Spontaneous Human Emotion Recognition from Video

Mounira Tlili

CMU-CS-19-124

August 2019

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Khaled Harras, Advisor Carnegie Mellon University Qatar
Bhiksha Raj, Advisor
Rita Singh, Advisor
Jeffery Cohon, University of Pittsburgh

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science.

Keywords: Deep Learning, Machine Learning, Emotion Recognition, Emotions, Facial Expressions, Micro Expressions, Facial Action Coding System, Action Units, Lie Detection, Psychology, MobileNet, Depthwise Convolutions, 3D Convolutions

Dedicated to the loving memory of my grandfather, and to my mother and father.

Abstract

From the computer vision perspective, the problem of automated facial expression analysis is a cornerstone towards high level human computer interaction. In this research we are interested in detecting true (non acted) human emotions for applications such as automatic lie detection, psychological diagnostics and the detection of malicious intents in public spaces. We use a guided emotion detection and identification approach based on facial muscle contractions and relaxations. Our method is composed of two main parts: (1) Action Unit[1] detection which reaches a very high average precision per Action Unit. (2) Action Unit to emotion mapping - we developed a highly accurate expert network that sees through fake emotions and detects even the slightest micro-expressions[2]. This results in a complete and accurate facial emotion recognition system.

[1] An Action Unit [AU] is a facial muscle contraction.

[2] Micro-expressions are discussed in Section 2.1

Acknowledgments

I'd like to thank my advisor Professor Bhiksha Raj and Professor Rita Singh for all the support they provided me both moral and educational throughout the entire year.

A very special gratitude goes out to my advisor Professor Khaled Harras, Program Director and Teaching Professor of the Computer Science department in Carnegie Mellon Qatar for his consistent help and support.

I would also like to thank Professor Jeffrey Cohn, Professor of Psychology at University of Pittsburgh for explaining to me all the psychological facts needed and providing me with psychology books to support my research. He provided me with an excellent emotion dataset that is very well labelled which I used throughout my entire research.

I am also grateful to president Farnam Jahanian for funding my thesis partially as part of a NeuroHackathon prize.

I want to finally thank my friends Mohammed Fituri, Asma Sahli and Salma Moustfa for their constant love and support.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

From the computer vision perspective, the problem of automated facial expression analysis is a cornerstone towards high level human computer interaction, and its study is a long tradition within the community. Various fields benefit from an improvement in human emotion recognition. These fields include, but are not limited to,

- Automatic lie detection
- Psychological diagnostics
- Smart homes
- Education, particularly in understanding whether students are engaged, bored or confused
- public safety, such as detecting malicious intents in public spaces such that prompt against can be taken
- Healthcare, for example the ability to identify the emotional state of patients, even when they are not able to communicate verbally at all

1.2 What are emotions

Emotions are both a psychological and physiological state. On one hand, an emotion is a natural instinctive state of mind that derives from one's circumstances, mood, and relationships with others. On the other hand, an emotion is the perception of change in the body such as the heart rate, breathing rate, perspiration, and hormone levels.

1.3 How is emotion recognition done

The study of emotion has been around since 1872 when Charles Darwin theorized - in his book The Expression of the Emotions in Man and Animals - that emotions are evolved traits universal to the human species. Current emotion recognition is conducted either by hand, where psychologists manually interpret facial expressions and body language, or with the use of polygraphs,

where physiological vitals such as breathing rate and heart rate are read and interpreted, such as in lie detection for example. With the explosion of neural networks and machine learning, recent efforts have shifted towards automatic emotion detection from audio and video.

1.4 Our focus

In this work we focus on emotion recognition from facial expressions. This is quite challenging mainly due to the fact that facial expressions can be very easily faked. Therefore, we will be focusing on the problem of **differentiating between True Spontaneous emotion and Acted Fake emotion**.

Chapter 2

Related work

2.1 Without AUs

Researchers in [10] propose a method for recognizing emotions from video streams. Videos of 60 subjects and six emotions from the BU4DFE database were considered. Optical flow was estimated between the facial points in neutral and apex frames, and a feature vector from the optical flows is obtained and passed to a neural network for emotion classification. The paper reaches an accuracy of 70% to 75% on two different datasets.

Researchers at CMU developed a micro-expression recognition system that improves upon state of the art [8]. They used machine-learned features that treat the whole face as a canvas, in contrast to traditional hand-crafted features and techniques that search pre-defined areas of the face for facial action units. In their system, the machine-learned features are generated with a pre-trained convolutional neural network. Optical flow data is then combined with frame-by-frame pixel information to better incorporate temporal structure into the recognition model. They utilized several pre-processing techniques, such as video interpolation via graph embedding to improve and maintain accuracy while reducing runtime.

2.2 AU related

Some researchers focus on using deep neural networks for AU detection and have managed to get better accuracies than conventional methods. In [14], the authors develop a system where they crop out facial images by using morphology operations including binary segmentation, connected components labeling and region boundaries extraction. Then for each type of AU, they train a corresponding expert network by specifically fine-tuning the VGG-Face network on cross-view facial images, so as to extract more discriminative features for the subsequent binary classification. They achieved an AU detection accuracy of 77.8%. The authors of [4] proposed to jointly address three issues: spatial representation, temporal modeling and AU correlation by using a hybrid network. They adopted CNNs to extract spatial representation. Long Short-Term Memory (LSTM) was employed to model temporal dependencies. They have improved the accuracy for classification of the different Action Units.

Applications such as Affectiva have also incorporated Action Units for their emotion recognition, although they do not disclose their exact architecture. They report highly accurate emotion recognition.

Chapter 3

Background

3.1 Micro-expressions

Micro-expressions, also referred to as micro-emotions, are the result of a voluntary emotion conflicting with an involuntary emotion. This occurs when a person attempts to conceal his/her true emotion towards a stimulus. The result is a brief expression of the involuntary emotion (true emotion), which is then suppressed by the voluntary emotion (fake emotion). The duration of a micro-expression is a fraction of a second, as opposed to an ordinary involuntary emotion which can be sustained for seconds.

3.2 Delving into neurology

In this section I would like to mainly summarize the findings of the article *The Neuropsychology of Facial Expression: A Review of the Neurological and Psychological Mechanisms for Producing Facial Expressions* [11] which I believe to be a great source for my thesis topic. The main idea of this article is that non-spontaneous induced movements of the face use different pathways than those used for emotionally induced movements (spontaneous). Biologically, impulses for non-spontaneous movements start from the cortical motor strip and arrive at the facial nucleus (a collection of neurons in the brain stem that belong to the facial nerve; cranial nerve VII) through the pyramidal tract (the upper motor neurons that originate in the cerebral cortex and terminate in the spinal cord).

On the other hand, impulses for spontaneous emotional facial movements arise from the extrapyramidal motor system, which is not actually a unitary system but a group of highly interactive neural circuits, each of which contributes its own specialized influences to the final motor response. (Although neuroanatomists generally include some cells of the frontal and prefrontal cortex in the extrapyramidal system, the system involves mostly subcortical nuclei, and its influences are conveyed to the facial nucleus through pathways other than the pyramidal tract.)

Four main proofs

First line of evidence An observation of patients with lesions of the cortical motor strip gives us great proof that the spontaneous and non-spontaneous facial expressions come indeed from different pathways. These patients' faces are what is called hemiparalyzed, namely they cannot smile symmetrically to command; they can usually pull one corner of the lips but not both. However surprisingly, the same patients are commonly seen to smile bilaterally when something strikes them as amusing. It's important to note that the muscles resulting in a both smiles (commanded and spontaneous) are the same in both cases. Those muscles are considered paralyzed on one side for voluntary control but in the case of the the non-voluntary smile it is typically just as pronounced on the paralyzed side as on the non paralyzed side. Often, in fact, the smile is exaggerated on the paralyzed sideprobably because of an absence of normal cortical inhibitory influences [3] (See figure 3.1 below).

