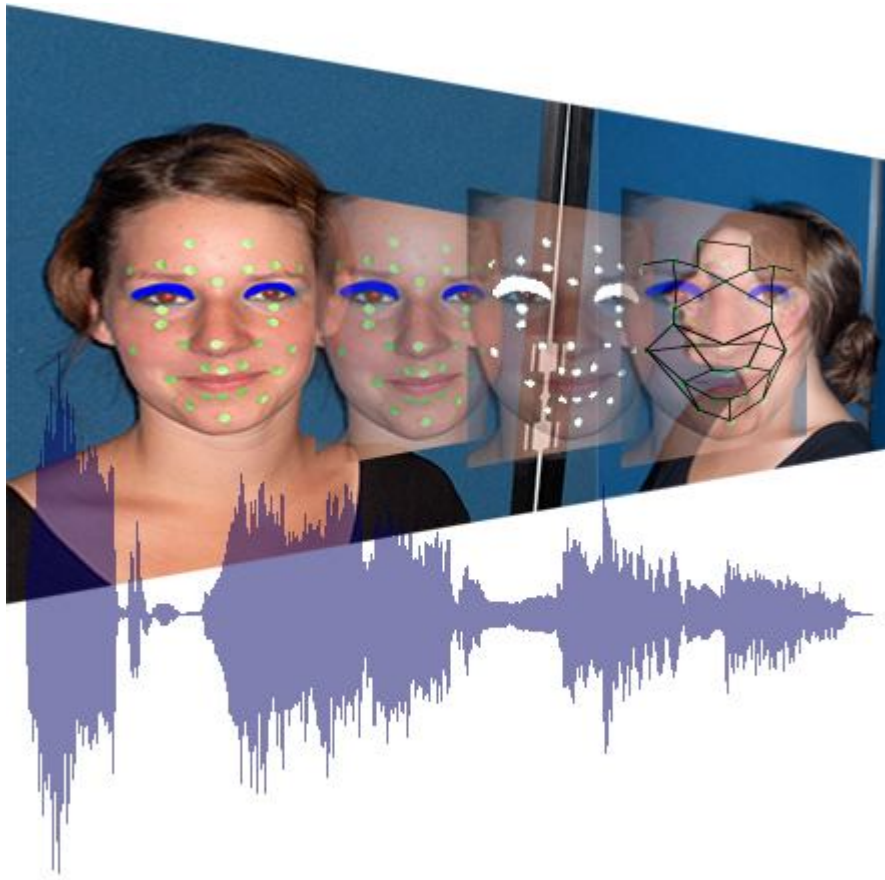


Analysis and Recording of Multimodal Data

Version of September 10, 2008



Analysis and Recording of Multimodal Data

In partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Mathijs van Vulpen
born in Brouwershaven, The Netherlands

September 2008



Delft University of Technology

Man Machine Interaction group

Delft University of Technology

Faculty of Electrical Engineering, Mathematics, and Computer Science

Mekelweg 4

2628 CD Delft, The Netherlands

<http://www.ewi.tudelft.nl>

Analysis and Recording of Multimodal Data

Emotions are part of our lives. Emotions can enhance the meaning of our communication. However, communication with computers is still done by keyboard and mouse. In this human-computer interaction there is no room for emotions, whereas if we would communicate with machines the way we do in face-to-face communication much information can be extracted from the context and emotion of the speaker. We have proposed a protocol for the construction of a multimodal database and a prototype that can be trained on this database for multimodal emotion recognition.

The multimodal database consists of audio and videos clips for lip reading, speech analysis, vocal affect recognition, facial expression recognition and multimodal emotion recognition. We recorded these clips in a controlled environment. The purpose of this database is to make it a benchmark for the current and future emotion recognition studies in order to compare the results from different research groups.

Validation of the recorded data is done online. Over 60 users scored the apex images (1.272 ratings), audio clips (201 ratings) and video clips (503 ratings) on the valence and arousal scale. Textual validation is done based on Whissell's Dictionary of Affect in Language. A comparison is made between the scores of all four validation methods and the results showed some clusters for distinct emotions, but also some scatter for certain emotions which depend mainly on the context. Context is not always available.

We created a prototype that can extract and track the facial feature points, this prototype is based on the system of Anna Wojdel. The prototype is designed in Matlab and is able to separate the audio from the video clip, extract frames and perform 5 different classifiers on the audio and video stream separately. For the auditory channel we have trained three classifiers: one for all 21 emotions, one for positive and negative emotions and one for active and passive emotions. For the visual channel we have trained two classifiers: one based on the found facial feature points and one based on AU activation. The classification results from our prototype are promising, considering we have 21 different emotions and trained the auditory classifiers on two persons and the visual classifiers on one person. Better results can be established if we have access to more samples from various people. The average classification rate for the three auditory classifiers is 38%, 36% and 59% respectively, for the two visual classifiers 2% and 0% respectively.

Thesis Committee:

Chair:	Prof. Dr. Drs. L. J. M. Rothkrantz, Faculty EEMCS, TU Delft
University supervisor:	Ir. H. Geers, Faculty EEMCS, TU Delft
Committee Member:	Dr. Ir. P. Wiggers, Faculty EEMCS, TU Delft
Committee Member:	Ir. A. G. Chitu, Faculty EEMCS, TU Delft

Preface

Eight years ago I started a new chapter in my life, “Delft”. I came to Delft as a young man who just finished high school and was ready for the next step. The challenge was to finish my computer science education in five years, but soon enough I learned that “Delft”, or studying at the TU Delft, is more than just studying! The Delft life contains lots of extracurricular activities and opportunities to grow and develop yourself into an international respected entrepreneur.

After two years of doing a lot of extracurricular activities I met my girlfriend, Ewine. Together with her I picked up my study where I left it two years ago (at the start). I followed most courses together and I could always talk to her about problems that we had, or vice versa.

We both did our master courses and along the way looked for suitable and interesting Masters Projects. The question that arises then is whether or not you are going to graduate at the TU Delft or at a company. My choice fell on graduating here, at the TU Delft. The opportunities that were presented were all challenging and some of them even had a sort of cooperation with international companies, like Philips. This meant for me that I did not have to work from eight till five, after so many years, the student habits are strong and difficult to get rid of.

Working interactively with a computer, where the computer understands what you want, is the real challenge for me. I briefly touched the concept of emotion recognition in my Bachelors Project; “Audio and Video Emotion recognition”. This research area suited me well and I started talking to the professor responsible for this subject in Delft, Prof. Leon Rothkrantz. We both were very enthusiastic and found a suitable subject quickly. We were to record our own multimodal database and would try to recognise emotions from these clips.

Furthermore, I would like to thank all the master students from the MMI and CG lab, for their warm company during hard days and the lunches we all had together. These people helped me get through the ‘hard’ days of the graduating process. I would like to thank my study guild, W.I.S.V. ‘Christiaan Huygens’ for the wonderful time in my life they offered me. The chances to explore myself outside the study, going abroad, taking seat in important committees, organising the 10th Lustrum and finally becoming a member of honour of this study guild, has made me the person I am today. These experiences have also given me many friends. I can’t name them all here otherwise the list would be too long, but you know who I mean guys!

Besides the aforementioned, my gratitude goes out to Leon Rothkrantz for offering the research assignment and providing support and advice in constructing this thesis. Besides his help I would like to thank him for the endless conversations we had about our study guild, W.I.S.V. ‘Christiaan Huygens’, for the eye-opening talks about emotions and for all the jokes we made either in the lab, the hallway or in-between recordings. Besides my daily supervisor I would like to thank Alin Chitu for sharing his experience with me and putting a lot of effort into our recordings too. Being a PhD student of Leon Rothkrantz gives you more responsibilities than you hoped for, but I know these are all worth it.

Last but certainly not least I am grateful for the support, motivation and advice from my parents, family, friends, and my girlfriend Ewine. Thank you all for the support and believe you all have in me.

Mathijs van Vulpen
Delft, The Netherlands
September 5, 2008

Content

PREFACE	VII
CONTENT	IX
LIST OF FIGURES	XIII
LIST OF TABLES	XV
1 INTRODUCTION	1
1.1 Problem Overview.....	2
1.2 Research Goals.....	3
1.3 Project Approach.....	4
1.3.1 Research Assignment	4
1.3.2 Data Recording	5
1.3.3 Data Validation	5
1.3.4 Model.....	6
1.3.5 Prototype.....	6
1.3.6 Results	6
1.4 Thesis Structure.....	7
2 OVERVIEW OF RELATED WORK	9
2.1 The Cohn-Kanade Database	9
2.1.1 The Cohn-Kanade Approach	9
2.1.2 The Cohn-Kanade Drawbacks	10
2.2 The e'NTERFACE'05 Audio-Visual Emotion Database	10
2.2.1 The e'NTERFACE Approach.....	10
2.2.2 The e'NTERFACE Drawbacks	11
2.3 A New Emotion Database	11
2.3.1 The New Emotion Database Approach	12
2.3.2 The New Emotion Database Drawbacks	12
2.4 State of the Art Multimodal Emotion Recognition	12
3 EMOTIONS AND EMOTION RECOGNITION	15
3.1 Emotions	15
3.1.1 Face-to-Face Communication	16

3.1.2	Visualisation of Emotions	16
3.2	Facial Expressions Analysis	17
3.2.1	The Facial Action Coding System (FACS)	17
3.2.2	Facial Expression Recognition	18
3.3	Vocal Affect Recognition	19
3.3.1	Automatic Vocal Affect Recognition	20
3.4	Multimodal Emotion Recognition	20
3.4.1	Fusion of Different Modalities	20
3.4.2	Decision-Level Fusion	21
3.4.3	Feature-Level Fusion	22
4	CONSTRUCTION OF A MULTIMODAL DATA CORPUS	25
4.1	Emotional Content	25
4.1.1	Spontaneous or Posed Data	26
4.2	Recording Protocol	26
4.2.1	The Room Setting	26
4.2.2	The Environment Conditions	26
4.3	The Procedure of the Experiment	27
4.4	Technical Aspects	28
4.4.1	Audio and Video Devices	29
4.4.2	Storage	30
4.4.3	Metadata	30
4.5	Multimodal Database Availability	31
4.5.1	Standards	31
4.5.2	Security	31
4.5.3	Search Queries	32
5	VISUAL DATA INSPECTION	33
5.1	Evaluation Space	33
5.2	Data Validation	34
5.2.1	Image Based Validation	35
5.2.2	Audio Based Validation	35
5.2.3	Video Based Validation	35
5.2.4	Text Based Validation	36
5.3	Validation Results	40
5.3.1	Image Validation Results	40
5.3.2	Auditory Validation Results	41
5.3.3	Multimodal Validation Results	41
5.3.4	Textual Validation Results	41
5.4	Comparison of Validation Results	42
6	MODEL, PROTOTYPE AND DATA PRE-PROCESSING	45
6.1	The Model	45
6.2	Prototype Design	46
6.2.1	Implementation Details	46
6.3	Audio Feature Extraction	46
6.3.1	Praat	49
6.4	Facial Feature Points Localisation	49
6.4.1	Image Segmentation	49
6.4.2	Facial Measurement Model	51
6.4.3	Detection of Facial Feature Points	52
6.4.4	Point Validation	53

7	FEATURE VECTOR CLASSIFICATION	55
7.1	Classification Model.....	55
7.2	Feature Selection and Noise Reduction.....	57
7.2.1	Feature Selection by Boosting Algorithms	57
7.2.2	Audio Feature Selection	57
7.2.3	Image Feature Selection	58
7.2.4	Principal Component Analysis for Facial Feature Vectors.....	58
7.3	Facial Feature Point Translation	59
7.4	Feature Vector Comparison	60
7.5	Classification	62
8	CLASSIFICATION RESULTS.....	65
8.1	Audio Feature Classification Results	65
8.2	Facial Feature Classification Results.....	67
8.2.1	Number of Correct Processed Frames	68
9	CONCLUSIONS AND FUTURE WORK	69
9.1	Concluding Remarks	70
9.2	Future Work.....	70
	BIBLIOGRAPHY	73
	APPENDICES	81
A:	List of Action Units	83
B:	Points to AUs Translation.....	87
C:	AU Constraints for AU Activation.....	93
D:	Participant’s Instructions	97
E:	Text Used for Recordings	99
F:	Image Validation Results.....	107
G:	Audio Validation Results.....	111
H:	Video Validation Results	115
I:	Table with Resulting Classification Labels for All Clips	119
J:	Normalised Distance Matrix for 2 Persons for all 21 Feature Vectors.....	123
K:	Consent Document	125
L:	Building a Dutch Multimodal Corpus for Emotion Recognition	127

List of Figures

Figure 1: Face-to-face interaction with different modalities.	2
Figure 2: Screenshot of the prototype.	6
Figure 3: Architecture of an “ideal” automatic analyser of human affective feedback.	13
Figure 4: Duration of different emotional states.	16
Figure 5: Activation-Evaluation space of the six basic emotions.	17
Figure 6: Overview of the room setting.	27
Figure 7: Situation and reactions to elicit “disappointment”.	28
Figure 8: The Pike F032C camera built by AVT.	29
Figure 9: The NT2A Studio Condensators.	29
Figure 10: Different emotions on the valence and arousal scale.	34
Figure 11: The valence 9-point scale Self Assessment Mannekin (SAM).	34
Figure 12: Explanation of flash content with controls.	36
Figure 13: The developed DAL Web Interface.	36
Figure 14: Example validations of an image (left) an audio clip (middle) a video clip (right).	37
Figure 15: Example of displayed images for validation of participant P1.	38
Figure 16: Example of displayed images for validation of participant P2.	39
Figure 17: Comparison of different validation methods.	44
Figure 18: System overview.	45
Figure 19: Waveform with 10ms time window.	47
Figure 20: The original image (a), the detection of only 26 of the 29 green stickers (b) and the detection of the blue eye make-up (c).	50
Figure 21: The detection of feature points 10 and 11.	50
Figure 22: Placement of the 31 points we tracked.	51
Figure 23: Yaw, Pitch and Roll visualised for the face.	52

Figure 24: Processing flow of facial features tracking.....	53
Figure 25: Large model view of multimodal emotion recognition.....	56
Figure 26: Visualisation of obsolete points.....	58
Figure 27: Plots of the first two principal components of the X-coordinates (left) and Y-coordinates (right) of all 31 facial feature points.	59
Figure 28: Facial differences for Participant P1 (top) and Participant P2 (bottom).	61
Figure 29: Two examples for the Pearson Correlation.	63

List of Tables

Table 1: Recognition rates for recent attempts of multimodal emotion recognition	14
Table 2: List of used emotions.	25
Table 3: Metadata fields with corresponding values.	30
Table 4: Validation results for image validation.	40
Table 5: Validation results for auditory validation.....	41
Table 6: Validation results for multimodal validation.....	41
Table 7: Whissell scores^[1] for the used emotions (9-point scale).	42
Table 8: Overview of position of emotions per quadrant.....	43
Table 9: Results of feature selection done by GentleBoost for audio features.	57
Table 10: Implemented AUs.	59
Table 11: Normalised distances between facial feature vectors from P1 and P2.	61
Table 12: All different trained classifiers from PRTools	62
Table 13: 13 Template expressions.....	62
Table 14: Mean error rates per classifier per dataset.....	65
Table 15: Number and percentage of correct classified clips.	66
Table 16: Number and percentage of correct classified clips.	67
Table 17: Percentage of correct processed frames per emotion.	68

CHAPTER 1

Introduction

Emotions are an integral part of our daily rational decision making and communication. To approach the naturalness of face-to-face interaction machines should be able to emulate the way humans communicate with each other. It is the human face that conveys most of the information about our emotions to the outside world. Considerable research in social psychology has shown that besides speech alone, non-verbal communicative cues are essential to synchronise the dialogue, to signal certain emotions and intentions and to let the dialogue run smoother and with less interruptions [1].

Non-verbal communication is a process consisting of a range of features including body gesture, posture and touch or paralingual cues, often used together to aid expression. The combination of these features is often a subconscious choice made by native speakers, and interpreted by the listener. Of all different non-verbal communication means, facial expressions are the most important means for interpersonal communication [2]. We learn to recognise faces and facial expressions early in life, long before we learn to communicate verbally. A human face can supply us with important information.

- A human face gives us primary information about the identity of the person and provides information about sex and age of the subject.
- The appearance of a human face performs an active role in speech understanding [3]. It is shown that even normal-hearing people use, to some extent, lip reading in order to better understand the speaker. This means that the intelligibility of speech is higher when the speaker's face is visible [4]. When the amount of sound sources, or noise, increases, visual information can be very helpful in understanding the message.
- Appropriate facial expressions and/or body gestures also provide additional communicative functions [5]. Often unconsciously, people use nonverbal language (facial expressions, hand gestures, eye gaze etc.) to enrich their dialogue. The other way around, people unconsciously read the nonverbal queues to emphasise what the speaker is communicating. Facial expressions can even be used as a replacement for specific dialogue acts (such as confirmation or spatial specification).

It is natural for humans to try to understand in depth what the real meaning of the message, behind the verbal part of communication, is. Body gestures, speech and even written text have a hidden layer that corresponds to the emotional state of the speaker. Knowledge about this layer is used in perceiving and performing acts of communication. On the perception side, we can learn something more about the other person if we can consciously interpret the signals that reveal the emotional background (e.g. a hidden agenda or outright lies). On the other hand, if we are aware of the effect that

our body language or facial expressions can have on the other person, we can control them in such a way that the communication results in the most efficient way that is most beneficial for us [6].



Figure 1: Face-to-face interaction with different modalities.

For years, human-computer interaction has been dominated by keyboard and mouse. This is not a natural way for humans to communicate. Therefore, it would be much easier if we could communicate with computers in the same way as with other people. Hence, as soon as computers started to become multimodal communication devices, the need for robust facial expression analysis, vocal affect analysis, speech recognition, (body)gesture recognition and context awareness became apparent. In order to do so and obtain a natural way of communicating with computers, the computer should recognise and understand the users' emotional state, and at the same time, be able to communicate in an understandable way for humans.

In face-to-face interaction, humans employ different modalities like facial expression and paralingual cues simultaneously and in combination, using one to complement and enhance the other. The importance of facial expression and non-verbal cues is also acknowledged in cartoons like 'The Flintstones', see Figure 1. Fred and Barney act like real people. They make expressions just like real people do. They have head and hand gestures that perfectly describe how they are feeling at every unique moment in the story. Many different prototypes of expression recognition systems have been developed using visual or prosodic (the acoustic properties of speech) features. However, it remains very difficult to compare the performance of these prototypes due to the lack of common databases and protocols. Machine understanding of expressions could revolutionise human-machine interaction and has, therefore, become a hot topic in computer-vision research.

1.1 Problem Overview

The question of how many emotional states we use in our daily communication has yet to be answered. Very little research has been done to locate and recognise emotions other than the six archetypal emotions as described by Ekman [7]. Therefore most approaches to automatic facial expression analysis attempt to recognise this small set of archetypal emotional expressions. This practice may follow from the work of Darwin [8] and more recently Ekman [7], who has performed extensive studies on human facial expressions. Ekman found evidence to support universality in facial expressions. These "universal (also referred to as 'archetypal' or 'basic') facial expressions" are those representing the six archetypal emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise.

The recognition of these six archetypal emotions has been done separately for every modality. The best results are achieved by looking at facial expressions, a good comparison of results is given in [9]. Further unimodal research has been done for audio, text and gesture recognition. However all these approaches use mostly their own, self created, datasets. This means that comparison between the different recognition and classification algorithms is difficult. As a consequence, there have been very little attempts to combine different modalities and approaches in order to fuse these single modalities into one multimodal emotion recognition system.

Multimodal fusion is the integration of the information present in these individual input modalities which carry information about the expression. Integrating, or fusing, these modalities requires understanding of how people use their various senses to perceive and interact with the world around them. It will depend on knowledge of the natural integration patterns that typify people's combined use of different input modes. This means that the successful design of multimodal systems will require guidance from cognitive science on the coordinated human perception. However, there has not yet been any consensus about the fusing method, the literature is inconclusive and the results differ not that much in order to say that one method is significantly better. Psychologists themselves are not sure how we as humans fuse the different channels of information.

The main challenge in multimodal emotion recognition is achieving a high recognition rate in various environments under different circumstances. This means that there can be a lot of noise present in a channel, e.g. occlusion in the face by a hand or glasses, the head is rotated in such a way that it is not perfectly visible or multiple persons are speaking at the same time. The advantage of more modalities present helps in better recognising the expression based on the fact that these different modalities can enhance each other. A dataset containing all these various environments and different circumstances is yet to be developed.

The lack of a widely used multimodal database with data suitable for emotion recognition for unimodal and multimodal systems made us gather data and develop such a database ourselves. The content from this database should be suitable for multimodal emotion recognition systems as well as unimodal emotion recognition systems, lip reading and vocal affect recognition. This to make it easier for different research groups to compare their results with other research groups.

Besides emotional content, a great deal of recordings for lip reading should be added to the database. These recordings should also help better recognise speech due to the high frame rate used [10]. Visual information seems to be the most natural source of additional information for speech recognition. In the case of continuous speech, when the speech rate tends to increase, a recording rate in the range of 24 to 30 frames per second is definitely insufficient. The performance decreases because interpolation is needed.

All the steps from the creation till the adding of content to the database should be carefully considered in order to get the best basis for all emotion recognition research groups. High speed cameras and sensitive microphones should be used to get the best quality recordings. Reducing the quality of the recordings can be done in a later stage, if necessary. The recording protocol should be suitable for the content to be added and the people being recorded should feel comfortable and relaxed while being recorded. Recording peoples response in a foreign language might not be a good approach. The response would be unnatural and the speech part is not fluently spoken, this can cause errors in training speech recognisers. The approach used in [11] showed us that non-native speakers give bad recordings as they have to think before responding. Besides this mental issue, the pronunciation of non-native English speakers can cause a difference in pitch variation due to the fact that their native language uses different levels of pitch or intonation. Therefore the database is, for now, filled with native Dutch speakers only, as there are no large numbers of native English speakers available.

Furthermore, it is interesting to see whether humans rate images, audio clips and video clips from the same recording as alike. Differences between modalities should be eliminated when rating multimodal information. These hypothesis should be proven by letting people rate the different modalities and then compare the results. Confusion matrices should tell which emotions are alike and which of them are easy recognisable by humans as well as emotion recognition systems.

1.2 Research Goals

The main goal of this thesis is to find the most appropriate method to develop a multimodal database and use it to classify multimodal emotions. This area of research is currently very active, as context sensitive reasoning systems are increasingly popular. Human-Computer Interaction (HCI)

often suffers from a lack of context. Emotion recognition is a powerful way to determine context, and thus to enhance HCI.

Research on multimodal emotion recognition emphasises the need for a common multimodal database. Speech and signal processing groups are often working independent of each other and few joint audio-visual studies have been done so far. Here the lack of an existing multimodal database is the bottle neck for researchers. A recent attempt to fill the need for a multimodal emotion database was done by Douglas-Cowie et. al [12].

Although their approach focuses on a variety of different emotions, a database containing the six archetypal emotions defined by Ekman [13] is still preferred, as most of the existing systems aim at recognising this set of emotions.

Recognising emotions depends on extracting features from the channel we are working with. As mentioned before these channels are not always noise-free. So correctly extracting all required features can be troublesome. In this thesis we enhanced the visual channel by placing prominent markers on the face in such a way that we could extract the facial features correctly and in case of mismatched features we could use the symmetry of the face to reconstruct the missing features. We used special audio software for analysis of the audio signal and computation of audio features.

We propose a method for recording multimodal video clips and basic multimodal emotion recognition based on these recordings. The work is based on the work of Anna Wojdel [14] and we will use her recordings, supplemented with our own made recordings. We extract the image of the apex of the facial expression, the whole audio clip and the video clip and let users rank these on the valence and arousal scale. A comparison is made with the ranking of the textual labels from the Dictionary of Affect in Language (DAL). Special research questions in this proposal are:

- Can we successfully create a protocol for the creation of a multimodal database?
- Can we rank 21 different emotions on the valence and arousal scale?
- Can we automate the classification of 21 different emotions?
- Do users agree on the ranking of multimodal emotions compared to the textual ranking?

1.3 Project Approach

The approach to finish graduation consists of the following explained stages. The completion of a research assignment, recording data as a first step to fill a multimodal database, data validation applied to the afore mentioned recordings, explanation of the model and prototype used, results obtained by using this model and prototype and comparison of the results with the results obtained from the validation methods. These stages are described in the next few paragraphs.

1.3.1 Research Assignment

Multimodal emotion recognition can help computers to understand the complex intentions of humans when using human-computer interaction (HCI). Multimodal emotion recognition also provides humans with a natural way of communicating with machines. The research assignment gave a theoretical overview of the construction of a multimodal database which can be used as a basis for multimodal emotion recognition; ideally this multimodal database will become a benchmark for future research on multimodal emotion recognition.

The research assignment discusses unimodal emotion recognition and the corresponding recognition rates and compares these to multimodal emotion recognition and their recognition rates. The focus is then shifted on fusing the different modalities, in order to fully recognise multimodal emotions. As a conclusion, a comparison of the results of both approaches is given as well as

recommendations to extend the research from the six archetypal emotions to a set consisting of more than ten different emotions.

At this moment, comparing performances between emotion recognition systems is difficult which is partly due to the lack of shared datasets and shared evaluation standards. The future of emotion recognition is to shift the research from the unimodal approaches to a combined, multimodal, approach for emotion recognition, taking into account the modalities in which emotions can be expressed and unobtrusively be processed, knowing: Facial Expression Recognition (FER), Vocal Affect Recognition (VAR), Automated Speech Recognition (ASR) and Gesture Recognition (GR) systems and fusing the different modalities into one large emotion recognition system that can classify more than the six archetypal emotions.

1.3.2 Data Recording

As mentioned before the idea is to create a database which can be used by unimodal research groups as well as multimodal research groups. We asked native Dutch speakers to sit in front of our cameras and express emotions by facial expressions and speech. To get into the mood they had to pretend being in a situation presented by a short story or scenario. Two high speed cameras and sensitive microphones recorded the participant while expressing emotions. This data is stored in the database and should be accessible via web access.

Every participant signed a consent document, informing them of what would be done with the gathered data. Participants can decline to have their data be used for publishing, in a paper of presentation, or even decline to participate at any time. A minimum of 50 different participants from various backgrounds should provide us with the basis data for this database.

Every participant was asked to express all of the 21 different emotions. Besides these 21 different expressions the participants were asked to perform, just the facial expression belonging to the desired emotion. This resulted in 21 different facial expressions. These emotions were selected after the product emotions described in [15]. The stories which should bring the participants in the mood of the emotion were based on the protocol used in [11]. As this approach only used stories or scenarios for the six archetypal emotions some of these were translated and used in our protocol. Many of the stories or scenarios and possible answers are made especially for this experiment and carefully checked if they were correct and contained enough phoneme coverage for the Dutch language by specialists.

For the selection of participants we decided to use both actors and students or university personnel. In this case you get the real exaggerated expressions and the more timid ones. There was no wrong way of expressing oneself, but data validation must be done on the gathered data in order to have some consensus over the labels given to the data.

1.3.3 Data Validation

Validation of the gathered data was done online. Users were asked to visit a web site where they could give scores for valence and arousal on a 9-point scale to images, audio and video clips. For every emotion there were extracted images showing the apex of the expressed emotion, audio clips and video clips present. These obtained results were then compared with the results given by the 'Dictionary of Affect in Language' (DAL) database [16] for every word describing the emotion. Both scales were transformed to a 9-point scale. If a word came up empty a comparing word was searched for until a result was given (e.g. 'Amusement' → 'Amusing' and 'Fascination' → 'Fascinating').

1.3.4 Model

We detect 31 predefined facial feature points. If we do not find all points, the missing points can be automatically generated according to the position of other points located in a certain position from the missing point. We can determine the rotation of the face by calculating the angle between two key points, as these two points lay in the same plane and should always be perpendicular to the horizon for a straight face. We then normalise the points such that the distance between two key points, feature point 1 (the eye point) and feature point 20 (the nose point), is 50 pixels. In this case it does not matter how big the image is. All pre-process steps are visualised in Figure 24.

1.3.5 Prototype

The developed prototype is designed as a simple tool for recognising 21 different multimodal emotions from audio and facial expressions. Matlab is used as the programming environment, because of the fast handling of matrices and the use of many toolboxes available. A screenshot of the prototype is shown in Figure 2.

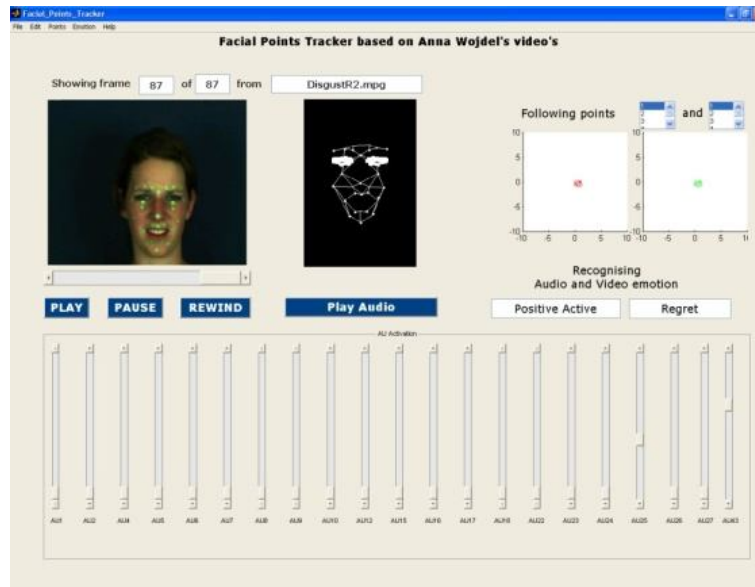


Figure 2: Screenshot of the prototype.

This prototype localises and tracks a number of facial feature points. In total we extract 31 facial feature points from the face, all marked with stickers, to ease the recognition process. Each point can be tracked separately and the displacement can be visualised. The recognised emotions come from five different recognition methods. We try to classify the emotion according the three audio classifiers and two video classifiers. We make a distinction between 21 classes, positive and negative classes and active and passive classes for training each of the three audio classifiers. For the video classifiers we make use of the Pearson distance for classification directly applied to the extracted facial feature points and AU activation for emotion classification based on AU activation.

1.3.6 Results

The results from the prototype are compared to the results acquired from the data validation methods. We expect the results to be similar for most emotions. However, some emotions, like contempt could

give derogatory scores, as multiple channels give contrary information. The comparison of all results should give a clear picture of which emotions are really dependant on multimodal information.

Furthermore we measured the number of correct processed frames. A correct processed frame is a frame in which the facial feature points are located correctly and we could extract useful information from these locations. For most clips the prototype processed the frames correctly, but for 8 of the 105 clips the processing rate was below 25%.

1.4 Thesis Structure

In the first two chapters of this thesis, we introduce knowledge related to the research presented further. In Chapter 2 we will briefly give an overview of the related research in this area. We give an example of a widely used database which you might say is a benchmark database for Facial Expression Recognition (FER). After that we will give two examples of attempts to create a benchmark database for Multimodal Emotion Recognition (MER). As computers get more and more computational power the challenge to recognise emotions not only from one modality, like facial expressions, but from multimodal channels is the real challenge for the future. Several methods have proven to give good results, but the lack of a common database to compare these results make the comparison between these methods impossible.

In Chapter 3 we will explain the theory that describes emotions and emotion recognition. We start with an introduction to emotions. What are emotions and how do they influence our communication? We continue with facial analysis and present the Facial Action Coding System (FACS). Then we discuss Vocal Affect Recognition (VAR), emotion recognition from sound. Our voice is the main channel of information during face-to-face communication. Next, we discuss various functions of facial expressions in face-to-face communication. At the end of this chapter, we will give an overview of the current research into multimodal emotion recognition. We briefly handle the different fusing methods and emphasise again the need for a commonly used multimodal database which can serve as a benchmark for future research.

Chapter 4 covers the creation of such a multimodal database. The content of this database should be clearly described as well as the protocol to record the different video clips, the storage and offering of the recorded data to the public. The question to use genuine or acted emotions is answered and the different emotions recorded are summed up. A detailed description of the recording protocol is given, as this is the main building block of our database. Chapter 5 is all about the validation of the recorded data. This chapter describes the evaluation space, the data validation methods and concludes with the validation results. Three different validation approaches have been tested. Two unimodal approaches, image only and audio only and one multimodal approach, the whole video clip.

The research described in Chapter 6 explains the working and implementation of the multimodal emotion recogniser. This system uses 31 pre-defined points on the face. These points are tracked and feature vectors are constructed. Chapter 7 continues with data and noise reduction. After that the emotion can be recognised using the rules extracted beforehand from the training samples. This chapter also includes a description of the methods for transforming the changes on the facial surface, resulting from changes that occur when expressing an emotion, into activation of different parameters (AUs).

The results obtained from the previous chapter are discussed in Chapter 8. The methods used are described and the recognition rates are given. Furthermore, these results are compared with the results obtained from the data validation from Chapter 5.

In Chapter 9 we evaluate the whole project. Here we will give some conclusions and future work recommendations. Besides that we will also briefly evaluate the effort that we put into the recordings for the lip reading research done at the TU Delft.

CHAPTER 2

Overview of Related Work

In this chapter we will give an overview of related work to our research. In the following paragraphs we will discuss the previous attempt to fulfil the need for a common multimodal database which can be used in a wide range of research areas. The most well known example is the Cohn-Kanade database [17].

2.1 The Cohn-Kanade Database

The Cohn-Kanade database is the most common database used in emotion recognition from facial expressions. We cover this database because we think that the creators did very well in their approach to offer the database to various groups of researchers and thus made this dataset a common benchmark for current facial expression systems. The continuing adding of videos and images should ensure that this database will be around for quite a while and if the administrators keep alert they can even make the step to multimodal emotion recognition.

2.1.1 The Cohn-Kanade Approach

Recent attempts in facial expression recognition try to recognise a small set of prototypic expressions. However, to capture the subtlety of human facial expression a fine-grained description of facial expressions is needed. The FACS system [18] is a human-observer based system designed to detect subtle changes in facial features. A very common assumption in earlier research is that expressions are singular and begin and end with a neutral position. Transitions between action units may involve no interleaving neutral state. In order to be comprehensive a database should include individual action units and both additive and non-additive combinations.

The chance of having spontaneous emotions in the database is small, as most facial expression databases ask subjects to perform an expression. These ‘acted’ expressions may differ in appearance and timing from spontaneous occurring emotions. Thus, fine-motor control of deliberate facial expressions is often inferior and less symmetric to spontaneous expressions. Therefore, a comprehensive database should include both.

Occlusion or partial occlusion of the face can be the case with beards, eyeglasses or jewellery. These differences can have great consequences for face analysis. To develop robust algorithms that can handle individual differences it is essential to include a large sample of varying of ethnic background, age, sex, (partially) occluded faces and even clinically impaired individuals. Even face orientation towards the camera, or the presence of other people in the background, are factors that can

also influence face analysis. Most researchers assume that face orientation is limited to in-plane variation. In reality, large out-of-plane variation in head position is common and often accompanies a change in expression. Image data where facial expression changes in combination with limited planar change is therefore needed. The same holds for interaction on the background. Reality shows that there are almost always people present on the background. To get robust results, this variation must be included in the training data.

This approach, asked experts to label the performed expressions, as they cannot assume that the given expression is genuine. Asking the subject to perform a given action is no guarantee that they will. To allow for comparative tests of alternative approaches to facial expression analysis, appropriate data must be made available to the face analysis community. All subjects were filmed with two camera's both connected to a VCR with a Horita Synchronised time-code generator. One of the cameras was located directly in front of the subject, the other at a 30 degrees angle. Different lighting circumstances were used, one third was lit by ambient room lighting by means of a high-intensity lamp and the other two third had two high-intensity lamps with reflective umbrellas. The subjects were asked to perform a series of 23 facial displays; these included single action units and combinations of action units.

2.1.2 The Cohn-Kanade Drawbacks

Although the Cohn-Kanade database has succeeded in creating a comprehensive image/video database for facial expression recognition purposes, it still does not handle auditory or gesture modalities which are very useful in multimodal emotion recognition. While the authors are still extending the database with more data they are also adding data on behalf of emotion research. This can be promising for the facial expression recognition community that does not depend upon or work with the FACS system. The added value of genuine emotions in this database makes it the current benchmark for all image based facial expression recognition groups. If this database will proceed to the next level of emotion recognition it has to add multimodal recordings to their dataset. Unfortunately multimodal emotion recognition systems have little or no use from the Cohn-Kanade database, yet.

2.2 The e'NTERFACE'05 Audio-Visual Emotion Database

Driven by the idea that the speech and image signal processing communities are often working independently from each other, and relatively few joint audio-visual studies of emotions have been conducted so far, the e'NTERFACE attempts to create a database that can provide a future basis for multimodal emotion recognition [11]. The approach used to construct this database is a useful one. It emphasises to make subjects express 'genuine' emotions in a controlled way.

2.2.1 The e'NTERFACE Approach

The basis for the recordings of the e'NTERFACE database was to get as close as possible to spontaneous emotions, while keeping at the same time a fully controlled recording environment. Allowing the subject to react in its own language has a main drawback: the prosodic features (such as pitch variations, speaking rate, etc...) largely depend on the language itself. For example, the speaking rate is typically higher for an Italian than for a French speaking subject. As one of the goals of the database is to have prosodic features that depend only on the emotion that is expressed, the choice to conduct all experiments in English was made.

Furthermore, the fact that the subjects had to think about what their reactions should be and then translate those reactions into English led to totally unnatural utterances. Therefore, it was chosen to use pre-defined answers for each situation. The subject was first asked to listen carefully to a short

story and to ‘immerge’ into the situation. Once ready, the subject may read, memorise and pronounce (one at the time) the five proposed utterances as reactions to the given situation. The subjects are asked to put as much expressiveness as possible into their reaction, producing a message that contains only the emotion to be elicited. Expressing emotions ‘on demand’ is a very difficult task. In addition, none of the subjects were an actor. When asked to express their selves emotively, some subjects performed pretty well while other totally failed to express the requested emotions.

To take this variety of outcomes into account and still end up with a high-quality database, it was necessary to apply a post processing step to the set of original recordings: human experts had to examine carefully each recorded sample.

The database is currently filled with recordings expressing the six archetypal emotions. These recordings were taken from 46 subjects and these subjects were asked to react to six different situations, each of them eliciting one of the following emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise. As a data validation step there were two human experts who decided whether or not the subject had expressed itself in such a way that an untrained human observer could without ambiguity recognise the emotion present in the reaction.

2.2.2 The e’NTERFACE Drawbacks

Although the e’NTERFACE database is a good attempt to construct a multimodal corpus, the usage of non-native speakers resulted in the opposite of what a multimodal database should look like. The visual expressions can be genuine, however, the vocal utterances are poorly expressed and result in a poor basis for vocal affect recognition and thus for multimodal emotion recognition.

Besides the poor basis for vocal affect recognition the limitation to only six archetypal emotions can eventually lower the researchers’ interest for this database. As there is increasingly more research done towards other emotions and even blended emotions. A set containing only six emotions is too limited to work with and does not represent the emotional states users commonly express.

2.3 A New Emotion Database

The aim of the European PHYSTA project is to develop a system that will recognise emotion from facial and vocal signs. The rationale behind the fact that most pre-existing databases consist of examples representing a few archetypal states is rarely spelled out, but the only obvious way to justify it is to postulate that the whole space of emotional signs can be reconstructed from information about a few cardinal types. This hypothesis is called the benign interpolation hypothesis. To achieve ecological validity this database used four considerations [12]. Keep in mind that the database’s main concern is with speech.

- 1 **Genuine emotions** The core design was to use material generated by people experiencing genuine emotions. Here acted emotions cannot be a sufficient basis for conclusions about the expression of emotion.
- 2 **Emotion interaction** This database focuses on examples derived from people engaged in human interactions.
- 3 **Gradation** Ecological validity entails sampling situations where emotion is mixed or controlled in the ways that typically occur in everyday life.
- 4 **Richness** Ecological validity entails collecting samples which make it possible to study whether those elements are effectively independent or interactive, and how they evolve in time.

Existing databases have not generally been developed with ecological considerations in mind. Theory and practice perhaps reinforce each other. For example if one is dealing with only a few clear-

cut primary emotions, then it is natural to think of generating that kind of clear-cut data in artificial contexts.

2.3.1 The New Emotion Database Approach

The database should contain genuine emotional states including archetypal and other states, involve both modalities (audio and video) and allow exploration of emotion over time. Two main sources were used for this, one were studio recordings and the other recorded programs from British television. The studio recordings consisted of recordings of a conversation between postgraduate students on topics that provoked strong feelings. This approach was not further pursued because subjects' behaviour was generally very constrained. A second attempt was done recording one to one interactions. Each session lasted about 1-2 hours in an informal setting as possible. The recordings were on purpose very long, because subjects tend to relax after an hour and speak more freely. The interviewer in this case knew prior knowledge of each subject to get the conversation going. Most of the useful recordings were labelled mild; they rarely showed dramatic signs of emotion.

After watching a range of television programs over a period of several months the authors identified four useful types of programs, knowing: chat shows, religious programs, reality shows and current affairs programs. Chat shows provided the most obvious emotional material, but the emotions tended to be limited to negative emotions. Religious programs are a source of positive emotions. Current affairs programs provided material related to deaths resulting from food poisoning. The reality show "The Village" gave emotional material of both positive and negative emotions. The target recordings should start with an emotional neutral state evolving in a emotion for a substantial period of time. This included emotional states that were not that extreme. Mixed emotions were included as well, when the signs were strong enough. These recording were the basis for the clips in the database, which were no longer in length than 60 seconds. The clips were saved as MPEG-files, separate sound files as WAV-files.

