuracy of each component separately. The work resulted in OpenPose, the first open source real-time system to detect the 2D locations of multiple people, including key points on the body, legs, arms, and face.

The OpenPose model, Fig. 1.3, has two branches with multiple stages using extracted features of pretrained Mobilenet, ResNet or VGG19. Top branch predicts body parts confidence map and bottom branch predicts affinity fields, representing the degree association between body parts, Fig. 1.4. In the first stage, the model predicts the initial states of body part of left hand with part confidence S and affinity field L. Upcoming stages produce more refined predictions of original image feature F. The detected body parts including in the face are shown in Fig. 1.5.

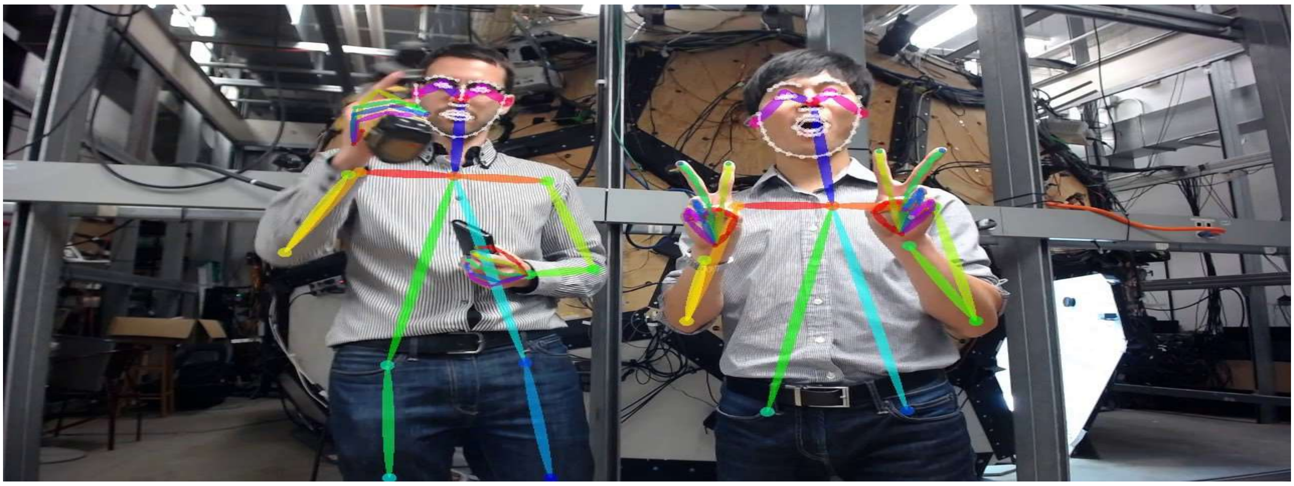


Figure 1.5: Facial and body keypoints in sample https://bit.ly/3DntJcT.

The purpose is finding the relationship between body expression and Arousal-Valence. It is possible to recognize emotion more accurately with the assistance of body posture.

1.2.3 MediaPipe Face Mesh 468

Face Mesh [10] is a face geometry resolution that estimates 468 vertices 3D face landmarks in a time period even on small and mobile devices like smartphones. It employs machine learning (ML) to infer the 3D surface, requiring solely one camera input while no zealous depth device. Utilizing light-weight model architectures on the side of GPU acceleration throughout, it delivers a time period performance essential for live experiences. The Face Mesh model is shown in Fig. 1.6.

Additionally, the answer is bundled with the Face geometry module that bridges the gap between the face landmark estimation and helpful period of time Augmented Reality (AR) applications. It establishes a metric 3D area and uses the face landmark screen positions to estimate face pure mathematics inside that area. The face pure mathematics information consists of common 3D pure mathematics primitives, together with a face cause transformation matrix and a triangular face mesh. Underneath the hood, a light-weight applied mathematics analysis technique referred to as mythical being analysis is used to drive a sturdy, performance and transportable logic. The analysis runs on central processing unit and encompasses a bottom speed/memory footprint on high cubic centimetre model logical thinking. The 3D face landmarks utilized transfer learning and trained a network with many objectives: the network at the same time predicts 3D landmark coordinates on artificial rendered information and second linguistics contours on annotated real-world information.

A clipped video frame is sent into the 3D landmark network without any extra depth information. The model returns the 3D point coordinates as well as the likelihood that a face is present and adequately aligned in the input. An inference example is shown in Fig.1.7. Predicting a 2D heat-map for each landmark is a typical option, but it is not suitable to depth prediction and has large processing costs for so many points. Figure 1.8 shows the flattened face mesh with the visualization shown in Fig. 1.9. Repeated bootstrapping and refining predictions are undertaken to increase the model's accuracy and resilience. As a result, the dataset may expand to include more difficult scenarios like grimaces, oblique angles, and occlusions [11].

The Face Mesh trained on a globally sourced dataset of around 30K in-the-wild mobile camera photos taken from a wide variety of sensors in changing lighting conditions. In this work Face Mesh model is used as feature extractor of action units. Furthermore, the results are compared to Face Mesh, OpenPose and OpenFace features used models.

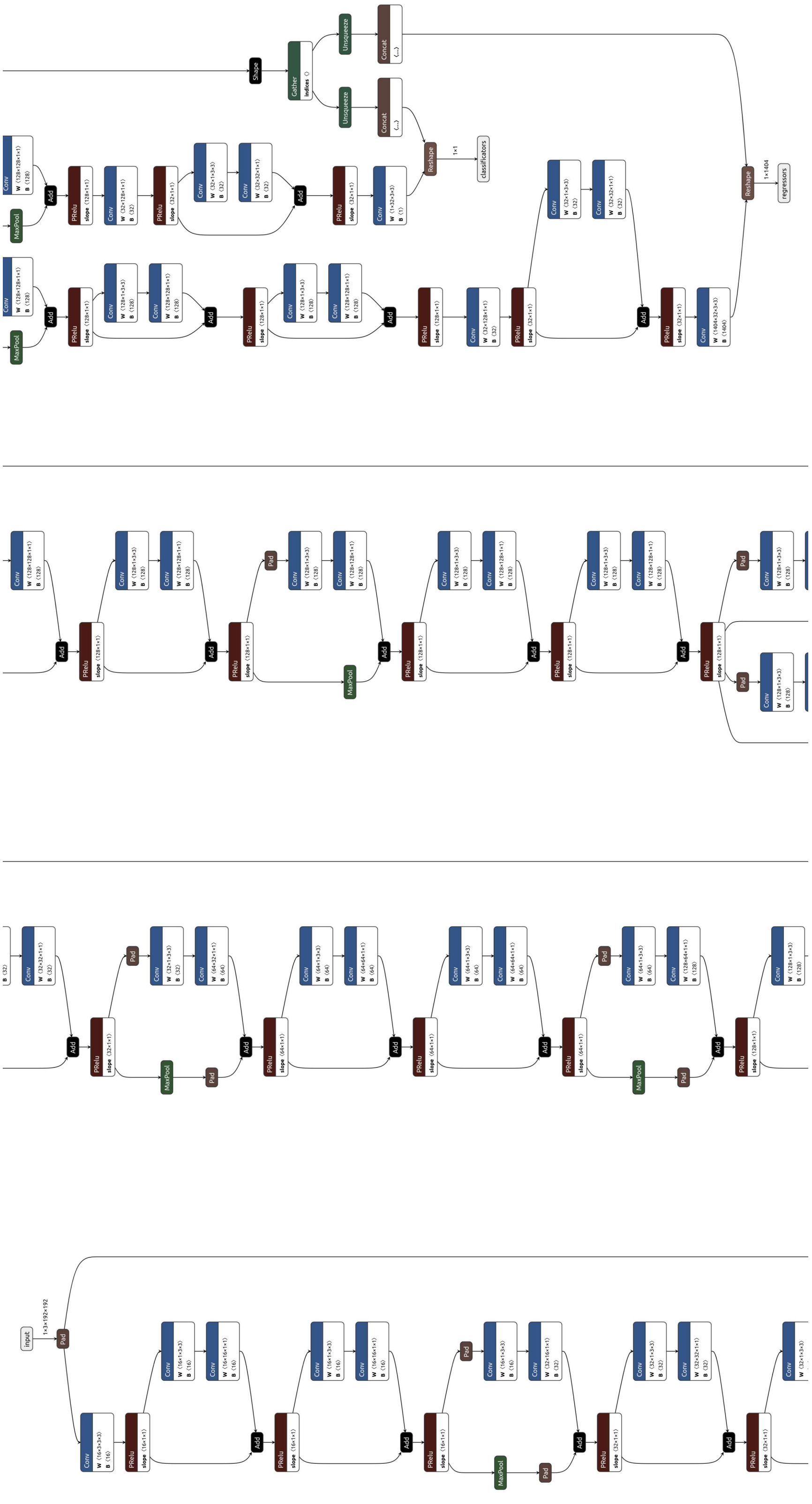


Figure 1.6: Face Mesh model visualization (right side view).

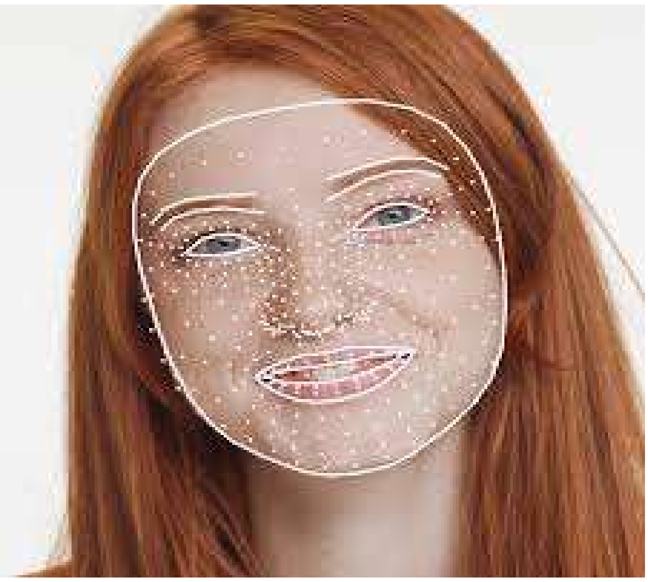


Figure 1.7: Mediapipe Face Mesh inference [11] https://bit.ly/3GcXrDg.

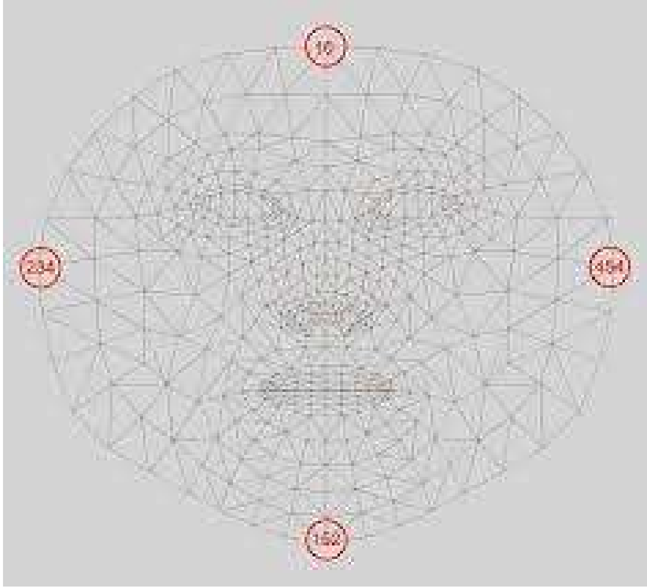


Figure 1.8: Flattened Face Mesh.

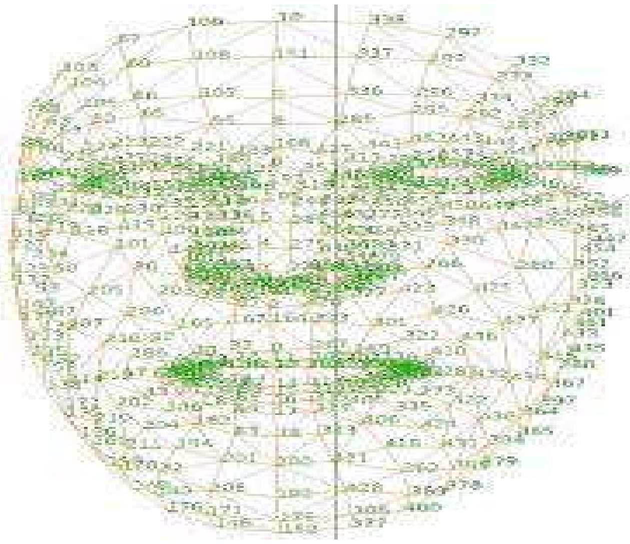


Figure 1.9: 3D visualization.

1.2.4 Mel spectrogram

Frequency Mel scale describes the perceptual distance between pitches of different frequencies. A classical approximation is to define the frequency-to-mel transform function for a frequency f as

$$m = 2595 * log_{10}(1 + \frac{f}{700}) \tag{1.1}$$

The Mel frequency Spectrogram used as input to the linear layers is generated as detailed by Promod et al. [12]. To generate spectrograms from windowed audio or speech signals sampled at 22050Hz, Short Term Fourier Transform (STFT) is applied . A "Hann" window of length 2048 is used with an STFT hop-length equal to 512. Using the Mel scale, the resulting magnitudes (128 coefficients per window are use) then mapped to get Mel spectrograms. The lower end of the frequency spectrum are emphasized over the higher ones, thus, imitating the perceptual hearing capabilities of humans. In speech emotion recognition, Chan et al. [13] previously researched this topic. A sample Spectrogram corresponding to Aff-Wild training 110.avi audio is shown below in Fig.1.10.

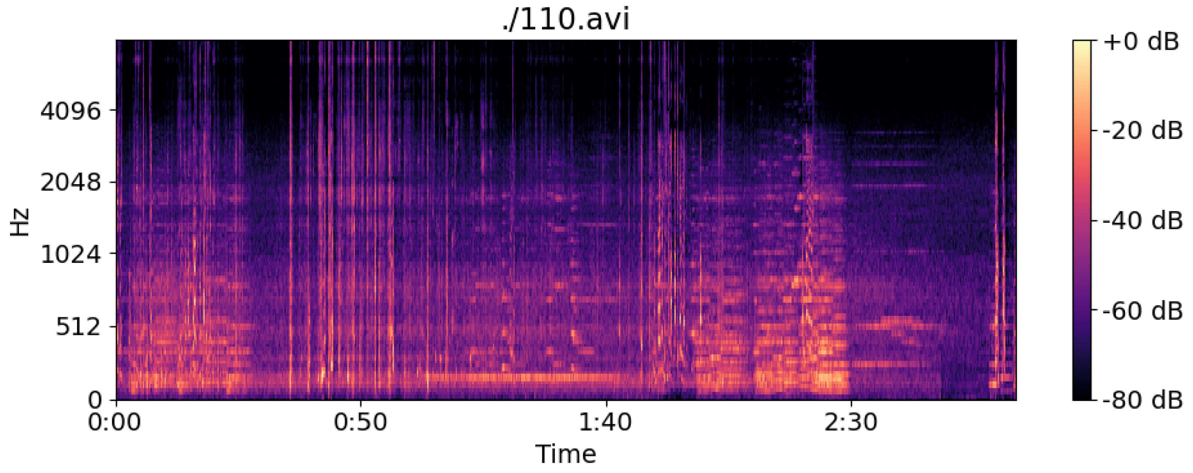


Figure 1.10: Aff-Wild training 110.avi audio spectrogram.

1.2.5 Eulerian Video Magnification

Eulerian Video Magnification (EVM) is a set of simple and robust algorithms that can reveal and analyze tiny motions effectively. It is a new type of microscope, not made of optics, but of software taking an common video as input and producing one in which the temporal changes are larger. It reveals a new world of tiny motions and color changes showing us hidden vital signs, body and face part muscle activity due to emotional reaction control signal from Hypothalamus [14]. To explain the n^{th} relationship between temporal processing and motion magnification, let $I(x,t)$ imply the image strength at position x and time t. Since the image undergoes translation motion, the observed strengths with respect to a displacement function $\sigma(t)$, such that $I(x,t) = f(x + \sigma(t))$ and $I(x,0) = f(x)$ can be expressed. The purpose of magnifying the motion is to integrate the signal

$$\hat{I}(x,t) = f(x + \alpha\sigma(t)) \tag{1.2}$$

for some amplification value α.

Assuming the image can be approximated by a first-order Taylor series expansion, the image at time t, $f(x+\sigma(t))$ in a first-order Taylor expansion about x, can be expressed as

$$I(x,t) \approx f(x) + \sigma(t)\frac{\Delta f(x)}{\Delta x} \tag{1.3}$$

Let $B(x,t)$ be the result of applying a broadband temporal band-pass filter to $I(x,t)$ at every position x (picking out everything except $f(x)$ in Eq.1.3). For now, let assume the motion signal, $\sigma(t)$, is within the pass-band of the temporal bandpass filter.

$$B(x,t) = \sigma(t)\frac{\Delta f(x)}{\Delta x} \tag{1.4}$$

to amplify that bandpass signal by α and add it back to $I(x,t)$, then result is following.

$$\hat{I}(x,t) = I(x,t) + \alpha B(x,t) \tag{1.5}$$

The full version of motion amplified signal is as follows.

$$\hat{I}(x,t) \approx f(x) + \alpha\sigma(t)\frac{\Delta f(x)}{\Delta x} \tag{1.6}$$

the simplified version is

$$\hat{I}(x,t) \approx f(x + \alpha\sigma(t)) \tag{1.7}$$

This shows that the processing magnifies motions, the spatial displacement $\sigma(t)$ of the local image $f(x)$ at time t, has been amplified to a magnitude of α.

The Eulerian approach to motion magnification, Fig. 1.11, is robust and fast, but works primarily when the motions are small. If the motions are large, this processing can introduce artifacts. However, one can detect when this happens and suppress magnification in this stabilized video case. There are limits to how well Spatio-temporal filtering can remove noise and amplified noise can cause image structures to move incoherently. The EVM motion magnifier as “Action Units” magnifier is used in this work.

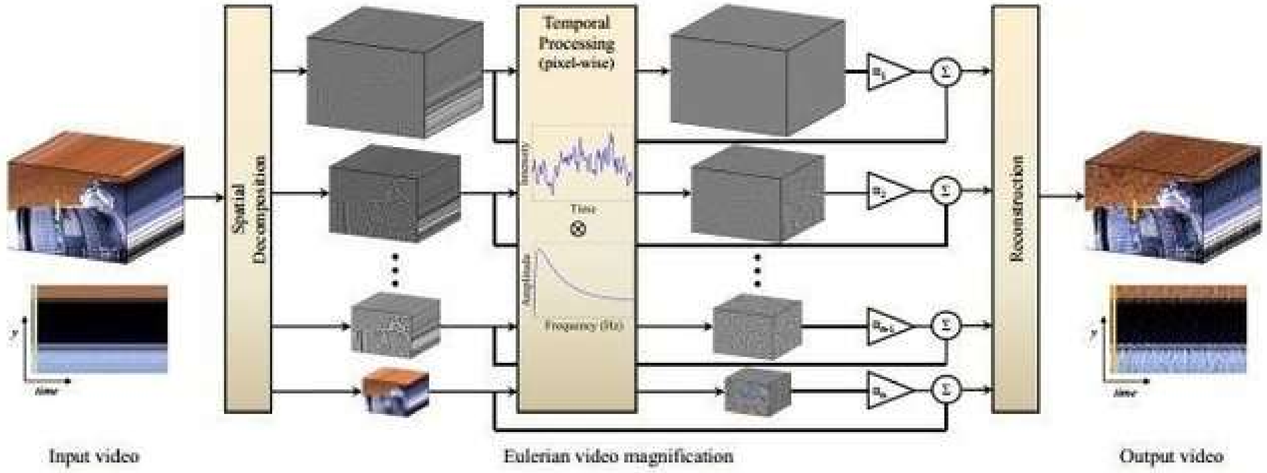


Figure 1.11: Eulerian Video Magnification Method https://bit.ly/32Nm5fH.