Figure 3.1: Patient with cortical motor strip lesions expressing voluntary vs spontaneous smile (on the right)

Second line of evidence A second line of evidence comes from patients with mimetic facial paralysis which is a condition in which the patient is able to move facial muscles to verbal commands but is not able to have any spontaneous emotional facial movements. This paralysis appears in patients with Parkinson's disease (a neurological disorder affecting the neurotransmitter systems of the basal ganglia). In particular in the Masked Face syndrome of Parkinsonism, the patient shows a marked diminution of expressive gestures of the face, including brow movements that accompany speech, and emotional facial expression. This condition is also seen, in patients with strokes, tumors, and traumatic lesions of the basal ganglia. The Masked Face is commonly produced on only one side of the face and is a neurological motor disorder. This sparing of volitional movements in mimetic facial paralysis also suggests separate motor systems for spontaneous and non spontaneous expressions of emotions. See figure 3.2

Third line of evidence A third line of evidence comes from a surgical procedure called facial nerve anastomosis or facial reanimation. An operation that restores some of the facial movements that are paralyzed due to a lesion of the facial nerve at some point prior to its emergence onto the face. In this procedure, the motor root of the facial nerve is surgically severed. A few fibers

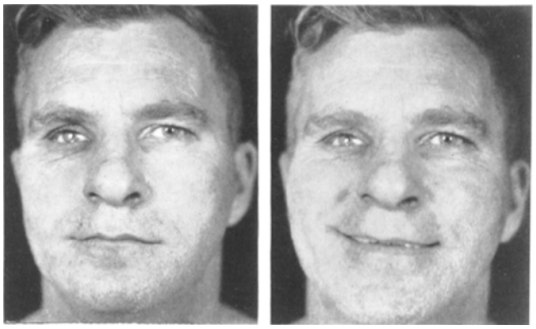

Figure 3.2: Patient with Parkinson Masked Face syndrome spontaneous smile (on the right) vs voluntary smile

are teased out of another cranial motor nerve, usually the spinal accessory nerve (which supplies the muscles that move the shoulder). These fibers are then spliced into the distal stump of the facial nerve so that impulses coursing through the spinal accessory nerve will now innervate the facial muscles as well as the muscles of the shoulder [11]. After the surgery the patient gradually takes control over his facial muscles and is able to experience a good range of voluntary facial movements. However, the patient will still be unable to experience spontaneous reactions on the paralyzed side of their face (despite them being able to voluntarily move that side of their face). Hence they can only experience genuine smiles on the non paralyzed side of the face. Reportedly, they frequently find this embarrassing and avoid such expressions.The most likely explanation for the absence of emotional movements on the affected side is that the motor centers for emotional movements continue to send their impulses to the now disconnected stump of the facial nerve. The behavioral plasticity of the cerebral cortex allows it to learn to use the new pathway through the cortical shoulder representation and the spinal accessory nerve. The more primitive motor centers for emotional movement do not have this degree of flexibility. [11]

Figure 3.3: Before and after dynamic smile reconstruction in facial paralysis (Wikipedia)

Fourth line of evidence A fourth line of evidence comes from observations of non-emotional involuntary laughing and/or weeping often seen in cases of pseudobulbar palsy. Pseudobulbar

palsy is a medical condition characterized by the inability to control facial movements (such as chewing and speaking) and is caused by a variety of neurological disorders. Patients experience difficulty chewing and swallowing, have increased reflexes and spasticity in tongue and the bulbar region, and demonstrate slurred speech (which is often the initial presentation of the disorder), sometimes also demonstrating uncontrolled emotional outbursts [Wikipedia]. Biologically, Pseudobulbar palsy results from lesions of the corticobulbar pathways. About 50% of patients of palsy get outbursts of crying or laughing sometimes with no apparent provocation at all. Once triggered, these emotions are uncontrollable till they fade on their own. These situations are usually hard to distinguish from real crying or laughing from spontaneous emotional responses (including respiratory and vocal responses). The difference however is that the patients report not experiencing the emotions displayed. Sometimes, they would even be experiencing contradictory emotions (laughing while angry or sad).

Patients with pseudobulbar palsy typically also have at least some degree of voluntary facial paralysis. Apparently, since these patients have a paralysis on the voluntary abilities, it seems like they have no control over inhibiting these outbursts. Thus, a double dissociation between voluntary and emotional facial movements can be demonstrated by the fact that either can be interrupted or disturbed by neurological damage while the other remains intact. [1]

According to evidence put forth, spontaneous and voluntary displays of emotion appear independently on the face. Therefore, if they have contrasting characteristics, **fake emotions should be detectable**

3.3 Questions

- Are the emotions displayed spontaneous
- Are the emotions displayed acted
- Are all emotional displays controllable by humans
- Are there unconscious displays of emotion that reveal more about what a person is experiencing than what they are showing

3.4 Paul Ekman's findings

Dr Paul Ekman is an American psychologist and professor emeritus at the University of California, San Francisco who is a pioneer in the study of emotions and their relation to facial expressions. He has created an "atlas of emotions" with more than ten thousand facial expressions, and has gained a reputation as the best human lie detector in the world. He was ranked 59th out of the 100 most cited psychologists of the twentieth century. Ekman conducted seminal research on the specific biological correlations of specific emotions, demonstrating the universality and

[1]Note: I gathered this information from readings that I have done from the internet (Wikipedia and other resources)

Figure 3.4: Spontaneous (on the left) vs voluntary emotion

discreteness of emotions in a Darwinian approach[18]

Dr. Paul Ekman, developed a system called the Facial Action Coding System. FACS is a model that analyzes facial expressions to measure emotions. It encodes the movements of the facial muscles and changes in their patterns (the contraction or relaxation of facial muscles). Trained FACS experts can see through the difference between an insincere and voluntary smile (Pan-AM) and a sincere and involuntary smile (Duchenne). This is a more accurate way to determine and measure a persons emotion, for which a system of facial action recognition has to be implemented.

3.5 From Action Units to emotion

FACS provides us with a manual that maps different sets of Action Units to emotions. An example of that is seen in 3.7.

3.6 What do we look for

- If we define a facial expression of emotion as a simultaneous contraction of a set of muscles (Action Units), to fake the emotion one has to fake all the muscle contractions of that emotion simultaneously
- Even in the case where a person is able to perfectly fake their emotions, micro-emotions will appear at the surface at some point, which are basically a set of muscular contractions on the face that do not match the fake emotion that is being attempted to be portrayed.

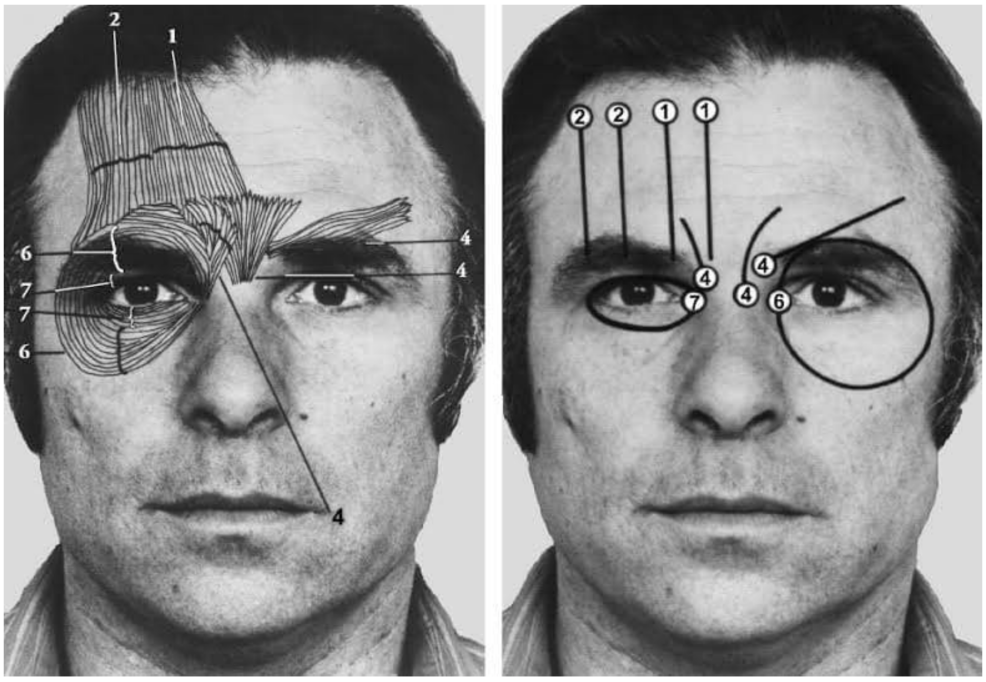

Figure 3.5: FACS decomposes facial movements into component actions. Sample Action Units are illustrated by numbers on the upper facial muscles.

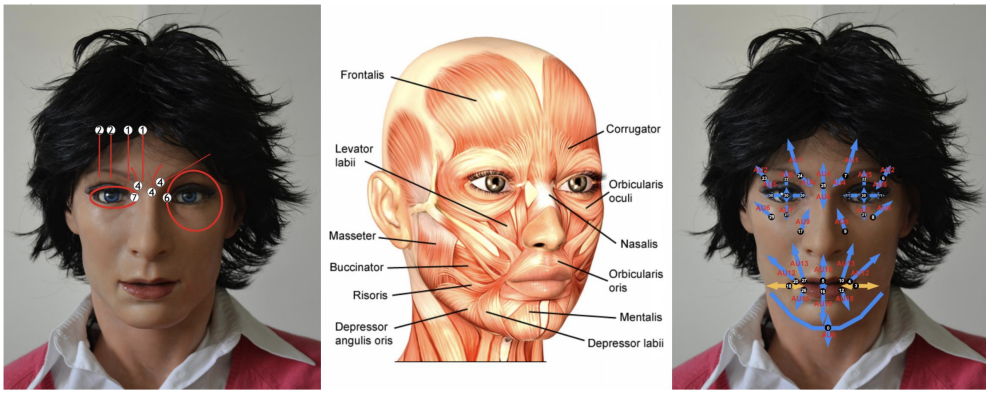

Figure 3.6: Left: Areas of contraction as seen on the face. Middle: the underlying muscles contributing to this apparent muscle contraction. Right: The Action Unit names associated to these facial muscle contractions

