Syna

emotion recognition based on spatio-temporal machine learning

Shahrokhian, D.

Award date:
2017

Awarding institution:
Royal Institute of Technology

DEGREE PROJECT IN INFORMATION AND COMMUNICATION TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, *SWEDEN 2017*

Syna: Emotion Recognition based on Spatio-Temporal Machine Learning

DANIYAL SHAHROKHIAN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

TRITA TRITA-ICT-EX-2017:139

Declaration

I hereby certify that I have written this thesis independently and have only used the specified sources and resources indicated in the bibliography.

Stockholm, 17. July 2017

. . . . . .

Daniyal Shahrokhian

Abstract

The analysis of emotions in humans is a field that has been studied for centuries. Through the last decade, multiple approaches towards automatic emotion recognition have been developed to tackle the task of making this analysis autonomous. More specifically, facial expressions in the form of Action Units have been considered until now the most efficient way to recognize emotions. In recent years, applying machine learning for this task has shown outstanding improvements in the accuracy of the solutions. Through this technique, the features can now be automatically learned from the training data, instead of relying on expert domain knowledge and hand-crafted rules. In this thesis, I present *Syna* and *DeepSyna*, two models capable of classifying emotional expressions by using both spatial and temporal features. The experimental results demonstrate the effectiveness of Syna in constrained environments, while there is still room for improvement in both constrained and *in-the-wild* settings. *DeepSyna*, while addressing this problem, on the other hand suffers from data scarcity and irrelevant transfer learning, which can be solved by future work.

Keywords
Emotion recognition, spatio-temporal machine learning

Sammanfattning

Mänsklig känsloigenkänning har studerats i århundraden. Det senaste årtiondet har mängder av tillvägagångssätt för automatiska processer studerats, för att möjliggöra autonomi; mer specifikt så har ansiktsuttryck i form av Action Units ansetts vara mest effektiva. Maskininlärning har dock nyligen visat att enorma framsteg är möjliga vad gäller bra lösningar på problemen. Så kallade features kan nu automatiskt läras in från träningsdata, även utan expertkunskap och heuristik. Jag presenterar här Syna och DeepSyna, två modeller för ändamålet som använder både spatiala och temporala features. Experiment demonstrerar Synas effektivitet i vissa begränsade omgivningar, medan mycket lämnas att önska vad gäller generella sådana. DeepSyna löser detta men lider samtidigt av databristproblem och onödig så kallad transfer learning, vilket här lämnas till framtida arbete.

Nyckelord
Känsloigenkänning, spatio-temporal maskininlärning

Acknowledgments

This thesis has been possible with the assistance of both people that I know personally, and people that I have never met. Researchers Abubakrelsedik and Erik helped me with advice on technical aspects of my thesis. Magnus and Daniel arranged supervision during writing. The Open Source community as a whole has given me the opportunity of not having to implement everything from scratch, and focus on what really matters.

Indirectly, there are other persons that helped me up to this point. My mother Azar taught me on the importance of hard work, my old supervisor back in Spain, Sergio, showed me how important it is to be knowledgeable, and my colleague Robin inspired me on being passionate about my work.

I thank every single one of them.

Stockholm, July 2017

Daniyal Shahrokhian

Contents

List of Figures

Chapter 1

1 Introduction

Emotions are an incredible tool in the nature of humans and other organisms. As a mechanism developed through evolution, emotions have played a crucial role in the survival of species. Whether it is fear or laughter, they have different purposes that benefit the organisms on their environment.

Not long ago, emotions were a complete black box to science. There was not much understanding on why humans experienced feelings, and what were their purpose. Their nature was first attributed to divinity, as a tool that extended the consciousness with the sole purpose of honoring and glorifying the Creator.

Today, scientists have developed several theories on how emotions are generated. What first were hypotheses based on external observation in the field of psychology, have later been verified by the study of the brain in neuroscience. If one feels pain, this is due to an environmental change that negatively affects its self-being. In the same manner that organisms developed companionship as a means to increase their chances of survival, emotions mentally prepare an organism to face the changes in the environment that affect it. There is even a relation between these two, and this can be seen through several examples. If someone relative to you dies, you feel pain because that decreases the chances of your survival, since there is one less person to take care of you. If you are in a situation that hazards your survival, the levels of adrenaline increase in order to prepare you for what Cannon [14] called the fight-or-flight response.

As the brain complexity increases, so does the complexity of the individual's behavior. This also involves the aspect of emotions. Human emotions are way more sophisticated than the ones that can be seen in other species, mainly thanks to the development of the Amygdala [15]. Given this high complexity, scientists focused in studying simpler forms. As the neuroscientist Jaak Panksepp once said in one of his talks, *"If we understand the emotions of other animals, then maybe we will be able to understand our own emotions"* [16].

When focusing on humans, there are two main branches of study: the ones that focus on what can be externally seen (physical appearance, social behavior), and the ones that focus on what cannot be externally seen (neuroscience, analysis of

the brain). Neuroscience has played a huge role in understanding the causes and effects of emotions, while the physical appearance focuses solely on a subset of such effects.

1.1 Background

As with many other tasks, humans perform emotion recognition flawlessly. Our brains are fine-tuned machines over millions of years of evolution to perform this function without consciously realizing the process that it involves.

As my personal belief, even though there is a theoretical background on human actions, they do not realize this underlying theory. They just behave in freedom of act, with an internal inertia to apply most laws unconsciously. In contrast to this, machines follow specific instructions, applying theory as they receive it. A computer just understands assembly, and does not realize the final purpose of its actions. A computer simply understands that it is adding two numbers, or that it must change the Program Counter to the address FFFF:0000. It is a duty for us humans to use those provided instructions to synthesize a program that carries out the desired result.

The previous analogy might seem unrelated to this thesis, but it is not. In classical Artificial Intelligence (AI), building an intelligent system involves domain knowledge in the specific problem that needs to be solved. The programmer (or someone else in other levels of the hierarchy) needs deep knowledge in the problem in order to define a solution, and translate this solution into an algorithm. This defines a bottleneck in the pipeline, both in performance and cost. Performance because the developed algorithm will go as far as the field expertise is available. Cost because improving over state-of-the-art established rules requires deep research and field expertise.

This thesis requires some notions on how emotion recognition is performed in the field of psychology. This technique is primarily based in analyzing facial expressions, which are one of the effects of emotions in human beings.

1.1.1 Swedish Institute of Computer Science

The Swedish Institute of Computer Science (SICS) is a leading research institute for applied information and communication technology in Sweden. SICS is non-profit and carries out advanced research in strategic areas of computer science, in close collaboration with Swedish and international industry and academia. The research creates cutting-edge technology, invigorating companies beyond their own R&D.

The thesis was carried out at the Decisions, Networks and Analytics Laboratory, where I was provided with the necessary environment and equipment.

1.1.2 Department of Psychology Stockholm University

Aside of the main work at SICS, I had a few meetings to discuss and monitor my research at the department of psychology of Stockholm University. Given the orientation of my thesis towards Computer Science and Engineering rather than psychology, these interactions were at the beginning.

1.2 Problem

The problem tackled in this thesis involves creating an AI that is capable of simulating this unconscious process humans employ to recognize emotions. In particular, using Machine Learning to make this AI learn Emotion Recognition by itself.

Apart from solving the main problem, this thesis tries to tackle it in a different way from what can be seen in state-of-the-art. The main contribution is adding temporal units to current single-frame state-of-the-art methods, so an extra dimension of the data can be learned. For instance, when a person is smiling, the facial expression of the individual does not instantly go from neutral to smiling. This temporal information can improve recognizing emotions by analyzing the changes of the face through time.

1.3 Purpose

This area of research is growing in interest every day, with multiple competitions being held [12, 17–20] and a variety of products being released to market [1, 21–24]. The relevance of correct emotion recognition extends to a variety of applications.

First of all, there is User Experience and Marketing Research. Companies spend large amounts of money in Customer Relationship Management (CRM) [25], with special focus on detecting how users react to small changes in their products, or in the release of completely new ones. This process could be potentially automated by the implementation of an agent that automatically infers the level of satisfaction of the customers based on their emotion.

Second, there is Social Robotics. Computers have always been a tool that make our life easier, but there is always the barrier of interaction with the machine.

Human-Computer Interaction is a branch of Computer Science that focuses on this intent. With the latest improvements in AI, many of the graphical interfaces can be substituted by direct actions of the computers, based on explicit orders or user behavior. It is in the latter where automatic emotion recognition comes into play, with the ability of a robot to execute certain actions when, for instance, the person is feeling down.

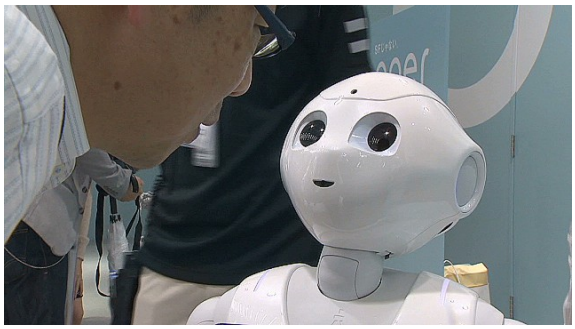

Figure 1.1: Pepper, a robot capable of reading human emotions and react to them [1].

Third, there is security. One example of such is crowd monitoring in certain areas, such as airports. When a subject shows indications of fear, this might be because of a potentially harmful action to be taken. Detecting this behavior could potentially prevent acts such as terrorism or drug trafficking, while at the same time drop the costs of security by automating the whole process with intelligent computer vision through cameras.

Fourth and finally, there is Adaptive Technology. If it is games or educational services, analyzing the frustration of the users can give insights to adapt the content to their engagement.

1.4 Goals

The main goal of this project is to determine whether Syna, which uses the combination of machine learning methods for spatial and temporal features, has the potential to compete with other state-of-the-art when performing emotion recognition.

To fulfill this goal, I will design and implement two new approaches to the problem. The code for these will be openly available on GitHub, so that anyone can replicate the results.

There are different datasets to test the given solutions. These datasets are also used as part of the solution for the learning aspect of the algorithm. Nevertheless, to test the algorithm, I am only allowed to use samples that have not been included in the

training process. The main metric to measure the performance is the accuracy when predicting the category of the expressed emotion.

1.5 Benefits, Ethics and Sustainability

As most research projects done with test subjects, there are the issues of anonymity and privacy. The test subjects used to train the model can be identified, given they appear in the video feed. In addition to this, some other consequences may arise if the technology is not used ethically. For instance, if used in public areas, the emotion analysis of subjects invade to certain degree the privacy of the individuals. In addition to this, the algorithmic bias inherent from the data may falsely identify emotions on individuals when used in certain applications, such as surveillance in airports. An innocent person can be falsely accused due to a false positive. These ethical challenges lead to the same problem that was once described back in 1984 by Orwell as *Surveillance Society* [26].

When it comes to benefits, Syna shares the same benefits that can be found in emotion recognition software. These primarily enable users of the technology to understand the emotional state of other individuals, for a variety of purposes which would not be possible without an automated solution that gathers data ubiquitously. An extensive survey on the benefits of such applications can be found at [27].

1.6 Methodology

I want to explore new alternatives to the task of emotion recognition in video feed. For doing so, I propose two different new approaches that take into account the time dimension for the classification. Since certain datasets have already shown close to 100% accuracy, I will also target those in which the accuracies are lower. The datasets used can be arranged by difficulty of prediction: easy (CK+) and hard (AFEW).

The two new approaches of this thesis both explore the usage of Recurrent Neural Networks (RNNs), which are a class of artificial neural networks where connections between units form a directed cycle, which allows for temporal behavior. In particular, I use what is known as Long Short-Term Memory (LSTM) networks, which solve some of the issues inherent from using RNNs. This will be explained in detail in Section 5.2.

Even though both approaches use LSTMs, this is applied close to the last layers of the network. The initial layers, which are closer to the input data, differ a lot with respect to the type of features to be learned by the algorithms.

The first approach uses a Constrained Local Neural Fields (CLNF) model [28], which efficiently extracts Facial Landmarks in unseen lighting conditions and in the wild. The second approach uses face frontalization for decreasing the variance in the frames and 3D Convolutional Networks (C3D) for extracting both spatial and temporal features. These will be explained in detail in Section 3.

To formally describe the research methods, the 'portal' provided by Hakansson [29] is used. This guideline addresses the need for applying methods before the actual research, along with the reasons for doing so. The main methods chosen for this project are:

- Quantitative and qualitative research methods: The quantitative method is used as this thesis involves comparing the accuracy to other methods as a measurement of effectiveness.
- Philosophical assumption: The interpretivist approach is taken given the social origin of emotion recognition. The basic emotions are categories defined from the study of human psychology, and therefore constitute a social abstraction of reality.
- Research approach: The deductive approach is best suited for this research given the premise of testing whether the accuracy of Syna and DeepSyna are comparable to state-of-the-art.
- Research strategies / designs: This thesis involves experimental research as main strategy, but it is limited by the amount of data available. Statistics such as accuracy are the determining metric used for testing the hypothesis.
- Data collection methods: Experiments are the main source of information for this research. This data is collected and labeled by experienced researchers in the field.

1.7 Delimitations

The research explores only datasets containing video-feed with labels corresponding to emotions. There are various means to test the new approaches, but the chosen one is the performance in terms of prediction accuracy.

The amount of labeled data available for Emotion Recognition is limited when compared with other Computer Vision tasks. This is not just a limitation on the number of datasets, but also on the number of samples they contain. There is even more scarcity when the content needs to be video-feed given the necessity of the temporal dimension. In total, I was granted access to two different datasets, which will be described in Chapter 6.

Some of the Neural Networks used in this thesis are hard to train. In fact, one of these (C3D) would take about 3 weeks to train in a high-end computer. Even with more computing resources, the limited size of my datasets would end-up making my model overfit the training data. Therefore, I am forced to use a technique that allows me to use networks trained on other tasks. This technique will be explained in Section 3.2.3.

1.8 Outline

The rest of the content in this thesis has the following structure:

Chapter 2 explains the historical and theoretical background of the methods used for emotion recognition. This ranges from the origins of the field in Psychology and Neuroscience, to the computer algorithms that automate the classification.

Chapter 3 contains the methods used for extracting facial features. This includes a description of two main approaches: one that relies on domain knowledge, using a pipeline that is typical for automatic facial expression recognition; and one that does not require any prior knowledge, relying on automatic feature construction from raw video data.

Chapter 4 describes the methodology for capturing temporal information from the facial features extracted in the previous step. It is based on a habitual architecture used in the field of ML that has been shown to be best suited for this task.

Chapter 5 describes the work of the thesis in detail. The architecture of the models, what the data looks like and how it is processed, and how the models are trained.

Chapter 6 defines the conducted experiments for determining the effectiveness of the approach. Further tests are conducted to compare Syna with other state-of-the-art methods and established baselines.