The other findings of EVM is the “Color magnification”. Performing temporal processing on each spatial band and considering the time series corresponding to the value of a pixel in a frequency band and applying a band-pass filter to extract the frequency bands of interest. For example, users might select frequencies within $0.4 - 4Hz$, corresponding to $24 - 240$ beats per minute, if magnifying a pulse is needed. When possible to extract the pulse rate, then using a narrow band around that value is recommended. The temporal processing is uniform for all spatial levels, and for all pixels within each level. In the next step, multiply the extracted band-passed signal by a magnification factor α. This factor can be specified by the user, and may be attenuated automatically. Possible temporal filters are discussed in Section 4 [14]. Next, add the magnified signal to the original and collapse the spatial pyramid to obtain the final output. Since natural videos are spatially and temporally smooth, and because filtering is performed uniformly over the pixels, the method implicitly maintains Spatio-temporal coherency of the results.

This research used the method’s blood circulation and action unit amplifier of face. The input of this method is the OpenFace output of the facial images. Output of this method is shown following figures.

Figure 1.12 shows, the first twenty frames of Aff-Wild training video 447 split ten by ten. The first ten show high blood pressure and rest of them

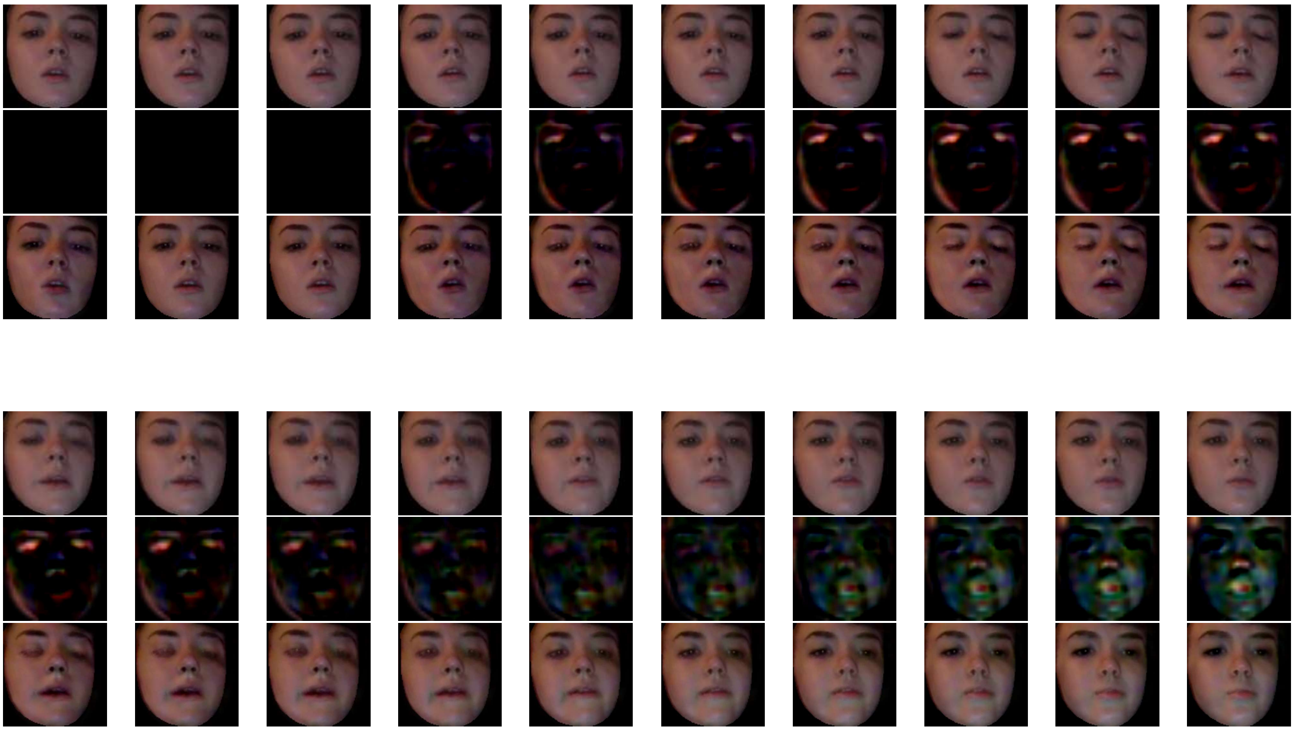


Figure 1.12: Eulerian video magnification color amplifier example on Aff-Wild training video 447.

show low blood pressure scenes. In the first ten, first row shows output of the OpenFace method, the second row shows the difference of the output of the Eulerian video magnification and the OpenFace output. The last row shows the output of the EVM color amplification method for low frequency set to $0.8Hz$, high frequency to $3Hz$, and the amplifier α as 5.

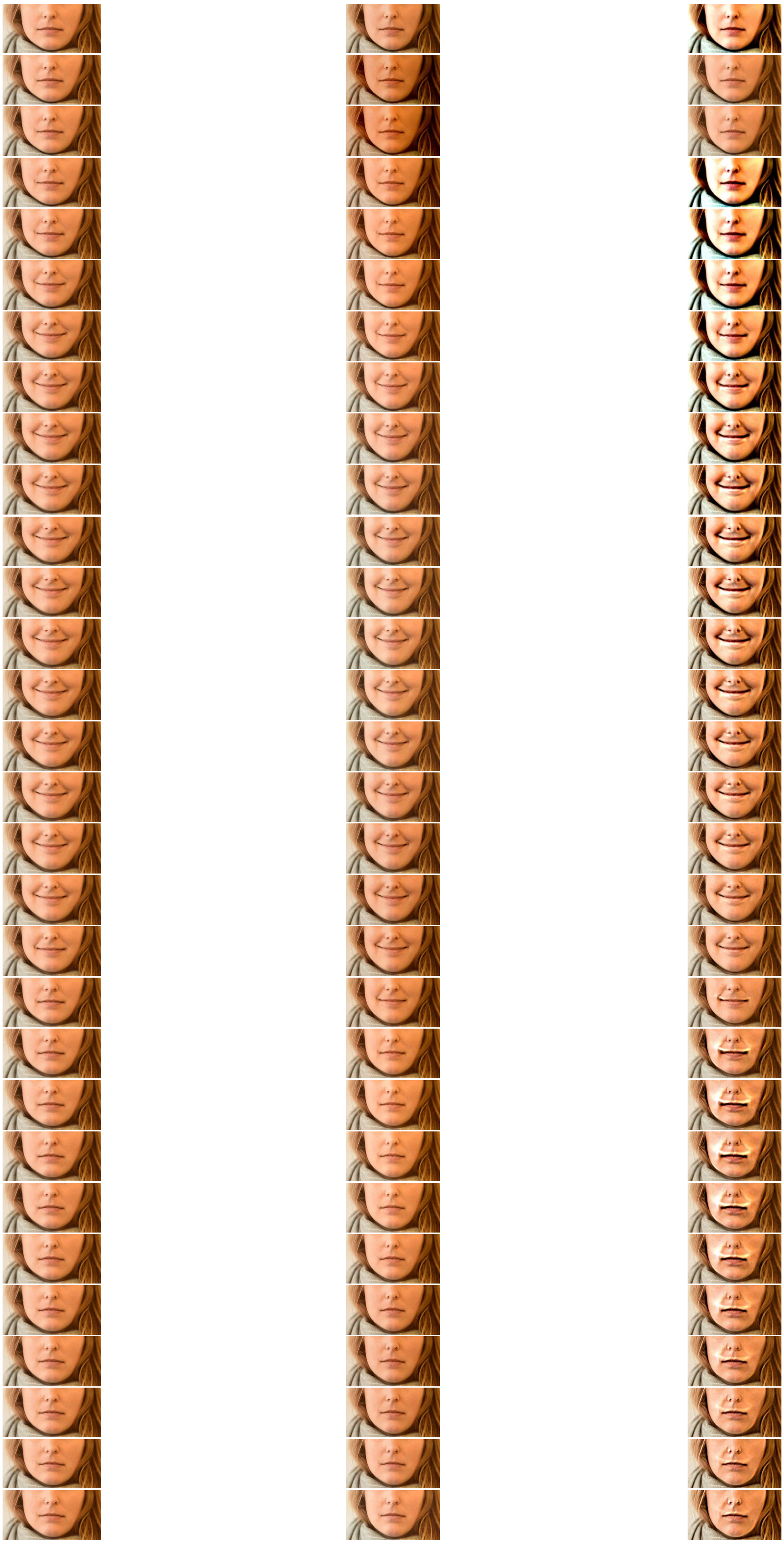


Figure 1.13: Eulerian video magnification motion amplifier on AU12 FACS.

In Figure 1.13, the EVM motion magnification of “Facial Action Coding system’s” action unit 12 is examined. The 1st column shows the original action unit 12 which represents “Lip corner puller”. The 2nd column is the color magnification for the same settings as for Figure 1.12. In the 3rd column motion, magnification settings are $\alpha = 5$, $f_{max} = 5Hz$, $f_{min} = 0.8Hz$ for the band-pass filter. As a result, the last ten frames are shown as AU12 amplified. If no motion or movement is observed in the video output (the first three frames of the last column). Therefore, in last column of Figure 1.13, the starting ten frames are not a related motion of AU12.

1.3 Related works

1.3.1 Emotion classification on image

Emotions often mediate and facilitate interactions between human beings. Thus, understanding emotion often brings context to seemingly bizarre and complicated social communication. Emotion can be recognized through a variety of means such as voice intonation, body language, and more complex methods such electroencephalography (EEG) [15]. For mankind, arousal and valences of emotions plays an important role in the human brain. Emotion Recognition is a branch of Activity Recognition. However, the easier, more practical method is to examine facial expressions. There are six basic types of human emotions shown to be universally recognizable across different cultures: anger, disgust, fear, happiness, sadness, surprise and sub-emotions, identified by Paul Ekman [16]. However, other researchers in emotion recognition also include contempt or neutral as an emotion. Moreover, even for complex expressions where a mixture of emotions could be used as descriptors, cross-cultural agreement is still observed. The task of emotion recognition is particularly difficult for three reasons:

1. A large database of training videos does not exist yet.

2. Classifying emotion can be difficult depending on whether the input video has static or transitive frame in the facial expression.

3. Training database has some noisy labeled items in the train and test sets.

The first issue is particularly difficult for real-time detection where facial expressions vary dynamically. Most algorithnms of emotion recognition examine static images of facial expressions [17], [18]. CNN models, for example, VGG, Resnet, Squeeze Extraction Resnet and Densenet are used. In addition, accounting for variations in lighting and subject position in a real-life environment is challenging. Over the last two decades, researchers have significantly advanced human facial emotion recognition with computer vision techniques, concentrating on improving accuracy rate in the wild environment, to read video stream or recorded video and classifying human emotions using effective deep neural network. Historically, there have been many approaches to this problem, including using pyramid histograms of gradients (PHOG) [19], AU aware facial features [20], boosted Local Binary Pattern (LBP) descriptors [21], and RNNs [22]. However, recent submissions [23] to the 2018 Emotions in the Wild (EmotiW 2018) contest for video frames used Recurrent Neural Networks (RNN), generating up to 60.64% test accuracy on the AFEW test set. Moreover, the winner of EmotiW 2018 [24] improved the baseline results by 28%. In their result on the EmotiW 2018 test set, they experimented over an ensemble of seven networks, which were different variants of OpenFace, OpenPose, split dataset, Local Binary Pattern (LBP) and Convolution 3D (C3D) with LSTMs. Overall the best network of the ensemble was a mixture of OpenFace, OpenPose and new split dataset. A recent development by G. Levi et al. [25] and Jianfei Yang et al. [24] shows significant improvement in facial emotion recognition using a Deep Neural Network (DNN). The authors addressed two specific problems:

1. A small amount of data available for training deep CNNs.
2. Appearance variation usually caused by variations in illumination.

They used LBP to transform the images to be illumination invariant. This data preprocessing was applied to various publicly available models such as VGG S [26]. The model was then re-trained on the large CASIA Web Face dataset [27] and transfer learned on the Static Facial Expressions in the Wild (SFEW) dataset, which is a smaller database of labeled facial emotions released for the EmotiW 2015 challenge [28]. Results showed a test accuracy up to 54.56%, an improvement of 15% over baseline scores.

Direct experimental method for recognizing facial emotions with CNN's, state of art networks known as Resnet [29], Seresnext [30] and Densenet [31] are also used. In the future, well trained network could be used in pretrained network of the Long-term Recurrent Convolutional Networks (LRCN) network. In [32], a survey of facial expression recognition most researcher used, more successful results are obtained in facial emotion recognition systems, using MMI and FER2013 datasets. The Facial Emotion Recognition dataset (FER2013) is used to avoid overfiting. In the FER2013 dataset, all images have been registered and resized to 48*48 pixels after rejecting wrongfully labeled frames and adjusting the cropped region. FER2013 contains 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels (anger, disgust, fear, happiness, sadness, surprise and neutral), which are validated by 10 persons' average. Moreover, the authors addressed the FER2013 dataset trained models to be well suited by the pretrained model.

1.3.2 Continual learning on emotion classification

Systems that can learn new information without catastrophic forgetting prior knowledge are vital. For this reason,this work was also extended to the field of continual learning. Moreover, due to low class availability of FERplus and FER2013 datasets, the broadly used CIFAR100 dataset is used. Living beings have the ability to continually investigate, adapt, and to transfer knowledge and skills throughout their lifetime. Therefore, continual lifelong learning [33] has achieved much interest in recent deep learning studies.

In this section, training of a model competent of handling unknown sequential tasks while keeping prior knowledge without much loss on once trained tasks is detailed. In continual lifelong learning, the training data of previous tasks are assumed non-available for the newly available tasks. However, the trained model can be used as a pre-trained model and fine-tuning it for the new task will force the model parameters to train on the new data, which causes catastrophic forgetting [34] on once trained tasks.

The state of the art study CPG (Compacting, Picking and Growing for Unforgetting Continual Learning) [35] has addressed catastrophic forgetting problem and solution based on the progressive learning method [36]. Moreover, the result achieved their goal. On the other hand, the cost of achievement was expensive, because of picking well trained parameter and growing. Hung, Ching-Yi et al. proposed model expansion at 1.5 on each task in scratch learning. The model size on the drive was expanded from 435Mb to, after 20 task of CIFAR100 dataset, 5.3Gb. However, if expansion parameter reduced to 1.2, the method didn't reduce catastrophic forgetting. Other models include, CondConvContinual [37], PEANet [38] and PackNet [39]. The results of the related works are shown in Table 1.3.

Table 1.3: Related research result comparison.

Models	Network growth @times	Network size @mega	Accuracy @percent
CPG	1.5	278	80.9
CondConvContinual	1.6	98	77.4
PEANet	1.2	98	77.1
PackNet	2	112	67.5

In this research some types image processing transformations on classifier weights were investigated. Firstly, implementing continual learning with the image processing alpha blending method was performed. After that, another method of pruning weights using histogram near zero values set to zero was applied. The last method researched was histogram equalization. All of our research work was conducted on a dataset based on CIFAR-100 [40] dataset.

1.3.3 Emotion recognition from video

To understand each other, emotions have played an important predictive role. Emotions are extremely complex brain function that reacts to previous actions based on the human brain functional system, called the "Limbic System", section 1.1.1. Recently, there has been an enormous interest in this emotion recognition field.

Recognizing emotion from image and video in wild setting is one of the challenging research field in machine learning. In the wild setting, Aff-Wild, AFEW-VA and Aff-Wild2 datasets exists, collected using pure emotion without acting or wearing a neutral emotion mask in an every day life situation.

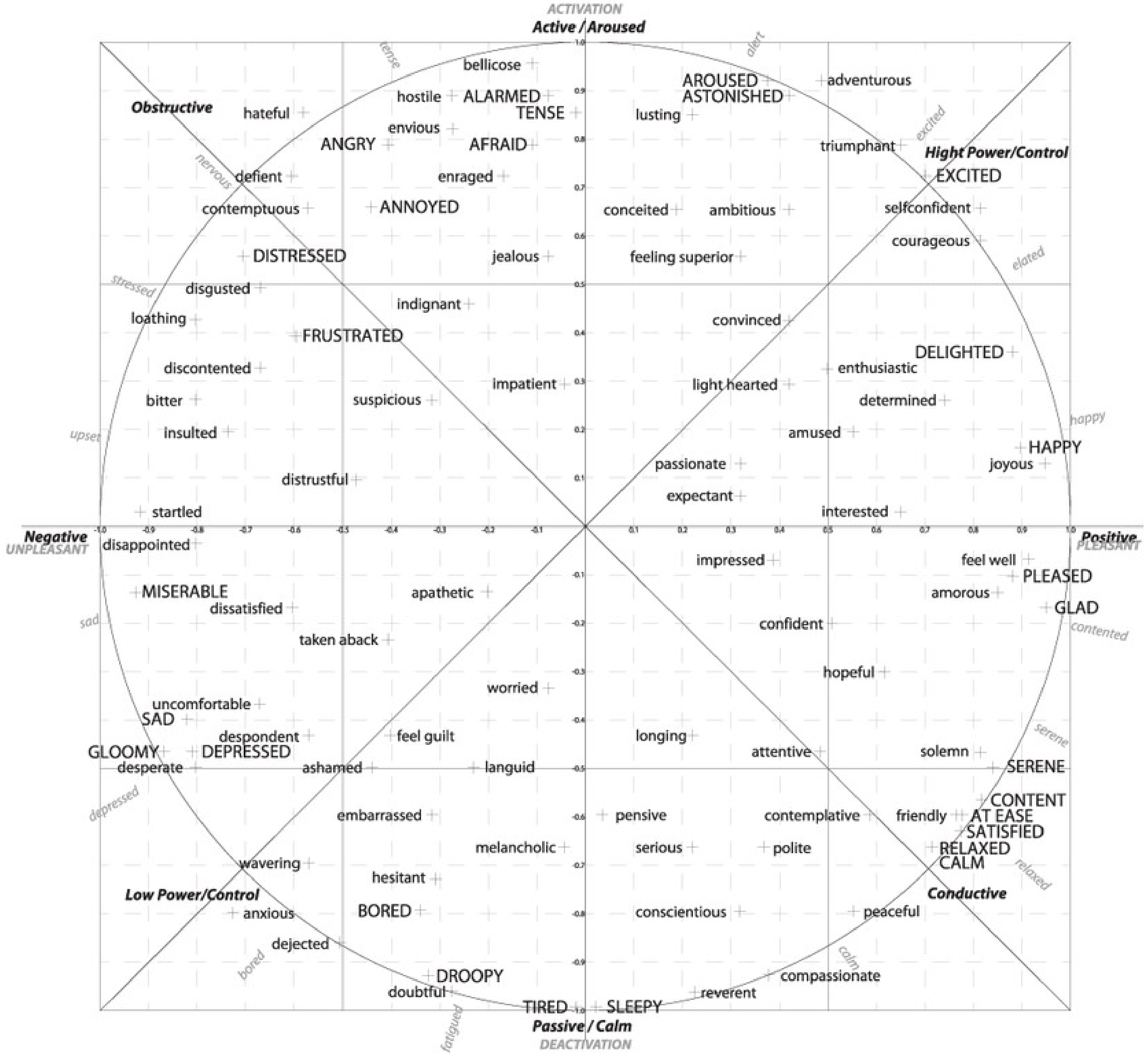


Figure 1.14: Illustration of relationship between emotion using Valence and Arousal. (Fig.1 [41]).

Facial emotion can be visualized by the expansion and contraction of the muscles located around the mouth, nose, and eyes [42]. Moreover, Action Units

(AUs) [43], [44], [20] were proposed to model facial behavior, and the combination of AUs could also be utilized for facial expression recognition. Most studies are based on the seven basic categories [45], [18], [21], [46], [28], [47] with some researchers using triplet expression recognition [48].

Consequently, facial expression recognition has attracted renewed attention owing to recent advanced network architectures. In facial expression recognition, real-time automated analysis of facial expression in video plays an essential role in implementing human-computer interaction interfaces.