Category	AUs	Category	AUs
Happy	12, 25	Sadly disgusted	4, 10
Sad	4, 15	Fearfully angry	4, 20, 25
Fearful	1, 4, 20, 25	Fearfully surprised	1, 2, 5, 20, 25
Angry	4, 7, 24	Fearfully disgusted	1, 4, 10, 20, 25
Surprised	1, 2 , 25, 26	Angrily surprised	4, 25, 26
Disgusted	9, 10, 17	Disgusted surprised	1, 2, 5, 10
Happily sad	4, 6, 12, 25	Happily fearful	1, 2, 12, 25, 26
Happily surprised	1, 2, 12, 25	Angrily disgusted	4, 10, 17
Happily disgusted	10, 12, 25	Awed	1, 2, 5, 25
Sadly fearful	1, 4, 15, 25	Appalled	4, 9, 10
Sadly angry	4, 7, 15	Hatred	4, 7, 10
Sadly surprised	1, 4, 25, 26	-	-

Figure 3.7: Action Units mapping to emotions from FACS manual

Chapter 4

System

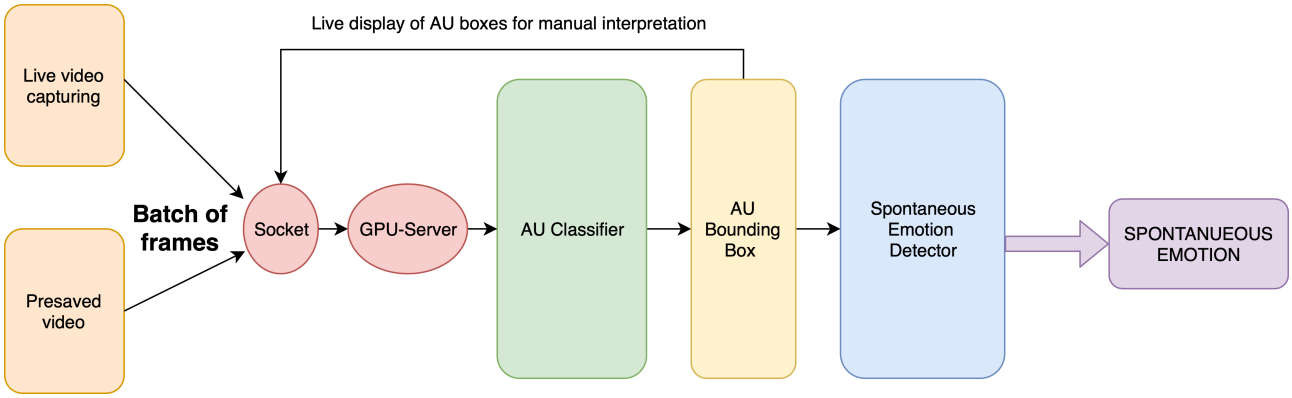

Figure 4.1: Sample video frames from EB+ [6]

We developed an end to end system 4.1 that when fed batches of frames, returns the emotion that is experienced by the participant.

1. For data acquisition we run our system either on a prerecorded video or live video from a webcam with the condition of having the participant's face centered in the video frame.
2. For the sake of our experiments, we transfer the data through a websocket one batch at a time to the GPU-server.
3. At the server side, the data passes through the AU classifier which is one of our pretrained models that outputs the Action Units that are present in the video frames.
4. Bounding boxes are drawn on the person's face where the AU is present.
5. At this point, we feed the Action Units that have been produced to a Spontaneous Emotion Classifier that outputs one of 9 different emotions.

Chapter 5

EB+ Dataset [6]

Figure 5.1: Sample video frames from EB+

The EB+ [6] (Expanded BP4D+) is currently one of the largest and most challenging benchmarks available for the task of Action Unit detection. EB+ [6] contains 2D and 3D videos of spontaneous facial expressions in young people. All of them were recorded in non-acted scenarios, as the emotions were always elicited by experts. This dataset contains around 360k individual

frames from 200 subjects annotated by trained psychologists.

5.1 Data collection

In this dataset, psychologists designed tasks for authentic emotion induction. These tasks include social interviews between a naive subject and a professional actor/director, planned activities (e.g., games), film clip watching, a cold pressor test for pain elicitation, a social challenge to elicit anger followed by reparation, and olfactory stimulation to elicit disgust.

The emotions induced in these tasks are: Happiness, Sadness, Surprise, Skepticism, Embarrassment, Fear/Nervousness, Pain, Anger, Disgust.

Every participant has 9 videos, one per experiment. Every video had two experienced FACS-certified coders independently code onsets and offsets of 27 action units per the 2002 edition of FACS. Onsets are moments where the participants start showing signs of emotional reaction to the situation, followed by a peak moment where the emotion is very clear on their faces, followed by the emotion fading out until it is no longer visible. The disappearance moment is called an offset. The duration of an onset-offset varies between 10-15 seconds per video.

We have around 1165 videos, with every video containing around 260-300 labeled video frames.

5.2 Data annotation

Well-experienced, certified FACS coders annotated the videos. This serves as the frame-level ground-truth for facial actions.

5.3 Annotation verification

The effectiveness of the eliciting methods have been verified by subject self-report and FACS analysis. Also, an alternative subjective evaluation and validation was conducted by human observer ratings.

5.4 The product

A set of meta-data, including AU codes, emotions, tracked 3D/2D features, and head poses is provided. Videos were taken at a rate of 25-30 fps at a resolution of 1092x768.

Chapter 6

Action Unit Classifier

6.1 Data preprocessing

Out of the 370k frames, we divide the dataset into half training, quarter evaluation and quarter testing. Each frame has been resized to 224x224 pixels.

6.2 Chosen evaluation metrics

mAP: Mean Average Precision Every frame can contain multiple AUs at the same time, hence we have a multiclass classification problem where the distribution of these classes is not necessarily uniform. Therefore, a simple accuracy-based metric will introduce biases. Thus, there is a need to associate a confidence score with each AU detected and to assess the model at various levels of confidence.

To calculate the AP, for a specific class, we compute the area under the precision recall curve, which is the precision in function of the recall, then divide by the number of instances to get the Average Precision. Lastly we average the AP value for each class [17].

Loss Train and test losses are logged. We use their corresponding curves to determine whether the loss is decreasing smoothly, whether overfitting has occurred, and whether additional training is required. If any of these cases apply, the parameters of the network are tweaked. The loss used for this task is Binary Cross Entropy loss 6.1.

$$\ell_c(x, y) = L_c = \{l_{1,c}, \dots, l_{N,c}\}^\top, \quad l_{n,c} = -w_{n,c} \left[p_c y_{n,c} \cdot \log \sigma(x_{n,c}) + (1 - y_{n,c}) \cdot \log(1 - \sigma(\right.$$

where c is the class number ($c > 1$ for multi-label binary classification, $c = 1$ for single-label binary classification), n is the number of the sample in the batch and p_c is the weight of the positive answer for the class c.

Figure 6.1: Binary Cross Entropy loss for binary multilabel classification task.[9]

Recall To justify the use of this metric, let us take a real life scenario as an example. A criminal is being questioned for a crime during a one hour video recording. Assuming the video is recorded at 60 frames per second, we have about 216k frames to analyze. Assume our model detects fear or lying in 100 of those frames, but in reality only 60 of those actually contain fear or lying. This means that in this case we have a recall of 100% but a precision of 60%. The redundant frames could be reanalyzed and dismissed by a psychologist with no high stakes. So in our scenario it makes sense to denote the recall and make sure to get a high recall value. Combined with other metrics, we can properly evaluate the learning of our model.

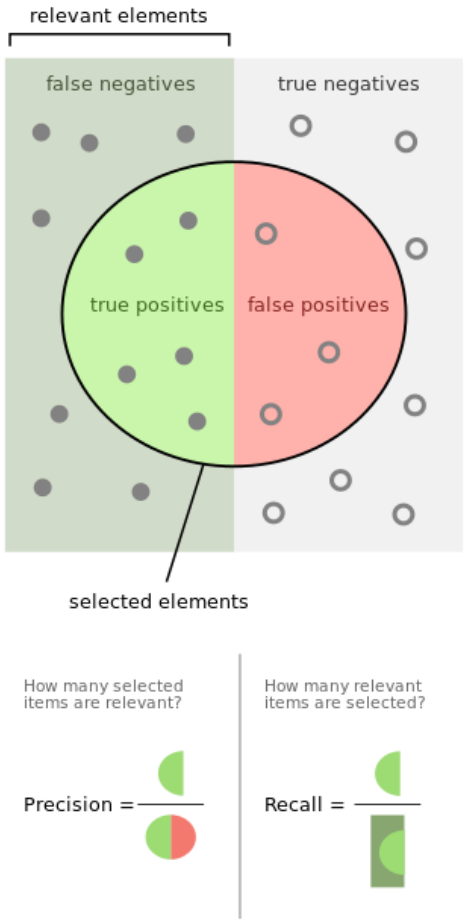

Figure 6.2: Visual explanation of recall, one of our evaluation metrics for the Action Unit classifier[16]