The database incorporates two types of description for the emotional content of each clip, these are dimensional and categorical. The dimensional type describes the emotional content in terms of activation-evaluation space. Categorical labels are given in two forms, one being the desired expressed emotion (e.g. Anger) the other being a list of associated words to the emotion the user gave that clip. The intensity was also rated on a 1-3 scale.

2.3.2 The New Emotion Database Drawbacks

The approach to fill the database with television recordings is a good approach, but it is very time consuming. The recordings of the one to one sessions are a good and controlled alternative for the television recordings. As long as there will occur full blown emotions in the one to one recordings, these can be as real as in everyday life. The inclusion of separate audio files gives this database a broader range of use.

The given metadata to the clips is extensive, which makes searching the clips easier, but more or better annotation of the recordings could benefit this database. The lack of use of the FACS system does not encourage researchers to use this data for emotion recognition, as FACS still is the best way to model facial changes and thus recognise facial emotions.

2.4 State of the Art Multimodal Emotion Recognition

Here we will focus on reviewing the efforts toward audio-visual emotion recognition, especially those done in the last few years. The first studies for audio-visual emotion recognition include Chen [19] [20]Huang [21] and De Silva [22] [23]. In the past few years, there are an increasing number of reports investigating the integration of emotion-related information from audio and visual channels in

order to improve the recognition performance [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34]. An overview of the results of recent attempts on multimodal emotion recognition is shown in Table 1. The unimodal recognition rates before the fusion are also shown.

The fusion methods differ with the approach used. An example fusion at feature-level fusion is shown in Figure 3 [35]. The results obtained with different fusion methods do not differ very much. There does not seem to be a fundamental basis to choose for one fusion approach, as shown in [36] the fusion method seems to be context dependent.

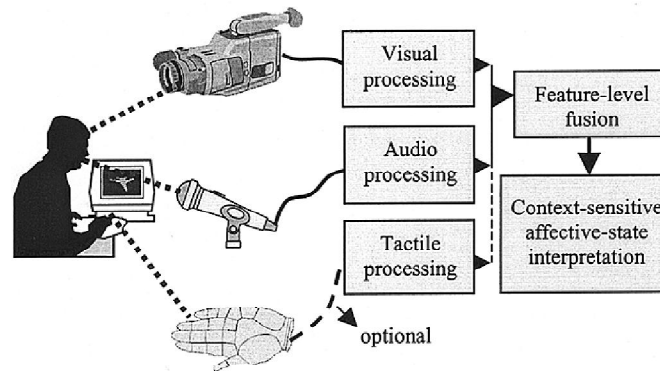


Figure 3: Architecture of an “ideal” automatic analyser of human affective feedback.

The question which modality is the stronger one does not hold for multimodal emotion recognition. There is evidence that some emotions are expressed stronger in audio and others in facial expressions [20]. Anger, happiness, surprise and dislike are better recognised visually and sadness and fear are better recognised in audio. Since some modalities may carry complementary information and since humans also make use of multimodal information, the logical conclusion is to fuse different modalities, which may eventually lead to higher classification accuracies. A comparison of fusion methods is given in [19].

Table 1: Recognition rates for recent attempts of multimodal emotion recognition

Paper	Database (subjects M/F)	Method	Fusion level	Emotions	Testing method	Recognition Rate (in %)	Comments
[24]	Unknown (10/10)	kMeans	Audio only	6 basic emotions	Person dependent	93.3/93.3	Male/Female
		LDA	Video only			90/95	
		Codebook	Decision			95/98.33	
[25]	Unknown (10/10)	HMM	Audio only	ang, bore, conf, dis, fea, frus, hap, int, neu, sad, sur	Person Independent	62.89	
		HMM	Video only			38.64	
		MFHMM	Feature			83.64	
[26]	Unknown (10/10)	HMM	Audio only	ang, bore, conf, dis, fea, frus, hap, int, neu, sad, sur	Person Independent	61.82	
		HMM	Video only			38.64	
		MFHMM	Feature			80.61	
[27]	Unknown (10/10)	HMM	Audio only	ang, bore, conf, dis, fea, frus, hap, int, neu, sad, sur	Person Independent	64.50	
		HMM	Video only			39	
		MFHMM	Decision			75	
[28]	Unknown (10/10)	?	Audio only	Pos/Neg	Person Dependent	?	Measured emotion on activation-evaluation axes
		?	Video only			?	
		Fisher Boosting	Feature			85.5	
[29]	Unknown (1/1)	HMM	Audio only	Pos/Neg emotions	Person Dependent	65.15/75.09	Male/Female
		LPP HMM	Video only			84.85/87.50	
		AdaBoost MHMM	Feature			89.39/90.36	
[30]	Unknown	HMM	Audio only	6 basic emotions + neu	Person dependent	68.42	
		HMM	Video only			82.52	
		THMM	Feature			90.35	
[31]	Unknown (1 actress)	SVM	Audio only	ang, hap, neu, sad	Person dependent	70.9	
			Video only			85	
			Feature/Decision			89.1/89.0	
[32]	Unknown (7)	SVM	Audio only	Pos/Neg + Neu	Person Dependent	86.8	
			Video only			66.8	
			Decision			90.7	
[33]	Unknown (8)	LDA	Audio only	6 basic emotions	Person Independent	66.43	
			Video only			49.29	
			Decision			82.14	
[34]	Unknown (38)	DBN	Audio only	ang, bore, conf, dis, fea, frus, hap, int, neu, sad, sur	Person dependent	56	
			Video only			45	
			Feature			90	

ang: Anger; bore: Boredom; conf: Confused; dis: Disgust; fea: Fear; frus: Frustrated; hap: Happiness; int: Interest; neu: Neutral; sad: Sadness; sur: Surprise.

CHAPTER 3

Emotions and Emotion Recognition

Emotions are a ritual of our daily routine. They are our main motivators. An emotion gives a human face a very complex structure. Emotions give strength to our way of communicating. Facial expressions are not always a reflection of our emotional state, but can give extra meaning to the content of the message. The human body [5], and specifically the human face [37], provide a lot of conversational information. Facial expressions can supplement text; add an emotional state to the information which helps us to understand a message according to the intention of the speaker. The same holds for vocal affect. Vocal affect can give an emotional load to words or sentences, the intonation or intensity of the spoken utterance gives us clues about the emotional state of the speaker.

Paragraph 3.1 presents basic knowledge about emotions, facial communication and emotions that play a role in face-to-face communication. Besides that, some facial gestures can even replace words as e.g. an act of nodding the head can replace a verbal confirmation. In fact, not only the speaker uses facial expressions, but the listener can give nonverbal feedback via facial expressions too. Paragraph 3.2 tells us something about Facial Expressions Recognition and the different approaches to recognise facial expressions and emotions in general. In this paragraph we also present the Facial Action Coding System (FACS) [18]. FACS is a facial expressions annotation system which is used in both analysis and synthesis of facial expressions and is used by most researchers who cover this area of research. Paragraph 3.3 describes the process of Vocal Affect Recognition. Emotion classification from audio clips has been done since the 1930's. Finally, Paragraph 3.4 goes into Multimodal Emotion Recognition. The two fusion methods of the different unimodal channels are explained.

3.1 Emotions

Humans use a daunting number of labels to describe emotion [38]. Therefore most approaches to automatic facial expression recognition attempt to recognise a small set of archetypal emotional facial expressions. This practice may follow from the work of Darwin [8] and more recently Ekman [7], who has performed extensive studies of human facial expressions. Ekman found evidence to support universality in facial expressions [13]. These “universal facial expressions” are those representing the six ‘archetypal emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise.

Labelling the emotions in discrete categories, such as the six archetypal emotions sometimes is too restricted. One problem with this approach is that the stimuli may contain blended emotions. The choice of words may be too restrictive or culturally dependent too.

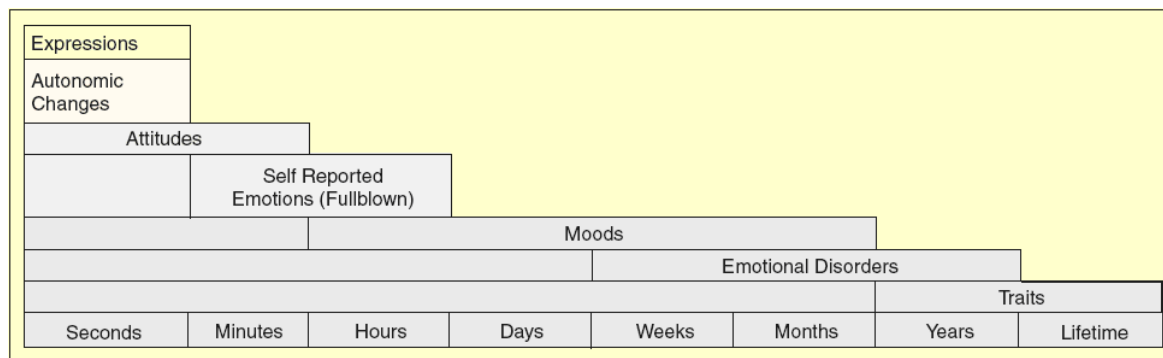


Figure 4: Duration of different emotional states.

There is little agreement about a definition of emotion. Many theories of emotion have been proposed. Some of these could not be verified until recently when measurements of some physiological signals become available. In general, emotions are short-term, whereas moods are long-term, and temperaments or personalities are very long-term (see Figure 4) [38].

3.1.1 Face-to-Face Communication

In everyday face-to-face communication the human face changes all the time. As explained before people use their face to emphasise the meaning of the message, as a consequence people show a large variety of facial expressions – not only the ones corresponding to the archetypal emotions described in Paragraph 3.1. Sometimes, the exact same changes in facial appearance that are related to some emotion can fulfil also other communicative functions. It is important to remember that the same facial expression, used in different contexts, will have various meanings.

3.1.2 Visualisation of Emotions

Research from Darwin forward has recognised that emotional states involve dispositions to act in certain ways. The various states that can be expressed are simply rated in terms of the associated activation level. Instead of choosing discrete labels, observers can indicate their impression of each stimulus on several continuous scales. Two common scales are valence and arousal. Valence describes the pleasantness of the stimuli, with positive (or pleasant) on one end, and negative (or unpleasant) on the other. The other dimension is arousal or activation, see Figure 5 [39]. The vertical axis shows activation level (arousal) and the horizontal axis evaluation (valence). A circumplex can be viewed as implying circular order, such that variables that fall close together are more related than variables that fall further apart on the circle, with opposite variables being negatively related and variables at right angles being unrelated (orthogonal).

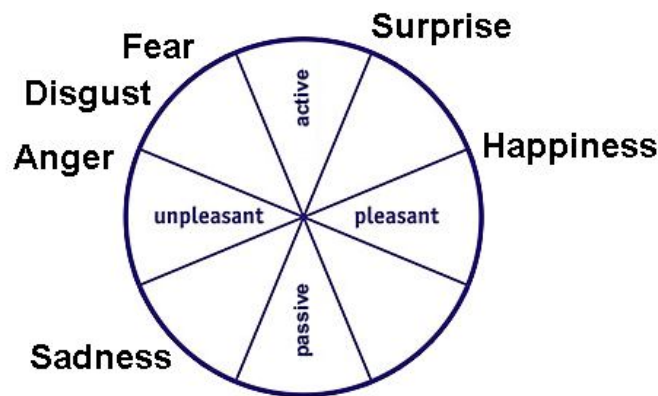


Figure 5: Activation-Evaluation space of the six basic emotions.

We can identify the centre as a natural origin or the neutral state. The neutral state is a condition of readiness to respond. Emotional strength can be measured as the distance from the origin to a given point in the activation-evaluation space. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation. A related extension is to think of the six archetypal emotions as cardinal points in that space.

3.2 Facial Expressions Analysis

In order to detect and analyse facial expressions we need first to define how to describe these emotional states expressed on our face. The prime example of the codified version of facial expressions is the Facial Action Coding System (FACS), widely used by psychologists. Another way of describing facial activity is to give some quantitative description in terms of geometrical changes of the face. Such geometry changes can be described by using an MPEG-4 standard with its Facial Animation Parameters (FAPs) [40] [41].

Other prominent components of the face are the eyes, teeth and tongue. It is the eyes to which people pay a lot of attention during a conversation. The eyeball is generally white with a black pupil positioned in the centre of the visible part of the eyeball, and surrounded by a colourful iris. Every iris is different; it varies in colour and structure between individuals. Its task is to regulate the amount of light passing through the lens by controlling the size of the pupil. The teeth and tongue play a minor role in everyday face-to-face communication. The teeth are visible only when the mouth is open, and although they do not attract as much attention as the eyes, they are very important objects in speech processing.

3.2.1 The Facial Action Coding System (FACS)

The most commonly used system for measuring and describing facial behaviours is the Facial Action Coding System (FACS) [18]. Ekman and Friesen developed the original FACS in the 1970s by determining how the contraction of each facial muscle (singly and in combination with other muscles) changes the appearance of the face. This facial activity is described in terms of visually observable facial muscle actions (i.e., action units, AUs). With FACS, a human observer decomposes a shown facial expression into one or more of in total 32 AUs that produced the expression. As an addition 12 dispositions have been added to the FACS, these include the head and eye movements.

FACS is a structure-based coding, closely connected to the anatomy of the face. The obtained facial expression scoring is universal across a broad spectrum of faces. Therefore FACS is widely used by psychology researchers, and it is also very common among researchers that work with facial expression analysis by machines [42] [43]. Currently FACS is a leading method used in behavioural

investigations of emotion, cognitive processes, and social interaction. Over 20 years of psychological research on the relationship between action units and facial expressions provided a lot of data about facial behaviour expressed in terms of facial action codes. FACS was used, for example, to analyse differences between facial expressions of people lying and telling the truth [6] or to demonstrate facial signals of interest and boredom. And, what is very interesting for us, there is also a lot of information available about nonverbal conversational signals that, for example, emphasise the verbal part of speech, or regulate the flow of conversation.

3.2.2 Facial Expression Recognition

Every facial expression has its own features. By knowing these features and the locations of the major changes in the face, caused by a certain expression, it is easier to link a face to an emotion. Contractions of facial muscles change the appearance of permanent and transient facial features. Here permanent facial features are the visible components that make up the face such as eyebrows, eyes and mouth. Their shape and location can alter immensely while expressing an expression. Transient facial features are any facial wrinkles that did not become permanent with age but only appear within an expression [42]. These features are most prominent when the emotion is at its peak: the apex of the expression. There is more information in video than just these static features: the direction and the speed of the movements of the feature points (and the face itself) give considerable information about the expressed emotion.

There are many different methods for recognising facial expressions. The approaches can be roughly divided into two main groups, the feature based methods and the template based methods. The feature based methods use a set of texture and geometrical information as features where the template based methods use 2D or 3D head and facial models as templates for facial expression recognition. Many different research groups follow different approaches and claim to get the best results, but most of the methods have only been tested on a single facial expression database.

One of the feature based methods are Dynamic Bayesian Networks (DBNs). A Bayesian network (or a belief network) is a probabilistic model that represents a set of variables and their probabilistic interdependencies. The papers [44] [45] [46] [47] [48] [49] [50] [51] follow this approach. Another method used are the Hidden Markov Models (HMMs). A HMM is a statistical model in which the system being modelled is assumed to be a Markov process with un-observational states. The challenge here is to determine the hidden parameters from the observable parameters. A HMM can be considered as the simplest DBN. This method is pursued in [51] [52], [53] [54] [55].

Probably the most used method for facial expression recognition is the Support Vector Machines (SVMs). SVMs are a set of related supervised learning methods used for classification and regression. They belong to a family of generalised linear classifiers. SVMs map input vectors to a higher dimensional space where a separating hyperplane is constructed. The separating hyperplane is the hyperplane that maximises the distance between the two closest support vectors of each class. The larger the distance is to the separating hyperplane from the support vectors the smaller the generalization error of the classifier will be. SVMs are used in [56] [57] [58] [59] [60] [61] [62]. Derived from the SVMs are the Relevance Vector Machines (RVMs). A RVM is a machine learning technique that uses Bayesian theory to obtain sparse solutions for regression and classification. The RVM is based on the SVM, but provides probabilistic classification. In [63] a RVM is used to classify facial expressions from still images.

A whole different method is the Gabor wavelet or Gabor filter. A wavelet is a mathematical function used to divide a given function into different frequency components. A wavelet transform is the representation of a function by wavelets. The copies, so called “daughter wavelets”, are scaled and translated copies of the “mother wavelet”, a finite-length or fast-decaying oscillating waveform. Gabor filters are directly related to Gabor wavelets. There is no expansion applied to Gabor Filters as this is very time-consuming. Therefore a filter bank consisting of Gabor filters with various scales and rotations is created beforehand. These methods are used in [64] [65] [66] [67].

A Neural Network (NN) is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing. In most cases a NN is an adaptive system that changes its structure, or weights, based on external or internal information that flows through the network. These NNs are used in [67] [68] [69] [70].

Linear Discriminant Analysis (LDA) is a statistical method used in statistics and machine learning to find the linear combination of features which separate two or more classes best. The resulting combination may be used as a linear classifier. LDA is used in [71] [72]. Principal components analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA involves the calculation of the eigenvalue decomposition or Singular value decomposition of a data set. PCA is used in [71] [73]. A Linear Programming (LP) method is used in [74].

An example of a template based method is the Active Appearance Model (AAM), AAM is a statistical-based successful method for matching a combined model of shape and texture to new unseen faces. This model is used in [75]. Template matching is also done in [76], here in combination with LBPs. The used methods in [77] [78] [79] [80] [81] make use of 3-D models in order to recognise facial expressions. The templates used can either be a 2-D or 3-D model of the face. The method used in [35] is exceptional and uses two 2-D face models (frontal and profile view).

Not all classifiers perform even well; this difference can be assigned to the fact that not every database is the same. Training one classifier on one single database can make this classifier perform worse on another database. Besides the fact of differences in databases there are strong indications that some classifiers are better than the others. But here, also, there are a few exceptions [71].

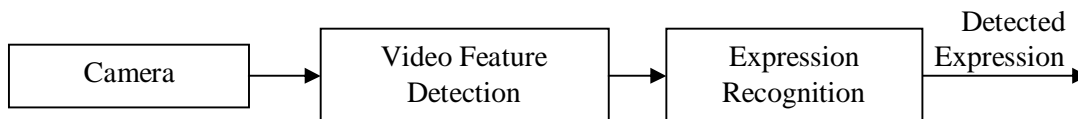


Diagram 1: Simple example of a unimodal image based emotion recognition system.

3.3 Vocal Affect Recognition

The vocal aspect of a communicative message carries various kinds of information. If we disregard the manner in which the message was spoken and consider the verbal part (e.g., words) only, we might miss the important aspects of the pertinent utterance and we might even completely misunderstand what was the meaning of the message. Nevertheless, in contrast to spoken language processing, which has recently witnessed significant advances, the processing of emotional speech has not been widely explored.

To estimate a user's emotion by the speech signal one has to carefully select suited features. Common features include the values of pitch, intensity and spectral features and all of these within a 10 ms or 15 ms time window.

The main energy source in speech is vibration of the vocal cords. The vocal cord can vibrate at any given time with any given rate. The rate at which vocal cords vibrate determines the fundamental frequency of the acoustic signal, the so called 'F0'. Traditional as well as most recent studies in emotional contents have used "prosodic" information which includes the pitch, duration, and intensity of the utterance. Variations in voice pitch are considered to have a linguistic function. Pitch features are statistical properties of the pitch contour. The set of spectral features is comprised by statistical properties of the first 4 formants and the energy below 250 Hz. A formant is a concentration of acoustic energy around a particular frequency in the speech wave which results from the resonant frequencies of any acoustical system. Formants occur roughly at 1000 Hz intervals. Each formant corresponds to a resonance in the vocal tract. The speech intensity depends primarily on the amplitude of vocal cord vibrations which is related to the pressure of the air stream. The larger the expiratory effort, the larger the intensity. An extended list of vocal features is given by Ververidis in [82].

3.3.1 Automatic Vocal Affect Recognition

There has been less work on recognising human vocal emotions by computers than there has been on recognising facial expressions by machine. Studies in [83] extracted five features from speech and used a multilayered neural network for the classification. For 20 test sentences, they were able to correctly label all three categories. In [84] they used 17 features and compared different classification algorithms and feature selection methods. They achieved a recognition rate of 79.5% with only 4 categories and 5 speakers speaking 50 short sentences per category. A comparison in [85] showed that human and machine recognition of emotions in speech achieve similar recognition rates (around 65%). In that work, 30 subjects spoke 4 sentences, with each sentence repeated 5 times, once for each emotion category. In this study, a large-scale study using 14 professional actors, he extracted as many as 29 features from the speech. According to [86], human ability to recognise emotions from purely vocal stimuli is about 60%. Looking at the findings we can say that the best recognition rates come from sadness and anger, followed by fear and joy. Disgust has the lowest recognition rates.



Diagram 2: Simple example of a unimodal speech based emotion recognition system.

3.4 Multimodal Emotion Recognition

The studies in facial expression recognition and vocal affect recognition have been done largely independent of each other. The aforementioned works in facial expression recognition used still photographs or video sequences where the subject expresses only facial expression without an emotionally loaded utterance. Similarly, the works on speech based emotion detection used only the audio information. There are situations where people would speak and exhibit facial expressions at the same time. For example, “he said hello with a smile”, here facial expression recognisers may fail due to the fact that the mouth movements may not fit the description of a pure “smile”. For computers to be able to recognise emotional expression in practical scenarios, these cases must be handled.

3.4.1 Fusion of Different Modalities

Multimodal fusion is the integration of the information present in individual input modalities, here audio and video features carrying information about emotions. Integrating modalities requires understanding how people use their various senses to perceive and interact with the world around them. It will depend on knowledge of the natural integration patterns that typify people’s combined use of different input modes. This means that the successful design of multimodal systems will require guidance from cognitive science on the coordinated human perception. However, exactly how the fusion of audio and visual information takes place in human perception is not yet answered. In particular, the existing studies do not agree on the central question of at which stage the fusion occurs. The various existing models can be roughly categorised into two groups:

- Early integration models.
- Late integration models.

In the perspective of the early integration models, or feature-level fusing, fusion takes place before the recognition stage. These methods typically use only one classifier, combining the features from the different modalities as its input. The other class of methods carries out fusion in the decision-level, which are referred as late integration or decision fusion methods. In these methods the likelihood scores of the single-modality classifiers of each modality are calculated independently from one another, and the fusion is carried out by combining the decisions given by the parallel channels.

3.4.2 Decision-Level Fusion

Decision fusion (late integration) is most commonly found in HCI and can be described as fusion of independent mode interpretations (or decisions). This level of fusion is mostly applied to modalities that differ in the time scale characteristics of their features. Timing plays an important role and hence all fragments of the modalities involved are time-stamped and further integrated in conformity with some temporal neighbourhood condition. In its classic form, decision-level fusion is based on the naïve Bayes paradigm. It makes the assumption that the channels are conditionally independent given the class label, thus the joint likelihood of the observations from the audio and visual channels can be factorised as follows:

$$P(X_a \cdot X_v | C) = P(X_a | C) \cdot P(X_v | C) \quad (1)$$

where variables X_a and X_v are the observations from the audio and the visual channels respectively and C is the class label. The most commonly used statistical approaches for late fusion are [87] [88]:

- Product rule.
- Sum rule.
- Max / Min / Median rule.
- Majority vote.
- Adaptation of weights.

Difficulties

There are some difficulties in using decision-level fusing. First of all, the assumed independency of the channels is often not true. The modalities might not be independent of each other, i.e. speech and lip movements or speech and manual gestures are known to be highly correlated. This is of course also true with emotions in audio and video.

Second, to combine audio and video to classify emotions, one needs to have a great knowledge of cognitive science and psychology. For instance, to define a set of rules and give weights to the different modalities is not an easy task to do. Some emotions are better identified with audio such as sadness and fear, and others with video, such as anger and happiness [89]. Computer scientists often lack this knowledge, sometimes existing studies do not even agree on these facts.

The third difficulty in using decision-level fusing is redundancy. In multimodal systems, complementary input modalities provide the system with no redundant information whereas redundant input modalities allow increasing both the accuracy of the fused information by reducing overall uncertainty and the reliability of the system in case of noisy information from a single modality [90]. Complementary modalities need to be merged to result in a meaningful command. Redundant modalities produce the same command either combined or taken separately. If the modalities are parallel they might introduce redundancy to the system. In this case the fusion component needs rules

that identify those redundant hypotheses. One of two user's actions must be ignored if not processed simultaneously.

Redundancy is not a problem in this system, as it wants to classify only one emotion. It is ok if the system recognises the emotion multiple times, that will even make it more robust. But in multimodal systems that need to classify input into actions, this is a problem as the action should not be performed twice.

3.4.3 Feature-Level Fusion

In feature-level fusion (early fusion), the fusion takes place before the recognition stage. This type of fusion is performed by concatenating the feature vectors from each modality and using a single classifier, which uses the combined information to assign likelihoods to the recogniser's hypotheses. Typically, early integration architectures assume a strict time synchronicity of the modalities. To minimise the errors, some adaptation strategy can be adopted (e.g. weighting coefficients) [91]. Early fusion enables usage of some relationship between the different channels, for classification. Kapoor and Picard [92] claim that this level of fusion can be useful in cases like audio-video fusion. However, when fusing the multimodal information at the feature-level, feature sets can get quite large. Therefore, this level of fusion requires a large amount of data for the training and has high computational costs [90]. It is necessary to use a feature selection technique to find the features from all modalities that maximises the performance of the classifier.

Difficulties

The difficulties concerning early fusing are often so severe that scientists choose decision-level fusion. Problems are synchronisation of time and different entities, computational power and flexibility in modelling.

First, with synchronisation the empirical evidence reveals that multimodal signals often do not co-occur temporally at all during human-computer or natural human face-to-face communication. Therefore, computationalists should not count on conveniently overlapped signals in order to achieve successful processing in the multimodal architectures they build. Still, even if for instance the emotion is expressed at exactly the same time in every modality, how to choose a time frame? A video frame can be extracted at that moment, but audio features are calculated over time, not every moment. However, late integration allows asynchronous processing of the available modalities.

Besides the difficulties with synchronisation, when fusing the multimodal information at the feature-level, the feature set can get quite large. Because of this high dimensional data space, making a large multi-modal database necessary for robust statistical model training is necessary. Late integration provides greater flexibility in modelling. With late integration, it is possible to train different classifiers on different data sources and can be integrated without retraining. Off-the-shelf recognisers can be utilised for single modalities like e.g. speech. All of this is impossible using feature-level fusion.

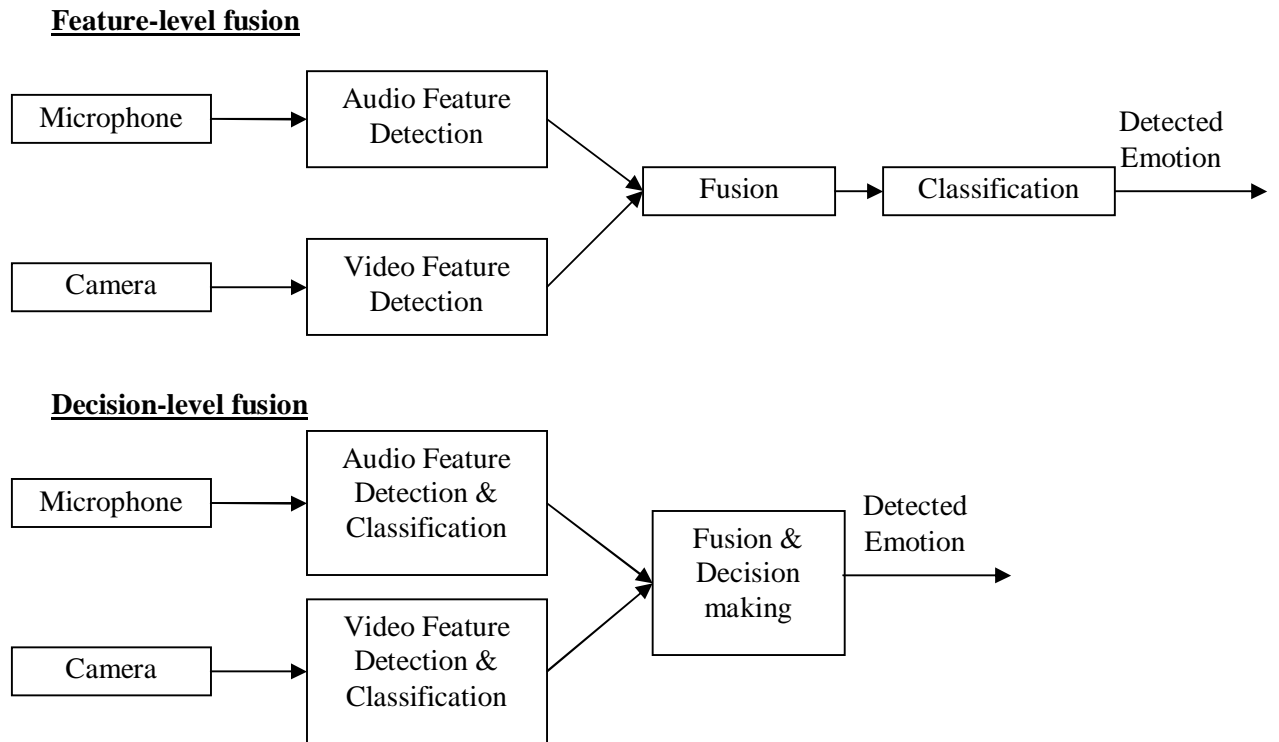


Diagram 3: The two aforementioned fusion methods, shown in a simple diagram.

CHAPTER 4

Construction of a Multimodal Data Corpus

Data corpora are an important building block of any scientific study. The data corpus should provide the means for understanding all the aspects of a given process, direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a good data corpus (i.e. well designed, capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results. Knowing this we decided to build such a data corpus. A good data corpus should have a good coverage of the process it's going to be investigated such that every aspect should get a fair slice.

In this chapter we present in detail the process of building an advanced multimodal emotion data corpus for the Dutch language. We strongly believe that sharing our experiences is the first step for understanding the issues around building a reliable data corpus. We envision a future standard for data corpora that combines the views of the entire scientific community.

In Paragraph 4.1 we explain which emotions are captured in the data corpus, where Paragraph 4.2 covers the settings in which the recordings will take place. The procedure followed to record the utterances is explained in Paragraph 4.3, this protocol is based on the approach described in [11]. Finally, Paragraph 4.4 gives an overview of the technical aspects that were used to gather data. Having high quality recordings makes training easier, but it also leaves the possibility open to degrade the recordings to a lower standard of quality for future research.

4.1 Emotional Content

The emotional content of the database should include at least the six archetypal emotions described by Ekman, some more blended emotions and a neutral recording of each subject must be present. A summation of the proposed emotions is listed below in Table 2. These emotions are retrieved from [15] and are believed to be product emotions. These emotions should represent the way people feel when judging products.

Table 2: List of used emotions.

Admiration	Desire	Fear	Satisfaction
Amusement	Disappointment	Furious	Surprise (Pleasant)
Anger	Disgust	Happiness	Surprise (Unpleasant)
Anger (Surprise)	Dislike	Indignation	
Boredom	Dissatisfaction	Interest	Neutral
Contempt	Fascination	Sadness	

4.1.1 Spontaneous or Posed Data

Another interesting topic is how the researchers managed to obtain data for observation. Some people used posers, including professional actors and non-actors. Others attempted to induce emotional reactions by some clever means. For example, Ekman showed stress-inducing film of nasal surgery in order to get the disgusted look on the viewers' faces. Other examples of rigorous elicitation of emotions are dumping water on the subjects or fired blank shots to induce surprise, while others used clumsy technicians who made rude remarks to arouse fear and anger. Obviously, some of these are not practical ways of acquiring data. After studying acted and natural expressions, Ekman concluded that expressions can be convincingly portrayed [93]. Besides the findings of Ekman, the researchers in [94] studied the spectrograms of real emotional speech and compared them with acted speech. They found similarities which suggest, again, that the use of acted data is allowed.

Therefore, following the above mentioned conclusions and keeping in mind that acquiring acted data is much easier and less time consuming, we decided to add recorded data portrayed by actors to our database.

4.2 Recording Protocol

In this paragraph we will briefly describe the protocol which we will follow in order to capture the expressions expressed by the participants. We describe the settings for the room, the environment and the manner in which the stories are told. Furthermore we will give an example of a story in which the participants have to express one particular emotion. The original stories are in Dutch; however they have a lot of similarities to the stories from the e'NTERFACE approach.

4.2.1 The Room Setting

In the room there is a chair on which the participants will take a seat. Right in front of this chair there is a camera, about two meters from the participant. We want to give the participant some degree of freedom in moving their head, but this freedom is limited to moving the head within the resolution of the camera. Figure 6 gives a good view of how the situation is [95]. We decided to use two synchronous cameras and replaced the mirror with a profile view camera, set at the same height as the frontal view camera and capturing images with the same speed. The second camera is placed on the left side of the participant, giving the camera a good silhouette view of the participants face. The chair is placed not too far from the wall, so the background can be covered by a large, dark coloured, piece of cloth. We choose to use two cameras because using a mirror gives slightly out of focus results due to the fact that the distance from the camera via the mirror to the participant is larger than the distance to the face for the frontal view. Besides this issue the freedom for the movement of the head was very limited, this is because the mirror is very narrow.

4.2.2 The Environment Conditions

The experiment is done in a closed room with good lighting conditions. A good lighting condition means that there is enough diffuse light to leave no shadows on the participants face. The camera is focused on the participant and the height of the camera is just right for lip reading. A consequence of this is that the height of the camera must change for every participant in order to get a perfect frontal view, but this is easily solved by placing the participant higher or lower in the chair. The background behind the participant is covered with a dark colour, preferably blue or green.



Figure 6: Overview of the room setting.

In the case of video data recording there are a larger number of important factors that control the success of the resulted data corpus. Hence, not only the environment, but also the equipment used for recording and other settings is actively influencing the final result. The environment where the recordings are made is very important since it can determine the illumination of the scene, and the background of the speakers. We use a mono-chrome background so that by using a “chroma keying” technique the speaker can be placed in different locations, inducing in this way some degree of visual noise.

4.3 The Procedure of the Experiment

The database should ideally contain only genuine expressions of emotions. However, as the database should also consist of high-quality video samples (with constant illumination, background, head pose, etc...) to be useful for practical applications, the choice that was made was to get as close as possible to spontaneous emotions, while keeping at the same time a fully controlled recording environment. To achieve this goal it was chosen to use pre-defined answers for each situation.

After a consent form was signed, the experiment could start. The average time the experiment lasted was 45 minutes. For each emotion the participant was asked to listen carefully to a short story and to ‘immerge’ themselves into the situation. Once ready, the participant may read, memorise and pronounce (one at the time) the five proposed utterances, which results in five different reactions to the given situation. The participants are asked to put in as much expressiveness as possible, producing a message that contains only the emotion to be elicited.

To obtain the participant’s facial expressions the experimenter should instruct the participant to perform the pure facial expressions. The procedure of the experiment is as follows:

- 1 The participant is told about the short stories to which he or she has to react to. The corresponding emotions are explained beforehand, so the participant knows which emotion to express.
- 2 The participant gets a list with responses corresponding to the expected expressed emotion.
- 3 The participant is asked to sit in front of the frontal camera, while the profile camera is positioned to the left of the participant. The distance between the cameras and the participant is about two meter.
- 4 The participant listens to or reads the short story and is asked to imagine being in this situation.

- 5 The experimenter gives the order to capture the emotional expressions of the participant.
- 6 The participant is then asked to react with each of the five pre-defined sentences.
- 7 If the participant’s performance is not ideal, repeat step 4 to 6. The repetition times are under control.
- 8 After the experiment is done, thank the participant for the cooperation.

Repeat step 4 to 7 for the next designated emotional expression, until the participant has displayed all of the emotional expressions.

With this procedure we let our participants take multiple sessions. This way we can fill the database quickly with many different recordings. The goal is to create a balanced database with respect to gender and age. A constant adding of recordings to the database will ensure the expansion of the data corpus and will add more diversity to the recordings. An example of a recording session is showed in Figure 7.

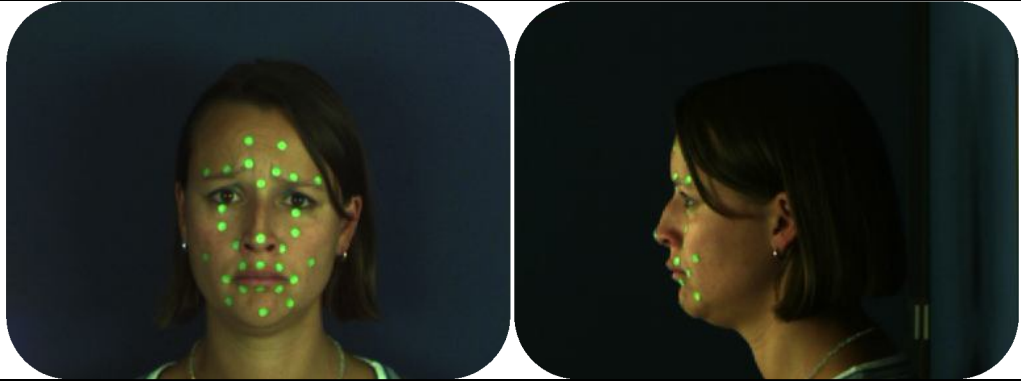
Teleurstelling (Disappointment)

<p>Scenario: “Je laatste eindexamen was zeker twee weken geleden en de cijfers moeten reeds bekend zijn. Vanmiddag zal je worden gebeld en iemand zal je vertellen wat de uitslag is en dan kan je gaan genieten van je zomervakantie. Je hebt zo hard je best gedaan, maar je weet niet zeker of je geslaagd bent. Als de telefoon over gaat ben je als eerste bij de telefoon om hem op te nemen. Je hoort een stem aan de andere kant zeggen dat je gezakt bent.”</p>
Reacties:
R1: Ik ben gezakt.
R2: Jammer, volgende keer beter.
R3: Dat had ik niet verwacht.
R4: En wat moet ik nu?
R5: Nee hè!

Figure 7: Situation and reactions to elicit “disappointment”.

4.4 Technical Aspects

The audio and video quality is an important issue to be covered. An open question is for instance, what is the optimum sampling rate in the visual domain? Current standard for video recording frame rate ranges from 24 up to 30 frames per second, but is that enough? A first problem with high sample rate, and the most intuitive, is the difficulty in handling the increased amount of data, since the bandwidth needed is many times larger. A second problem is a technical problem and is related to the techniques used for fusing the audio and video channels. Since it is common practice to sample the audio stream at a rate of over 100 feature vectors per second, in the case when the information is fused in an early stage, we encounter the need to use interpolation to match the two data sampling rates. A

third issue, that actually convinced us to use a high speed camera, is related to the coverage of the visemes during recording, namely the number of frames per visemes. A viseme is a facial image used to represent a phoneme or group of phonemes.

4.4.1 Audio and Video Devices

While talking with experts from the brain and speech domain we learned that recording at 125 Hz should cover almost every movement on a person's face. There are, however, movements like the lips vibration when the air is pushed with high speed through the loosely closed lips that require some 400 Hz for exact recording. Therefore we decided to use a high speed camera for video recordings.

When one goes outside the range of consumer devices, things become extremely more complicated and definitely more expensive. The quality of the sensors and the huge bandwidth necessary to stream high speed video to the PC makes high speed video recording very restrictive. We used for recording a Pike F032C camera built by AVT (Figure 8). The camera is capable of recording at 105 Hz when using the chroma subsampling ratio 4:2:2 while capturing at maximum resolution 640x480. We subsampled the two chroma components at half the sample rate. Now the horizontal chroma resolution is halved and this reduces the bandwidth of the video signal by one-third with little to no visual difference.



Figure 8: The Pike F032C camera built by AVT.

For recording the audio signal we used NT2A Studio Condensators (Figure 9). We recorded a stereo signal using a sample rate of 48 kHz and a sample size of 16 bits. The data was stored in PCM audio format. The recordings were conducted in a controlled laboratory environment. We considered that it is more advantageous to have very good quality recordings and degrade them in a post process as needed.



Figure 9: The NT2A Studio Condensators.

Unfortunately, we did not have an exact way to synchronise the audio and video stream. We had to rely on Windows that the starting delays will not be that far apart. The current synchronisation of the two streams is synchronised up two some milliseconds. Although this is for the human eye more than enough, for the mathematical models (HMMs) this might not be. So we should try to synchronise them manually.

4.4.2 Storage

Using a high speed camera increases the storage needs for the recordings. It is almost impossible to record everything and then during the annotation process, cut the clips to the required lengths. Therefore we let the participant control the beginning and end of the recording of a clip through the mouse buttons of a wireless mouse that was taped on the arm of the chair. The result was synchronised audio and video clips already cropped to the exact length of the utterance. After a series of trials we conclude that this level of control is sufficient and not very disruptive for the speaker. The size of a recording increases fast, the transfer rate of the data to the hard disk should therefore not be a limiting factor. Two high speed cameras connected to one computer gives a very large data stream, to solve the problem of writing this data to the hard disk in real-time we used two 250 GB SATAII hard disk in RAID 0.

RAID stands for “Redundant Array of Inexpensive Disks” (as named by the inventor) or occasionally known as “Redundant Array of Independent Disks” (a name which later developed within the computing industry). RAID is a technology that employs the simultaneous use of two or more hard disk drives to achieve greater levels of performance, reliability and/or larger data volume sizes [96]. RAID 0, or Striped set of hard disks without parity, provides improved performance and additional storage but no fault tolerance. Any disk failure destroys the array, which becomes more likely with more disks in the array. A single disk failure destroys the entire array because when data is written to a RAID 0 drive, the data is broken into fragments. The number of fragments is dictated by the number of disks in the drive. The fragments are written to their respective disks simultaneously on the same sector. This allows smaller sections of the entire chunk of data to be read off the drive in parallel, giving this type of arrangement huge bandwidth. RAID 0 does not implement error checking so any error is unrecoverable. More disks in the array means higher bandwidth, but greater risk of data loss.