Chapter 7 describes the results of the thesis. The main metric to measure the performance is the accuracy when determining the category of the expressed emotion.

Chapter 8 includes a discussion over the content of this thesis. This is done in a critical way, providing a peripheral view on the shortcomings and what could have been improved.

Finally, chapter 9 provides the conclusions of the thesis. It describes the contributions of the thesis to research and development, together with the future work.

Chapter 2

2 Towards Emotional Intelligence: History and Theory

This chapter summarizes the theoretic background and history that makes automatic emotion recognition possible. The overall work connects the multiple disciplines of Neuroscience (Emotions), Psychology (Facial Expressions), Computer Science (Artificial Intelligence), and Data Science (Machine Learning).

2.1 The Science of Emotions

2.1.1 What is an Emotion

When hearing the word emotion, most people tend to think of happiness, love, hate, or fear. Those are the strong emotions that are experienced through life, consciously classifying them as good or bad. This is because our brain is designed to look for threats and rewards. When one of these is detected, the feeling part of the brain alerts us by the release of chemical messages. At the end, emotions are interpreted as the effects of these chemical messages.

For instance, in the case of a threat, our brain releases the stress hormones adrenaline and cortisol, which prepares us for a fight-or-flight response [14]. On the other hand, when perceiving a reward, our brain releases dopamine, oxytocin, or serotonin, which are the chemicals that make us feel good and motivated to continue such behavior.

In these instances of emotion, the feeling part of the brain reacts way before the thinking part does. Sometimes, the reactions of the feeling brain are so strong that dominates our behavior, preventing us from using the thinking part. This can prevent us from thinking rationally, in such a manner that emotions somehow hijack our brain.

Even though most of our emotional responses happen unconsciously, there are methods in which our thinking can control those emotions. Just thinking of something threatening, like presenting in front of a large crowd, can trigger a negative emotional response. It is in such cases where one can control the emotion by conscious thinking, which in this case could be reducing the importance of the audience, or strong confidence that the delivered presentation will be good. There is an entire research field addressing this methodology, shaped by Herbert Benson's Relaxation Response [30].

2.1.2 Components of Emotions

"An emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response" [31].

Subjective Experience While experts accept the universality of the basic emotions, the experience of these emotions in individuals is highly subjective. Even though there are broad labels for certain emotions such as anger or happiness, the manifestation of these in individuals can vary a lot. While anger might mean mild annoyance for someone, it can be a blinding rage for somebody else. Plus, one usually does not experience single emotions, but mixed. An easy example can be starting a new job, in which case one can feel both excited and nervous, in different levels depending on the individual.

Physiological Response Emotions can cause strong physiological reactions. Anxiety can cause sweaty palms, racing heartbeat, or even stomach lurch. Early studies attributed these reactions to the sympathetic nervous system, a section of the autonomic nervous system which controls blood flow and digestion. Nevertheless, recent research targets the brain's role in emotions, especially the amygdala. This almond-shaped structure has been shown to be linked to motivational states such as hunger or thirst, as well as memory and emotion. Researchers have shown that under threat, the amygdala becomes activated, and that damages to this structure can impair fear response.

Behavioral or Expressive Response The main component taken into account for this thesis is the actual expression of the emotion. Humans have the ability of interpreting emotional expressions in the people around them, something that psychologists refer to as emotional intelligence. Many of these expressions are considered universal (e.g. a smile indicating happiness), while cultural roles tend to provide variety in the expressions (e.g. people from Japan have been discovered to mask displays of certain emotions [32]).

2.1.3 Classifying Emotions

Taking into account the three components, describing human emotion can be done with two different approaches.

The Categorical Description of Affect intends to classify emotions into a determined set of classes. Everyone has heard the words happy or sad, as they have been used at least from the 19th century. From 1972, this approach was heavily influenced by the work of Paul Ekman [33–36], who believed that humans universally express a set of six basic emotions: happiness, sadness, fear, anger, disgust and surprise. In 1999, he expanded this list to include embarrassment, excitement, contempt, shame, pride, satisfaction, and amusement [37, p. 301–320].

The Dimensional Description of Affect places a particular emotion into a space with a limited set of dimensions [38, 39]. There are certain variations when determining what the dimensions are, but all include valence (how pleasant or unpleasant the emotion is), arousal or activation (how likely is the person to take action under this emotional state) and control (the sense of control over the emotion). Combining different sets of values in these dimensions can generate more complex emotions.

Out of these two approaches, the Categorical Description of Affect is the one explored in affective computing, given its simplicity and universality claim. The richness of the space in the Dimensional Description is more difficult to automate since it is hard to map expressive responses to certain values of these dimensions.

2.2 The Science of Facial Expressions

Facial expressions study the variations of an individual's appearance due to facial movements under the skin. A facial movement, in turn, is the movement of one or more facial muscles. The mapping between facial movements and facial muscles is many-to-many, which means that one facial movement may involve more than one facial muscle, and one facial muscle can be involved in more than one facial movement. If this last statement seems confusing, think of it in the following way. For certain facial movements, two or more facial muscles need to be contracted. On the other hand, one of those same facial muscles may be contracted in different facial movements.

There is a long history of philosophers and researchers trying to conceive the origin and purpose of facial expressions, within branches such as Creationism, Neuroscience or Psychology.

Facial expressions were first studied in the context of physiognomy and creationism, in which they tried to link a person's character by their looks, especially the face [40]. Leonardo Da Vinci was one of the first to refute such claims, stating that they

were without scientific support [41]. Forward in the 19th century, Sir Charles Bell, influenced by Creationism, investigated their role in the sensory and motor control [42]. He attributed their purpose to solely human communication, endowed by the Creator. Later on, the french neurologist Duchenne studied the body's neuromuscular system and how facial expressions are produced by electrically stimulating facial muscles [43] (See Figure 2.1).

Figure 2.1: Experiments conducted by Duchenne de Boulogne in the 19th century. Adapted from Cambridge University Library.

When studying their origin, facial expressions were first attributed to God [44], and later to evolution. In the 19th century, Charles Darwin stated that Facial Expressions were evolved behaviors for expressing emotion [45]. Darwin's claims were later supported by the research of Adam Anderson [46].

Up until now, there is an ongoing debate on what is the true purpose of facial expressions, and how they increased the chances of survival in the species that used them. On the one hand, there is the role in social communication, specifically in the context of signaling systems. This theory states that the role of facial expressions is a form of nonverbal communication [47], that expressions can communicate everything from pleasure or displeasure to surprise or boredom. On the other hand, sensory regulation considers them as functional adaptations of more direct benefit to the expresser [46, 48]. When experiencing surprise, humans widely open their eyes, not to communicate such expression, but to enhance their field of vision. In the same way, constricting the nose in disgust reduces the inhalation of harmful substances.

2.2.1 Parametrization of Facial Expressions: Facial Action Coding System (FACS) and Action Units (AUs)

When recognizing facial expressions, the first task involves defining a coding scheme for such facial expressions. There are two main classes of coding schemes. Descriptive coding schemes focus on what the face can do based on surface properties, while

judgmental coding schemes describe facial expressions in terms of the latent emotions that generate them.

The most well known example of descriptive coding is the Facial Action Coding System (FACS) [49] developed by Ekman and Friesen, which was later improved in FACS 2002 [50]. The purpose of this scheme is to represent all facial expressions as a combination of facial muscles. Facial expressions are coded in action units (AUs), which represent the contraction of one or more facial muscles (see Figure 2.2). FACS also provides the rules for visual detection of AUs and their temporal segments, which are the ordinal intensity of the AU (onset, apex, offset) from when the facial expression emerges until it fades. A complete description of FACS can be found in [51]. Having this set of rules, a human can analyze a shown facial expression and subdivide it into specific AUs and their temporal segments. A great survey in the history, trends and approaches for Facial Expression Recognition can be found at [52].

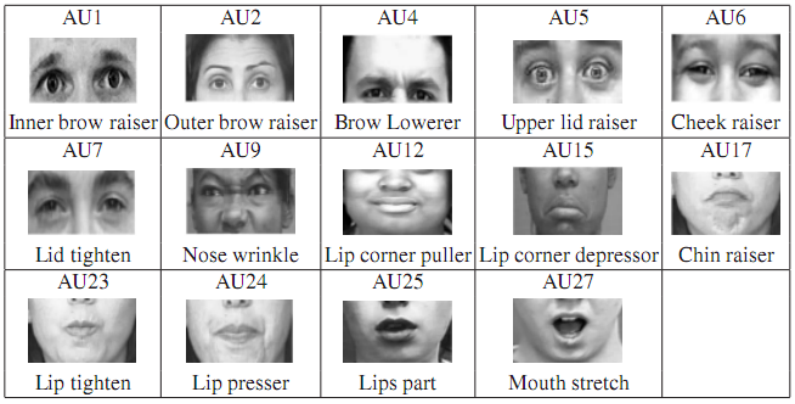

Figure 2.2: Examples of some action units extracted from CK+ database [2].

2.3 Conveying Emotions from Facial Expressions

The question is: what differentiates facial expressions from emotions? In the one hand, facial expressions involve the variations in an individual's face based on different muscles. As mentioned earlier, an emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response. As a result, facial expressions are considered an expressive response of emotions. This relation between facial expressions and emotions heavily relies in the Universality Hypothesis. This hypothesis assumes that certain facial expressions are signals of six basic emotional states (happiness, sadness, anger, fear, surprise and disgust) that are recognized by people everywhere, regardless of culture or language. The truth of this hypothesis has remained one of the longest standing debates in the biological and social sciences. One example of

such is the disclaim made by Jack *et al.* [53] which is supported by the result of a survey targeting different cultural groups. Despite these claims, implementations of these methods have shown decent level of generalization and accuracy, which is the reason a generalized solution to recognizing emotions is possible. One of the main contributions to this relation between facial expressions and emotions was developed by Ekman and Friesen, called Emotion FACS (EMFACS) [54], which scores facial actions relevant for the six basic universal emotions. This can be considered an hybrid of descriptive and judgmental coding schemes.

2.4 Emotion Recognition through Machine Learning

The previous sections showed how psychology describes an approach to perform Emotion Recognition. Nevertheless, from the perspective of a human, this task may seem feasible after some training, but it is laborious when it comes to its formalization. There is a big gap between the visual features that can be captured through a camera (in the format of pixels) and the required processed attributes that are used in Emotion Recognition (Landmarks and AUs).

To overcome this problem, in the past the algorithms used visual hand-crafted features, such as Dense SIFT or Histogram of Oriented Gradients (HOG). More recently, Deep Learning enables to automatically infer a hierarchy representation of this visual information. One of the main methods for doing so is what is known as Convolutional Neural Networks (CNNs), a subclass of Neural Networks inspired by the study of the visual cortex in the brain [55]. The visual information would be encoded into a hierarchy through the layers, with the first layers encoding low-level features such as edges, while later layers can build high-level features such as eyes or mouth.

In the same manner solutions for the task of image classification have to handle a wide variety of objects within each category, an emotion classifier has to handle how different the face of each person can be. There are some instances of CNNs for emotion recognition that have shown competitive results. Burkert *et al.* [56] achieves high accuracies in the datasets of Extended Cohn-Kanade (CK+) and MMI Facial Expression Database (99.6% and 98.63%, respectively). It is worth mentioning that this solution relies on individual images. It can be used in these video datasets since the emotions are shown from neutral (first frame) to the highest peak of the emotion (last frame). They simply pick the first frame as a neutral emotion and the last frames as the labeled emotion.

The learning process of the emotions can be tackled from two different perspectives. On the one hand, an approach could be to train the network to detect those characteristics that are known to be related to emotional states, mainly the FACS [49] system. On the other hand, one could ignore all the domain knowledge about

Emotion Recognition, and let the machine learn the most suitable features that determine an emotion. Either approach can be equally valid, since recent research has shown strong correlation between both methods. Khorrami *et al.* [3] experimented on whether deep neural networks learn Facial Action Units when doing Expression Recognition, and the results in Figure 2.3 show that CNNs trained to do emotion recognition model high-level features that strongly correspond to Facial Action Units.

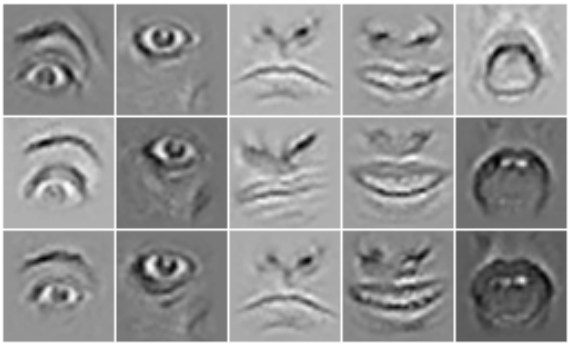

Figure 2.3: Sample of the learned features by a CNN when performing emotion recognition. From left to right, maximally activating images of fear, disgust, sadness, happiness, and surprise. For instance, note how in the case of surprise, there is a strong activation when subjects have their mouths open, which corresponds to AU 27. Adapted from [3, p. 25].

Chapter 3
3 Facial Feature Extraction

This chapter provides an explanation of the techniques and building blocks that will be used for developing a system that automatically infers facial features of an individual from visual footage. There are two main approaches for this task. On the one hand, there is Automatic Facial Expression Recognition, which relies on prior knowledge, adopting concepts from the study of Human Psychology and Neuroscience. On the other hand, there are 3D Convolutional Neural Networks, which are a subclass of CNNs applied to analyzing visual imagery while also capturing the motion information encoded in multiple adjacent frames.

3.1 Automatic Facial Expression Recognition

There are three main steps when tackling the problem of Automatic Facial Expression Recognition:

1. Face localization in the image
2. Feature extraction from the face
3. Classification/Regression from facial features

For each of these steps, a state-of-the-art technique is applied. King [57, 58] is the author of the face detector that can be found in the dlib library. Baltrušaitis *et al.* proposed Constrained Local Neural Fields (CLNF) for robust facial landmark detection in the wild [7], and also Cross-dataset learning and person-specific normalisation for automatic Action Unit detection [59].

3.1.1 Constrained Local Model

CLNF uses what is known as Constrained Local Model (CLM) framework, so it is briefly explained in this section. CLM was coined by Cristinacce and Cootes [5], and it is a class of methods for modelling deformable objects that posses a

distinct set of features (see Figure 3.1). This can be applied to a setting in which there is a face (deformable object) and one wants to detect the facial landmarks (features).

