Individual and Inter-related Action Unit Detection in Videos for Affect Recognition

THÈSE N^{O} 6837 (2016)

PRÉSENTÉE LE 19 FÉVRIER 2016

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE TRAITEMENT DES SIGNAUX 5

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Anıl YÜCE

acceptée sur proposition du jury:

Dr J.-M. Vesin, président du jury

Prof. J.-Ph. Thiran, Prof. D. Sander, directeurs de thèse

Prof. M. F. Valstar, rapporteur

Prof. H. K. Ekenel, rapporteur

Dr S. Marcel, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse

2016

Canım aileme ...
To my family ...

Acknowledgments

These past 6 years that I have spent at LTS5 has been an important period of my life, where I have learned a lot, met great people and had the chance to experience how our lives have evolved together. I would like to thank these great people who have not only helped with the thesis but have also been an important part of my life.

The first and biggest thanks goes to Prof. Jean-Philippe Thiran. Thank you Jean-Philippe, first, for accepting me in your group and then for your constant support and supervision during all these years. It is the greatest comfort to know that your advisor will always be on your side no matter what happens and that he will always have a solution for whatever problem you might have. The wonderful atmosphere in the lab is also thanks to you and I feel very lucky to have been a part of it.

I would also like to thank my co-supervisor Prof. David Sander. David, I have always felt privileged to have such a great researcher as my co-advisor and I thank you for introducing me to your domain and for always giving me new ideas and fundamental research questions that have given me directions for the thesis. Another important mentor for me was Dr. Jean-Marc Vesin, who was also the president of my thesis jury. Thank you Jean-Marc not only for this, but also for being available all these years for whatever question I had, always taking your time to answer them and more importantly for your friendship and showing me your (fun) way of doing research. I am grateful to all the other members of my thesis committee Professors Michel Valstar, Hazim Ekenel and Sébastien Marcel. I really appreciate how you have so carefully read the manuscript and all the constructive comments that have shaped the final version of my thesis. It was a great honor and pleasure to have had you in my jury. I would like to thank Michel also for kindly accepting me in his lab for three months in 2013 and for his supervision and our collaboration that has continued afterwards.

Next acknowledgment is for my hidden co-supervisor Dr. Hua Gao. Hua, this thesis literally wouldn't have happened if it wasn't for you. Thank you for your supervision, for our hard work together and for pushing me forward and motivating me. I want to also thank my first supervisor at this field and good friend, Dr. Matteo Sorci, who introduced me the world of faces. An important part of the thesis came out from our work with PSA and Valeo and I would like to thank Estelle Chin, Olivier Pajot, Patrick Bonhoure, Stéphanie Dabic and Julien Moizard for our fruitful collaboration. Thanks also to Fabien Ringeval, Daniel Dukes, Patrick Schoettker and Christophe Perruchoud, with whom I have collaborated at some point during my PhD. A special thanks to my dear friends Frank, Eleni, Gabriel, Murat, Christina and Eren for proof-reading, translating, giving me feed-back on and making better parts of this manuscript.

The best part of being at EPFL and Lausanne was certainly the great friends that I have been with. Firstly, one can rarely pick and even less often get along so well with his office-mates and I feel extremely lucky to have had three fantastic ones who have also become three of my best friends. Frank, Murat, Christina, I probably haven't been the easiest guy to have around at all times (especially towards the end) but you have certainly been the best ones someone can ever imagine. Thanks to all three of you, one by one, for being such great friends.

I have had amazing colleagues who have made being at EPFL everyday so fun and with whom I have shared many great memories. Thanks firstly to the *face group*: Gabriel, Marina, Christophe, Damien. My awesome friends Eleni, Ale, Anna and all the girls of the lab Elda, Francesca, Alia, Laura; thank you for all the fun times and the moral support at difficult times. Jeanne, Thomas & Charlotte, Benoit, thanks for making me feel like I have a french family. Thanks to the rest of LTS5 and all my corridor buddies from ASPG, LTS4, LTS2, MMSPG and LIONS; Martin, Andrea, Ashkan, Leila, Luca, Tom, Pınar, Mario, Ana, Dorina, Sofia and all the others that I fail to list here. Thank you Rosie, as well, for your help during many years.

A big part of my life in Lausanne was also spent with my *non-LTS* friends, to whom I am so grateful for the amazing moments. Gökhan, Emrah, Damla, Ali G., Gözen, Ebru, Tuğba, Mustafa, Cansaran, Marili, Onur thank you so much. Also, my friends who have been physically away but were always present: Emre, Eren, Yaprak, TOY, Muratcan, Erinç, Cansu, Eymen, Zeynep and many many others; I thank you for our long and enduring friendship that I always value so much. Particularly, I would like to thank my girlfriend Ezgi. You have been so wonderful, supporting and understanding during this period, I cannot even imagine how it would have been without you. Thank you so much for everything.