4.4.3 Metadata

It is important to provide the data with the correct information in the form of metadata. This metadata is a description of the real data. The metadata should include the following fields with corresponding values:

Table 3: Metadata fields with corresponding values.

Field name	Fields value(s)	Occlusion	(Sun)Glasses/Beard/Hand/Hat/Object
User-nr	0001	Emotion	Admiration/Amusement/.../Surprise
Session-nr	0001	Spontaneous	YES/NO
Date	YYYY-MM-DD	AU	1/2/3.../28/43
Gender	M/F	View	Frontal/Silhouet
Age	XX	Text	The spoken text
Ethnic	Caucasian/Hispanic/...	Remarks	Some remarks
Type	Lip/Emotion/...		

4.5 Multimodal Database Availability

In order to make this database a benchmark for multimodal emotion recognition and lip reading research groups we need correct metadata associated with each database object. Note that it takes over an hour to manually label the 300 or 400 images and audio samples corresponding to a few seconds of high speed recordings. The ground truth of the metadata should always be guaranteed. Validation processes should be available in order to let users validate the data and give a value to the metadata of the corresponding database object. The methods explained in Chapter 5 are good ways to keep the metadata of these objects up to date.

Besides the crucial metadata a method to provide these database objects to the afore mentioned research groups is needed. The best solution would be a web-based direct-manipulation application. Possible research groups could subscribe for a login for this web site and would be granted access to the data. These users should then be able to download different database objects and they should also validate other database objects too. This ensures that the metadata is always renewed and that field experts validate the database objects.

4.5.1 Standards

The data that is to be added to the database should comply with a standard. This standard should be the minimal requirement for the data. Using a higher standard is allowed, data degeneration can always be performed in a later stage. For images, e.g., a different standard is required than for audio or video objects.

For images we should keep a standard of a true colour image (24 bit), a minimal size of half PAL (384 by 288) and a compression using the well-known jpeg compression technique. The audio samples should be sampled with a sampling rate of at least 44,100 Hz, 16 bit. The format of these audio samples should be in Pulse-Code Modulation (PCM) format. The minimal standards for video should be split in two. For high speed recordings, such as the ones we make, the minimal frame size should be half PAL (384 by 288). A frame rate of at least 100 frames per second for high speed recordings and for lower speed recordings the standard should be at least 25 frames per second. The compression should be lossless for both video recordings, preferably in the Audio Video Interleaved (AVI) format.

4.5.2 Security

Access to the multimodal database is restricted to authorised users only. Users can apply for a subscription to the database at any time. After agreeing with the end user license agreement (EULA), access is usually granted.

Different permission levels should be present. Users have user rights and group rights, the user should inherit all the rights from the group the user belongs to and group rights are stronger than user rights. Different groups with different permission should be created. Every new user is automatically placed in a 'new user' group. This group has the read only privilege but not the write privilege. An admin should be able to change the permissions of a user or a group and should be able to add or remove users to different groups. All registered users are able to download objects from the database. If a user has permission to write data to the database, this user can change the metadata belonging to a database object. A special group should be created for users who should be able to upload new database objects to the database. Adding data to the database should only be possible if metadata is present.

4.5.3 Search Queries

Most multimodal databases, if publicly available at all, offer only the possibility to obtain the entire database by, e.g. a download or by mailing CD's. This multimodal database should provide the users with a search environment to specifically search the entire database. There should be two search options, simple and advanced search. All search entries should be case-insensitive.

Simple search should provide the user with an input field in which the user can type search strings, e.g. "anger video caucasian". When using simple search each search query entry is linked to a metadata field and then the translation to a sql statement is done. This way it does not matter if a user types 'happy' or 'happiness'. The search query should then look in the database for objects which metadata fields hold the same values as the search query. All found objects should be returned in an orderly manner.

Advanced search gives the user the possibility to search the database for all the fields present in the metadata of database objects, thus providing the user with pre-defined search criteria. The user can then select different search variables from all available metadata fields present in the database. The user now cannot type 'happy' when searching for all objects with metadata field happiness, but can select the correct emotion from a selection of emotions present in the metadata emotion field.

CHAPTER 5

Visual Data Inspection

Data with incorrect labels attached to it are useless to work with. We cannot conclude anything from it. So in order for the data to be correct to work with, the data should be validated. This chapter explains the validation process applied to the recorded data of Chapter 4. If humans do not have consensus on what is expressed how can we expect computers to achieve at least the same result? To train a system on this data a lot of people should annotate the data with the same information. Data validation, as described in Paragraph 5.1, therefore, is an essential component in the process of setting up an (multimodal) emotional database. The methods used are described in Paragraph 5.2, these methods include validation by experts as well as validation done by naïve users. Validation results of the different types of validation are discussed in Paragraph 5.3.

5.1 Evaluation Space

Ekman claims that six emotions are universal, but most emotions are dependent on the context, words, character, etc. We hypothesised that *valence* and *arousal* are more or less universal. This will be verified. To label 21 emotions with 21 labels is postponed to the future.

The term *valence* usually refers to the positive and negative character of an emotion and/or of its aspects (e.g. happiness is high in valence). The English word *valence* was introduced in psychology in the 1930s [97], but not immediately within emotion theory. This use of *valence*, however, was not unambiguous and it contained the seeds of subsequent uses of the term, including its relation to emotion. At present, *valence* is often used as *affect valence* [98] [99] it refers to how good or bad an emotion experience, or affect, feels. Russel [39] has always positioned pleasantness-unpleasantness as a fundamental dimension of emotion experience. This dimension has gradually become ‘the valence dimension’. Another dimension of the emotional space is ‘the arousal dimension’. “Arousal” stands for the level of activation of the emotion, and it is characterised as a range of affective responses extending from “passive” to “active”. Meaning the subject is in a condition of sensory alertness and readiness to respond.

The valence-arousal model is used in most studies, because of these reasons:

- **Simplicity:** it is easy to express an emotion in terms of arousal and valence, whereas it is much more difficult to decompose an emotion into basic emotions.
- **Universality:** there is little controversy about the first two dimensions of the model. Valence and arousal are natural terms to use, when speaking about emotion, and they are understood between all cultures.

- Analysed data of emotional expressions with Principal Component Analysis (PCA) results in the first/second component being the valence/arousal axes.

An example of the valence-arousal space is given in Figure 10 below. Here some emotions are drawn in the space spanned by these dimensions.

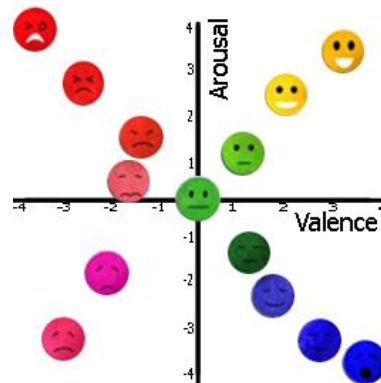


Figure 10: Different emotions on the valence and arousal scale.

There is a third dimension, but this dimension is not always used, and when it is used, it is not always the same. Sometimes dominance is used as the third dimension [100]. Other times it is motor activation (approach or avoid) that is on the third axis.

Both dimensions are represented by a 9-point scale. This 9-point scale was represented by a Self Assessment Mannekin (SAM) [101]. Users had to select the best corresponding state, of the 9-point scale, that described the expression best. An example of a 9-point SAM, for the valence scale, is given in Figure 11.

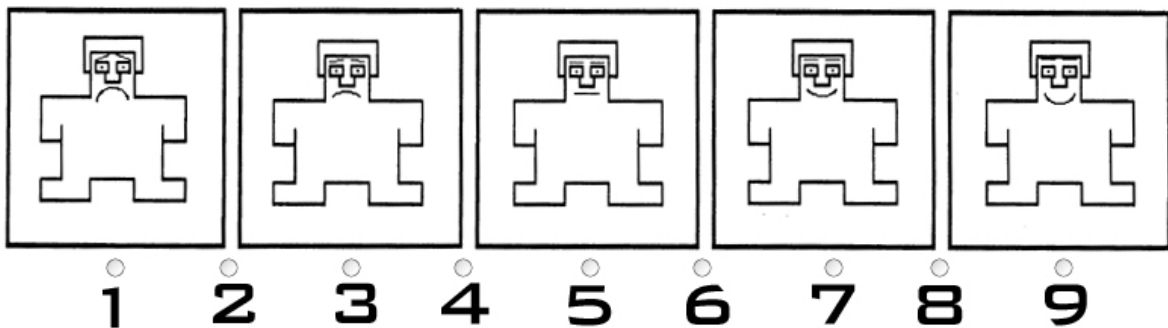


Figure 11: The valence 9-point scale Self Assessment Mannekin (SAM).

These 9-point scales ran from negative to neutral to positive, for the valence scale and from passive to neutral to active, for the arousal scale.

5.2 Data Validation

All recorded data has to be validated and there are several methods to do so. In the next paragraphs we will explain how we offered the data for validation and how we processed the results. We split up the validation into three different parts, this to see if there are differences in the validation of the various modalities. We offered from the recordings, the image showing the apex of the emotion, we offered only the audio clip from the recording and we offered the whole recording as a video clip.

We made a website where users could validate the data on two dimensions, valence and arousal. In more detail we requested the users to score a sample image, sound clip or video clip on these two scales. Note that also non-native Dutch speakers could, and did, validate the images. We therefore

made the language of this web site English. How many non-native Dutch users there were is not known. We ensured the users privacy and did not record any personal data, except their IP address. This in order to know how many distinct users participated in our validation process. Most users are naïve users and therefore they had the opportunity to look at an example validation before starting their own validation process. Examples of the validation processes of images, audio clips and video clips are visualised in Figure 14 below.

There are several methods to validate and annotate the recorded clips. Experts can annotate the clips to their best knowledge of the domain, but naïve users can give labels to the recordings too. As long as there are enough different people who annotated the clips, the ‘consensus’ of the broad public can be extracted.

5.2.1 Image Based Validation

In order to get a lot of ratings in a short time the validation of the images was done on-line. In order to quickly start validating the recorded clips, images showing the apex of the expressed emotion from the clips we extracted. These images were then displayed at a special developed website for validation. Users could click on ‘validate Images’ to start validating images right away. There was no limited amount of images shown and the displayed images were selected in a random order from all available images in the database. All 21 images that were validated are shown in Figure 15. These images are all extracted from the recordings of one person. In a matter of days more than 30 people had been rating images and the number of ratings exceeded 1200.

The blended emotion, like Anger-Surprise, was one for which we could not extract a single apex image from the clip. This is because the clip clearly showed two emotions being blended very fast after one another. The surprise expression of the emotion came after the anger expression. A list containing all of these emotions can be found in Paragraph 4.1, note that the image for Anger-Surprise was not validated and that there was an image present expressing neutral face.

5.2.2 Audio Based Validation

Validation of the various audio clips was done on the same website too. Here the users first listened to an audio clip taken from a recording and then they had to rate what they thought about the expressed emotion on the 9-point scales. We presented a number of 105 different audio clips, divided into 21 emotions and per emotion 5 different audio clips. These audio clips were all taken from one participant.

The screen presented to the user differs not much from the screen presented for the validation of images, this in order not to confuse the user. If a user can validate and rate an image, they know how to do this for an audio clip too. The user is in control of the playback of the audio clip. If necessary the user can play the audio clip again and again. When the page was loaded the audio clip automatically started to play.

5.2.3 Video Based Validation

For the video based validation the same principal was used as for the afore mentioned image and audio validation. Here the users could control the playback of the video. When the video was loaded the user could start playing it at any time.

The videos were constructed from the images recorded with the high speed camera to a format presentable for average online access speeds. We then converted the movie to a flash movie. These so called swf files are suitable for online presentation and commonly used in websites like youtube.com.

An overview of the showed video clip and its controls is shown in Figure 12 below.



Figure 12: Explanation of flash content with controls.
 1: The video; 2: Play button; 3: Timeline and buffer progress bar;
 4: Time indicator; 5: Volume slider; 6: Full screen button

5.2.4 Text Based Validation

For the text based validation we used the Whissell database [16] to look up the corresponding scores for the 21 different displayed emotions. We selected 21 emotions, with corresponding labels. We assume that the answers retrieved from the database show corresponding outcomes with the ones retrieved from the validation process. The whole database consists of more than 10.000 words and every word is scored by many respondents. We looked up the scores of our selected 21 emotional labels and where necessary substituted unknown words to the database with similar ones (e.g. ‘Amusement’ → ‘Amusing’ and ‘Fascination’ → ‘Fascinating’). The Whissell scores are based on a 3-point scale so these scores were converted to the 9-point scale used in the previous explained validation processes.

The text based validation was not done by the developed website. A second web interface was created to extract the values from the database. An example, here “Interest”, is shown in Figure 13.

Type your word or sentence here:

Look Up

You looked for

Interest

Found words:

interest 2.6667 1.25 1.6 5.52 5.49 56

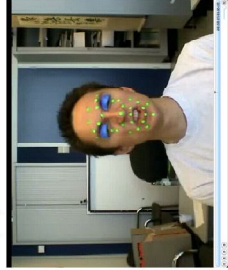
Interest has an evaluation of: 2.6667 and an activation of: 1.25.

On the 9-point scale *Interest* has an evaluation of: 7.6668 and an activation of: 2.

Figure 13: The developed DAL Web Interface.

Please score this video clip according to what you think the video clip is displaying.

You can stop validating any time you like, the past results are saved.



Please score this audio clip according to what you think the audio clip is displaying.

You can stop validating any time you like, the past results are saved.



Please rate both scales (valence and arousal)

Rate both scales (valence and arousal)

NEGATIVE NEUTRAL POSITIVE

PASSIVE ACTIVE

Please rate both scales (valence and arousal)

Rate both scales (valence and arousal)

NEGATIVE NEUTRAL POSITIVE

PASSIVE ACTIVE

Please score this image according to what you think the image is displaying.

You can stop validating any time you like, the past results are saved.



Please rate both scales (valence and arousal)

Rate both scales (valence and arousal)

NEGATIVE NEUTRAL POSITIVE

PASSIVE ACTIVE

Figure 14: Example validations of an image (left) an audio clip (middle) a video clip (right).

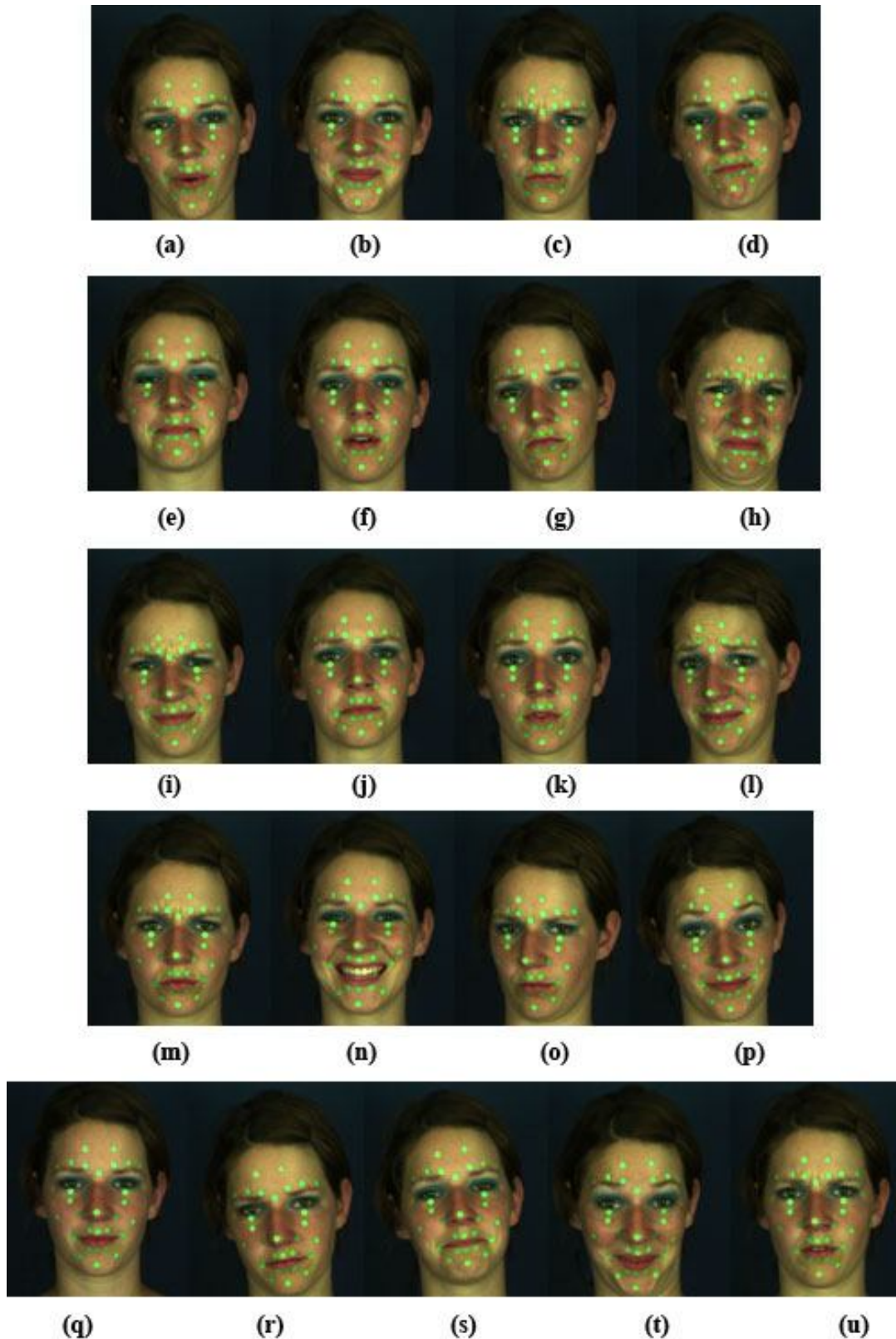


Figure 15: Example of displayed images for validation of participant P1.
 (a): Admiration, (b): Amusement, (c): Anger, (d): Boredom, (e): Contempt, (f): Desire, (g): Disappointment, (h): Disgust, (we): Dislike, (j): Dissatisfaction, (k): Fascination, (l): Fear, (m): Furious, (n): Happiness, (o): Indignation, (p): Interest, (q): Neutral, (r): Sadness, (s): Satisfaction, (t): Surprise (Pleasant), (u): Surprise (Unpleasant).

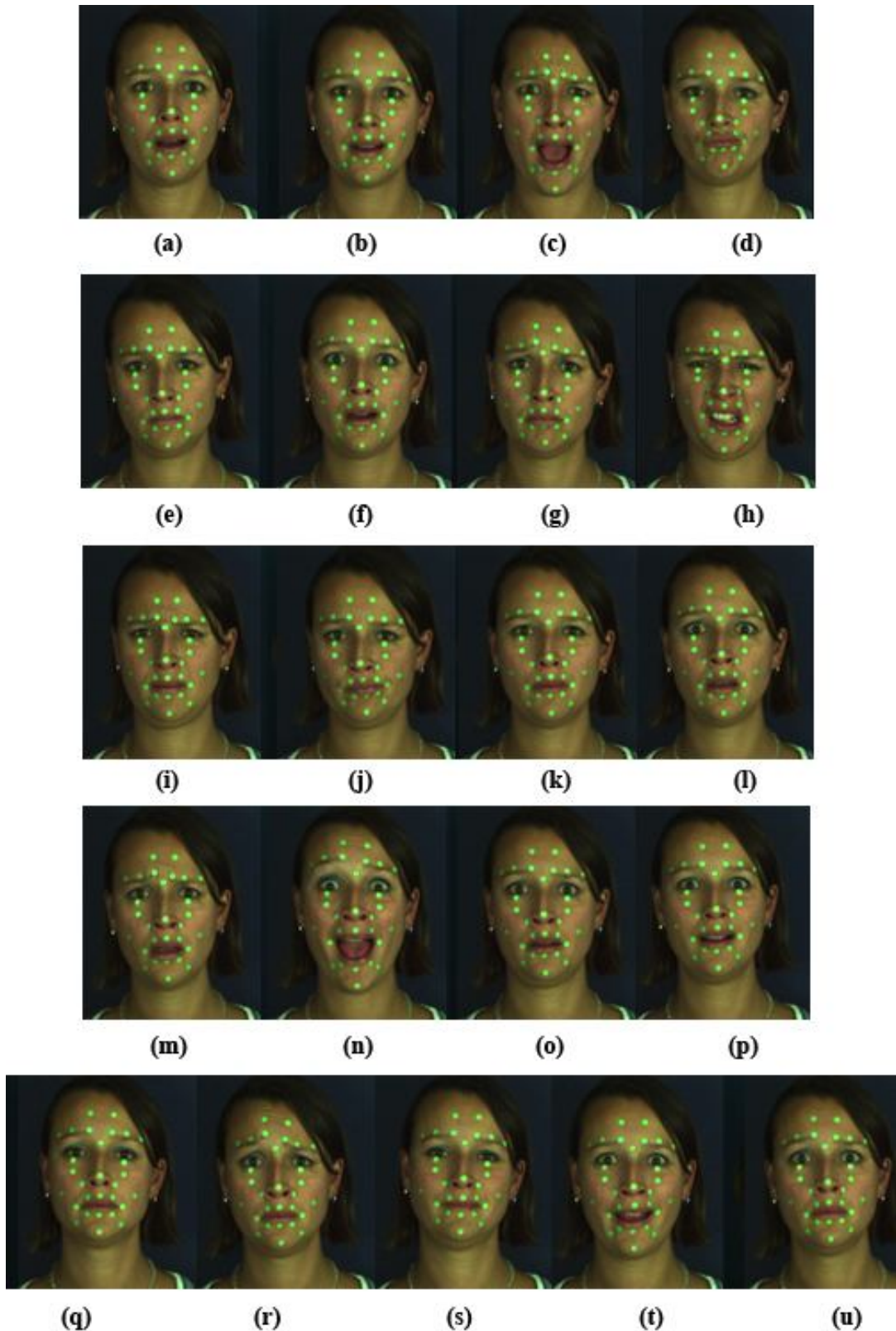


Figure 16: Example of displayed images for validation of participant P2.

(a): Admiration, (b): Amusement, (c): Anger, (d): Boredom, (e): Contempt, (f): Desire, (g): Disappointment, (h): Disgust, (i): Dislike, (j): Dissatisfaction, (k): Fascination, (l): Fear, (m): Furious, (n): Happiness, (o): Indignation, (p): Interest, (q): Neutral, (r): Sadness, (s): Satisfaction, (t): Surprise (Pleasant), (u): Surprise (Unpleasant).

5.3 Validation Results

The used validation approach was to annotate the images/clips with a value for valence (positive vs. negative) and a value for arousal (passive vs. active). These validation results could then be drawn on the two axes. We can look at the validation result in several ways.

1. We can look at the different modalities separately.
2. We can look at the average of all modalities combined in the video.

If we look at the modalities separately we expect users to score the displayed emotion incorrectly, due to the lack of extra information we usually get via other modalities. However, if we look at the validation of the videos, where all modalities are present, we expect the results to be similar. More information is present, such that users can give a more precise score for the emotion. A more detailed description of validation results per modality is given in the next four paragraphs. Validation tables are given for the validation scores per modality separately too. A total of 62 unique users participated in the validation of the different media displayed on the web page. As mentioned before, not every user that validated images, validated audio and video. This is because the language of the spoken text is Dutch and we also asked foreign users to validate the images.

5.3.1 Image Validation Results

A total of 49 unique users validated the images present on the website (see Figure 15). As can be seen from Figure 15 and Figure 16 we can clearly distinguish the 21 different facial expressions. If the users score these expressions widely spread across the valence and arousal axis the prototype should be able to correctly classify these 21 expressions. A distance matrix is given in Appendix J, here the distances between the facial feature vectors of both persons are compared. The shorter the distance, the more alike the feature vectors and the similar the facial expressions.

However, the results from the image validation were not what was expected. The expressed emotions were not over expressive, so users hardly rated the expression shown in the images as very aroused or activated. The expression of happiness has a very good (or what was expected) validation, however the validation of contempt is way off and is even validated as a positive expression. A comparison between the different validation methods is visualised in Figure 17. All the results of the image validation can be found in Appendix F, an overview is given in Table 4, below.

Table 4: Validation results for image validation.

		Valence	Arousal			Valence	Arousal
<i>1</i>	Admiration	0.81	0.19	<i>12</i>	Fascination	0.03	-0.74
<i>2</i>	Amusing	1.92	0.58	<i>13</i>	Fear	-1.17	0.23
<i>3</i>	Anger	-2.16	0.39	<i>14</i>	Furious	-2	-0.06
<i>4</i>	Anger (Surprise)	Not validated		<i>15</i>	Happiness	2.99	2.12
<i>5</i>	Boredom	-0.68	-1.02	<i>16</i>	Indignation	-1.47	-0.2
<i>6</i>	Contempt	0.7	-0.65	<i>17</i>	Interest	1.38	0.2
<i>7</i>	Desire	0	-0.69	<i>18</i>	Neutral	-0.12	1.4
<i>8</i>	Disappointment	-1.42	-1.1	<i>19</i>	Sadness	-0.69	-0.98
<i>9</i>	Disgust	-2.83	1.64	<i>20</i>	Satisfaction	0.79	-0.34
<i>10</i>	Dislike	-1.61	0.44	<i>21</i>	Surprise (Pleasant)	2.48	1.52
<i>11</i>	Dissatisfaction	-0.77	-1.25	<i>22</i>	Surprise (Unpleasant)	-1.54	0.09

5.3.2 Auditory Validation Results

A total of 9 unique users validated the audio clips present on the website, till now. The presented audio clips were the audio tracks from the video clips present at the website too. A comparison between the different validation methods is visualised in Figure 17. All the results of the auditory validation can be found in Appendix G, an overview is given in Table 5, below.

Table 5: Validation results for auditory validation.

		Valence	Arousal			Valence	Arousal
<i>1</i>	Admiration	2.5	1	<i>12</i>	Fascination	0.75	1.25
<i>2</i>	Amusing	1.43	0.57	<i>13</i>	Fear	-1.86	2.57
<i>3</i>	Anger	-2.91	2.82	<i>14</i>	Furious	-3.29	3.14
<i>4</i>	Anger (Surprise)	-1	2.33	<i>15</i>	Happiness	2.5	2.25
<i>5</i>	Boredom	-1	-1.67	<i>16</i>	Indignation	-0.33	-1.17
<i>6</i>	Contempt	-0.4	-0.6	<i>17</i>	Interest	2	1.75
<i>7</i>	Desire	2.25	2.5	<i>18</i>	Neutral	Not validated	
<i>8</i>	Disappointment	-2	-0.83	<i>19</i>	Sadness	-2.71	-0.43
<i>9</i>	Disgust	-1.75	2	<i>20</i>	Satisfaction	3	-3
<i>10</i>	Dislike	-2.17	0.17	<i>21</i>	Surprise (Pleasant)	3	3
<i>11</i>	Dissatisfaction	-1	0	<i>22</i>	Surprise (Unpleasant)	0.67	1.33

5.3.3 Multimodal Validation Results

A total of 14 unique users validated the video clips present on the website, till now. Remarkable is the fact that these users gave the emotion ‘fear’ a negative arousal and all the other validation method scored a positive arousal for fear. A comparison between the different validation methods is visualised in Figure 17. All the results of the multimodal validation can be found in Appendix H, an overview is given in Table 6, below.

Table 6: Validation results for multimodal validation.

		Valence	Arousal			Valence	Arousal
<i>1</i>	Admiration	1.5	0.43	<i>12</i>	Fascination	1.24	0.33
<i>2</i>	Amusing	3.65	3.12	<i>13</i>	Fear	-3.09	-1.64
<i>3</i>	Anger	-3.41	2.53	<i>14</i>	Furious	-2.77	2.46
<i>4</i>	Anger (Surprise)	-2.86	1.07	<i>15</i>	Happiness	3.83	3.33
<i>5</i>	Boredom	-1.31	-2.13	<i>16</i>	Indignation	-0.45	-0.27
<i>6</i>	Contempt	0.6	1.4	<i>17</i>	Interest	2.47	1.42
<i>7</i>	Desire	2.61	1.56	<i>18</i>	Neutral	Not validated	
<i>8</i>	Disappointment	-1.93	-1.67	<i>19</i>	Sadness	-2.43	-1.79
<i>9</i>	Disgust	-3.55	0.09	<i>20</i>	Satisfaction	2.57	1.93
<i>10</i>	Dislike	-2.44	0.38	<i>21</i>	Surprise (Pleasant)	2.92	2.25
<i>11</i>	Dissatisfaction	-1.35	-0.41	<i>22</i>	Surprise (Unpleasant)	-3.29	1.07

5.3.4 Textual Validation Results

A total of over 200 unique users validated the words in the Whissell database. The Whissell database consists of over 10.000 words for which these users gave a score for the valence, arousal and imaginary scale. In our research we do not use this imaginary scale.

The Whissell database gave the following coordinates for the evaluation and arousal space, showed in Table 7 below. Note that the Whissell scale is a 3-point scale, rating from 1 to 3, where 1 is the lowest value and 3 the highest. To compare these Whissell values we have transformed these outcomes to the 9-point scale we used. The numbers in front of the emotions in the table correspond to the red number shown in Figure 17.

**Table 7: Whissell scores^[1] for the used emotions (9-point scale).
If a certain emotion is not present in the Whissell database, the words between brackets is looked up.**

		Valence	Arousal			Valence	Arousal
1	Admiration	2.00	-0.89	12	Fascination(Fascinating)	3.20	2.00
2	Amusement(Amusing)	4.00	2.50	13	Fear	-4.00	1.45
3	Anger	-4.00	3.56	14	Furious(Fury)	-2.40	2.29
4	Anger (Surprise)	Not validated		15	Happiness	4.00	3.20
5	Boredom	-4.00	-2.55	16	Indignation(Resentment)	-2.67	-0.44
6	Contempt	-1.71	-1.00	17	Interest	2.67	-3.00
7	Desire	0.57	1.50	18	Neutral	Not validated	
8	Disappointment	-1.33	-0.57	19	Sadness(Sad)	-2.50	-2.29
9	Disgust	-2.5	-0.50	20	Satisfaction	2.22	-3.00
10	Dislike	-3.43	-1.71	21	Pleasant Surprise	2.91	1.00
11	Dissatisfaction(Disapproval)	-2.55	-0.67	22	Unpleasant Surprise ^[2]	-2.91	1.00

[1]: Note that Whissell uses ‘evaluation’ and ‘activation’ as terms for the two dimensions.

[2]: Here the value for valence is the opposite to the value for Pleasant Surprise.

5.4 Comparison of Validation Results

If we compare the results acquired with the different validation methods used and described in the previous paragraphs we can draw intermediate conclusions. A simple overview can be generated if we combine all four previous tables into one figure. Each validation method is represented with its own colour; this is the same colour as the numbering in the previous four tables. The resulted figure, Figure 17, is shown below.

In the ideal case we should see clusters of the same numbers close to each other. But this is not the case for many of the 21 emotions. For example, the number 6, representing the emotion ‘contempt’ is scattered over three quadrants of the axes. The image validation gives this emotion a positive valence and a negative arousal, the auditory validation a negative valence and a negative arousal, the video validation a positive valence and a positive arousal, and the textual validation gives contempt a negative valence and a negative arousal. This example shows that a lack of information, due to the separation of the unimodal channels can cause the user to misplace the intended emotion.

A positive example is, for example, the number 8. This number represents the emotion ‘disappointment’. It can be clearly seen that all four number eights lay close together. This means that the validation of this emotion does not depend much on multimodal information. Other clear clusters are the numbers 15, 21, 3 and 14, representing happiness, pleasant surprise, anger and furious accordingly. These very active emotions were validated very closely by all validation methods, showing that the distinction between active positive and negative emotions is not as subtle as with other emotions.

Table 8: Overview of position of emotions per quadrant.
RED: Image validation, GREEN: Audio validation, BLUE: Video validation and PURPLE: Text validation

		1 st Quadrant	2 nd Quadrant	3 rd Quadrant	4 th Quadrant
1	Admiration	IAV			T
2	Amusement	IAVT			
3	Anger		IAV		T
4	Anger (Surprise)		AV		
5	Boredom			IAVT	
6	Contempt	V		AT	I
7	Desire	AVT			I
8	Disappointment			IAVT	
9	Disgust		IAV	T	
10	Dislike		IAV	T	
11	Dissatisfaction			IAVT	
12	Fascination	IAVT			
13	Fear		IA	VT	
14	Furious		IAVT		
15	Happiness	IAVT			
16	Indignation			IAVT	
17	Interest	IAV			T
18	Neutral		I		
19	Sadness			IAVT	
20	Satisfaction	V			IAT
21	Pleasant Surprise	IAVT			
22	Unpleasant Surprise	A	IVT		

In Table 8 the red **I**, the green **A**, the blue **V** and the purple **T** correspond with the image validation, the audio validation, the video validation and the textual validation respectively.

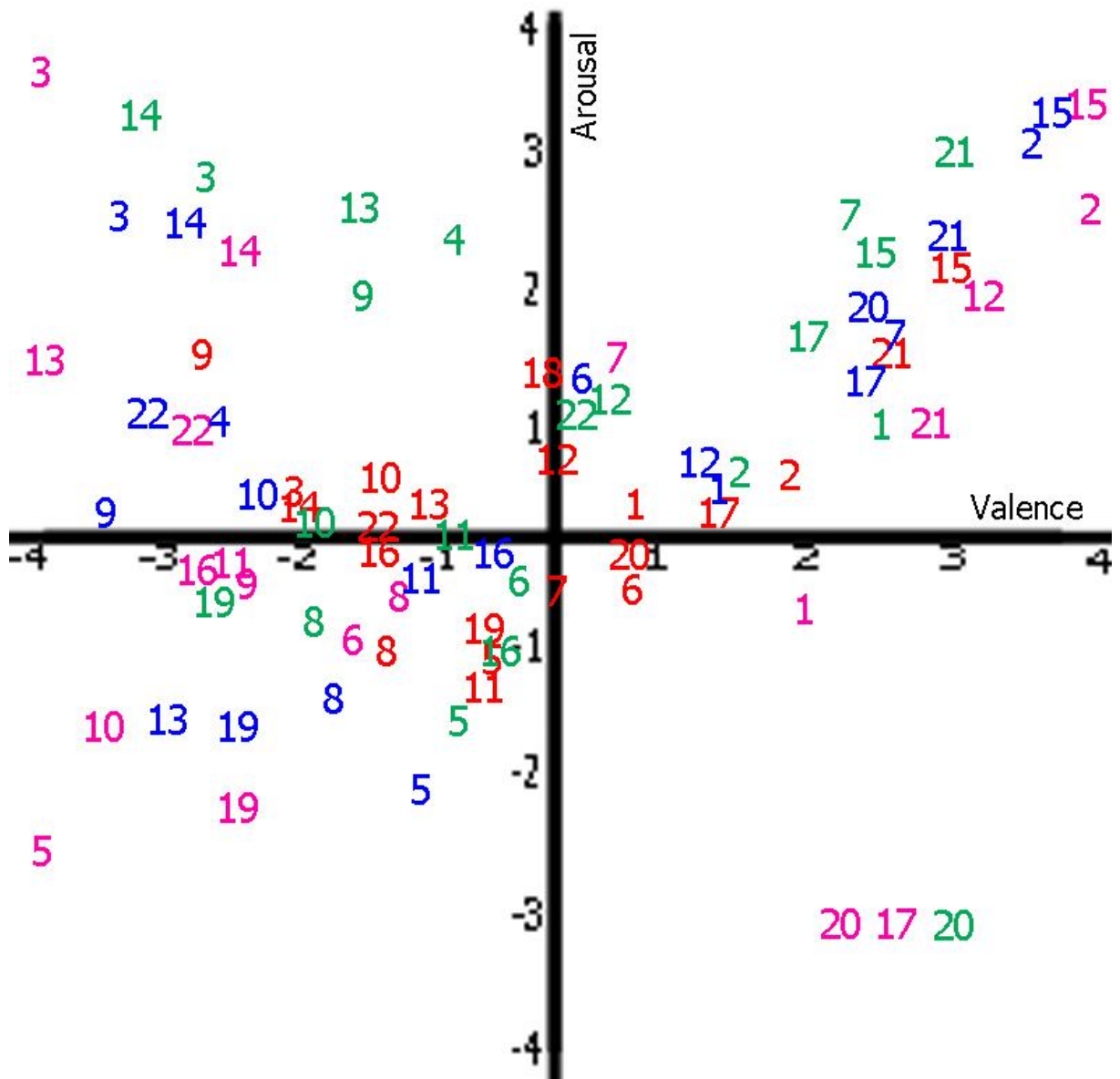


Figure 17: Comparison of different validation methods.
RED: Image validation, GREEN: Audio validation
BLUE: Video validation and PURPLE: Text validation

CHAPTER 6

Model, Prototype and Data Pre-processing

This chapter describes the model and the prototype we have created to track facial feature points. This prototype should be able to identify and track these pre-defined points. Knowing the locations of these points the prototype should be able to translate these locations to activation of different AUs (see Paragraph 7.2) from which the prototype should then be able to recognise emotions.

Besides the prototype description this chapter describes in detail the extraction process of the facial feature points for each frame and the auditory features for the audio clip. Furthermore, we will discuss these pre-processing steps taken in order to further process the data. For a correct extraction of features this is a necessary step. Data pre-processing steps for the auditory channel are given in Paragraph 6.3. Pre-processing steps for the extraction of the facial feature points are described in Paragraph 6.4.

6.1 The Model

We designed our prototype according to the model as described in Figure 18. It is clear that this model consists of two different modules, an audio part (green) and a video part (red). These are the basic parts of the multimodal emotion recogniser.

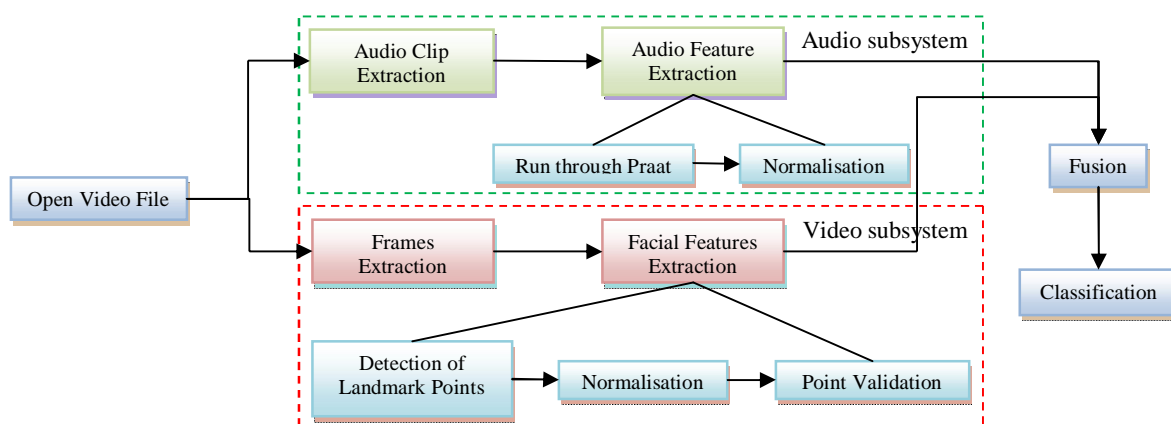


Figure 18: System overview.

The video part of the prototype extracts the frames from the video clip. It then extracts the feature vector from the frame to work with. For the audio part we process the whole audio clip at once. The designed prototype tries to recognise an emotion. Taking into account that some emotions are visually dominant and some are auditory dominant. Emotion recognition can be done per frame, giving the user some idea freedom.

6.2 Prototype Design

The prototype is designed as a simple tool for recognising emotions from audio and facial expressions. Our prototype is composed of several modules (see Figure 25). We discuss the modules in more details in Paragraph 6.3 and 6.4. A screenshot of the prototype can be seen in Figure 2.

The prototype depends on the extraction of 31 facial feature points, marked on the face for easy segmentation, and 91 audio features. The schematic design of the prototype is presented in Figure 18. This prototype can handle both avi and mpeg files. After the program has started the user can import a video file. Once selected, the prototype reads in the whole audio sample and the first frame. The import of a whole movie can cause Matlab to stop responding, therefore we only import the first frame. From the first frame we extract the position of the points and we assume that the first frame is always showing a neutral face. The extraction of the audio features is done immediately; the facial features are computed per frame.

6.2.1 Implementation Details

Matlab is used as the programming environment, because of the fast handling of matrices and the use of many toolboxes available. The version used was MATLAB Version 7.3.0.267 (R2006b) on a Microsoft Windows machine running: Microsoft Windows XP Version 5.1 (Build 2600: Service Pack 3). The used toolboxes are:

- Pattern Recognition Tools Version 4.0.23 [102].
- Image Processing Toolbox Version 5.3 (R2006b).
- Mmread Version 2008-04-28 [103].

Especially for the graphical processing, Matlab was a convenient environment to work with. The easy handling of matrices and images made the programming to process these much easier. For the extraction of the audio features Matlab and “*Praat*” are used. Praat is a software program for speech analysis [104]. This choice was necessary in order to correctly calculate the list of auditory features as described in Paragraph 6.3. The interaction between Matlab and Praat did not go directly, but via written text files which each program had to read or write.

6.3 Audio Feature Extraction

We extract several values of pitch, intensity and spectral features with a 10 ms time window, see Figure 19. Other values that are extracted are: jitter, intensity range, intensity variability and high frequency energy, making the list with 91 features complete. For the implementation of features that could not be calculated by Praat, MATLAB was used to calculate the remaining features and add these to the feature matrix.

