Automatic data collection for facial expression recognition

Valentina Bollini

Automatic data collection for facial expression recognition

by

Valentina Bollini

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended on Tuesday September 4, 2018 at 13:00.

Student number:	4513665	
Project duration:	December, 2017 – September, 2018	
Thesis committee:	Prof. dr. H. Hung,	TU Delft, supervisor
	Prof. dr. ir. P. Jonker,	TU Delft, supervisor
	Dr. ir. C. Heemskerk,	HiT
	Prof. dr. ir. J. van den Dobbelsteen,	TU Delft

Abstract

Facial expression recognition today is a widely-researched topic with some pertinent applications in human-computer interaction (HCI), surveillance, etc. The diverse range of expressions makes data collection expensive in terms of time and money, making task-specific data collection inconvenient. This thesis project investigates a possible method for quickly and cheaply collecting this data, especially for a Human Computer Interaction (HCI) application. In particular, the focus is on *Pepper*, a robot meant for interacting with humans through conversation. For collecting such data, emotions should be triggered in participants and their faces should be video-recorded. The triggering method used was to show videos selected for triggering specific emotions, letting participants watch them in pairs, thus enabling mutual interactions to better simulate the social environment in which the robot will operate. After watching each video, participants were asked to rate their feelings through the *AffectButton*, a tool for intuitively describing emotions in a dimensional way. Video selection was based on a questionnaire in which people were asked to rate emotions a video was triggering in them. Recordings obtained from people watching the triggering video were then compared to a model of a neutral expression performed by the same participant, in order to select the frames in which expressions were shown, ignoring neutral and transition frames. The pictures obtained were included in a dataset, which is used for training a linear regressor and a Convolutional Neural Network (CNN). These were then tested on naturalistic data taken during conversations, in order to investigate whether the proposed data collection method could build a useful dataset and the results showed this method to be promising.

Acknowledgments

Firstly I would like to thank the whole graduation committee for giving their time on a day so important for me. I want to thank Hayley, for her guidance, support and tips: for being able to always point me in the right direction and to understand where I was struggling more. I want to thank Pieter, for putting time and effort in my project despite his retirement.

I want to thank Cock, for his patience and support. For welcoming me in his HiT family, for replying to questions not with plain answers, but giving me the tools I needed for reaching them myself. I want to thank everyone at HiT, for the friendly and helpful work environment. For the pleasant talks at lunch and for teaching me more and more about this beautiful country.

I want to thank all my friends for being there for me. A special thank goes to Petros and Salil, for solving my doubts and giving me the push I needed to overcome the most panicking moments. I want to thank all the Aan 't Verlaat crowd: my family in the Netherlands. You have always been there for me, just like a family, and you made these years unforgettable. For the dinners together, the nights spent talking in the corridors, every single smile and hug you donated me. I also want to thank all my other friends, listing all the people that have been important for me during these years would take pages and pages, so I will just limit to a collective thank you and a virtual hug.

Last but not least, I want to thank my real family, my mom, my dad and my brother, for their support from day 0. For supporting my choice of coming to the Netherlands, for the Skype calls in the middle of the night, for giving me the strength of going through even the hardest moments. If I reached up to here, it is mainly thank to you, for making me the strong woman you wanted me to be I also want to thank all my other friends, in Delft and back home, for their constant support, for cheering me and for always believing in me.

Contents

1
Introduction

Facial expression recognition is a growing field, mainly due to its wide range of applications like surveillance [1], mental state identification [28], education, aid for disability, safety (for example detecting tiredness and distraction in drivers) [39], lie detection [14], etc. Facial expression recognition refers to the techniques which enable a computer to automatically recognize human emotions via taking their facial data as input.

1.1. Motivation

Together with the possible applications for facial expression recognition already indicated, we can mention human-computer interaction (HCI). HCI includes all the methods we have in which we interact with computers. Researchers are studying different kinds of interfaces trying to make the communication more and more natural and spontaneous.

In face-to-face human communication facial expression delivers 55% of the message (in contrast with 38% of vocal tone and just 7% for spoken words) [33]. Considering this, it is easy to comprehend why facial expression recognition can play a big role also when interacting with a computer.

In this particular case, the research has been carried out at Heemskerk Innovative Technology (HiT) on Pepper: a robot designed for interactions with humans that will be introduced in hospitality and health-care environments. At the moment Pepper has a weak facial expression recognition algorithm that is able to distinguish between positive, negative, neutral and unknown emotions, which is much less detailed than what we would expect to be recognized in a conversation between humans.

1.2. Problem statement

Facial expression recognition is far away from being perfect. Many challenges still have to be solved mainly to increase its robustness. Some of them are listed below [26] [25]:

- Fast and **reliable in-the-wild** performances (using videos recorded in uncontrolled environments) are really difficult to achieve because most of the datasets used for training algorithms have been collected in a controlled environment (same point of view, same light condition etc). A second main drawback of many datasets is that they collect acted expressions which can result in being very different from spontaneous ones. Moreover, algorithm designers tend to optimize their performances on datasets they are testing on, leading to a dataset bias[26]: this means that algorithms that have a high accuracy on paper may result in terrible in-the-wild performances.

- There is a very **wide range** (up to 7000) **of** combinations of Action Units (**AUs**) and Action Descriptors (ADs) (sets of facial movements) that people can perform every day. These combinations will be different between individuals, which makes it even more difficult to have a dataset that is representative for everyone [26].

- A more fundamental challenge is **defining which emotions should be recognized**. Experts do not agree on how many emotions exist. Some say there are six [20], some say there are four, some say there are many more or that they are not identifiable [31]. This topic will be more exhaustively discussed in section 2.1.1. Different applications can require more precision on different emotions. For examples in a hospitality applications we might not need to precisely distinguish between fear and anger since we will probably require the intervention of an operator in case such an extremely negative emotion was detected. In some other application, such as security, both those emotions would be extremely important while we might not be that interested in milder emotions such as boredom or sadness.

- Datasets' **labels** can be inherently **subjective** since different annotators may have different opinions on which emotion the subject is expressing. Ambiguity in labeling is also a problem regarding time continuity [26] and frames corresponding to transitions between different expressions are those that are most affected by subjectivity. An other labeling ambiguity involves situations in which people fake an expression. It can be argued that the label should follow the displayed expression or the real feeling of the subject.

Many of these improvement areas for facial expression recognition are dataset related.
In his work [26] Kawulock suggests creating task specific datasets. A solution like this will face big cost related issues: manually collecting and labeling data requires many working hours of specialized operators that would cost time and money. This leads to our problem statement: **how to optimize the dataset collection process to improve the efficiency and reduce the process duration?**.

1.3. Research question, goal and requirements

As explained in section 1.2 creating personalized datasets would be inconvenient most of the times, unless the collection of the dataset could be done almost completely automatically. The goal of this thesis would be the **development of a method for a fast and unsupervised collection of labeled dataset for facial expression recognition**. A possible solution for automatically building datasets would be triggering emotions. Knowing the emotion we triggered using a certain method, we could automatically label the video frames, cutting on the time needed for labeling. For the recognition to perform better, the data contained in the dataset should be as similar as possible to those that will be encountered in the real application. This means that the emotions triggered during the dataset collection should be the same as those that then will have to be recognized by the algorithm. In literature, different triggering methods are described and they can grouped in the following categories [17]:

- **Through simulated conversations**: Data are collected from subjects interacting with an Artificial Intelligence (AI). A relevant AI used for this scope is SAL, the same method that has been used for collecting SEMAINE dataset. Subjects were having conversations with SAL who was replying based on different personalities triggering this way emotions in its interlocutor. A wide variety of emotions could be collected, including not very intense emotions [18].

- **Through experience**: Subjects are exposed to different situations designed for triggering specific emotions. More and less controlled environments have been used for this scope: from outdoor activities to tasks performed in a simulator. This is the method that has been used by BP4D dataset.

- **Showing videos**: Data are collected recording participants when viewing videos selected for triggering certain emotions[29]. This method has been used by DISFA dataset.

The first two methods, have the capability of better replicating situations in which a certain algorithm could be used. Their drawback is that they require a lot of surveillance and manual labeling, which is not going to solve the problem of data collection being too expensive. For this reason, the choice of drawing more attention to video triggering has been made, in order to determine whether it can be an effective triggering method for reproducing emotions felt during conversations.
This leads us to the research question:
Can we collect a useful dataset triggering emotions through videos?.
A problem related to this method would be that emotions felt in a passive situation (watching a video) would be different from those felt in a social environment such as during a conversation. This problem could be solved showing the triggering videos to more people at the same time and having them interacting with each other: this way the environment would become social and the triggered emotions could be more similar to those felt during a conversation. First, it will be explained how this would solve the challenges explained in section 1.2 and then the requirement such a method should have will be described.

1.3.1. Advantages of a fast self-building dataset

Before focusing on the requirements of such a system, the reasons of its usefulness should be explained. A customized dataset, for each application, would contribute on addressing the following challenges:

In the wild performance: being able to collect a dataset in the same environment in which the facial expression recognition will be performed, would in most of the cases result in a very similar distribution of collected data and real data. In our application, the algorithm would be performed on Pepper, that is located always in the same indoor environment and approximately having the same point of view. If it could collect its own dataset, it would be consistent with the distribution of the data it would encounter in its application.

Wide rage of AUs: With a fast way of collecting data, if the algorithm is expected to be encountering the same individuals more than once, it would be quick and cheap to build a specific classifier for each subject. In our application, we expect

Pepper to meet the same people more than once. If it was able to recognize when it meets a person it already collected data from, it could build a classifier with those data. Otherwise, in case a new person is met, all the collected data should be use on a person independent classifier.

Defining which emotions should be recognized: given that the emotions definition can also be task dependent, having a easy way to form a classifier also makes it possible to select which emotions the user will actually need in their application.

Labels subjectivity: Triggering emotions could give the opportunity to ask directly to the subjects what emotion they were feeling in a certain moment. This would both solve the problem related to labels subjectivity and cut on the amount of time required for labeling.