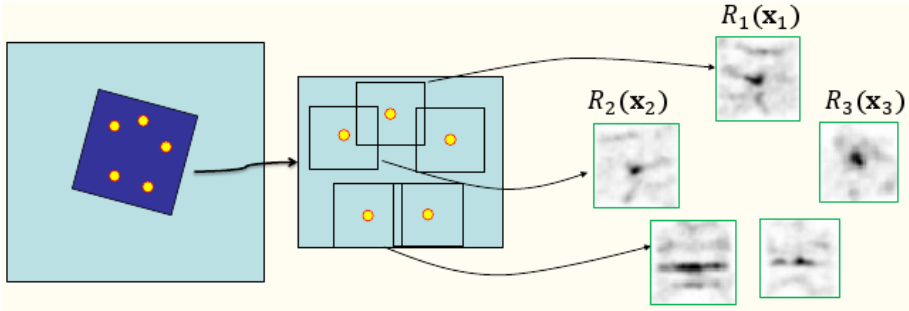

Figure 3.1: Given a detected object on the image (left), a set of features locations are predicted (middle) and a "response image" $R(x)$ is generated for each location (right) [4].

It all starts by providing an estimate on where the location of the features are within the image. In the case of the face, a template of the landmarks seen from a frontal view over the area from a face detector is the first estimate. This is adjusted through multiple iterations until convergence (see Figure 3.2). The overall workflow can be subdivided in three main components: a point distribution model, patch experts and a fitting approach.

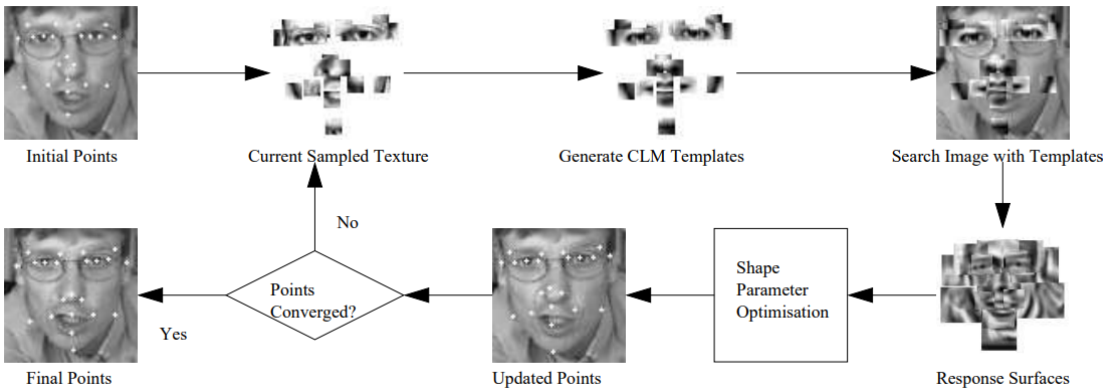

Figure 3.2: CLM Search Algorithm [5, p. 5].

Point Distribution Model

The point distribution model (PDM) represents the mean geometry of a shape through a set points, which are inferred after performing training from a set of labeled shapes. In our particular case, it models the location of facial landmarks in the image using a non-rigid shape and rigid global transformation parameters[1].

[1]In geometric terms, rigid parameters belong to the types of transformations that do not change the shape of an object, while nonrigid parameters do.

The location of the ith landmark is represented as $x_i = [x_i, y_i, z_i]^T$ and controlled through the parameters of the PDM:

$$xi = s \cdot R_{2D} \cdot (\bar{x}_i + \Phi_i q) + t \tag{3.1}$$

where $\bar{x}_i$ is the mean value of the ith landmark, Φ_i is a $3 \times m$ principal component matrix, q is an m dimensional vector of parameters controlling the non-rigid shape. The rigid shape parameters can be defined using 6 scalars: a scaling term s, a translation $t = [tx, ty]^T$, and an orientation $w = [wx, wy, wz]^T$. Rotation parameters w control the rotation matrix R_{2D} (the first two rows of the 3×3 rotation matrix R). Thus, the shape parameters can be described by the vector $p = [s, t, w, q]$.

Patch Experts

Patch experts evaluate the probability of a landmark being aligned at certain pixel location. There is one patch expert per landmark, and the response of the ith patch expert π_{x_i} at the location x_i in the image I is defined by:

$$\pi_{x_i} = C_i(x_i; I),$$

where C_i is the regressor for the ith landmark, and its output can be modelled using values from 0 (no alignment) to 1 (perfect alignment).

There have been multiple methods proposed as patch experts, but the most popular for a long time has been a linear Support Vector Regressor (SVR) in combination with a logistic regressor [60, 61]. An example of its response maps can be seen in Figure 3.3.

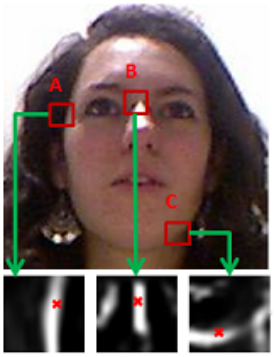

Figure 3.3: Logistic regressor response maps of three patch experts: (A) face outline, (B) nose ridge and (C) part of chin. The red cross represents the ground truth position. Adapted from [6, p. 1].

Fitting Approach

The fitting approach is used to estimate the optimal rigid and non-rigid parameters $p*$ that fit the underlying image best:

$$p^* = \underset{p}{\operatorname{argmin}}[R(p) + \sum_{i=1}^{n} D_i(x_i; I)],$$

where R is a penalization term for overly complex models or unlikely shapes and D_i is the measurement of misalignment of the ith landmark. After an initial estimate p_0 is provided, an update parameter Δ_p it's required for approaching optimal solution $p*$:

$$p^* = \underset{\Delta p}{\operatorname{argmin}}[R(p_0 + \Delta p) + \sum_{i=1}^{n} Di(x_i; I)]$$

There is a variety of fitting strategies applied to CLMs, but a popular technique is the Regularised Landmark Mean Shift (RLMS) [60]. It uses the least squares method to fit the following function[2]:

$$p^* = \underset{\Delta p}{\operatorname{argmin}}(||p_0 + \Delta p||^2_{\wedge^{-1}} + ||J\Delta p_0 - v||^2), \tag{3.2}$$

where J is the Jacobian of the landmark locations respect to the parameter vector p; $\wedge^{-1}$ is the prior matrix of the parameter p, in such a manner that the non-rigid parameters follow the Gaussian distribution $\mathcal{N}(q|0, \wedge)$ and the rigid parameters follow a uniform distribution; and $v = [v_1, ..., v_n]^T$ is the mean-shift vector over the patch responses using a Gaussian Kernel Density Estimator:

$$v_i = \sum_{y_i \in \Psi_i} \frac{\pi_{y_i} \mathcal{N}(x_i^c | y_i, \rho I)}{\sum_{z_i \in \Psi_i} \pi_{z_i}(x_i^c | z_i, \rho I)} - x_i^c$$

Finally, the update rule is derived using Tikhonov regularised Gauss-Newton method and it is computed iteratively until convergence[3]:

$$\Delta p = -(J^T J + r \wedge^{-1})^{-1}(r \wedge^{-1} p - J^T v)$$

[2] $||\cdot||_W$ refers to a weighted $l2$ norm.
[3] r is the regularisation term.

3.1.2 Constrained Local Neural Field

The Constrained Local Neural Field (CLNF) model is an instance of the CLM framework that includes a novel Local Neural Field (LNF) patch expert and a novel Non-uniform RLMS fitting technique. CLNF outperforms other state-of-the-art techniques when estimating landmarks in unseen lighting conditions and *in the wild settings*. The content of this section significantly relies in the theory that can be found in Baltrušaitis' PhD thesis [62].

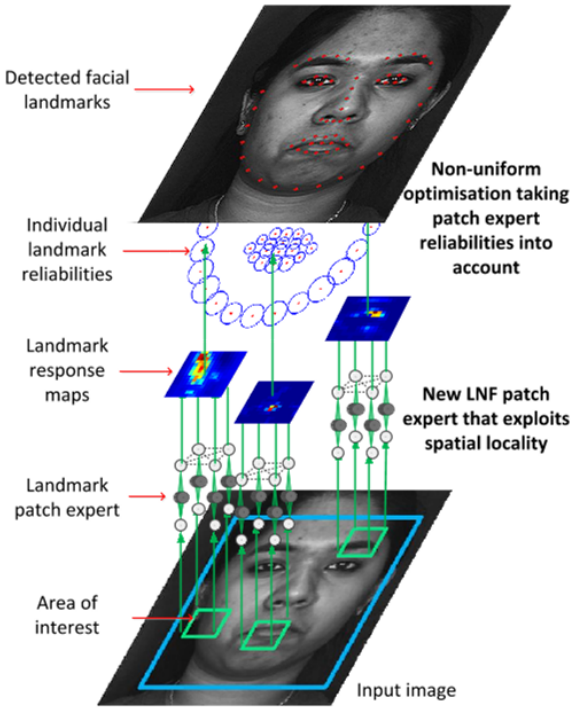

Figure 3.4: Overview of the CLNF model (showing only three patch experts) [7, p. 1].

Local Neural Field patch expert

One of the issues of CLM models is that the prevailing patch experts bottleneck the performance of the landmark detection in complex settings. It is primarily because linear SVRs or logistic regressors fail to learn non-linear relationships between pixel values and response maps. For instance, they are capable of real-time tracking but perform very poorly on illumination-invariant landmark detection. Furthermore, the alternative more complex methods (such as RBF kernel SVRs) are too slow (under one frame-per-second), which limits their usage and training time on big datasets.

To solve these issues, Baltrušaitis *et al.* introduced the Local Neural Field (LNF) patch expert, which combines the non-linearity of Conditional Neural Fields [63]

with the flexibility and continuous output from Continuous Conditional Random Fields [64]. LNF captures two types of spatial relationships: similarity (pixels nearby should have similar alignment probabilities) and sparcity (only one peak in the whole area of a patch expert). This can be seen in Figure 3.5 when comparing the two variants of LNF to SVR.

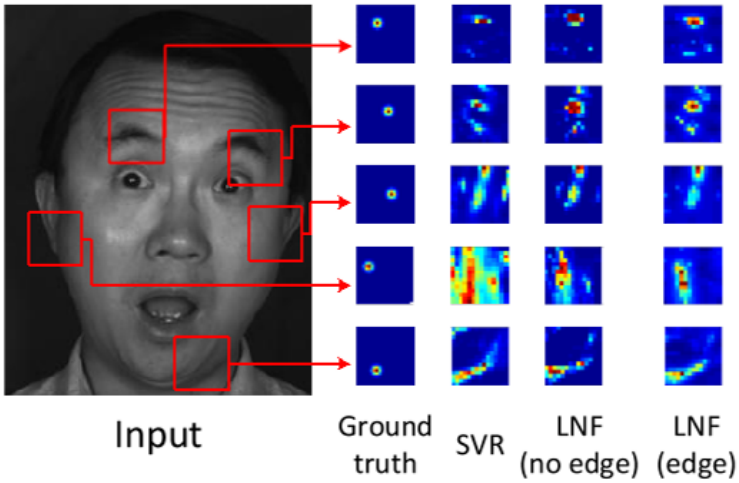

Figure 3.5: Sample of the response maps from four patch experts using different response techniques. Notice the noisiness of the SVR patch expert when compared to LNF. Also, adding edge features leads to a smother response [7, p. 2].

The model training consists of a set of observations $X = \{x_1, x_2, ...x_n\}$ and a set of ground truth landmark locations $y = \{y_1, y_2, ...y_n\}$, where $x_i \in \mathbb{R}^m$ represents the vector of pixel intensities in a patch expert region (e.g. $m = 121$ for an 11×11 support region), and $y_i \in \mathbb{R}$ is a scalar prediction at location i. The process will now be explained through its components.

First, the LNF model is an undirected graph that models a conditional probability distribution with the following probability density:

$$P(y|X) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) dy} \tag{3.3}$$

where $\int_{-\infty}^{\infty} \exp(\Psi)$ is a normalization function for the probability distribution, making it sum to 1. Second, the kernel of the model is defined as follows:

$$\Psi = \sum_{i} \sum_{k=1}^{K1} \alpha_k f_k(y_i, X, \theta_k) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j) + \sum_{i,j} \sum_{k=1}^{K3} \gamma_k g_k(y_i, y_j) \tag{3.4}$$

where $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_{K1}\}$, $\theta = \{\theta_1, \theta_2, ..., \theta_{K1}\}$, $\beta = \{\beta_1, \beta_2, ..., \beta_{K2}\}$, $\gamma = \{\gamma_1, \gamma_2, ... , \gamma_{K3}\}$ are the learned parameters. There are three individual kernels within the kernel function: vertex features f_k, edge features g_k and edge features l_k.

$$f_k(y_i, X, \theta_k) = -(y_i - h(\theta_k, x_i))^2 \tag{3.5}$$

$$h(\theta, x) = \frac{1}{1 + \exp(-\theta^T x)} \tag{3.6}$$

$$g_k(y_i, y_j) = -\frac{1}{2} S_{i,j}^{(g_k)} (y_i - y_j)^2 \tag{3.7}$$

$$l_k(y_i, y_j) = -\frac{1}{2} S_{i,j}^{(l_k)} (y_i + y_j)^2 \tag{3.8}$$

Vertex features f_k represent a 1-layer CNN that maps the input x_i to output y_i, and θ_k is the weight vector for a particular neuron k.

Edge features g_k represent the similarities between observations y_i and y_j, enforcing smoothness on connected landmarks through the neighborhood measure $S^{(g_k)}$. In particular, $S^{(g_1)}$ returns 1 when two nodes i and j are horizontal/vertical neighbors in a grid, and 0 otherwise. $S^{(g_2)}$ returns 1 when two nodes i and j are diagonal neighbors in a grid, and 0 otherwise.

Edge features l_k represent sparsity constraints. For instance, it penalizes the model when both y_i and y_j are high, but it's not penalized when both are zero. This has the unwanted consequence of slightly penalizing y_i for just being high, but the penalization is way bigger when both are high. The neighborhood measure $S^{(l_k)}$ allows to define regions where sparsity is enforced, in such a manner that the neighborhood region $S^{(l)}$ returns 1 when two landmarks i and j are between 4 and 6 edges apart in a grid layout, and 0 otherwise. These bounds have been shown empirically to work best.

Finally, the parameters $\{\alpha, \beta, \gamma, \theta\}$ are estimated such that they maximize the conditional log-likelihood of LNF on the training sequences:

$$L(\alpha, \beta, \gamma, \theta) = \sum_{q=1}^{M} logP(y^{(q)}|x^{(q)}) \tag{3.9}$$

$$(\bar{\alpha}, \bar{\beta}, \bar{\gamma}, \bar{\theta}) = \underset{\alpha,\beta,\gamma,\theta}{\operatorname{argmax}}(L(\alpha, \beta, \gamma, \theta)) \tag{3.10}$$

It is worth mentioning that the probability density function (Equation 3.3) is converted into a multivariate Gaussian form because it helps with the derivation of the partial derivatives of Equation 3.9. For the sake of brevity, the reader is referred to [62, Chapter 6, Section 1.2] for the full explanation on this conversion.