Finally, the biggest thanks is for my big, loving family and especially for my parents and my sister, who have supported me all my life in every possible way and also during these many years that I have spent away from them. I have always felt extremely lucky to have you as my family. Your presence and love have been the greatest support.

Anıl Yüce, Lausanne 2015

Abstract

The human face has evolved to become the most important source of non-verbal information that conveys our affective, cognitive and mental state to others. Apart from human to human communication facial expressions have also become an indispensable component of human-machine interaction (HMI). Systems capable of understanding how users feel allow for a wide variety of applications in medical, learning, entertainment and marketing technologies in addition to advancements in neuroscience and psychology research and many others. The Facial Action Coding System (FACS) has been built to objectively define and quantify every possible facial movement through what is called *Action Units* (AU), each representing an individual facial action.

In this thesis we focus on the automatic detection and exploitation of these AUs using novel appearance representation techniques as well as incorporation of the prior co-occurrence information between them. Our contributions can be grouped in three parts. In the first part, we propose to improve the detection accuracy of appearance features based on local binary patterns (LBP) for AU detection in videos. For this purpose, we propose two novel methodologies. The first one uses three fundamental image processing tools as a pre-processing step prior to the application of the LBP transform on the facial texture. These tools each enhance the descriptive ability of LBP by emphasizing different transient appearance characteristics, and are proven to increase the AU detection accuracy significantly in our experiments. The second one uses multiple local curvature Gabor binary patterns (LCGBP) for the same problem and achieves state-of-the-art performance on a dataset of mostly posed facial expressions. The curvature information of the face, as well as the proposed multiple filter size scheme is very effective in recognizing these individual facial actions.

In the second part, we propose to take advantage of the co-occurrence relation between the AUs, that we can learn through training examples. We use this information in a multi-label discriminant Laplacian embedding (DLE) scheme to train our system with SIFT features extracted around the salient and transient landmarks on the face. The system is first validated on a challenging (containing lots of occlusions and head pose variations) dataset without the DLE, then we show the performance of the full system on the FERA 2015 challenge on AU occurrence detection. The challenge consists of two difficult datasets that contain spontaneous facial actions at different intensities. We demonstrate that our proposed system achieves the best results on these datasets for detecting AUs.

The third and last part of the thesis contains an application on how this automatic AU detection system can be used in real-life situations, particularly for detecting cognitive distraction. Our contribution in this part is two-fold: First, we present a novel visual

database of people driving a simulator while being induced visual and cognitive distraction via secondary tasks. The subjects have been recorded using three near-infrared camera-lighting systems, which makes it a very suitable configuration to use in real driving conditions, i.e. with large head pose and ambient light variations. Secondly, we propose an original framework to automatically discriminate cognitive distraction sequences from baseline sequences by extracting features from continuous AU signals and by exploiting the cross-correlations between them. We achieve a very high classification accuracy in our subject-based experiments and a lower yet acceptable performance for the subject-independent tests. Based on these results we discuss how facial expressions related to this complex mental state are individual, rather than universal, and also how the proposed system can be used in a vehicle to help decrease human error in traffic accidents.

Keywords: facial analysis, facial expressions, action units, local binary patterns, Gabor wavelets, multi-label embedding, driver monitoring, cognitive distraction

Résumé

Le visage humain a évolué jusqu'à devenir la source d'information non-verbale la plus importante permettant de communiquer notre état affectif, cognitif ou mental. Outre la communication interpersonnelle, nos expressions faciales sont également devenues une composante indispensable des interactions homme-machine (human-machine interaction, HMI). Les systèmes capables de comprendre l'état d'esprit d'un utilisateur ouvrent un large champ d'applications dans les domaines des technologies médicales, de l'apprentissage, du divertissement et du marketing parmi tant d'autres et permettent des avancées en neurosciences ou en psychologie. Le Facial Action Coding System (FACS) a été développé pour permettre de définir objectivement et de quantifier tous les micro-mouvements possibles du visage à l'aide d'unités d'action (action units, AU), chacune représentant un changement distinct d'apparence faciale.

Dans cette thèse, nous nous concentrons sur la détection et l'utilisation automatique de ces AU à l'aide de nouvelles techniques de représentation de l'apparence ainsi que de l'intégration de l'information préalable de cooccurence de celles-ci. Nos contributions peuvent être divisées en trois parties.