Figure 1.14 shows the 2-D Emotion Wheel [41], [49], illustrating valence ranging from extremely positive to negative and arousal ranging from extremely active to passive. Based on the data from the activity a person is engaged in, discriminating information on the valence of a person's emotions may be known. Emotion sensing parameters can be distinguished as to whether they provide information on qualification of the valence aspect or the arousal aspect. There are many related works in this field depending on the dataset used and proposed methods [50], [32]. There are also previous studies on emotion recognition in videos [51], [52]. Valence and Arousal (V-A) are not separated values; binding these two parameters describe an emotion.

For most emotion recognition methods so far, the Concordance Correlation Coefficient (CCC) loss and the Mean Squared Error (MSE) loss are widely used during training. In FATAUVA-Net, Chang et al. [52], has provided the best-confirmed results using mean CCC and mean MSE for valence and arousal. The authors concentrated on the connection between V-A estimation and Action Units (AUs) such as the face and its parts. Moreover, their research environment was based on a wild setting. Yang et al. [24] concentrated on feature extraction. A network was assembled that extracted features with a Recurrent unit and was trained on MSE loss. Vielzeuf et al. [23] trained audiovisual ensemble network on emotion video classification. MIMAMO Net [51] trained a spatial, temporal network with CCC loss.

Two criteria were measured for evaluating the performance of the networks; The Valence and Arousal fractional range is between $[-1, 1]$. The main problem is that there is no loss function that can quickly train a given network on less data. Moreover, Valence and Arousal are not separated values, and it is usually considered important to train them together in a coordinate system points.

Chapter 2

Objectives and Datasets

Currently, there are no large enough emotion dataset to completely and accurately recognize emotions. The existing ones in the audio recognition field, include, LibriSpeech [53], and VoxCeleb [54] datasets, with over thousand hours of audio speech available. However, in emotion recognition, the biggest dataset is Aff-Wild2 with 2.8 million frames, and around 26 hours of video. This is not enough to recognize emotions to the same standards as the Audio recognition systems. Moreover, collecting an emotion dataset in the wild setting is complicated, and after that, a long time is required to annotate it.

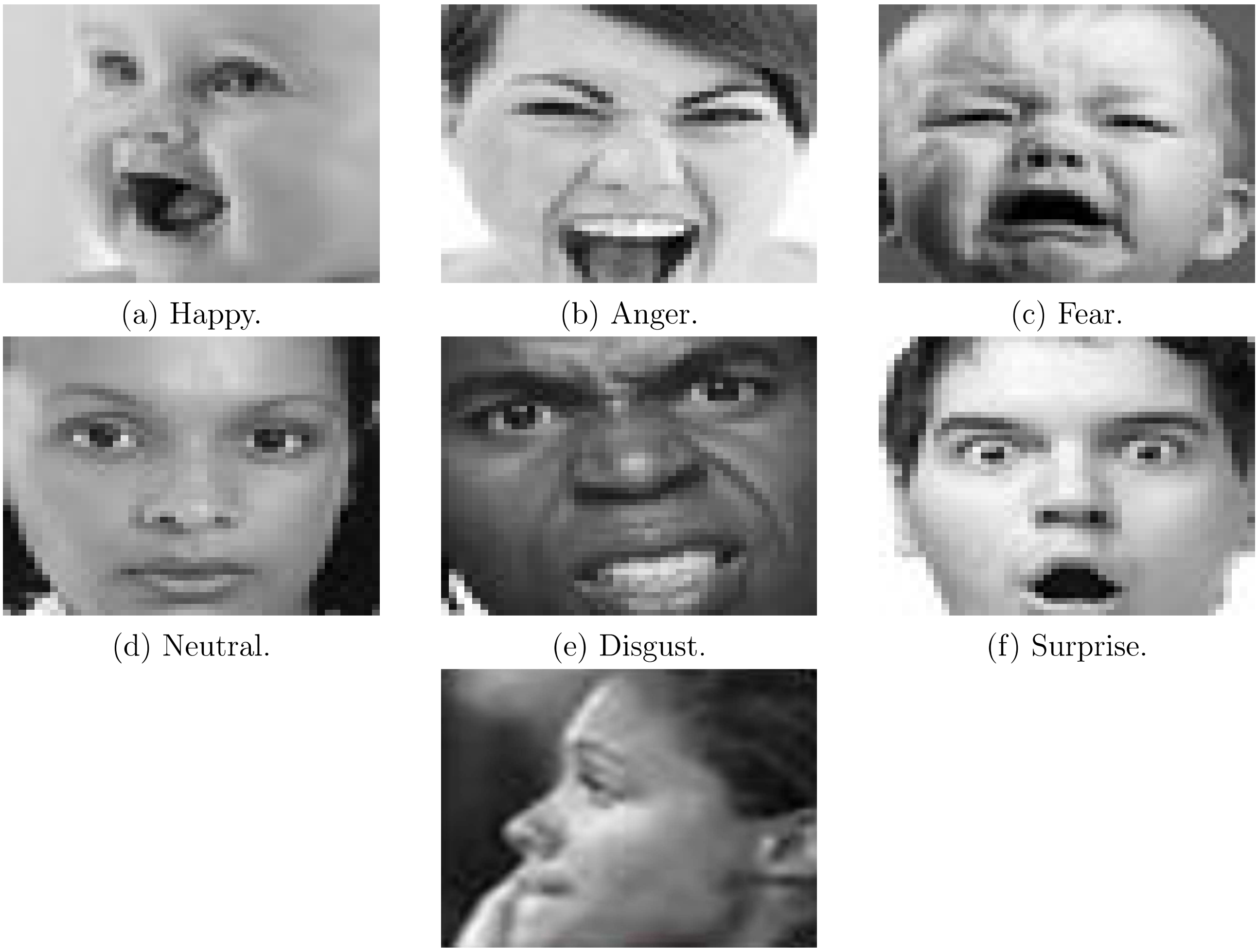
Another way to collect a dataset is using Wearable Technology. However, it has not yet been implemented for collecting data in brain signal reading field in a wild setting. On the other hand, instead of collecting a large dataset, another way to recognize emotion from video is reinforcement learning or continual learning using the limited data available. Continual learning field is investigated in section 3.3.

In the following sections, we will introduce and discuss the dataset used in this work.

2.1 FERPlus

The Kaggle Facial Expression Recognition Challenge dataset FER2013 consists of 48x48 pixel grayscale images of faces. After researchers' inspection, the relabeled FER2013 dataset was renamed FERPlus. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The aim is to categorize each face based on the emotion shown in the facial expression to one of basic seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 facial expression images. The public test set used for accuracy calculation consists of 3,589 examples.

This dataset was prepared by Pierre-Luc Carrier and Aaron Courville, as



(a) Happy.

(b) Anger.

(c) Fear.

(d) Neutral.

(e) Disgust.

(f) Surprise.

(g) Sad.

Figure 2.1: FER2013 sample images.

part of an ongoing emotion prediction on image research. Example images from the dataset are shown in Fig. 2.1. Researchers mentioned that the FERPlus dataset is unbalanced, Fig. 2.2. Due to this unbalance, targets balanced sampler was used.

2.1.1 Imbalanced data sampler

In several machine learning applications, across dataset some categories of information could also be coaching quite different categories. Take identification of rare diseases for instance, in the square measure there are most likely a lot of traditional samples than malady ones. In these cases, they have to be compelled to certify that the trained model isn't biased towards the category that has a lot of knowledge. As an associate in a nursing example, think about a dataset with five malady pictures and twenty traditional pictures. If the model predicts all pictures to be traditional, its accuracy is 82%, and F1-score of such a model is 88%. Therefore, the model has high tendency to be biased toward the 'normal' category.

To solve this drawback, a wide adopted technique is termed resampling. It consists of removing samples from the bulk category (under-sampling) and / or adding a lot of examples from the minority category (over-sampling). Despite the advantage of leveling categories, these techniques have their weaknesses. The best implementation of over-sampling is to duplicate random records from the minority category, which may cause overfitting. In under-sampling, the best technique involves removing random records from the bulk category, which may cause loss of data. The FER2013 target Imbalanced sampler is shown in Fig.2.3.

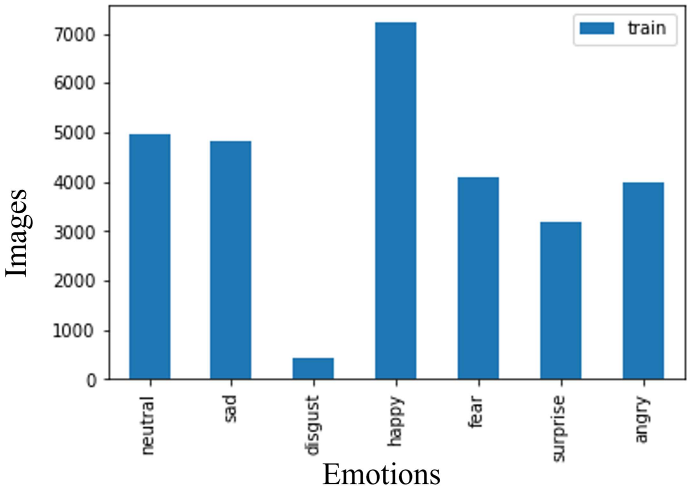


Figure 2.2: FER2013 target training set distribution.

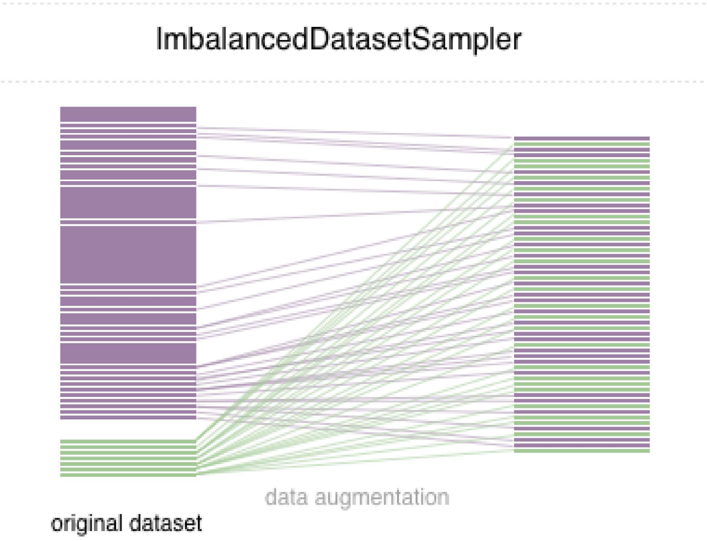


Figure 2.3: FER2013 target Imbalanced sampler.

2.2 AFEW-VA

The Acted Facial Expressions in the Wild (AFEW-VA) [55] dataset is a collection of highly accurate per-frame annotations levels of valence and arousal, along with per-frame annotations of 68 facial landmarks for 600 challenging video clips. AFEW contains video clips collected from different movies with spontaneous expressions, various head poses, occlusions and illuminations. Moreover, temporal and multimodal database provides vastly different environmental conditions in both audio and video. Samples are labeled with seven discrete expressions: anger, disgust, fear, happiness, sadness, surprise, neutral and continuous Valence, Arousal space. The annotation of expressions have been continuously updated, and reality TV show data have been continuously added. Sample frames from the database are shown in Figure 2.4. Figure 2.5 shows an annotation example.

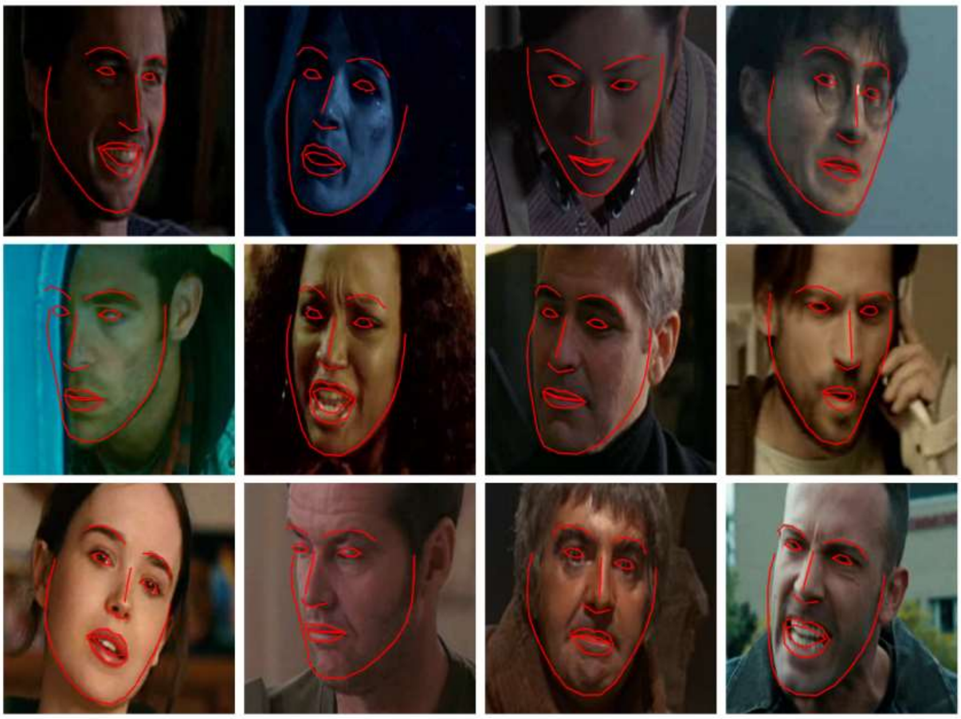


Figure 2.4: Sample frames of facelandmark.

Since the goal is recognizing true feeling, no model was trained on AFEW-VA dataset.

2.3 Aff-Wild2

Affective computing has been largely limited in terms of available data resources. The need to collect and annotate diverse in-the-wild dataset has become apparent with the rise of deep learning models, as the default approach to address any computer vision task. Some in-the-wild databases have been recently proposed. However: i) their size is small, ii) they are not audiovisual, iii) only a small part is manually annotated, iv) they contain a small number of subjects, or v) they are not annotated for all main behavior tasks (valence-arousal

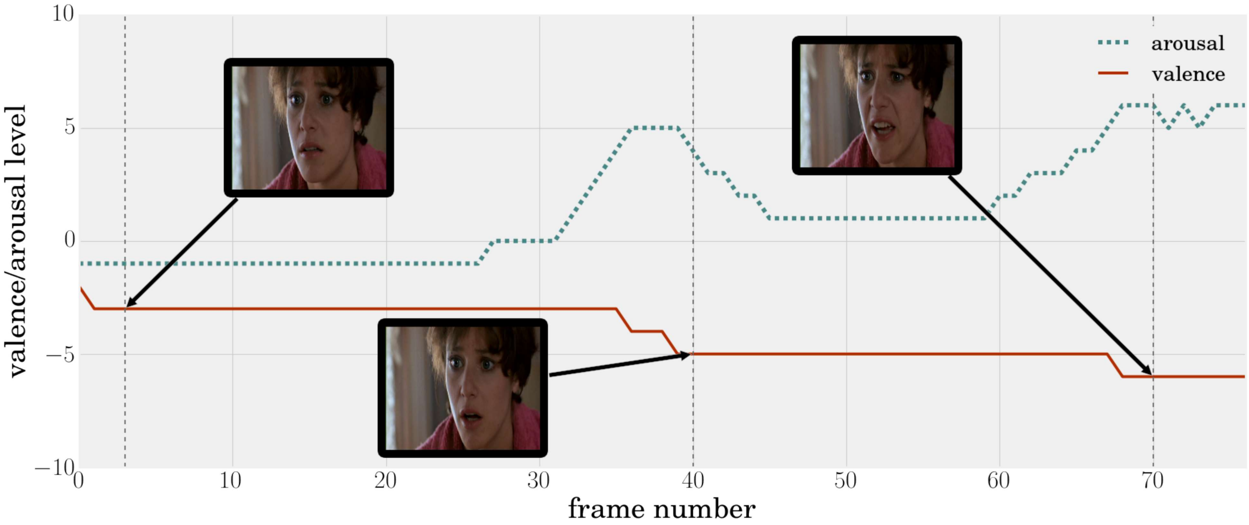


Figure 2.5: AFEW-VA dataset, Video demo annotation.

estimation, action unit detection and basic expression classification). To address these issues, the largest available in-the-wild database (Aff-Wild) to study continuous emotions such as valence and arousal was substantially extended. Furthermore, parts of the database with basic expressions and action units are annotated. This database was termed Aff-Wild2. In total, Aff-Wild2 contains 558 videos with around 2.8 million frames. To the best of our knowledge, Aff-Wild2 is the first large scale in-the-wild database containing annotations for all 3 main behavior tasks. It is also the first audiovisual database with annotations for AUs. All AU annotated databases do not contain audio, but only images or videos [56].

The new videos are a fresh dataset of 260 films totaling 1,413,000 frames in 13 hours and 5 minutes duration. 11 of the 260 films had two subjects, all of which were annotated. The new dataset includes 258 participants, 149 of whom are male and 109 of whom are female. Figure 2.6 shows example images from the Aff-Wild2 dataset.

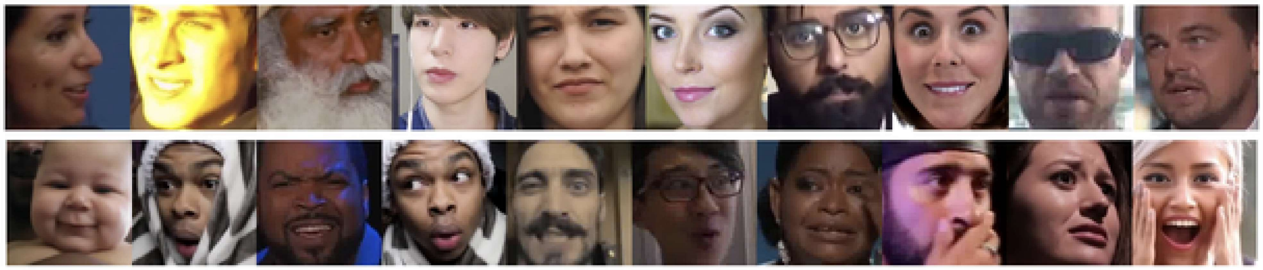


Figure 2.6: Example images from the Aff-Wild2 dataset.

Chapter 3

Proposed methods

In general, recognition models are trained using loss functions. The most common ones are the Mean Squared Error (MSE) and the Concordance Correlation Coefficient (CCC). The two loss functions produce acceptable results but with training and accuracy issues. This section will discuss our proposed new loss function and show its superiority to the existing losses using the proposed network models evaluated in static and sequential images.

Additionally, this study explored the continual learning field, which in the future will enable life-long learning of unobserved emotions, as shown in Figure 1.14, without losing previously gained knowledge (catastrophic forgetting).

3.1 Loss functions

Video emotion recognition system broadly use loss functions. This section will discuss the two existing loss functions and introduce the proposed new loss function. Before discussing the proposed non-inertial loss function, for comparison purposes, the two other largely used loss functions are briefly introduced below.

3.1.1 Mean Squared Error (MSE)

The first most comparative metric criterion is the Mean Squared Error (MSE), defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{3.1}$$

where x_i are predictions values, y_i are annotations (valence-arousal), and N is the total number of samples. The MSE calculates an approximate indication of how the training model is performing. A small value of MSE is desirable.

3.1.2 Concordance Correlation Coefficient (CCC)

The second one is the Concordance Correlation Coefficient (CCC) [57]. It is widely used to measure the performance of dimensional emotion recognition methods, e.g., in the Aff-Wild challenge [58]. CCC calculates the similarity between two time-series (e.g., all video annotations and predictions) by scaling their correlation coefficients with their mean square difference. The predictions that are well correlated with the annotations but shifted in value are penalized in proportion to the deviation. CCC values are in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes discordance. The higher the value of the CCC, the better the fit between annotations and predictions. The mean value of CCC for valence and arousal estimation is used as the main evaluation criterion.