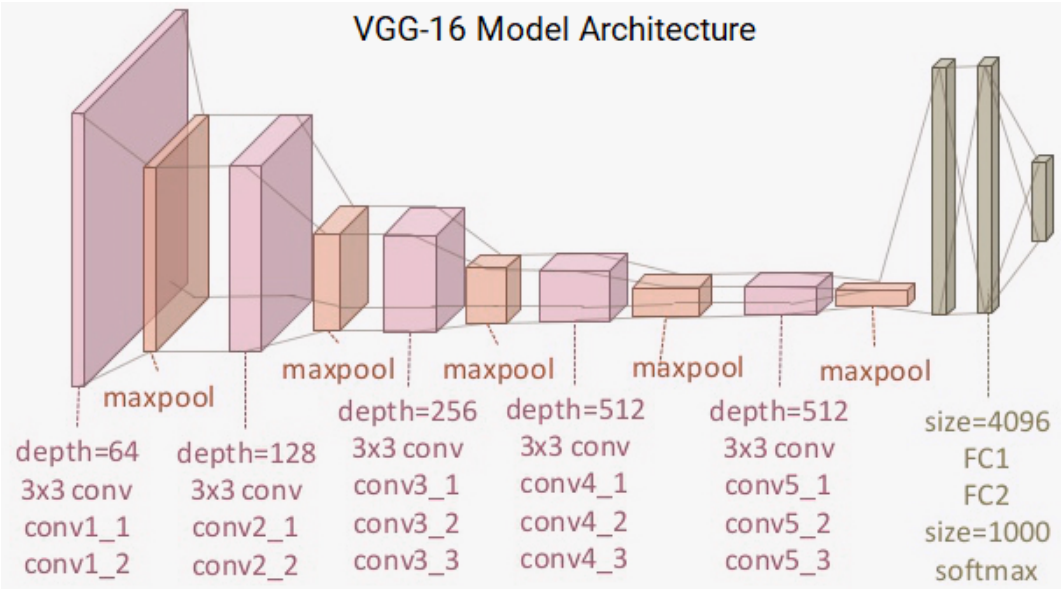

Figure 6.3: VGG16 architecture [2]

6.3 VGG16 (Baseline)

6.3.1 Why VGG16

VGG16 [13] was invented by the Visual Geometry Group at University of Oxford. It is a network that won the ImageNet [5] challenge in 2014. The model has been used to solve multiple problems including face recognition. In 2015 the Visual Geometry group trained VGG16 on 2.6 million faces for face recognition task. The weights of the trained model have been released to the public. In [13], the authors of the paper proposed a system that directly analyzes information on a whole human face in order to predict the presence of specific action units. This method bypasses landmark localization on the standard benchmarks for this task, and therefore predicts the presence or absence of a specific AU in a single face image as holistic binary classification. This was one of the state of the art papers on the task of Action Unit classification. Hence, we used the same architecture on our dataset. We used VGG-Face with its trained weights and fine tuned it on our EB+ Dataset [6]. At every time step, we sample a batch of random frames and try to identify the existing Action Units. Below are the training and evaluation loss, mAP and recall curves:

6.3.2 Summary

Both training and evaluation losses drop and stabilize pretty quickly, with the same happening to the recall curve as it stabilizes between 37% and 40%. The mAP values stabilize at around 0.42 and 0.48. Training was conducted for about 20 epochs. All of our curves show a quick initial learning that is due to the fact that the network was pre-trained thoroughly on faces. Followed by an early stabilization, due to the fact that the network architecture is not necessarily a suitable

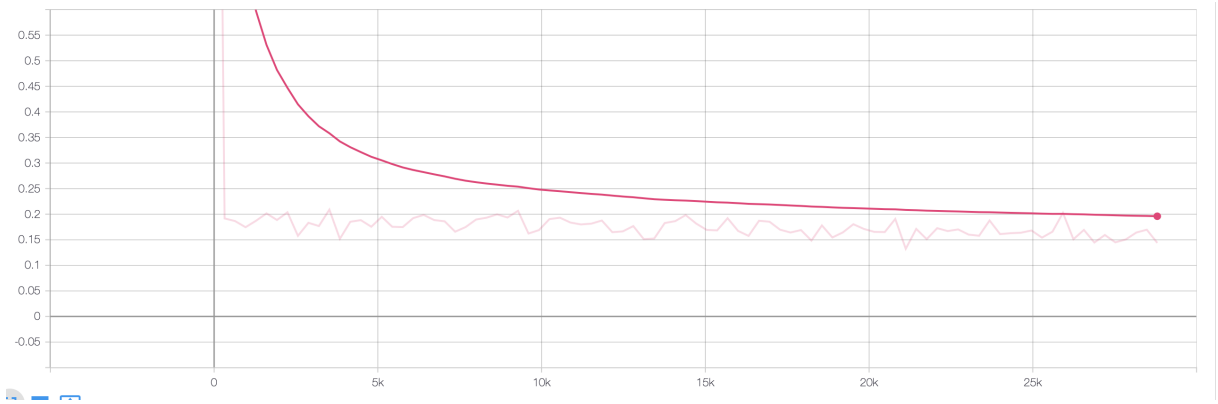

Figure 6.4: Train Loss curve: This curve shows a quick initial drop due to the fact that the model has been trained on faces before, followed by an early stabilization.

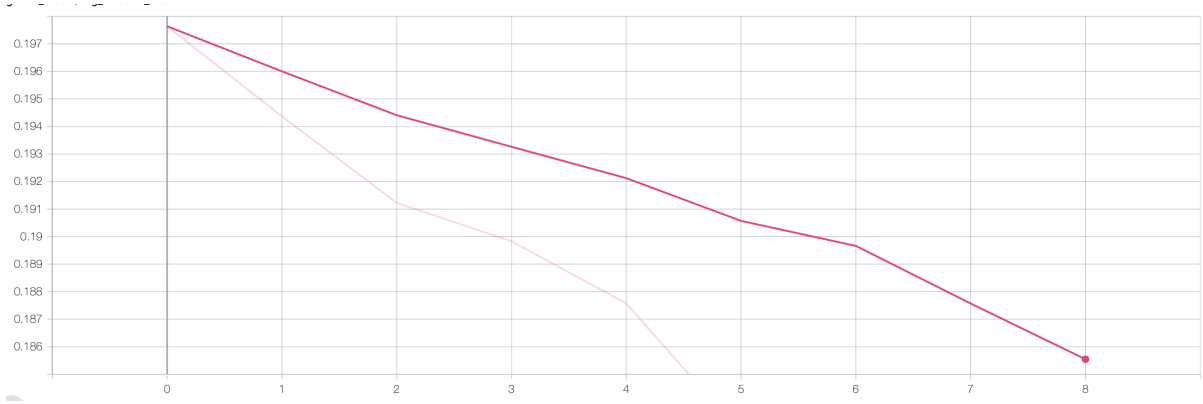

Figure 6.5: Evaluation Loss curve: This curve shows a quick initial drop (just like the train loss curve) that is followed by an early stabilization.

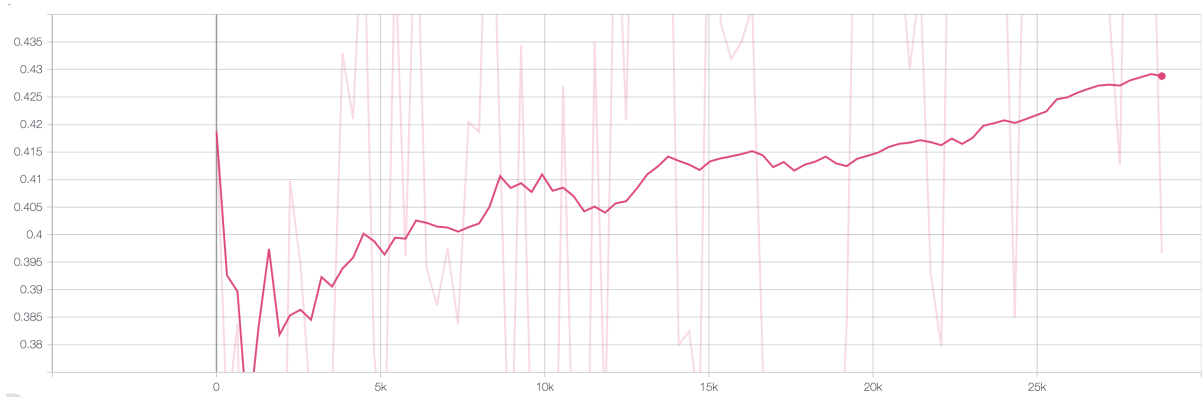

Figure 6.6: Train Recall curve: Recall values have had a slow improvement during training reaching a value of 0.43.