Dans la première partie, nous proposons d'améliorer le taux de détection des attributs d'apparence basées sur des motifs binaires locaux (local binary pattern, LBP) pour la détection d'AU dans des vidéos. A cet effet, nous proposons deux approches méthodologiques. La première fait usage de trois outils fondamentaux du traitement d'image comme étape de pré-traitement avant le calcul des LBP sur la texture du visage. Chacun de ces outils améliore la capacité descriptive des LBP en mettant l'accent sur différentes caractéristiques transitoires de l'apparence. Il en résulte une augmentation significative du taux de détection des AU dans nos expériences. La seconde utilise plusieurs motifs binaires de courbure de Gabor locaux (local curvature Gabor binary patterns, LCGBP) pour le même problème et atteint les performances de l'état de l'art sur un ensemble d'images d'expressions faciales principalement posées. L'information de courbure du visage ainsi que le schéma de filtres multiples que nous proposons sont particulièrement efficaces dans la reconnaissance des actions faciales individuelles.

Dans la seconde partie, nous proposons de mettre à profit les relations de cooccurance entre les AU que nous pouvons apprendre à l'aide d'exemples d'entrainement. Nous utilisons cette information dans un système multi-étiquettes discriminant de plongement laplacien (discriminant Laplacian embedding, DLE) pour entraîner notre système avec des attributs SIFT extraits autour de points caractéristiques saillants du visage. Le système est d'abord validé sans le DLE sur un ensemble de vidéos difficiles (contenant beaucoup d'occlusions et de variations dans la pose de la tête). Les performances du système entier sont ensuite évaluées sur la base de données du challenge FERA 2015 de

détection d'occurences des AU. Ce challenge est constitué de deux ensemble de données très difficiles qui contiennent des actions faciales spontanées de différentes intensités. Nous montrons que notre système atteint les meilleurs résultats sur ces ensembles de données de détection d'AUs.

La troisième et dernière partie de cette thèse contient une application démontrant comment ce système automatique de détection d'AU peut être utilisé dans des situations de vie réelle, en particulier pour détecter une distraction cognitive. Notre contribution dans cette partie est double : premièrement, nous présentons une nouvelle base de données de sujets conduisant un simulateur tout en étant soumis à des distractions visuelles et cognitives sous forme de tâches secondaires. Les sujets ont été enregistrés à l'aide d'un système de trois caméras et d'éclairage proche-infrarouges, le rendant parfaitement adapté pour un usage dans des conditions de conduite réelles, soit avec de grands mouvements de la tête et des variations de la lumière ambiante. Deuxièmement, nous proposons un système original permettant de différencier automatiquement les séquences de distraction cognitive des séquences de base en extrayant certaines attributs des signaux continus d'AU et leur corrélation croisée. Nous atteignons une très grande précision de classification lorsque nous considérons chaque sujet individuellement et des performances moins bonnes, mais néanmoins acceptables, lors de tests indépendants du sujet. Sur la base de ces résultats, nous discutons l'individualité des expressions faciales liées à cet état mental complexe et leur manque d'universalité. Nous discutons également comment le système proposé peut être utilisé dans un véhicule afin de réduire les accidents de voiture dus à des erreurs humaines.

Mots clefs : analyse de visages, expressions faciales, unité d'action, motifs binaires locaux, ondelettes de Gabor, plongement multi-étiquette, surveillance de conducteur, distraction cognitive

Contents

List of Acronyms

2D	2-dimensional
3D	3-dimensional
AAM	Active Appearance Models
ADAS	Advanced Driver Assistance Systems
ADHD	Attention deficit hyperactivity disorder
ADs	Action Descriptors
AFEA	Automatic Facial Expression Analysis
ANS	Autonomic Nervous System
ASM	Active Shape Models
AU	Action Unit
AUC	Area Under Curve
BDI	Beck's Depression Index
BL	Baseline
CCC	Concordance Correlation Coefficient
CLM	Constrained Local Model
CNS	Central Nervous System
COG	Cognitive Distraction
CPM	Component Process Model
CRF	Conditional Random Fields
DBN	Dynamic Bayesian Networks

DCT	Discrete Cosine Transform
DLE	Discriminant Laplacian Embedding
DoG	Difference of Gaussians
ECG	electrocardiography
EDA	electrodermal activity
EEG	electroencephalography
EMG	electromyography
FACS	Facial Action Coding System
FDR	False discovery rate
FER	Facial Expression Recognition
fMRI	functional Magnetic Resonance Imaging
FPR	False Positive Rate
GPU	Graphics Processing Unit
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HMI	Human-Machine Interfaces
HoG	Histogram of Oriented Gradients
HR	heart rate
ICC	Intraclass Correlation Coefficient
LBP	Local Binary Patterns
LCS	Lane Change Sequence
LED	Light Emitting Diode
LGBP	Local Gabor Binary Patterns
LPQ	Local Phase Quantization
MDD	Major Depressive Disorder
MRF	Markov Random Fields
NES	Neuro-Endochrine System