1.3.2. Requirements

For being usable in real applications, the algorithm for data collection should be fulfilling some basic requirements:

- The emotions that are being triggered, should be as similar as possible to those that will occur during the application.
- It should require minimum or zero intervention from operators for labeling and processing the videos.
- Participating to the data collection should not require particular abilities from the subject. The collection itself should be designed in such a way that no intervention from an operator is needed.

2

Background and related work

2.1. Background

Facial expression recognition is a growing field, mainly due to its wide range of applications like surveillance [1], mental state identification [28], education, aid for disability, safety (for example detecting tiredness and distraction in drivers) [39], lie detection [14], etc. Facial expression recognition refers to the techniques which enable a computer to automatically recognize human emotions via taking their facial data as input. Two fundamental choices characterize a facial expression recognition algorithm: which emotions should be recognized and how should this recognition happen. In the following sections (2.1.1 and 2.1.2) these two main choices will be described further.

2.1.1. Emotions definition

Defining which emotions are present and should be recognizable is not a trivial task. Two strategies for describing emotions can be distinguished: categorical and dimensional description [14].

> **Categorical description**: A categorical description defines a finite number of categories that represent some sort of building blocks for composing any possible emotions. An example of categorical description is shown in figure 2.1, in which six basic emotions are represented by anger, fear, surprise, sadness, joy and disgust. However, there is currently no agreement on the precise categorization of basic emotions. Some researchers claim there are six of them [20], some say there are four, some say there are many more or that they do not exist at all [31].
> While naming a emotion is intuitively easier for humans, a major limitation of this approach is that it considers emotions as discrete elements, without analyzing all the shades and levels they can have.
>
> **Dimensional description**: A dimensional description, on the other hand, defines a multidimensional space and describes emotions as coordinates in that space. Some examples of the dimensions are i)valence: how pleasant or unpleasant the feeling is, ii) activation: how likely is the person to take action during the emotional state, iii) control: how much in control of the emotion they are, iv) intensity of the emotion, v)dominance, etc. An example of a dimensional description is shown in figure 2.2

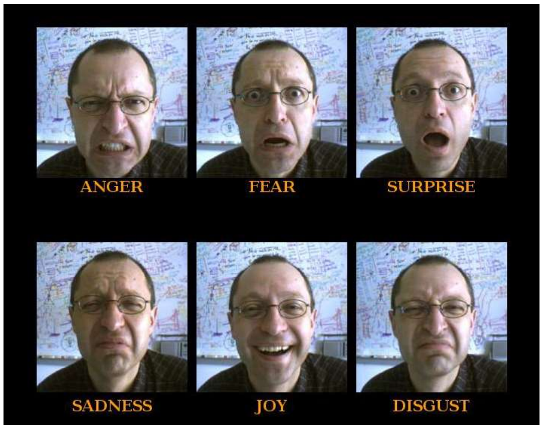


Figure 2.1: An example of categorical description of emotions: six basic emotions represented by anger, fear, surprise, sadness, joy and disgust [36].

The major advantage of this approach is that it considers all the shades in between emotions, having a continuous description instead of a discrete one. The main drawback is that such a description is not intuitive for us, since we are used to describe our emotions in a categorical way. This problem is partially solved by the *AffectButton* [7], a tool which is described in detail in the following section.

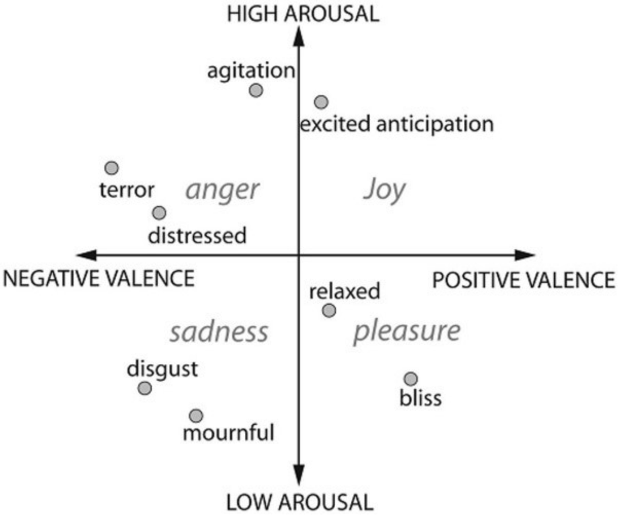


Figure 2.2: An example of dimensional description of emotions. This is not the same model that will be used throughout the thesis.

A third description model that can be mentioned is the **Plutchik wheel** [32]. It has been designed as a combination between a categorical and a dimensional description. The eight sectors represent eight primary emotions and the cone's vertical represents their intensities. This intensity dimension is an improvement from a categorical description because it includes different levels within the same primary emotion, but it still does not have the same continuity as a dimensional description. A figure of the Plutchik wheel can be found in figure 2.3 [32].

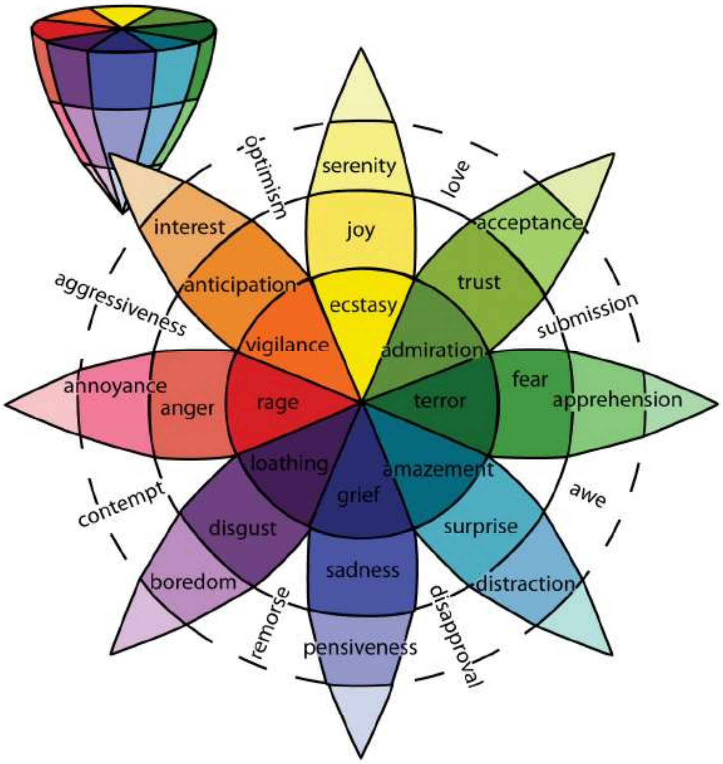


Figure 2.3: Plutchik wheel of emotions: an intermediate description model between categorical and dimensional description. The sectors represent eight basic emotion and the vertical their level of intensity [32].

The AffectButton

A dimensional description is made more intuitive by the AffectButton, designed by Broekens and Brinkman [7]. It uses a tri-dimensional model in which dimensions are represented by valence, dominance and arousal and their value can go from -1 to +1.

The AffectButton appears as a yellow emoticon in a squared frame as shown in figure 2.4. The vertical dimension represents the dominance (more dominant on top and less dominant on the bottom), while the horizontal dimension represents the valence (more pleasant on the right and less pleasant on the left). The user can move the mouse on the emoticon, that will change expression based on the cursor's coordinates. While dominance and valence are independent variables, arousal is computed as a result of valence and dominance. When clicking, the coordinates representing the expression shown by the emoticon are given. Examples of these expressions with the corresponding mouse positions are shown in figure 2.4.

The choice of having the third dimension (arousal) dependent on the other two has been made considering that a three-dimensional system is more difficult to use. Since most of the times high arousal is associated with extreme emotions, its independence can be dropped without losing critical information. The relation between arousal and the other two dimension is shown in figure 2.5. The area of AffectButton is divided in sectors and in each of them the arousal assumes a different behavior. In the central area, the arousal is really low (close to -1) Then, on the medium square, the Arousal goes from -1 at the inner border to +1 at the outer border. In the outer region, the Arousal is fixed at +1.

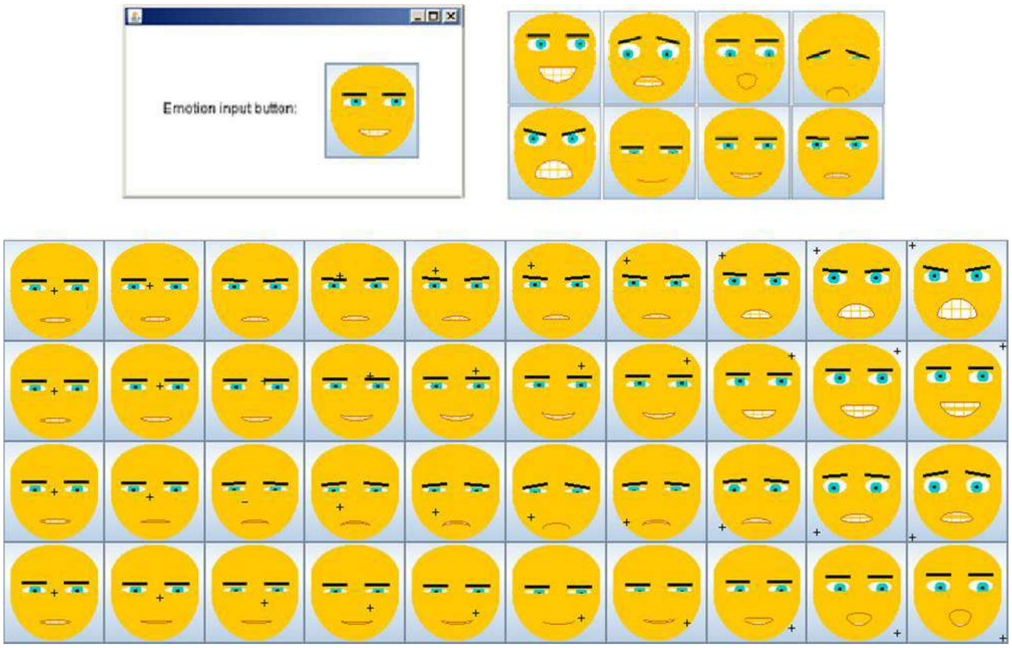


Figure 2.4: The affect button, showing different expression with the corresponding mouse position. Vertical movements of the mouse affect the dominance while horizontal movements affect the valence [7].