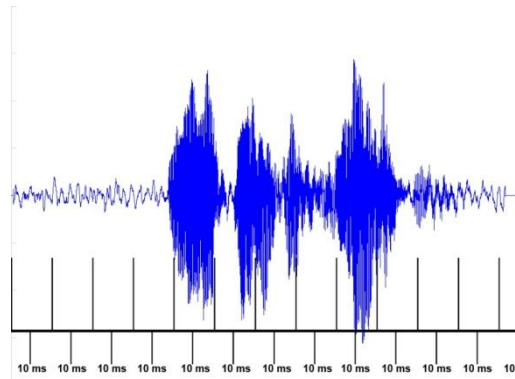


Figure 19: Waveform with 10ms time window.

The set of spectral features is comprised by statistical properties of the first 4 formants and the energy below 250 Hz. A formant is a concentration of acoustic energy around a particular frequency in the speech wave which results from the resonant frequencies of any acoustical system. It is most commonly invoked in phonetics or acoustics involving the resonant frequencies of vocal tracts. There are several formants, each at a different frequency. Formants occur at roughly 1000 Hz intervals. Each formant corresponds to a resonance in the vocal tract. By definition, the information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. Formants are the characteristic partials that identify vowels to the listener. The main energy source in speech is vibration of the vocal cords. The vocal cord can vibrate at any given time with any given rate. The rate at which vocal cords vibrate determines the fundamental frequency of the acoustic signal F0. Variations in voice pitch are considered to have a linguistic function. Pitch features are statistical properties of the pitch contour. The speech intensity depends primarily on the amplitude of vocal cord vibrations which is related to the pressure of the air stream. The greater the expiratory effort, the greater the intensity. As a small addition some higher order features are added, making a total of 91 features.

Spectral related features:

1. Energy below 250 Hz
2. - 5. Mean value of the first, second, third, and fourth formant
6. - 9. Maximum value of the first, second, third, and fourth formant
10. - 13. Minimum value of the first, second, third, and fourth formant
14. - 17. Variance of the first, second, third, and fourth formant

Pitch related features:

18. - 22. Maximum, minimum, mean, median, interquartile range
23. Pitch existence in the utterance expressed in percentage (0-100%)
24. Maximum duration of plateaux at minima
25. Mean duration of plateaux at minima
26. Mean value of plateaux at minima
27. Median duration of plateaux at minima
28. Median value of plateaux at minima
29. Interquartile range of plateaux at minima
30. Interquartile duration of plateaux at minima
31. Maximum duration of plateaux at maxima
32. Mean duration of plateaux at maxima
33. Mean value of plateaux at maxima
34. Median duration of plateaux at maxima
35. Median value of plateaux at maxima
36. Interquartile range of plateaux at maxima
37. Interquartile duration of plateaux at maxima
38. Upper limit (90%) of duration of plateaux at maxima

39. Maximum duration of rising slopes
40. Mean duration of rising slopes
41. Mean value of rising slopes
42. Median duration of rising slopes
43. Median value of rising slopes
44. Interquartile range of rising slopes
45. Interquartile duration range of rising slopes
46. Maximum duration of falling slopes
47. Mean duration of falling slopes
48. Mean value of falling slopes
49. Median duration of falling slopes
50. Median value of falling slopes
51. Interquartile range of falling slopes
52. Interquartile duration range of falling slopes
53. Number of inflections in F0 contour

Intensity (Energy) related features:

54. - 58. Maximum, minimum, mean, median, interquartile range
59. Maximum duration of plateaux at minima
60. Mean duration of plateaux at minima
61. Mean value of plateaux at minima
62. Median duration of plateaux at minima
63. Median value of plateaux at minima
64. Interquartile range of plateaux at minima
65. Interquartile duration range of plateaux at minima
66. Maximum duration of plateaux at maxima
67. Mean duration of plateaux at maxima
68. Mean value of plateaux at maxima
69. Median duration of plateaux at maxima
70. Median value of plateaux at maxima
71. Interquartile range of plateaux at maxima
72. Interquartile duration range of plateaux at maxima
73. Upper limit (90%) of duration of plateaux at maxima
74. Maximum duration of rising slopes
75. Mean duration of rising slopes
76. Mean value of rising slopes
77. Median duration of rising slopes
78. Median value of rising slopes
79. Interquartile range of rising slopes
80. Interquartile duration range of rising slopes
81. Maximum duration of falling slopes
82. Mean duration of falling slopes
83. Mean value of falling slopes
84. Median duration of falling slopes
85. Median value of falling slopes
86. Interquartile range of falling slopes
87. Interquartile duration range of falling slopes

Higher order features:

88. Intensity range
89. Intensity variability
90. High frequency energy
91. F0 perturbation (jitter)

This list is retrieved from [82], and contains this many features because a boosting program like GentleBoost can later on always choose a feature set with the largest discriminating capabilities.

6.3.1 Praat

Praat is a standalone product, which has the advantage that it can execute Praat-scripts. This allows fully automated analyses of the audio signals. One other reason we choose Praat above other programs is that it has a very large user base, and an active forum. A lot of audio research is done by researchers with Praat. Praat is mentioned in papers explaining how Praat can easily compute the necessary features from audio signals and provide the user with a highly detailed plot of these features.

It was quite hard to combine Praat with a MATLAB program. The interaction between Praat and MATLAB was difficult to achieve, but an easy solution was found. The program comes with a sub program, which can interact with Praat called 'sendPraat.exe'. This tool is the communication between other programs and Praat, and is executed from the DOS prompt. We could just attach the name of the script to be executed as a parameter to sendPraat.

Executing sendPraat with a script is the way to let Praat manipulate the audio files. These audio files have first been written from MATLAB to a known directory and with a standard file name, so Praat can cope with this. After reading in the movie, we extract the audio file from the video file and save it to the directory where Praat is running as 'audio.wav'. We then execute sendPraat via the DOS command from within Matlab to let Praat analyse the audio file with our script the results are saved on the disk from which Matlab then reads the results back in. The remaining features are calculated with Matlab.

6.4 Facial Feature Points Localisation

This paragraph describes in detail the extraction process of the facial features for each frame. There have been numerous attempts to track faces and their facial features but not all perform very well under different circumstances or on different data [105] [106]. Therefore, in order to easily obtain a feature vector for each frame of the recording, we used well-known tracking techniques and did take recordings in a very controlled environment.

6.4.1 Image Segmentation

In order to correctly track the feature points we first need to detect these points in each frame. We start by extracting the face from the extracted frame. For the localisation of the face in the frame we used the Machine Perception Toolbox (MPT) [107]. This piece of software finds the face in the frame and returns the cut out face to Matlab. This face-only image is then pre-processed in order to find the green stickers and the blue eyes. As mentioned before we placed green stickers on the participant's face to make the detection of these points easier.

For each frame we start with a fixed colour based filter applied to the pseudo hue space of the image. Pseudo hue colour segmentation [108] gave better results than HSV colour segmentation. The pseudo hue component can be calculated as follows:

$$P_h = \frac{R}{R + G} \quad (2)$$

where P_h is the *pseudo hue* component, R the *red* component and G the *green* component of the image. We then extract the part of the image that corresponds to the green stickers as follows:

$$F_{(x,y)} = \begin{cases} 1, & P_h < 0.1 \\ 1, & v > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $F_{(x,y)}$ is the value of the binary image on location (x,y) , v the *value* (HSV) component of the image. An example of extracted green stickers is shown in Figure 20b below. Finally we extract the blue part of the image that corresponds to the blue make-up as follows:

$$F_{(x,y)} = \begin{cases} 1, & h > 0.55 \\ 1, & v > 0.6 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $F_{(x,y)}$ is the value of the binary image on location (x,y) , h the *hue* (HSV) component of the image. An example of extracted blue eye make-up is shown in Figure 20c below.

Feature point 31 (chin point) is always visible and labelled point 1 by the algorithm, which works from left to right and labels the blobs accordingly. We rotate the image 90 degrees clockwise and perform the labelling algorithm, labelling the chin point with label 1. We then rotate the image back to the original orientation. Other points may not always get the same label this can be the case when some points are missing, the rotation of the head causes points not to align, or there is noise in the image and we find false positive points.

The feature points 10 and 11, both placed on the lower part of the upper eye lid, were detected by making use of the blue make-up. These points are supposed to be at the lowest 'on' pixel below the centre of mass of the detected blue eye blob (4). An example of the detection of the points 10 and 11 can be seen in Figure 21 below. The red points indicate the centre of mass while the blue points indicate the lowest 'active' or 'on' pixel below the centre of mass.

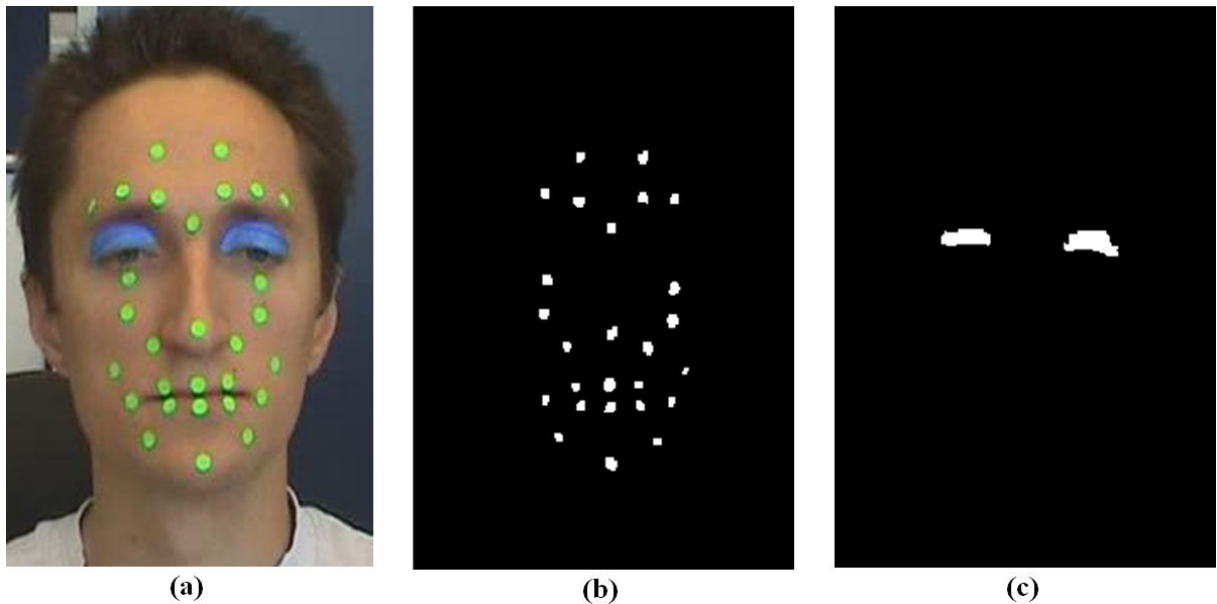


Figure 20: The original image (a), the detection of only 26 of the 29 green stickers (b) and the detection of the blue eye make-up (c).

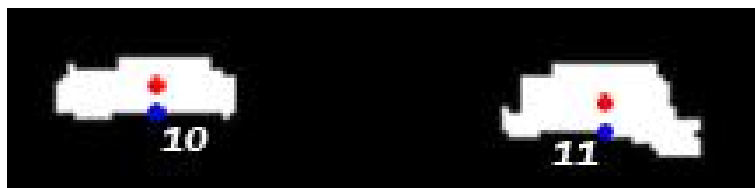


Figure 21: The detection of feature points 10 and 11.

We use the blue eye make-up in order to avoid sticking a green sticker on the upper eye lid. This can be uncomfortable for the participant and when blinking the sticker can fall off. Therefore this less rigid method is used:

$$P_{eye} = \left(\frac{1}{n} \sum_{i=1}^n x_{i, \max(y_i)} \right) \quad (5)$$

where P_{eye} is the lowest ‘on’ pixel of the eye blob, corresponding to the blue point on the eye lid. When we have detected the blue eye lids we can also find feature point 1 (eye point), which always lies in between both eyes. The point closest to the centre of mass of both eyes must be feature point 1. All other points can be derived from the location of these two points, and as more and more points are known, the labelling of unknown points becomes much easier.

6.4.2 Facial Measurement Model

We worked with a point-based facial model. During the recordings all feature points from the model (or areas used for automatic determination of needed points) were marked with easy distinguishable colours. These colours differ from the natural colour of the skin, hair and background. Besides that, we also took care about reasonably good illumination conditions (see Paragraph 4.2.2), and the recorded persons were asked not to wear green nor blue. These conditions ensured that an easy colour based segmentation of the facial points in each frame was successful.

Notice that we used stickers on the face to identify the facial points. Earlier recordings have been made in this way and were readily available so we continued this way. This means that the implementation is limited to the segmentation of green stickers and blue eye shadow on the face. In the future it is possible to adjust the system to accept facial points found by other algorithms than the colour segmentation we used.

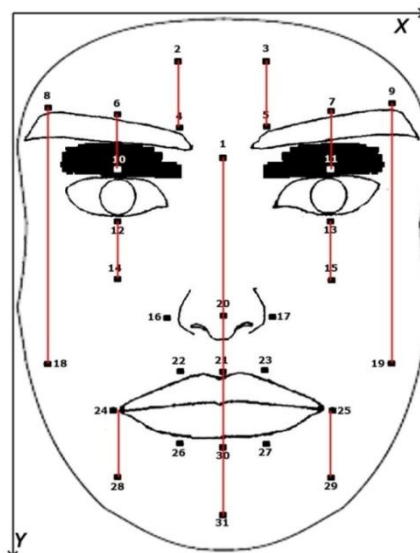


Figure 22: Placement of the 31 points we tracked.

The facial model we used to collect information about facial features is a 2D point-based model. Defined in this model are 31 feature points on the face (see Figure 22). The model is based on the work of Anna Wojdel [14]. This model was designed with the intentions that:

- Feature points can be easily marked (to simplify the process of tracking them).

- Facial movements related to facial expressions should effect in the displacement of feature points.

The most feature displacements take place around the mouth and the eyes. Because we recorded participants expressing emotions not only visually, but also auditory, the movement of the mouth makes the points in that region move more frequently than without speech considered. Frequent movement of these points is thus not limited to emotions.

6.4.3 Detection of Facial Feature Points

We detect the 31 predefined points on the face according to the labelling, as mentioned before. All found points, it can happen we do not find all 31 correctly, are stored in a vector. Luckily it is always the case that feature point 31, the chin point, is labelled '1'. This ensures that the first point in the point vector '*points*' corresponds to the chin point.

The following figure shows the processing of the different regions of the face were the points should be located. We start with feature point 1, the eye point. This point should be closest to the centre of mass of the labelled eyes and can be easily obtained. The next point is feature point 20, the nose point. In the same way we locate the nose point as the point closest to the middle of the image. The nose is always located at the centre of the image. And we know that feature point 31, the chin point, is always the first point in the points vector. Now that we have located these 'key' points from the points vector we can see if the head was rolled. For an explanation of what movements the head can make, look at Figure 23 below.

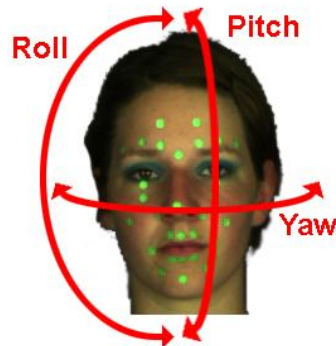


Figure 23: Yaw, Pitch and Roll visualised for the face.

We can determine the roll of the face by calculating the angle between the eye point and the chin point, as these two points lay in the same plane and should always be perpendicular to the horizon for a straight face. If the absolute roll or rotation of the head is greater than 1 degree we rotate the image back to a straight position. We then label the green stickers again, because it helps to locate future points better and find the key points again.

We then normalise the points such that the distance between the eye point and the nose point is 50 pixels. In this case it does not matter how large the image is, images with a smaller height than 200 pixels cannot be pre-processed correctly. All pre-process steps are visualised in Figure 24 below.

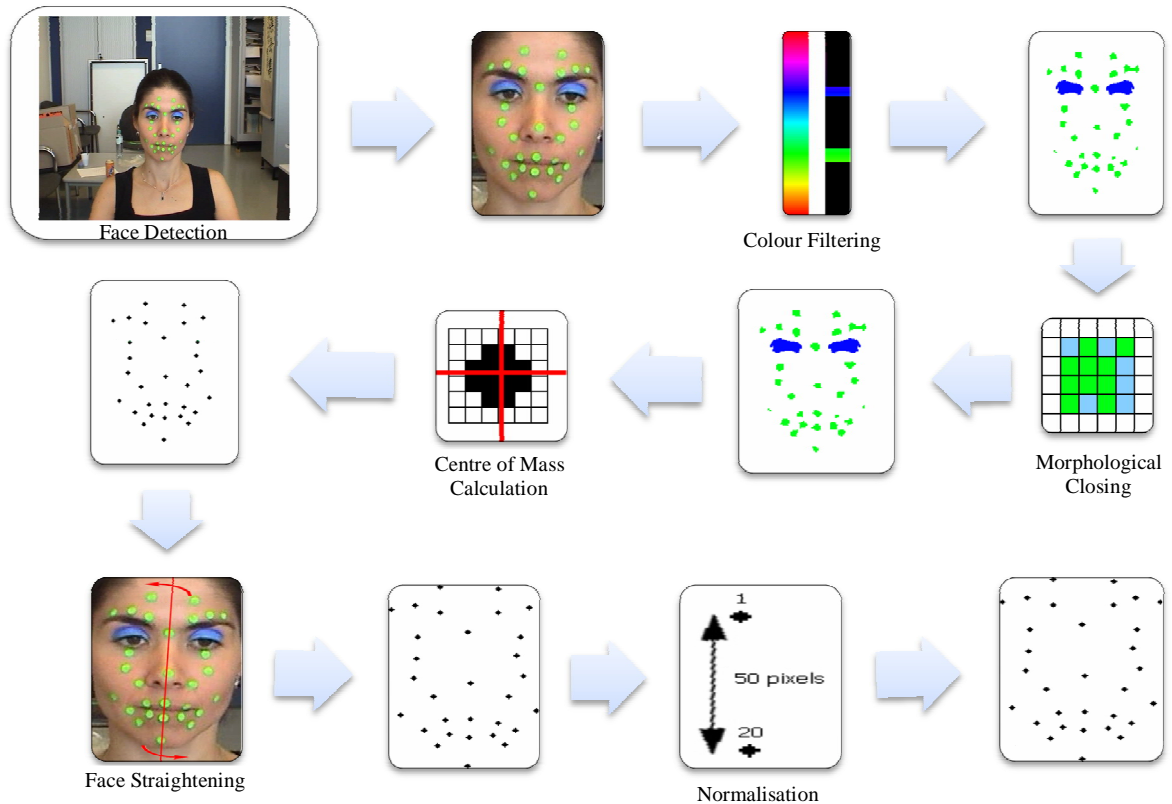


Figure 24: Processing flow of facial features tracking.

Every point of our model should be occupied with an (x,y) coordinate now. The model can now be checked by the validation rules to make sure that all the coordinates are linked to the correct position in our model. These rules are straight forward and can be linked to the red lines in Figure 22 and the symmetry of the face.

6.4.4 Point Validation

Unfortunately, while doing the experiments we discovered that, in some cases, the segmentation method was not sufficient enough for proper facial feature point localisation. This can have two separate causes. First, in some images some markers were not found at all due to not perfect illumination of the face, or when the participant presses their lips to each other, and the stickers placed around the mouth join together, two separate stickers are covered by one big blob. In order to improve the accuracy of the found facial feature points, we used basic knowledge about the coordinates of these feature points in the facial model. After the positions of blobs are calculated, we perform a validation of the found points.

First, we solve the problem of missing feature points by using the placement of neighbouring feature points. Most points have a relation with other points, see Figure 22, where already some red lines indicate dependencies. Besides vertical dependencies, horizontal dependencies can be established too and some points are not supposed to be located above, or below other points. We always start with the assumption that the key points are found correctly. These points are the base of the model and easy to locate. This way, building dependencies onto these points are assumed to be correct.

Second, if any of the locations in our model cannot be linked to a detected point we use the symmetry of the face to copy the symmetrical feature point, opposite of the line through feature point 1, feature point 20, feature point 21, feature point 30 and feature point 31, to the empty location. We start with the assumption that at least one of the mirroring points is found. Although, it can occur that

both mirroring points are missing, e.g. feature points 2 and 3. Might this be the case than we try to copy the location of these points from the previous found location and as a last resort we try to reconstruct these points based on the knowledge of surrounding points and their mutual distances.

Finally, there exists the problem of detecting two blobs corresponding to one feature point, we checked whether found blobs are distant enough to treat them as separate blobs, or they represent the same marker. If the distance between two blobs is shorter than a predefined limit (here 10 pixels), we merge them and calculate a position of a new blob as a middle point of unified blobs. This condition may seem at the first sight to amplify the problem of merging markers around the lips, but it does not. In our recordings, the size of the typical blob is usually between 5 and 10 pixels in each direction. Even after normalisation of the points based on the found key points, the distance between typical blobs, even when they are separated only by one pixel, is longer than the predefined limit (10 pixels).

Validation of the accuracy of found positions of feature points is difficult to perform as we do not have the real positions of the feature points available for comparison. However, on the basis of visual inspection, we estimate that the localisation error for correctly found blobs corresponding to stickers (that means blobs that cover only one sticker), in most of the frames, is on the level of human performance and remains below 2 pixels. For blobs that cover more than one sticker or blobs corresponding to painted eyelids the approximated localisation error is slightly higher. The higher error for facial feature points 10 and 11 (on the eyelids) is mainly a consequence of bad illumination in this place (it is overshadowed by eyebrows and eyelashes) and the fact that during the recordings, as the result of blinking, the make-up from eyelids was partially wiped off.

CHAPTER 7

Feature Vector Classification

In this chapter we present a method for classification of the extracted feature vectors, applied to the recordings made as described in Chapter 3. The presented classification method is able to localise the face and its facial features. It also processes the facial feature vector and the audio feature vector and classifies the emotion the person is expressing. This procedure is not done real-time, because of the heavy pre-processing steps. An overview of the part that we use to classify emotions is explained in Paragraph 7.1 and the larger picture is visualised in Figure 25. This chapter deals with two aspects of multimodal emotion recognition:

- Feature vector extraction (from the audio clip and the video frames).
- Classification of emotions.

This process of feature selection is discussed in Paragraph 7.2. Besides feature selection, we reduce the feature vector size by removing points we assume to be noise and (key) points that always are located at the same position. The extracted feature vectors are very large, which forced us to reduce the data. Noise reduction is needed too, due to the fact that the head moves and the centre of mass can shift some pixels after pre-processing the image. Paragraph 7.2 presents these methods to reduce the data complexity and noise reduction. Paragraph 7.3 describes the translation of the facial feature points to AU activation. Paragraph 7.4 discusses the differences between the facial feature vectors for all of the 21 emotions. Paragraph 7.5 concludes with the methods we used to classify emotions based on the feature vectors.

7.1 Classification Model

The model of 31 facial feature points is very simple, as discussed before in Paragraph 6.4.2. FACS is a universal, general accepted, scoring method. Therefore our method connects the 31 facial points to different AUs and can classify which AUs are active and thus which emotion is being displayed. The movements of the points indicates facial changes, the direction of the moving points indicates activation of one or more AUs. Combination of several active AUs can define an emotion.

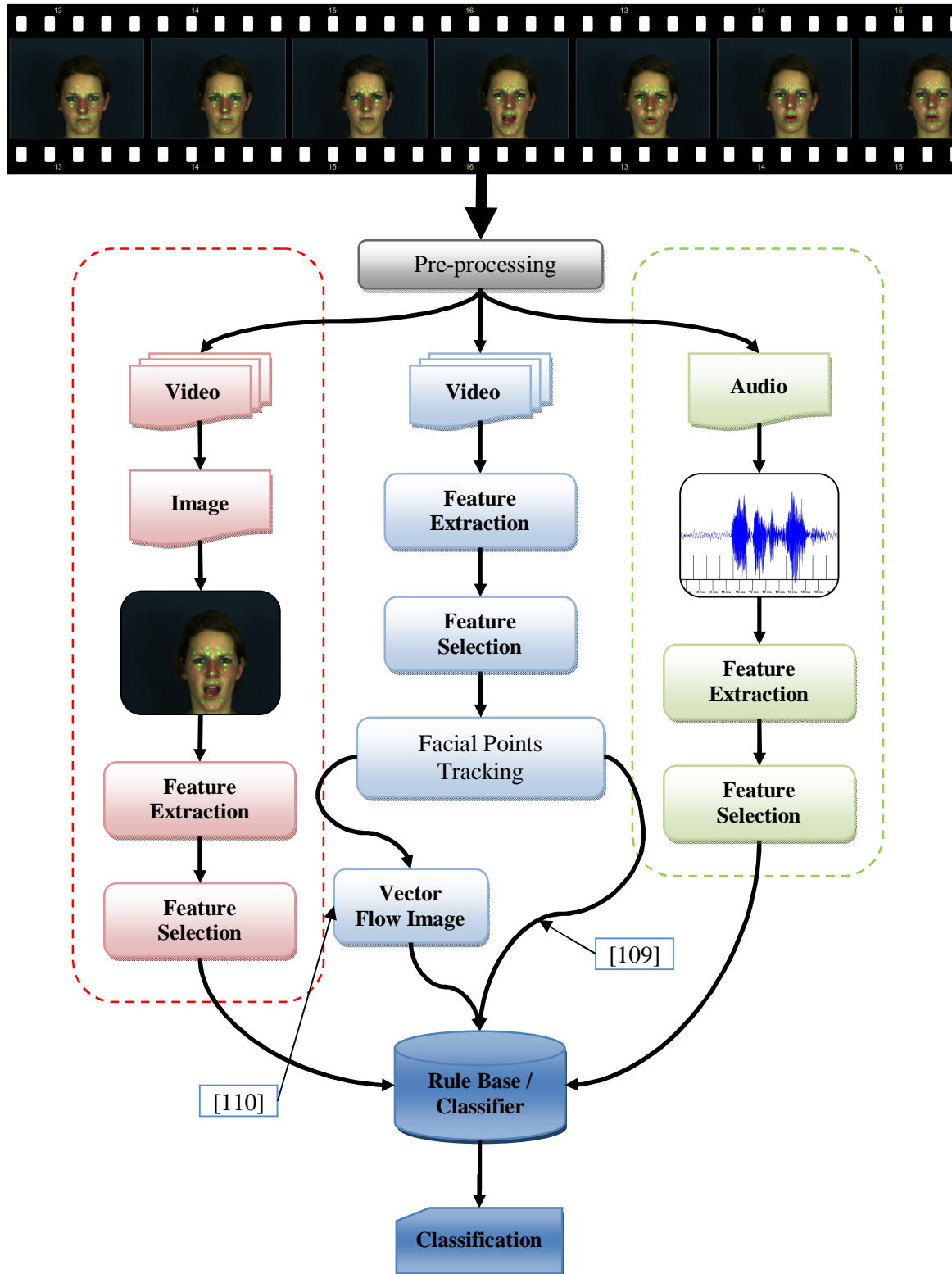


Figure 25: Large model view of multimodal emotion recognition.

Because of the limited time we have we did not focused on point tracking, though there has been recent research to this field of study [109] [110]. This approach resembles the blue path visualised in Figure 25 above. We decided to follow the approach represented within the red and green boxes. We recognise emotions using a Rule Base approach and the Pearson Correlation for the red path (images) and a multiple classifiers (see Table 14) for the green path (audio).

7.2 Feature Selection and Noise Reduction

Because we work with an audio feature vector of 91 features and a facial feature vector of 62 features the data complexity is high. We need to reduce this to a much lower dimension. Feature selection can help to do this. Furthermore we need to remove noise from the channels. Noise reduction can be done by subtracting the recordings from the second microphone (background microphone) from the recordings made by the speakers' microphone. Unfortunately, due to the time constraint we did not have the time to do so. We begin by introducing the boosting technique, applied to the audio feature vectors. The results for the feature selection done by GentleBoost are displayed in Table 9 below. The number of the features corresponds to the numbers of the list from Paragraph 6.3.

7.2.1 Feature Selection by Boosting Algorithms

When a lot of features are available, classification tends to be too costly and time-consuming. Also when the number of samples in our train set is small compared to the number of features, there is a high risk of over fitting to this train data resulting in an inaccurate classification system. Probably, only a very small number of these features can be combined to form an effective classifier. The main challenge is to find these features. A boosting program can be used as a feature selection algorithm. Boosting is a popular learning technique, proposed in the computational learning theory literature by Schapire and Freund [111] [112] [113]. Boosting is a general method for improving the accuracy of any given learning algorithm. Boosting combines the performance of many weak classifiers to produce a strong classifier.

7.2.2 Audio Feature Selection

Audio feature selection is done by the GentleBoost algorithm. This boosting algorithm selects the best possible features to make the best distinction between classes. The results are given below, notice that the numbers of the emotions are green again. This is to emphasise that we deal with audio features here.

Table 9: Results of feature selection done by GentleBoost for audio features.

	Emotion	Selected Best n Features
<i>1</i>	Admiration	29,14,33,25,78,55,23,19,17
<i>2</i>	Amusement	53,56,33,87
<i>3</i>	Anger	58,54,43,44,46,3,36,85,23,1
<i>4</i>	Anger (Surprise)	90,89,87,56,43,71,34,51,52,17,85,15,32,57,91,77,21,26,61,16,13,22,76,64,74,84,83
<i>5</i>	Boredom	37,20,22,57,54,91,23,78,2,79,46,73,61,30,11,35,60,32,49,13,25,75,87,33,10,21,86,84,48,52,53,26
<i>6</i>	Contempt	75,56,74,41,61,71,90,67,37,15,32,54,81,58
<i>7</i>	Desire	51,50,47,5,21,89,76,61,45,80,10,26,78,69
<i>8</i>	Disappointment	47,85,82
<i>9</i>	Disgust	78,76,54,37,60,51,10,52,80,81
<i>10</i>	Dislike	57,83,13,30,5,48,76,1,65,23,35,20,75,87,42,66,90,61
<i>11</i>	Dissatisfaction	91,32,78
<i>12</i>	Fascination	51,1,83,82,63,81,4,40,23,41,48,72,49,29,31,74,13,21,76,90,45,20,71
<i>13</i>	Fear	53,50,85,27,14
<i>14</i>	Furious	9,54,76,36,16,5,40,74,44,26,63,77,38,87,47,90,34
<i>15</i>	Happiness	18,76,27,5,54,20,72,13,87,64,16
<i>16</i>	Indignation	56,31,43,1,46,82,52,75,4,33,65,10,2,58

17	Interest	77,43,31,5,47,90,36,29,80,64,57,28,62,16,20
18	Sadness	51,33,60,78,16,5,64,49,41,65,81
19	Satisfaction	21,54,26,76,19,75,71,58,10,84,16,33,80,18,34,82,77,66
20	Pleasant Surprise	63,90,6,71,79,64,84,54,48,86,87,74,65,14,61,72,60,4,58,23,39
21	Unpleasant Surprise	64,83,87,78,20,89,11,43,72,26,50,34,47,5,40,52,67,81,23,54,41

The number of selected features is not always the same. The algorithm also calculates certain weights per feature and if the weights no longer improve, the algorithm stops.

7.2.3 Image Feature Selection

Selecting features from the facial feature points that can be omitted is very simple. As mentioned before we assume that the points around the mouth can be considered ‘noise’. If we do so, we do not need these points represented in our feature vector. The same holds for feature point 20, the nose point and feature point 1, the eye point. These points are always located in the same position and never shift more than 1 or 2 pixels. This is because we use these points in the normalisation process.

The reduced feature vector now consists of 18 features. The obsolete points are visualised with a red circle around them in Figure 26 below.

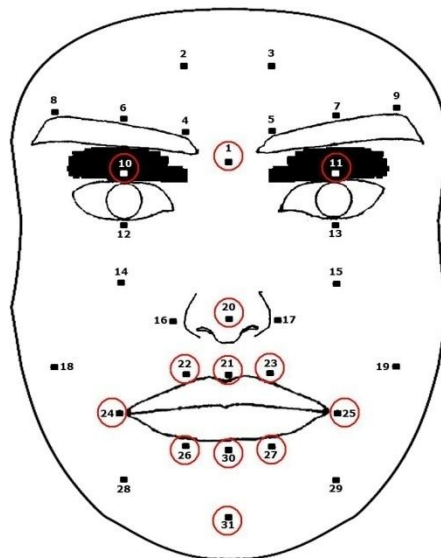


Figure 26: Visualisation of obsolete points.

7.2.4 Principal Component Analysis for Facial Feature Vectors

It is known that the first principal component of Principal Component Analysis of the coordinates of the 31 feature points corresponds to valence and that the second principal component corresponds to arousal. If we plot those two components for both the X-coordinate and Y-coordinate we expect similar patterns as displayed in Figure 17. However if we look at Figure 27 there is no distinction possible between the 21 emotions. Notice that due to the fact that Matlab can only handle a maximum of 20 different data types in its legend, there is no textual label for data type 21 (which corresponds to ‘Unpleasant Surprise’). However there are 21 different markers plotted.

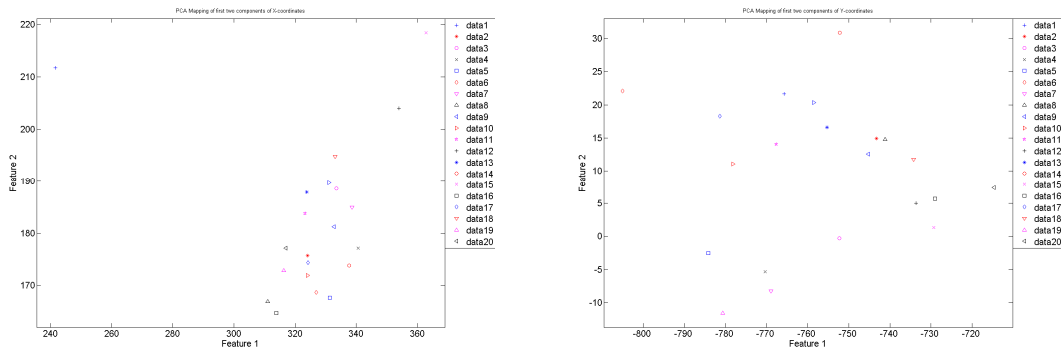


Figure 27: Plots of the first two principal components of the X-coordinates (left) and Y-coordinates (right) of all 31 facial feature points.

7.3 Facial Feature Point Translation

An AU represents the simplest visible facial movement, which cannot be decomposed into more basic ones. Each AU is controlled by contraction or relaxation of a single muscle or a small set of strongly related muscles. Activation of an AU is described by observable changes in the face caused by activity of the underlying muscles. In order to show facial expressions people usually activate more than just one AU. The implementation of the points to AUs translation is given in Appendix B.

Not all of the AUs can be scored independently; however, there are restrictions on how different AUs interact with each other or whether they are allowed to occur together at all. To make the translation from the feature points to the AUs we need to understand these restrictions and rules for AU activation. The implemented AUs are summarised in Table 10. In order to score AUs appearing in combinations, a FACS coder has to know how they influence each other. Ekman and Friesen introduced five different generic co-occurrence rules that describe the way in which AUs combine and influence each other [114]. The complete constraints implementation is found in Appendix C

Table 10: Implemented AUs.

AU	Description	AU	Description
AU1	Inner Brow Raiser	AU17	Chin Raiser
AU2	Outer Brow Raiser	AU18	Lip Puckerer
AU4	Brow Lowerer	AU20	Lip Stretcher
AU5	Upper Lid Raiser	AU22	Lip Funneler
AU6	Cheek Raiser	AU23	Lip Tightener
AU7	Lid Tightener	AU24	Lip Presser
AU9	Nose Wrinkler	AU25	Lips Part
AU10	Upper Lip Raiser	AU26	Jaw Drop
AU12	Lip Corner Puller	AU27	Mouth Stretch
AU15	Lip Corner Depressor	AU43	Eyes Closed
AU16	Lower Lip Depressor		

Most of combinations are *additive combinations*. This means that the facial changes are just the sum of all changes caused by each AU scored separately. The evidence of each activated AU from the combination is recognisable, and none of the facial changes, due to separate AUs, is modified in combination. Additive combinations are e.g AU5 ‘Upper Lid Riser’ and AU26 ‘Jaw Drop’.

The second rule for combining AUs is applied when one AU *dominates* the other. In combinations which involve dominance, a dominant AU can completely cancel the facial changes due to the subordinate AU or can make the evidence of scoring a subordinate AU very subtle and difficult to detect. To avoid errors in detecting subordinate AUs, Ekman and Friesen established the rule that prohibits scoring the subordinate AU in some particular combinations. For example AU6 ‘Cheek

Riser' dominates AU7 'Lid Tightener', and while describing the face with these two AUs activated, we should score only one AU6.

Besides these, only one AU should be scored when a combination is an *alternative combination*. The difference between this, and the previous rule is that in dominance both AUs could be activated, but only one was clearly visible (and thus scored), while in an alternative combination it is not possible to activate both AUs simultaneously. When two of the alternative AUs give similar appearance, a choice has to be made, which one should be scored. The reasons, why given AUs are alternative to each other can be as follow:

- Anatomy of our face doesn't allow scoring both AUs in the same time (e.g. AU5 'Upper Lid Raiser' and AU7 'Lid Tightener').
- It is impossible to discriminate one of the AUs from the occurrence of both simultaneously (e.g. AU41 – Lid Drop and AU43 – Eyes Closed Optional – if the upper lid is dropped it cannot be described as closed).
- The logic of FACS prohibits the scoring of both AUs at the same time.

The next kind of combination is called *substitution*. Substitution occurs when two combinations are so similar, that they can be scored in the same way. Usually, a scored combination is the one, which is notationally simpler. This can be seen with the combinations of AU13 'Sharp Lip Puller' and AU14 'Dimpler'; these must be scored just as a single AU12 'Lip Corner Puller'.

Finally all combinations which do not belong to any of the above described groups are called *different combinations*. In this kind of combinations, combination of given AUs involves new distinctive facial changes, which do not occur for those AU scored separately. The changes in the face are not just the sum of the changes caused by AUs scored separately but a result from their joint action. Sometimes, for example, one AU cancels one of the effects of another AU. In other cases all facial changes from scoring AUs separately are preserved and there are added new, distinctive changes which occur only in the combination.

For every AU we let zero activation, or deactivation, correspond to the neutral position of the point(s) that influence the AU. For every emotion we look at the maximal displacement of these points and from this we can calculate the percentage of displacement of a point. If the displacement of a certain point linked to an AU exceeds some threshold, we activate the AU. In this way we have linked most of the points to the implemented AUs described above. For a complete overview of which points correspond to which AU, we refer to Appendix A. Here a list of implemented AUs is given as well as the corresponding points per AU. The complete implementation of the AU activation can be found in Appendix C.

7.4 Feature Vector Comparison

As mentioned before we have compared the feature vectors of both participants (P1 and P2). We could easily distinguish 21 different expressions from Figure 15 and Figure 16, however the displayed expression from the two participants for the same emotion was not always the same. Visually we can see differences between some emotions, while others show many similarities. Examples of similar and different facial expressions are given in Figure 28 below. The similarity matrix is given in Appendix J, here the distances between all the feature vectors are displayed. We normalised the matrix and took the absolute values for every element. Normalisation was done according to (6).

$$N_{(i,j)} = \frac{M_{(i,j)} - \mu_i}{\sigma_i} \quad (6)$$

Where N is the normalised matrix, M is the matrix to be normalised, i the column, j the row, μ_i the mean of column i , σ_i the standard deviation of column i .

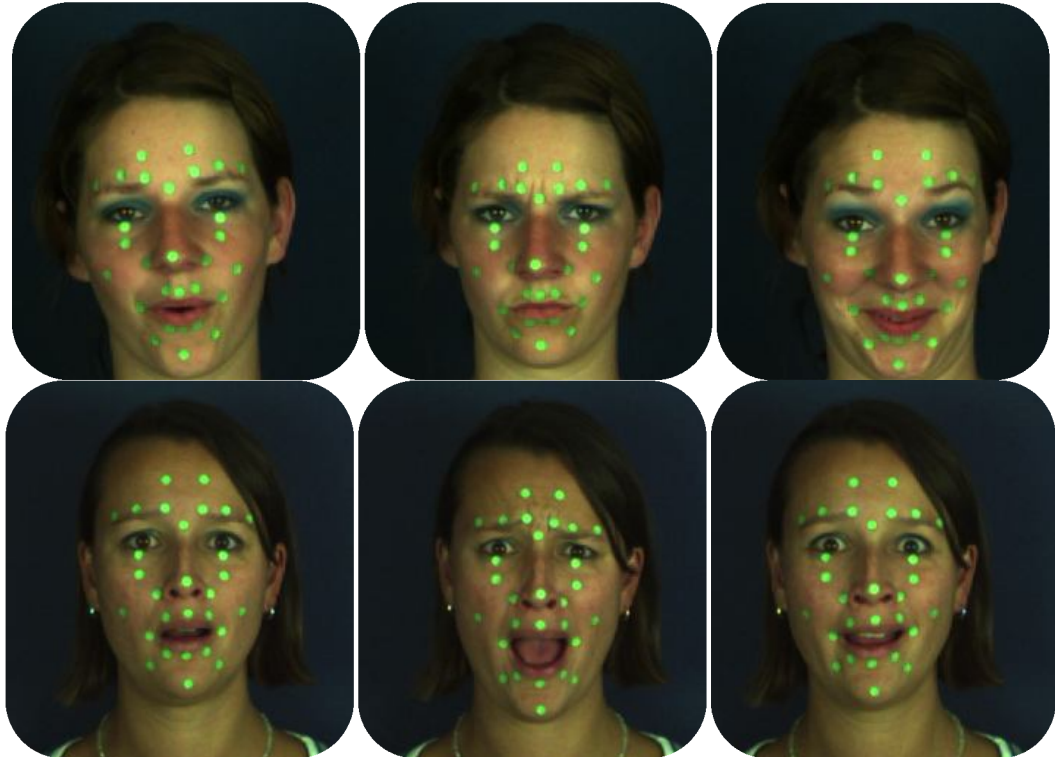


Figure 28: Facial differences for Participant P1 (top) and Participant P2 (bottom). Left: Admiration (2.5691), middle: Anger (1.9911) and right: Surprise Pleasant (0.0088).