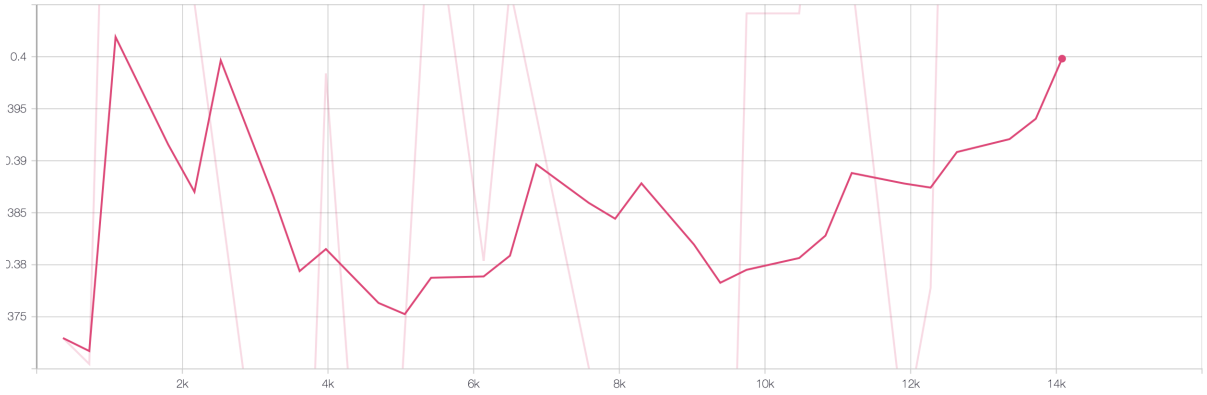

Figure 6.7: Evaluation Recall curve: Evaluation recall fluctuated between 0.37 and 0.4.

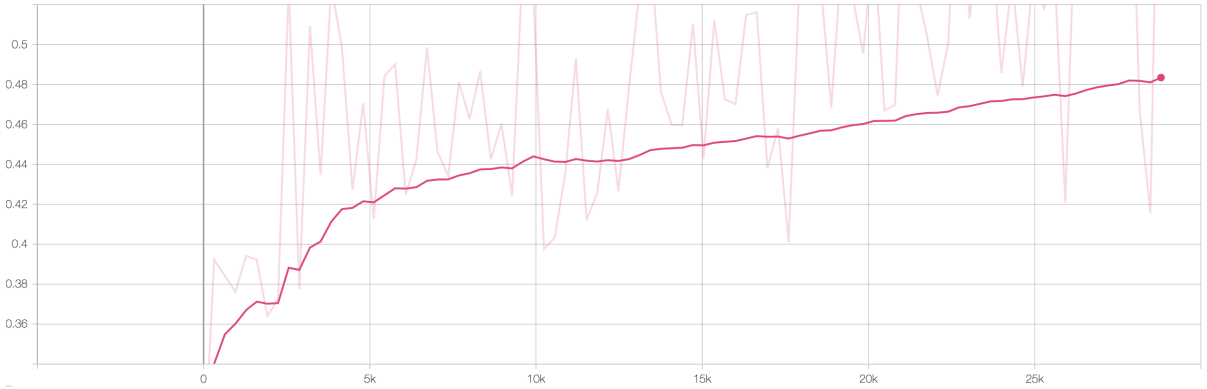

Figure 6.8: Train mAP curve: This curve improved slowly but steadily until reaching a value of 0.48.

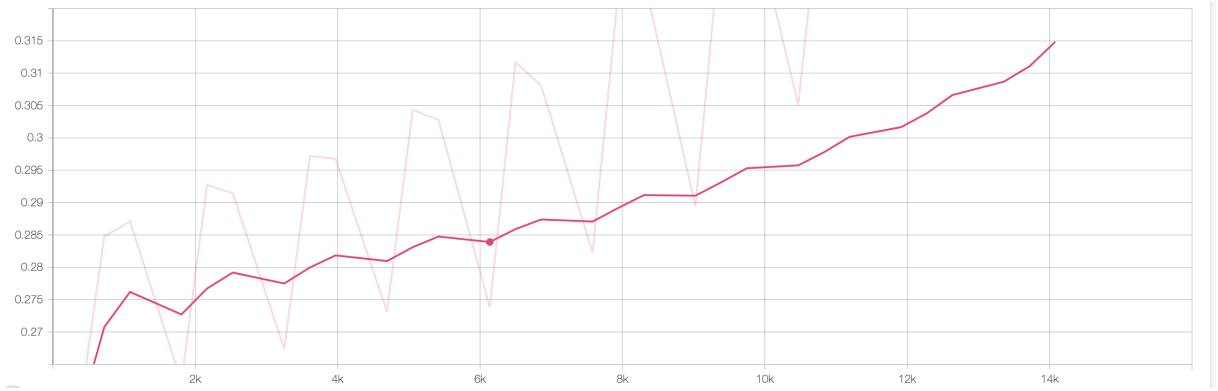

Figure 6.9: Evaluation mAP curve: This curve also improved steadily until reaching a little above 0.3 mAP.

architecture for this task.

6.3.3 VGG Face Conclusions

- Emotions are dynamic temporal patterns
- We need to capture spatial temporal features, these features are three dimensional
- State of the art techniques for this type of data are 3D convolutions
- However, the number of parameters in 3D convolutions grows by a factor of O(n^3)
- Long-term temporal features require larger kernels
- The computational requirement for training and inference are not scalable
- Hence we used depth-wise convolutions

6.4 3D Convolutions-MobileNet

6.4.1 MobileNet

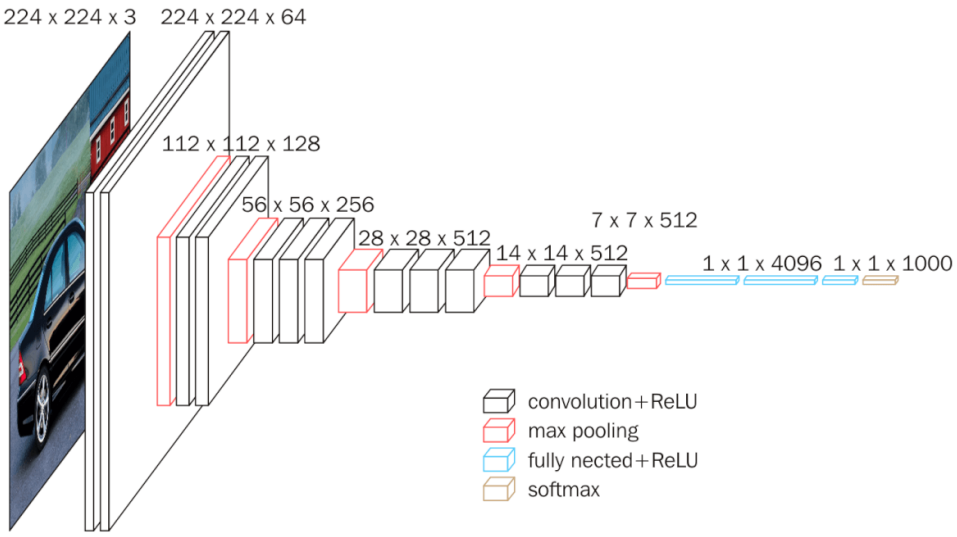

Figure 6.10: MobileNet architecture [1]

In this work, we re-implemented a variant of MobileNet [1] which is a light wight network developed by Google that performs 2D depthwise convolutions on images using Depthwise-separable convolutions. We adapted this architecture to our problem to incorporate the spacio-temporal features in 3D depthwise-separable convolutions as opposed to just spacial features.

6.4.2 Depthwise separable convolutions:

The depthwise separable convolution (as explained in [7]) consists of a depthwise-convolution, i.e. a spatial convolution performed independently over every channel of an input, followed by

a pointwise convolution, i.e. a regular convolution with 1x1 windows, projecting the channels computed by the depthwise convolution onto a new channel space.

$$\begin{aligned}
\text{PointwiseConv}(W, y)_{(i,j)} &= \sum_{m}^{M} W_m \cdot y_{(i,j,m)} \\
\text{DepthwiseConv}(W, y)_{(i,j)} &= \sum_{k,l}^{K,L} W_{(k,l)} \odot y_{(i+k,j+l)} \\
\text{SepConv}(W_p, W_d, y)_{(i,j)} &= \text{PointwiseConv}_{(i,j)}(W_p, \text{DepthwiseConv}_{(i,j)}(W_d, y))
\end{aligned}$$

Figure 6.11: Equations of Depthwise separable convolutions as explained in [7] note the $\odot$ represents element-wise product.

6.4.3 Normal vs Depthwise convolutions

A standard convolution both filters and combines inputs into a new set of outputs in one step. Depthwise separable convolutions split this into two layers, a separate layer for filtering and a separate layer for combining. This factorization has the effect of drastically reducing computation and model size.