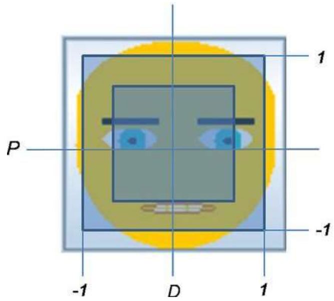


Figure 2.5: Division of the areas on the AffectButton regarding the behavior of the dimension arousal [7].

2.1.2. Recognition algorithm

Once the classification strategy is defined, the algorithm itself can be designed. Most of the facial expression recognition algorithms share a similar template shown in figure 2.6.

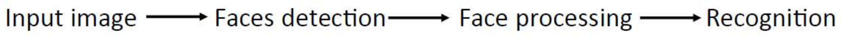


Figure 2.6: Pipeline of a facial expression recognition algorithm. First faces are detected from the input image. Then they are processed and fed into the recognition algorithm.

1. **Faces detection**: Pictures (or videos) included in a dataset contain images taken in different conditions: they might include the whole body or just the face,

might be taken from different points of view, faces might be partially occluded etc. As a consequence, faces need to be detected in order to be presented as input to the algorithm. The same raw image can contain more than one person: in this step all the faces are detected and afterwards considered as separate images.

2. **Face processing**: This step includes all the operations that are necessary to be performed in order to make an image usable by the algorithm. In other words, it takes care of transforming the image/video into an input data in the form that the algorithm is expecting to receive. This phase can include tasks such as landmarks or features detection.
   Landmarks are important points of the face that are easily detectable and whose movements are meaningful for showing emotions. Examples of these landmarks could be the angles of the eyes, the center point of upper and lower lid, mouth angles etc.
   Features are relevant elements of an image that carry an important amount of information for recognition. Different kinds of features can be distinguished. *Appearance features* encode pixels intensity information. In other words, they encode the appearance of the face, such as the presence of wrinkles. *Geometric features* encode relations between landmarks' positions. They are easily computed once the landmarks are detected and they are invariant to illumination.
   Features can be either manually extracted or learned. In case they are manually extracted, the algorithm is designed explicitly indicating which features will be detected. In case they are learned, an algorithm (such as an autoencoder) needs to be trained in order to learn which features would be relevant for the task.

3. **Machine learning**: Machine learning is the core of the algorithm: it is the part that actually trains a classifier which is able to distinguish different facial expressions. Training happens by means of a dataset: a collection of labeled images/videos serving as an example for the algorithm to learn how it is supposed to label unknown images. Many different machine learning algorithms are present in literature [39]. Some of them are usable exclusively in a categorical or in a dimensional domain, some others in both.
   It is possible to distinguish between person specific and generic models. Person specific models train a classifier for each person they encounter, while generic ones are trained using data from different people and can be used on unknown subjects. It is proved that person specific models are better than generic ones [34], but they need many images of every single person, so they are not usable in all of those situations in which strangers need to be analyzed or in which it is not possible to collect big amount of data from a specific person.

2.2. Related work

Different strategies have been followed in order to collect datasets. In this section both datasets meant for facial expression recognition and for AUs detection will be considered, since the methodology for collecting the data is quite similar.

The most quick and inexpensive way is just recording people when asking them to perform some facial expressions. Two widely used datasets that have been collected this way are the Extended Cohn-Kanade Dataset (CK+) and its earlier version CK. Smiles are an exception in CK+, since it happened during the test that the subjects

spontaneously smiled to the operator and those smiles have been recorded as data [27]. As already stated, this strategy is often not the most accurate one because acted emotions can be very different from spontaneous ones. Moreover, algorithms trained on acted datasets are less robust to face movements due to speech or blinking [41]. In order to overcome this, two main strategies have been applied: dataset collection through triggering emotions and dataset collection in the wild, labeling already existing videos.
Triggering emotions also partially solves the definition of the ground truth (related to the subjectivity of labels), if we assume that the subjects will react accordingly to the stimuli they received. Belonging to this category we find the following datasets.

> **BP4D dataset**: participants went through through tasks designed for triggering particular emotions (happiness, amusement, embarrassment, fear, physical pain, anger and disgust) [43].
>
> **SEMAINE dataset**: uses conversation with Sensitive Artificial Listener (SAL), which simulates 4 types of people (in terms of personality) the subject has to interact with. This dataset has been manually labeled for AUs [38].
>
> **DISFA dataset**: triggered emotions through videos, all participants were in the same controlled conditions and alone during the test. AUs have been manually labeled [29].

Various realistic datasets have also been collected, harvesting videos or images from the Internet. Most of them take data from a single source (that could be the news or web). These videos would be more similar to those that would be encountered in a real life application, but labels subjectivity can still be a problem since it might not be possible to ask the subjects which emotion they were feeling in a certain moment. This means the truthfulness of labeling is subjected to experts' skills. Belonging to this category we find:

> **QUT FER dataset**: joins three sources (TV drama, news and web) and labels intensities for six basic emotions (anger, fear, happiness, surprise, disgust, sadness) plus positive negative or neutral. Cases of a frame containing more basic emotions at the same time have also been considered [42].
>
> **Belfast naturalistic database**: scenes taken from TV, in which subjects are in a static pose, mostly sitting. Seven raters labeled each clip both categorically (on 16 categories) and dimensionally [16].

3
Triggering emotions

The choice of the emotions triggers is crucial because it determines the usability of the dataset. For this reason, also the triggers' selection is an important step of the project.

3.1. Questionnaire

Group emotional contagion can be studied in social environments. Having people watching videos together makes them a group and consequently subjected to emotional contagion. For this reason, videos' selection is crucial for finding stimuli that will trigger similar emotions to both the subject, otherwise they might influence each other and the given label could be not aligned with the emotion they actually showed. Triggering videos have taken from YouTube and then they have been tested on people through a questionnaire. 17 videos were included. People were shown the videos one by one and they were asked to rate the emotion they felt through the AffectButton. This means for each video a label was given, representing a point in a bi-dimensional space having as horizontal coordinate valence and as vertical coordinate dominance. According to the AffectButton's structure, coordinates could take values between -1 and 1. Videos' order was randomized, people were warned that some of them might have been disturbing and not everyone completed the survey. The questionnaire has been spread through my and other people's networks. 29 participants between 22 and 54 years old have been reached, coming from different countries. Based on the questionnaire's responses, 7 videos have been selected for becoming the triggers for building the dataset.

3.1.1. Selected videos

Videos have been selected considering two factors: minimizing the variance of the triggered emotions and covering as much as possible of the area of the AffectButton. Figure 3.1 shows for each proposed video the corresponding variances on valence and dominance. On the x axis we find the titles of the videos that where hidden to the participants not to give them expectations on what they would have seen.

Seven videos have been selected:

Baby: a video of a baby laughing when his father rips a piece of paper in front of him [9].

Candle: a toddler trying to blow his birthday candle [22].

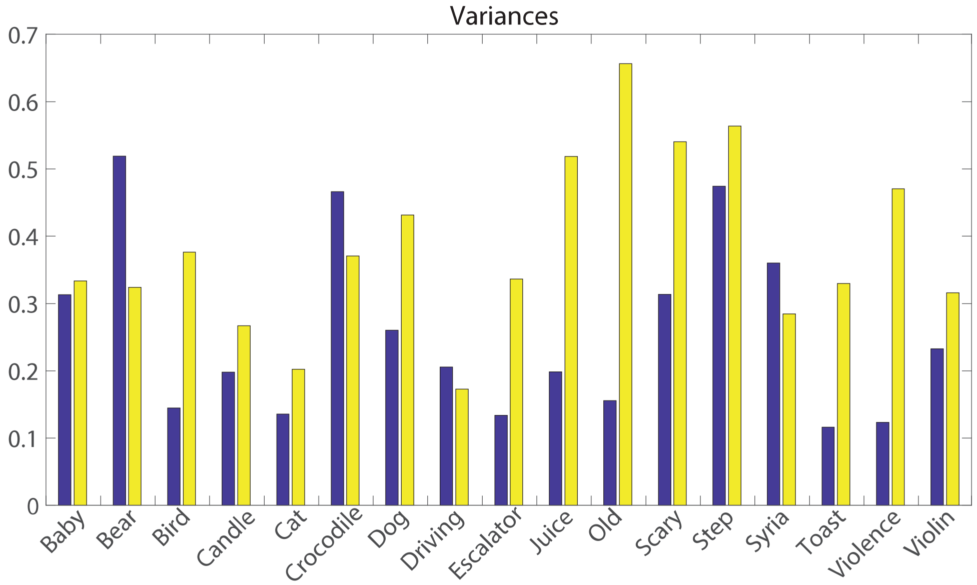


Figure 3.1: Videos showed in the questionnaire with the variances of the emotions they triggered. In blue we find variance on valence, in yellow on dominance.

Cat screen: a cat seeing a squirrel in a laptop's screen. When the squirrel goes out of the framing the cat searches for it behind the screen [23].

Driving: a video from a public service campaign for increasing awareness about safe driving. It shows examples of car crashes and what led to them [10].

Old woman: a very old woman (80 years old) dancing acrobatic salsa [8].

Syria: children in Syria, injured or dead from attacks [11].

Violence: a video taken from a security camera portraying an episode of domestic violence [12].

Since the video of the cat and the toddler have similar distributions and they are very short, we decided to show them in a row, to give people time to build the emotion and show it.
In figure 3.2 we find how the responses corresponding to the selected videos cover the area of the AffectButton. From the distribution, we can see the areas representing anger and surprise are more empty, probably because these emotions are more difficult to trigger through videos. More about this consideration will be found in the discussion section.