CCC is defined as follows:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\hat{x} - \hat{y})^2} = \frac{2s_x s_y \rho_{xy}}{s_x^2 + s_y^2 + (\hat{x} - \hat{y})^2} \tag{3.2}$$

where ρ_{xy} is the Pearson Correlation Coefficient (PCC), s_x and s_y are the variances of all valence or arousal video annotations and predicted values, respectively and s_{xy} is the corresponding co-variance value, $\hat{x}$ and $\hat{y}$ are mean values of predictions and annotations.

3.1.3 Proposed Non-Inertial loss

Two criteria are measured for evaluating the performance of networks. The Valence and Arousal's fractional range is between $[-1, 1]$. The problem is that there is no loss function to quickly train the networks on less data. Moreover, Valence and Arousal are not separated values, and it is usually considered important to train them together in a coordinate system of points. This work uses the unit circle map of V-A. This new definition is named the "Non-Inertial loss".

The new loss function for the V-A is defined by the following equations.

$$\Delta l_c = \sqrt{x_c^2 - y_c^2} \tag{3.3}$$

$$\begin{aligned} \Delta l_{x,y} = (&\sqrt{(x_c(v) - x_p(v))^2 + (x_c(a) - x_p(a))^2} \\ &- \sqrt{(y_c(v) - y_p(v))^2 + (y_c(a) - y_p(a))^2})/t \end{aligned} \tag{3.4}$$

where v and a are the valence and arousal values respectively.

$$\begin{aligned} \Delta\alpha &= tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} \\ &= abs(a_x * v_y - a_y * v_x) \end{aligned} \tag{3.5}$$

$$loss_t = mean(\Delta l_c + \Delta l_{x,y} + \Delta\alpha) \tag{3.6}$$

Equation (3.3) is referred to as norm 2, Euclidean distance or Root Mean Squared Error (RMSE) of the annotations and predictions, which calculates how far away in V-A unit map the prediction is. A small value for the distances is desired. Equation (3.4), where t is time, $x_c(v)$ is the current prediction valence, and $x_p(a)$ is the previous prediction arousal, calculates the difference of prediction and annotation V-A velocities. To avoid the network parameters exploding, $t = 1$ is assumed. Velocity will be enhanced in the recurrent section of the training model, which in network is the LSTM layer.

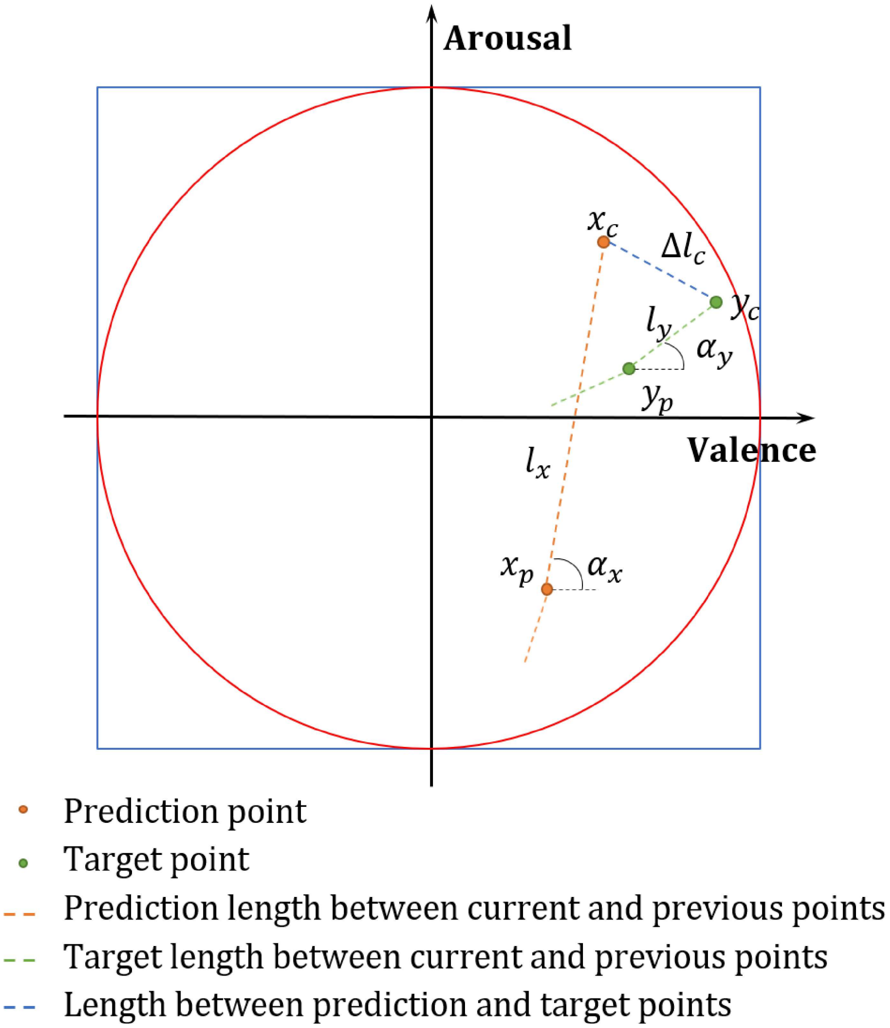

Arousal
x_c
Δl_c
y_c
l_y
α_y
y_p
Valence
l_x
x_p
α_x
● Prediction point
● Target point
-- Prediction length between current and previous points
-- Target length between current and previous points
-- Length between prediction and target points

Figure 3.1: Non-inertial loss function illustration.

Equation (3.5) is the direction. It calculates the differences of angles. The difference between annotation and prediction angles are expected to become zero; a_x is the projection of prediction arousal and v_y, is the projection of annotation valence. Differences of angles are converted to avoid subtraction because of zero denominators and a value more than 1 is overlooked in the arcsin function. The proof of the Equations (3.5) is in Appendix A.

The non-inertial loss function works with the form $[batch, sequence, 2]$, which is the annotation size. If the annotation size is more than 2, then an acceleration parameter can be added and the circle of angle difference can be extended from the first column to the end.

Fig.3.1 shows an illustration of an example loss function annotation and prediction sequence in arousal and valence map.

3.2 Proposed models for loss functions evaluation recognition

3.2.1 Emotion recognition from image models

Emotion recognition in images is still a challenging task in Computer Vision. In this work, we performed experiments to find the most effective static emotion recognition based on several models on the FERPlus dataset. However, the FERPlus dataset contains small images with low resolution. Therefore, we decided to increase the image size because of the difficulty of comparing the models. The new image shape to feed the model is finally 224×224.

The main model selected is a Residual neural network (ResNet) [29], Fig. 3.2. It is an Artificial Neural Network (ANN) that is based on a constructed famed from pyramidic cells within the brain cortex. Residual neural networks attempt this by utilizing skip connections, or shortcuts to leap over some layers using self multiplication. Sometimes it can be the add function. Typical ResNet models are square measure enforced with one to triple layer skips that contain non-linearities (ReLU) and batch standardisation in between. An extra weight matrix is also accustomed to learn the skip weights; these models square measure is referred to as HighwayNets. Models with many parallel skips square measure are called DenseNets [31]. In parallel to shortcuts added squeeze and extraction layers in the multiple way are called SeResNexts [30]. The following figure shows simplified ResNet version of emotion classification on FERPlus dataset.

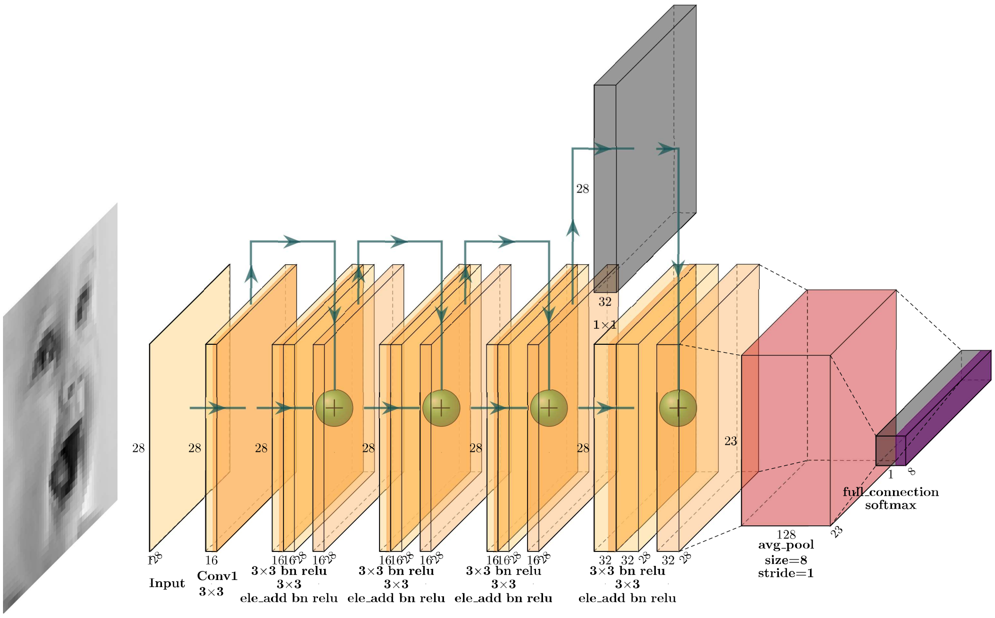


Figure 3.2: ResNet model simplified.

3.2.2 Video emotion recognition model

The proposed method on video emotion recognition overall model (Full Net) is illustrated in Fig.3.3. It consists of three stages.

1. The first stage inputs the OpenPose features,
2. The second stage consists of the Linear layers, and
3. The last stage is a Recurrent stage LSTM, which is widely used and already proven effective in Action Recognition [51].

The first stage extracts a feature representation of a snippet, which consists of a facial RGB image, face landmarks, body pose, and Mel spectrogram features. The RGB image is fed into the ResNet50 network. The sequence of extracted features of OpenPose outputs was fed directly into the second stage linear layers. The output of the linear layers feeds the last stage of the LSTM network and classifier. For low-cost training, features were extracted using a pre-trained ResNet50 in our previous study [59].

In this work, only OpenPose and pre-trained ResNet50 are not trainable. All other parameters are trained. Moreover, to test the proposed loss functions in different network parameter sizes, the Full network is split into 3 additional models;

1. “Without Mel Network”,
2. “Audio Network”, and

3. “Pose Network”.

For interrupted data network, “Without Mel Network” and “Pose Network”, are used because of input data interruption, that is, when there is no detected pose estimation and face features from OpenPose.

For “Full Network” and “Audio Network”, representation of continuous networks is assumed.

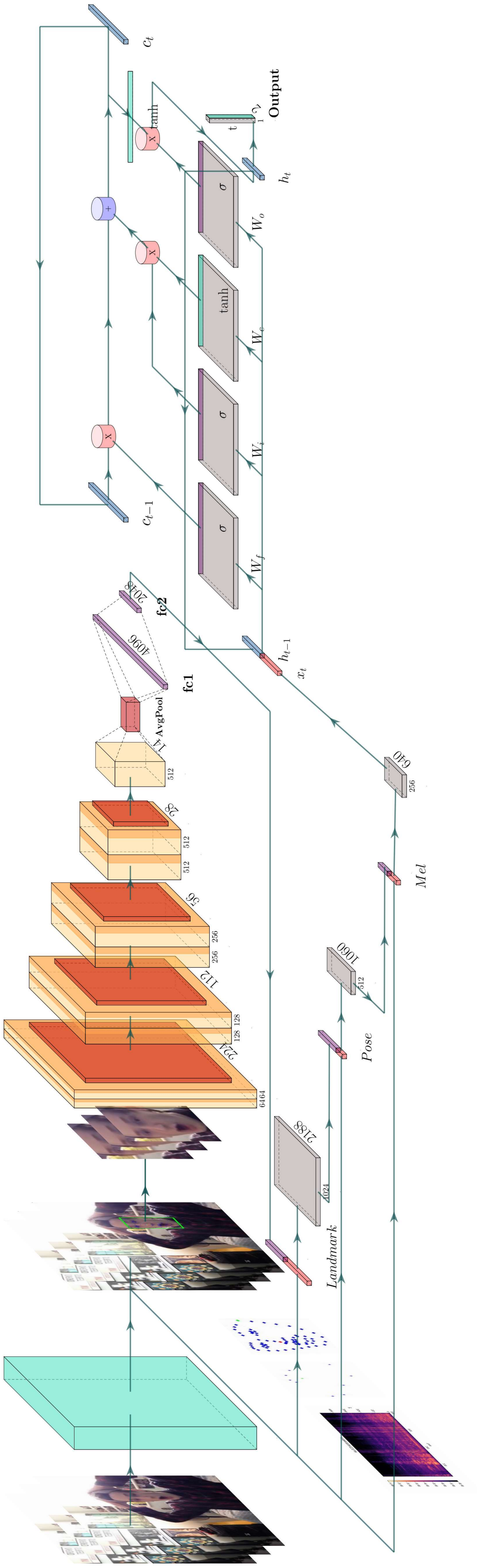


Figure 3.3: RNN network architecture (right side view).

The four networks used for evaluation of the proposed Non-inertial loss function are implemented as follows:

1. Full Network: The network architecture is shown in Fig.3.3. The flow is as follows. The Input(Batch(B), Sequence(S), Height(H), Width(W), Channel(C)) feed the OpenPose to get pose estimation and face features. Face feature coordination is used to calculate and pre-process the face with a bounding box and then feed into ResNet50 network. The previous layer output is concatenated with the face landmarks $(B \times S, 2048 + 140)$ and fed into the linear layer named landmark. The landmark layer directly feeds into the Pose layer $(B \times S, 1024 + 36)$. After the Pose layer, Mel layer continuous input is $(B \times S, 512 + 128)$. Mel layer output is $(B \times S, 256)$ which is directed as input into the LSTM layer $(B, S, 256)$. The LSTM layer requires additional Sequence dimension, the dimension of Batch, Sequence, and output of previous layers. At the end of the classifier layer $(B, S, 2)$ the final outputs are calculated.

2. Without Mel Network: This network does not include continuous input Mel spectrogram and Mel layer. Other parameters are the same as the previous Full Network.

3. Audio Network: $Input(B \times S, 128)$ continuous audio data is fed into Mel layer and output is $(B \times S, 256)$. Before the LSTM layer, a new dimension of sequence $(B, S, 256)$ is added. The end Classifier linear layer $(B, S, 2)$ classifies arousal and valence.

4. Pose Network: As non-continuous data, Pose data $Input(B \times S, 36)$ is fed into the Pose layer. In addition, the dimension of sequence $(B, S, 256)$ is added and fed into LSTM $(B, S, 256)$ layer. This is the same as the previous networks. The classifier layer $(B, S, 2)$ also classifies arousal and valence.

All the network configurations used batch normalization 1D and dropout 0.5 after every fully connected layer.

3.2.3 Transformer model

The transformer model has proved to be more effective in sequential data such as in the Natural language processing field. Moreover, some researchers also experimented in the image recognition field [60]. Authors split images to $(N \times Channel \times 16 \times 16)$, tokens and feed to Transformer model. Instead of using their method, this work use features that are as small as possible and feed it to model. Since, our goal is using sequence images of Valence and Arousal, it is not possible to use the classification method in their method. If we split our facial image,the attention method on the decoder side would be meaningless. Moreover, for the Visual transformer model instead of using multi-head attention decoder they used multi-layer perceptron (MLP).

Previously, we implemented the RNN architecture and split the full model to four parts based on the data flow as shown in Fig 3.3. In this section, instead of using LSTM in the RNN part replaced it with the Transformer. Moreover, we replaced OpenPose with Google MediaPipe (Section 1.2.3) face-mesh detection model, which is fast and light-weight for detecting face and its mesh. After detecting the face-mesh,we applied the Eulerian Video magnification method (Section 1.2.5) and fed color magnified video into the ResNet50 model. The transformer network has an advantage in the multi-head attention part, which can use image data as its input. On the other hand, image size will explode the model size if higher resolution image are used. Due to this issue, we should extract features using previously experimented ResNet50 model as feature extractor without using the Classifier layer. Extracted features output shape should be ($Sequence \times Batch \times Channel \times Feature$) (note Transformer model requires Sequence first). Using this model this study can train the end to end model and also split the model into Mesh-Transformer, EVM-Transformer, Feature-Transformer and Multi-Transformer.

The implemented Transformer model shown in Fig 3.4.

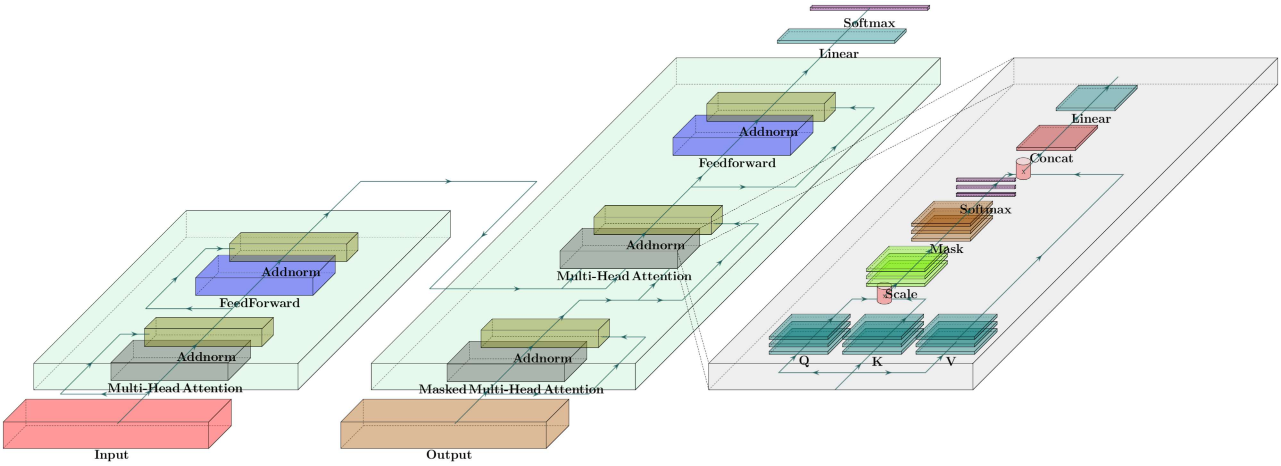


Figure 3.4: Transformer network architecture.

In the transformer model we used the standard encoder and decoder parts, but our input feature and output of the model shape is different. Input ($Sequence \times Batch \times Channel \times Feature$) and Output ($Sequence \times Batch \times 2$). Therefore, we used additional Linear layer to decode and adapt between the input and output shapes. Therefore, for the Mesh-Transformer, and the Feature-Transformer models, instead of using tiny images, we used normal transformer training settings.

3.3 Continual learning proposed model

In the future, unobserved or splitting emotions according to Figure 1.14 could happen. Therefore, continual learning experiments were conducted. Multiple classifier mixing methods for continual learning including concatenation of weights are proposed. Furthermore, experiments on how to train an universal model and add new tasks continuously were performed. The universal model, ResNet18 [29], is pre-trained on ImageNet dataset while freezing the convolution layers. Moreover, wefine tuned all classes of the CIFAR-100 dataset and network classifier weight as shown in Figure 3.6. After training of the universal network, a new task is added into the linear layer. An additional unknown class on each task training is added and the alpha blending method used to merge the trained linear layers. The proposed method illustration is shown in Figure 3.5.