Non-uniform Regularised Landmark Mean Shift

One of the problems inherent from the fitting performed in CLMs is that each patch expert is equally trusted. This is specially problematic when the output of certain response maps are noisy, such as the SVR response maps in Figure 3.5. To tackle this issue, CLNF uses a simple modification to CLM's objective function:

$$\underset{\Delta p}{\operatorname{argmin}}(||p_0 + \Delta p||^2_{\wedge^{-1}} + ||J\Delta p_0 - v||^2_W), \tag{3.11}$$

where W is the diagonal weight matrix representing the trust on each patch expert. Notice that when comparing to RLMS in Equation 3.2, this formula is exactly the same when W is an identity matrix. Then Tikhonov Regularization is applied to the update rule:

$$\Delta p = -(J^T W J + r\wedge^{-1})^{-1}(r \wedge^{-1} p - J^T W v) \tag{3.12}$$

Finally, to construct W the correlation coefficients are calculated using holdout validation separately for each view and scale.

3.1.3 Head Pose Estimation

The estimated facial landmarks can be used to estimate the head pose. This is done by using a 3D representation of these and projecting them to the image using orthographic camera projection, solving a Perspective-n-Point problem [65].

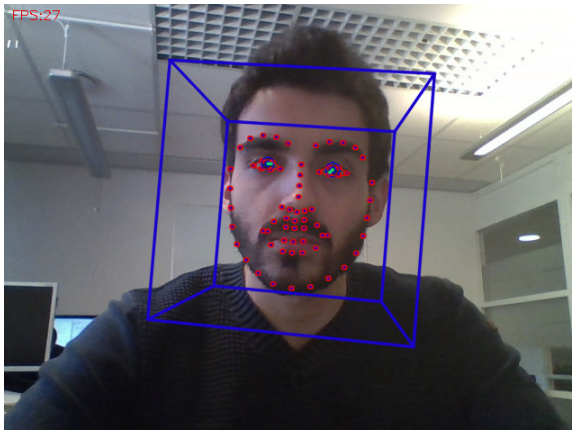

Figure 3.6: Visualization of landmarks and head pose estimation.

3.1.4 Action Unit Recognition

As it was described in Chapter 2, AUs can be recognized from facial landmarks. Nevertheless, recent work has shown better performance when landmarks only assist in face alignment while texture features are used for AU detection. Baltrušaitis *et al.* [59] published a real-time AU intensity estimation and occurrence detection system based on Histogram of Oriented Gradients. By using CLNF for landmark estimation, it has been shown to outperform the Facial Expression Recognition and Analysis challenge (FERA) 2015 [17] baselines. An overview of the system can be seen in Figure 3.7, and is briefly explained below.

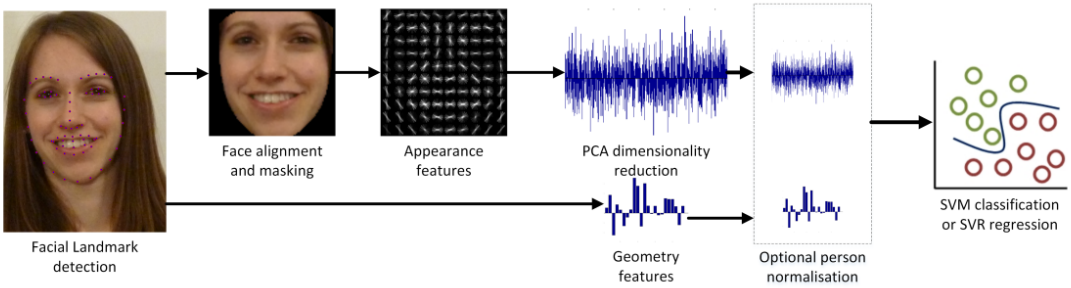

Figure 3.7: Overview of the AU detection or intensity estimation pipeline.

In order to analyze the texture of the face, this needs to be mapped into a common reference frame to avoid useless variance inherent from position and rotation with respect to the camera. This can be done by applying a similarity transform from the currently detected landmarks to a representation of frontal landmarks in a neutral expression. This was done by applying the Procrustes superimposition that minimized the mean square error between aligned pixels, using only a subset of landmarks that are the most stable ones across all facial expressions (these are the ones located at the nose, under the eyes and by the sides of the face). Finally, masking is applied to remove non-facial information using convex hull surrounding the aligned landmarks. The result is a 112×112 pixel image with 45 pixel interpupilary distance.

To extract appearance features the author applies Histogram of Oriented Gradients, using blocks of 2×2 cells of 8×8 pixels, resulting in 12×12 blocks of 31 dimensional histograms. At the end, there is a 4464 dimensional vector describing the face. Principal Component Analysis (PCA) is applied to reduce the dimensionality down to 1379 dimensions after using a wide variety of facial expression datasets: CK+ [11], DISFA [66], AVEC 2011 [19], FERA 2011 [18], and FERA 2015 [17]).

In addition to appearance features, a set of geometry-based features from CLNF are extracted. In specific, the non-rigid shape parameters q (representing the top 23 dimensions, responsible for 95% variance in the training landmark data) and landmark

locations from Equation 3.1. As a result, there are an additional $23 + 3 \times 68 = 227$ dimensions.

Since neutral expressions vary between individuals, it is sometimes mistaken with showing certain emotion. For instance, some people are more smiley or more frowny even though their faces are at rest [67]. For this reason, the descriptors are normalized per sample, by subtracting the median value of the face of the person in the video.

Finally, Support Vector Machines (SVM) is used for AU occurrence detection and Support Vector Regression (SVR) is used for AU intensity estimation. The kernels used are linear, since complex kernels do not improve performance while significantly slowing down the training process.

3.2 Automatic Feature Construction

The previous section explained a technique which targets the modeling of an emotion classifier based features developed from the study of emotions in the field of psychology. This section will explore an alternative that does not use any domain knowledge for constructing these models. It will focus on explaining 3D Convolutional Neural Networks (3D CNNs), and for brevity it is required that the reader already understands the notion of standard CNNs despite a brief introduction being provided below.

3.2.1 3D Convolutional Neural Networks

In the last couple years, CNNs have become the *de facto* technique for a wide range of Computer Vision tasks. This is due to their large performance margin they exhibit with respect to other methods and their capacity for constructing a hierarchical representation of visual footage. Early layers in the network learn to identify small components such as edges, while layers close to the output can identify entire objects. This learned representation can be demonstrated through a process of deconvolution [68], which allows to visualize which features are learned on each layer of the network.

The main limitation of CNNs is that they only consider spatial information at individual frame level. Tasks in Computer Vision that involve video feed do not adapt so well when using CNNs, since they do not take into account implicit motion data. It is for this reason that Ji *et al.* introduced 3D CNNs for human activity recognition [69], which captures motion information encoded in multiple adjacent frames.

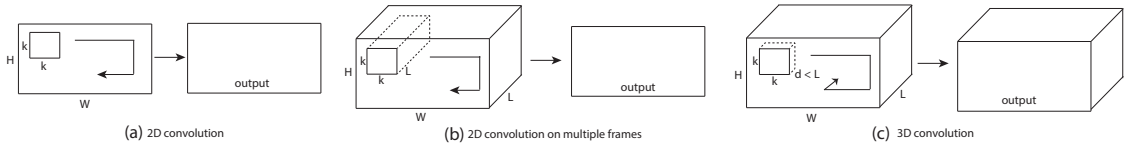

Figure 3.8: 2D and 3D convolution operations. (a) Applying 2D convolution on an image results in an image. (b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. (c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal [8].

The core difference between 2D and 3D CNNs relies on how convolution and pooling are performed. In addition to CNNs exploiting spatial neighborhood between pixels, 3D CNNs utilize temporal neighborhood among frames (see Figure 3.8). More formally, the resulting pixel from a 2D convolutions can be described as:[4]

$$y_{ij} = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} w_{ab} x_{(i+a)(j+b)} \tag{3.13}$$

where w_{ij} is the value at the position (i,j) from a kernel of size $m \times n$, and x_{ij} is the pixel at position ij in the input image. In contrast, the formula of 3D convolutions is:

$$y_{ijk} = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} \sum_{c=0}^{p-1} w_{abc} x_{(i+a)(j+b)(k+c)} \tag{3.14}$$

where w_{ijk} is the value at the position (i,j,k) from a kernel of size $m \times n \times p$, and x_{ijk} is the pixel at position ij in the frame k in the input video. The same concept is applicable when comparing 2D and 3D pooling.

3.2.2 C3D

One of the tasks in which 3D CNNs have been proven very successful is in classification of videos. In particular, Tran *et al.* [8] proved the effectiveness of their C3D architecture by outperforming all previous state-of-the-art methods in 4 different benchmarks. They were also the first ones to empirically prove that 3D CNNs were more suitable for spatio-temporal feature learning compared to CNNs.

The architecture of the network can be seen in Figure 3.9. It consists of $3 \times 3 \times 3$ convolution kernels, 8 convolution layers, 5 pooling layers, followed by two fully connected layers, and a softmax output layer. The training was done on the Sports-1M [70] dataset.

[4] In CNNs, a non-linearity such as tanh is also present

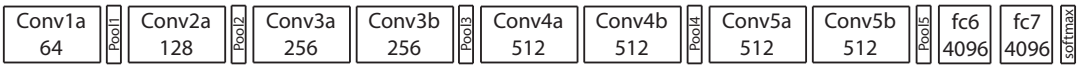

Figure 3.9: C3D network architecture. All convolution kernels are $3 \times 3 \times 3$*, while all pooling kernels are* $2 \times 2 \times 2$ *except for pool1, which is* $1 \times 2 \times 2$*. The stride in all dimensions is 1 [8].*

3.2.3 Transfer Learning for Emotion Recognition

The C3D architecture is used for the variant of Syna which is based on Automatic Feature Construction. Unfortunately, given the depth of the architecture and the limited amount of emotion data available in video format, it is unrealistic to try training the network from scratch. One could argue that reducing the depth of the network can help reducing the needed data, but even by doing so the data is still too limited. C3D was trained over a million videos, while in the case of emotions there are only sequences of hundreds taking into account all the datasets used in this thesis.

To solve this issue, the usage of transfer learning [71] is explored. The pre-trained C3D model on the Sports-1M dataset is used, removing the last fully connected layers in charge of the classification. Therefore, the 4096 intermediate output features are passed to the temporal classification step in the pipeline. During training, the gradients of classification are not back-propagated to C3D, again because of data scarcity. Thus, the training process will focus on learning how to handle C3D output features to predict emotions.

Chapter 4
4 Temporal Classification

The previous chapter explained the methodology for extracting facial features, which corresponds to the first stage of the pipeline implemented on this thesis. CLNF relies on domain knowledge and is limited to spatial information, while C3D is a general method for capturing spatial and short-term temporal information from image sequences. Without regard to which of these techniques is used, the last steps of the pipeline remain the same. In particular, they capture the temporal information present in the sequence of features provided by CLNF and C3D, and infer a final emotion category for the entire sequence. It is in this domain where a particular class of Machine Learning techniques is useful for dealing with sequence information, capturing both short and long-term relationships in the data.

4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of Neural Networks specialized in sequential processing. While in Feed-Forward Neural Networks the inputs and outputs are fixed in size and independent among samples, RNNs' inputs and outputs can be of arbitrary size and depend on previous observations. In other words, the training/inference step t depends also on step $t-1$, and subsequently on every previous step (see Figure 4.1). This enables RNNs to capture temporal correlations in the form of using the weights as a mechanism for persistence.

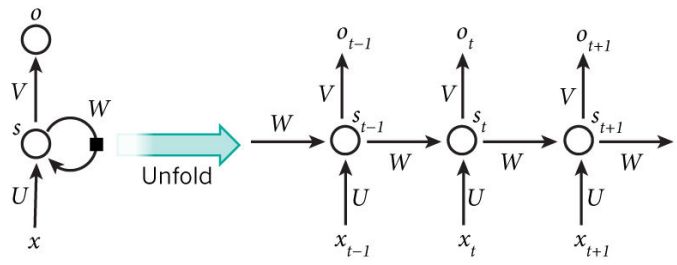

Figure 4.1: The unfolding in time from a recurrent neural network during forward computation [9].

One of the main issues that emerged when using RNNs is what is know as the *vanishing gradient problem* [72]. In Feed-Forward Neural Networks, the effects of this phenomenon increase as the number of hidden layers in the model increase. In RNNs, this becomes even more problematic given the backpropagation through time (BPTT), which makes each layer backpropagate within itself and thus making the gradient vanish at a greater pace. In other words, in Feed-Forward Neural Networks, the gradient is propagated back to the input of the model, while in RNNs the gradient is backpropagated both within the same neuron to previous timesteps, and also to the previous layers. This limited the usage of RNNs until better architectures supporting BPTT were developed. Britz *et al.* [73] provide a detailed explanation of both RNNs and the vanishing gradient problem.

4.2 Long Short-Term Memory

To solve the gradient short-comings of RNNs, Hochreiter *et al.* introduced a new method titled Long Short-Term Memory [74]. While in RNNs the repeating module has a very basic structure (usually just a tanh layer), LSTMs posses a more complex structure.

Figure 4.2 depicts an overview of the repeating structure in LSTMs. From the perspective of the data interacting in this unit, C_t is the cell state acting as memory, h_t represents the output of the module and X_t represents the input. There are four main computational steps within the module, which will be described from left to right.

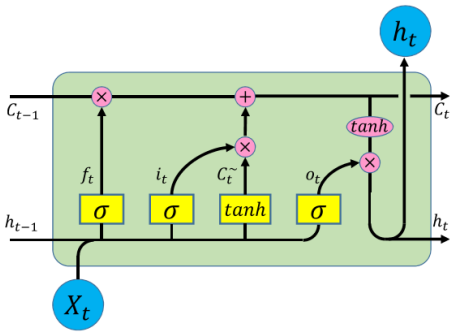

Figure 4.2: The repeating module in LSTMs, where blue circles represent input/outputs of the module at timestep t, yellow rectangles represent neural network layers and pink circles represent pointwise operations [10].

The first component in the LSTM is the "forget gate" and determines the information that is desired to be thrown away from the cell state, and it's modelled according to:

$$f_t = \sigma(W_f \cdot [h_{t_1}, x_t] + b_f) \tag{4.1}$$

where W_f represents the learned weight parameter, h_{t_1} is the previous output, x_t is the new input and b_f is the learned bias parameter.