3.2. Differences with DISFA dataset

An other dataset collected through video triggering is Denver Intensity of Spontaneous Facial Action (DISFA) database [29]. There are few key differences between DISFA and the work described in this report. First of all, DISFA is meant for studying AUs. Videos are meant for triggering spontaneous facial movements, without specifically focusing on which emotions are triggered. Considering this, there was no need for people to watch the videos in pairs and as a consequence videos did not

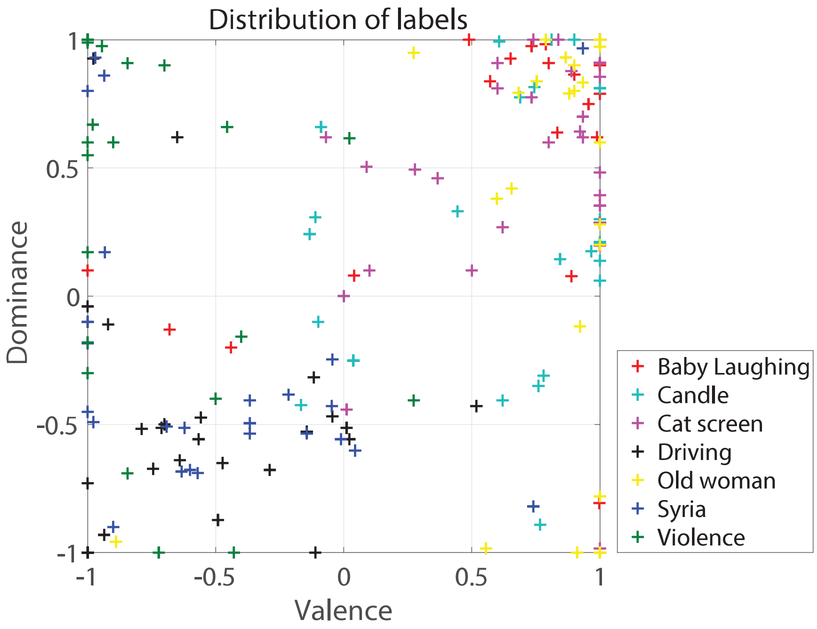


Figure 3.2: Distribution of the labels given to the selected videos by the survey participants. Labels here are defined as value pairs of (valence, dominance) represented by the x and y axes respectively.

need to trigger consistent responses in different subjects. Moreover, they were always shown in the same order with a few seconds in between different videos. Such a short break between them did not give the subjects enough time for having their emotions fading before receiving the next trigger. These differences are consistent with the different scope of the dataset.

3.3. Data collection

3.3.1. Setup

Data have been collected in a time span during a week, involving 41 participants of different age, gender and ethnicity, always indoor and with artificial lightning. Subjects were sitting in front of a screen, beneath which a camera was positioned. All the subjects had the same distance from the screen.

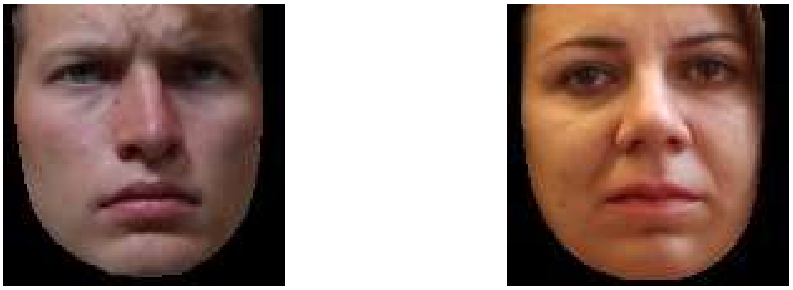
Before starting the data collection, they were asked to record a 10 seconds video with a neutral expression. In some cases, it required a few attempts before participants managed to keep their face neutral. Giving more attempts does not contradict the required automatic nature of the system because in case of an automatic setup, the subjects themselves could be put in charge of recording a neutral video and they could be allowed to delete failed attempts.

Videos were shown in a random order. Participants themselves pointed out that their reactions were dependent on the videos they saw before, specifically regarding sad or upsetting videos. They have been warned that some videos could have been disturbing and no one refused to watch them. However, a couple of people were upset up to the end of the experiment because of a video they saw as second. This might have compromised the data, but it has been ignored because it happened to two people with a peculiar personal history. Their data have been labeled and included in the dataset like everyone else's.

Participants were instructed to watch the videos keeping their head frontal to the screen and that they could interact with each other. They were also asked not to occlude their faces, but many failed to do that for example covering their mouth with a hand when seeing something upsetting. In the case of this work, this is a drawback because it was focused on facial expressions, but it would be useful in case emotion recognition was happening also considering body language and hand gestures.
After each video, participants were asked to rate how they were feeling. They were rating through the AffectButton and they could not see their partner's grade not to be influenced.
Afterwards recordings have been cut in order to comply with the start and the end of the video shown. For two videos, recording started about half a minute after the video started because the interesting stimulus was happening later.

3.3.2. Findings from the data collection

Some conclusions could already be drawn after the data collection.
In one video, reactions during the data collection were really different from those that have been reported in the questionnaire. The video affected by this was showing car accidents. Some people considered the events shown as unrealistic and as a consequence funny, but in the questionnaire it has been rated mostly as sad. An explanation for this could be that people were afraid of being judged and they lied in the questionnaire, but during the data collection they could not do it because smiles and laughs were recorded. This suspicion has been confirmed when, talking to some participants after the experiment, they confessed they were unsure on how to rate the video in question for that same reason. This compromised the effectiveness of the questionnaire which was meant for having people with contradictory reactions watching the video together. The data corresponding to this video have been kept again for fulfilling the automability requirement.



(a) Reaction of subject 10 to video 6

(b) Reaction of subject 24 to video 6

Figure 3.3: Examples of different reactions given very similar rating. Subject 24 gave as valence -0.7587, subject 10 gave -0.7569

An other finding has been that people have their own response style when it comes to rating their emotions [5]. Some are more prone to give extreme grading, while other prefer to stay closer to the middle. An example of this phenomenon would be the reaction and rating of the video about Syria (video 6) from subjects 10 and 24. They gave almost the same unpleasantness rating, but it was evident that subject 24 was more deeply affected by the images because of her personal history. This extreme reaction can not be noticed by her rating that is similar to many others. She explained her choice stating that she has seen much worse than that.

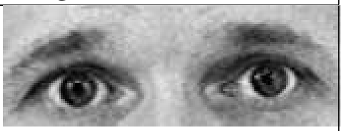
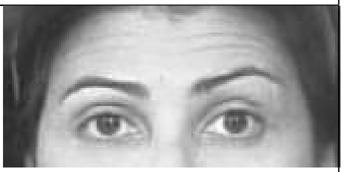
4 Frames selection

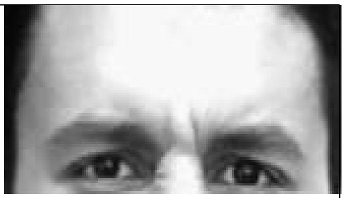
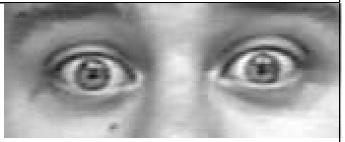
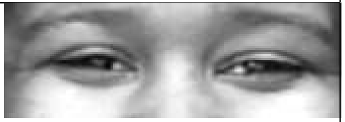
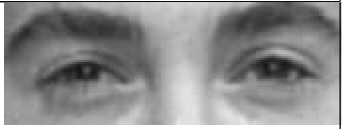
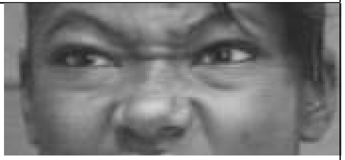
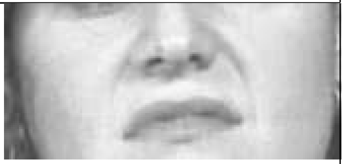
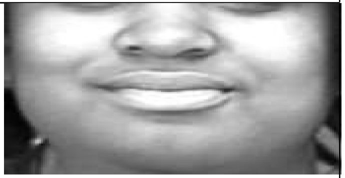
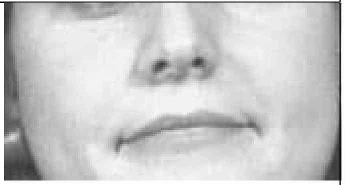
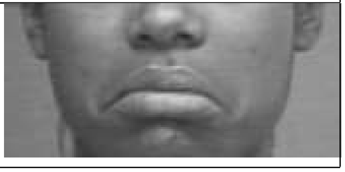
Most of the time spent by people in front of a screen happens assuming a neutral face [13]. If all the video frames recorded from the subjects were directly becoming part of the dataset, numerous neutral frames would end up being labeled as showing emotions. For this reason, a reliable neutral detector is needed in order to sort out neutral images from the labeled set. A neutral face detector has been designed using functions from OpenFace [4], an tool designed for computer vision and machine learning researcher studying facial behavior analysis.