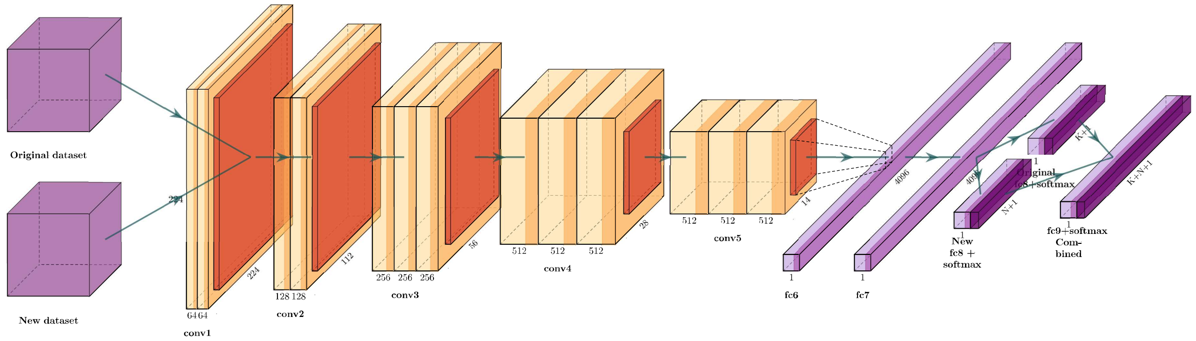


Figure 3.5: Proposed base model architecture.

In Figure 3.5, the original dataset representation is any kind of dataset on which classses are to be extended. Continual dataset is represented with any kind of dataset set needed to be added to the original dataset. In the example figure, from $conv1$ to $fc7$, the selected model is any kind of model that should continuously train. $Fc8$ can be of two types. One is original model trained classifier layer and the other one is new fully connected layer. After the mixture of original and new fc layer, $fc9$ will become the new original fully connected layer.

Figure 3.6 shows the trained classifier weights of ResNet18 on CIFAR-100 dataset, which is used in the trained model as the universal model.

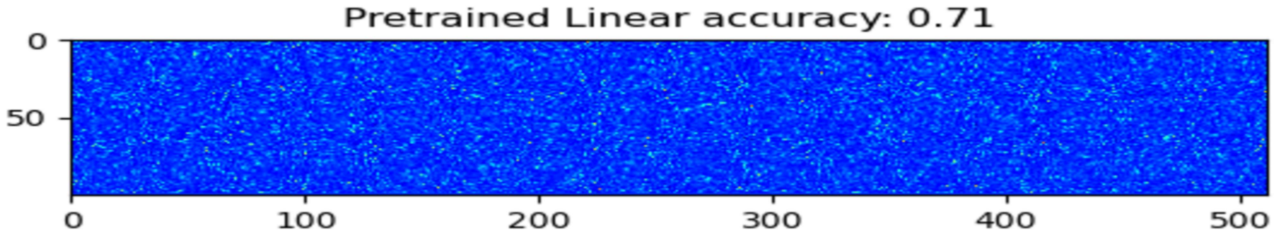


Figure 3.6: ResNet18 trained model classifier weight on 100 classes with 71% accuracy rate.

3.3.1 Continual learning proposed method: Alpha blending on linear layer weights

Fully connected layers are represented mathematically using the following equation.

$$y_i^n = \sigma[W_{i,j}^n y_j^{n-1} + B_i^n] \tag{3.7}$$

where, i and j are linear weight dimensions.

The equation represents the output of the n^{th} linear layer. σ is the representation of the activation function. W^n is the weight of n^{th} linear layer and y^{n-1} is the previous layer output, B^n is bias of the n^{th} linear layer, Equation (3.7). One class is added in both the original $fc8$ and the new $fc8$ layer. This unknown knowledge of the classes, in other words, a class of the other objects in each dataset, represents those not included in the original or new dataset. Thus,

Equation (3.7) to follows:

$$y_{K+1}^n = \sigma[W_{K+1,j}^n y_j^{n-1} + B_{K+1}^n] \tag{3.8}$$

$$y_{N+1}^n = \sigma[W_{N+1,j}^n y_j^{n-1} + B_{N+1}^n] \tag{3.9}$$

where, N and K are trained classifier class sizes.

After mixing procedure of new and original fc8 layer, Equation (3.8) and Equation (3.9) change as follows:

$$y_{N+1+K+1}^n = \sigma[W_{N+1+K+1,j}^n y_j^{n-1} + B_{N+1+K+1}^n] \tag{3.10}$$

$$y_{N+K+1}^n = \sigma[W_{N+K+\alpha+(1-\alpha),j}^n y_j^{n-1} + B_{N+K+\alpha+(1-\alpha)}^n] \tag{3.11}$$

In Equation (3.11) size of weight parameter is the concatenated new and original layers. Both will use the alpha blending method. This procedure is the same as in bias parameter of fc9 layer. In addition, alpha parameter will be very large due to mixing fc layer's weight size. In other words, $\alpha = N/(N+K)$.

3.3.2 Continual learning proposed method: Histogram equalization transform on linear layer weights

Initially, the classifier weights are pruned after a training task. After that, the weights are concentrated and then the test model is tested on the dataset. We set threshold range from weight's as $[min(W_{task}) + 0.3, max(W_{task}) - 0.3]$. Another histogram equalization method that is broadly used is Contrast Limited Adaptive Histogram Equalization (CLAHE), which avoids amplifying noise. The CLAHE equalizer parameter clip limit and tile grid size was set to 100 and $[2, 2]$ respectively. The histogram equalization is shown in the equation 3.12.

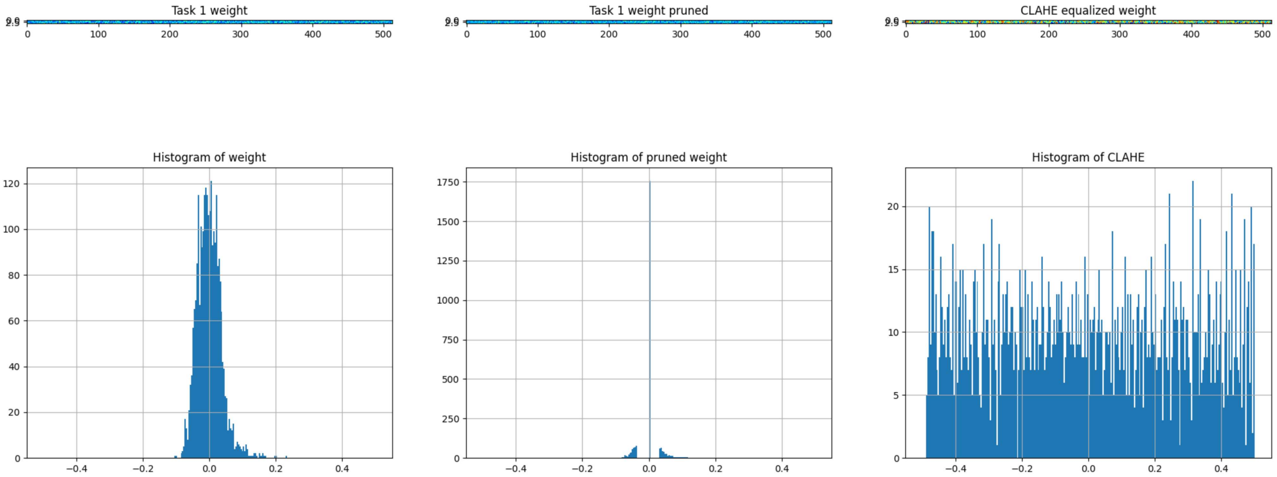


Figure 3.7: Task 1 trained weight, pruned and CLAHE applied histograms.

$$h(v) = round\left(\frac{cdf(v) - cdf_{min}}{(M \times N) - cdf_{min}} \times (L - 1)\right) \tag{3.12}$$

where *cdf* Cumulative distribution function, $M \times N$ gives the image's number of pixels and L is the number of levels. Figure 3.7 shows the CLAHE process.

Chapter 4

Experiments and Results

4.1 Facial image emotion classification models results

Before training, the preprocessing of images from 48*48 to 256*256 pixels was performed and 2 channels added using OpenCV. Training images recommended minimum size for the Densenet and Seresnext models is 224*224 pixels, random cropping, normalization method to avoid overfitting and imbalanced datasampler previously mentioned in section 2.1.1. In training of some models, faced with out of memory errors, the training batch size was reduced to 8 images. Training loss and Validation loss are shown in Figure 4.1 and 4.2. Test results are shown in Table 4.1. Validation and Test average losses are almost the same as shows in Figure 4.2.

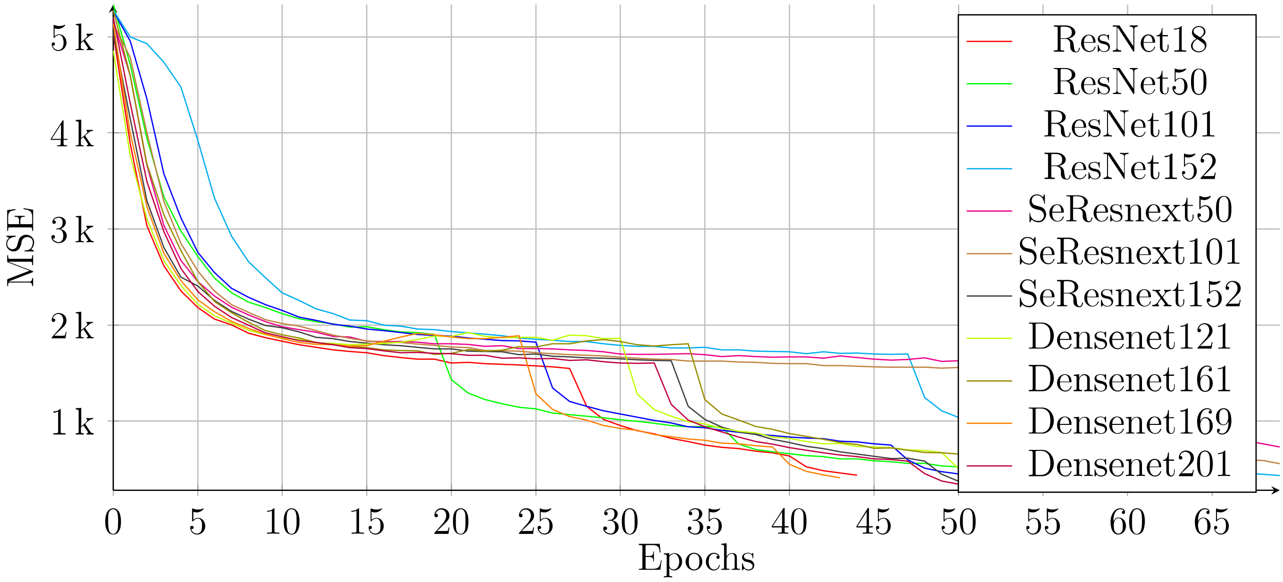


Figure 4.1: Training loss.

By carefully observing these two Figures, training losses are continuously

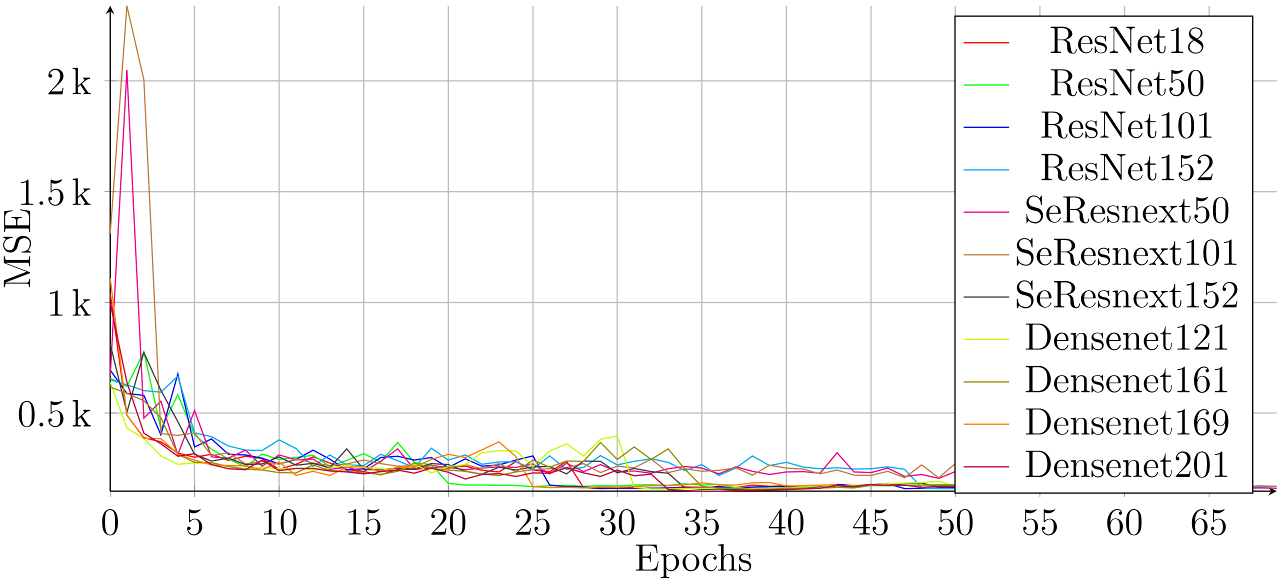


Figure 4.2: Validation loss.

decreasing, but tangents of the test loss lines increased. Moreover, in Table 4.1 the results are almost same. However, latest result improved about 2%. Furthermore, dataset labels may have some noise in the future. Therefore, as shown in Table 4.1, pretrained CNN network models such as LRCN were used.

Table 4.1: Deep models results.

Model Name	MSE	Test accuracy (Correct/Total %)
ResNet18	173.0	85.874% (2693/3136)
ResNet50	177.1	86.288% (2706/3136)
ResNet101	194.9	87.309% (2738/3136)
ResNet152	180.0	86.639% (2717/3136)
SeResnext50	166.0	86.161% (2702/3136)
SeResnext101	193.4	85.268% (2674/3136)
SeResnext152	182.1	87.277% (2737/3136)
Densenet121	206.3	85.491% (2681/3136)
Densenet161	197.3	85.810% (2691/3136)
Densenet169	197.3	86.511% (2713/3136)
Densenet201	201.6	86.607% (2716/3136)

Figure 4.3 shows confusion matrix of the average network and Figure 4.4 shows the trained models wrongly predicted facial images, in which true and predicted emotions are confused, a problem also addressed in [61]. Furthermore, emotion recognition can be confused in recognition using only facial images. Therefore, a model trained, including time dimension and well understood facial scene identity could produce better results.

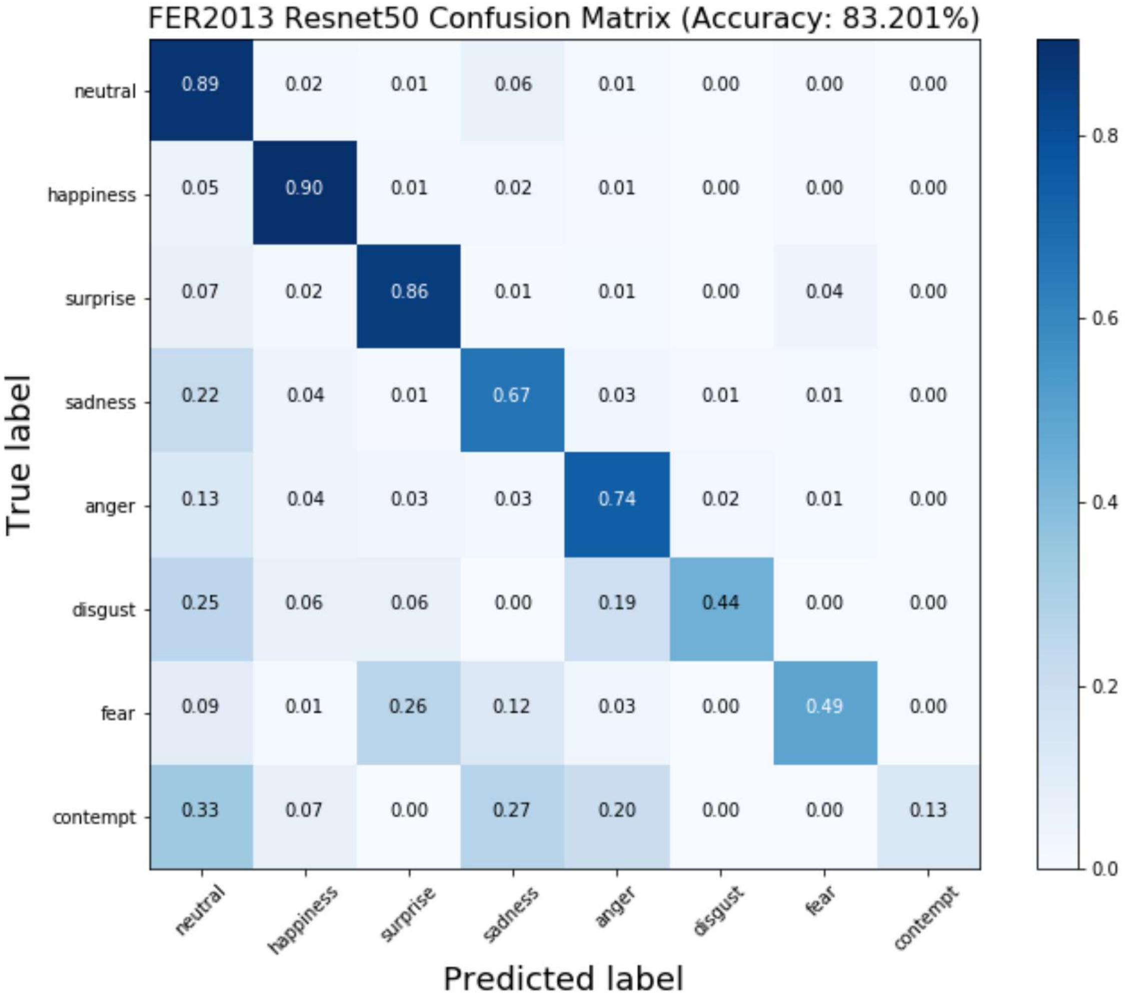


Figure 4.3: Confusion matrix on FERPlus dataset.

4.2 Video emotion recognition results

Due to comparison difficulty, the base unit in the CCC loss score was set as the base, which in training, $VA = -2$ was the best loss, and in validation, 2 was the best score. The 3 loss functions trained on the Aff-Wild dataset was observed in 2 ways; 5 times trained for stability check in 4 networks and split/trained candidates of the 4 networks in training and validation data set proportions 70% : 30% and 90% : 10% among the configuration sequence length.