The differences in the ‘Anger’ images are much greater than in the other image columns, however, the numbers show it otherwise. The distance between both facial feature vectors is smaller than the one from the ‘Admiration’ images. This is because for anger most of the differences are happening around the mouth and these points are discarded and thus the numbers give a different impression.

Table 11 shows the normalised distances between the 21 similar emotions. These numbers are taken from the larger similarity matrix from Appendix J. If we inspect the six archetypal emotions in Table 11 we can see that the interpretation, or expression, for some emotions is different per person. For example the difference between the middle figures in Figure 28 show a different interpretation of ‘Anger’ and the figures on the right have much similarity for the expression ‘Surprise (Pleasant)’.

Table 11: Normalised distances between facial feature vectors from P1 and P2.

Facial Feature Vector	Distance (from P1 to P2)	Facial Feature Vector	Distance (from P1 to P2)
Admiration	2.5691	Fear	0.3371
Amusement	0.0201	Furious	0.4618
Anger	1.9911	Happiness	0.5460
Boredom	0.2452	Indignation	0.0059
Contempt	0.2987	Interest	0.2309
Desire	1.3526	Neutral	0.7811
Disappointment	1.1719	Sadness	1.3069
Disgust	1.1406	Satisfaction	0.1860
Dislike	0.3240	Surprise (Pleasant)	0.0880
Dissatisfaction	0.1830	Surprise (Unpleasant)	0.9254
Fascination	0.2843		

7.5 Classification

The prototype uses decision-level fusion in the classification process, because of the multiple options of classification. The results for audio and video can easily be compared and conclusions can be drawn immediately.

For the audio classifier we have tested a lot of different classifiers (see Table 12). All classifiers are standard classifiers which are available in the PRTools toolbox. Most classifiers have been tested with standard parameters, some classifiers, like the SVC and the NNC, were given other than standard parameters. This was done because most classifiers in this toolbox tend to optimise the parameters themselves.

Table 12: All different trained classifiers from PRTools .

1	Support Vector Classifier (RBF kernel, C=1)
2	Support Vector Classifier (RBF kernel, C=20)
3	Radial Basis Support Vector Classifier
4	Parzen Classifier
5	K Nearest Neighbour Classifier
6	Fisher Classifier
7	Linear Discriminant Classifier
8	Nearest Mean Classifier
9	Quadratic Classifier
10	Neural Network Classifier (hidden layer with 5 neurons)
11	Neural Network Classifier (hidden layer with 15 neurons)

The audio classifier is trained on 935 audio samples from which 35 samples were considered positive and all other samples were regarded negative for each of the 21 emotions. Besides this classifier we trained two more classifiers, one for the classification of positive and negative samples (a valence classifier) and one for the classification of active and passive samples (an arousal classifier).

As described in Paragraph 7.2, we translate the facial feature points to AU activation. We then try to classify the current activation, or de-activation, of the AUs as an emotion. The rules describing which AUs correspond to which emotions, is retrieved from [14], an implementation of these rules is shown in Appendix C. These rules described the following emotions:

Table 13: 13 Template expressions.

Anger	Grief	Sadness
Astonishment	Happiness	Satisfaction
Disbelief	Irony	Surprise
Disgust	Regret	Understanding
Fear		

As an alternative classification method we calculated the Pearson Correlation to measure the distance for a given input facial feature vector to a known facial feature vector corresponding to an emotion. The Pearson Correlation measures the similarity in shape between two profiles. However it can also capture inverse relationships.

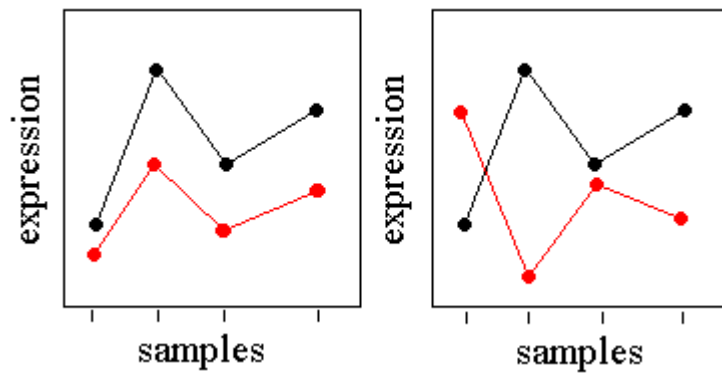


Figure 29: Two examples for the Pearson Correlation.

In Figure 29 on the left, the black profile and the red profile have an almost perfect Pearson correlation despite the differences in basal expression level and scale. These samples would cluster together with the Pearson Correlation. In Figure 29 on the right, the black and red profiles are almost perfectly opposing. These samples would be placed in remote clusters using Pearson Correlation. The formula for the Pearson correlation is

$$d = 1 - \frac{Z(x) \cdot Z(y)}{n} \quad (7)$$

where $Z(x)$ and $Z(y)$ are the normalised values of the feature vector and n is the number of coordinates in the feature vector.

CHAPTER 8

Classification Results

In this chapter we discuss the classification results of our prototype. Because we used decision-level fusion we are able to separately handle the classification method for audio and video. Both classification results are showed and a selection between different classifiers can be made. The Matlab toolbox PRTools comes with multiple implementations of various classifiers.

First we will describe the audio classification results and the comparison of classification results for various classifiers applied to the audio feature vectors. These findings are discussed in Paragraph 8.1. Second we will describe the classification of the facial feature vectors and the correctly processing of frames. Not all frames were processed correctly, so from these frames we could not get any classification. These findings are discussed in Paragraph 8.2.

8.1 Audio Feature Classification Results

The classification error rates, given in Table 14 below, are obtained by training the corresponding classifier 5 times and averaging the overall error rate. The training set consists of 80% random samples as negative samples and the training set holds the remaining 20% as positive samples.

Table 14: Mean error rates per classifier per dataset.

Classifier:	21 Emotions	Positive vs. Negative	Active vs. Passive	AVERAGE
SVC (RBF, c=1)	0.95095	0.37881	0.42857	<i>0.5861</i>
SVC (RBF, c=20)	0.80476	0.38000	0.42857	<i>0.5378</i>
RBSVC	0.77000	0.39952	0.42857	<i>0.5327</i>
PARZENC	0.76786	0.38643	0.39167	<i>0.5153</i>
KNNC	0.76810	0.40881	0.38452	<i>0.5205</i>
FISHERC	0.85476	0.41119	0.46786	<i>0.5779</i>
LDC	0.84952	0.41095	0.46786	<i>0.5761</i>
NMC	0.84429	0.40667	0.45000	<i>0.5670</i>
QUADRATICC	0.85476	0.41119	0.46786	<i>0.5779</i>
NNC (5 hidden neurons)	0.95976	0.38833	0.42262	<i>0.5902</i>
NNC (15 hidden neurons)	0.93630	0.38595	0.44762	<i>0.5900</i>
AVERAGE	<i>0.85100</i>	<i>0.39710</i>	<i>0.43510</i>	

Table 14 shows that the Parzen classifier gives the best results for the classification of an unknown sample into one of 21 classes. Here the classifier is trained as a multiclass classifier. The difference

with a single class classifier is that this classifier is trained with 21 classes, where the other one is trained with only 2 classes (one positive and the rest are considered negative samples). For the positive versus negative classifier the SVC, with parameters Kernel='Radial Basis Function' and C=20, gives the best results. The best classifier for the active versus passive classification is the K-Nearest Neighbour Classifier. These two classifiers are examples of a single class classifier.

Table 15: Number and percentage of correct classified clips.

Emotion	Number of correct classified clips (from 5 clips)		
	21 classifiers	Positive vs. Negative	Active vs. Passive
Admiration	4 (80%)	1 (20%)	3 (60%)
Amusement	1 (20%)	3 (60%)	5 (100%)
Anger	4 (80%)	1 (20%)	2 (40%)
Anger (Surprise)	2 (40%)	2 (40%)	1 (20%)
Boredom	1 (20%)	1 (20%)	2 (40%)
Contempt	3 (60%)	0 (0%)	5 (100%)
Desire	3 (60%)	1 (20%)	3 (60%)
Disappointment	2 (40%)	2 (40%)	2 (40%)
Disgust	2 (40%)	1 (20%)	3 (60%)
Dislike	1 (20%)	1 (20%)	4 (80%)
Dissatisfaction	1 (20%)	1 (20%)	3 (60%)
Fascination	2 (40%)	2 (40%)	4 (80%)
Fear	2 (40%)	1 (20%)	0 (0%)
Furious	3 (60%)	2 (40%)	1 (20%)
Happiness	3 (60%)	2 (40%)	5 (100%)
Indignation	2 (40%)	2 (40%)	4 (80%)
Interest	4 (80%)	2 (40%)	2 (40%)
Sadness	0 (0%)	4 (80%)	1 (20%)
Satisfaction	0 (0%)	4 (80%)	5 (100%)
Pleasant Surprise	0 (0%)	3 (60%)	2 (40%)
Unpleasant Surprise	0 (0%)	2 (40%)	5 (100%)
Average	38%	36%	59%

The results in Table 15 show clearly that it was hard to train all three classifiers for the recognition of 21 emotions. The classifier that was trained on 21 classes did not even recognise at least one sample from the last four classes. The positive versus negative classifier should perform a lot better, according to the error rate from Table 14, but the outcomes show a different result. Even for the most expressive emotions, like Anger, Fear, Furious and Happiness, this classifier could not get a recognition percentage above 80%. Although, according to Table 14, the active versus passive classifier should perform slightly worse than the positive versus negative classifier, Table 15 shows better classification results. The 'bad' outcomes of these results can be the result of several causes, namely:

1. Overfitting.
The classifier can be trained to perfectly fit the training data, but when new samples are classified that are just outside the decision boundary, these can be wrongly classified.
2. Lack of samples.
If there are too few samples of each class, therefore, the classifier is not sure of the interpretation or the classes have very similar and/or overlapping features and cannot be classified successfully.
3. Lack of expressiveness.
The training of the audio classifiers is based upon the samples of two participants. This means that the decision boundary is set by the interpretation of only these two participants. A more variety of participants should ensure a more generalisable classifier.

8.2 Facial Feature Classification Results

Facial feature classification is done in two ways. We did a point based classification based on the found facial feature points and an AU based classification based on the activation or de-activation of the implemented AUs.

For the point based classification we used the Pearson Correlation. This correlation measures the similarity between samples. For our facial feature vectors we used the facial feature vectors corresponding to the 21 emotions also used for the image validation. We used 105 different samples, each emotion represented by 5 different video clips. Every clip was played once and the emotion was based on the most frequent emotion throughout the whole clip.

For the AU classification we decided to measure the Euclidean distance between different AU activation vectors. An AU activation vector is a vector where each entry correspond to an AU, this entry can either be on (activated or 1) or off (de-activated or 0). Pre-defined AU activation vectors are constructed from the research from [14].

Table 16: Number and percentage of correct classified clips.

Emotion	Number of correct classified clips (all processed frames)		Number of correct classified clips (only middle half of processed frames)	
	Point Based	AU Based	Point Based	AU Based
Admiration	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Amusement	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Anger	0 (0%)	0 (0%)	1 (20%)	0 (0%)
Anger (Surprise)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Boredom	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Contempt	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Desire	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Disappointment	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Disgust	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Dislike	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Dissatisfaction	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Fascination	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Fear	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Furious	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Happiness	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Indignation	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Interest	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Sadness	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Satisfaction	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Pleasant Surprise	0 (0%)	0 (0%)	2 (40%)	0 (0%)
Unpleasant Surprise	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average	0%	0%	3%	0%

These results are extraordinary, because the prototype cannot classify any emotions from the video frames. The cause for the misclassification can be the fact that we look for the most frequent emotion in all the processed frames. The clips all should start and end with a neutral expression, so the most frames should contain a neutral emotion. Adjusting the classification protocol from all frames to just the middle half of the clips, this is the part where we assume the apex of the expression is, also did not solve this problem.

8.2.1 Number of Correct Processed Frames

As a secondary measure we measured the percentage of correct processed frames. A correct processed frame is a frame in which the facial feature points are located correctly and we could extract useful information from these locations. The prototype processes almost every frame from the clips correct. The last column in Appendix I shows the percentage of correct processed frames for that clip. If most frames of the video clip cannot be processed the facial feature classification results are not trustworthy. Here only the percentages are given in Table 17.

Table 17: Percentage of correct processed frames per emotion.

Emotion	Number of correct processed frames
Admiration	39.4
Amusement	67.2
Anger	32.8
Anger (Surprise)	75.2
Boredom	34.4
Contempt	98.0
Desire	98.8
Disappointment	99.8
Disgust	98.4
Dislike	96.0
Dissatisfaction	94.8
Fascination	99.4
Fear	99.2
Furious	96.8
Happiness	99.6
Indignation	96.8
Interest	98.4
Sadness	100
Satisfaction	99.6
Pleasant Surprise	96.4
Unpleasant Surprise	99.6
Average	86.7

The average number of correct processed frames is 86.7 %. This means that for some emotions the correct processing of frames is low. This can also be seen in Table 17. Even so there are cases of a 100% correct processing percentage. The lower percentages can be the cause of movement of the head, or such facial deformations that too many green stickers were invisible or too faint to process.

The emotions for which the correct number of processed frames is very low are: Admiration, Anger and Boredom. The classification results for these emotions cannot be very trustworthy and should be omitted from the overall results.

CHAPTER 9

Conclusions and Future Work

In this thesis, the recording and analysis of multimodal data has been discussed. We have presented a protocol for the recording of 21 different emotions and we have developed a prototype in order to recognise emotions from these multimodal recording. In Chapter 1 we formulated a number of research questions to answer during this thesis project. In this chapter we try to give answers to these questions. These answers are given below.

✓ *Can we successfully create a protocol for the creation of a multimodal database?*

We believe that we not only have succeeded in the creation of a successful protocol for recordings for a multimodal database, but we also have successfully setup a new multimodal database. These findings are described in Chapter 4. The translation of the protocol to other languages can be done easily. It is recommended that the same scenarios should be used, but improvements or additions can be necessary for new languages. This database and its content should be made available to as many researchers and research groups as possible in order to really compare the results these different groups achieve. A description of an approach to do so is extensively covered in Paragraph 4.5.

✓ *Can we rank 21 different emotions on the valence and arousal scale?*

If we look at the validation scores people gave we can clearly see some clusters for very expressive and active emotions. It seems that there is not much information about the context needed to give these expressions the same scores for images, audio, video and text. On the other hand these results point out once more that contextual information is of great importance for emotion recognition of some emotions.

✓ *Can we automate the classification of 21 different emotions?*

The recognition of 21 different emotions was a difficult task. Most emotions have a lot of overlap and therefore a classifier could not be trained well enough. The classification of the audio samples resulted in an average error rate for 21 emotions for the tested classifiers from Table 14 of 0.85100; this means an average classification of $0.14900 \approx 14.9\%$, but a 23.21% of correct recognition by the parzen classifier. Similar we can calculate the best recognition rates for the other two classifiers too. These are 62% with a SV classifier for the positive vs. negative samples and 61.55% with a parzen classifier for the active vs. passive samples.

✓ *Do users agree on the ranking of multimodal emotions compared to the textual ranking?*

Concluding from the results we showed in Chapter 5, Figure 17, we clearly can distinguish some clusters where all validations are more or less the same.

9.1 Concluding Remarks

The lack of a multimodal database that could serve as a benchmark for future multimodal emotion recognition systems forced us to develop our own multimodal database. We know that in this case we cannot compare our results with the results accomplished by other multimodal researchers, but at least we try to make our data as public to these researchers as possible. This database has great potential and should be used as a benchmark database for lip reading systems as well as (multimodal) emotion recognition systems.

The used data corpus for this project was of great importance. If we did not have this data we had to search for other multimodal corpora who had and we haven't come across one who did. Unfortunately we only had the recordings of one participant to work with. This is not a good basis to work with, but because we did not have that much time we decided to go along with it. The protocol, as described in this thesis, is a suitable protocol for recording multimodal clips. This protocol takes into consideration the environmental influences, the setup of equipment and participant and it describes clearly the steps to follow in order to come to suitable recordings. The e'NTERFACE approach is a step in the right direction, but our improvements to this approach are the next step in the evolution of emotion recognition to expand from the six archetypal emotions to more useful emotions which we use in our daily lives.

The validation of the data gave a good impression on what to expect from the program, using this data. The validation also gave some interesting results considering the validation of one emotion for different unimodal channels and all combined. It is clearly shown that humans need contextual information too in order to recognise all emotions correctly. This can be concluded from the 'weird' scores that the image validation gave to the emotion 'contempt'. At first the facial expression looks like it is showing some sort of smile, but without the context one would never know that the intended emotion was negative. Other emotions were scored equally well over all three validation processes. Emotions showing a lot of arousal, e.g. very active emotions like happiness, disgust and furious, are easily distinguishable no matter what validation method was used.

9.2 Future Work

Making recordings does not take long, but the post-processing of the recorded data is a time consuming task. Therefore we recommend that in the future a steady flow of recordings is being made, annotated carefully and added to the database, to make sure the database keeps increasing in its size. The database now is very small and unorganised too. The way we will store and access this data will be via a web based application. This application has to be developed and users have to be granted access to work with the data. In order to get as many correctly labelled data as possible, there is the possibility to let users provide unlabelled data with metadata first before they can download whatever they are looking for in the database.

The validation web site we created can be a part of the new web application. This web site is an easy way to let users validate data, however people can download these clips from the internet, so access rights should be provided or validation should only be limited for those who are given access to the database.

The training of the audio classifiers can be significantly improved if we had more time to record more people and made sure we had male and female recordings available. More expressiveness in the expressed emotions can help, but if this will give the required results is something we are not sure about. The usage of real actors for recordings can help solve this problem, as they can over-express themselves and they should be able to invoke whatever emotion possible. Sometimes it was difficult for the participant to express the required emotion, just because they did not know how to express this correctly.

Anxiety can be an obstacle for multimodal recordings. People who cannot express themselves anymore, or only in one way are of no use for the database. For the video approach we hope that in the near future the points extraction does not depend on coloured stickers anymore and an implementation

of facial feature points is used as a solid basis for the points tracker. These methods should then be applied to accurate (facial) expressions. If we have recordings of real extreme cases of emotional expressions we have an idea where these emotions are located in time, as well as in the valence and arousal space. Other future work can include vector flow classification. Research has been done in this direction and the implementation into the model should not be that hard.

Furthermore, the high speed recordings could be quite suitable for synchronising the video frames with the audio features extracted with a 10 ms window. In this case each window will correspond to one frame and the feature vectors could be concatenated into one large feature vector. This way feature-level fusion can be performed. Here too, the time constraints forced us to set aside this approach but we expect that in the near future this approach will be followed and that the results are higher than our accomplished results.

Bibliography

- [1] E. Boyle, A. H. Anderson, and A. Newlands, "The effects of visibility on dialogue and performance in a co-operative problem solving task," *Language and Speech*, vol. 37, pp. 1-20, 1994.
- [2] J. A. Russel and J. M. Fernandez-Dols, "The psychology of Facial Expression," vol. 9, no. 3, pp. 185-211, 1990.
- [3] K. S. Benoit and C. Mohamadi, "Audio-visual intelligibility of French speech in noise," *Journal of Speech and Hearing Research*, vol. 37, pp. 1195-1203, 1994.
- [4] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215, 1954.
- [5] J. A. Graham and M. Argyle, "A cross-cultural study of the communication of extra-verbal meaning by gestures," *International Journal of Psychology*, vol. 10, pp. 67-67, 1975.
- [6] P. Ekman, *Telling lies: Clues to deceit in the marketplace, marriage and politics*. New York: Berkeley Books, 1985.
- [7] P. Ekman, *Emotions Revealed*. New York: Times Books, 2003.
- [8] C. Darwin, *The Expression of the Emotions in Man and Animals*. 1872.
- [9] M. v. Vulpen, E. A. P. Smits, and M. Grootveld, "AVE: Using Audio and Video for Emotion recognition," TU Delft Bachelor Thesis, 2005.
- [10] A. G. Chitu and L. J. M. Rothkrantz, "The Influence of Video Sampling Rate on Lipreading Performance," in *Proceedings of the 12th International Conference on Speech and Computer (SPECOM'2007)*, Moscow, Russia, 2007.
- [11] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database," in , Atlanta, Apr. 2006.
- [12] E. Douglas-Cowie, D. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 39-44, 2000.
- [13] P. Ekman, "The argument and evidence about universals in facial expressions of emotion," in *Handbook of social psychology*. Chichester, England: Wiley, 1989, pp. 143-164.
- [14] A. Wojdel, "Knowledge Driven Facial Modelling," PhD Thesis, Delft University of Technology, Delft, 2005.
- [15] P. M. A. Desmet, "Designing Emotions," MSc Thesis, Delft University of Technology, Delft, 2002.

- [16] C. M. Whissell and M. J. Dewson, "The Dictionary of Affect in Language," in *Emotion: Theory, Research, and Experience*. New York, USA: Academic Press, 1989, vol. 18, p. 113–131.
- [17] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition (AFGR 2000)*, Grenoble, France, 2000, pp. 46-53.
- [18] P. Ekman and W. V. Friesen, *The facial action coding system: A technique for measurement of facial movement*. 1978.
- [19] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, Nara, Japan, 1998, p. 366.
- [20] L. S. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2000)*, New York City, USA, 2000, pp. 423-426.
- [21] T. S. Huang, L. S. Chen, H. Tao, T. Miyasato, and R. Nakats, "Bimodal Emotion Recognition by Man and Machine," in *proceeding of ATR Workshop on Virtual Communication Environments*, Kyoto, Japan, 1998.
- [22] L. C. D. Silva, T. Miyasato, and R. Nakatsu, "Facial Emotion Recognition using Multi-modal Information," in *Proceedings of the 1st International Conference on Information, Communications and Signal Processing (ICICS 1997)*, Beijing, China, 1997, pp. 397-401.
- [23] L. C. DeSilva and P. C. Ng, "Bimodal Emotion Recognition," in *Proceedings of the 4th IEEE International Conference on Automatic Face & Gesture Recognition (AFGR 2000)*, Arnoldstein ,Austria, 2000, pp. 332-335.
- [24] H. Go, K. Kwak, D. Lee, and M. Chun, "Emotion Recognition from the Facial Image and Speech Signal," in *Proceedings of the 42nd Annual Conference of the Society of Instrument and Control Engineers (SICE 2003)*, vol. 3, Fukui, Japan, 2003, pp. 2890-2895.
- [25] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T. S. Huang, "Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition," in *Proceedings of the 7th ACM International Conference on Multimedia*, Singapore, 2005, pp. 65-68.
- [26] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, and T. Zhang, "Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI," in *Proceedings of the 5th International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, USA, 2005, pp. 967-972.
- [27] Z. Zeng, J. Tu, M. Liu, and T. S. Huang, "Multi-stream Confidence Analysis for Audio-Visual Affect Recognition," in *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII 2005)*, Beijing, China, 2005, pp. 946-971.
- [28] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, "Audio-visual Affect Recognition in Activation-evaluation Space," in *Proceedings on the 6th International Conference on Multimedia & Expo (ICME 2005)*, Amsterdam, Netherlands, 2005, pp. 828-831.
- [29] Z. Zeng, et al., "Audio-Visual Spontaneous Emotion Recognition," in *Artificial Intelligence for Human Computing*, 2007, pp. 72-90.
- [30] M. Song, J. Bu, C. Chen, and N. Li, "Audio-Visual Based Emotion Recognition - A New Approach," in *Proceedings of the 4th Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, Washington, USA, 2004, pp. 1020-1025.
- [31] C. Busso, Z. Deng, and S. Yildirim, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in [83] C. Busso, Z. Deng, S. Yildirim, et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information", in *Proceedings of ACM 6th International Conference on Mutlmodal Interfaces (ICMI 2004)*, State College, USA, 2004.

- [32] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, USA, 2005, pp. 1085-1088.
- [33] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, USA, 2005, pp. 1125-1128.
- [34] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, China, 2006, pp. 1136-1139.
- [35] M. Pantic and L. J. M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881-905, 2000.
- [36] C. Busso, et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of ACM 6th International Conference on Multimodal Interfaces (ICMI 2004)*, State College, Pennsylvania, USA, 2004, pp. 205-211.
- [37] J. Jiang, A. Alwan, E. Auer, and L. Bernstein, "Predicting visual consonant perception from physical measures," in *Proceedings of Eurospeech*. Aalborg, Denmark, 2001, pp. 179-182.
- [38] R. Cowie and E. Douglas-Cowie, "Emotion Recognition in human-Computer-Interaction," *IEEE Signal Processing Magazine*, Jan. 2001.
- [39] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1167-1178, 1980.
- [40] J. Ostermann, "MPEG-4 overview," in *Circuits and Systems in the Information Age*, 1997, pp. 119-135.
- [41] T. Sikora, "The MPEG-4 video standard and its potential for future applications," in *Proceedings of the conference on International Symposium on Circuits and Systems (ISCAS 1998)*, Hong Kong, China, 1997.
- [42] M. Pantic and L. J. M. Rothkrantz, "Facial Action Recognition for Facial Expression Analysis from Static Face Images," *IEEE Transactions on Systems, Man and Cybernetics – Part B*, vol. 34, no. 3, pp. 1449-1461, Jun. 2004.
- [43] M. S. Bartlett, G. Littlewort, B. Braathen, T. J. Sejnowski, and J. R. Movellan, "A prototype for automatic recognition of spontaneous facial actions," in *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press, 2003, pp. 1271-1278.
- [44] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 699-714, May 2005.
- [45] A. Colmenarez, B. J. Frey, and T. S. Huang, "A probabilistic framework for embedded face and facial expression recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 592-597, 1999.
- [46] C. Shan, S. Gong, and P. W. McOwan, "Dynamic Facial Expression Recognition Using A Bayesian Temporal Manifold Mode," *Computer Vision Publications*, vol. 1, pp. 297-306, Sep. 2006.
- [47] I. Cohen, N. Sebe, F. G. Cozman, and T. S. Huang, "Semi-Supervised Learning for Facial Expression Recognition," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, Berkeley, California, 2003, p. 281.
- [48] I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang, "Learning Bayesian Network Classifiers for Facial Expression Recognition with both Labeled and Unlabeled data," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Toronto, Canada, 2003, pp. 595-601.

- [49] J. Hoey, "Hierarchical unsupervised learning of facial expression categories," in *Proceedings of ICCV Workshop on Detection and Recognition of Events in Video*, Vancouver, Canada, 2001, pp. 99-106.
- [50] Z. Hammal, A. Caplier, L. Couvreur, and M. Rombaut, "Facial Expression Recognition Based on The Belief Theory: Comparison With Different Classifiers," in *International Conference on Image Analysis and Processing (ICIAP 2005)*, Cagliari, Italy, 2005.
- [51] I. Cohen, N. Sebe, S. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modelling," in *Computer Vision and Image Understanding*, vol. 91, 2003, p. 160–187.
- [52] I. Cohen, A. Garg, and T. S. Huang, "Emotion Recognition using Multilevel-HMM," in *NIPS Workshop on Affective Computing*, Colorado, 2000.
- [53] J. J. Lien, "Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity," MSc Thesis.
- [54] J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Automated facial expression recognition based on face action units," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition (AFGR 1998)*, 1998, pp. 390-395.
- [55] J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Detection, tracking, and classification of action units in facial expression," *Robotics and Autonomous Systems*, vol. 31, pp. 131-146, 2000.
- [56] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences using Geometric Deformation Features and Support Vector Machines," *IEEE Transactions on Image Processing*, vol. 1, no. 16, pp. 172-187, 2007.
- [57] I. Kotsia and I. Pitas, "Real time facial expression recognition from image sequences using support vector machines," in *Proceedings of the International Conference on Image Processing (ICIP 2005)*, 2005, pp. 966-969.
- [58] I. Kotsia, N. Nikolaidis, and I. Pitas, "Facial Expression Recognition in Videos Using Novel Multi-Class Support Vector Machines Variant," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, Hawaii, 2007.
- [59] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, vol. 26, Vancouver, Canada, 2003.
- [60] P. Michel and R. El Kaliouby, "Real Time Facial Expression Recognition in Video using Support Vector Machines," in *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI 2003)*, Vancouver, Canada, 2003, pp. 258-264.
- [61] I. Kotsia, N. Nikolaidis, and I. Pitas, "Fusion of Geometrical and Texture Information for Facial Expression Recognition," in *Proceedings of the International Conference on Image Processing (ICIP 2006)*, Atlanta, USA, 2006, pp. 2649-2652.
- [62] M. S. Bartlett, et al., "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, USA, 2005, pp. 568-573.
- [63] D. Datcu and L. J. M. Rothkrantz, "Facial expression recognition with relevance vector machines," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands, 2005, pp. 193-196.
- [64] Y. L. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity," in *Proceedings of the 5th IEEE International Conference Automatic Face and Gesture Recognition (AFGR 2002)*, Los Alamitos, USA, 2002, pp. 218-223.

- [65] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition (AFGR 1998)*, Nara, Japan, 1998, pp. 200-205.
- [66] W. Liu and Z. Wang, "Facial Expression Recognition Based on Fusion of Multiple Gabor Features," in *18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, China, 2006, pp. 536-539.
- [67] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron," in *3rd International Conference on Face and Gesture Recognition (FG98)*, Nara, Japan, 1998, pp. 454-461.
- [68] L. Franco and A. Treves, "A Neural Network Facial Expression Recognition System using Unsupervised Local Processing," in *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis (ISPA 2001)*, Pula, Croatia, 2001.
- [69] B. Fasel, "Multiscale Facial Expression Recognition Using Convolutional Neural Networks," in *Proceedings of the third Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2002)*, Ahmedabad, India, 2002.
- [70] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Network*, vol. 16, no. 5-6, p. 555-559, Jun. 2003.
- [71] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, Feb. 2001.
- [72] D. Yang and L. Jin, "Facial Expression Recognition with Pyramid Gabor Features and Complete Kernel Fisher Linear Discriminant Analysis," *International Journal of Information Technology*, vol. 11, no. 9, pp. 91-100, 2005.
- [73] W. Sun and Q. Ruan, "Two-Dimension PCA for Facial Expression Recognition," in *8th International Conference on Signal Processing (ICSP 2006)*, Guilin, China, Nov. 2006.
- [74] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognition and Image Analysis*, vol. 15, no. 2, pp. 546-548, 2005.
- [75] B. Abboud and F. Davoine, "Facial expression recognition and synthesis based on an appearance model," *Signal Processing: Image Communication, Elsevier*, vol. 19, no. 8, pp. 723-740, Sep. 2004.
- [76] X. Feng, "Facial Expression Recognition Based on Local Binary Patterns and Coarse-to-Fine Classification," in *Proceedings of the Conference on Instructional Technologies (CIT 2004)*, New York, USA, 2004, pp. 178-183.
- [77] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 9, pp. 757-763, 1997.
- [78] Y. Zhan, J. Ye, D. Niu, and P. Cao, "Facial Expression Recognition Based on Gabor Wavelet Transformation and Elastic Templates Matching," *International Journal on Image Graphics*, vol. 6, no. 1, pp. 125-138, 2006.
- [79] K. Karpouzis, G. Votsis, G. Moschovitis, and S. Kollias, "Emotion Recognition Using Feature Extraction and 3-D Models," in *Proceedings of International Multiconference on Circuits and Systems Communications and Computers (CSCC 1999)*, Athens, Greece, 1999.
- [80] I. A. Essa and A. P. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *The International Conference on Computer Vision (ICCV 1995)*, Cambridge, USA, 1995.











- [81] M. S. Bartlett, G. Littlewort, B. Braathen, T. J. Sejnowski, and J. R. Movellan, "An approach to automatic analysis of spontaneous facial expressions," in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR 2002)*, Washington, USA, 2002.
- [82] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, 2004, pp. 593-596.
- [83] C. C. Chiu, Y. L. Chang, and Y. J. Lai, "The analysis and recognition of human vocal emotions," in *Proceedings of the International Computer Symposium*, Hsihchu, Taiwan, 1994, p. 83-88.
- [84] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, USA, 1996, p. 1970-1973.
- [85] V. A. Petrushin, "How well can people and computers recognize emotions in speech?," in *Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium (AAAI 1998)*, Orlando, Florida, 1998, p. 141-145.
- [86] K. R. Scherer, "Adding the affective dimension: A new look in speech analysis synthesis," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, USA, 1996, pp. 1808-1811.
- [87] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 20, no. 3, Mar. 1998.
- [88] I. Ludmila and A. Kuncheva, "Theoretical Study on Six Classifier Fusion Strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, Feb. 2002.
- [89] L. C. D. Silva, T. Miyasato, and R. Nakatsu, "Facial Emotion Recognition Using Multimodal Information," in *Proceedings of the 1st International Conference on Information, Communications and Signal Processing*, Singapore, 1997, pp. 397-401.
- [90] A. Corradini, M. Mehta, N. O. Bernsen, and J. C. Martin, "Multimodal Input Fusion In Human Computer Interaction On The Example Of The On-Going Nice Project," in *Proceedings of the NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, Yerevan, Armenia, 2003.
- [91] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 854-866, Jul. 2002.
- [92] A. Kapoor, R. W. Picard, and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04) - Volume 3*, Cambridge, United Kingdom, 2004, pp. 969-972.
- [93] P. Ekman, *Emotion in the Human Face*, 2nd ed., P. Ekman, Ed. New York, NY: Cambridge University Press, 1982.
- [94] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *Journal of the Acoustic Society of America*, vol. 52, no. 4, p. 1238-1250, 1972.
- [95] A. G. Chitu, M. van Vulpen, P. Takapoui, and L. J. M. Rothkrantz, "Building a Dutch multimodal Corpus for Emotion Recognition," in *Proceedings of the 2nd International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation (LREC 2008)*, vol. 2, Marrakech, Morocco, May 2008, pp. 53-56.
- [96] D. A. Patterson, G. A. Gibson, and R. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," in *Proceedings of the 14th Conference of the ACM Special Interest Group on Management of Data (SIGMOD 1988)*, Chicago, USA, Jul. 2008, pp. 109-116. [Online]. http://en.wikipedia.org/wiki/RAID
- [97] E. C. Tolman, *Purposive Behavior in Animals and Man*. New York, USA: Century, 1932.

- [98] L. C. Charland, "Emotion experience and the indeterminacy of valence," in *Emotions: Conscious and Unconscious*. New York, USA: Guilford Press, 2005.
- [99] L. C. Charland, "The heat of emotion: Valence and the demarcation problem," *Journal of Consciousness Studies*, vol. 12, pp. 82-102, 2005.
- [100] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament," *Journal of Current Psychology*, vol. 14, no. 4, pp. 261-292, Dec. 1996.
- [101] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49-59, Mar. 1994.
- [102] R. P. W. Duin, et al. (2007) PRTools4.1, A Matlab Toolbox for Pattern Recognition. [Computer program].
- [103] M. Richert. (2008, Mar.) MATLAB Central File Exchange - mmread. [Online]. <http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=8028&objectType=FILE>
- [104] P. Boersma and D. Weenink. (2008,) Praat: doing phonetics by computer (Version 5.0.32). [Computer program].
- [105] A. Kapoor and R. W. Picard, "Real-Time, Fully Automatic Upper Facial Feature Tracking," in *Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition (FGR 2002)*, Washington D.C., USA, 2002, pp. 10-15.
- [106] V. Vezhnevets, "Face and Facial Feature Tracking for Natural Human-Computer Interface," in *Proceedings of the 12th International Conference on Computer Graphics (GraphiCon 2002)*, Nizhny Novgorod, Russia, 2002.
- [107] P. Viola and M. Jones, "Robust real-time object detection," in *Proceedings of the ICCV 2nd International Workshop on Statistical and Computational Theories of Vision-Modelling, Learning, Computing, and Sampling*, Vancouver, Canada, 2001.
- [108] A. C. Hurlbert and T. A. Poggio, "Synthesizing a Color Algorithm from Examples," *Science*, vol. 239, pp. 482-485, Jan. 1988.
- [109] D. Datcu and L. J. M. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," in *Proceedings of the European Multidisciplinary Society for Modelling and Simulation Technology*, Porto, Portugal, 2005, pp. 58-65.
- [110] A. G. Chitu, L. J. M. Rothkrantz, and P. Wiggers, "Comparison between different feature extraction techniques for audio-visual," *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 7-20, Mar. 2007.
- [111] R. E. Schapire, "The Strength of Weak Learnability," *Machine Learning Journal*, vol. 5, no. 2, pp. 197-227, 1990.
- [112] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and Computation Journal*, vol. 121, no. 2, pp. 256-285, Sep. 1995.
- [113] Y. Freund and R. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *Journal of Computational System Sciences*, vol. 55, no. 1, p. 119-139, 1997.
- [114] P. Ekman and W. F. Friesen, *Facial Action Coding System*. Palo Alto, California, USA: Consulting Psychologists Press, Inc., 577 College Avenue, Palo Alto, California, 1978.







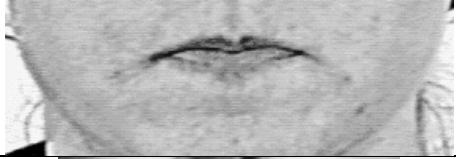


Appendices



A: List of Action Units

List of all implemented Action Units (number and name). List retrieved from <http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm>.