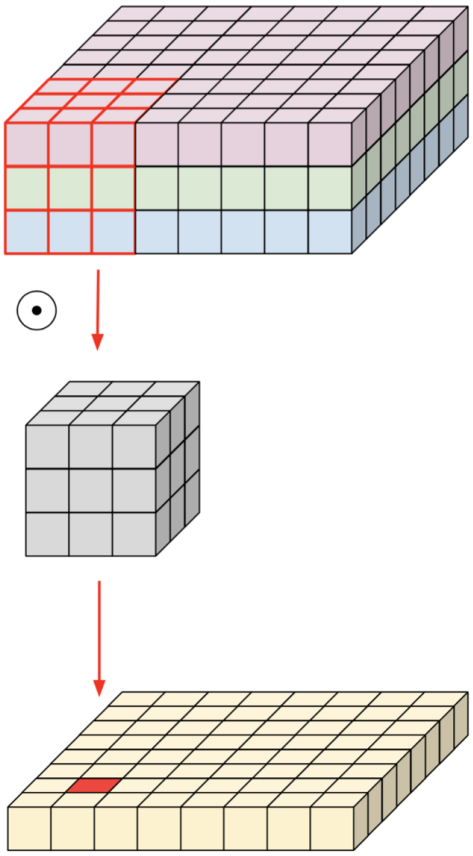

Figure 6.12: Normal Convolutions: Filters and combines input into a new output in one step

6.4.4 Network parameters

As previously mentioned, we have re-implemented the MobileNet architecture (V2 [12]) and adapted it to our problem by incorporating 3D Depthwise separable convolutions. We used Binary Cross Entropy loss and Adam Optimizer with a learning rate starting at 0.05 with Cosine Learning Rate Scheduler.

6.4.5 Evaluation

Note: The curves reported below are for the first 3 days of training. We interrupted the training and resumed it for about 3 more days.

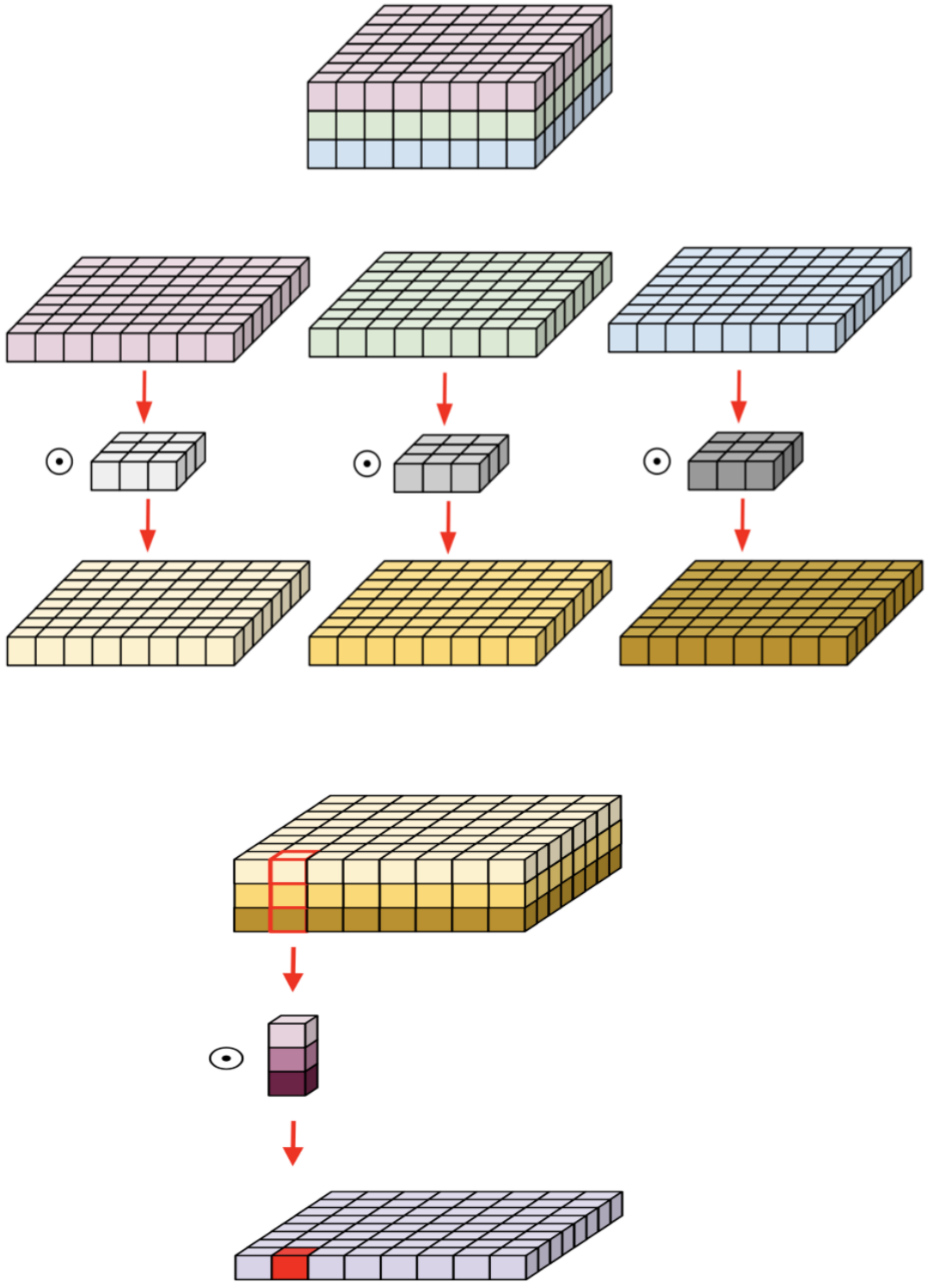

Figure 6.13: Depth Convolutions: Filters separately then combines filtered outputs into a new merged output

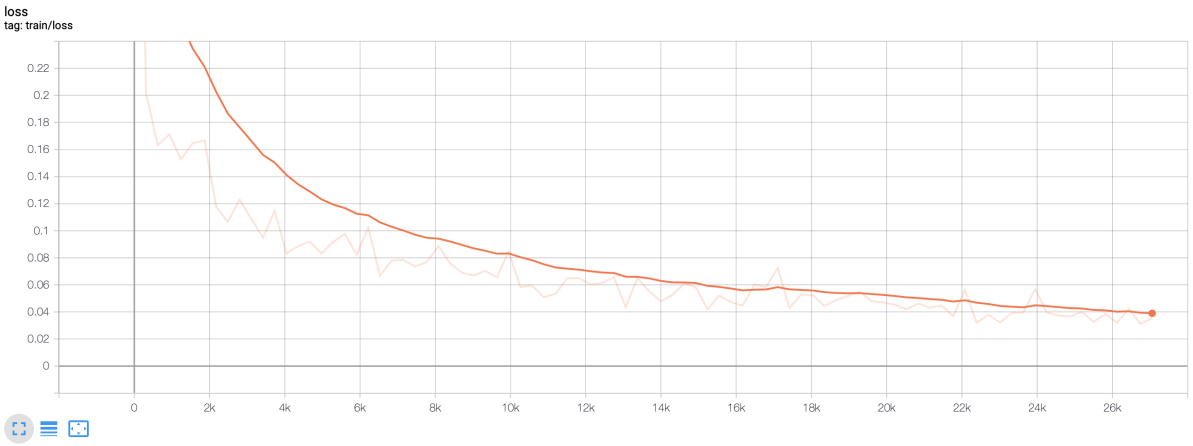

Figure 6.14: Train Loss Curve. Loss went down smoothly as training progressed

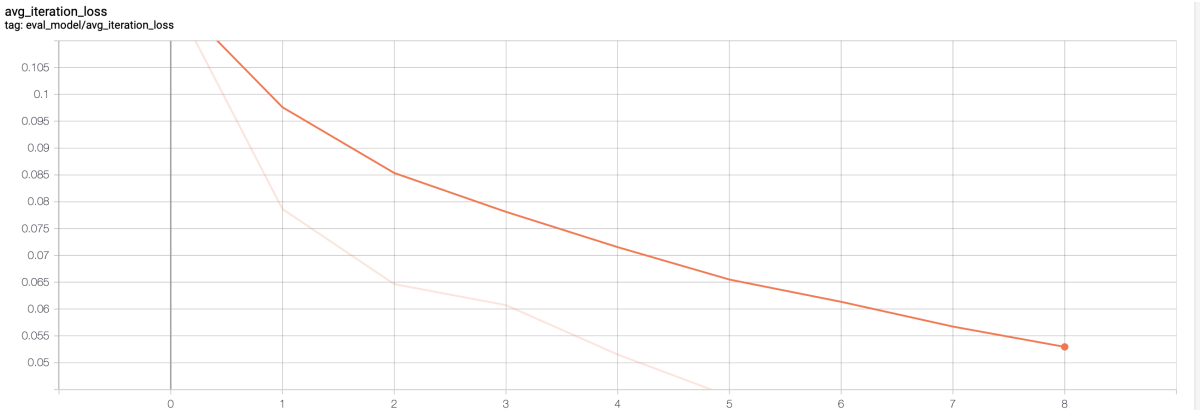

Figure 6.15: Evaluation Loss Curve. Just like training, loss went down smoothly as training progressed