NIR	Near Infrared
OA	Overall Accuracy
PCA	Principal Component Analysis
PDM	Point Distribution Model
RBF	Radial Basis Function
RF	Random Forests
RMSE	Root Mean Squared Error
ROC	Receiver Operator Characteristics
SDM	Supervised Descent Method
SIFT	Scale Invariant Feature Transform
SEC	Stimulus Evaluation Check
SNS	Somatic Nervous System
SSP	Social Signal Processing
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SURT	Surrogate Reference Task
TOP	Three Orthogonal Planes
TPR	True Positive Rate
VIS	Visual Distraction
WSR	Wilcoxon signed-rank statistical test

List of Tables

List of Figures

1 Preface

1.1 Context and Motivations

"A man's face as a rule says more, and more interesting things, than his mouth, for it is a compendium of everything his mouth will ever say, in that it is the monogram of all this man's thoughts and aspirations." said philosopher Shopenhauer. Shakespeare, on the other hand, stated how difficult it is to read those things: *"there is no art to find the man's construction in the face."* Faces and what they show and hide have always been of great interest to humanity probably even before the emergence of languages.

In terms of affective computing the importance of facial expressions and their automatic detection is reviewed in the further chapters of the thesis, both for theoretical and experimental sciences and application-wise. In this opening chapter we would rather talk about specific questions that the thesis focuses on. The main focus of the thesis is automatic detection and analysis of action units (AU) as defined by the Facial Action Coding System (FACS) (Ekman & Friesen, 1977), which are the most *basic* and objective way of explaining what is in a face. Here, we use the word *basic* not in the sense of being simple or incompetent but rather as unitary and primary.

As detailed in the next section, the thesis consists of three main subjects around this primary focus: Individual AU detection, Multi-label AU detection and Driver's cognitive distraction detection using AUs. Each part aims to address different issues in the domain and proposes solutions with a novel approach.

Although great advancement has been achieved, especially in the last decade, the automatic AU detection problem is still far from being considered solved. One of the biggest issues is how to handle differences in individual's facial configuration. Even in the neutral case (no expression) people differ in terms of the size and placement of the facial components (e.g. eyes, nose, mouth) with respect to each other. A possible solution to this problem is to apply a subject based normalization either in the feature extraction level or at the decision level, i.e. normalizing the continuous AU output values by those obtained from the neutral face of the same subject. In this thesis we apply both techniques within the different contributions. Although we propose to use manually annotated neutral frames of subjects, there exist methods to do this automatically, e.g.

via averaging on many frames of that subject, or projecting the face in a test frame onto a lower dimensional manifold obtained from neutral faces of many other training subjects.

Another important issue, even though it is not directly addressed as a major contribution in this thesis, is how to handle pose variations when using the systems in natural environments with users moving freely. These movements cause a change of appearance with respect to the camera acquisition regardless of the facial action. The relatively recent and growing trend is to use 3D acquisition systems as they become more available. However, in addition to requiring special equipment, this is also a computationally expensive approach. In this thesis we only use 2D videos and corresponding solutions to the head-pose problem, and in a specific application use a three camera acquisition system whose captured images we then use to construct a virtual frontal view and apply AU detection on this virtual image.

Individual AU detection refers to systems that do not use the joint AU information, i.e. treat each AU as an independent image label. A lot of research has already been performed on this issue, yet there is still room for improvement in terms of facial representation that can more efficiently detect the appearance changes. The use of filtering for pre-treatment of images is rarely encountered in facial image analysis. The more common practice is to directly use image transformations with varying configurations. With our first major contribution we aim to show that the filtering methods commonly used in other domains are also effective in improving the descriptive properties of facial appearance features in terms of facial action recognition. With the second one we introduce the use of curvature based appearance features in order to better represent these facial features. The face indeed consists of curved components and these are more accented when we consider the variation between the neutral face and the one that contains muscle contractions (facial actions). Curvature features had already been shown to be effective in face recognition and motivated by these facts we aimed at demonstrating their impact in AU detection.