The second component decides what new information needs to be added to the cell state. This is further subdivided into two main parts: the "input gate" decides the values to be updated (using a sigmoid to determine from 0 to 1 how much each value should be updated), while the tanh layer creates a vector of new candidate values $\tilde{C}_t$. Combining these two with a pointwise multiplication allows us to create a cell state update. This whole process can be formalized as:

$$i_t = \sigma(W_i \cdot [h_{t_1}, x_t] + b_i) \tag{4.2}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t_1}, x_t] + b_C) \tag{4.3}$$

The first and second components are applied to the cell state C_{t_1} in order to generate the new cell state C_t. First, it is multiplied by f_t forgetting what is considered unnecessary, and then the new candidate values scaled according to how much each state value should be updated are added. The resulting new cell state is:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4.4}$$

Finally, the output of the unit is computed. This output represents a filtered version of the cell state, and the applied filter becomes another parameter to be learned:

$$o_t = \sigma(W_o \cdot [h_{t_1}, x_t] + b_o) \tag{4.5}$$

$$h_t = o_t * tanh(C_t) \tag{4.6}$$

4.3 Capturing Temporal Features in Emotions

For the task of determining the emotion shown in a video, one could think that the straightforward approach could be to just calculate the average emotion among frames. This wouldn't work in most cases, since the shown emotion usually lasts for a very short period of time in comparison with the neutral expression. This makes habitual image-based methods such as CNN inapplicable directly to the task at hand.

The application of capturing temporal information in the task of emotion recognition is two-folded. First, it can hypothetically enable constructing a model which automatically reports a single prediction for the entire sequence of images. Second, it can use the implicit temporal information that is present in the video sequence. One can easily hypothesize on how this implicit information can be useful for a learning algorithm. Emotions do not appear and disappear instantaneously, they emerge and fall through a short period of time. This motion can help identify key emotions further from still images. For instance, the evolution of the face from a neutral expression to being happy has a very specific pattern (see Figure 4.3). Also, some emotional expressions appear more spontaneously than others, such as the case of a surprise.

Figure 4.3: Temporal pattern embedded in the expression of happiness [11].

Therefore, LSTMs are used for capturing this temporal dependencies present in evolution of emotions through time. The input facial features depends on the technique used for facial feature extraction, which at the same time varies the suited hyperparameters of the LSTM. Hypothetically, the LSTM will enhance the classification when the evolution of the input facial features corresponds strongly with the learned evolution of certain emotion. It is also important to notice that LSTMs have the capacity of providing real-time predictions (after each frame), which also enables the system to perform live prediction. Last, a fully-connected layer together with a softmax layer provide probabilities for the estimated emotions. A diagram of the proposed temporal classifier can be seen in Figure 4.4.

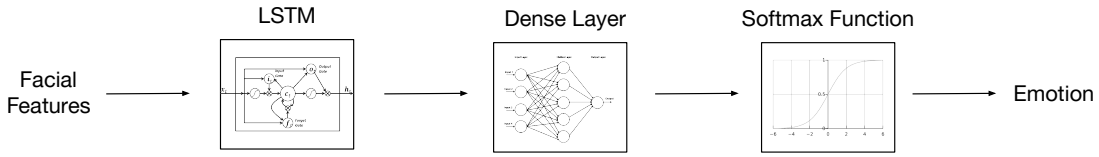

Figure 4.4: Diagram of the proposed temporal classifier. An LSTM layer captures the temporal information, while the fully-connected layer and softmax function provide the emotion estimates.

Chapter 5

5 Spatio-Temporal Emotion Recognition

This chapter explains the two new approaches used in this thesis. Chapters 3, 4 described that the system is divided in two main parts: a facial feature extractor and a temporal classifier. Figure 5.1 depicts the main diagram of the architecture. The two new approaches share the same face detector [58] and temporal classifier described in Section 4.3, but differ in how facial features are extracted. It is important to notice that sharing architecture components does not imply that these share the same hyperparameters. In other words, the system architectures are the same, but the hyperparameters used in the learning stages are different, according to the different extracted features of these models.

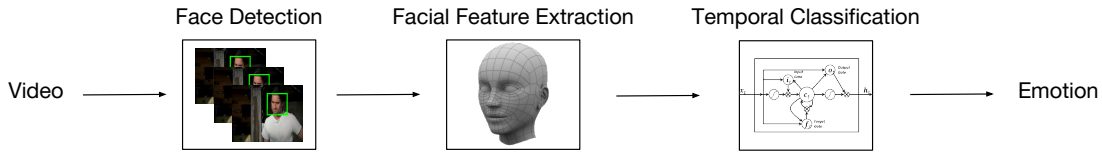

Figure 5.1: The main components of all the variants of Syna. First, the faces are detected from the frames of the input video. Second, the facial features are extracted from the detected face. Third, the facial features are classified through time, and an estimate for the emotion in the entire video is provided.

There is previous work on emotion recognition capturing temporal information. Jung *et al.* [75] proposed a combination of two deep neural networks: a deep temporal appearance network and a deep temporal geometry network. The former is based on 3D CNNs applied to cropped faces for detecting facial movements, while the later is a DNN that tracks the trajectories of facial landmarks. More similar to the approach employed in this work, Fan *et al.* [76] proposed an hybrid network with two main parts. The first one captures visual feed through two pipelines, one using CNN-RNN and the other 3D CNN. The second is in charge of capturing audio features using an SVM. The prediction from the three models is later used to build an hybrid

classifier which predicts the emotion. Unlike the previous work, Syna and DeepSyna are composed of a single pipeline model.

5.1 Syna

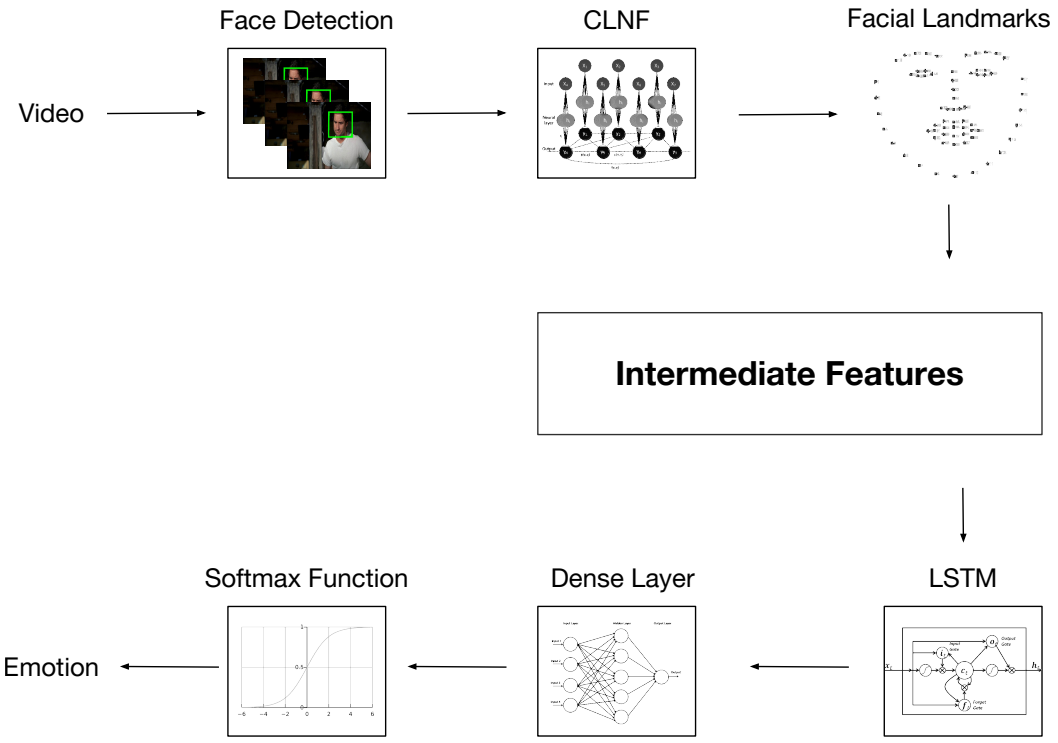

Figure 5.2: System diagram of Syna. The detected faces in each frame are fed to CLNF, from which the facial landmarks are extracted. Then, these landmarks are used for capturing intermediate features: either normalized landmarks, AU occurrences or AU intensities. These intermediate features are later feed into the temporal classifier.

The first variant of the system uses the methods described in Section 3.1 for automatic facial expression recognition. To avoid confusion, this particular variant will be referred to as Syna from here on. The cropped face is feed into CLNF which outputs the landmark estimates. These landmarks are further processed into intermediate features, which can be normalized landmarks, AU occurrences or AU intensities. These intermediate features are finally feed into the LSTM and a final 1-fully-connected layer with a softmax function that provides the final emotion classification. All three intermediate feature alternatives are explored, as described below.

Normalized facial landmarks as intermediate features The straightforward feature to be used are the facial landmarks produced by CLNF (see Figure 5.3). These landmarks are provided in camera coordinates, which introduces a lot of useless variance in the data, depending on where the person is located with respect to the camera lens. To tackle this issue, one can convert the camera coordinates into local coordinates with respect to the head pose. This can be done by using the facial landmarks to estimate the head pose as described in Section 3.1.3, and then converting the landmarks to local coordinates. Figure 5.3 illustrates a diagram of this process.

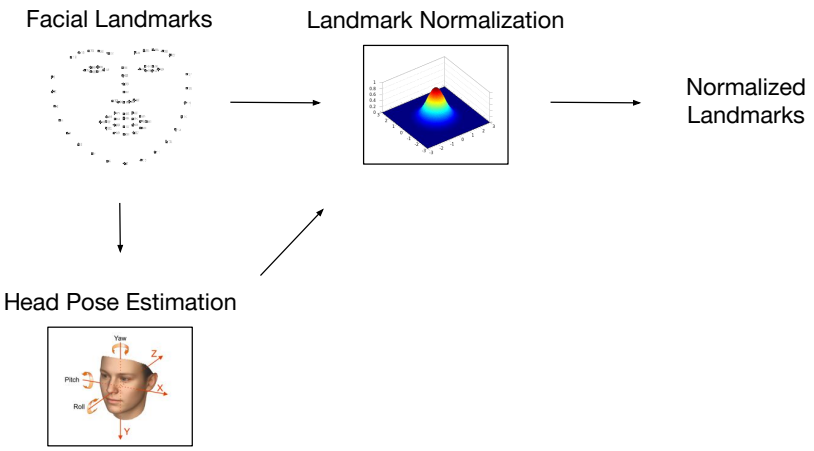

Figure 5.3: Facial Feature Extraction system diagram based on facial landmarks intermediate as features.

Action Units' Occurrence as intermediate features The second alternative is to use the AU occurrences metric as intermediate feature (see Figure 5.4). It follows the method described in Section 3.1.4, using an SVM as classifier.

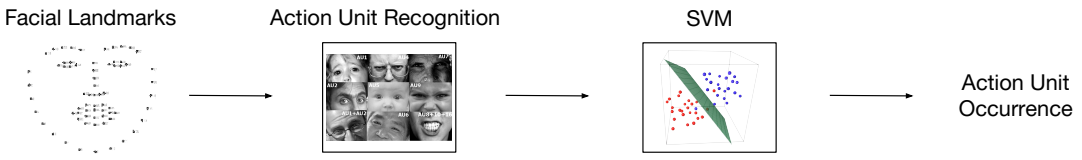

Figure 5.4: Facial Feature Extraction system diagram based on AU occurrences as intermediate features.

Action Units' Intensity as intermediate features The third and last intermediate feature under test is the intensity of AUs (see Figure 5.5). Again, it follows the method explained in Section 3.1.4, but in this case an SVR is applied to retrieve intensities instead of occurrences.

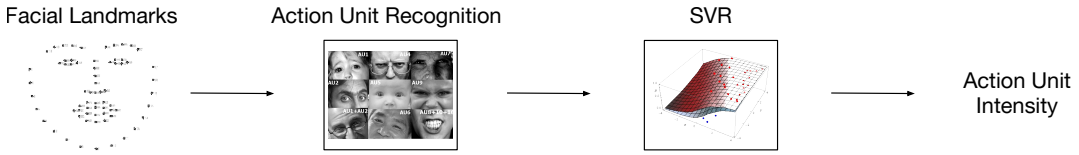

Figure 5.5: Facial Feature Extraction system diagram based on AU intensities as intermediate features.

Normalization of the data prior to applying the learning algorithm is often essential to improve the overall training performance. This improvement affects both training convergence and accuracy [77]. For this reason, Gaussian Normalization is applied so the intermediate features (landmarks, AU occurrences and AU intensities) range from -1 to 1 with unit standard deviation. Finally, these intermediate features are feed to the temporal classifier explained in Section 4.3.

5.2 DeepSyna

The second variant of the system, referred to as DeepSyna, uses the methods described in Section 3.2 for automatic feature construction. In specific, it uses the pre-trained C3D model. A diagram of this system can be seen in Figure 5.6. The detected face is frontalized using the method described in [78]. Next, the pixel values are normalized to the range $[-1, 1]$ with unit standard deviation and the frames are pre-processed for C3D input. This pre-processing involves zero padding the sequences so they fit the 16-frame architecture, subtracting the mean cube and center-cropping the image to 112×112 format. Next, 3D convolutions are applied as explained in Section 3.2.2, and the 4096 intermediate features are feed to the LSTM. Finally, as in Syna the output of the LSTM is feed to a 1-fully-connected layer with a softmax function as output for the emotion prediction.

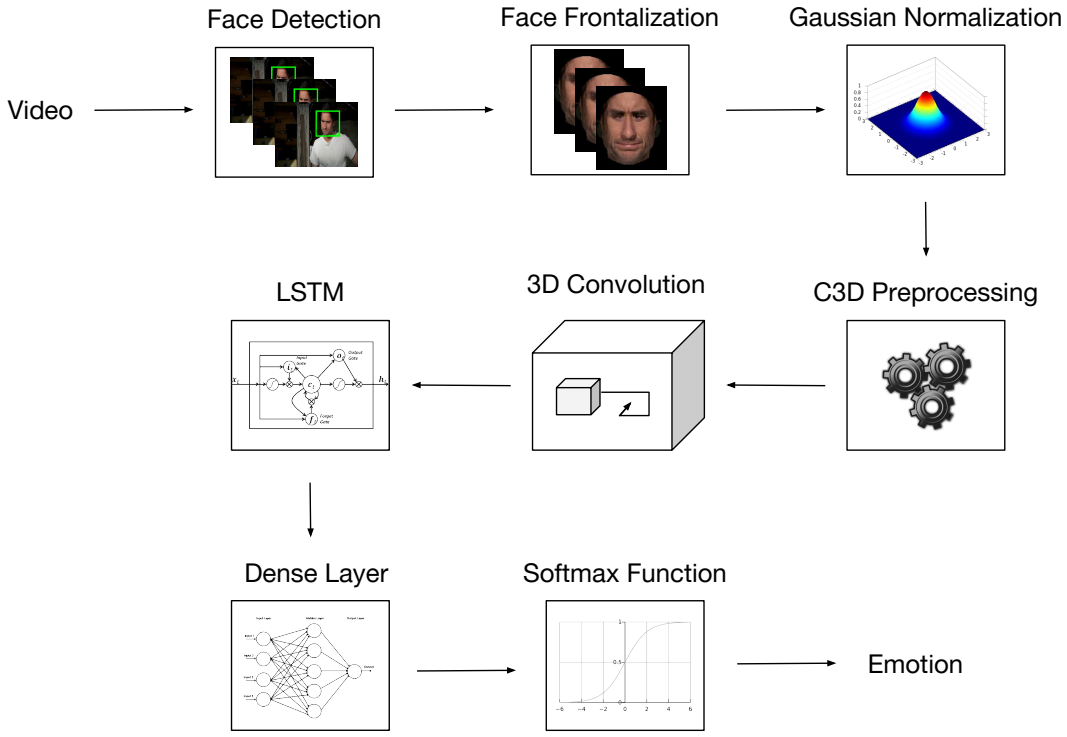

Figure 5.6: System diagram of DeepSyna.