AU	Description	Point Activation	Example image
1	Inner Brow Raiser	4,5	
2	Outer Brow Raiser	6,7,8,9	
4	Brow Lowerer	4,5	
5	Upper Lid Raiser	10,11	
6	Cheek Raiser	14,15	
7	Lid Tightener	12,13	
9	Nose Wrinkler	14,15,16,17	
10	Upper Lip Raiser	22,23	
12	Lip Corner Puller	24,25	
15	Lip Corner Depressor	24,25	

List of Action Units

16	Lower Lip Depressor	26,27,30	
17	Chin Raiser	31	
18	Lip Puckerer	24,25,30	
20	Lip stretcher	24,25	
22	Lip Funneler	21,22,23,24, 25,26,27,30	
23	Lip Tightener	24,25	
24	Lip Pressor	21,22,23,26,27,30	
25	Lips part	30	
26	Jaw Drop	31	

27	Mouth Stretch	24,25,26,27, 28,29,30,31			
43	Eyes Closed	10,11			

B: Points to AUs Translation

```
% AU1 Inner Brow Raiser
AU1_1 = (neutralPoints(4,2)-neutralPoints(2,2))-(grid_struct.p4(2)-
grid_struct.p2(2));
AU1_2 = (neutralPoints(5,2)-neutralPoints(3,2))-(grid_struct.p5(2)-
grid_struct.p3(2));
if(AU1_1>0 && AU1_2>0)
    if(mean([AU1_1;AU1_2])/10 < 1)
        grid_struct.AU1 = mean([AU1_1;AU1_2])/10;
    else
        grid_struct.AU1 = 1;
    end
else
    grid_struct.AU1 = 0;
end

% AU2 Outer Brow Raiser
AU2L_1 = (neutralPoints(6,2)-neutralPoints(2,2))-(grid_struct.p6(2)-
grid_struct.p2(2));
AU2L_2 = (neutralPoints(8,2)-neutralPoints(3,2))-(grid_struct.p8(2)-
grid_struct.p3(2));
AU2R_1 = (neutralPoints(7,2)-neutralPoints(2,2))-(grid_struct.p7(2)-
grid_struct.p2(2));
AU2R_2 = (neutralPoints(9,2)-neutralPoints(3,2))-(grid_struct.p9(2)-
grid_struct.p3(2));
if(mean([AU2L_1;AU2L_2])>0 && mean([AU2R_1;AU2R_2])>0)
    if(mean([AU2L_1;AU2L_2;AU2R_1;AU2R_2])/7 < 1)
        grid_struct.AU2 = mean([AU2L_1;AU2L_2;AU2R_1;AU2R_2])/7;
    else
        grid_struct.AU2 = 1;
    end
else
    grid_struct.AU2 = 0;
end

% AU4 Brow Lowerer
AU4_1 = (neutralPoints(1,2)-neutralPoints(4,2))-(grid_struct.p1(2)-
grid_struct.p4(2));
AU4_2 = (neutralPoints(1,2)-neutralPoints(5,2))-(grid_struct.p1(2)-
grid_struct.p5(2));
AU4_3 = grid_struct.p6(2)-neutralPoints(6,2);
AU4_4 = grid_struct.p7(2)-neutralPoints(7,2);
if(AU4_1>0 && AU4_2>0 && AU4_3>0 && AU4_4>0)
    if(mean([AU4_1;AU4_2;AU4_3;AU4_4])/6 < 1)
        grid_struct.AU4 = mean([AU4_1;AU4_2])/6;
    else
        grid_struct.AU4 = 1;
    end
else
    grid_struct.AU4 = 0;
end

% AU5 Upper Lid Raiser
AU5_1 = neutralPoints(10,2)-grid_struct.p10(2);
AU5_2 = neutralPoints(11,2)-grid_struct.p11(2);
```

```
if(AU5_1>5 && AU5_2>5)
    grid_struct.AU5 = mean([AU5_1;AU5_2])/10;
else
    grid_struct.AU5 = 0;
end

% AU6 Cheek Raiser
AU6_1 = neutralPoints(14,2)-grid_struct.p14(2);
AU6_2 = neutralPoints(15,2)-grid_struct.p15(2);
if(AU6_1>5 && AU6_2>5)
    if(mean([AU6_1;AU6_2])>1)
        grid_struct.AU6 = 1;
    else
        grid_struct.AU6 = mean([AU6_1;AU6_2]);
    end
else
    grid_struct.AU6 = 0;
end

% AU7 Lid Tightener
AU7_1 = (grid_struct.p10(2)-neutralPoints(10,2))/25;
AU7_2 = (grid_struct.p11(2)-neutralPoints(11,2))/25;
if(mean([AU7_1;AU7_2])>0.6)
    grid_struct.AU7 = mean([AU7_1;AU7_2]);
else
    grid_struct.AU7 = 0;
end

% AU9 Nose Wrinkler
AU9L_1 = grid_struct.p14(2)-neutralPoints(14,2);
AU9L_2 = grid_struct.p16(2)-neutralPoints(16,2);
AU9R_1 = grid_struct.p15(2)-neutralPoints(15,2);
AU9R_2 = grid_struct.p17(2)-neutralPoints(17,2);
if(AU9L_1>5 && AU9L_2>5 && AU9R_1>5 && AU9R_2>5)
    grid_struct.AU9 = mean([AU9L_1;AU9L_2;AU9R_1;AU9R_2])/50;
else
    grid_struct.AU9 = 0;
end

% AU10 Upper Lip Raiser
AU10_1 = neutralPoints(22,2)-grid_struct.p22(2);
AU10_2 = neutralPoints(23,2)-grid_struct.p23(2);
if(AU10_1>5 && AU10_2>5)
    if((mean([AU10_1;AU10_2])/15)>1)
        grid_struct.AU10 = 1;
    else
        grid_struct.AU10 = mean([AU10_1;AU10_2])/15;
    end
else
    grid_struct.AU10 = 0;
end

% AU12 Lip Corner Puller
AU12_1 = neutralPoints(24,1)-grid_struct.p24(1);
AU12_2 = neutralPoints(24,2)-grid_struct.p24(2);
AU12_3 = neutralPoints(25,1)-grid_struct.p25(1);
AU12_4 = neutralPoints(25,2)-grid_struct.p25(2);
if(AU12_1>5 && AU12_2>5 && AU12_3>5 && AU12_4>5)
    if((mean([AU12_1;AU12_2;AU12_3;AU12_4])/10)>1)
```



```

    grid_struct.AU12 = 1;
else
    grid_struct.AU12 = mean([AU12_1;AU12_2;AU12_3;AU12_4])/10;
end
else
    grid_struct.AU12 = 0;
end

% AU15 Lip Corner Depressor
AU15_1 = grid_struct.p24(2)-grid_struct.p26(2)+5;
AU15_2 = grid_struct.p25(2)-grid_struct.p27(2)+5;
if(AU15_1>0 && AU15_2>0)
    if((mean([AU15_1;AU15_2])/7)>1)
        grid_struct.AU15 = 1;
    else
        grid_struct.AU15 = mean([AU15_1;AU15_2])/7;
    end
else
    grid_struct.AU15 = 0;
end

% AU16 Lower Lip Depressor
AU16_1 = grid_struct.p30(2)-neutralPoints(30,2);
AU16_2 = grid_struct.p26(2)-neutralPoints(26,2);
AU16_3 = grid_struct.p27(2)-neutralPoints(27,2);
if(mean([AU16_1;AU16_2;AU16_3])>3)
    if(mean([AU16_1;AU16_2;AU16_3])/3>1)
        grid_struct.AU16 = 1;
    else
        grid_struct.AU16 = mean([AU16_1;AU16_2;AU16_3])/3;
    end
else
    grid_struct.AU16 = 0;
end

% AU17 Chin Raiser
AU17 = neutralPoints(31,2)-grid_struct.p31(2);
if(AU17>10)
    if((AU17/20)>1)
        grid_struct.AU17 = 1;
    else
        grid_struct.AU17 = AU17/20;
    end
else
    grid_struct.AU17 = 0;
end

% AU18 Lip Puckerer
AU18_1 = grid_struct.p24(1)-neutralPoints(24,1);
AU18_2 = neutralPoints(25,1)-grid_struct.p25(1);
AU18_3 = grid_struct.p30(2)-neutralPoints(30,2);
if(AU18_1>3 && AU18_2>3 && AU18_3>3)
    grid_struct.AU18 = mean([AU18_1;AU18_2])/10;
else
    grid_struct.AU18 = 0;
end

% AU20 Lip Stretcher
AU20_1 = neutralPoints(24,1)-grid_struct.p24(1);

```

```

AU20_2 = grid_struct.p25(1)-neutralPoints(25,1);
if(AU20_1>10 && AU20_2>10)
    grid_struct.AU20 = mean([AU20_1;AU20_2])/15;
else
    grid_struct.AU20 = 0;
end

% AU22 Lip Funneler
AU22_1 = neutralPoints(24,1)-grid_struct.p24(1);
AU22_2 = grid_struct.p25(1)-neutralPoints(25,1);
AU22_3 = neutralPoints(21,2)-grid_struct.p21(2);
AU22_4 = grid_struct.p30(2)-neutralPoints(30,2);
if(AU22_1>0 && AU22_2>0 && AU22_3>0 && AU22_4>0 && AU22_4<15)
    grid_struct.AU22 = mean([AU22_1;AU22_2])/5;
else
    grid_struct.AU22 = 0;
end

% AU23 Lip Tightener
AU23_1 = grid_struct.p24(1)-neutralPoints(24,1);
AU23_2 = neutralPoints(25,1)-grid_struct.p25(1);
if(AU23_1>5 && AU23_2>5)
    grid_struct.AU23 = mean([AU23_1;AU23_2])/10;
else
    grid_struct.AU23 = 0;
end

% AU24 Lip Funneler
AU24_1 = neutralPoints(24,1)-grid_struct.p24(1);
AU24_2 = grid_struct.p25(1)-neutralPoints(25,1);
AU24_3 = neutralPoints(21,2)-grid_struct.p21(2);
AU24_4 = grid_struct.p30(2)-neutralPoints(30,2);
if(AU24_1>0 && AU24_2>0 && AU24_3>0 && AU24_4>0 && AU24_4<15)
    grid_struct.AU24 = mean([AU24_1;AU24_2;AU24_3;AU24_4])/5;
else
    grid_struct.AU24 = 0;
end

% AU25 Lips Part
AU25 = (grid_struct.p30(2)-grid_struct.p21(2))-(neutralPoints(30,2)-
neutralPoints(21,2));
if(AU25>5)
    grid_struct.AU25 = AU25/25;
else
    grid_struct.AU25 = 0;
end

% AU26 Chin Drop
AU26 = grid_struct.p31(2)-neutralPoints(31,2);
if(AU26>10)
    grid_struct.AU26 = AU26/30;
else
    grid_struct.AU26 = 0;
end

% AU27 Mouth Stretch
AU27_1 = grid_struct.p7(2)-neutralPoints(7,2);
AU27_2 = grid_struct.p8(2)-neutralPoints(8,2);
AU27_3 = grid_struct.p26(2)-neutralPoints(26,2);

```

```
AU27_4 = grid_struct.p27(2)-neutralPoints(27,2);
AU27_5 = grid_struct.p28(2)-neutralPoints(28,2);
AU27_6 = grid_struct.p29(2)-neutralPoints(29,2);
AU27_7 = grid_struct.p30(2)-neutralPoints(30,2);
AU27_8 = grid_struct.p31(2)-neutralPoints(31,2);
if(AU27_1>0 && AU27_2>0 && AU27_3>0 && AU27_4>0 && ...
    AU27_5>0 && AU27_6>0 && AU27_7>0 && AU27_8>0)

if((mean([AU27_1;AU27_2;AU27_3;AU27_4;AU27_5;AU27_6;AU27_7;AU27_8])/30)>1)
    grid_struct.AU27 = 1;
    else
        grid_struct.AU27 =
mean([AU27_1;AU27_2;AU27_3;AU27_4;AU27_5;AU27_6;AU27_7;AU27_8])/30;
    end
else
    grid_struct.AU27 = 0;
end

% AU43 Eyes Closed
distance = mean([(neutralPoints(12,2)-neutralPoints(10,2)), ...
    (neutralPoints(13,2)-neutralPoints(11,2))])-10;
curr_distance = mean([(grid_struct.p12(2)-grid_struct.p10(2)), ...
    (grid_struct.p13(2)-grid_struct.p11(2))]);
if(1/(curr_distance/distance) >= 1)
    grid_struct.AU43 = 1;
elseif(1/(curr_distance/distance) <= 0)
    grid_struct.AU43 = 0;
else
    grid_struct.AU43 = 1/(curr_distance/distance);
end
```


C: AU Constraints for AU Activation

```
% look which AU should be activated, according to the rules made by Anna
% Wojdel in her thesis "KNOWLEDGE DRIVEN FACIAL MODELLING".
% AU1
if(grid_struct.AU1 > grid_struct.AU4 && grid_struct.AU1 < grid_struct.AU9
|| (grid_struct.AU1>0 && grid_struct.AU2>0))
    grid_struct.AU1 = grid_struct.AU1;
else
    grid_struct.AU1 = 0;
end

% AU4
if(grid_struct.AU4 > grid_struct.AU1)
    grid_struct.AU4 = grid_struct.AU4;
else
    grid_struct.AU4 = 0;
end

% AU5
if(eliminateAU(grid_struct.AU5,grid_struct.AU43)>0)
    grid_struct.AU5 = eliminateAU(grid_struct.AU5,grid_struct.AU43);
else
    grid_struct.AU5 = 0;
end

% AU7
if(grid_struct.AU6 > grid_struct.AU7 && grid_struct.AU7 < grid_struct.AU9)
    grid_struct.AU7 = grid_struct.AU7;
else
    grid_struct.AU7 = 0;
end

% AU10
if(grid_struct.AU9 > grid_struct.AU10 && grid_struct.AU10 <
grid_struct.AU12)
    grid_struct.AU10 = grid_struct.AU10;
else
    grid_struct.AU10 = 0;
end

% AU12
if(grid_struct.AU12 < grid_struct.AU15 && grid_struct.AU12 <
grid_struct.AU6+grid_struct.AU15)
    grid_struct.AU12 = grid_struct.AU12;
else
    grid_struct.AU12 = 0;
end

% AU15
if(grid_struct.AU12 < grid_struct.AU15 && ...
    grid_struct.AU15+grid_struct.AU24 == grid_struct.AU15+grid_struct.AU17)
    grid_struct.AU15 = grid_struct.AU15;
else
    grid_struct.AU15 = 0;
end
```

```

% AU16
if(grid_struct.AU16 < grid_struct.AU12+grid_struct.AU18 && ...
    grid_struct.AU16 < grid_struct.AU12+grid_struct.AU20 && ...
    grid_struct.AU16 < grid_struct.AU12+grid_struct.AU22 && ...
    grid_struct.AU16 < grid_struct.AU12+grid_struct.AU23 && ...
    grid_struct.AU16 < grid_struct.AU18+grid_struct.AU20 && ...
    grid_struct.AU16 < grid_struct.AU18+grid_struct.AU23 && ...
    grid_struct.AU16 < grid_struct.AU20+grid_struct.AU22 && ...
    grid_struct.AU16 < grid_struct.AU20+grid_struct.AU23 && ...
    grid_struct.AU16 < grid_struct.AU22+grid_struct.AU23)
    grid_struct.AU16 = grid_struct.AU16;
else
    grid_struct.AU16 = 0;
end

% AU18
if(grid_struct.AU18 < grid_struct.AU20+grid_struct.AU23)
    grid_struct.AU18 = grid_struct.AU18;
else
    grid_struct.AU18 = 0;
end

% AU20
if(eliminateAU(grid_struct.AU20,grid_struct.AU12)>0 && ...
    eliminateAU(grid_struct.AU20,grid_struct.AU15)>0 && ...
    eliminateAU(grid_struct.AU20,grid_struct.AU18)>0)
    grid_struct.AU20 =
min([eliminateAU(grid_struct.AU20,grid_struct.AU12);eliminateAU(grid_struct
.AU20,grid_struct.AU15);eliminateAU(grid_struct.AU20,grid_struct.AU18)]);
else
    grid_struct.AU20 = 0;
end

% AU23
if(grid_struct.AU23 < grid_struct.AU18 && ...
    grid_struct.AU20 < grid_struct.AU28)
    grid_struct.AU23 = grid_struct.AU23;
else
    grid_struct.AU23 = 0;
end

% AU24
if(grid_struct.AU24 < grid_struct.AU9+grid_struct.AU17 && ...
    grid_struct.AU24 < grid_struct.AU10+grid_struct.AU17 && ...
    grid_struct.AU24 < grid_struct.AU12+grid_struct.AU17 && ...
    grid_struct.AU24 < grid_struct.AU17+grid_struct.AU22 && ...
    grid_struct.AU24 < grid_struct.AU17+grid_struct.AU23 && ...
    grid_struct.AU24 < grid_struct.AU18 && ...
    grid_struct.AU24 < grid_struct.AU20 && ...
    grid_struct.AU24 < grid_struct.AU23)
    grid_struct.AU24 = grid_struct.AU24;
else
    grid_struct.AU24 = 0;
end

% AU43
if(eliminateAU(grid_struct.AU43,grid_struct.AU5)>0)
    grid_struct.AU43 = eliminateAU(grid_struct.AU43,grid_struct.AU5);

```

```
else
    grid_struct.AU43 = 0;
end

% custom elimination function for AU's that have opposite movements and
% thus fase eachother out
function AU = eliminateAU(AU1,AU2)
if(AU1-AU2<0)
    AU = 0;
else
    AU = AU1-AU2;
end
```


D: Participant's Instructions

Instructies voor de testpersoon

Welkom bij ons experiment. Luister of lees aandachtig het korte scenario door en stel je voor dat jij in deze situatie beland bent. Als je klaar bent willen wij graag dat je vijf maal de gevraagde emotie uit, elke keer door middel van een andere zin. De gevraagde zinnen komen in beeld op het scherm voor je.

Ga rustig zitten in de stoel die voor de camera staat. Ontspan je en zorg dat jij je comfortabel voelt.

Luister of lees aandachtig het korte scenario door om in de gewenste stemming te komen. Mocht het nodig zijn mag je altijd het scenario inzien of nogmaals lezen. Stel je vervolgens voor dat jij in deze situatie zit. Het herinneren van een situatie met de gevraagde emotie kan helpen.

Er worden per emotie vijf verschillende zinnen gebruikt en je moet deze alle uitten. Als je goed in de stemming bent voor de gevraagde emotie kan je de vijf verschillende zinnen, met een korte tussenpauze, achter elkaar uitten. Er wordt niet gewisseld tussen verschillende emoties, als je eenmaal in de stemming bent wordt de gevraagde emotie geheel afgerond. Hierdoor hoef jij je maar een keer in te leven in de gevraagde emotie.

Je wordt verzocht de getoonde zin te uitten in de gevraagde emotie. Mocht je niet weten hoe je de emotie moet uitten dan kan je altijd de deskundige om hulp vragen.

Als je klaar bent willen we graag dat je een vragenlijst invult met informatie over jezelf. Deze informatie wordt strikt geheim gehouden en niet aan derden verstrekt.

Instructions for the participant

Welcome to our experiment. We ask you to listen to or read the scenario carefully and imagine being in this situation. When you are ready you are asked to express the desired emotion five times with a different sentence each time. The sentences are shown on the screen in front of you.

Please take a seat in the chair in front of the camera. Relax and make sure you are at ease.

Listen to or read the short scenario, just to come in the desired mood for the expression. You are allowed to read the scenario again if you like. Imagine yourself to be in this situation. Recalling an experience with the desired emotion from the past can help you do this.

You are asked to express five different sentences for every emotion. Once you are comfortable with the emotion you may pronounce all of the shown sentences, we do not switch between different emotions. Once you have successfully expressed the five sentences we move on to the next emotion. This way you only have to be in one emotion once.

You are asked to express the desired emotion while pronouncing the displayed sentence. If you do not know how to express the desired emotion you can ask the experimenter for help.

Afterwards you are asked to fill in some questions about yourself, these answers are kept confidential at all times.

E: Text Used for Recordings

Admiration

“Je loopt samen met een vriend/vriendin door een dure winkelstraat in Amsterdam en ziet in de etalage een jas hangen die je altijd al had willen hebben. Je droomt over wat je zou doen als je het geld had om deze jas te kopen. Je gaat voor de etalage staan en denkt...”

Reacties

R1: Oooohhh..	o
R2: Dat ziet er goed uit.	dat zit @r Gut Y+t
R3: Die zou ik graag willen hebben.	di zA+ Ik GraG wIl@ hEb@
R4: Was die maar van mij.	wAs di mar van mE+
R5: Zodra ik mijn geld heb, is die jas van mij.	zodra Ik mE+n (m@n) GElt hEp Is ti jAs fAn mE+

Amusement

“Jij loopt met een vriendin op straat langs een kiosk. Opeens zie jij de foto van je favoriete zanger op de cover van een tijdschrift. Onder de foto staat dat hij binnenkort in een film van een beroemde regisseur de hoofdrol zal spelen. Jij vindt het werk van die regisseur ook heel goed. O, wat wil je die film graag zien.”

Reacties

R1: Dat zal een goeie film worden.	dAt zal en Guj@ film wOrd@
R2: Die film wil ik zeker zien.	di film wIl Ik zek@r zin
R3: Hij maakt altijd goede films.	hE+ makt AltE+t Gud@ films
R4: Wanneer gaat deze film in première?	wAner Gat dez@ flIm In pr@miE:r@
R5: Dat wordt kijken geblazen!	dAt wOrt kE+k@ G@blaz@

Anger (Surprise)

“Je zit thuis op de bank lekker televisie te kijken als er plots wordt aangebeld. Je loopt naar de deur en doet open. Voor je staat een vreemde vrouw die nogal nerveus over komt. Ineens barst ze in tranen uit en vertelt ze dat zij tijdens het parkeren jouw auto heeft aangereden! Je dacht dat de auto veilig in de garage stond, maar langzaam herinner je dat je vanmiddag nog snel even bent wezen winkelen en de auto voor je huis hebt laten staan.”

Reacties

R1: Wat??? Nee, nee dat kan niet!	wAt ne ne dAt kAnit
R2: Dat gaat je geld kosten!	dAt Gat j@ GElt kOst@
R3: Weet je zeker dat het mijn auto is!?	wet j@ zek@r dAt @t mE+n A+to Is
R4: Zie je wel, vrouwen kunnen niet rijden!	zi j@ wEl vrA+Yn k@n@ nit rE+d@
R5: Niet mijn auto? Het is niet waar!	nit mE+n A+to hEt Is nit war

Anger

“Je vriend(in) neemt je mee naar een duur restaurant om jullie 5 jarig samen zijn te vieren. Samen met je vriend(in) genieten jullie van de avond en het restaurant. Nog geen idee wat jullie zullen bestellen komt de ober vragen of jullie vast een aperitiefje willen. Je stemt in en laat vast een stokbrood met kruidenboter komen en twee biertjes. Het brood is fantastisch en zulke lekkere kruidenboter heb je nog nooit geproefd, je kan niet wachten tot de soep er is. Als de soep eenmaal arriveert laat de ober per ongeluk de gloeiend hete soep over je heen vallen. Je hele avond is verpest en je humeur wordt zeker niet beter.”

Reacties

R1: AUW! Klootzak!	A+ klotsAk
R2: Pas op met die hete soep!	pAs op mEt di het@ sup
R3: Hier kom ik dus nooit meer!	hir kOm Ik d@s nojt mer
R4: Nee! Mijn nieuwe broek verpest!	ne mE+n niw@ bruk v@rpEst
R5: Kom, we gaan!	kOm w@ gan

Boredom

“Je zit alleen op de bank op je vrije zaterdagmiddag. Vroeger ging je nog wel een naar het sportveld, maar tegenwoordig niet meer. Je hebt eigenlijk weinig te doen en je verveelt je een beetje. Je moeder komt binnen en vraagt ‘Verveel je je?’. Waarop jij antwoord:”

Reacties

R1: Mwah...	mwAh
R2: Laat me maar...	lat m@ mar
R3: Huh?	h@
R4: Hmmmmmm. Kweenie.	hm kweni
R5: Neuh..	n@

Contempt

“Van een goede vriend hoor je dat je vervelende buurman, die altijd loopt te roddelen over je, van de trap gevallen is. Hij heeft zijn neus gebroken.”

Reacties

R1: Veel slechter kan het niet worden.	vel sleGt@r kAn @t ni wOrd@
R2: O, het is Jaap maar.	o hEt Is jap mar
R3: O die, ja dat dacht ik wel.	o di ja dAt dAGt Ik wEl
R4: Met zijn neus recht in de boter!	mEt z@n_n2s In d@ bot@r
R5: Net goed.	nEt Gut

Desire

“Je loopt langs een elektronica winkel en ziet in de etalage een prachtige 50inch breedbeeld LCD televisie staan. Zo een heb je altijd al willen hebben, maar hij is natuurlijk veel te duur voor je. Je blijft stil staan en kijkt er nog eens goed naar.”

Reacties

R1: Aaaahhh..	a
R2: Dat ziet er goed uit!	dAt zit@r Gut Y+t
R3: Die zou ik graag willen hebben.	di zA+ Ik GraG wIl@ hEb@
R4: Had ik er ook maar een.	hAt Ik Er ok mar en
R5: Wóóów.	wow

Disappointment

“Je laatste eindexamen was zeker twee weken geleden en de cijfers moeten reeds bekend zijn. Vanmiddag zal je worden gebeld en iemand zal je vertellen wat de uitslag is en dan kan je gaan genieten van je zomervakantie. Je hebt zo hard je best gedaan, maar je weet niet zeker of je geslaagd bent. Als de telefoon over gaat ben je als eerste bij de telefoon om hem op te nemen. Je hoort een stem aan de andere kant zeggen dat je gezakt bent.”

Reacties

R1: Ik ben gezakt.	Ik bEn G@zAkt
R2: Jammer, volgende keer beter.	jAm@r vOIG@nd@ ker bet@r
R3: Dat had ik niet verwacht.	Dat hAt Ik nit v@rwAGt
R4: En wat moet ik nu?	En wAt mut Ik ny
R5: Nee hè!	ne hE

Disgust

“Vrolijk loop je over straat op weg naar je nieuwe baan. In de verte zie je een vage bekende. Je kijkt nog eens goed en probeert te herinneren wie het is. Als je even niet op let stap je in een grote, vers gelegde, hoop poep. Je goede schoenen zitten er volledig onder, zo kan je toch niet verschijnen op je sollicitatiegesprek. Je baalt en kijkt naar de viezigheid beneden.”

Reacties

R1: Ieeuuww, poep.	we@w pup
R2: Wat is dat vies.	wAt Is dAt vis
R3: Hè, bah!	hE bA
R4: Niet nu.	nit ny
R5: Gatver, mijn schoenen vies.	GAtf@r mE+n sGun@ vis

Dislike

“Je moeder heeft voor je verjaardag een, volgens haar, leuk cadeau gekocht. Ze neemt je mee naar buiten en laat een gloed nieuwe tent zien. Je denk ‘ze weet toch dat ik niet van kamperen hou.’ En kijkt haar een beetje teleurgesteld aan.”

Reacties

R1: Een groene?!	@n Grun@
R2: Je weet toch dat ik niet van kamperen hou.	j@ wet tOG dAt Ik nit vAn kAmper@ hA+
R3: Wat moet ik hier nu weer mee?	wAt mut Ik hir ny wer me
R4: Dit vind ik toch niet leuk.	dIt vInt Ik tOG nit l2k
R5: Ik vind hem niet eens mooi!	Ik vInt hem nit ens moj

Dissatisfaction

“Die mooie jas die je laatst had gekocht is stuk gegaan en je gaat er mee terug naar de winkel. Je wilt je geld terug, maar de mevrouw achter de kassa zegt dat je wel een tegoed bon kan krijgen, maar niet je geld terug. Je bent het er eigenlijk niet mee eens, maar accepteert de bon toch.

Reacties

R1: Dan neem ik deze maar.	dAn nem Ik dez@ mar
R2: Ach, het is beter dan niets	AG hEt Is bet@r dAn nits
R3: Toch liever mijn geld gehad.	tOG liv@r m@n GElt G@hAt
R4: Jammer, toch bedankt.	jAm@r tOG b@dAnkt
R5: Ik snap niet waarom ik mijn geld niet terug krijg	Ik snap nit warOm Ik mE+n GElt nit_tr@G krE+G

Fascination

“Je kijkt naar een documentaire over het heelal. Je merkt hoe onvoorstelbaar groot het heelal is en hoe klein onze kleine aarde. Na de afloop van het programma denk jij...”

Reacties

R1: Wow wat een verhaal.	wA+w wAt @n v@rhal
R2: Dat wist ik allemaal niet.	dAt wist Ik Al@mal nit
R3: Het was echt een interessant verhaal.	het was EGt en Int@r@sAnt vErhal
R4: Het is echt indrukwekkend.	het Is EGt Indr@kwEk@nt
R5: ohh, fascinerend!	o fasiner@nt

Fear

“Je bent allen thuis en ligt in bed. Je slaapkamer is boven in het huis en jij bent de enige persoon die thuis is. Plotseling hoor je een geluid beneden. Je weet zeker dat er niemand zou thuiskomen vanacht. Je houdt je stil en weet zeker dat er iemand beneden is. Waarschijnlijk een dief, of zelfs een moordenaar! Je hoort hem de trap op lopen en je wordt heel erg bang.”

Reacties

R1: Oh mijn God, er is iemand binnen!	o mE+n GOt Er Is imAnt bIn@
R2: Er komt iemand naar boven.	Er kOmt imAnt nar bov@
R3: Vermoordt me alsjeblijft niet...	vErmort m@ AIS@blift nit
R4: Help!	hElp
R5: Alsjeblijft, doe me niks!	AIS@blift du m@ nIks

Furious

“Je loopt in een onbekende stad en je hebt €200 nodig om terug naar huis te komen. Er is maar één bank in deze stad. Je moet het geld echt vandaag nog hebben. De pinautomaat buiten is buiten gebruik en je kent verder niemand hier. De bank is tot vijf uur open en je stapt om kwart over vier binnen. Je ziet een enorme rij voor de balie staan en gaat achteraan staan. Na drie kwartier wachten als je eindelijk vooraan staat, de bankbediende vraagt of je morgen terug kunt komen, omdat hij nu zijn koffiepauze gaat houden voordat hij de bank sluit en naar huis gaat. Je legt je situatie uit, de bank moet nog zeker 15 minuten open zijn en je hebt het geld echt vandaag nodig, maar de man wil er niets van weten. Hij blijft maar herhalen dat het niet zijn probleem is en nu een koffiepauze gaat houden. Je bent woedend en zal niet thuis komen.”

Reacties

R1: Wat??? Nee, nee, nee, luister! Ik moet dit geld vandaag hebben!

R2: Die koffie kan me niets schelen, help mij liever!

R3: Ik wil je baas spreken, NU!

R4: Is je koffie belangrijker dan mij helpen?

R5: Je krijgt betaalt om te werken, niet om koffie te drinken!

wAt ne ne lY+st@r Ik mut dIt GElt vAndaG heb@

di kofi kAn m@ nit sGel@ hElp mE+ liv@r

Ik wIl j@ bas_sprek@ ny

Is j@ kOfi b@lANrE+k@r dAn mE+ hElp@

j@ krE+Gt b@talt Om t@ wErk@ nit Om kOfi t@ drINk@

Happiness

“Je gaat een avond stappen in Holland Casino en besluit om vlak voordat je naar huis gaat nog even je laatste geld in de fruitautomaat te werpen. Vol verwachting blijf je wachten als het eerste kroontje verschijnt. Drie kroontjes is de jackpot, en de tweede valt. Je staat bijna op springen als je gestaag op het derde wiel wacht. Ja! Ook een kroon, je hebt de jackpot gewonnen en bent hartstikke blij.”

Reacties

R1: JAAAAAHHH!!!

R2: Gewonnen!

R3: Nu kan ik lekker op vakantie!

R4: O, dankjewel!

R5: Wat heerlijk.

ja

G@wOn@

ny kAn Ik lEk@r Op vakAntsi

o dANkj@wEl

wAt herl@k

Indignation

“Je loopt over straat naar een vriend. Het is best druk en je moet goed opletten, anders loop je nog tegen iemand aan. Als je telefoon gaat haal je deze uit je zak en let je even niet op. Je loopt tegen iemand aan die gestrekt op de grond valt. Als de man overeind komt en direct begint te schreeuwen weet je eerst niet hoe je moet reageren. Hij scheldt je uit voor van alles en nog wat.”

Reacties

R1: Euhh, sorry.

R2: Pardon, ik lette niet op.

R3: Rustig maar, u heeft niets gebroken hoor.

R4: Nou zeg, ik deed het niet expres.

R5: Het was niet mijn bedoeling om u omver te lopen.

@ sOri

pArdOn Ik lEt@ nit Op

r@stIG mar y heft nits G@brok@ hor

nA+ zEG Ik det hEt nit EksprEs

hEt wAs nit mE+n b@dullIN Om y OmvEr

t@ lop@

Interest

“Op een feestje van je een vriendin, sta je met iemand te praten. Je kent hem niet zo goed, maar hij houdt een mooi verhaal. Hij vertelt jou dat hij zweefvliegles geeft. Je wilde altijd al zweefvlieglessen nemen. Hij heeft een oranje lestoestel. Je denkt”

Reacties

R1: Ohh, wat leuk een oranje.	O wAt l2k en orAJ@
R2: Ik wilde altijd al zweefvliegles nemen.	Ik wIld@ AltE+t Al zwefvliGIEs nem@
R3: Zweefvliegen wilde ik altijd nog eens doen.	zwefvliG@ wIld@ Ik AltE+t nOG ens dun
R4: Geeft jij echt zweefvliegles?	Gef jE+ EGt zwefvliGIEs
R5: Kan je het mij leren?	kan j@ het mE+ ler@

Pleasant surprise

“Als je ‘s middags thuis aan het stofzuigen bent wordt er aangebeld. Je verwacht geen bezoek en hebt ook niemand aan horen komen. Als je open doet zie je daar Ron Brandsteder jr. Staan met een cameraploeg voor je deur. ‘Gefeliciteerd’ zegt hij, ‘U heeft 100.00 euro gewonnen’. Met stomheid geslagen sta je in de deuropening en kijkt hem aan.”

Reacties

R1: Heb ik gewonnen?	hEb Ik G@wOn@
R2: Voor mij?	vor mE+
R3: Eindelijk heb ik geluk!	E+nd@l@k hEb Ik G@l@k
R4: Héél erg bedankt!	hel ErG b@dANkt
R5: Wahoo, Dat had ik nooit gedacht!	wahu dAt hAt Ik nOjt G@dAGt

Sadness

“Je komt thuis na een dag hard werken en ploft op de bank. Je wilt lekker ontspannen als plotseling de telefoon gaat. Het is je moeder, en ze klink nogal erg verontrustend. Ze verteld je dat je vader is overleden. Eerst wil je het niet geloven, maar na een tijdje dringt het tot je door wat dit voor je betekend. Als je terug denkt aan alle fijne momenten samen en dan bedenkt dat alles voortaan anders is voel je je erg verdrietig.”

Reacties

R1: Het is niet waar, alsjeblieft.	hEt Is nit war AIS@blift
R2: NEEEEEE!	ne
R3: Ik snap het niet, hij was nog zo jong!	Ik snAp hEt nit hE+ wAs nOG zo jON
R4: Dit kan niet waar zijn.	dIt kAn_nit war zE+n
R5: Wat erg, wat moet ik nu.	wAt ErG wAt mut Ik ny

Satisfaction

“Op een warme zomerdag zit je op een terrasje en drink je een koud biertje. Terwijl je een slokje neemt geeft dit je een bevredigend gevoel. Wat is het leven toch goed.”

Reacties

R1: Dit is echt lekker bier.	dit Is EGt lek@r bir
R2: Wat een perfecte zomerdag.	wAt @n pErfEkt@ zom@rdAG
R3: Dit voelt echt goed.	dit vult EGt gut
R4: Ik kan hier geen genoeg van krijgen.	Ik kAn hir Gen G@nuG vAn krE+g@
R5: Wat wil je nog meer?	wAt wIl j@ nOG mer

Unpleasant surprise

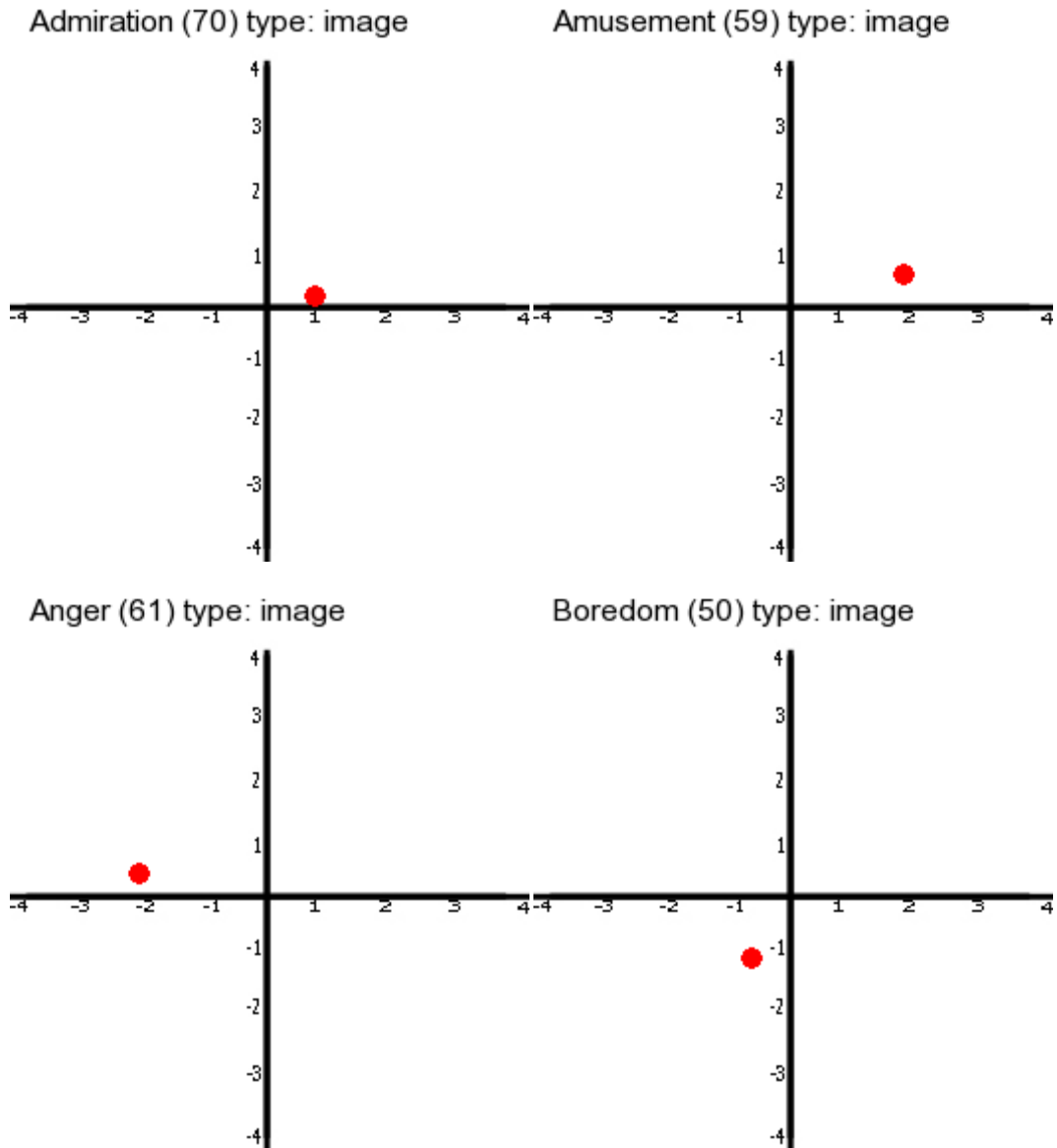
“Je partner zegt tegen je dat jullie eens serieus moeten praten over jullie relatie. Als jullie samen op de bank zitten zegt hij uit het niets tegen je dat hij homoseksueel is. Je bent erg verbaasd en had dit zeker niet verwacht.”

Reacties

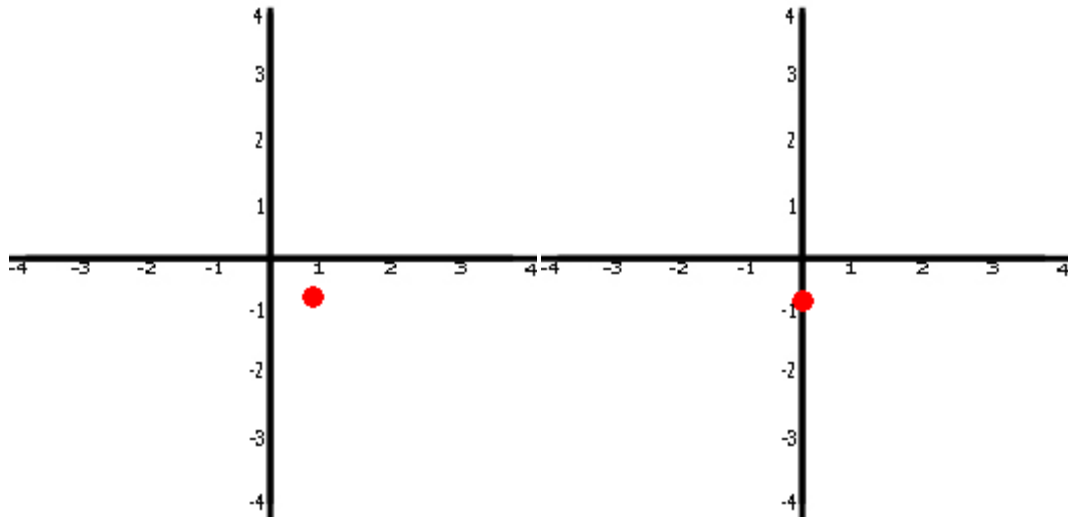
R1: Wat zeg je me nu!	wAt zEG j@ m@ ny
R2: Dat had ik niet verwacht!	dAt hAt Ik nit v@rwAGt
R3: Dit geloof je toch niet!	dIt g@lof j@ tOG nit
R4: Dat meen je niet!	dAt men j@ nit
R5: O mijn God, het is niet waar!	o mE+n GOt hEt Is nit war

F: Image Validation Results

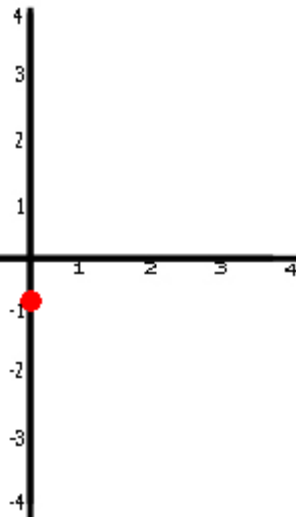
These are the results for the user validation of the 21 different images showed on the website. The number between the brackets indicates the number of validations for this particular item.