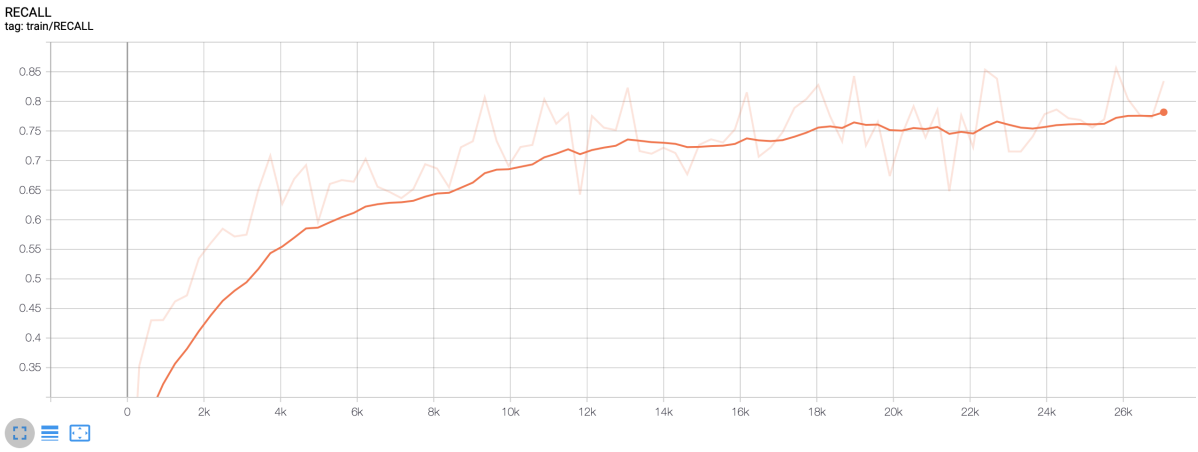

Figure 6.16: Train recall Curve. Recall increased steadily until reaching around 0.79

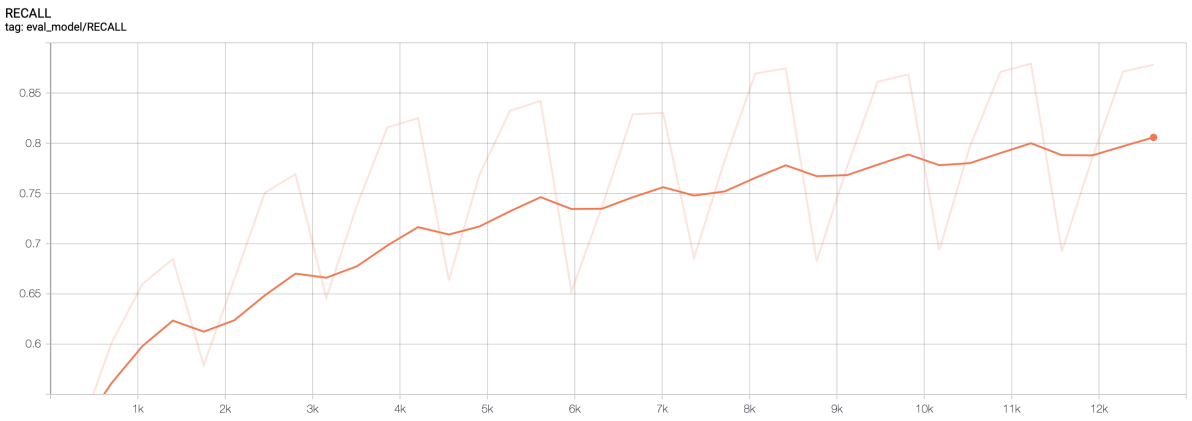

Figure 6.17: Evaluation recall Curve. Recall curves show an increase until reaching 0.81-0.84

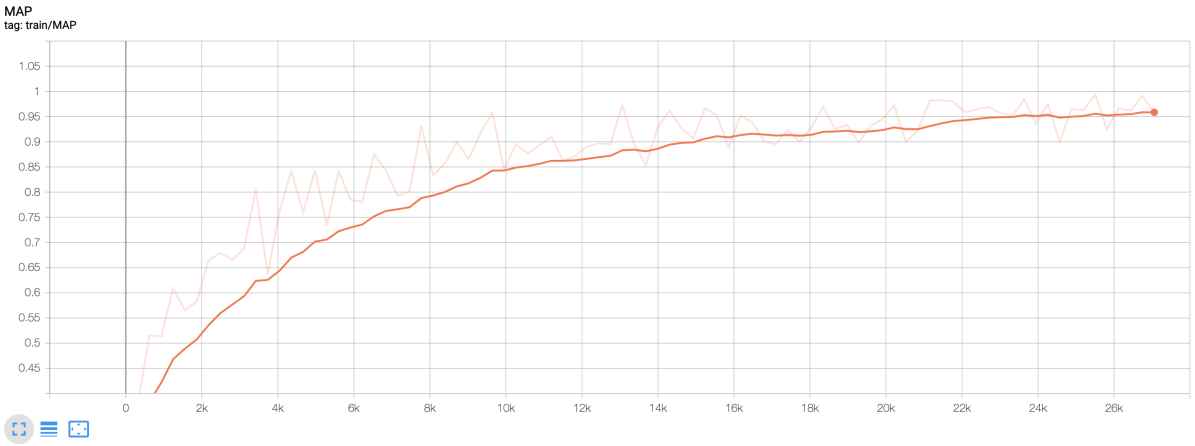

Figure 6.18: Train mAP Curve. Increased smoothly as training progressed

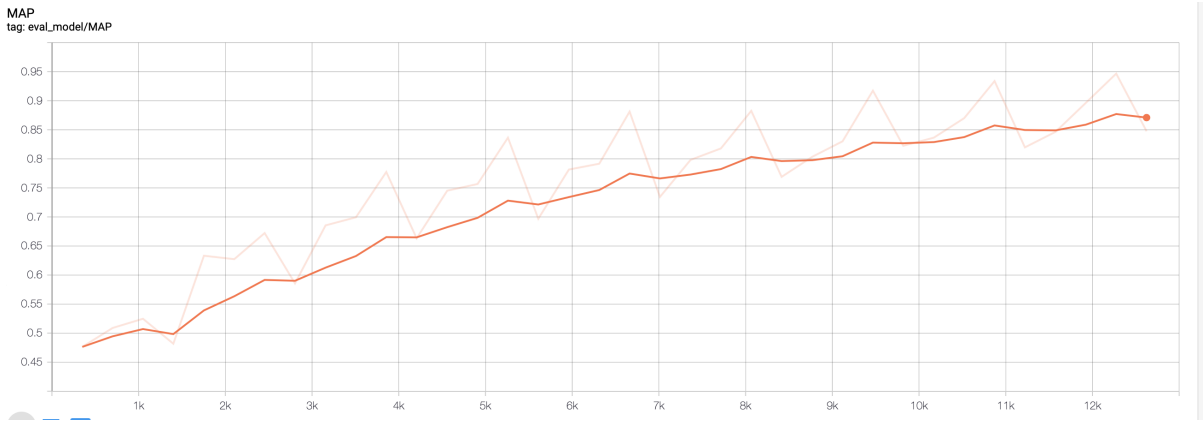

Figure 6.19: Evaluation mAP Curve. Increased smoothly until reaching 0.91

AU	FACS Name	Average Precision	Accuracy	OCC %	Muscular Basis
1	Inner brow raiser	0.9772	0.9820	11.9	frontalis (pars medialis)
2	Outer brow raiser	0.9737	0.9855	8.6	frontalis (pars lateralis)
4	Brow lowerer	0.9752	0.9849	8.8	depressor glabellae, depressor supercilii, corrugator supercilii
5	Upper lid raiser	0.8968	0.9973	0.7	levator palpebrae superioris, superior tarsal muscle
6	Cheek raiser	0.9978	0.9764	50.7	orbicularis oculi (pars orbitalis)
7	Lid tightener	0.9988	0.9784	67.4	orbicularis oculi (pars palpebralis)
9	Nose wrinkler	0.9630	0.9917	4.0	orbicularis oris
10	Upper lip raiser	0.9990	0.9811	64.7	levator labii superioris alaeque nasi
11	Nasolabial deepener	0.9967	0.9785	38.0	levator labii superioris, caput infraorbitalis
12	Lip corner puller	0.9984	0.9793	55.5	zygomaticus minor
13	Cheek puller	0.8164	0.9994	0.1	zygomaticus major
14	Dimpler	0.9974	0.9693	60.3	levator anguli oris (also known as caninus)
15	Lip corner depressor	0.9625	0.9725	13.0	buccinator
16	Lower lip depressor	0.9748	0.9484	27.7	depressor anguli oris (also known as triangularis)
17	Chin raiser	0.9564	0.9634	16.0	depressor labii inferioris
18	Lip pucker	0.7794	0.9973	0.5	mentalis
19	Tongue show	0.7163	0.9963	0.6	incisivii labii superioris and incisivii labii inferioris
20	Lip stretcher	0.9815	0.9754	17.0	
22	Lip funneler	0.8007	0.9914	1.7	risorius w/ platysma
23	Lip tightener	0.9481	0.9493	19.9	platysma
24	Lip pressor	0.9494	0.9882	4.4	orbicularis oris
27	Mouth stretch	0.8888	0.9993	0.2	orbicularis oris
28	Lip suck	0.9631	0.9957	1.9	orbicularis oris
30	Jaw sideways	0.3828	0.9996	0.0	depressor labii inferioris, or relaxation of mentalis or orbicularis oris
32	Bite	0.9668	0.9956	2.0	masseter; relaxed temporalis and internal pterygoid
38	Nostril dilator	0.9043	0.9991	0.2	pterygoids, digastric
39	Nostril compressor	0.9647	0.9981	1.0	orbicularis oris

Figure 6.20: Evaluation and interpretation of AUs

6.4.6 Comparison

In table 6.20, we report the Average Precision per Action Unit, the Accuracy per AU and the number of times we perceived that AU in the dataset.