Chapter 6
6 Experiments

6.1 Emotion Datasets

For the experiments of this thesis, the data used for training the systems need a set of prerequisites. The features in the data need to be 4-dimensional (3 dimensions from RGB images and the fourth being time). The labels must model the Categorical Model of Affect [36] representing the 7 basic emotions (happiness, sadness, fear, anger, contempt, disgust and surprise) plus a neutral expression, with a single label per video.

According to what was described in Chapter 3, some modules of the pipeline are pre-trained using other datasets. For the fairness of these experiments, I certify that no other emotion-labeled datasets have been used apart from the ones described in this section. In total, two different datasets were gathered and are described below.

6.1.1 Extended Cohn-Kanade Dataset

The Extended Cohn-Kanade Dataset (CK+) [11] is used as first step to determine whether the hypothesized system is feasible for the task of emotion recognition. This is a dataset that is setup in a strongly constrained environment, with the subject facing the camera and emotions being shown from neutral (first frame) to the highest peak of the emotion (last frame). There is a total of 327 videos[1] and the data is divided into 1-6 sequences from 123 subjects (university students) having the age range 18-30 and most of them being females.

[1] In reality, there are 593 videos, but only 327 are labelled with emotions.

Figure 6.1: Samples extracted from CK+ database [11].

6.1.2 Acted Facial Expressions in the Wild database 6.0

The Acted Facial Expressions in the Wild database (AFEW) 6.0 is the dataset involved in the EmotiW [12] challenge. It contains audio and video short clips labeled using a semi-automatic approach defined in [79]. As its name indicates, the feed represents an *in-the-wild* setting, with extreme head-pose and lighting conditions.

The huge variance introduced in this dataset makes most models easily overfit. When exploring winning solutions of previous years [76, 80], they all use additional datasets for pre-training the models. Within this thesis, AFEW is used to determine how feasible are Syna and DeepSyna in an environment of extreme visual difficulty.

Figure 6.2: Samples extracted from AFEW database [12].

Table 6.1: Attributes of AFEW database [79].

Attribute	Description
Length of sequences	300-5400 ms
No. of sequences	957
No. of annotators	1
Expression classes	Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise
Total No. of expressions	1259 (some seq. have multiple subjects)
Video format	AVI

6.2 Training

The model validation technique used in this thesis is Stratified K-fold Cross-validation for CK+ and train/test split in AFEW. Stratified K-fold in particular overcomes the differences in size of data depending on the displayed emotion. The used loss-function is the categorical cross entropy with an Adam optimizer [81] for learning parameters.

In reference to hyperparameter settings, Bayesian Global Optimization [82] is used as a mean to automate the entire tunning process. The hyperparameters required in the models are the number of hidden LSTM units, learning rate, learning rate decay and number of training iterations. In the following section, this process is described in detail.

Automatic Hyperparameter Tuning: Bayesian Global Optimization

There is particular interest in automatic approaches that can optimize the hyper-parameters to the problem at hand. A good choice for achieving this is Bayesian optimization [83], which has been shown to outperform other state-of-the-art techniques [84].

In recent years, Jasper *et al.* [82] explain how to practically apply Bayesian Optimization methods for performing hyper-parameter tuning. This technique has been shown to be highly effective in different areas of machine learning. The results show significant speed-up with respect to other methods, and surpassed the state of the art in data sets such as CIFAR-10.

Optimization techniques tend to have the final goal of minimizing a given function $f(x)$, but what makes Bayesian optimization different is the fact that it constructs a probabilistic model of $f(x)$ while taking into account the uncertainty. The optimization usually works by assuming an unknown function sampled from a Gaussian process and maintains a posterior distribution of this function as observations are made.

A Gaussian process (GP) is a prior distribution of functions in the form $f : X \to \mathbb{R}$. A GP is defined by the property that any finite set of N points $\{x_n \in X\}_{n=1}^{N}$ induces a multivariate Gaussian distribution on $\mathbb{R}^N$. These models assume that similar inputs give similar outputs. For completeness, Ramussen and Williams [85] provide an explanation in more detail about GPs.

In the case of hyper-parameter optimization, the hyper-parameters are chosen such that the improvement with respect to the best setting seen so far is big. The process can be formalized in a set of steps, and understood easier with a single hyper-parameter (See Figure 6.3).

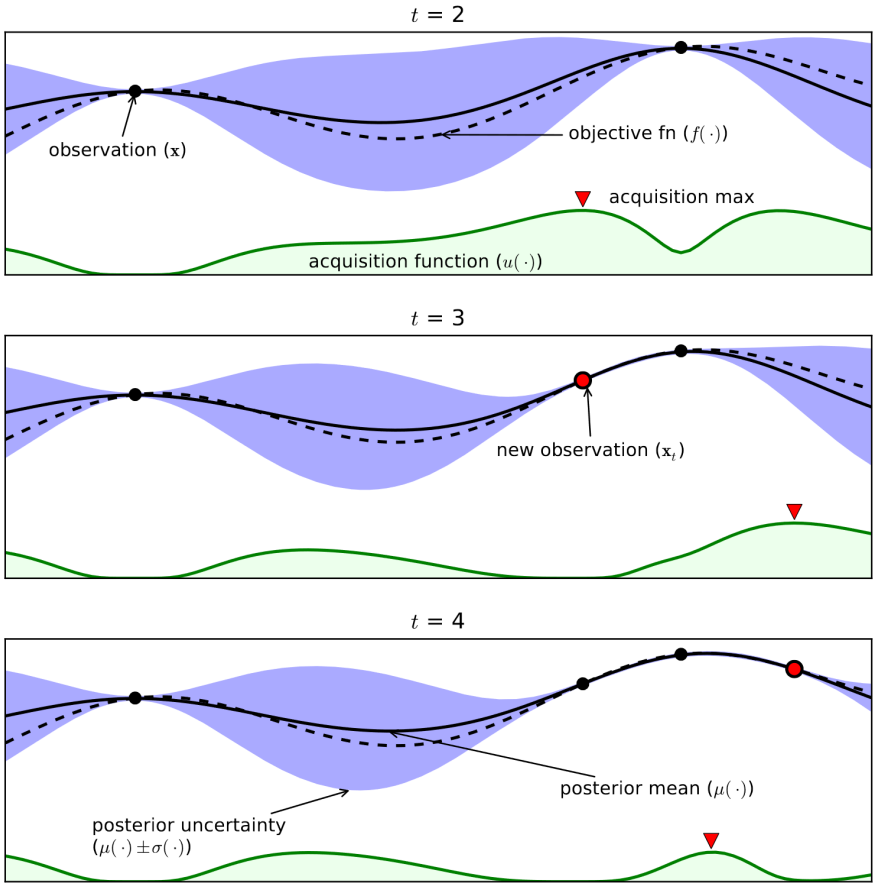

Figure 6.3: Illustration of the Bayesian optimization procedure over three iterations. The plots show the mean and confidence intervals estimated with a probabilistic model of the objective function. Although the objective function is shown, in practice, it is unknown. The plots also show the acquisition functions in the lower shaded plots. The acquisition is high where the model predicts a high objective (exploitation) and where the prediction uncertainty is high (exploration). Note that the area on the far left remains unsampled, as while it has high uncertainty, it is correctly predicted to offer little improvement over the highest observation [13].

First, an objective function and an acquisition function are defined. The objective function $f : X \to \mathbb{R}$ will react to the hyper-parameter setting and it is consid-

ered to be Gaussian distributed. The acquisition function $a : \mathcal{X} \to \mathbb{R}^+$ determines the strategy to maximize the probability of improving over the best current value. Expected Improvement (EI) has been shown to be better-behaved than other metrics.

$$a_{EI}(\mathbf{x}\,;\,\{\mathbf{x}_n, y_n\}, \theta) = \sigma(\mathbf{x}\,;\,\{\mathbf{x}_n, y_n\}, \theta)\,(\gamma(\mathbf{x})\,\Phi(\gamma(\mathbf{x})) + \mathcal{N}(\gamma(\mathbf{x})\,;\,0, 1)) \tag{6.1}$$

Second, a set of points are sampled from the objective function and an update is applied to fit the GP. As it can be seen in Figure 6.3, the variance of the GP decreases around the known points, acting as the uncertainty measurement. On the other hand, the EI decreases in those points.

Third, the point with the highest EI is selected. The objective function is applied and the GP is updated, repeating the process with certain convergence tolerance and then returning the best solution. There has been recent research on how to determine this convergence rate [86], and in general different settings in negative factors of 10 (10^{-n}) plus a given evaluation timeout results most practical.

Chapter 7
Results

This chapter shows the results of Syna and DeepSyna on the two datasets. In Sections 7.1 and 7.2, the reader can find detailed results for the training and prediction processes on both datasets. All the results shown here can be reproduced by using the code available at Syna's official Github repository [87].

7.1 Extended Cohn-Kanade Dataset

Table 7.1 describes the results of applying Bayesian Global Optimization to the model's hyperparameter space. These hyperparemeter settings are used to describe the results achieved on each system.

Table 7.1: Results of applying Bayesian Global Optimization to the hyperparameters' space using the CK+ dataset with 1-layer LSTM.

System / Best hyperparemeters	LSTM units	learning rate	decay	epochs
Syna (normalized landmarks)	40	0.0005	0	100
Syna (AU occurrence)	40	0.0005	0	100
Syna (AU intensity)	40	0.0005	0	100
DeepSyna	200	0.0005	0.0001	60

Table 7.2 shows the comparison of Syna and DeepSyna with other techniques, primarily a baseline and other state-of-the-art methods. It is worth mentioning that there are other techniques that achieve a higher level of accuracy but they rely on still images, exploiting the fact that the highest peak of an emotion in CK+ is in the last frame. Thus, these techniques are considered off-topic and are not mentioned.

The first version of the system errs from the fact that landmarks are not the most suited features to perform emotion recognition, and the system shows an overall accuracy of **79%**.

Table 7.2: The comparisons on video face based emotion dataset CK+.

Method	10-fold cross-validation accuracy (%)
3D SIFT [88]	81.4
3DCNN [89]	85.9
3D HOG [90]	91.4
3DCNN-DAP [89]	92.4
STRNN [91]	95.4
DeepSyna	**51.1**
Syna (normalized landmarks)	**79**
Syna (AU occurrence)	**87.5**
Syna (AU intensity)	**91.3**

When it comes to AU occurrence, the model shows a test loss (and confidence interval) of 0.62 (+/- 0.30), and test accuracy (and confidence interval) of **87.49%** (+/- 5.7 %). These values show that the emotions are well modelled and there is low variance in the classification performance. Figure 7.1 depicts the loss and accuracy curves when training over the number of epochs. It can be seen that both train loss and accuracy reach optimal half-way through the epochs, while the test loss tends to overfit after the 15th epoch. Nevertheless, the accuracy stays more or less stable with a noticeable amount of variance. This can be attributed to the use of Dropout as regularization technique, which works significantly better when compared to other techniques (e.g. Holdout). Syna using AU intensity estimates is the best model for the CK+ dataset. It achieves an overall accuracy of **91.3%**. The learning statistics are very similar to the ones shown when using AU occurrence.

Table 7.3: Statistics on the performance of AU occurrence model tested on the CK+ dataset.

	precision	recall	f1-score	support
Angry	0.79	0.69	0.74	45
Contempt	0.75	0.83	0.79	18
Disgust	0.82	0.86	0.84	59
Fear	0.87	0.80	0.83	25
Happy	0.93	0.99	0.96	69
Sad	0.81	0.75	0.78	28
Surprise	0.95	0.96	0.96	83
Average/Total	0.87	0.87	0.87	327

Finally, DeepSyna shows the lowest accuracy[1] (**51.1%**), which is still impressive if one takes into account the depth of the model, the fact that it is pre-trained in a

[1]It is worth mentioning that my machine exhausted from memory swapping during Bayesian

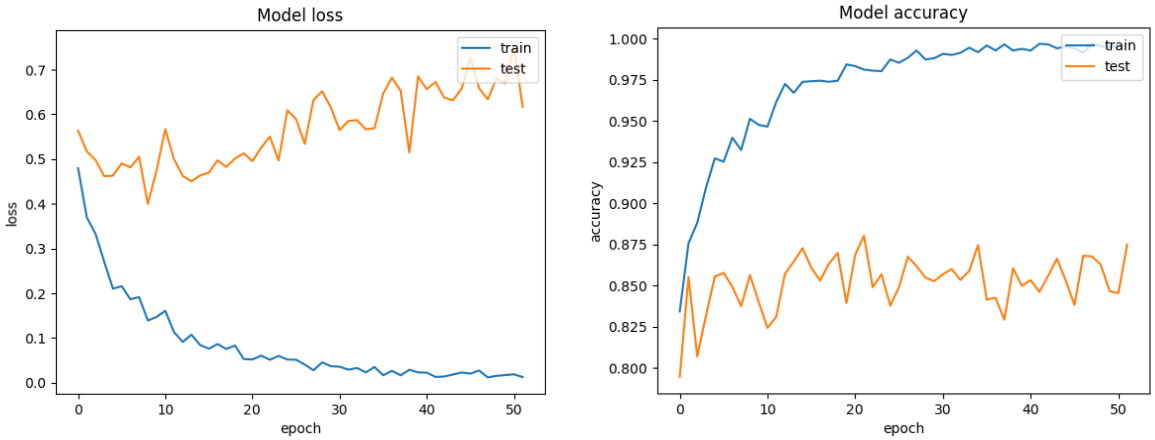

Figure 7.1: Loss (left) and accuracy (right) graphs over consequent epochs for the AU occurrence model tested on the CK+ dataset.