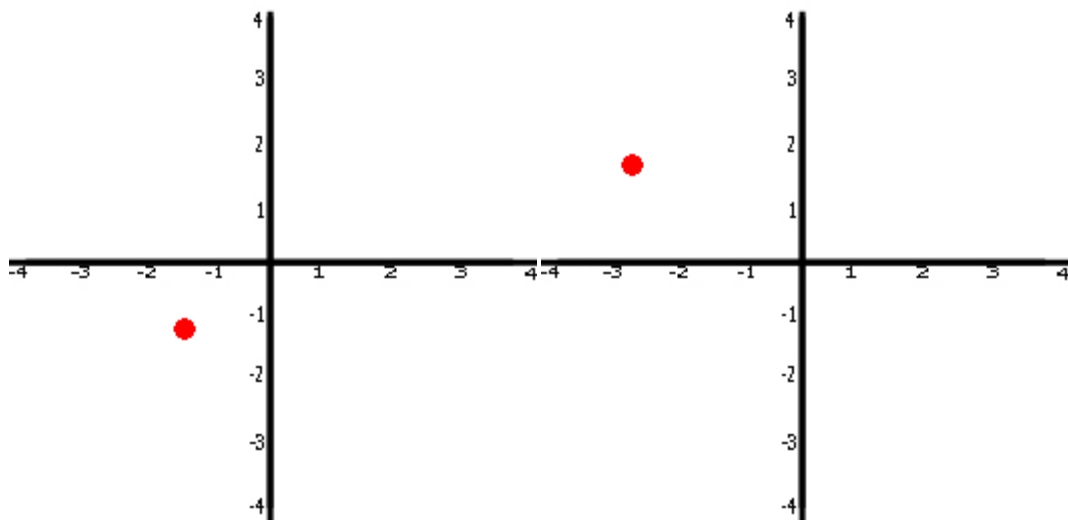
Contempt (71) type: image



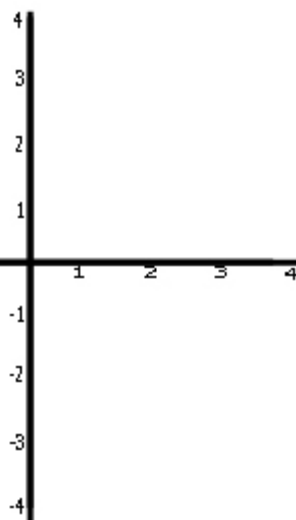
Desire (52) type: image



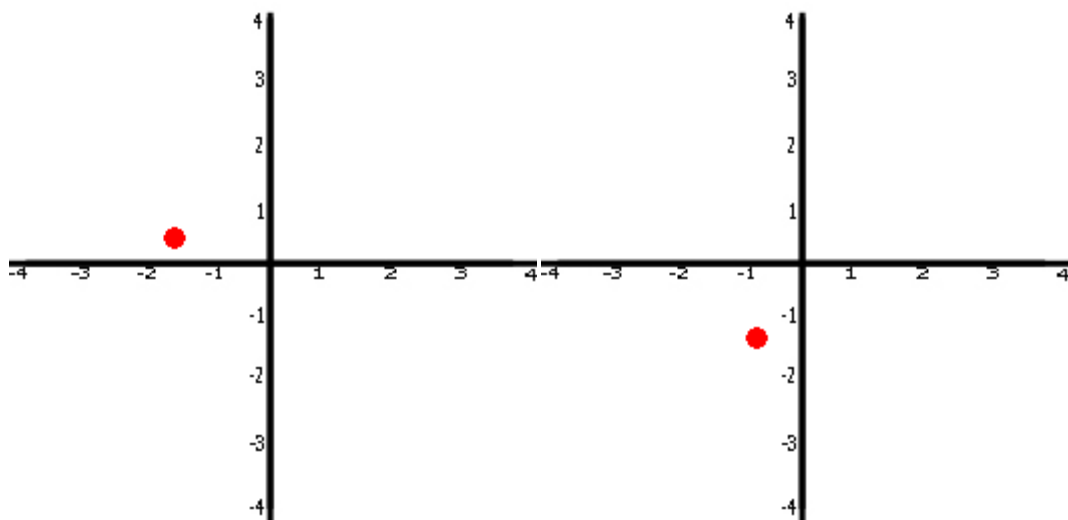
Disappointment (50) type: image



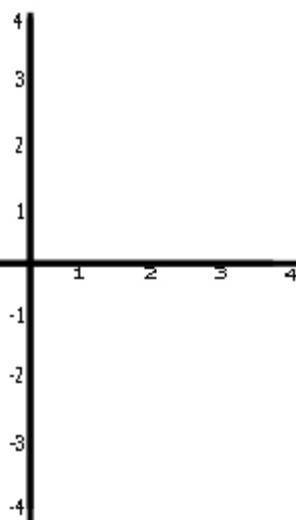
Disgust (70) type: image



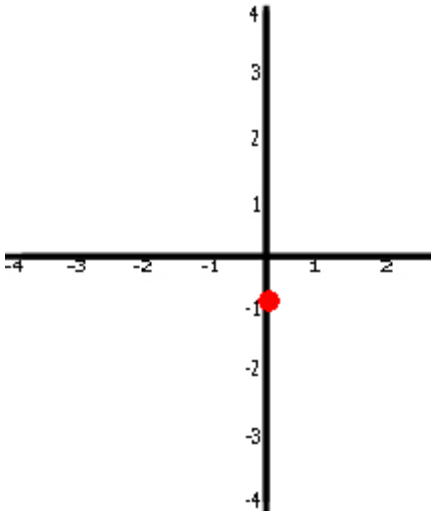
Dislike (61) type: image



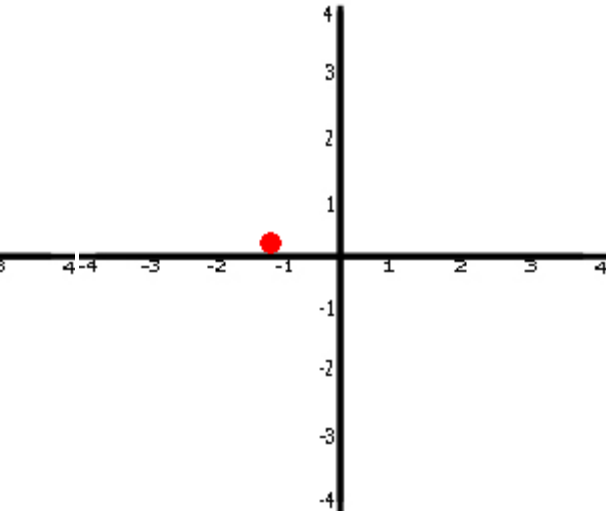
Dissatisfaction (57) type: image



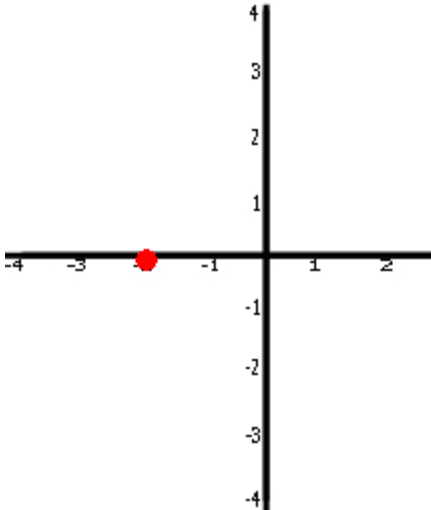
Fascination (61) type: image



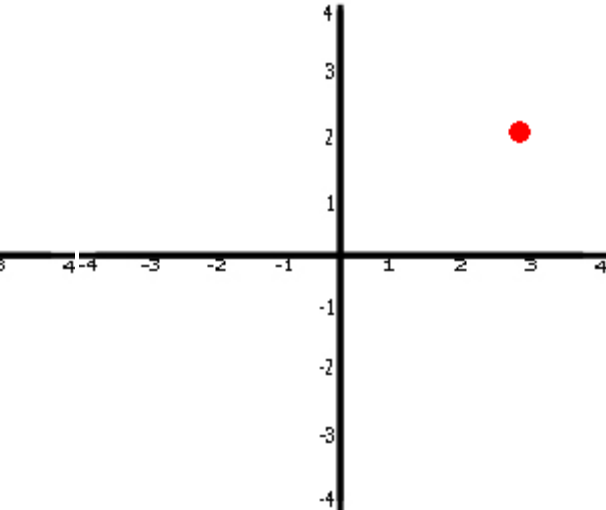
Fear (60) type: image



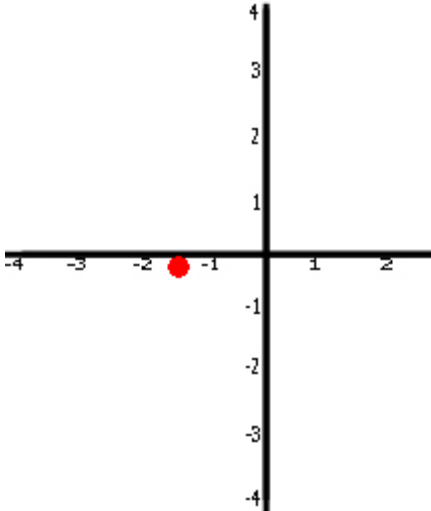
Furious (62) type: image



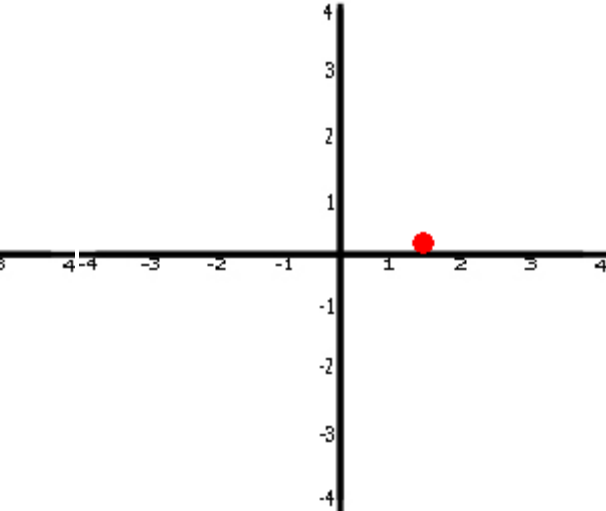
Happiness (78) type: image



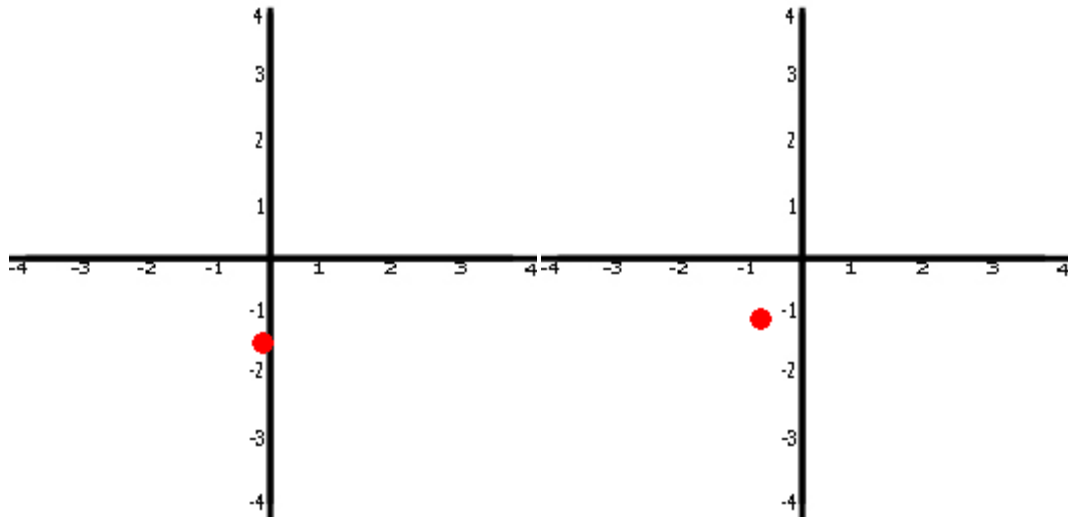
Indignation (52) type: image



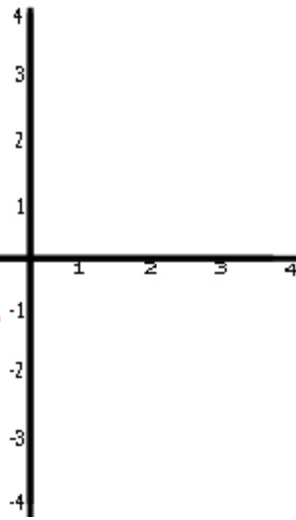
Interest (56) type: image



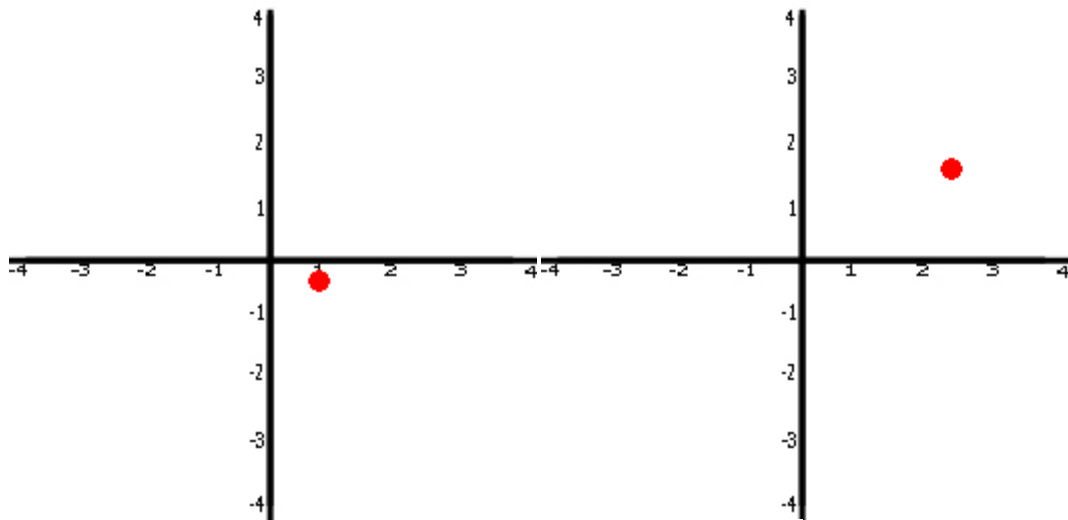
Neutral (73) type: image



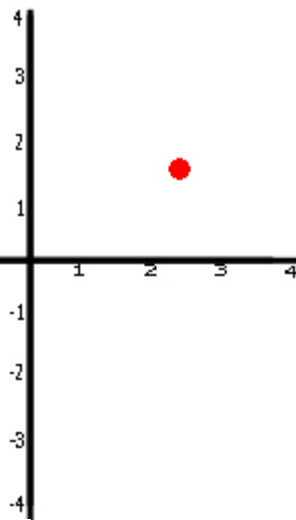
Sadness (56) type: image



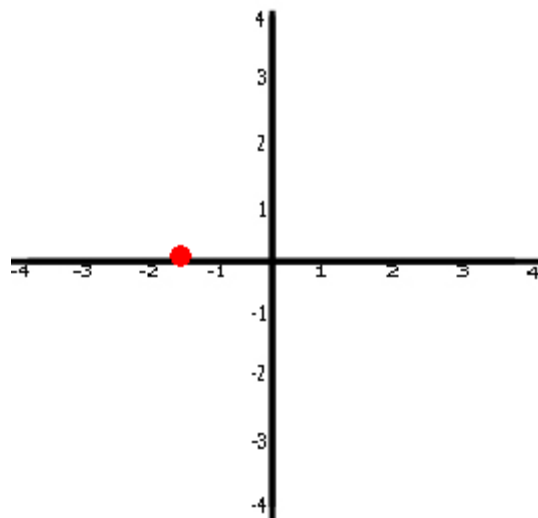
Satisfaction (67) type: image



Surprise_(Pleasant) (46) type: image

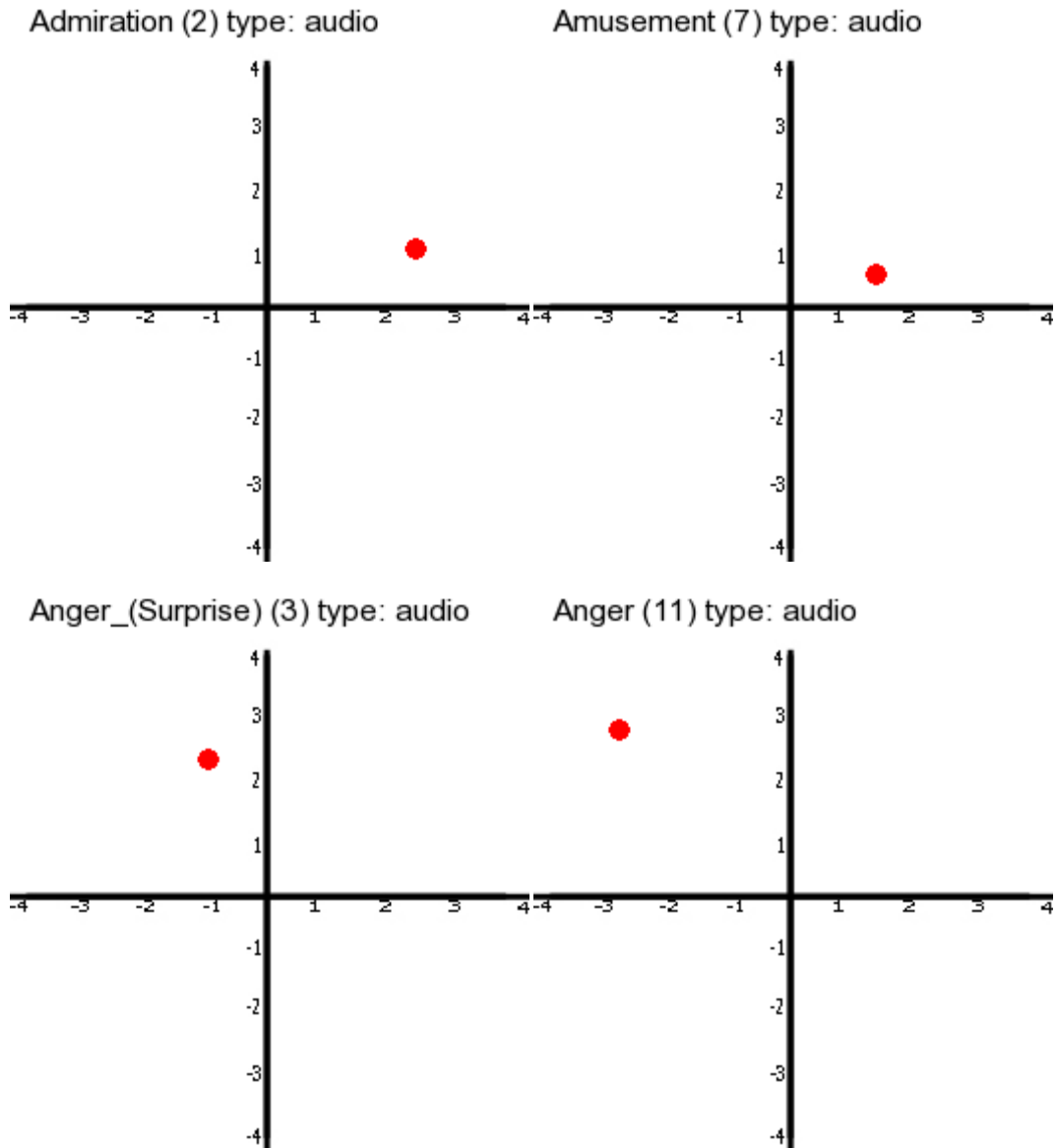


Surprise_(Unpleasant) (56) type: image

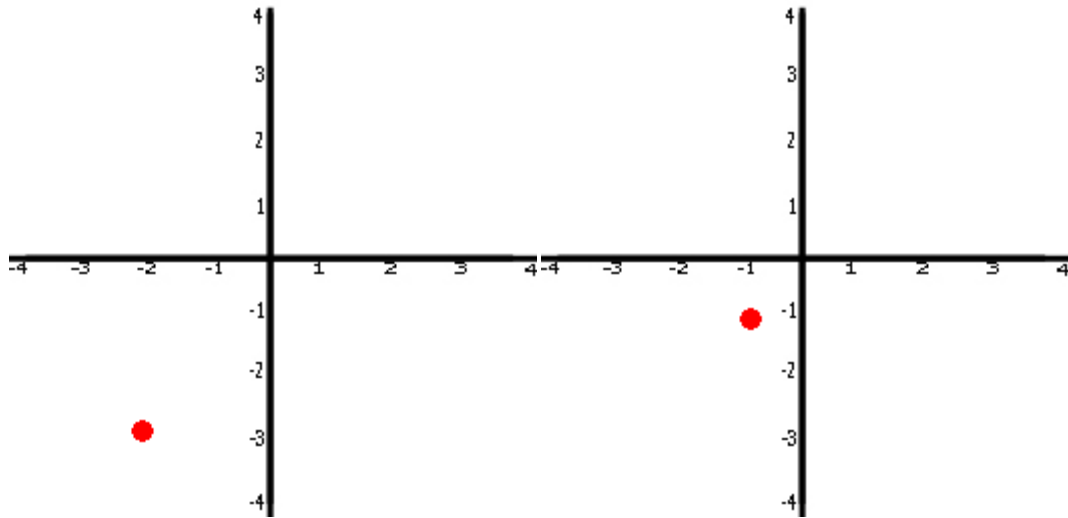


G: Audio Validation Results

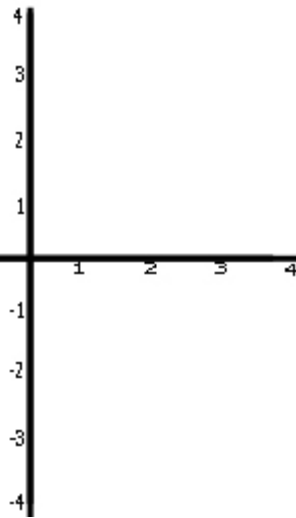
These are the results for the user validation of the 105 different audio clips played on the website. The number between the brackets indicates the number of validations for this particular item.



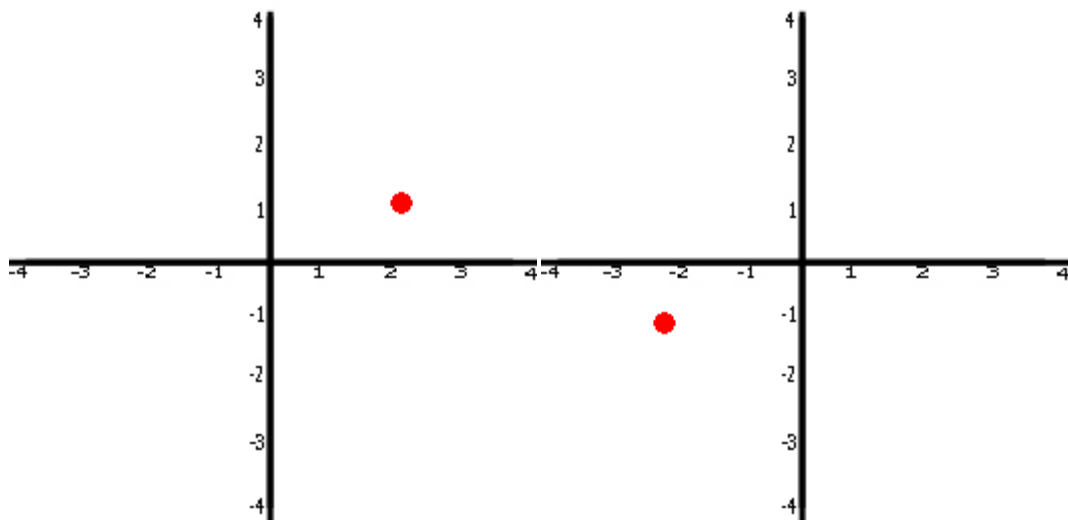
Boredom (8) type: audio



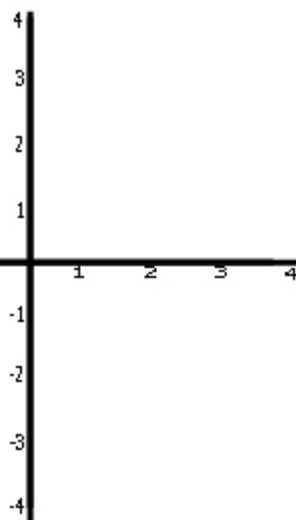
Contempt (7) type: audio



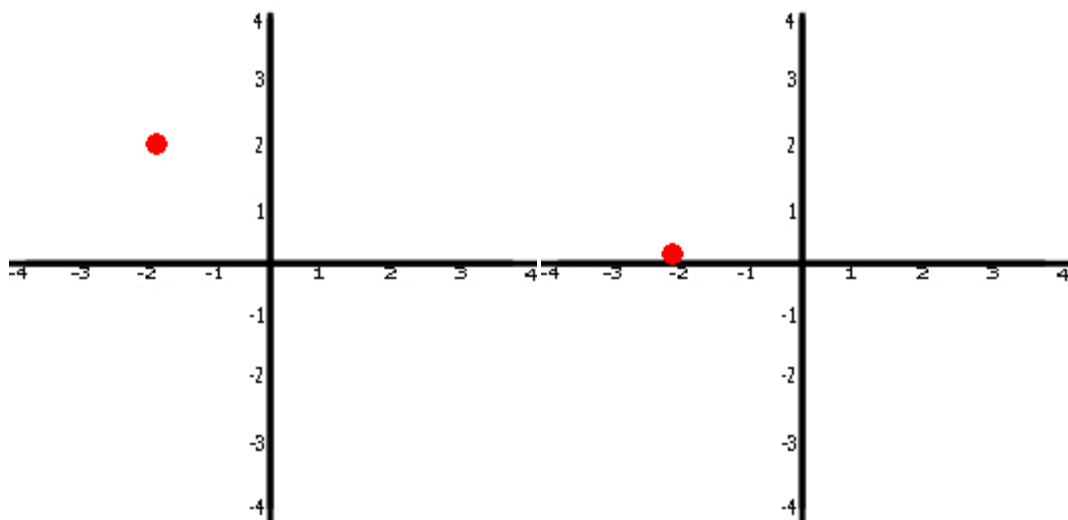
Desire (6) type: audio



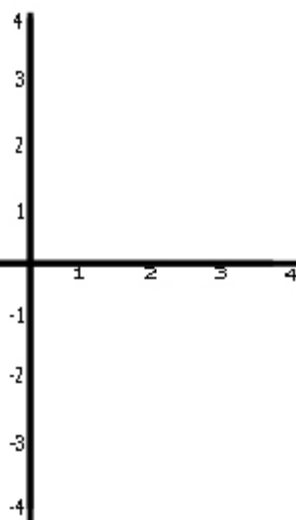
Disappointment (7) type: audio



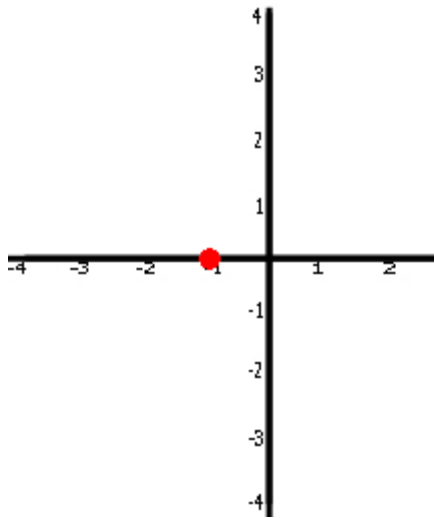
Disgust (9) type: audio



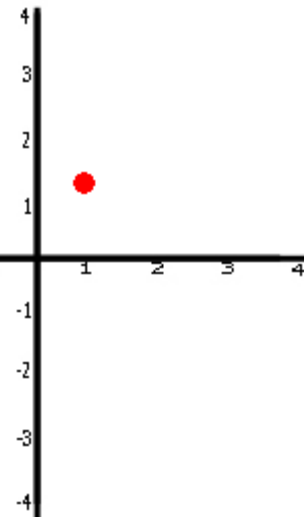
Dislike (6) type: audio



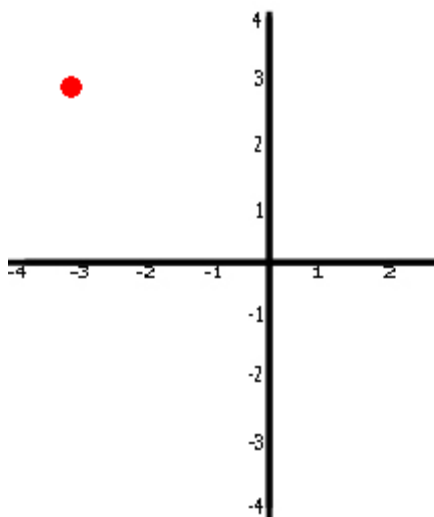
Dissatisfaction (6) type: audio



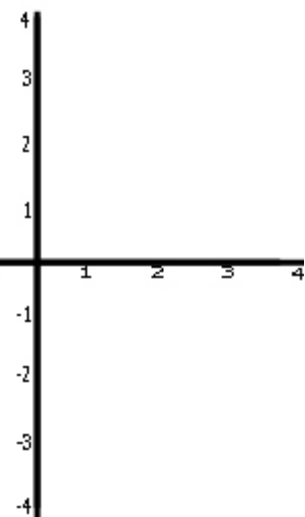
Fascination (4) type: audio



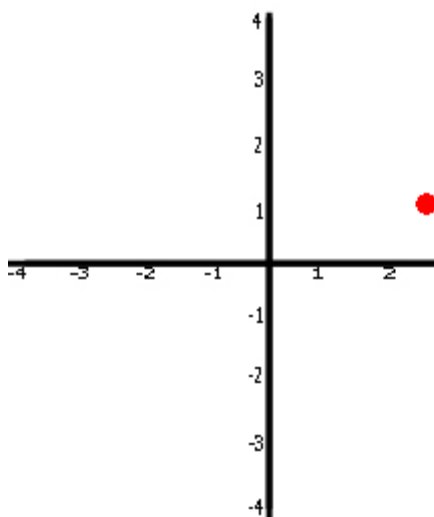
Fear (73) type: audio



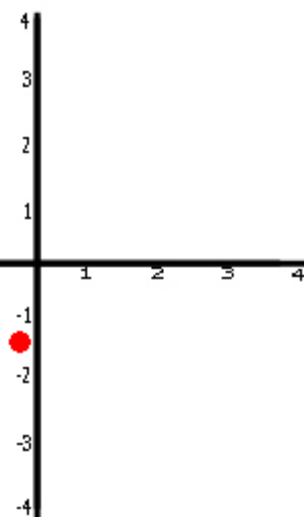
Furious (7) type: audio



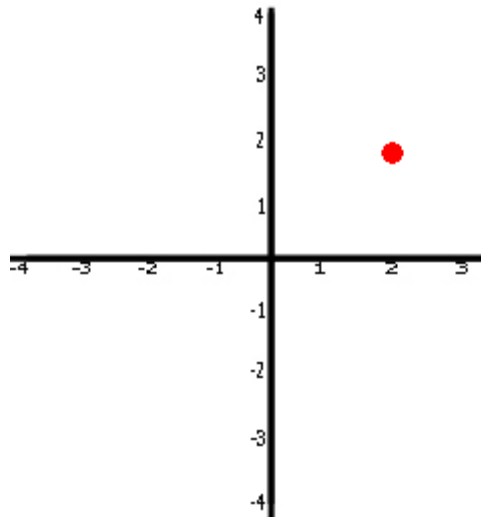
Happiness (5) type: audio



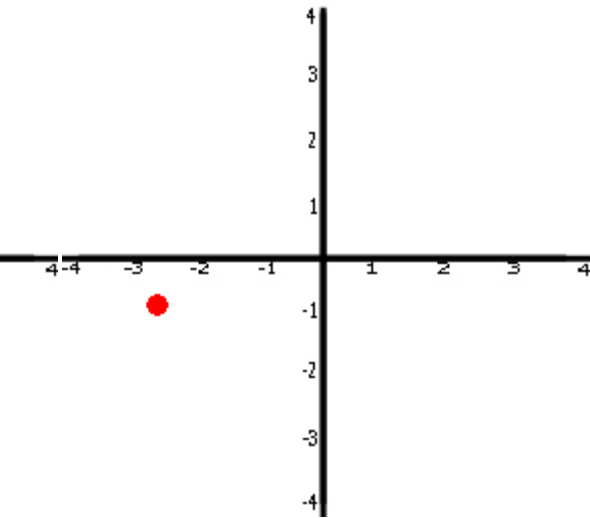
Indignation (7) type: audio



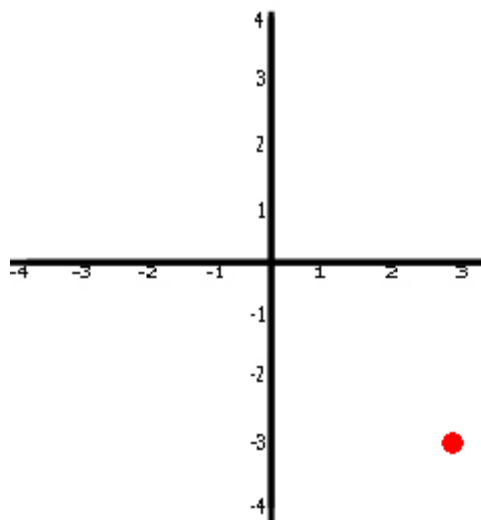
Interest (4) type: audio



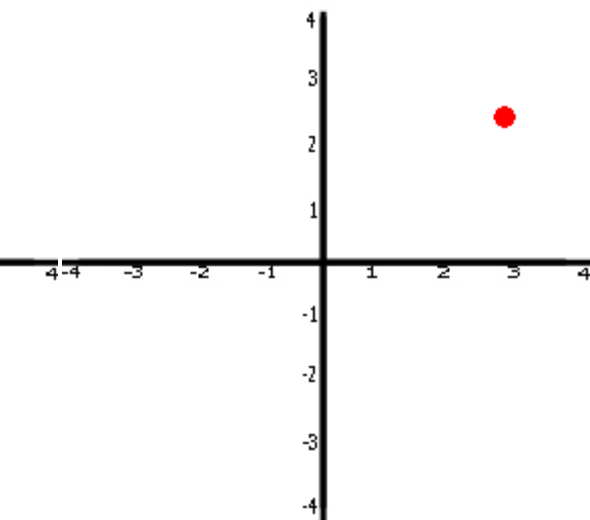
Sadness (8) type: audio



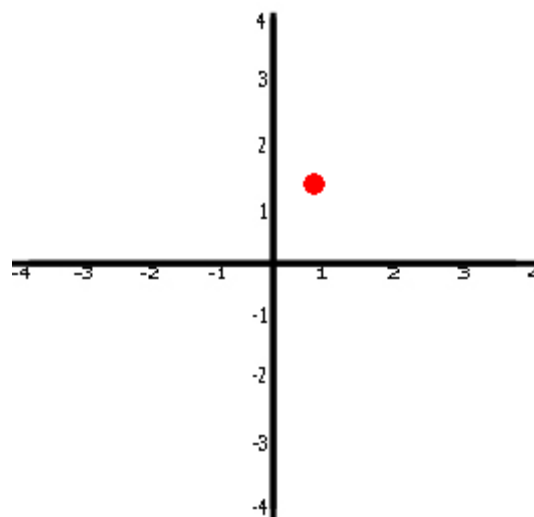
Satisfaction (9) type: audio



Surprise_(Pleasant) (9) type: audio

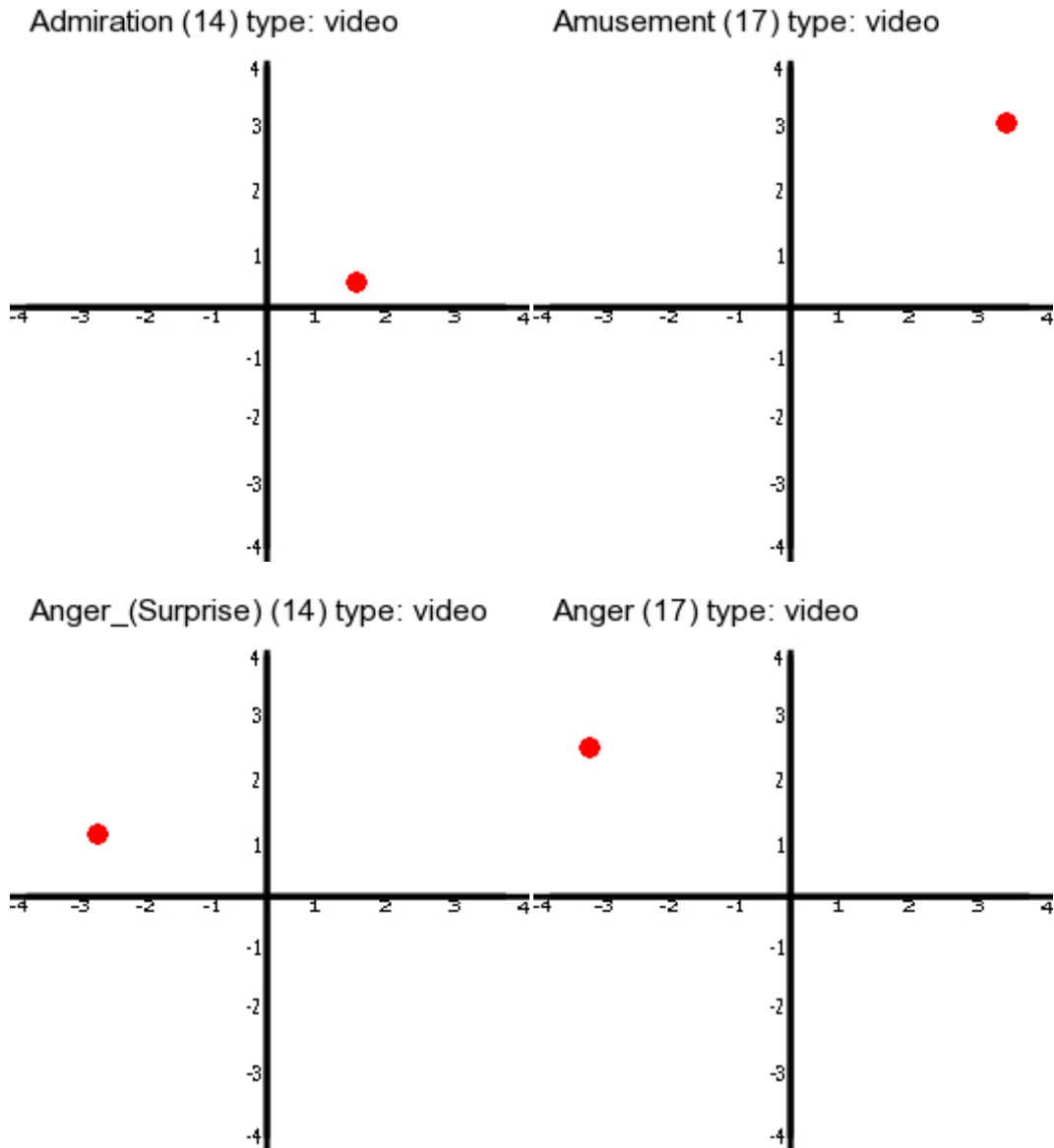


Surprise_(Unpleasant) (3) type: audio

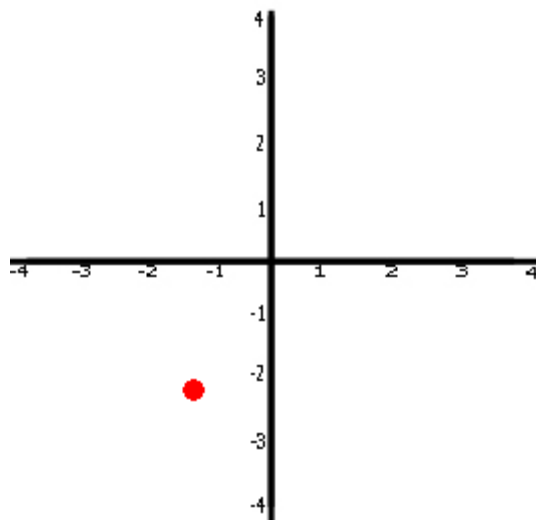


H: Video Validation Results

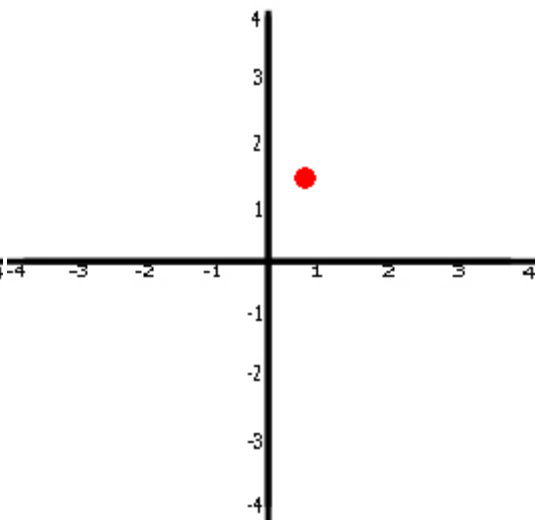
These are the results for the user validation of all the 105 different video clips showed on the website. The number between the brackets indicates the number of validations for this particular item.