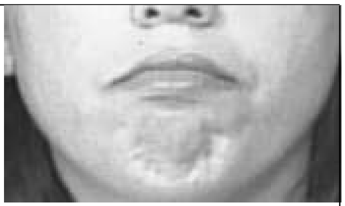
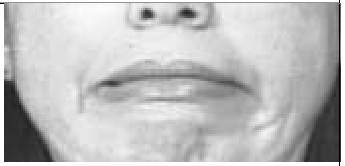
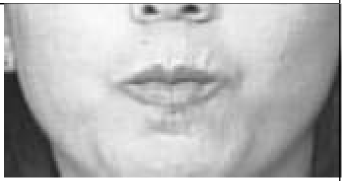
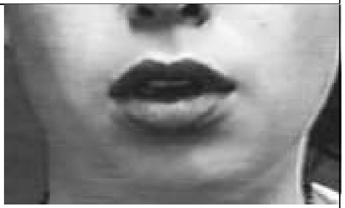
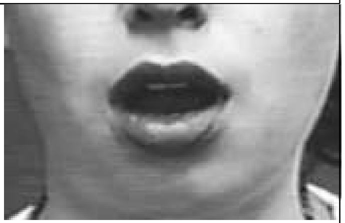
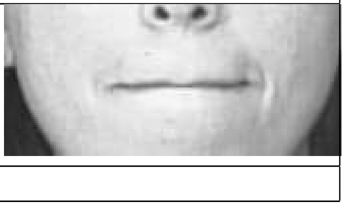
4.1. Action units detection

OpenFace can perform various facial analysis tasks: facial landmarks detection, facial landmark and head pose tracking, gaze tracking, facial features extraction and facial action units recognition. For this research this last function has been used.
Action Units (AUs) are part of the Facial Action Coding System (FACS) [19]: a tool for measuring facial expressions. FACS breaks down facial expression into individual components of muscle movement and labels them as AUs.
OpenFace recognizes 18 AUs. For 17 of them it reports their presence (0 or 1) and their intensity from 0 to 5. For just one action unit (Lip suck), intensity is not reported. Table 4.1 reports and explains the AUs detected.

Table 4.1: A list of the AUs detected by OpenFace [37]

Name	Description	Image
AU01	Inner brow raiser	
AU02	Outer brow raiser	

AU04	Brow lowerer	
AU05	Upper lid raiser	
AU06	Cheek raiser	
AU07	Lid tightener	
AU09	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU14	Dimpler	
AU15	Lip corner depressor	

AU17	Chin raiser	
AU20	Lip stretcher	
AU23	Lip tightener	
AU25	Lips apart	
AU26	Jaw drop	
AU28	Lip suck	
AU45	Blink	

OpenFace's AUs detection has been trained on DISFA [29], BP4D-spontaneous [43] and SEMAINE [38] datasets. They do not all share the same AUs. All the three datasets include AUs 2, 12 and 17. AUs 2, 12, 17, 25 are shared between SEMAINE and DISFA, while 1, 2, 4, 6, 12, 15, 17 are common to DISFA and BP4D. This means that for 10 AUs (5, 7, 9, 10, 14, 20, 23, 26, 28, 45) cross-dataset generalization was not possible.

A pipeline of AUs recognition performed by OpenFace can be found in figure 4.1. Two different procedures are followed for the presence and the intensity of AUs. Both methods have in common facial landmarks detection [2]. After landmarks are detected, for computing the presence of AUs, geometric features are used (as already explained, geometric features consider the distance between facial landmarks). Then

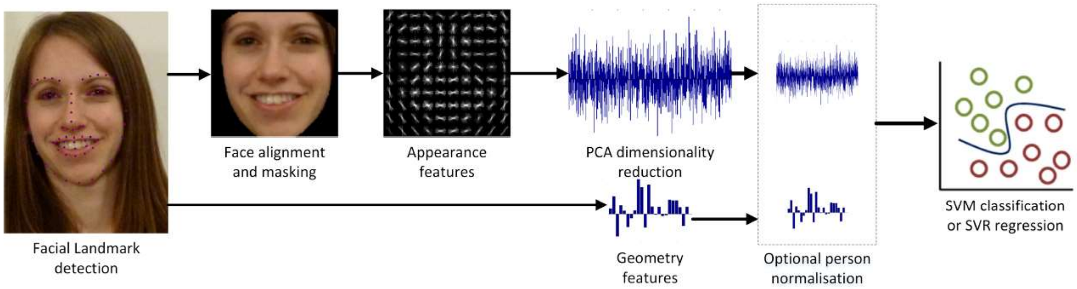

Facial Landmark detection
Face alignment and masking
Appearance features
PCA dimensionality reduction
Geometry features
Optional person normalisation
SVM classification or SVR regression

Figure 4.1: A pipeline of the action unit recognition as performed by OpenFace [3]

optionally a person normalization is performed and the results are fed into a Support Vector Machine (SVM) classifier. For the AUs' intensities the pipeline is slightly longer because it also extracts appearance features. Given the big number of features, Principal Component Analysis (PCA) needs to be performed in order to decrease that number.

4.2. Selection system

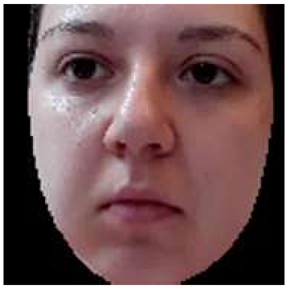


Figure 4.2: An example of face in which AUs are detected wrongly. The person in the picture is assuming a neutral face, but AUs 2 (outer brows raiser) and 4 (brow lowerer) are detected

Unfortunately OpenFace's AUs detection is not perfect and it often detects action units when they are not present. An example can be found in figure 4.2 where two AUs are detected when the subject was performing a neutral face. For this reason, a calibration is needed in order to personalize the neutral face detector. In order to do this, the neutral video recorded from each subject has been used.

The neutral face detector has been optimized analyzing 5 subjects. It has been noticed that all of them had certain AUs that were always present in neutral frames and their absence was corresponding with a facial expression.

For all the computations related to the neutral face detector, only the intensity values of the AUs have been used because they deliver more information. AU 45, which represents blinking, has been ignored because blinking is mostly unrelated to the emotional status. A neutral face detector has been designed, following the following steps:

1. For each frame belonging to the neutral recording, AUs are detected.
2. An array is built, representing for each AU how much it occurred among the neutral frames. This array will be called *percentage array*.
3. For each new frame to be recognized if it is neutral or not AUs, are detected and stored in an array. This array is then compared with the *percentage array* corresponding to the same person. A value that will be called *difference* is then computed, representing how much the new array differs from the percentage.
4. All those frames for which *difference* is bigger than a given threshold, will be considered as not neutral.

After neutral faces are detected, one more step is needed for sorting those frames that are actually useful. If all the not neutral frames were labeled according to the participant's response it would also include all the transition frames. Noticing this, a sorting algorithm needs to be designed. Given the not neutral frames, a threshold selects those that have biggest *difference* value. For each recording, all the frames detected as not neutral are taken with their *differences* computed. Then the frame with maximum distance *max* is considered and only that frame and those with distance above *max* - *t* will be labeled (in which t is a threshold). Through this method, the frames that will become part of the dataset will be those in which the participant is performing a given expression at its apex, ignoring all those frames that are not neutral, but they are just a transition leading to that apex.

4.3. Results

The neutral face detector has been tested on 5 people. For each of these people, a neutral video has been recorded for computing the percentage array. Afterwards, they have been asked to perform different facial expression for a second recording. They were not being told which expressions to perform, as long as they were alternating expressive moments to neutral faces. Then frames belonging to the expressive videos have been manually labeled as neutral or not neutral and the algorithm has been tested. The early phase of an expression's performance is difficult to be detected even from people. For this reason, the ambiguous frames have been ignored at this step, considering only the frames that were unequivocally neutral or not neutral. The video corresponding to subject 1 had 441 frames, subject 2 had 326, subject 3 had 244, subject 4 had 134 and subject 5 had 588. Table 4.2 shows the results on the five subjects. For each subject we can find how many frames have been classified wrongly as false neutrals and as false not neutrals in four different situations in which the neutrality threshold has been changed. Next to each number we find a roman number, representing how many clusters were the wrongly classified frames divided into. This is relevant considering that frames that are close to each other during the recording are very similar because the time passing between a frame and the next one are captured is not enough for the subject to completely change their facial expression.

Attempts		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Attempt 1	False not neutrals	1	0	17 (III)	8 (II)	7 (I)
	False neutrals	4 (I)	5 (II)	0	6 (II)	20 (V)
Attempt 2	False not neutrals	1	0	17 (III)	8 (II)	7 (I)
	False neutrals	4 (I)	5 (II)	0	6 (II)	20 (V)
Attempt 3	False not neutrals	1	0	16 (III)	8 (II)	7 (I)
	False neutrals	4 (I)	6 (III)	0	6 (II)	20 (V)
Attempt 4	False not neutrals	1	0	18 (III)	9 (II)	7 (I)
	False neutrals	4 (I)	3 (I)	0	5 (I)	20 (V)

Table 4.2: Misclassifications of neutral face detector

From testing the neutral faces detector, some conclusions could be drown:

- It works weakly on soft emotions. Especially wrinkles between eyebrows are very difficult to detect.
- Has terrible performance on certain specific people. All the neutral recordings

from the participants of the experiment have been tested as well and it could be noticed that most of the people had 0 or close to 0 misclassification, but few specific people had a high number of misclassified frames.

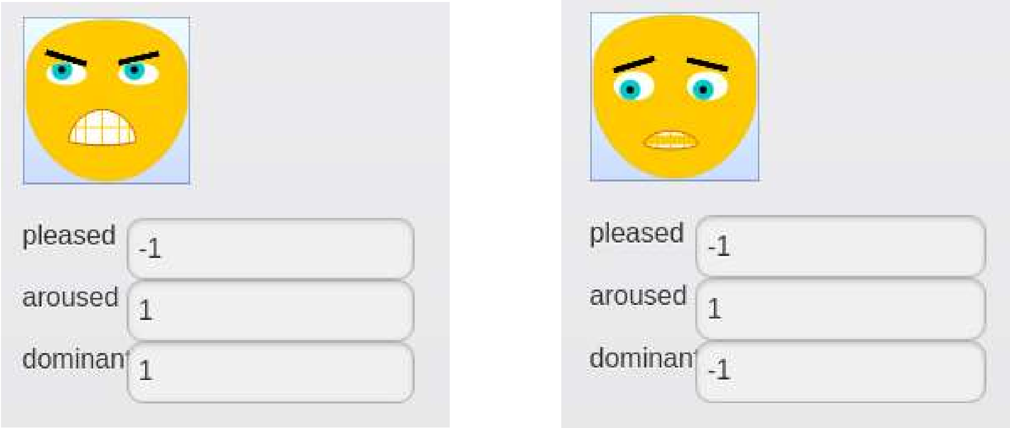
- Changing threshold does not make a big difference, probably because we rarely move muscles singularly in our faces and if more AUs change at the same time the difference with the percentage vector increases quickly.