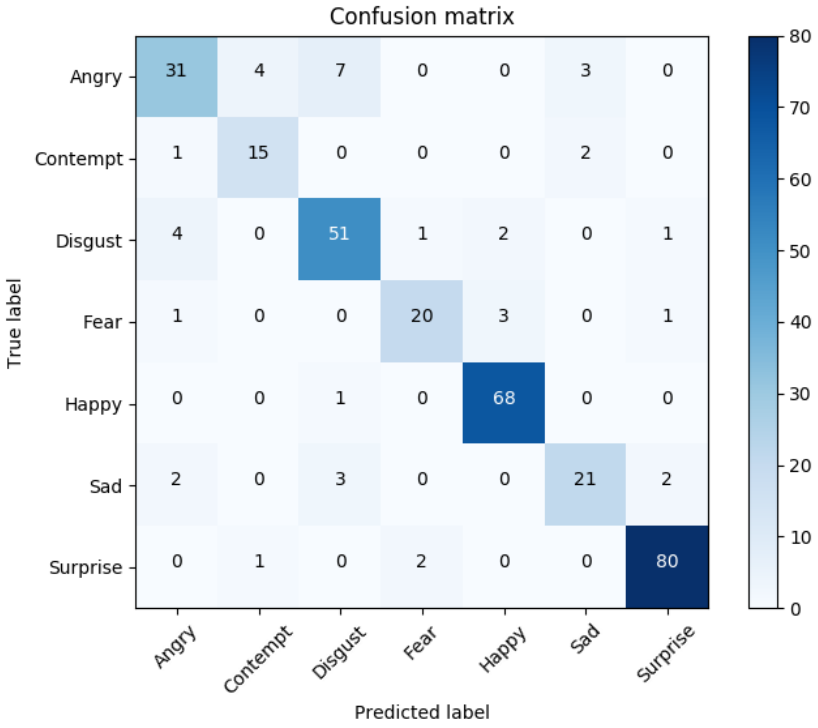

Figure 7.2: Confusion Matrix for the AU occurrence model tested on the CK+ dataset.

Optimization. Due to the fact that the whole process was halted midway through and took two days of computation, I decided to simply go with the best result after 15 iterations, which provides the shown accuracy. There is a possibility to obtain a better accuracy if the whole process is re-executed in a better machine.

complete different task, and the small amount of emotion-labeled data present in CK+.

7.2 Acted Facial Expressions in the Wild database

When it comes to emotion recognition *in the wild* environments, neither Syna nor DeepSyna achieve results comparable with state-of-the-art. One of Syna variants does compete with the baselines of the dataset, while DeepSyna fails to handle the high variance of the datasets given that it is exclusively trained on AFEW. Table 7.4 shows the results of applying Bayesian Global Optimization to the model's hyperparameter space. Using these hyperparemeters, the following paragraphs describe the results of each system in detail.

Table 7.4: Results of applying Bayesian Global Optimization to the hyperparameters' space using the AFEW dataset with 1-layer LSTM.

System / Best hyperparemeters	LSTM units	learning rate	decay	epochs
Syna (normalized landmarks)	71	0.003	9e-05	14
Syna (AU occurrence)	100	0.001	8e-5	100
Syna (AU intensity)	100	0.001	5e-5	100
DeepSyna	74	0.0026	0.0001	89

It is worth mentioning that in the case of AFEW dataset, there is the added feature of sound. This was not explored in this thesis, which makes the comparison with other techniques a little bit trickier. The current state-of-the-art for AFEW is Fan *et al.* [76], with an accuracy of 59.02%. The model is pre-trained using large image datasets such as FER2013[92] and also using audio features. All baselines shown in Table 7.5 share the usage of only video feed.

Table 7.5: The comparisons on video face based emotion dataset AFEW.

Method	Validation accuracy(%)
MoPs [93] (video only, AFEW 3.0)	27.5
MoPs [93] (video only, AFEW 4.0)	33.2
LBP-TOP [94] (video only, AFEW 5.0)	36.1
LBP-TOP [12] (video only, AFEW 5.0)	38.8
DeepSyna	**16.7**
Syna (normalized landmarks)	**21.1**
Syna (AU occurrence)	**26.6**
Syna (AU intensity)	**27.8**

As it happened with the CK+ dataset, landmarks are not optimal for emotion recognition. The accuracy of the model provides with just a **21.1%**. Using the AU occurrence, the model tends to overfit the data really fast, with a test loss (and confidence interval) of 2.09 (+/- 1.20) and accuracy (and confidence interval) of **26.6%** (+/- 44.20%). The high variance reflects how the learned features fail to model the shown emotion, given the high variance induced by the poor lighting condition and extreme pose variations. Figure 7.4 shows the confusion matrix, where it can be seen that the model tends to generalize the label of contempt (notice the vertical line in the predicted label-axis), and it performs best for the emotions of contempt and happiness. The AU intensity model is once again the best model, with an accuracy of **27.8%**. As it happened with AU occurrence, it over-generalizes the contempt emotion, and shares very similar statistics in training.

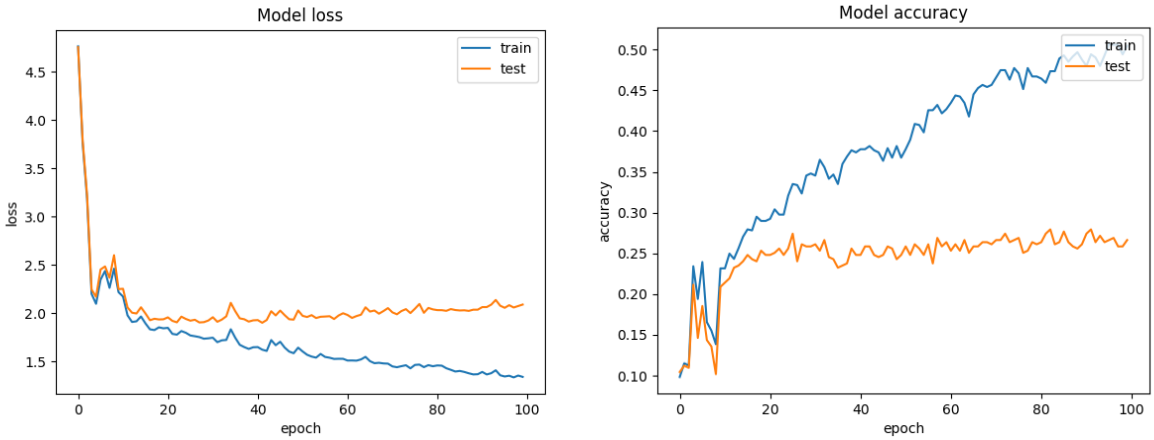

Figure 7.3: Loss (left) and accuracy (right) graphs over consequent epochs for the AU occurrence model tested on the AFEW dataset.

Table 7.6: Statistics on the performance of AU occurrence model tested on the AFEW dataset.

	precision	recall	f1-score	support
Angry	0.28	0.27	0.28	63
Contempt	0.21	0.62	0.32	64
Disgust	0.11	0.03	0.04	40
Fear	0.18	0.04	0.07	46
Happy	0.43	0.59	0.50	63
Sad	0.23	0.08	0.12	61
Surprise	0.00	0.00	0.00	46
Average/Total	0.22	0.27	0.21	383

The high variance in the feed prevents DeepSyna from properly learning any emotions. In addition to the shortcomings of learning the spatio-temporal features, the face frontalizer has to also cope with *in-the-wild* cases, which results in very poor data

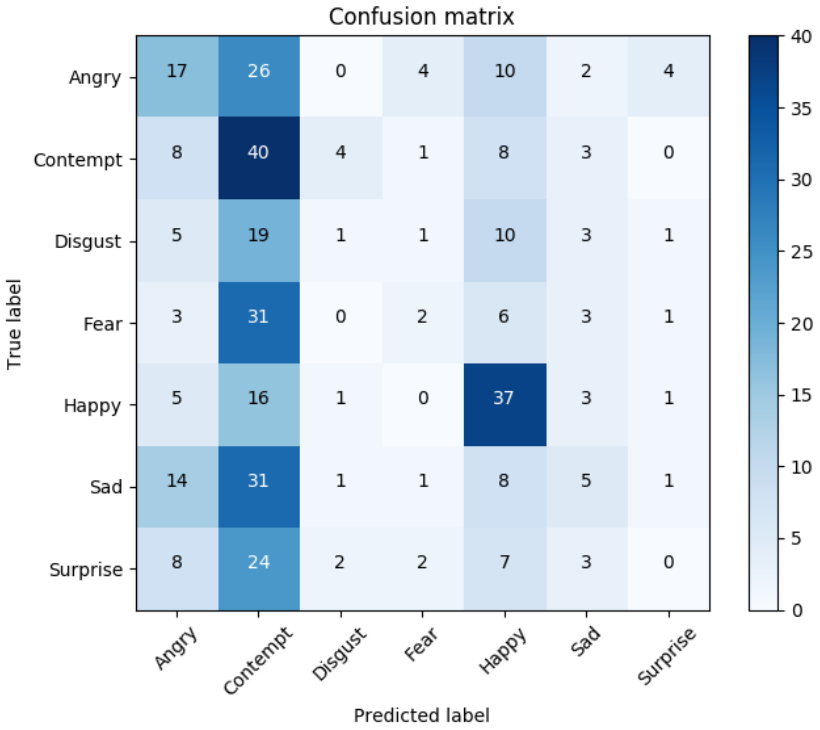

Figure 7.4: Confusion Matrix for the AU occurrence model tested on the AFEW dataset.

for intermediate features. It generalizes the contempt emotion even more, giving it as prediction about 95% of the time. Overall, the accuracy of DeepSyna is **16.7%**, which makes it too close to just random guessing (12.5%) and not comparable with the baselines. This will be further discussed in Chapter 8.

Chapter 8
8 Discussion

8.1 Are Syna and DeepSyna worth exploring in further research?

When I began my thesis, the main goal was to develop a system that could compete with state-of-the-art techniques in a variety of settings. The idea behind the implementation of Syna was to transform a state-of-the-art technique for single-frame facial expression analysis (CLNF) into state-of-the-art solution for video emotion recognition. In the case of DeepSyna, the premise was to use transfer learning from C3D, which already showed good transferability properties for other tasks [76, 95, 96]. Unfortunately, both approaches failed to beat current state-of-the-art techniques in terms of accuracy.

Given the results of this thesis, DeepSyna seems like a very poor solution for the task at hand. The data used for training C3D (sports) is too different for transferability to the target domain (emotions). It was a mistake to directly use the trained model without fine-tuning, as it has been done in the current state-of-the-art from EmotiW 2016 [76]. A reduction in the number of layers of the 3D CNN model, data-augmentation by image-rotation, and propagation of the gradients learned by training on emotion recognition to the C3D network seem like solutions for improving the performance.

Syna is still likely to increase in performance if its modules do so. Currently, there are better techniques that improve face detection [97], landmark estimation [98] and temporal dependencies [99]. The incorporation of these methods into Syna will most likely improve the performance, but the dependency on AU recognition is still present. Overall, the results of Syna on CK+ show that there is a high correlation between emotions and AUs. Nevertheless, this step is no better than end-to-end training on emotion recognition unless the AU prediction is 100% accurate and AU-emotion correlation is total. It is important to notice that the AU recognition technique relies on Histogram of Oriented Gradients, which might be considered obsolete in favor of

CNNs in a wide variety of computer vision tasks (even more in variable pose and light conditions, as it happens in the AFEW dataset).

8.2 Lack of Research Premises

The goal of trying to compete with state-of-the-art with a new approach is too simplistic for a master thesis. There is no clear explanation of why I used the methods that I used, or what is the contribution to research of my thesis. An instance of such is the rationale of DeepSyna, which is testing whether transferring a system which has been successful in other tasks can be also successful in emotion recognition. In specific, using C3D without fine-tuning makes no sense, since the dataset it is trained on (Sports-1M) lacks facial footage. Instead, an alternative could have been to use a pre-trained model on similar data, such as VGG-Face [100]. The main reason for not doing so is because the alternative of VGG + LSTM has already been explored in emotion recognition [76].

8.3 Unbalanced Level of Detail

The reader might question about the reason of the unbalanced level of detail among certain sections. On the one hand, I tried to provide a lot of detail on sections that are not mainstream in the field of ML. This means that I put a lot more effort in the detailed explanation of topics such as the study of emotions, CLM and CLNF as particular techniques. On the other hand, I skipped theoretical explanations of widely known topics in ML. These include DNNs, CNNs and RNNs, among others. For completeness, in most cases the reader is provided with sources for further content in case of necessity.

8.4 Doubtful Experiments

In the results of my thesis, I use CK+ as parts of my experiments. The reality is that CLNF is already trained on that dataset, so it might be interpreted as cheating. The difference lies in the learned labels. In the case of CLNF, CK+ is used for AU occurrence detection and intensity estimation, while Syna uses it for emotion recognition. To some extent, this is a case of transfer learning, in which I use the learned features from one task for a different, but related, problem. Nevertheless, the fact that the input features (videos) are the same, I can argue that the results of Syna might be inconclusive. This could be prevented by using the train-test split end-to-end, also by training CLNF together with the later stages in the pipeline.

8.5 Poor performance

Both CLNF and C3D performed decently in the task of facial feature extraction with the CK+ datasets. Nevertheless, the same cannot be said about the AFEW dataset. After achieving the accuracy that can be seen in Chapter 7, I extracted the intermediate features to take a look at what could have gone wrong. There are way too many instances in which CLNF fails to capture the facial landmarks that are required for further facial processing in the pipeline (see Figure 8.1). This wrong behavior includes cases in which the face is not even detected, the landmarks for the chin are mistaken by the landmarks of the mouth, and over-frontalizing the landmarks. Additionally, the face frontalization algorithm suffers from the same issue, given that the frontalization algorithm also relies in landmark estimation (see Figure 8.2).

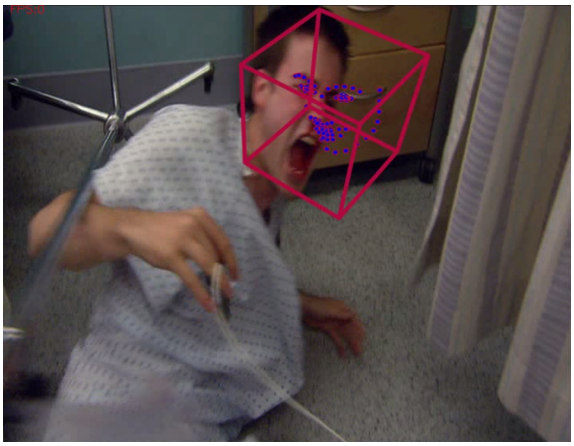

Figure 8.1: Instance of wrong landmark estimation from CLNF in AFEW dataset.

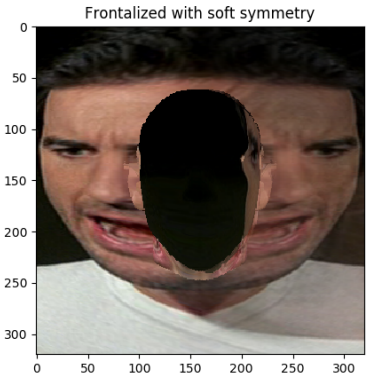

Figure 8.2: Instance of incorrect behavior for the face frontalization in AFEW dataset.