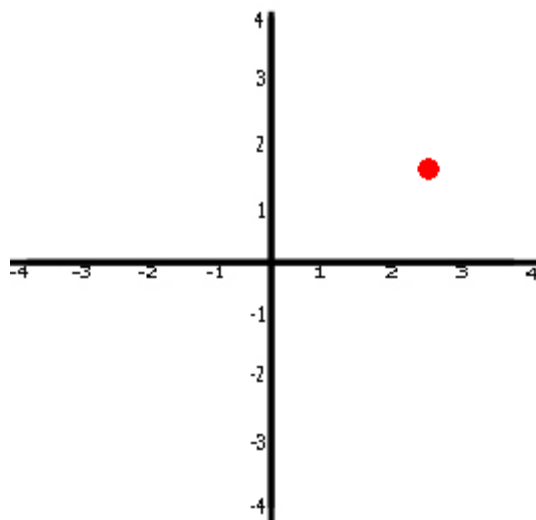
Boredom (16) type: video



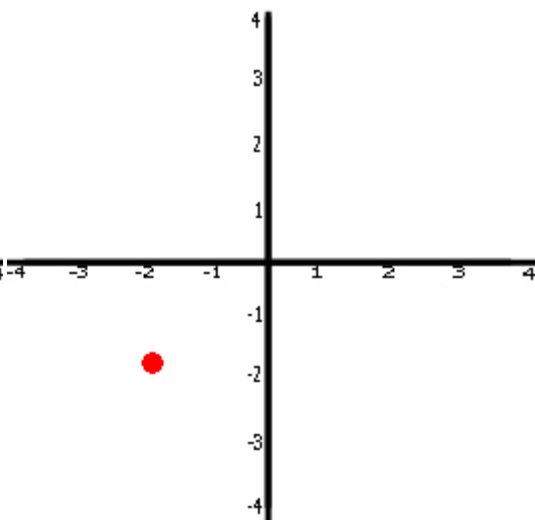
Contempt (10) type: video



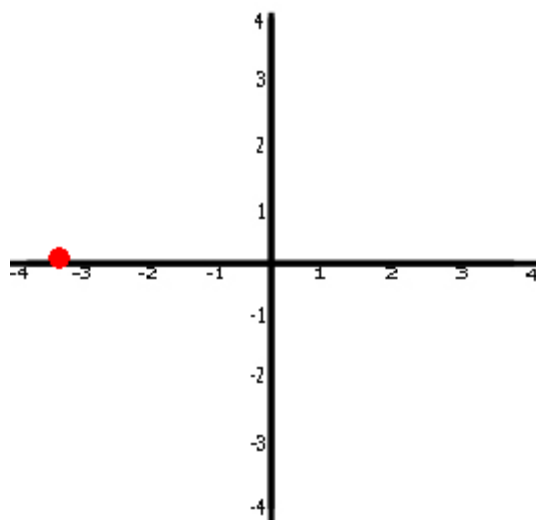
Desire (18) type: video



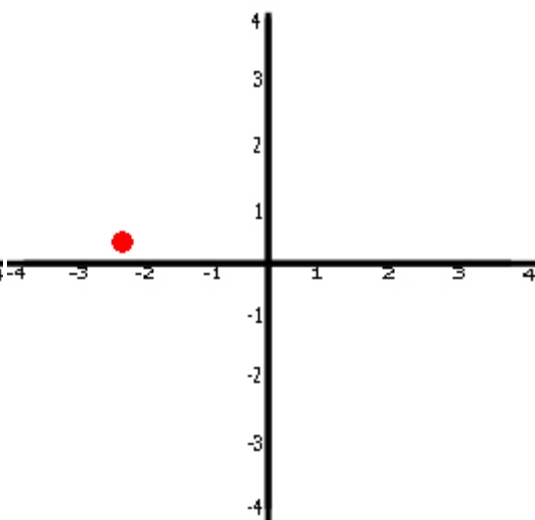
Disappointment (15) type: video



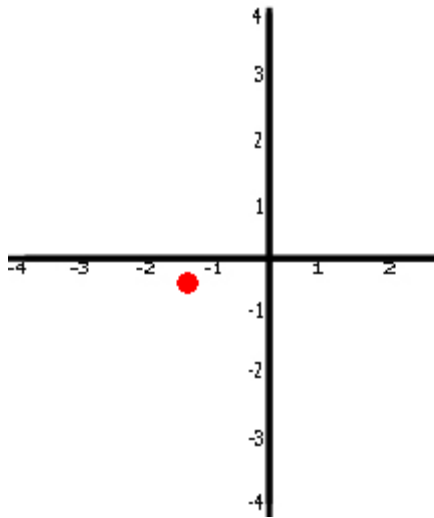
Disgust (11) type: video



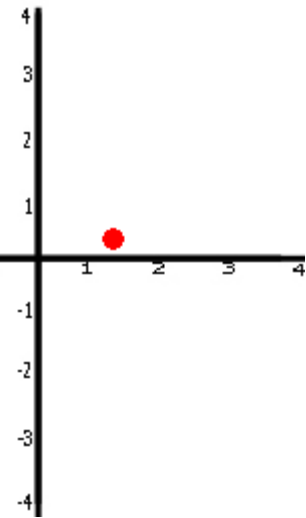
Dislike (16) type: video



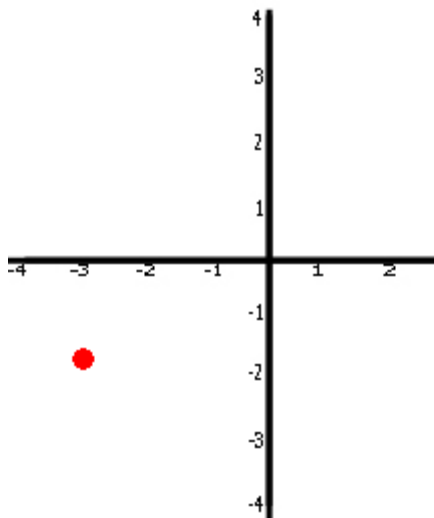
Dissatisfaction (17) type: video



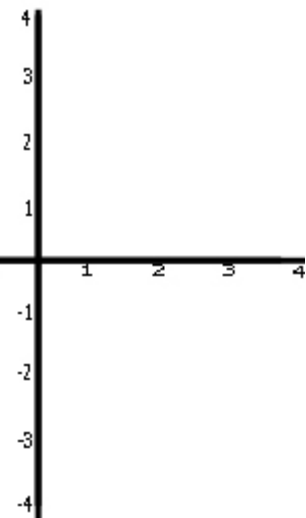
Fascination (21) type: video



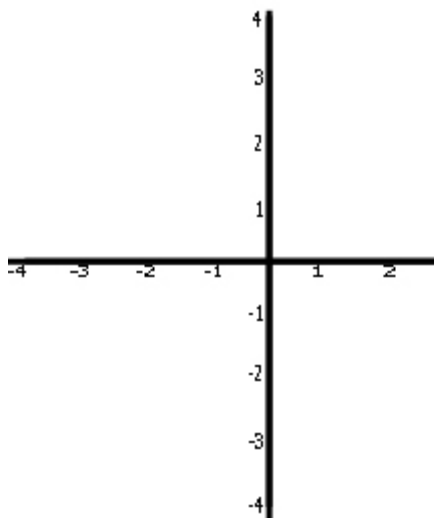
Fear (11) type: video



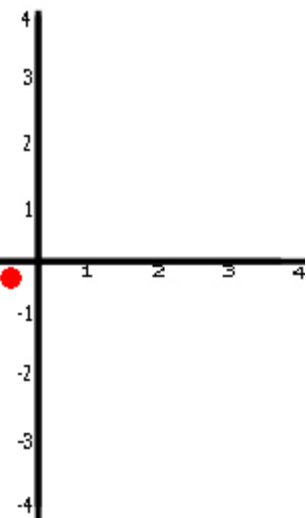
Furious (13) type: video



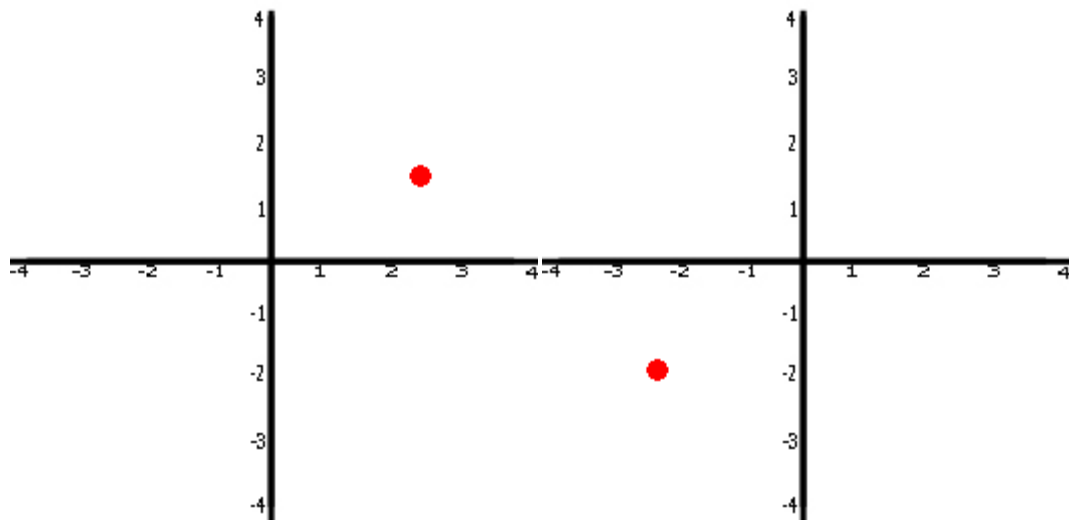
Happiness (12) type: video



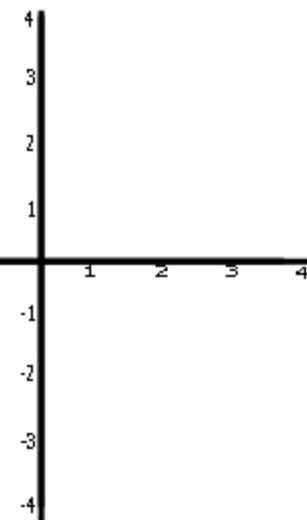
Indignation (11) type: video



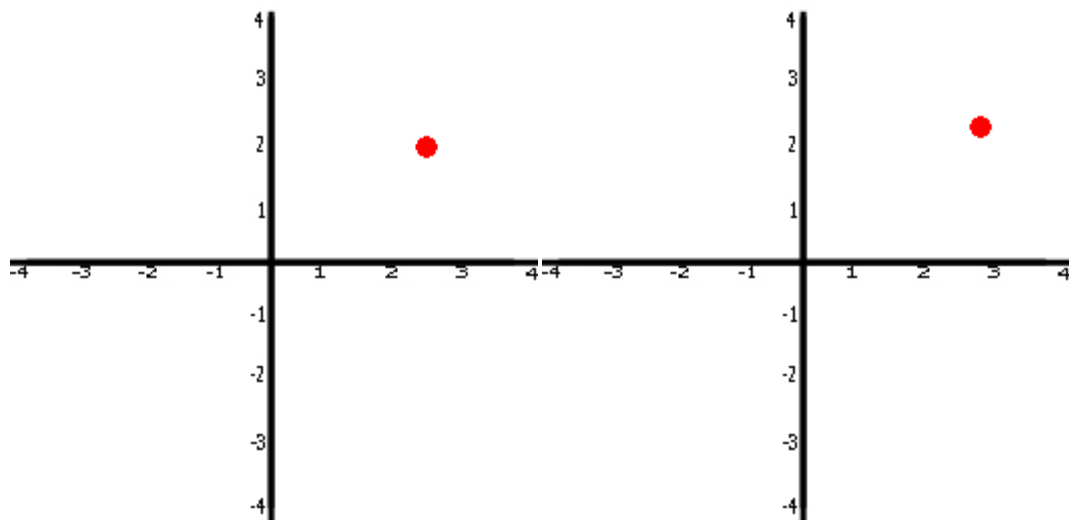
Interest (19) type: video



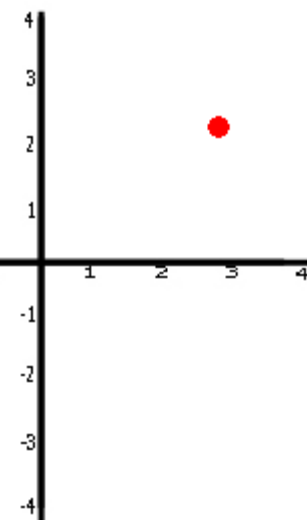
Sadness (14) type: video



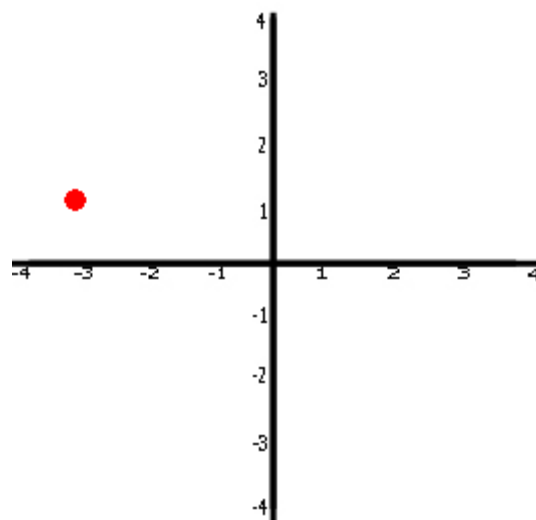
Satisfaction (14) type: video



Surprise_(Pleasant) (12) type: video



Surprise_(Unpleasant) (14) type: video



I: Table with Resulting Classification Labels for All Clips

Emotion	Audio Classifiers			Video Classifiers		% Correct processed frames
	21 Emotions	Positive vs. Negative	Active vs. Passive	Point based classification	AU based classification	
Admiration R1	Admiration	Negative	Active	Fascination	Neutral	74
Admiration R2	Admiration	Negative	Passive	Fascination	Neutral	50
Admiration R3	Admiration	Positive	Active	Fascination	Neutral	18
Admiration R4	Satisfaction	Negative	Passive	Neutral	Neutral	28
Admiration R5	Admiration	Negative	Passive	Neutral	Neutral	27
Amusement R1	Fear	Positive	Passive	Anger	Neutral	85
Amusement R2	Admiration	Negative	Passive	Fascination	Neutral	78
Amusement R3	Amusement	Positive	Passive	Fascination	Neutral	65
Amusement R4	Disappointment	Positive	Passive	Surprise (Pleasant)	Neutral	58
Amusement R5	Admiration	Negative	Passive	Surprise (Pleasant)	Neutral	50
Anger R1	Anger	Positive	Passive	Neutral	Neutral	53
Anger R2	Anger	Positive	Active	Admiration	Admiration	19
Anger R3	Furious	Negative	Passive	Surprise (Pleasant)	Neutral	22
Anger R4	Anger	Positive	Passive	Neutral	Neutral	45
Anger R5	Anger	Positive	Active	Anger	Neutral	25
Anger (Surprise) R1	Anger (Surprise)	Positive	Passive	Neutral	Neutral	81
Anger (Surprise) R2	Anger (Surprise)	Negative	Passive	Surprise (Pleasant)	Neutral	66
Anger (Surprise) R3	Desire	Positive	Active	Surprise (Pleasant)	Neutral	67
Anger (Surprise) R4	Furious	Negative	Passive	Surprise (Pleasant)	Neutral	80
Anger (Surprise) R5	Neutral	Positive	Passive	Fascination	Neutral	82
Boredom R1	Boredom	Positive	Active	Surprise (Pleasant)	Neutral	24
Boredom R2	Neutral	Positive	Active	Neutral	Neutral	20
Boredom R3	Disappointment	Negative	Active	Neutral	Neutral	59
Boredom R4	Dislike	Positive	Passive	Neutral	Neutral	24
Boredom R5	Surprise (Pleasant)	Positive	Passive	Neutral	Neutral	45
Contempt R1	Desire	Positive	Passive	Anger	Neutral	100
Contempt R2	Admiration	Positive	Passive	Neutral	Neutral	100
Contempt R3	Contempt	Positive	Passive	Neutral	Neutral	91
Contempt R4	Contempt	Positive	Passive	Neutral	Neutral	99
Contempt R5	Contempt	Positive	Passive	Neutral	Neutral	100
Desire R1	Desire	Negative	Passive	Neutral	Neutral	100
Desire R2	Anger	Negative	Active	Neutral	Neutral	99
Desire R3	Desire	Negative	Active	Neutral	Neutral	99
Desire R4	Admiration	Positive	Active	Neutral	Neutral	100

Table with Resulting Classification Labels for All Clips

Desire R5	Desire	Negative	Passive	Neutral	Neutral	96
Disappointment R1	Disappointment	Positive	Passive	Neutral	Neutral	99
Disappointment R2	Disappointment	Negative	Passive	Neutral	Neutral	100
Disappointment R3	Admiration	Positive	Active	Neutral	Neutral	100
Disappointment R4	Satisfaction	Positive	Active	Neutral	Neutral	100
Disappointment R5	Fear	Negative	Active	Neutral	Neutral	100
Disgust R1	Admiration	Positive	Active	Neutral	Neutral	100
Disgust R2	Fascination	Positive	Active	Neutral	Neutral	99
Disgust R3	Disgust	Negative	Active	Neutral	Neutral	100
Disgust R4	Disgust	Positive	Passive	Neutral	Neutral	100
Disgust R5	Contempt	Positive	Passive	Neutral	Neutral	93
Dislike R1	Disappointment	Negative	Active	Neutral	Neutral	97
Dislike R2	Dislike	Positive	Active	Neutral	Neutral	97
Dislike R3	Desire	Positive	Active	Neutral	Neutral	100
Dislike R4	Amusement	Positive	Active	Neutral	Neutral	97
Dislike R5	Fascination	Positive	Passive	Neutral	Neutral	89
Dissatisfaction R1	Dissatisfaction	Positive	Passive	Neutral	Neutral	95
Dissatisfaction R2	Desire	Negative	Passive	Neutral	Neutral	100
Dissatisfaction R3	Disappointment	Positive	Passive	Neutral	Neutral	100
Dissatisfaction R4	Admiration	Positive	Active	Neutral	Neutral	100
Dissatisfaction R4	Admiration	Positive	Active	Neutral	Neutral	79
Fascination R1	Indignation	Positive	Active	Neutral	Neutral	99
Fascination R2	Admiration	Positive	Active	Neutral	Neutral	99
Fascination R3	Fascination	Negative	Passive	Neutral	Neutral	99
Fascination R4	Amusement	Negative	Active	Neutral	Neutral	100
Fascination R5	Fascination	Negative	Passive	Neutral	Neutral	100
Fear R1	Fear	Positive	Passive	Neutral	Neutral	99
Fear R2	Fear	Positive	Passive	Neutral	Neutral	100
Fear R3	Amusement	Positive	Passive	Neutral	Neutral	99
Fear R4	Happiness	Positive	Passive	Neutral	Neutral	100
Fear R5	Fascination	Negative	Passive	Neutral	Neutral	98
Happiness R1	Anger (Surprise)	Negative	Active	Neutral	Neutral	100
Happiness R2	Furious	Negative	Passive	Neutral	Neutral	100
Happiness R3	Furious	Positive	Passive	Neutral	Neutral	87
Happiness R4	Dissatisfaction	Positive	Passive	Neutral	Neutral	99
Happiness R5	Furious	Positive	Passive	Neutral	Neutral	98
Indignation R1	Happiness	Positive	Active	Neutral	Neutral	99
Indignation R2	Happiness	Negative	Active	Neutral	Neutral	99
Indignation R3	Furious	Negative	Active	Neutral	Neutral	100
Indignation R4	Anger (Surprise)	Positive	Active	Neutral	Neutral	100

Table with Resulting Classification Labels for All Clips

Indignation R5	Happiness	Negative	Active	Neutral	Neutral	100
Interest R1	Indignation	Negative	Active	Neutral	Neutral	100
Interest R2	Indignation	Positive	Passive	Surprise (Pleasant)	Neutral	99
Interest R3	Disappointment	Positive	Passive	Neutral	Neutral	89
Interest R4	Furious	Negative	Passive	Neutral	Neutral	100
Interest R5	Dissatisfaction	Positive	Passive	Neutral	Neutral	96
Sadness R1	Interest	Negative	Passive	Neutral	Neutral	99
Sadness R2	Interest	Positive	Passive	Neutral	Neutral	100
Sadness R3	Admiration	Positive	Active	Neutral	Neutral	97
Sadness R4	Interest	Negative	Active	Neutral	Neutral	98
Sadness R5	Interest	Negative	Passive	Neutral	Neutral	98
Satisfaction R1	Disappointment	Positive	Passive	Neutral	Neutral	100
Satisfaction R2	Neutral	Negative	Active	Neutral	Neutral	100
Satisfaction R3	Furious	Negative	Active	Neutral	Neutral	100
Satisfaction R4	Fear	Negative	Active	Neutral	Neutral	100
Satisfaction R5	Neutral	Negative	Active	Neutral	Neutral	100
Surprise (Pleasant) R1	Amusement	Positive	Passive	Neutral	Neutral	100
Surprise (Pleasant) R2	Dissatisfaction	Positive	Passive	Neutral	Neutral	100
Surprise (Pleasant) R3	Disappointment	Positive	Passive	Sadness	Neutral	100
Surprise (Pleasant) R4	Dissatisfaction	Positive	Passive	Surprise (Pleasant)	Neutral	98
Surprise (Pleasant) R5	Fear	Negative	Passive	Neutral	Neutral	100
Surprise (Unpleasant) R1	Disappointment	Positive	Active	Disgust	Neutral	100
Surprise (Unpleasant) R2	Boredom	Positive	Passive	Neutral	Neutral	98
Surprise (Unpleasant) R3	Satisfaction	Positive	Passive	Surprise (Pleasant)	Neutral	100
Surprise (Unpleasant) R4	Fascination	Negative	Active	Neutral	Neutral	84
Surprise (Unpleasant) R5	Satisfaction	Negative	Passive	Surprise (Pleasant)	Neutral	100

J: Normalised Distance Matrix for 2 Persons for all 21 Feature Vectors

	Surprise (Unpleasant)	Surprise (Pleasant)	Satisfaction	Sadness	Neutral	Interest	Indignation	Happiness	Furious	Fear	Fascination	Dissatisfaction	Dislike	Disgust	Disappointment	Desire	Contempt	Boredom	Anger	Amusement	Admiration
Lotte	0.9272	0.7551	0.4035	0.0088	1.0331	0.4393	1.0106	1.1701	0.3419	0.0914	0.6132	0.0128	0.3686	1.6881	0.2566	0.3282	0.5216	1.355	1.5121	0.3183	2.5691
Maaike	0.5867	0.0453	0.2184	0.2939	0.688	0.2463	0.3917	0.502	0.2166	0.1801	0.1424	1.0352	0.4612	1.1172	1.1366	0.6659	0.8643	0.2603	2.2912	0.0201	2.8545
	0.6714	0.0551	0.0968	0.2505	1.1236	0.1756	0.4375	0.666	0.3222	0.1665	0.265	1.1981	0.0149	1.3124	1.0443	0.6686	0.9854	0.3257	1.9911	0.0969	2.7563
	1.0983	0.2381	0.1775	0.081	0.1488	0.7488	0.7159	0.0503	0.2125	0.2302	0.3301	0.9515	0.2207	1.538	1.1658	0.2834	0.791	0.2452	2.6162	0.1743	2.2527
	1.452	0.5272	0.1163	0.2105	0.0265	1.0487	1.0786	0.1673	0.4499	0.5429	0.4119	0.1131	0.4523	1.8717	0.8912	0.6183	0.2987	0.6174	2.3469	0.2978	1.9061
	0.9429	0.5553	0.1441	0.0238	0.7385	1.1898	1.1194	0.2847	0.8919	0.9677	0.2762	1.0817	0.2056	1.8338	0.1166	1.3526	0.3064	1.9715	0.9517	0.5376	1.4437
	0.9085	0.0381	0.053	0.1927	0.7607	0.5684	0.7895	0.4816	0.4319	0.1779	0.3835	0.8587	0.2245	1.5823	1.1719	0.0791	0.7884	0.1119	2.1459	0.0755	2.6637
	0.4112	0.0988	0.5095	0.6776	0.6947	0.1517	0.0164	0.5113	0.0421	0.4689	0.5758	1.0724	0.798	1.1406	0.7916	1.2133	0.9491	0.6679	2.0561	0.0926	2.6481
	0.8468	0.1111	0.0595	0.0782	0.5319	0.2272	0.3388	0.3339	0.1743	0.3465	0.345	1.151	0.324	0.3861	1.0833	0.6348	0.9149	0.8401	2.5564	0.0643	2.682
	0.9922	0.3252	0.1267	0.2755	0.2647	0.7705	0.9028	0.4108	0.6802	0.3912	0.2834	0.183	0.3326	1.9696	0.8197	0.0679	0.5098	0.9668	2.1031	0.1782	2.4749
	0.6916	0.1531	0.2327	0.2605	0.5519	0.412	0.7088	0.6153	0.4613	0.0614	0.2843	0.5757	0.4236	1.7738	0.9378	0.2492	0.7927	0.5343	1.8954	0.2176	2.9276
	0.2096	0.4982	0.0473	0.163	1.0441	0.0244	0.1236	0.4101	0.2147	0.3371	0.0277	1.6099	0.4695	1.0627	0.98	0.7915	1.1184	0.807	1.7392	0.0297	2.7907
	0.5795	0.0456	0.101	0.1435	0.8583	0.2469	0.4755	0.7324	0.4618	0.1338	0.2035	0.9261	0.6452	1.2454	0.9546	0.633	0.9001	0.1092	2.0073	0.0873	2.9669
	0.6216	0.0444	0.3078	0.0444	0.0251	0.5762	0.4629	0.546	0.3788	0.1155	0.138	0.6492	0.8638	0.5669	0.9602	0.2369	0.6283	0.1202	2.2978	0.3381	3.1974
	0.0494	0.6044	0.0059	0.0074	0.7781	0.1095	0.0059	0.2695	0.1084	0.4016	0.084	1.6954	0.5775	0.1476	0.9944	0.8891	1.1941	0.9783	1.4422	0.0956	3.0464
	0.7643	0.2314	0.125	0.3985	1.1068	0.2309	0.4833	0.6473	0.0591	0.222	0.0223	0.9204	0.0429	1.2285	1.1522	0.728	0.6166	0.5897	2.4101	0.1353	2.5363
	0.2603	0.2185	0.1046	0.1695	0.7811	0.0692	0.1946	0.3615	0.1981	0.2515	0.2454	1.4435	0.6108	1.1049	1.1293	0.6927	1.1218	0.4475	2.0297	0.0319	2.8083
	0.9462	0.138	0.741	1.3069	0.2348	0.0512	0.3094	0.0894	0.1641	0.3495	0.2634	0.818	0.0143	2.2597	0.3859	0.5964	1.1053	0.349	1.9549	0.7941	2.0912
	0.4209	0.0951	0.186	0.3487	1.091	0.0246	0.1476	0.5605	0.0202	0.3919	0.0267	1.2736	0.2055	0.7296	1.1124	1.0097	0.8578	0.9816	2.1302	0.1143	2.6967
	0.4653	0.088	0.1366	0.0471	0.5917	0.2876	0.4085	0.5433	0.5612	0.0511	0.2998	0.9549	0.8194	1.2716	0.9711	0.5343	1.0047	0.1749	1.9223	0.0012	3.0477
	0.9254	0.6773	0.052	0.0179	0.4052	0.8257	0.9572	0.1043	0.6126	0.5424	0.434	0.1665	0.5586	1.8112	0.7965	0.6756	0.3826	1.1916	2.0402	0.3891	2.3366

K: Consent Document

Het doel van dit experiment is het creëren van een multimodale database waarin de proefpersonen verschillende emoties vertonen, zowel auditief als visueel. Het vervolgonderzoek aan multimodale emotieherkenning zal, naar wij denken, veel gebruik gaan maken van deze nieuwe datacollectie.

Het experiment waar u aan meewerkt duurt ongeveer één uur per sessie. U kunt er voor kiezen om meerdere sessies mee te doen, hiervoor hoeft u maar eenmaal dit document in te vullen. De toegang tot deze data wordt gelimiteerd tot personen die onderzoek doen aan de TU Delft en andere opleidingen of instituten die onderzoek doen naar multimodale emotieherkenning. Zij krijgen pas na het invullen van een gebruikersovereenkomst toegang tot de data.

Tijdens het experiment maken wij opnames van uw stem en van een voor en zijaanzicht van uw hoofd. U kunt het onderzoek ten allen tijde stop zetten, dit heeft geen verdere gevolgen voor u. Uw data zal dan niet gebruikt worden voor onderzoek naar multimodale emotieherkenning.

Het is de bedoeling om zo veel mogelijk verschillende personen toe te voegen aan de verzameling opnames, een minimum aantal personen voor dit onderzoek is 50.

Indien u akkoord gaat met deelname aan het onderzoek, dient u een ondertekend exemplaar van dit document te overhandigen. Uw deelname aan dit onderzoek is vrijwillig. Weigering tot deelname of een besluit om van verdere deelname af te zien zal niet leiden tot strafmaatregelen of het verlies van uw aanspraken.

Door dit document te ondertekenen verklaart u dat het onderzoek, met inbegrip van de bovenstaande informatie, u mondeling is beschreven en dat u er vrijwillig aan deelneemt.

Gaat u akkoord met het feit dat:

Uw opnames gebruikt mogen worden in onderzoekspublicaties: **JA / NEE**.

Afbeeldingen uit uw opnames gebruikt worden in publicaties, onderzoekspresentaties en/of demo's: **JA / NEE**

Audio clips uit uw opnames gebruikt worden in publicaties, onderzoekspresentaties en/of demo's: **JA / NEE**

handtekening deelnemer

datum

Gebruikersgegevens

Wilt u zo vriendelijk zijn om alle gegevens correct in te vullen. De juistheid van de gegevens tijdens de opnames kan van belang zijn voor verder onderzoek.

Wij garanderen dat uw naam nooit publiekelijk bekend wordt gemaakt.

Naam: _____

Leeftijd: _____

Geslacht: MAN / VROUW

Beroep: _____

Etnische achtergrond: Aziatisch / Blank / Negroïde / Hispanisch

L: Building a Dutch Multimodal Corpus for Emotion Recognition

Building a Dutch Multimodal Corpus for Emotion Recognition

Alin G. Chitu, Mathijs van Vulpen, Pegah Takapoui and Leon J.M. Rothkrantz

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands

E-mails: {A.G.Chitu,L.J.M.Rothkrantz}@ewi.tudelft.nl, M.vanVulpen@student.tudelft.nl, pegahtak@gmail.com

Abstract

Multimodal emotion recognition gets increasingly more attention from the scientific society. Fusing together information coming on different channels of communication, while taking into account the context seems the right thing to do. During social interaction the affective load of the interlocutors plays a major role. In the current paper we present a detailed analysis of the process of building an advanced multimodal data corpus for affective state recognition and related domains. This data corpus contains synchronized dual view acquired using high speed camera and high quality audio devices. We paid careful attention to the emotional content of the corpus in all aspects such as language content and facial expressions. For recordings we implemented a TV prompter like software which controlled the recording devices and instructed the actors to assure the uniformity of the recordings. In this way we achieved a high quality controlled emotional data corpus.

1. Introduction

The affective state of a person is very important in human communication. During social interaction humans express their affective state through a large variety of channels, such as facial expressions, communicative gestures like body posture, emotional speech, etc. The semantic content of our communication is largely enriched by transmitting to the interlocutor our current affective state. The affective state influences the way we interact with our interlocutors, our actions and reactions to certain situations. Also, in the case of human computer interaction, it would greatly increase the quality of our experiences if the machine would be able to adapt to our affective state. We can imagine for instance that we are involved into a crisis situation and we use our PDA to communicate to and receive indications from a central crisis management center. Knowing the affective state of the user the system can adapt the content and layout of the messages to increase their receptivity. The system can do this transparently, for all users without requiring that the sender is aware of this. In this way we can optimize the search and rescue activities. There are many other applications of affective state recognition, to name a few more: children toys which can tailor to the children needs in each moment, any public kiosks, ATMs, driver safety systems, etc.

As in the case of speech recognition (McGurk and MacDonald 1976) people use context information acquired through different communication channels to improve the accuracy of the affective state recognition. For instance speech and emotion recognition are two much interconnected processes, which influence each other. The exact influence is not completely elucidated. Our speech influences the facial expressions and our facial expressions influence our speech. Of course the affective state of the speaker is largely transmitted

through prosody. Buchan et. al. (Buchan et. al. 2007) analyzed what the subjects are watching while trying to understand what people are saying or what facial expressions are they showing. They showed that the distribution of gaze is dependent on the distribution of information in the face and on the goals of the user. It was concluded as well that emotion related information is spread on the entire face. Notable is for instance the concentration of the gaze around the nose when the signal to noise ratio decreases.

Data corpora are an important building block of any scientific study. The data corpus should provide the means for understanding all the aspects of a given process, direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a good data corpus (i.e. well designed, capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results. Having this in mind we decided to build such a data corpus. A good data corpus should have a good coverage of the process it's going to be investigated such that every aspect should get a fair slice.

We present in this paper a detailed analysis of the process of building an advanced multimodal emotion data corpus for the Dutch language. We strongly believe that sharing our experiences is the first step for understanding the issues around building a reliable data corpus. We envision a future standard for data corpora that combines the views of the entire scientific community.

2. Recordings' settings

This section presents the settings used while compiling the data corpus. Figure 1 shows the complete image of the setup. We used a high speed camera, a professional

microphone and a mirror for dual view synchronization. The camera was controlled by the speaker, through a prompter like software. The software was presenting the speaker the next item to be uttered together with directions on the speaking style required. This provided us with a better control of the recordings.

2.1 Audio and Video devices

The audio and video quality is an important issue to be covered. An open question is for instance, what is the optimum sampling rate in the visual domain? Current standard for video recording frame rate ranges from 24 up to 30 frames per second, but is that enough? A first problem and the most intuitive is the difficulty in handling the increased amount of data, since the bandwidth needed is many times larger. A second problem is a technical problem and is related with the techniques used for fusing the audio and video channels. Since it is common practice to sample the audio stream at a rate of 100 feature vectors per second, in the case when the information is fused in an early stage, we encounter the need to use interpolation to match the two data sampling rates. A third issue, that actually convinced us to use a high speed camera, is related to the coverage of the visemes during recording, namely the number of frames per visemes. In the paper Chitu and Rothkrantz 2007 it was showed that the visemes coverage becomes a big issue when the speech rate increases. While talking with experts from the brain and speech domain we learned that recording at 125Hz should cover almost every movement on a person’s face. There are, however, movements like the lips vibration when the air is pushed with high speed through the loosely closed lips that require some 400Hz for exact recording. Therefore we decided to use a high speed camera for video recordings. As we aim to discover where the most useful information for emotion detection lies and we want to give the possibility for developing new applications we decided to include side view recordings of the speaker’s face in our corpus.

When one goes outside the range of consumer devices, things become extremely more complicated and definitely more expensive. The quality of the sensors and the huge bandwidth necessary to stream high speed video to the PC makes high speed video recording very restrictive. We used for recording a Pike F032C camera built by AVT. The camera is capable of recording at 200Hz in black and white, 139Hz when using the chroma subsampling ratio 4:1:1 and 105Hz when using the chroma subsampling ratio 4:2:2 while capturing at maximum resolution 640x480. By setting a lower ROI the frame rate can be increased. In order to increase the Field Of View (FOV), as we will mention later, we recorded in full VGA resolution. To be able to guarantee a fix and uniform sampling rate and to permit an accurate synchronization with the audio signal we used a pulse generator as an external trigger. A sample frame is shown in Figure 2. To acquire a synchronized dual view we used a mirror which was placed behind the speaker at 45 (see Figure 1).



Figure 30: The setup of the experiment.

In the case of video data recording there are a larger number of important factors that control the success of the resulted data corpus. Hence, not only the environment, but also the equipment used for recording and other settings is actively influencing the final result. The environment where the recordings are made is very important since it can determine the illumination of the scene, and the background of the speakers. We use monochrome background so that by using a “chroma keying” technique the speaker can be placed in different locations inducing in this way some degree of visual noise.

For recording the audio signal we used NT2A Studio Condensators. We recorded a stereo signal using a sample rate of 48kHz and a sample size of 16bits. The data was stored in PCM audio format. The recordings were conducted in controlled laboratory environment. We considered that it is more advantageous to have very good quality recordings and degrade them in a post process as needed. The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data. An example of such database is NOISEX-92 (Varga and Steeneken 1993). This dataset contains white noise, pink noise, speech babble, factory noise, car interior noise, etc.

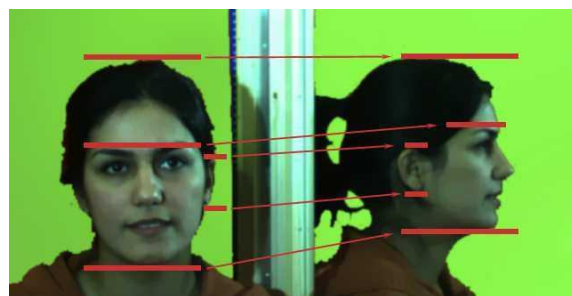


Figure 31: Sample frame with dual view.

2.2 The Prompter Tool

Using a high speed camera increases the storage needs for the recordings. It is almost impossible to record everything and then during the annotation process, cut the clips at the required lengths. One main reason is that

when recording in high speed high resolution the bandwidth limitation requires that the video be captured in the memory (e.g. on a RAM Drive). This makes the clips to have a maximum length of approximately 1 minute, depending on the resolution and color subsampling ratio used. However, we needed anyway to present the speakers with the pool of items required to be uttered. We build therefore a prompter like tool that provided the user the next item to be uttered together with some instructions about the speaking style and also controlled the video and audio devices. The result was synchronized audio and video clips already cropped to the exact length of the utterance. The tool provided the speaker the possibility to change the visual themes to maximize the visibility, and offer a better recording experience.



Figure 32: Prompter view during recordings.

The control of the software was done by the speaker through the mouse buttons of a wireless mouse that was taped on the arm of the chair. After a series of trials we conclude that this level of control is sufficient and not very disruptive for the speaker. The tool was also used to keep track of the user’s data, recording takes and recording sessions.

2.3 Emotional speech

There are two different approaches to collect data for an emotion database: by capturing real data or by inducing the emotional status to the actors. The first approach is almost impossible to be used because of all the ethical issues linked with trust and personal intimacy. Therefore we collected a set of stories which carried a strong emotional load. We asked each speaker to read each story and then transpose him/herself into the right affective state and utter a set of 5 appropriate sentences as a possible reaction to the particular story. Of course a good question regarding this approach would be whether the quality of the expressed emotions is preserved, or the recorded material contains artificial performances. In real life it is very difficult to select isolated emotions; usually people show an amalgam of emotions. The speakers were divided into two groups: professional actors and naive speakers. All speakers were native Dutch. This is very important for the case of emotional speech since the

performance of the speaker could get less genuine and definitely less spontaneous as result of the speaker spending more time in preparing his speech. However, it could be very interesting to analyze the cultural effect on expressing ones’ emotions through facial expressions and prosody. We recorded 21 emotions which are listed in Table 1. An example of the story and reactions used for recordings is given in Table 2.

#	emotion	#	Emotion
1	Admiration	12	Fear
2	Amusement	13	Fury
3	Anger	14	Happiness
4	Boredom	15	Indignation
5	Contempt	16	Interest
6	Desire	17	Pleasant surprise
7	Disappointment	18	Unpleasant surprise
8	Disgust	19	Satisfaction
9	Dislike	20	Sadness
10	Dissatisfaction	21	Inspiration
11	Fascination		

Table 18: List of emotions considered for recordings.

Dutch original
<p>Emotie: “Bewondering” Vertelling: “Je loopt samen met een vriend/vriendin door een dure winkelstraat in Amsterdam en ziet in de etalage een jas hangen die je altijd al had willen hebben. Je droomt over wat je zou doen als je het geld had om deze jas te kopen. Je gaat voor de etalage staan en denkt...” Reactie: R1: Oooohhh... R2: Dat ziet er goed uit! R3: Die zou ik graag hebben! R4: Was die maar van mij! R5: Zodra ik mijn geld heb, is die jas van mij!</p>
English approximative translation
<p>Emotion: “Admiration” Story: “You walk together with your friend/girlfriend in front of a fancy store in Amsterdam and you see in the store’s window a coat that you always wanted. You dream of what you would do if you have had the money to buy the coat. You stand in front of the window and think.... Reaction: R1: Oooohhh... R2: That looks so nice! R3: we would really want it! R4: That is for me! R5: As soon as we’ll have money, that coat is mine.</p>

Table 19: Story and possible reactions for “admiration”.

3. Demographic data recorded

As we specified in the introduction a proper coverage of the variability of the speakers is needed to assure the success of a data corpus. We also have seen that there is a language use difference between speakers. This can be used for instance to develop adaptive recognizers. Therefore we recorded for each speaker the following data: gender, age, education level, native language (as well as whether he/she is bi-lingual) and region where he/she had grown up. The last aspect is used to identify possible particular clusters in the pool of actors. The cultural background of the actors can play an important role in the expressions showed. Persons from different cultures might give different meaning to different gestures and expressions. In our case since we only collect data based on native Dutch speakers we expect that the cultural impact to be reduced. However, it is a matter that should be investigated anyway.

4. Research goals and usability of the resulted data corpus

As we specified in the introduction the presented corpus targets the domain of multimodal affective state recognition. However, we have a large interest in analyzing the degree in which the emotional content and the speech content interfere. Hence we would like to be able to describe the impact of the affective state on the visemes shown by the speaker.

We also envision that by analyzing the data recorded we will be able to develop a formal way for annotating and describing such affective data.

We also expect that the resulted data corpus will enable the analysis of the recording quality, especially of the video sampling rate on the recognition results.

5. Data corpus size

The duration of each recording session was approximately 45 minutes. Each session resulted in a number of 105 performances recorded by the actor. Hence each actor recoded approximately 15 minutes. We collected data from 25 persons, mainly students at our technical university (of course we also took advantage of the rest of the staff in our department). We would like however that our complete data corpus to contain data from at least 50 actors. We also have access to a number of professional actors which agreed to take part in our experiment. This set is particularly important because their performances are going to be used for assessing the quality of the acted emotions by the rest of the actors. Hence in total we expect to collect more than 5000 performances.

6. Conclusions

We presented in this paper our thoughts and investigations on building a good data corpus. We presented the settings used during the recordings, the language content and the recordings progression. The

new data corpus should consist of high speed recordings of synchronized dual view of speaker faces while uttering emotional speech and showing the appropriate facial expressions. It should provide a sound tool for training, testing, comparison and tuning a highly accurate affective state recognizer. There are still many questions to be answered with respect to building a data corpus. For instance which modalities are important for a given process, and moreover what is the relationship between these modalities. Is there any important influence between different modalities? A major issue to be addressed is the quality of the acted data. As we specified we plan to use the recordings of the professional actors to assess the quality of the rest of the naïve actors.

Our data corpus only contain recordings with individuals showing emotions triggered by reading some emotional stories, however we only consider scenes with single actors showing “clean” emotions. However, it has been shown that there are multiple situations in real life when people show in fact an amalgam of emotions. This issue should be address as well.

7. Acknowledgements

The work reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

8. References

- (Buchan et. al. 2007) Julie N. Buchan, Martin Paré and Kevin G. Munhall, „Spatial statistics of gaze fixations during dynamic face processing”, *Journal of Social Neuroscience*, 2007, vol 2, 1-13.
- (Chitu and Rothkrantz 2007) Alin G. Chitu and Leon J.M. Rothkrantz, "The Influence of Video Sampling Rate on Lipreading Performance", *12-th International Conference on Speech and Computer (SPECOM'2007)*, ISBN 6-7452-0110-x, pp. 678-684, Moscow State Linguistic University, Moscow, October 2007.
- (McGurk and MacDonald 1976) McGurk, H. & MacDonald, J. Hearing lips and seeing voices *Nature*, 1976, 264, 746 – 748.
- (Varga and Steeneken 1993) Varga, A. and Steeneken, H. 1993. “Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems.” *Speech Communication*, (vol. 12, no. 3, pp. 247-251, July).