5
Recognition

For answering the research question, an algorithm needs to be trained on the data collected through video triggering and tested on real life data. This chapter will describe how the testing data have been processed and the algorithms trained for performing the test.

5.1. Data for testing

Testing data have been taken from AffectNet dataset [30], collected harvesting in-the-wild pictures from Internet and labeled using a dimensional description.
The goal of the research, was to verify if the data collected by triggering emotions through videos in a social environment, can be used for training a facial expression recognition algorithm which can be useful in conversation applications. This means that the testing data should be taken from conversations. From AffectNet dataset, a set of images has been selected for fulfilling this requirement. Understanding if pictures of faces belong to a conversation can be challenging, for this reason a particular approach have been used for selecting them, for example selecting those in which a microphone or some particular background were visible assuming they were part of an interview. An other criterion for selecting usable pictures was regarding the demographics of the people in the dataset which, for example, did not have any children or elderly. Moreover, since during data collection positive emotions have been triggered more easily (more about this can be found in the discussion), 71% of the selected frames represent positive emotions. This ratio was maintained also selecting the pictures from AffectNet having 0.71 of the chosen pictures as positive.
The dimensions used by the AffectNet did not match the dimensions utilized by the AffectButton, which presented a problem. As already explained, the AffectButton gives three dimensions (valence, arousal and dominance), but only valence and dominance are independent. AffectNet, on the other hand, is labeled using only valence and arousal. In the AffectButton, a point can't be distinctly identified using valence and arousal because the same valence and arousal can correspond to different dominance, as shown in figure 5.1. For this reason, the test pictures had to be manually relabeled in order to have metrics corresponding to the training data. This did not affect the quality of the labels the labels in the original dataset were also given by external raters and not by the subjects themselves. Moreover, the AffectButton as a tool has its own reliability measure: using the same tool for rating testing and training data brings additional consistency to the approach.



pleased -1
aroused 1
dominan 1

pleased -1
aroused 1
dominan -1

(a) Appearance of the AffectButton with coordinates (-1,1,1)

(b) Appearance of the AffectButton with coordinates (-1,1,-1)

Figure 5.1: An example of why it is not possible to use as only coordinates pleasure as arousal. Both subfigures 5.5a and 5.5b have the same pleasure-arousal coordinates but they clearly represent different emotions

5.2. Regression algorithms

Two different algorithms were tested: a linear regression and a Convolutional Neural Network (CNN).

The linear regressor receives the AUs intensities as inputs and will give as an output the corresponding values for valence and dominance. It consists of a linear function with the AUs as variables. While training, the regressor receives AUs from a training frames and computes the output. Then the output is compared with the labels corresponding to the given frame and the regressor's parameters are adjusted according to the error. The main drawback linked to this method is that, the regressor being a linear function, it can't deal with non-linearities. The second method that has been tried is a CNN. A CNN is a particular kind of Neural Network (NN) specifically meant for processing images. An explanation on what a NN and a CNN are can be found in the appendix.

5.2.1. Linear regression

The linear regression was applied by taking as input the difference between AUs intensities in the given frame and AUs intensities corresponding to the neutral recording of the same person. As output it gave valence and dominance values. Optimization was used to minimize the two-dimensional distance between the point represented by valence and dominance given as output and the point represented by valence and dominance the frame was labeled with.

On my dataset

First, the regressor has been trained and then validated on the collected dataset. It has been split between training and validating set in order to do this. The results obtained are shown in table 5.1:

It needs to be clarified that these errors represent distances between the computed dimensions and the ones given in the label. This means that the errors on valence and distance would have a maximum value of 2 (since the possible values given to them by the AffectButton between -1 and 1). Following the same logic, the error

Error type	Magnitude
Distance	0.3608
Valence	0.2702
Dominance	0.1832

Table 5.1: Results in terms of error values

on distance is the geometric distance between the computed and the labeled point. Its maximum value equals the diagonal of the AffectButton which has length of $\sqrt{8}$. Considering this, normalized to 1 the errors are presented in table 5.2:

Normalized error	Magnitude
Distance	0.1276
Valence	0.1351
Dominance	0.0916

Table 5.2: Results in terms of normalized error values

On real data

The trained regressor was then tested on the testing data coming from AffectNet dataset, obtaining the following results, shown in tables 5.3 and 5.4:

Error type	Magnitude
Distance	0.5730
Valence	0.4176
Dominance	0.3920

Table 5.3: Results in terms of error values

Normalized error	Magnitude
Distance	0.2865
Valence	0.2088
Dominance	0.1960

Table 5.4: Results in terms of normalized error values

These results will be commented also taking into consideration the performance of the CNN.

5.2.2. CNN

Firstly, a description of the layers composing a CNN needs to be made [21]:

Convolutional layer: it applies a series of filters to the image. The same filter is applied more times to the images, moving it every time of a certain steps called stride. Three hyperparameters define this layer: the number of filters applied, their dimension and the stride.

Pooling layer: it downsamples the image, reducing its dimension and consequently the number of parameters needed for the following layers. It has two hyperparametes: the pooling size indicating the dimension of the region in which the pooling operation is applied and the stride. In this case, a max-pooling is used, performing a max operation on each region. An example of how a max-pooling layer works, can be found in figure 5.2

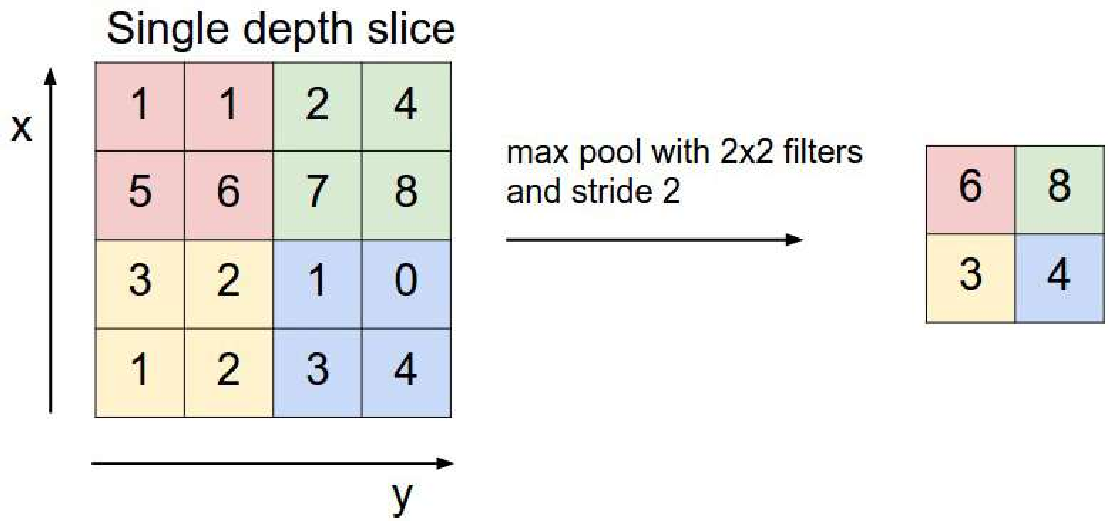


Figure 5.2: An example of a max-pooling layer. For each depth of the input volume, max-pooling operation is performed [21].

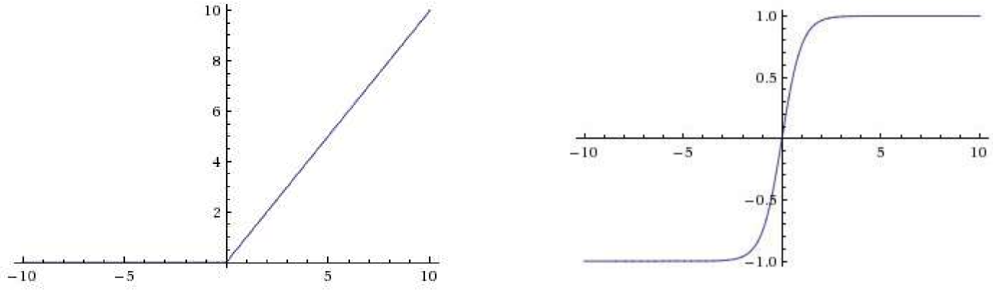
Fully connected layer: it is a layer in which all the neurons are connected with all the neurons of the previous layer. It is governed by a hyperparameter indicating the number of outputs.

Convolutional and fully connected layers also have an activation function, determining the appearance of its output. In the network described below, two different activation functions are used: a ReLU and a tanh function, visible in figure 5.3. The structure of the CNN has been adapted from an already existing network found in literature [35].
The composition of the layers is described below:

1. **Input**: 112*112 layer, receiving images as given by OpenFace, converted into gray-scale

2. **Convolution**: 10 filters with 5*5 dimension, stride 1 and ReLu activation function.

3. **Max pooling**: pool size of (2,2) and stride 2.



(a) ReLU activation function

(b) Tanh activation function

Figure 5.3: Plots from a ReLU activation function (subfigure A.2a) and tanh activation function (subfigure A.2c)

4. **Convolution:** 10 filters with 5*5 dimension, stride 1 and ReLu activation function.
5. **Max pooling**: pool size of (2,2) and stride 2.
6. **Convolution:** 10 filters with 3*3 dimension, stride 1 and ReLu activation function.
7. **Max pooling:** pool size of (2,2) and stride 2.
8. **Fully connected:** Giving 256 outputs, with ReLU activation function.
9. **Fully connected:** Giving 128 outputs, with ReLU activation function.
10. **Fully connected:** Giving 2 outputs, with tanh activation function.