Initially, for all the train and test sets, all body pose and facial landmarks were extracted using OpenPose [9]. From this output, objective personal data was obtained, using landmarks, by calculating a new facial bounding box expanded 5 pixels and then resizing it to 224x224 pixels. This facial bounding box was applied to extract features on the sequential image using the ResNet50 network. The ResNet50 network obtained the final linear layer output and saved 2048x1 features with facial landmarks (70x2) and body pose (18x2). Moreover, from the video, the audio signal was separated and converted to the Mel Spec-

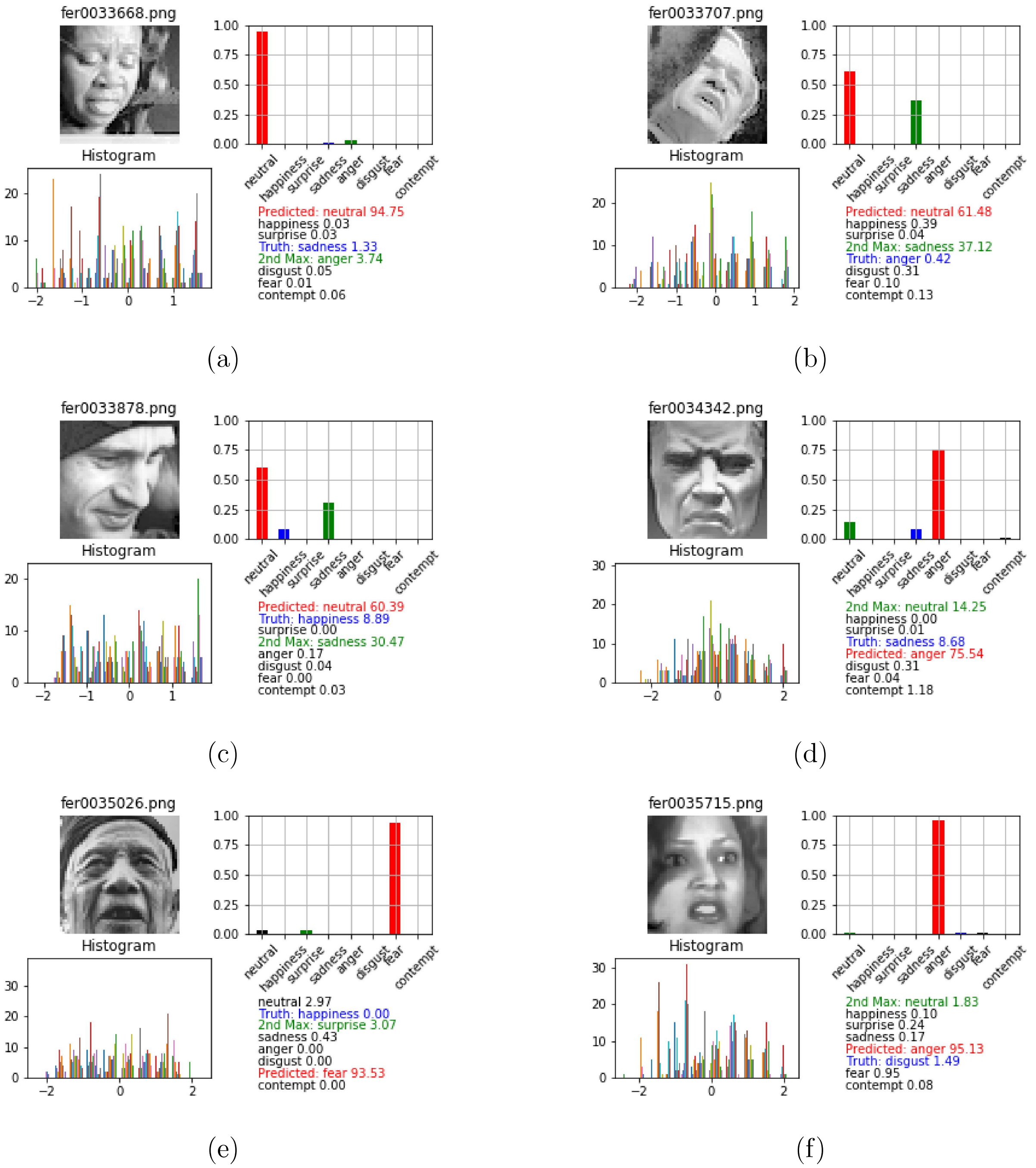


Figure 4.4: Common lost images.

trogram features. All training and validation data were normalized to between 0 and 1.

Training: All the following configurations were similar in the 4 candidate networks. The Aff-Wild dataset was split into 2 Train and Validation sets as 70%:30% and 90%:10% respectively to make sure that our loss function and models were accurate for both standard or less training and validation data proportions. During training, the 3 loss functions and the standard stochastic gradient descent optimizer, with momentum (0.9) and weight decay ($5e^{-4}$) were used. The number of epochs was set to 70. Early stopping (15 epochs) was used to prevent over-fitting. The Learning rate was 0.01 with the scheduler (patience 10 epochs). Batch size was set at 64, with the Sequence length varied in Validation:30% $[32, 64, 128, 256, 384]$ and Validation:10% $[128, 256, 512]$. However, because different experiments had different GPU consumption, the largest batch size that fits in Train and Validation set video length and our GPU memory (11GB) was chosen. However, as some of Validation:30% video sequences were less than 512, the max sequence length was reduced to 384. Iteration of data-loader, drop last, and shuffle parameters were set to False. Train and validation video sets were shuffled at the beginning of each epoch. The pre-trained ResNet50 model weights were fixed during training; all the other layers were trainable. Batch normalization and dropout were used after every fully connected layer.

Validation: The losses of MSE, Non-Inertial, and CCC were calculated in each video and sample output of arousal and valence saved using TensorBoard SummaryWriter. Additionally, all the best losses of the model were saved for the rest of the test set.

4.2.1 Validation 30%, 10%

In the train and validation, pipeline batch iteration took around $0.01sec$, which as expected, was fast enough even though it was the same when max sequence length 512 was used.

Fig.4.5a shows the comparison of our loss function, MSE, and CCC in various network parameter size space. From this result, the new loss function was sensitive to network parameter size just like the MSE function. In our candidate four networks, "Pose Net" input size is $[B, S, 36]$ normalized pose coordinates, which is small enough. The relation between pose and emotion is 0.163, which shows a very weak correlation.

Another expectation of the result in the "Audio Net" mean was 0.206, which shows intonation and emotion as weakly correlated. "Without Mel" and "Full Net" did not produce any notable difference in the CCC correlation. The results of these two networks' were 0.339 and 0.324, respectively. On the other hand, in these networks' training, the proposed loss function produced better results.

In Fig.4.6a, the four networks were trained using various sequence length $128, 256, 512$ and validation 10% proportion with the 3 loss functions. The result is the average of four candidate networks. From the result, the CCC loss shows the top result, because the CCC calculation uses $[std, mean]$ values of prediction, which is more advanced. On the other hand, these values showed some disadvantage if sequence lengths were not comfortable, as shown in Fig. 4.6a and 4.6b.

Next, the Validation:30% proportion was experimented on to verify that the result was the same as the previous experiment with additional sequence lengths. In Fig.4.5b, for the training mean sequence length in "Full Net" and "Without Mel", our loss function shows the best result and is 0.07 lower than 10% proportion. In Fig.4.6b the lower sequence length was extended to prove that the CCC loss has a disadvantage if sequence length in a small range is chosen. The loss would be worse than the other losses, which shows that in GPU memory a trade-off between sequence length with parameter size and accuracy may be necessary. Fig.4.5c shows the average of the previous two training sessions in different proportions of the Validation set. In the results of "Without Mel" and "Full Net" there is no big difference. In addition, the other two loss functions especially our new loss was shown to be stable and produced the best results among different sequence lengths.

4.2.2 5-fold training

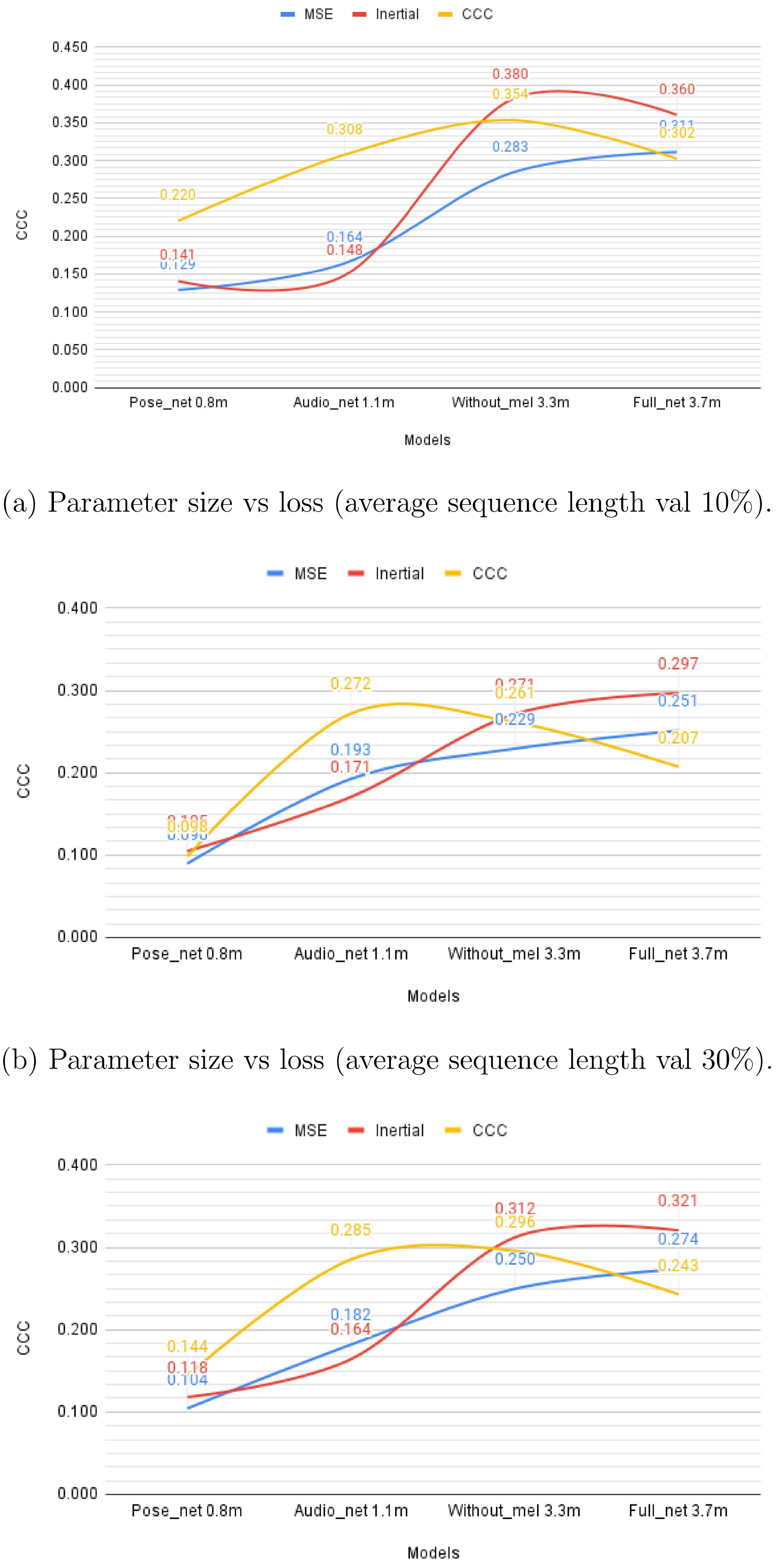
The Full network was trained five times, with a sequence length of 256, which was shown as optimal in previous training and the result of this training are shown in Table 4.2, where rows represent validation's main loss function and column shows parallel calculated losses. In Fig.4.7, the 5 training Average Line of Non-Inertial and MSE losses were nearly constant. It should be noted that the CCC loss average rapidly increased, which shows it is less stable than our new loss function.

Table 4.2: Full network loss of 5 training (sequence length 256).

Train Loss	Train	MSE	CCC	Non-Inertial	Average	Differences	Sum
MSE	1	0.209	0.331	0.105	0.210	0.0000014	
	2	0.211	0.346	0.107	0.210	0.0000005	
	3	0.207	0.366	0.104	0.210	0.0000116	
	4	0.225	0.337	0.113	0.210	0.0002326	
	5	0.199	0.333	0.100	0.210	0.0001291	0.000375
CCC	1	0.348	0.309	0.176	0.316	0.0000454	
	2	0.364	0.282	0.191	0.316	0.0011459	
	3	0.335	0.321	0.170	0.316	0.0000262	
	4	0.415	0.319	0.213	0.316	0.0000107	
	5	0.271	0.348	0.136	0.316	0.0010367	0.002265
Non-Inertial	1	0.202	0.411	0.102	0.103	0.0000025	
	2	0.206	0.355	0.104	0.103	0.0000006	
	3	0.204	0.414	0.103	0.103	0.0000000	
	4	0.202	0.415	0.102	0.103	0.0000016	
	5	0.209	0.376	0.105	0.103	0.0000047	0.000009

4.2.3 Example valence and arousal of full network

Figures 4.8a, 4.10a, and 4.9a show our new loss function training and sample outputs for arousal and valence. Figures 4.8b, 4.10b and 4.9b are the CCC loss training and sample output of the Validation video 158.avi. The best loss on "Full Net" in Validation was 0.86 for Arousal and 0.404 for Valence, for which back-propagation was done separately. Moreover, in Fig.4.10b Arousal was positively correlated but missed the constant value. In Figures 4.8a, and 4.8b training epochs was 70 but the training sessions did not reach 40 epochs in some models because of early stopping.

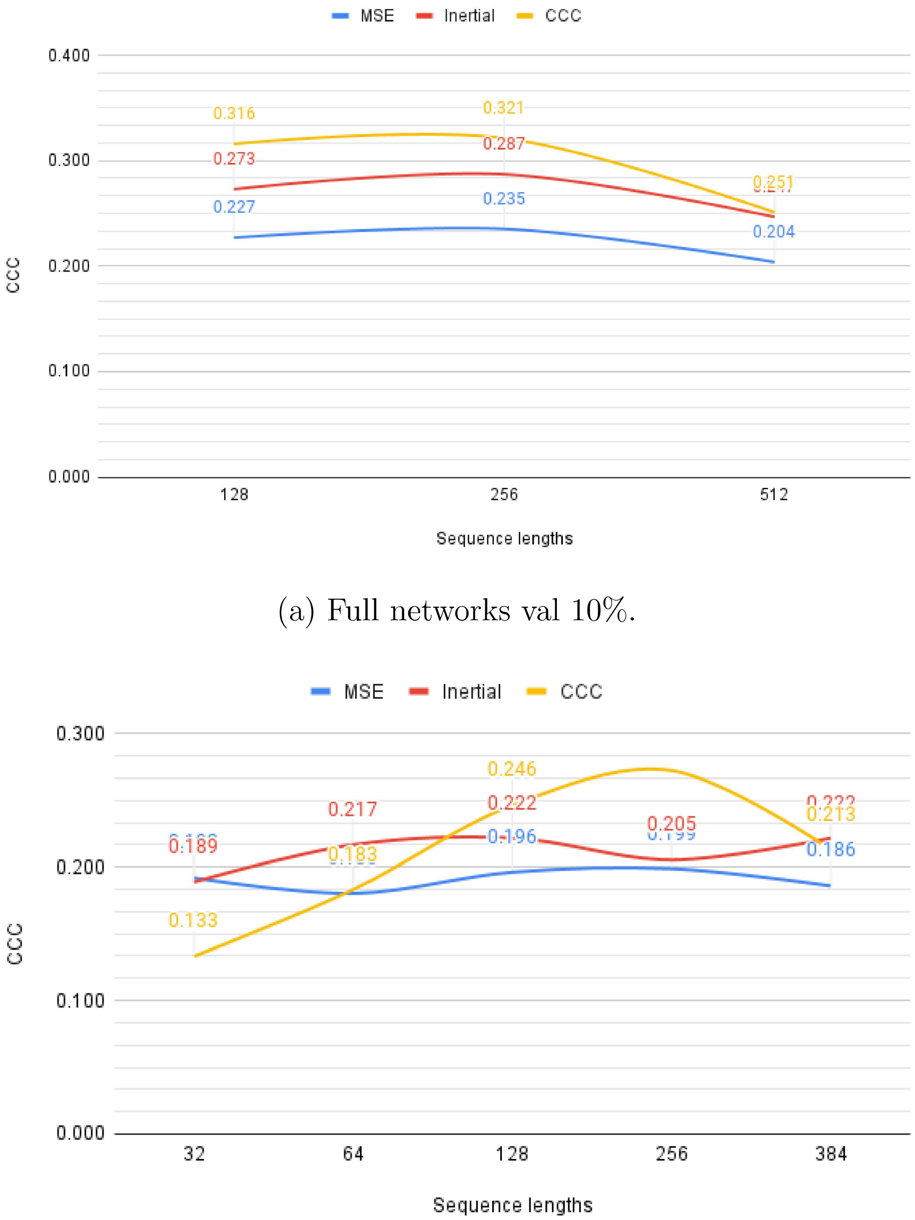


(a) Parameter size vs loss (average sequence length val 10%).

(b) Parameter size vs loss (average sequence length val 30%).

(c) Parameter size vs loss (total).

Figure 4.5: Comparison of validation size and model parameter size.



(a) Full networks val 10%.

(b) Full networks Val 30%.

Figure 4.6: Sequence length vs loss.

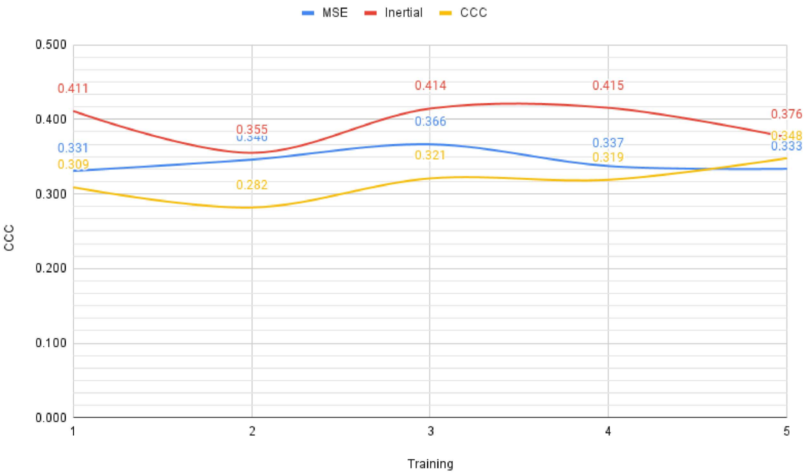
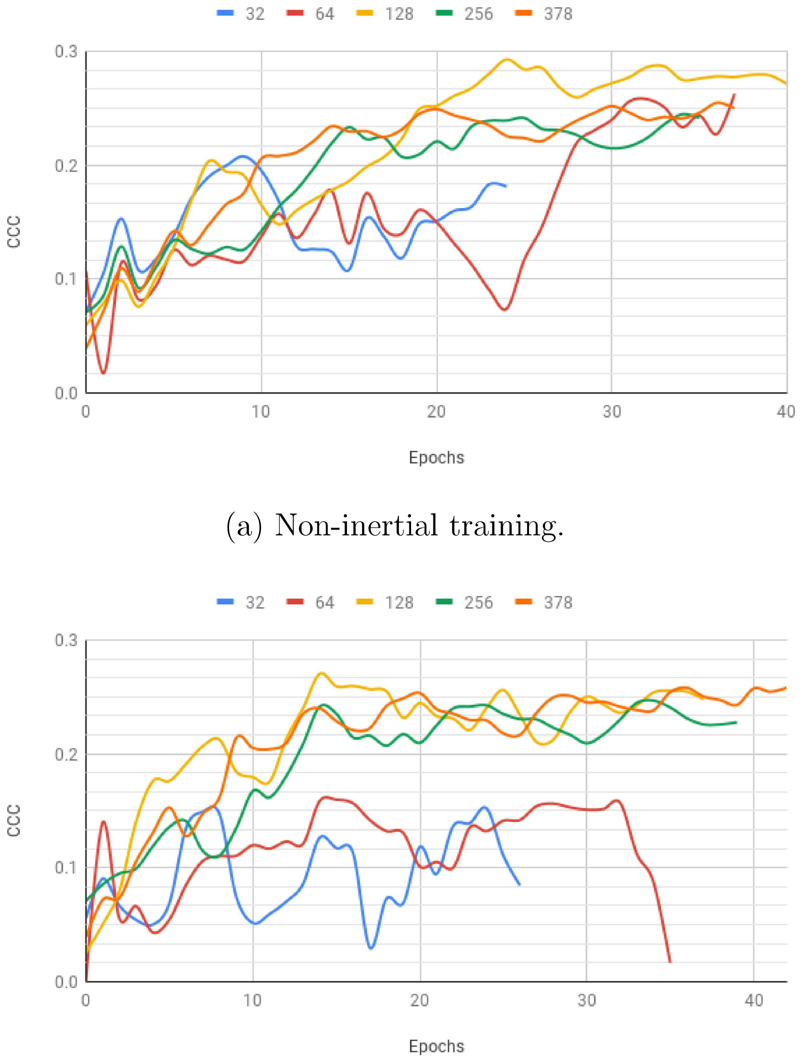


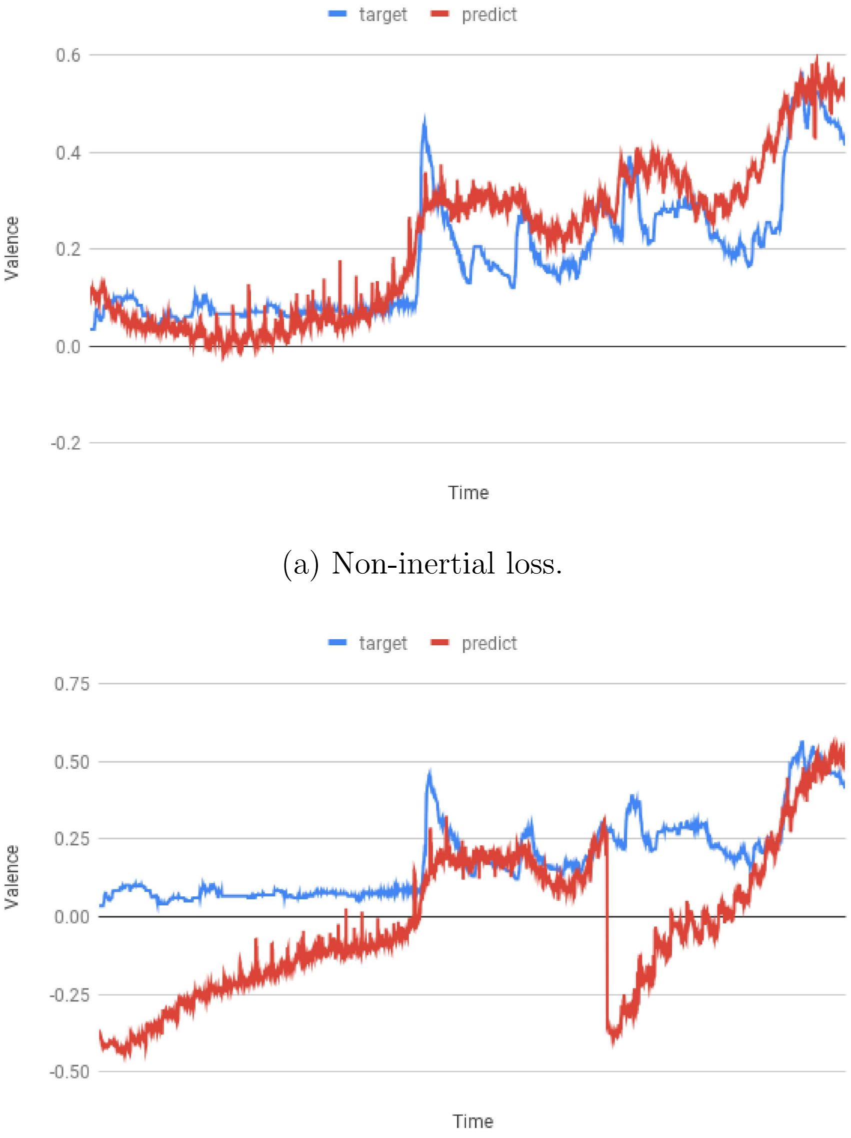
Figure 4.7: Full net 5 training CCC loss comparison (sequence 256 val 10%).