Chapter 9
9 Conclusion

In this thesis, I presented Syna and DeepSyna, two new approaches for the task of emotion recognition in video feed by learning both spatial and temporal features. The two approaches share the same architecture design, with two main components: a facial feature extractor and a temporal classifier. In Syna, the feature extraction is based on facial expression recognition through CLNF. In DeepSyna, the features merge both spatial and short-term temporal information encoded within adjacent frames, capturing facial motion features. These features are later fed to a temporal classifier that provides a classification for the entire video. The results achieved are close but non-competitive with state-of-the-art techniques in the CK+ dataset, and fall behind the baselines in the AFEW dataset. The most relevant contribution to research is the transformation of other state-of-the-art techniques into techniques applicable to random sequence lengths. The shortcomings of the models can be overcome by either pre-training the network with datasets related to other similar tasks or by augmenting the data feed to the learning algorithms.

9.1 Contributions

There are several contributions as outcome of this thesis. First, there is the explored two new approaches in research of spatio-temporal features for emotion recognition provided in this thesis. Second, the models along with the entire well-documented codebase have been made available online as Open Source [87]. Third, also as Open Source software I provide several Python toolkits for emotion datasets (CK+, BP4D-Spontaneous and AFEW), which makes handling and managing these datasets much easier. Fourth and last, I made small contributions to OpenFace and the adaptation of the face frontalization in Python used in DeepSyna.

9.2 Results

Even though the results of this thesis do not outperform state-of-the-art, there has been a huge evolution through the months. When I began my thesis, Syna was providing an accuracy of 45% in CK+. Through finding ways to make the data more usable, I managed to elevate those results to a 79%. Later, by doing tricks (video format used, data cleaning, various normalization trials, only using successful outputs from CLNF, applying Bayesian optimization, and training mini-batches, among others) I achieved the performance shown in the thesis. The same process applies to the usage of C3D, going from a 15% up to a 51%. All this evolution has been where I spent most of my time implementing the thesis, since ML was not my specialty when I started, but it became so at the end.

The fact that C3D performs so poorly is because of the limited amount of video footage that is present emotion recognition datasets. In the case of CLNF, it is pre-trained using a large number of datasets, most of these being composed by single images. Unlike CLNF, C3D by construction needs video feeds, and a large quantity given the depth in the architecture. To further prove this point, the reader can be referred to certain papers in which the authors show the performance of their models when no pre-training is performed. One of these papers is Fan *et al.*'s [76], which happens to be, at the time of writing this thesis, state-of-the-art for the AFEW dataset. In the paper, the authors mention that without pre-training on facial expression datasets, their model cannot even reach a 20% validation accuracy. The lack of data as the cause for low accuracies has already been explained and shown through training graphs in Chapter 7.

9.3 Future Work

9.3.1 CE-CLM for landmark estimation

When I began my thesis in February of 2017, CLNF was state-of-the-art for landmark tracking in constrained environments. By May of 2017, Zadeh *et al.* published Convolutional Experts Constrained Local Model (CE-CLM) [98], which brings the advantages of neural networks to CLMs. CE-CLM outperforms competitive state-of-the-art baselines (including CLNF) for facial landmark detection by a large margin. Given the timing with this thesis, I kept using CLNF.

It takes a decent amount of domain knowledge on the specific problem at hand to understand the underlying theory of CLNFs. In my particular case, it took me about two weeks of literature review. I am a supporter of the simplicity principle, which can be described as *"the simplest explanation for some phenomenon is more likely to be accurate than more complicated explanations"*. In the case of CE-CLM, its

understanding only requires a small subset of the theory employed in CLMs and regression-based approaches, in which landmark detection is directly performed on appearance without the need for a shape model.

9.3.2 Pipeline integration

At the time of writing this thesis, the pipeline of the system based on CLNF for feature extraction is segmented in two separated processes. The implementation of CLNF is written in C++, while the temporal classifier is written in Python. This segmentation does not allow for the system to perform classification in real-time (e.g. through a webcam), which would be a really useful feature. Therefore, an appealing improvement would be an integration of both components into the same process pipeline.

9.3.3 More Data

When exploring the results of both variants of the system, I argued that the low performance in the case of the general approach was due to the depth of the network and the low quantities of data available for training. In total, I found 5 datasets that possessed emotion labels with video features, and only got access to two of these. If more data is made available, I expect that the performance of the system will increase, as it has been argued in Chapter 7.

Unexplored dataset: BP4D-Spontaneous BP4D-Spontaneous [101] is a less constrained alternative for facial expression analysis. It is originally proposed as a source for 3D video feed of spontaneous facial expressions for exploring 3D spatiotemporal features. Luckily, these videos have labels corresponding to emotions, which makes it a valuable asset to test how well the temporal information is captured in a model. This is especially challenging, since the 41 subjects are are recorded for a prolonged amount of time (roughly between one or two minutes). This makes the classification process difficult, since the emotions emerge and fall in various sections of the videos, and in some cases, unrelated facial expressions appear.

Because it is not the main purpose, there is no baseline on this dataset regarding the classification accuracy of the shown emotion. This dataset can be used as a mean to determine how feasible are Syna and DeepSyna in an environment of extreme temporal difficulty.

9.3.4 Data Augmentation

One of the approaches that has been argued to improve the accuracy, especially from DeepSyna, is to augment the available data. A straight-forward solution could be to apply random transformation to the visual footage (e.g. image rotation).

9.3.5 Model Pre-training

A different approach to solve the problem of data-scarcity is to pre-train the models in similar tasks. In addition to the emotion-labelled datasets, there are a variety of other datasets used for Facial Expression Recognition. These datasets can be included into the pipeline to pre-train the model to recognize a task that is inherently similar to emotion recognition, which might become a significant improvement for Deep Syna. In the particular case of Syna, pre-training over single-frame emotion datasets such as FERA 2015 [17] is also possible.

9.3.6 Emotional State Visualization

Because of time constraints and preference to other more important tasks, a visualization and webcam live prediction are missing in Syna. Using the headpose estimation calculated with the landmarks, one can create a visualization that pins an emotional state to the face of the user, something in the lines of Figure 9.1. This work is already on development and will be delivered in the next release of Syna.

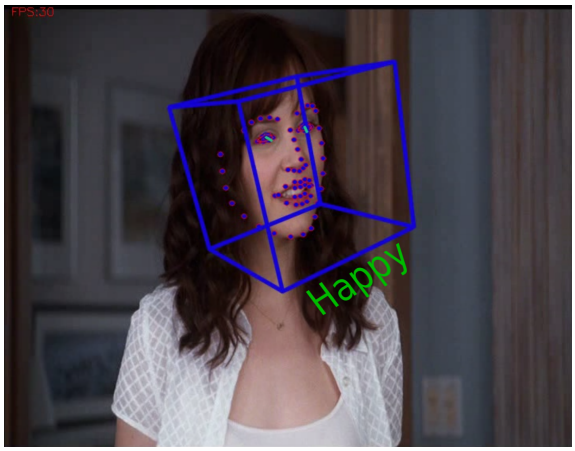

Figure 9.1: Visualization for emotional states in Syna.

References

[1] A. Singh, "'Pepper' the emotional robot, sells out within a minute." [Online]. Available: http://edition.cnn.com/2015/06/22/tech/pepper-robot-sold-out/

[2] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on.* IEEE, 2000, pp. 46–53.

[3] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 19–27.

[4] "Constrained local models," https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/Models/clm.html, (Accessed on 07/17/2017).

[5] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *BMVC*, vol. 1, no. 2, 2006, p. 3.

[6] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* IEEE, 2012, pp. 2610–2617.

[7] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 354–361.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] L. Sha, B. Chang, Z. Sui, and S. Li, "Reading and thinking: Re-read lstm unit for textual entailment recognition," in *COLING*, 2016.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.* IEEE, 2010, pp. 94–101.

[12] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 427–432.

[13] E. Brochu, V. M. Cora, and N. De Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *arXiv preprint arXiv:1012.2599*, 2010.

[14] W. B. Cannon, "Bodily changes in pain, hunger, fear and rage." 1929.

[15] E. A. Phelps, "Emotion and cognition: insights from studies of the human amygdala," *Annu. Rev. Psychol.*, vol. 57, pp. 27–53, 2006.

[16] J. Panksepp, *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004.

[17] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–8.

[18] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 921–926.

[19] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," *Affective Computing and Intelligent Interaction*, pp. 415–424, 2011.

[20] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 809–813.

[21] "Affectiva - emotion recognition software and analysis," https://www.affectiva.com/, (Accessed on 08/05/2017).

[22] "Artificial intelligence emotion recognition software | nviso," http://www.nviso.ch/, (Accessed on 08/05/2017).

[23] "Emovu emotion recognition software," http://emovu.com/e/, (Accessed on 08/05/2017).

[24] "Face recognition, emotion analysis & demographics | kairos," https://www.kairos.com/, (Accessed on 08/05/2017).

[25] V. Kumar and W. Reinartz, *Customer relationship management: Concept, strategy, and tools*. Springer Science & Business Media, 2012.

[26] G. T. Marx, "The surveillance society-the threat of 1984-style techniques," *Futurist*, vol. 19, no. 3, pp. 21–26, 1985.

[27] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Emotion recognition and its application in software engineering," in *Human System Interaction (HSI), 2013 The 6th International Conference on*. IEEE, 2013, pp. 532–539.

[28] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 354–361.

[29] A. Håkansson, "Portal of research methods and methodologies for research projects and degree projects," in *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013, p. 1.

[30] M. Herbert Benson and M. Z. Klipper, *The relaxation response*. Harper Collins, New York, 1992.

[31] D. H. Hockenbury and S. E. Hockenbury, *Discovering psychology*. Macmillan, 2010.

[32] D. Matsumoto and P. Ekman, "American-japanese cultural differences in intensity ratings of facial expressions of emotion," *Motivation and Emotion*, vol. 13, no. 2, pp. 143–157, 1989.

[33] P. Ekman, "The argument and evidence about universals in facial expres-sions," *Handbook of social psychophysiology*, pp. 143–164, 1989.

[34] ——, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique." 1994.

[35] ——, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[36] P. Ekman and H. Oster, "Facial expressions of emotion," *Annual review of psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[37] T. Dalgleish and M. Power, *Handbook of cognition and emotion*. John Wiley & Sons, 2000.

[38] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989.

[39] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[40] R. Highfield, R. Wiseman, and R. Jenkins, "How your looks betray your personality," *New Scientist*, vol. 201, no. 2695, pp. 28–32, 2009.

[41] C. D. O'Malley and J. B. d. C. M. Saunders, *Leonardo Da Vinci on the Human Body: The Anatomical, Physiological, and Embryological Drawings of Leonardo Da Vinci: with Translations, Emendations and a Biographical Introduction*. Gramercy Books, 1952.

[42] I. Loudon, "Sir charles bell and the anatomy of expression." *British medical journal (Clinical research ed.)*, vol. 285, no. 6357, p. 1794, 1982.

[43] G.-B. Duchenne and R. A. Cuthbertson, *The mechanism of human facial expression*. Cambridge university press, 1990.

[44] S. Williams, "Toward a Theology of Emotion," 2015. [Online]. Available: https://todahelohim.com/wp-content/uploads/2015/05/Toward-a-Theology-of-Emotion-Sam-Williams.pdf

[45] C. Darwin, P. Ekman, and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[46] J. M. Susskind, D. H. Lee, A. Cusi, R. Feiman, W. Grabski, and A. K. Anderson, "Expressing fear enhances sensory acquisition," *Nature neuroscience*, vol. 11, no. 7, pp. 843–850, 2008.

[47] M. Argyle, "Non-verbal communication in human social interaction." 1972.

[48] J. M. Susskind and A. K. Anderson, "Facial expression form and function," *Communicative & integrative biology*, vol. 1, no. 2, pp. 148–149, 2008.

[49] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[50] P. Ekman, W. V. Friesen, and J. C. Hager, "Facs investigator's guide," *A human face*, p. 96, 2002.

[51] P. Ekman, "FACS (Facial Action Coding System)." [Online]. Available: https://www.cs.cmu.edu/{~}face/facs.htm

[52] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[53] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[54] W. V. Friesen and P. Ekman, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983.

[55] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[56] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

[57] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[58] ——, "Max-margin object detection," *arXiv preprint arXiv:1502.00046*, 2015.

[59] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–6.

[60] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[61] Y. Wang, S. Lucey, and J. F. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[62] T. Baltrušaitis, "Automatic facial expression analysis," Ph.D. dissertation, University of Cambridge, 2014.

[63] J. Peng, L. Bo, and J. Xu, "Conditional neural fields," in *Advances in neural information processing systems*, 2009, pp. 1419–1427.

[64] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li, "Global ranking using continuous conditional random fields," in *Advances in neural information processing systems*, 2009, pp. 1281–1288.

[65] J. A. Hesch and S. I. Roumeliotis, "A direct least-squares (dls) method for pnp," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 383–390.

[66] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[67] D. Neth and A. M. Martinez, "Emotion perception in emotionless face images suggests a norm-based representation," *Journal of vision*, vol. 9, no. 1, pp. 5–5, 2009.

[68] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[69] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[70] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[71] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[72] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[73] D. Britz, "Recurrent neural networks tutorial, part 1–introduction to rnns," 2015.

[74] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[75] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, "Deep temporal appearance-geometry network for facial expression recognition," *arXiv preprint arXiv:1503.01532*, 2015.

[76] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.

[77] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Transactions on Nuclear Science*, vol. 44, no. 3, pp. 1464–1468, 1997.

[78] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[79] A. Dhall *et al.*, "Collecting large, richly annotated facial-expression databases from movies," 2012.

[80] A. Yao, J. Shao, N. Ma, and Y. Chen, "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*. ACM, 2015, pp. 451–458.

[81] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[82] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[83] J. Močkus, "On bayesian methods for seeking the extremum," in *Optimization Techniques IFIP Technical Conference*. Springer, 1975, pp. 400–404.

[84] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[85] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[86] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2879–2904, 2011.

[87] D. Shahrokhian, "Syna," 2017. [Online]. Available: https://github.com/dshahrokhian/syna

[88] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 357–360.

[89] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 143–157.

[90] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[91] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[92] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[93] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge (emotiw) challenge and workshop summary," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 371–372.

[94] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 423–426.

[95] A. Montes, A. Salvador, and X. Giro-i Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," *arXiv preprint arXiv:1608.08128*, 2016.

[96] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4207–4215.

[97] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[98] A. Zadeh, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017.

[99] K. Lee, O. Levy, and L. Zettlemoyer, "Recurrent additive networks," *arXiv preprint arXiv:1705.07393*, 2017.

[100] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[101] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.