Giving the last layer a tanh activation function, gives two values between -1 and 1 as output, corresponding with the possible values assumed by valence and dominance in the AffectButton. Training has been carried out by considering the mean square error as loss, using a Stochastic Gradient Descent (SGD) optimizer. A learning rate decay has been applied. This kind of solution allows the learning rate do decrease at each interaction. This gives the opportunity to adjust the weights by large margins at the beginning of the training and to add smaller variations later when the function is closer to optimal, since smaller variations enable the attainment of higher precision. Results obtained can be found in figure 5.4

5.3. Discussion

Firstly, a visual example (figure 5.5) are needed in order to show what those errors mean on the AffectButton. Some conclusions can be drawn from these results. It is evident how linear regression preforms worse than CNN. Probably, this is due to the fact that the input it receives is affected by the noise OpenFace adds while computing the AUs.

Results of the CNN its results need to be interpreted considering that this work was not meant to test the validity of the network, but the validity of the dataset. When designing the recognition algorithm, much better results are usually achieved for two main reasons. Firstly, this is a cross-dataset recognition, which means the algorithm is trained and tested on different datasets. The work from Goyal et al. in 2018 shows a series of results from cross-dataset recognition [24]. They obtained results between 84.7% and 41.3% of accuracy. The second element that needs to be considered is

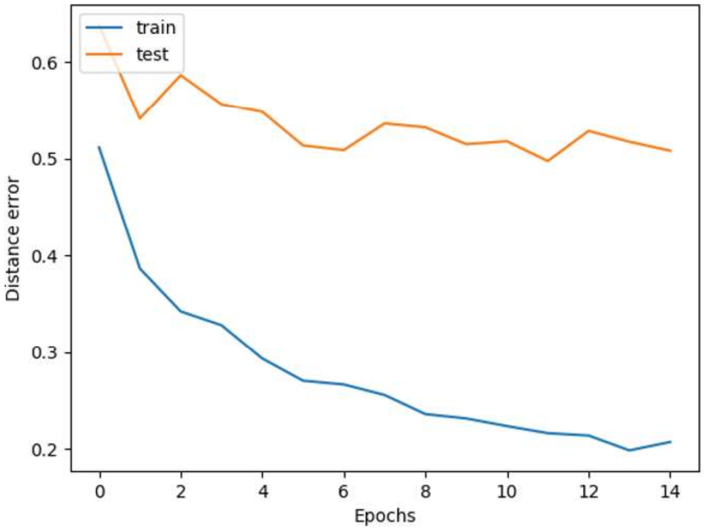
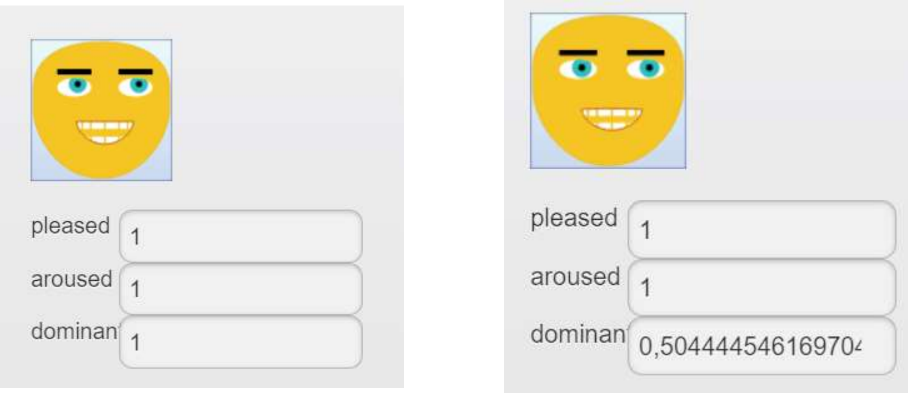


Figure 5.4: Error on training and testing sets. The error is represented as distance, which means it has 2 as maximum. Normalized, the error reached by the testing data is about 0.25.

that the AffectButton is a tool and with its own reliability, which is 0.81 for valence and 0.77 on dominance [7]. Reliability is what would be precision in measurement, it means we could expect of 0.19 and 0.23 of the measurement as error respectively. As our measures can go between -1 and 1, it leads to a difference of 2 in absolute value. This would make our ”allowed error” of 0.38 and 0.46. Considering this, reaching an error of 0.5 seems an acceptable result. On the other hand the error obtained in the training set appears to be too low, probably due to overfitting. Interrupting the training around the fourth epoch is likely solve the problem.
The linear regressor gives similar results on the testing data, but even lower on the training data. This is probably because an even higher problem of overfitting.



pleased 1

aroused 1

dominan 1

pleased 1

aroused 1

dominan 0,504444546169704

(a) Appearance of the AffectButton with coordinates (1,1,1)

(b) Appearance of the AffectButton with coordinates (1,1,0.5044)

Figure 5.5: An example of how much the appearance of the AffectButton changes with an error given by the dataset

6 Conclusion and future work

6.1. Summary

The aim of this project was addressing the challenges currently present in facial expression recognition by finding an alternative way of collecting data. One of the key obstacles to facial expression recognition is the wide variety of emotions felt and expressions acted by humans in different situations and, as a consequence, the difficulty on training an algorithm that would be useful in any situation. Moreover, data collection is expensive in terms of time and money, making task-specific data collection often inconvenient. This thesis project investigates a possible way for quickly and cheaply collecting data, especially for a Human Computer Interaction (HCI) application. In particular, this project is focused on Pepper: a robot meant for interacting with humans through conversation. For collecting data, emotions should be triggered in participants and their faces should be video-recorded. The fundamental characteristic training data should have is being as similar as possible to the one that will be found in the actual application. In this research, the triggering method utilized was showing videos selected for triggering specific emotions, letting participants watch them in pairs interacting with each other to better simulate the social environment in which the robot will have to operate. After watching each video, participants were asked to rate their feelings through the AffectButton: a tool for intuitively describing emotions in a dimensional way. Video's selection was assisted by a questionnaire in which people were asked to rate emotions they were triggering in them. This measure was important for avoiding videos that would have triggered various emotions in people because pairs would have influenced each other while interacting and the labels might have been compromised. Recordings obtained from people watching the triggering video have been then compared to a model of a neutral expression performed by the same participants, in order to select the frames in which expressions were shown, ignoring neutral and transition frames. The pictures obtained have been included in a dataset using for training a linear regressor and a Convolutional Neural Network (CNN), that have been then tested on naturalistic data taken during conversations, in order to discover if the collection method could build a useful dataset for real life application. The result showed that this is a promising direction to follow and some key elements for future improvements are provided in the following section.

6.1.1. Thesis contributions

The main contribution of this thesis is the opening to a new approach for data collection. As stated by several articles in the news regarding machine learning [40] [6] [15], power is represented not by designing the best algorithm, but possessing the biggest amount of data. The alternative method for collecting data described in this thesis could give the opportunity to universities and smaller companies to obtain big amounts of data based on their needs. The project described in this report, showed that showing people videos in pairs is a good direction to follow for triggering emotions and pointed out directions for further improvement on the topic. These improvements, together with suggestions for future work, will be described in the following section.

6.2. Remarks and future work

Some remarks have already been stated when giving interpretation for the results. One more factor that should be considered is that the data this work is referring to has been collected from a limited number of people (41). It should also be remembered that this method is meant for automating data collection. It would have the big advantage on reducing the time spent performing the task manually. Even if a worse performance was achieved by the automatic collection compared to a manual one, the decreasing amount of time and money spent would be an advantage in all of those situations in which time or money constraints are particularly strict.

Having a closer look at the frames that have been given a big error by the CNN, some more information about the functionality of the data collection can be provided. 0.39 of the test frames are given an output with an error bigger than the average (≥ 0.5). 60% of those show non-positive emotions. If we consider those with a bigger error (≥ 0.6), they represent 32% of the testing data and 67% of them are negative emotions. Considering that positive emotions compose 70% of the test data, it is possible to notice a big unbalance on the performance for positive and negative emotions. Given this, some suggestions for future work can be given:

- A further test could be run, having the recognition performed by Pepper in actual conversations in its application, comparing the performance with an algorithm trained on other datasets. In particular, an algorithm should be trained on a dataset obtained triggering emotions showing videos to participants not in pairs, to actually confirm if the social environment improves the usability of the data. Moreover, comparative studies with different triggering videos could be run, for finding those that would better comply the application's requirements.

- For the recognition to work properly on conversation, it should be robust to facial movements due to speech. For doing this, participants should be encouraged to talk to each other more during the data collection, and specific implementations of the algorithm should be considered.

- Negative emotions are more difficult to trigger and people display them less easily. This explains the gap between performances on positive and negative emotions and suggests to try with different triggering methods for those emotions. An example could be simple tasks or a video game. An other possible solution for making people more involved also negatively could be showing longer videos, but that would require more time from the participants.

- It has been mentioned the difference in the way participants rate their own feeling. A possible improvement on this point of view could be not basing the labels only on the subject's rating. An example of extra factors that could be considered are physiological signals such as hormonal levels, body temperature, heartbeat or blood pressure.

- More specifically on HiT's application, it could be studied what kind of triggers could give emotions useful in its application. For the health-care application, emotions such as pain, fear, despair or boredom should be considered.

A

Neural Networks

Neural networks (NN) are a model for machine learning. Their basic component is called **artificial neuron** and it appears as shown in figure A.1.