AU	FACS Name	verage Precisic	Accuracy	ACC-P	OCC %	Muscular Basis
1	Inner brow raiser	0.9772	0.9820	0.765	11.9	frontalis (pars medialis)
4	Brow lowerer	0.9752	0.9849	0.857	8.8	depressor glabellae, depressor supercilii, corrugator supercilii
6	Cheek raiser	0.9978	0.9764	0.794	50.7	orbicularis oculi (pars orbitalis)
7	Lid tightener	0.9988	0.9784	0.763	67.4	orbicularis oculi (pars palpebralis)
10	Upper lip raiser	0.9990	0.9811	0.832	64.7	levator labii superioris alaeque nasi
12	Lip corner puller	0.9984	0.9793	0.829	55.5	zygomaticus minor
14	Dimpler	0.9974	0.9693	0.683	60.3	levator anguli oris (also known as caninus)
15	Lip corner depressor	0.9625	0.9725	0.736	13.0	buccinator
17	Chin raiser	0.9564	0.9634	0.763	16.0	depressor labii inferioris
23	Lip tightener	0.9481	0.9493	0.754	19.9	platysma

Figure 6.21: Evaluation and interpretation of AUs

In 6.21 we put side by side the results from View-Independent Facial Action [14] Unit Detection paper and our results. We have successfully scored higher Accuracies across all of the AUs reported in [14].

Chapter 7

Emotion Classifier

7.1 FACS manual

Emotion	Criteria
Angry	AU23 and AU24 must be present in the AU combination
Disgust	Either AU9 or AU10 must be present
Fear	AU combination of AU1+2+4 must be present, unless AU5 is of intensity E then AU4 can be absent
Happy	AU12 must be present
Sadness	Either AU1+4+15 or 11 must be present. An exception is AU6+15
Surprise	Either AU1+2 or 5 must be present and the intensity of AU5 must not be stronger than B
Contempt	AU14 must be present (either unilateral or bilateral)

Figure 7.1: Emotion description in terms of Action Units by FACS Manual

When we first started this work, we planned on using Dr. Paul Ekman's FACS manual 7.1 to map AU sets to discrete emotions. However recent articles have been published illustrating results from various psychology papers regarding FACS manual translating like [15]. These articles claim that peoples' faces are very different and it is not possible to develop a mapping that works for all sorts of faces. As such, we arrived at the conclusion that if a mapping is to be learned, there must exist a machine learning solution that is capable of learning the mapping better and faster than human beings.

7.2 Automatic Action Unit detection

We have tried multiple machine learning and deep learning techniques including:

- LSTMs
- SVMs
- Naive Bayes

- MLPs
- ..etc

	PE Manual	EB+ participants
Happiness	12	5+9+10 | 7+8+11+14+19
Sadness	1+4+15 | 6+15	5+11+16+19 | 11+13 | 14+ 19
Embarrasment	12 | 24	0 + 5 + 22
Fear	1 + 2 + 4 | 1 + 2 + 5	14
Anger	23+24	16 + 19 +20
Disgust	9+10	9 + 12 + 15 + 19

Figure 7.2: Emotion description in terms of Action Units

Unfortunately none of these classifiers were able to exceed 50% accuracy on the test set. We looked more closely at the mappings between the emotions and AUs. The table 7.2 shows that the set of action units appearing per emotion does not necessarily follow the manual, confirming the findings of the recent psychology papers mentioned above.

In table 7.3 we notice that almost all of the Action Units appear in almost all of the emotions with different percentages, confirming again the fact that there are no emotion specific AUs when we're looking at different people.

Our solution would be to construct a table for mapping a highly co-occurring set of AUs to emotions (utilizing a basic thresholding technique) then a psychological exam is required to normalize/fine-tune the table. Once the mapping is fine-tuned the system runs autonomously end to end.

	Happiness	Sadness	Surprise	Embarrassment	Fear/Nervous	Pain	Angry	Disgust
1	21.11	0.0	12.39	30.47	34.85	19.35	22.49	22.84
2	20.52	0.0	3.67	25.67	31.75	10.13	18.42	11.21
4	8.58	26.09	28.44	12.41	17.99	36.47	6.22	31.33
5	0.5	4.35	7.8	4.35	6.05	0.71	0.0	0.16
6	50.97	0.0	38.53	50.71	50.94	36.58	33.25	56.31
7	76.45	39.13	57.34	76.16	73.39	58.76	67.46	77.49
9	6.06	0.0	1.38	4.92	9.83	12.06	1.2	16.65
10	74.1	0.0	49.54	77.3	71.73	63.53	54.55	76.67
11	43.4	0.0	21.56	39.74	45.05	33.94	16.03	39.57
12	73.68	0.0	42.66	69.81	67.88	41.84	59.33	57.21
13	1.09	0.0	0.92	1.09	0.23	1.01	0.0	0.08
14	69.13	52.17	29.36	77.82	71.58	63.93	51.91	42.04
15	33.05	0.0	16.06	36.54	33.11	20.36	23.92	39.82
16	38.52	4.35	27.06	44.08	24.64	25.94	39.0	32.89
17	30.28	34.78	33.49	32.53	31.67	23.51	24.88	40.4
18	2.02	0.0	0.92	2.74	0.83	1.62	3.83	2.06
19	5.47	8.7	5.5	4.52	4.38	3.34	4.55	1.32
20	17.33	4.35	8.26	20.75	15.65	14.89	9.57	10.72
22	10.6	0.0	2.29	17.9	6.58	4.56	12.44	9.15
23	43.9	47.83	31.65	46.6	40.06	35.97	50.48	33.55
24	10.43	8.7	5.96	6.35	12.09	17.63	7.89	7.91
27	0.42	0.0	0.92	0.29	0.23	2.23	0.96	0.08
28	9.25	4.35	10.55	5.72	7.94	15.3	6.94	1.65
30	0.42	0.0	0.46	0.23	0.23	0.3	0.48	0.08
32	12.28	0.0	2.29	10.12	8.16	16.31	5.98	2.97
38	0.42	0.0	0.46	0.23	0.3	1.22	0.0	0.08
39	3.03	0.0	2.29	0.63	3.78	6.08	1.67	0.99

Figure 7.3: Emotion description in terms of Action Units

Chapter 8

End to End system

Putting the system together, by passing batches of data frames to the network, passing them through our variant of MobileNet and using our generated table, we reach an accuracy of 74.48% on emotion detection without fine-tuning. We expect this number to become much higher once we're able to personalize the mapping per participant

Chapter 9

Conclusion

In this work:

- We developed a more accurate AU detection system
- We developed a new automated mapping between AU and human emotions
- The mapping is easily fine-tuned and personalized
- Overall, we developed an end to end system for guided spontaneous emotion detection and recognition from video

Bibliography

[1] mobileNet diagram. https://machinethink.net/blog/mobilenet-v2/. Accessed: 2019-08-17. (document), 6.10, 6.4.1

[2] VGG16 diagram. https://hub.packtpub.com/how-to-leverage-transfer-learning-using-pretrained-cnn-models-tutorial/ Accessed: 2019-08-17. (document), 6.3

[3] Brodal, a., neurological anatomy in relation to clinical medicine. third edition. new york, oxford university press, 1981, 1,053 pages, $35.00. *Annals of Neurology*, 10(6):584–584, 1981. 3.2

[4] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F. Cohn. Modeling spatial and temporal cues for multi-label facial action unit detection. *CoRR*, abs/1608.00911, 2016. 2.2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 6.3.1

[6] I. O. Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–8, May 2019. (document), 4.1, 5, 5, 6.3.1

[7] Lukasz Kaiser, Aidan N. Gomez, and François Chollet. Depthwise separable convolutions for neural machine translation. *CoRR*, abs/1706.03059, 2017. (document), 6.4.2, 6.11

[8] MultiMedia LLC. Micro-expressions: More than meets the eye, 2017. 2.1

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. (document), 6.1

[10] G. Patil and P. Suja. Emotion recognition from 3d videos using optical flow method. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 825–829, Aug 2017. 2.1

[11] William Rinn. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 95:52–77, 02 1984. 3.2, 3.2

[12] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. 6.4.4

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6.3.1

[14] C. Tang, W. Zheng, J. Yan, Q. Li, Y. Li, T. Zhang, and Z. Cui. View-independent facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 878–882, May 2017. 2.2, 6.4.6

[15] washington post. Emotion detection ai is a 20 billion industry. new research says it cant do what it claims, 2019. 7.1

[16] wiki. Recall visual explanation, 2019. (document), 6.2

[17] Wikipedia contributors. Evaluation measures (information retrieval) — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Evaluation_measures_(information_retrieval)&oldid=900146701`, 2019. [Online; accessed 17-August-2019]. 6.2

[18] Wikipedia contributors. Paul ekman — Wikipedia, the free encyclopedia, 2019. [Online; accessed 13-August-2019]. 3.4