(a) Non-inertial training.

(b) CCC training.

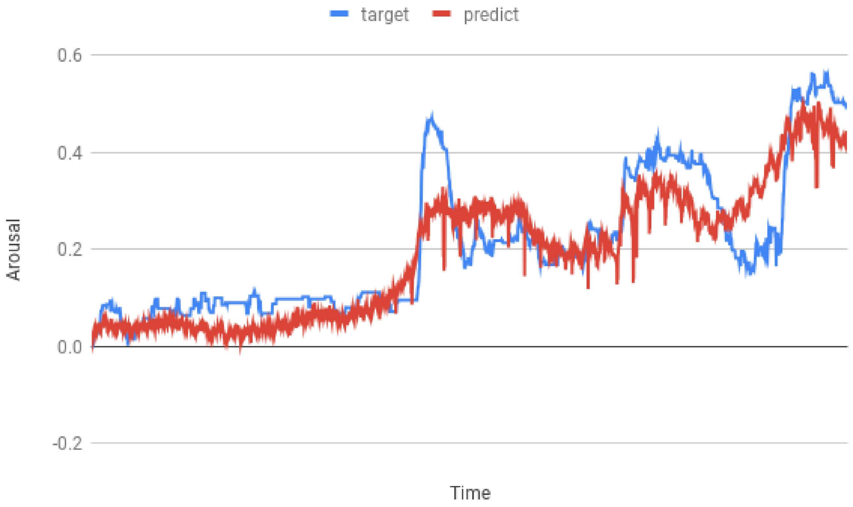
Figure 4.8: Full network training.



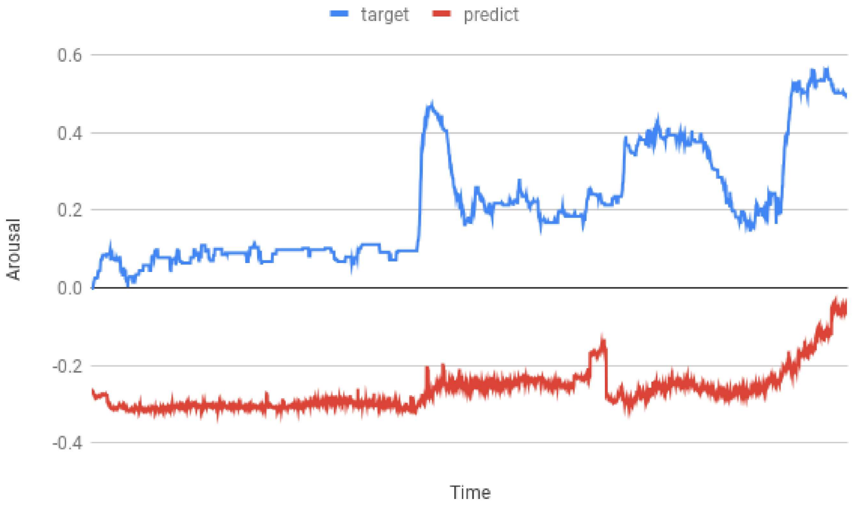
(a) Non-inertial loss.

(b) CCC loss.

Figure 4.9: 158.avi trained model valence comparison.



(a) Non-inertial loss.



(b) CCC loss.

Figure 4.10: 158.avi trained model arousal comparison.

4.3 Transformer Model results

The results of the training the transformer model are shown in Fig. 4.11. Initially, we compiled model training on Transformer model and obtained the results above. From the result MSE loss training epoch on 25 videos was 105 seconds for learning rate $1e^{-5}$, sequence length of 32, batch size of 8 and training epoch 50. The MSE loss showed bigger elbow shape on the graph and after 21 epoch training was stopped early, which was no success on the training. After training of the MSE loss in the Transformer model, the result showed 0.0363 on training set and 0.0315 on validation set, which meant good results were obtained. On the other hand, the implemented new loss function Non-Inertial had shown advantages on same settings of training. From the model training, new loss function shows far more advantages both in the training and the validation sets. In addition, its epoch training time was around 99 seconds, which is a faster calculation time. Training loss was same at 0.0363 for the training set, but rapidly decreasing to 0.0173 for validation set.

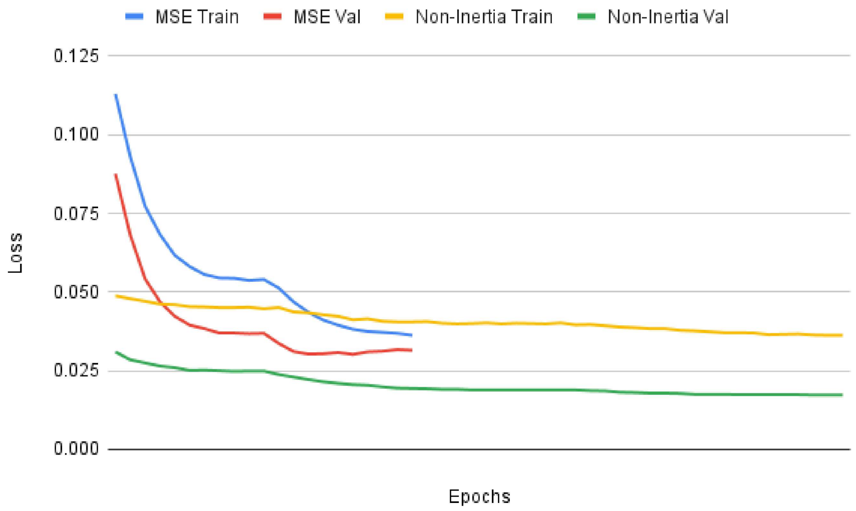


Figure 4.11: Transformer model training.

After compiling the model, we trained the transformer model in the same as the above configuration but set the sequence length to 128. In the previous experiment, sequence length 128 showed the best durability and effectiveness The trained transformer model results are shown in Table 4.3. From the table, CCC loss shows better results in all models, which proves the previous experiment in Fig. 1.1 that 128 sequence length shows better results than the new loss function. The Mesh Transformer model shows the best result in valence space because meshes represent “facial expression” (Table 1.1). Also, the Transformer model shows significant improvement from previous experiments when we compare “Feature Transformer” and “Full network” results. In comparison, two of the loss functions “EVM Transformer” model, shows better than others in Valence and Arousal spaces.

Table 4.3: Transformer model results

Models	Loss function	Epochs	Epoch duration @sec	Accuracy @2CCC	Valence	Arousal
Mesh Transformer	Ours	9	994	1.3311	0.7052	0.439
	CCC	10	882	1.16	0.6973	0.4981
Feature Transformer	Ours	10	1740	1.1768	0.5851	0.4313
	CCC	12	1428	1.2	0.6402	0.4848
EVM Transformer	Ours	11	3752	1.2286	0.5277	0.5807
	CCC	17	5104	1.24	0.6992	0.5767
Multi Transformer	Ours	24	6895	1.209	0.393	0.4497
	CCC	-	-	-	-	-

The trained transformer model produced the following results for a selected video in the database, Fig. 4.12. The true value was “suspicious” and the predicted value was “impatient” as shown on the figure on the left. The number indicates distance between emotional state and target or prediction state.

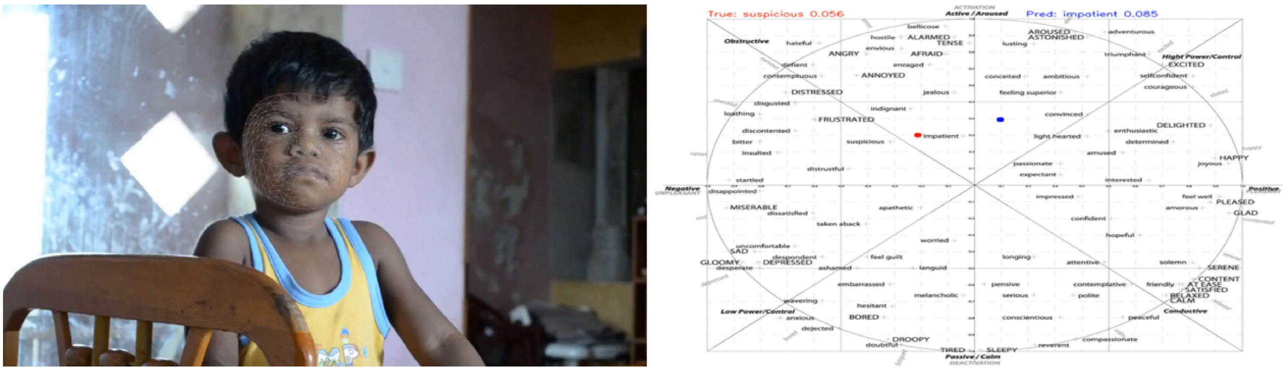


Figure 4.12: Transformer model inference.

4.4 Continual emotion classification results

At the beginning of the training, a model was trained on the original dataset until the best accuracy on test set was achieved. Carefully observing the trained classifier ResNet18 (Figure 3.6), the “hash” like weights, which in CIFAR-100 had some of classes with uniform distribution of feature space and some of classes were normal distribution. After training the universal model, we froze the base body of model, and added a parallel new $fc8$ layer to replace the n^{th} original $fc8$ layer. Then we trained the model on a new dataset with only the new $fc8$ layer. After that, we used the previous mixing fully connected layer method. The result of this method is shown in Table 4.4. By choosing α as balanced by mass. In 5 tasks of the testset, we achieved 68.7% accuracy, which reduced from 96.2%. In each task, the accuracy reduced about 8%. On the other hand, the training accuracy was stable around more than 95%. In this method model parameter size were maintained at only 605 parameters for each task.

Table 4.4: Method 1 training result on EMNIST letters dataset.

Original classes	New classes	Alpha	Test accuracy	Test Loss	Training accuracy	Trainable parameters	Total parameters
26	0		0.927	18.38	0.525	1,477,879	1,477,879
5	5	0.5	0.962	9.90	0.946	605	1,474,854
10	5	0.(6)	0.881	27.23	0.962	605	1,475,459
15	5	0.75	0.808	46.61	0.983	605	1,476,064
20	5	0.80	0.743	60.95	0.960	605	1,476,669
25	5	0.83	0.687	72.04	0.985	605	1,477,274

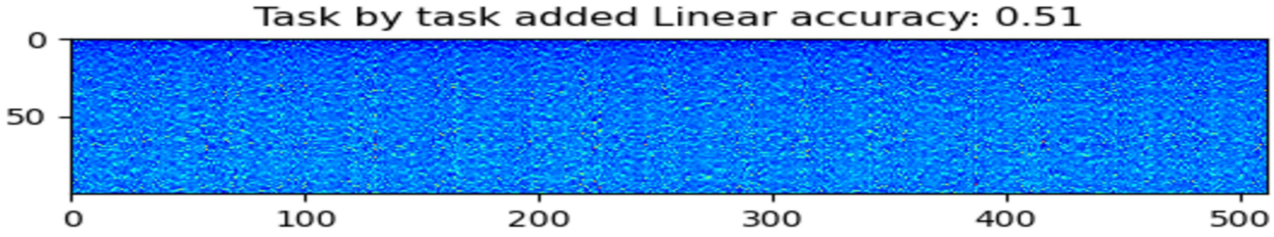


Figure 4.13: Task by task trained and added classifier's weight with 51% accuracy rate.

After this method, we experimented using the histogram equalization method. Firstly, we trained each task separately and concatenated. After concatenation, we set classifier weights as concatenated and tested the equalized weights. The result is shown in Figure 4.13. In the observation of added weights, tasks were

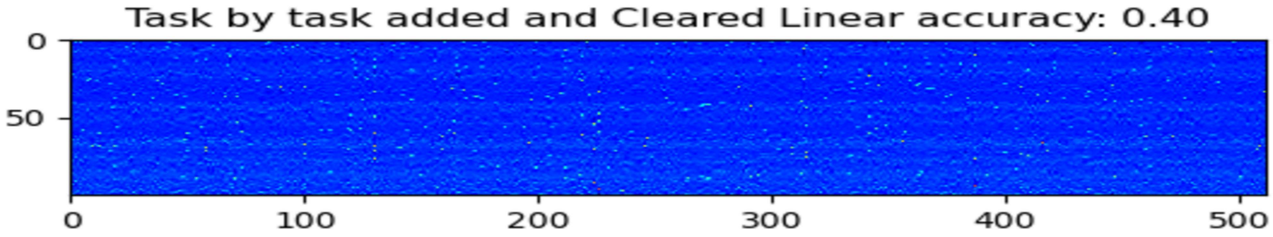


Figure 4.14: Task by task trained and added histogram cleaned classifier's weight with accuracy rate 40%.

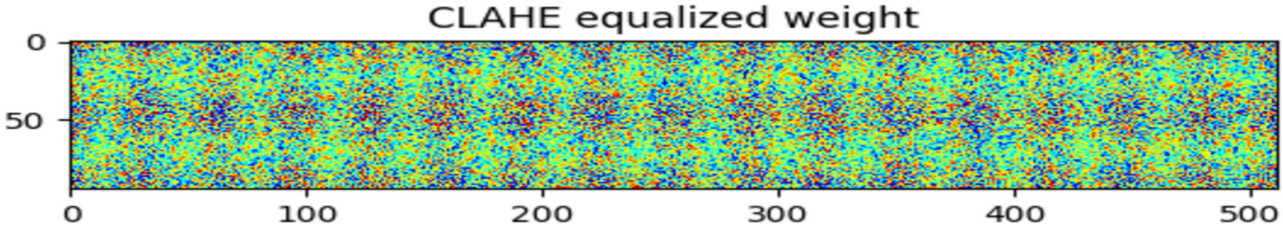


Figure 4.15: Task by task trained and added histogram equalized CLAHE classifier's weight with accuracy rate 2%.

trained by high contrast without consequence in the feature space and test accuracy was 51%. We also applied histogram equalization as a weight pruner from $[min(W_{task}) + 0.3, max(W_{task}) - 0.3]$ value range. Pruned weight results are shown in Figure 4.14.

Chapter 5

Conclusions and Future works

5.1 Conclusions on Image Emotion Recognition

In this thesis, a new loss function for emotion recognition was proposed and, using different proposed models, compared the training and result accuracies to two other existing loss functions. Emotion recognition can be recognized by not only using facial images, but also time dimension, body poses, facial landmarks and well understood the facial scene identity. Furthermore, we compared our work with LRCN and other state of art networks.

5.2 Conclusions on Video Emotion

A new loss function named “Non-Inertial Loss” was proposed and, its stability and effectiveness were proved using four networks designed with different sequence lengths, and validation proportion. We observed that, in data relation for continuous and interrupted data, if the network and data were low, the interruption effect was high. Conversely, if data and networks were large, the interruption effect was low.

The new loss function accuracy was limited by a smaller network parameter size, just like the other existing losses (CCC and MSE). However, in the sequence range the proposed loss function showed better results in the lower rank models, which allows better trade-off cost between the network size and sequence length.

The transformer model training on Non-Inertia loss function shows state-of-the-art results. Significantly, the EVM-Transformer model shows both arousal and valence were 0.5542. However, we are still training other loss functions

(MSE, CCC). On the other hand, the Multi-Transformer model requires more parameter settings with enormous parameter space, but the result is the lowest. The inference of Mesh-Transformer model shows promising result and movement of emotion is almost synchronized with ground truth. Reversely, scaling of prediction shows more implementation is need ed in the output section of the transformer model.

5.3 Conclusions on Continual Learning

We investigated image processing methods in continual learning field and proposed multiple methods such as alpha blending, histogram equalization and the pruner. The aim was dealing with the linear layer weights in continual learning. We build a universal network, using a previously trained one as a feature extractor. The methods got promising results as shown in the results. In the alpha blending method, we used a technique that separated knowledge from unknown knowledge. This method showed better results than the normal concatenation technique. The other method was applying histogram equalization. In this method, we used two kind of techniques; one was the equalizer as pruner and the other was the equalizer as enhancer.

The Cosine distance of the Pruner method trained weight was 0.23. However, the test set result showed a big difference. In CLAHE equalized weight, the enhanced weight hashes were mentioned before. The results of weight images showed no significant difference in critical features excluding CLAHE equalized weights, which meant that if feature extractor was the same, then training of continual learning separate or normal learning had no big difference.

5.4 Future Works

In the future, for emotion recognition, we will continue to test the new loss function with acceleration detailed in Appendix A.

We will use transfer learning for smaller models with low accuracy loss and implement emotion recognition end-to-end method. In addition, for dataset issues, most researchers aim to train a model using Self-supervised learning. Using Self-supervised learning, more interesting results may appear.

For continual learning, we aim to support more features and to clear up any consequence issue it may cause. If more training or weight space are needed, it is possible to extend network parameter sizes. Moreover, we will focus more on state of the art transfer learning experiments and regenerative memory replay methods.

In histogram equalization method we will look for a transformation function that solves the problem of catastrophic forgetting, an ongoing learning problem, by training using the *cdf* histogram equalizer.