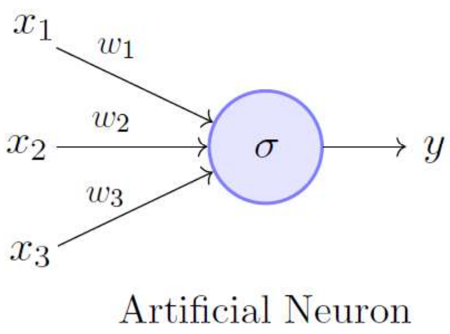


Figure A.1: An artificial neuron, basic component of an artificial neural network

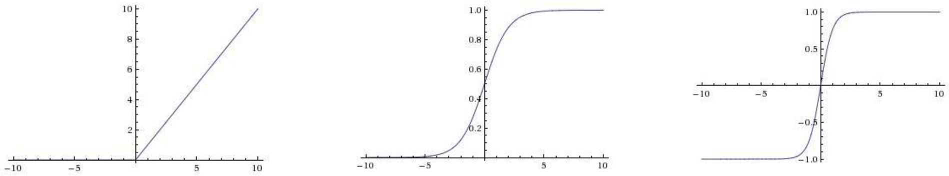
It takes a series of inputs $(x_1, ..., x_p)$ and computes an output v, according to the following formulas:

$$w_1 * x_1 + w_2 * x_2 + ... + w_p * z_p = z$$

$$\sigma(z) = v$$

$W_1, ...w_p$ are parameters. They are those that the training of the network addresses. σ is an activation function. It determines what output will be given by the neuron based on its computations. There are many different kinds of activation functions: A NN is composed by many neurons, organized in layers. Each neuron can receive inputs from other neurons, but the network can not have loops. The first layer is called input layer, the last is the output layer and the others are hidden layers. They are called hidden layers because their output. NNs have different dimensions based on their application, they can even have no hidden layers.

A Neural Network (NN) is trained using a dataset. A dataset is a series of data with the corresponding label. The label represent the output we expect the network to give when fed with that input. For each element present in the dataset, the NN computes



(a) Example of a ReLu activation function

(b) Example of a sigmoid activation function

(c) Example of a tanh activation function

Figure A.2: Examples of activation functions: a ReLu (subfigure A.2a), a sigmoid (subfigure A.2b) and a tanh (subfigure A.2c) [21].

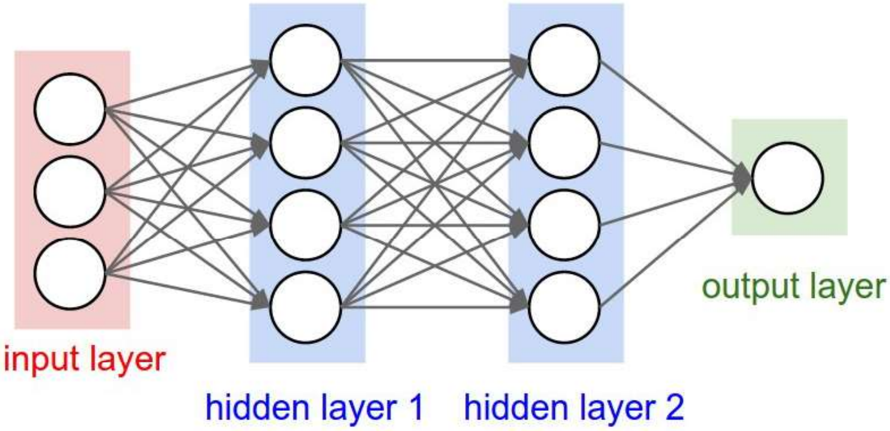


Figure A.3: Example of a NN composed by 2 hidden layers [21].

an output and then it compares it with its label. Based on the error, it adjusts its parameters (w_i) in order to improve its performance. This operation is repeated many times, until its error will decrease up to a desired value.

B

Glossary

List of acronyms

ADs	**A**ction **D**escriptor**s**
AI	**A**rtificial **I**ntelligence
AUs	**A**ction **U**nit**s**
CNN	**C**onvolutional **N**eural **N**etwork
DISFA	**D**enver **I**ntensity of **S**pontaneous **F**acial **A**ction database
FACS	**F**acial **A**ction **C**oding **S**ystem
HCI	**H**uman **C**omputer **I**nterface
HiT	**H**eemskerk **i**nnovative **T**echnology
PCA	**P**rincipal **C**omponent **A**nalysis
NN	**N**eural **N**etwork
SAL	**S**ensitive **A**rtificial **L**istener
SGD	**S**tocastic **G**radient **D**escent
SVM	**S**upport **V**ector **M**achine

Dictionary

Action Units	Anatomic facial muscle actions, each of them corresponds to a single muscle or group of muscles.
Dataset	A set of labeled examples fed to a machine learning algorithm for training.
Principal Component Analysis	Method used in machine learning for decreasing the number of features.
Support Vector Machine	Supervised machine learning model.

Bibliography

[1] Ashraf Abbas A. Al-modwahi, Onkemetse Sebetela, Lefoko Nehemiah Batleng, Behrang Parhizkar, and Arash Habibi Lashkari. Facial Expression Recognition Intelligent Security System for Real Time Surveillance. *World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP'12)*, pages 1–8, 2012. URL `http://elrond.informatik.tu-freiberg.de/papers/WorldComp2012/CGV2255.pdf`.

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[3] Tadas Baltrušaitis, Marwa Mahmoud, and Peter 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*, volume 6, pages 1–6. IEEE, 2015.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 59–66. IEEE, 2018.

[5] Hans Baumgartner and Jan-Benedict EM Steenkamp. Response styles in marketing research: A cross-national investigation. *Journal of marketing research*, 38(2):143–156, 2001.

[6] Kirk Borne. The real power of ai.

[7] Joost Broekens and Willem Paul Brinkman. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human Computer Studies*, 71(6):641–667, 2013.

[8] Britain's Got Talent (Youtube channel). Spectacular salsa - paddy & nico - electric ballroom | britain's got talent 2014. *https://www.youtube.com/watch?v=hjHnWz3EyHs&t=195s*, .

[9] BruBearBaby (Youtube channel). Baby laughing hysterically at ripping paper (original). *https://www.youtube.com/watch?v=RP4abiHdQpc*, .

[10] Giuseppe Romano (Youtube channel). Evanescence - my immortal (embrace life). *https://www.youtube.com/watch?v=ESJU903XfwM*, .

[11] Sapien Medicine (Youtube channel). Heart touching children in syrian civil war: Share if you care. *https://www.youtube.com/watch?v=jdKHVnHTXkU&t=124s&has_verified=1*, .

[12] Zen News (Youtube channel). Domestic violence caught on camera. *https://www.youtube.com/watch?v=1mzjeeG3B1E&has_verified=1*, .

[13] Pojala Chiranjeevi, Viswanath Gopalakrishnan, and Pratibha Moogi. Neutral face classification using personalized appearance models for fast and robust emotion detection. *IEEE Transactions on Image Processing*, 24:2701–2711, 2015.

[14] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F. Cohn, and Sergio Escalera 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*, 38(8): 1548–1568, 2016.

[15] Ben Croning. Artificial intelligence | data is power in the time of machine learning. *Sport Business*.

[16] Ellen Douglas-Cowie, Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Suzie Savvidou, Sarkis Abrilian, and Cate Cox. Multimodal databases of everyday emotion: Facing up to complexity. 2005.

[17] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer, 2007.

[18] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amier, and Dirk KJ Heylen. The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. In *LREC Workshop on Corpora for Research on Emotion and Affect*. ELRA, 2008.

[19] P Ekman and WV Friesen. Facial coding action system (facs): A technique for the measurement of facial actions. 1978.

[20] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200, 1992.

[21] Notes from course CS231n from Stanford university. http://cs231n.stanford.edu/.

[22] America's funniest home videos (Youtube channel). Boy pulls out all the stops to blow out candle. *https://www.youtube.com/watch?v=6n3IbUQvPWs*, .

[23] America's funniest home videos (Youtube channel). Confused cat thinks the computer squirrel is real. *https://www.youtube.com/watch?v=JUAv1jaocN8*, .

[24] Samta Jain Goyal, Arvind K Upadhyay, RS Jadon, and Rajeev Goyal. Real-life facial expression recognition systems: A review. pages 311–331, 2018.

[25] Hatice Gunes and Hayley Hung. Is automatic facial expression recognition of emotions coming to a dead end? the rise of the new kids on the block. 2016.

[26] Michal Kawulok, M. Emre Celebi, and Bogdan Smolka. *Advances in face detection and facial image analysis*. 2016. doi: 10.1007/978-3-319-25958-1.

[27] P. Lucey, Jeffrey F. Cohn, Takeo Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotionspecified expression. *Cvprw*, (July): 94–101, 2010.

[28] M K Mandal, R Pandey, and A B Prasad. Facial Expressions of Emotion and Schizophrenia: A Review. *Schizophrenia Bulletin*, 24:399–412, 1998.

[29] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[30] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[31] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[32] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

[33] M. Usman S. Ali Khan, A. Hussain. Facial expression recognition on real world face images using intelligent techniques: A survey. *Optik*, 127(15):6195–6203, 2016.

[34] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. We are not All Equal: Personalizing Models for Facial Expression Analysis with Transductive Parameter Transfer. *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 357–366, 2014.

[35] Sefik Serengil. Cnn model. *https://sefiks.com/2018/01/01/facial-expression-recognition-with-keras/*.

[36] Andrei State. Six basic emotions. *http://www.cs.unc.edu/ andrei/expressions/*.

[37] Carnegie Mellon university website. Descritpion of the action units. *https://www.cs.cmu.edu/ face/facs.htm*.

[38] Michel F. Valstar, Timur Almaev, Jeffrey M. Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F. Cohn. FERA 2015 - second Facial Expression Recognition and Analysis challenge. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015. doi: 10.1109/FG.2015.7284874. URL `http://ieeexplore.ieee.org/document/7284874/`.

[39] C. Vinola and K. Vimaladevi. A survey on human emotion recognition approaches, databases and applications. *Electronic Letters on Computer Vision and Image Analysis*, 14(2):24–44, 2015. ISSN 15775097. doi: 10.5565/rev/elcvia.795.

[40] Boris Wertz. Data, not algorithms, is key to machine learning success. *Machine intelligence report.*

[41] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Facial expression recognition experiments with data from television broadcasts and the World Wide Web. *Image and Vision Computing*, 32(2):107–119, 2014. URL `http://dx.doi.org/10.1016/j.imavis.2013.12.008`.

[42] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Facial expression recognition experiments with data from television broadcasts and the world wide web. *Image and Vision Computing*, 32(2):107–119, 2014.

[43] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.