Appendix A

Loss Function Details and Full Results

Extended proposed loss function

$$Distance = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{1}$$

$$Velocity = |\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\partial x_{i,j}}{\partial t} - \frac{\partial y_{i,j}}{\partial t}| \tag{2}$$

$$\begin{gathered}Alpha = tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} = |\frac{1}{N(M+1)} \\ \sum_{i=1}^{N}\sum_{j=1}^{M+1}(\frac{\partial x_{i,j}}{\partial t}\frac{\partial y_{i,j+1}}{\partial t} - \frac{\partial x_{i,j+1}}{\partial t}\frac{\partial y_{i,j}}{\partial t})|\end{gathered} \tag{3}$$

$$Acceleration = |\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\partial^2 x_{i,j}}{\partial^2 t} - \frac{\partial^2 y_{i,j}}{\partial^2 t}| \tag{4}$$

$$loss = \frac{D + V + A + Ac}{4} \tag{5}$$

Proof of Equation 3.5:

$$\begin{aligned}
\Delta\alpha = tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} &=> \\
\lim_{\Delta\alpha \to 0} \Delta\alpha = 0 \quad &=> \\
\Delta\alpha = 0 \quad &=> \\
0 = tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} \quad &* / tan \\
tan(0) = \frac{a_x}{v_x} - \frac{a_y}{v_y} \quad &=> \\
0 = \frac{a_x * v_y - a_y * v_x}{v_x * v_y} \quad &=> \\
0 = a_x * v_y - a_y * v_x \quad &=> \\
\Delta\alpha = a_x * v_y - a_y * v_x &
\end{aligned} \tag{6}$$

Proof of Equation 3.5 in an example:

$$\begin{gathered}
assume \\
\angle annotation = 90^\circ \\
\angle prediction = 30^\circ \\
\Delta\alpha = 90^\circ - 30^\circ = 60^\circ => \\
a_x = \sin(\pi/2) = 1 \quad v_x = \cos(\pi/2) = 0 \\
a_y = \sin(\pi/6) = 1/2 \quad v_y = \cos(\pi/6) = \sqrt{3}/2 \\
\sin(A - B) = \sin(A) * \cos(B) - \cos(A) * \sin(B) \\
\Delta\alpha = \arcsin(a_x * v_y - a_y * v_x) \\
\Delta\alpha = \arcsin(1 * \sqrt{3}/2 - 0 * 1/2) \\
\Delta\alpha = \arcsin(\sqrt{3}/2) = 60^\circ \\
60^\circ = 60^\circ
\end{gathered} \tag{7}$$

Training tables

Table 1: Validation: 10% training.

Network	Train Loss	Sequence Length	MSE	CCC	Non-Inertial
Full Network	MSE	128	0.233	0.334	0.117
		256	0.245	0.312	0.123
		512	0.229	0.288	0.115
	CCC	128	0.293	0.307	0.155
		256	0.303	0.338	0.161
		512	0.380	0.262	0.195
	Non-Inertial	128	0.216	0.404	0.109
		256	0.214	0.334	0.108
		512	0.217	0.343	0.109
Without mel Spectrogram	MSE	128	0.217	0.319	0.109
		256	0.226	0.277	0.115
		512	0.223	0.254	0.112
	CCC	128	0.287	0.426	0.147
		256	0.321	0.342	0.163
		512	0.330	0.292	0.182
	Non-Inertial	128	0.214	0.430	0.108
		256	0.210	0.375	0.106
		512	0.214	0.335	0.108
Audio Network	MSE	128	0.213	0.151	0.107
		256	0.199	0.202	0.100
		512	0.218	0.139	0.110
	CCC	128	0.231	0.300	0.116
		256	0.252	0.349	0.126
		512	0.256	0.274	0.129
	Non Inertial	128	0.208	0.131	0.105
		512	0.213	0.165	0.107
Pose Network	MSE	128	0.234	0.103	0.118
		256	0.232	0.150	0.117
		512	0.222	0.134	0.112
	CCC	128	0.335	0.230	0.180
		256	0.329	0.255	0.179
		512	0.457	0.176	0.247
	Non-Inertial	128	0.222	0.126	0.112
		256	0.229	0.151	0.115
		512	0.225	0.144	0.113

Table 2: 5 Training (sequence 256 val 10%).

Network	Train Loss	Fold	MSE	CCC	Non-Inertial
Full Network	MSE	1	0.209	0.331	0.105
		2	0.211	0.346	0.107
		3	0.207	0.366	0.104
		4	0.225	0.337	0.113
		5	0.199	0.333	0.100
	CCC	1	0.348	0.309	0.176
		2	0.364	0.282	0.191
		3	0.335	0.321	0.170
		4	0.415	0.319	0.213
		5	0.271	0.348	0.136
	Non-Inertial	1	0.202	0.411	0.102
		2	0.206	0.355	0.104
		3	0.204	0.414	0.103
		4	0.202	0.415	0.102
		5	0.209	0.376	0.105
Without mel Spectrogram	MSE	1	0.209	0.328	0.105
		2	0.225	0.354	0.115
		3	0.227	0.304	0.115
		4	0.235	0.326	0.118
		5	0.227	0.304	0.114
	CCC	1	0.332	0.342	0.176
		2	0.344	0.310	0.176
		3	0.300	0.353	0.153
		4	0.383	0.339	0.199
		5	0.292	0.330	0.156
	Non-Inertial	1	0.209	0.398	0.105
		2	0.214	0.370	0.108
		3	0.206	0.402	0.104
		4	0.206	0.379	0.104
		5	0.205	0.375	0.103
Audio Network	MSE	1	0.214	0.180	0.108
		2	0.216	0.250	0.109
		3	0.219	0.193	0.110
		4	0.219	0.172	0.110
		5	0.217	0.211	0.109
	CCC	1	0.246	0.300	0.124
		2	0.239	0.288	0.120
		3	0.246	0.314	0.124
		4	0.239	0.321	0.120
		5	0.239	0.319	0.120
	Non-Inertial	1	0.215	0.221	0.108
		2	0.214	0.175	0.108
		3	0.212	0.209	0.107
		4	0.214	0.187	0.108
		5	0.215	0.182	0.108
Pose Network	MSE	1	0.226	0.097	0.114
		2	0.226	0.075	0.114
		3	0.228	0.071	0.115
		4	0.223	0.071	0.112
		5	0.233	0.077	0.117
	CCC	1	0.466	0.210	0.239
		2	0.389	0.205	0.203
		3	0.453	0.143	0.235
		4	0.360	0.173	0.187
		5	0.360	0.176	0.191
	Non-Inertial	1	0.228	0.069	0.115
		2	0.233	0.084	0.117
		3	0.227	0.075	0.114
		4	0.228	0.091	0.115
		5	0.226	0.085	0.113

Table 3: Validation: 30% training.

Network	Train Loss	Sequence Length	MSE	CCC	Non-Inertial
Full Network	MSE	32	0.203	0.258	0.102
		64	0.201	0.251	0.101
		128	0.204	0.265	0.103
		256	0.198	0.251	0.100
		384	0.206	0.230	0.104
	CCC	32	2.470	0.000	1.235
		64	0.371	0.177	0.192
		128	0.291	0.289	0.148
		256	0.274	0.287	0.144
		384	0.281	0.283	0.149
	Non-Inertial	32	0.202	0.284	0.101
		64	0.204	0.346	0.103
		128	0.193	0.322	0.097
		256	0.194	0.260	0.098
		384	0.203	0.271	0.102
Without mel Spectrogram	MSE	32	0.203	0.249	0.102
		64	0.217	0.233	0.110
		128	0.202	0.241	0.102
		256	0.202	0.223	0.102
		384	0.208	0.201	0.105
	CCC	32	0.270	0.250	0.137
		64	0.359	0.229	0.180
		128	0.325	0.262	0.168
		256	0.268	0.290	0.140
		384	0.287	0.275	0.152
	Non-Inertial	32	0.208	0.243	0.105
		64	0.202	0.305	0.102
		128	0.191	0.285	0.096
		256	0.194	0.234	0.097
		384	0.195	0.290	0.098
Audio Network	MSE	32	0.213	0.202	0.107
		64	0.205	0.205	0.103
		128	0.201	0.184	0.101
		256	0.194	0.192	0.098
		384	0.200	0.180	0.101
	CCC	32	0.249	0.238	0.125
		64	0.244	0.252	0.123
		128	0.231	0.278	0.116
		256	0.251	0.300	0.126
		384	0.234	0.293	0.118
	Non-Inertial	32	0.202	0.184	0.102
		64	0.201	0.160	0.101
		128	0.199	0.162	0.100
		256	0.191	0.177	0.096
		384	0.195	0.171	0.098
Pose Network	MSE	32	0.229	0.059	0.115
		64	0.213	0.032	0.107
		128	0.209	0.094	0.105
		256	0.204	0.129	0.103
		384	0.207	0.134	0.104
	CCC	32	0.373	0.045	0.194
		64	0.432	0.076	0.225
		128	0.365	0.157	0.193
		256	0.350	0.213	0.179
		384	2.265	0.001	1.133
	Non-Inertial	32	0.212	0.046	0.106
		64	0.207	0.056	0.104
		128	0.205	0.118	0.103
		256	0.204	0.151	0.103
		384	0.202	0.154	0.101

Bibliography

[1] J. Westerink M. Krans and Martin. *Sensing Emotions: The Impact of Context on Experience Measurements Table 2.1.* Reading, Kindle: Springer, 2011.

[2] Alyssa Khoo. *The Emotional Brain.* 2020. URL: https://ysjournal.com/the-emotional-brain/.

[3] Elizabeth A. Clark et al. "The Facial Action Coding System for Characterization of Human Affective Response to Consumer Product-Based Stimuli: A Systematic Review". In: *Frontiers in Psychology* 11 (2020), p. 920. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.00920. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2020.00920.

[4] P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS).* Series in Affective Science. Oxford University Press, 2005. ISBN: 9780199792726. URL: https://books.google.co.jp/books?id=UXapcWqt0-sC.

[5] Izard C. "The maximally discriminative facial movement coding system". In: (1979).

[6] Silvan S. Tomkins Paul Ekman Wallace V. Friesen. *Facial affect scoring technique: a first valid study.* 1971. URL: https://bit.ly/3pvwfsC.

[7] Louis G Tassinary, John T Cacioppo, and Thomas R Geen. "A psychometric study of surface electrode placements for facial electromyographic recording: I. The brow and cheek muscle regions". In: *Psychophysiology* 26.1 (1989), pp. 1–16.

[8] Tadas Baltrušaitis Amir Zadeh Yao Chong Lim and Louis-Philippe Morency. "OpenFace 2.0: Facial Behavior Analysis Toolkit". In: 2018.

[9] Zhe Cao et al. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *CoRR* abs/1812.08008 (2018). arXiv: 1812.08008. URL: http://arxiv.org/abs/1812.08008.

[10] Yury Kartynnik et al. "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs". In: *CoRR* abs/1907.06724 (2019). arXiv: 1907.06724. URL: http://arxiv.org/abs/1907.06724.

[11] Google AI. *MediaPipe*. 2020. URL: https://google.github.io/mediapipe/.

[12] Promod Yenigalla et al. "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding". In: *INTERSPEECH*. 2018, pp. 3688–3692.

[13] Woo Yong Choi, Kyu Ye Song, and Chan Woo Lee. "Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data". In: *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 28–34. DOI: 10.18653/v1/W18-3304. URL: https://aclanthology.org/W18-3304.

[14] Hao-Yu Wu et al. "Eulerian Video Magnification for Revealing Subtle Changes in the World". In: *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31.4 (2012).

[15] Priyanka Abhang et al. "Emotion Recognition Using Speech and EEG Signal? A Review". In: *International Journal of Computer Applications* 15 (Feb. 2011), pp. 37–40. DOI: 10.5120/1925-2570.

[16] Paul Ekman et al. "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion". In: *Journal of personality and social psychology* 53 (Nov. 1987), pp. 712–7. DOI: 10.1037/0022-3514.53.4.712.

[17] Dawood Al Chanti and Alice Caplier. "Spontaneous Facial Expression Recognition using Sparse Representation". In: *VISIGRAPP 2017 - 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. Vol. 5. Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 64-74. Porto, Portugal, Feb. 2017, p. 11. DOI: 10.5220/0006118000640074. URL: https://hal.univ-grenoble-alpes.fr/hal-01485279.

[18] Ivona Tautkute and Tomasz Trzcinski. "Classifying and Visualizing Emotions with Emotional DAN". In: *CoRR* abs/1810.10529 (2018). arXiv: 1810.10529. URL: http://arxiv.org/abs/1810.10529.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[20] Anbang Yao et al. “Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI ’15. Seattle, Washington, USA: Association for Computing Machinery, 2015, pp. 451–458. ISBN: 9781450339124. DOI: 10.1145/2818346.2830585. URL: https://doi.org/10.1145/2818346.2830585.

[21] Caifeng Shan, Shaogang Gong, and Peter Mcowan. “Facial expression recognition based on Local Binary Patterns: A comprehensive study”. In: *Image and Vision Computing* 27 (May 2009), pp. 803–816. DOI: 10.1016/j.imavis.2008.08.005.

[22] Samira Ebrahimi Kahou et al. “Recurrent Neural Networks for Emotion Recognition in Video”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI ’15. Seattle, Washington, USA: Association for Computing Machinery, 2015, pp. 467–474. ISBN: 9781450339124. DOI: 10.1145/2818346.2830596. URL: https://doi.org/10.1145/2818346.2830596.

[23] Valentin Vielzeuf et al. “An Occam’s Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets”. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI ’18. Boulder, CO, USA: Association for Computing Machinery, 2018, pp. 589–593. ISBN: 9781450356923. DOI: 10.1145/3242969.3264980. URL: https://doi.org/10.1145/3242969.3264980.

[24] Jianfei Yang et al. “Deep Recurrent Multi-Instance Learning with Spatio-Temporal Features for Engagement Intensity Prediction”. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI ’18. Boulder, CO, USA: Association for Computing Machinery, 2018, pp. 594–598. ISBN: 9781450356923. DOI: 10.1145/3242969.3264981. URL: https://doi.org/10.1145/3242969.3264981.

[25] Gil Levi and Tal Hassner. “Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI ’15. Seattle, Washington, USA: Association for Computing Machinery, 2015, pp. 503–510. ISBN: 9781450339124. DOI: 10.1145/2818346.2830587. URL: https://doi.org/10.1145/2818346.2830587.

[26] Ken Chatfield et al. “Return of the Devil in the Details: Delving Deep into Convolutional Nets”. In: *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014* (May 2014). DOI: 10.5244/C.28.6.

[27] Dong Yi et al. “Learning Face Representation from Scratch”. In: (Nov. 2014).

[28] Abhinav Dhall et al. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark". In: Nov. 2011, pp. 2106–2112. DOI: 10.1109/ICCVW.2011.6130508.

[29] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[30] Jie Hu et al. "Squeeze-and-Excitation Networks". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 42.8 (Aug. 2020), pp. 2011–2023. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2019.2913372. URL: https://doi.org/10.1109/TPAMI.2019.2913372.

[31] Gao Huang et al. "Densely Connected Convolutional Networks". In: July 2017. DOI: 10.1109/CVPR.2017.243.

[32] Shan Li and Weihong Deng. "Deep Facial Expression Recognition: A Survey". In: *IEEE Transactions on Affective Computing* (2020), pp. 1–1. DOI: 10.1109/TAFFC.2020.2981446.

[33] German Ignacio Parisi et al. "Continual Lifelong Learning with Neural Networks: A Review". In: *CoRR* abs/1802.07569 (2018). arXiv: 1802.07569. URL: http://arxiv.org/abs/1802.07569.

[34] James Mcclelland, Bruce Mcnaughton, and Randall O'Reilly. "Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory". In: *Psychological review* 102 (Aug. 1995), pp. 419–57. DOI: 10.1037/0033-295X.102.3.419.

[35] Ching-Yi Hung et al. "Compacting, Picking and Growing for Unforgetting Continual Learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13647–13657.

[36] Andrei A. Rusu et al. "Sim-to-Real Robot Learning from Pixels with Progressive Nets". In: *CoRR* abs/1610.04286 (2016). arXiv: 1610.04286. URL: http://arxiv.org/abs/1610.04286.

[37] Cheng-Hao Tu, Cheng-En Wu, and Chu-Song Chen. "Extending Conditional Convolution Structures For Enhancing Multitasking Continual Learning". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 1605–1610.

[38] Steven CY Hung et al. "Increasingly Packing Multiple Facial-Informatics Modules in A Unified Deep-Learning Model via Lifelong Learning". In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM. 2019, pp. 339–343.

[39] Arun Mallya and Svetlana Lazebnik. *PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning.* 2018. arXiv: 1711.05769 [cs.CV].

[40] Alex Krizhevsky. *Learning multiple layers of features from tiny images.* Tech. rep. Toronto University, 2009.

[41] Dimitrios Kollias et al. *Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond.* Apr. 2018.

[42] Dawood Al Chanti and Alice Caplier. "Spontaneous Facial Expression Recognition using Sparse Representation". In: *CoRR* abs/1810.00362 (2018). arXiv: 1810.00362. URL: http://arxiv.org/abs/1810.00362.

[43] P. Ekman and E. Rosenberg. "What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (FACS)". In: 2005.

[44] Lin Zhong et al. "Learning Multiscale Active Facial Patches for Expression Analysis". In: vol. 45. June 2012, pp. 2562–2569. ISBN: 978-1-4673-1226-4. DOI: 10.1109/CVPR.2012.6247974.

[45] Daniel D Duncan and Gautam Shine. "Facial Emotion Recognition in Real Time". In: 2016.

[46] Emad Barsoum et al. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* ICMI '16. Tokyo, Japan: Association for Computing Machinery, 2016, pp. 279–283. ISBN: 9781450345569. DOI: 10.1145/2993148.2993165. URL: https://doi.org/10.1145/2993148.2993165.

[47] Selvarajah Thuseethan and Sinnathamby Kuhanesan. "Eigenface Based Recognition of Emotion Variant Faces". In: *Computer Engineering and Intelligent Systems* 5 (June 2014). DOI: 10.2139/ssrn.2752808.

[48] Raviteja Vemulapalli and A. Agarwala. "A Compact Embedding for Facial Expression Similarity". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5676–5685.

[49] R. Plutchik. *Emotion, a Psychoevolutionary Synthesis.* Harper & Row, 1980. ISBN: 9780060452353. URL: https://books.google.co.jp/books?id=G5t9AAAAMAAJ.

[50] A. Mollahosseini, B. Hasani, and M. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10 (2019), pp. 18–31.

[51] Didan Deng et al. “MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (Apr. 2020), pp. 2621–2628. DOI: 10.1609/aaai.v34i03.5646.

[52] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. “FATAUVA-Net : An Integrated Deep Learning Framework for Facial Attribute Recognition, Action Unit Detection, and Valence-Arousal Estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. July 2017.

[53] Vassil Panayotov et al. “LibriSpeech: an ASR corpus based on public domain audio books”. In: *(ICASSP)* (2015), pp. 5206–5210.

[54] Arsha Nagrani et al. “Voxceleb: Large-scale speaker verification in the wild”. In: *Computer Science and Language* (2019).

[55] Jean Kossaifi et al. “AFEW-VA database for valence and arousal estimation in-the-wild”. In: *Image and Vision Computing* 65 (2017). Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing, pp. 23–36. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2017.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0262885617300379.

[56] Dimitrios Kollias and Stefanos Zafeiriou. “Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition”. In: *CoRR* abs/1811.07770 (2018). arXiv: 1811.07770. URL: http://arxiv.org/abs/1811.07770.

[57] L. Lin. “A concordance correlation coefficient to evaluate reproducibility.” In: *Biometrics* 45 1 (1989), pp. 255–68.

[58] Dimitrios Kollias et al. “Deep Affect Prediction In-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond”. In: *Int. J. Comput. Vision* 127.6–7 (June 2019), pp. 907–929. ISSN: 0920-5691. DOI: 10.1007/s11263-019-01158-4. URL: https://doi.org/10.1007/s11263-019-01158-4.

[59] Jargalsaikhan Orgil and Stephen Karungaru. “Facial emotion recognition using deep learning”. In: *SAMCON2020 IEEJ Tokyo, Japan*. Mar. 2020.

[60] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929. URL: https://arxiv.org/abs/2010.11929.

[61] Emad Barsoum et al. “Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution”. In: *CoRR* 1608.01041 (2016). arXiv: 1608.01041. URL: http://arxiv.org/abs/1608.01041.