The multi-label structuring of the AU detection problem is an approach that is much less investigated. Considering it as an image labeling problem, faces contain combinations of active or non-active AUs at all times, i.e. a binary label vector of dimension the same as the number of AUs under question. These labels are not necessarily independent of each other, although some are less correlated than others. This fact overseen by many, probably due to the added complexity, is not only useful for exploitation of the AU information but also on the detection side. AUs are indeed individual units that work in combination with a certain delay in time to construct what we perceive as a facial expression. Their combined dynamic evolution has been studied by some researchers, using learning models such as Dynamic Bayesian Networks (DBN) or Conditional Random Fields (CRF). Within the scope of this thesis we stay in the static case (for detection only) and apply a discriminant multi-label embedding scheme to learn bases for AU detection that promote closeness of commonly occurring AUs and separation of differently labeled data points on the projection space.

On the application side, driver monitoring systems keep getting more commonly encountered in our daily lives, mainly due to the emergence of automatic and semi-automatic vehicles. Inferring certain states of the driver (fatigue, distraction, emotional state) has the possibility to be used for various in-car systems that aim at augmenting

safety and comfort. To list a couple, cars could have alerting systems that notify the driver when he/she is in a state that will be dangerous for driving, or a semi-automatic vehicle can decide to switch between the automatic and manual driving modes depending on the driver's alertness level.

In this thesis, we address the problem of automatically detecting driver's cognitive distraction, which is a very complex cognitive state, especially in terms of facial repercussions. As the case with most affective and cognitive states, this specific type of distraction, which can also be called mind wandering, is a very personal experience in terms of the mechanisms lying behind. In addition, it has no one-to-one correspondence to any AU or AU combination in theory, as is the case for some of the emotions for example. These facts make the problem we try to tackle very challenging and explains the low number of similar works in the literature. The application that we present in this thesis is within the context of driving, yet the methodology is valid for any situation in general that involves cognitive distraction or similar complex states. Within this application we once again aim to make use of inter-AU relations, but this time in a dynamic manner and for another level of inference than AU detection. We build a hypothesis that not only the existence of certain AUs is informative for detecting cognitive distraction, but also different synchronization patterns between them.

In the following section, we detail how we approach all these issues with an emphasis on our contributions to the related literature in the field.

1.2 Major contributions and organization of the manuscript

We organize this thesis in four main parts, grouped in a way to represent the necessary background information and the three fundamental subjects on which we propose multiple contributions. The first part (Part I) consists of two background chapters. In **Chapter 2** we first review the different approaches proposed in order to define, categorize and understand the mechanisms behind emotions. We focus on three particular approaches, namely the basic emotions, dimensional and appraisal based approaches. These models are important within the context of this thesis, as they provide a theoretical rationale behind certain propositions of the thesis, if not are related directly to them. Then we introduce the Facial Action Coding System (FACS) and the Action Units (AUs) also giving their relations to the aforementioned emotion models. Needless to say, AUs form the skeleton of the rest of the dissertation and this chapter contains the necessary background information on them.

Chapter 3 contains the technical background on face tracking, feature extraction and machine learning methods used for facial analysis systems in general, and AU detection in particular. We specifically detail the methods which are used in latter parts of the thesis, and review other notable methods in the literature. We also provide a review of the existing facial expression databases and applications of facial expression analysis to complete this survey of the state-of-the-art in AU detection and facial expression recognition research.

In Part II we present our two contributions on individual AU detection, i.e. where

the inter-AU relations are not taken into account in the framework. In both contributions we investigate novel feature extraction methods, that are based on Local Binary Patterns (LBP), in order to achieve very high precision AU detection in video frames. Our first contribution, presented in **Chapter 4**, uses the edge-preserving bilateral filter and morphological opening and closing operations by reconstruction as a preprocessing method to LBP feature extraction in order to enhance certain appearance properties and eliminate irrelevant ones on the face image. We also introduce extracting LBP from overlapping windows of varying sizes, which results in an enriched representation of the facial texture. These advancements result in a significant performance increase in our tests using the well-established CK+ dataset of facial expressions (Lucey *et al.*, 2010). In **Chapter 5** we investigate the use of curvature Gabor wavelets within the AU detection problem once again in combination with LBP. The Gabor wavelets extracted in multiple orientations, scales, curvature degrees and filter sizes bring about a rich set of features that describe the facial texture in different levels. Comparing the different combinations of these filters we demonstrate the efficacy of the proposed system over existing ones. The very high accuracy results obtained show the effectiveness of the proposed novelty especially in dataset (or subject) specific applications.

Part III of the thesis, or the associated **Chapter 6**, explores the added-value of using the co-occurrence of AUs in natural (spontaneous) situations for their detection, as its main contribution. First, we propose a new real-time AU detection system that is based on Scale Invariant Feature Transform (SIFT) features, which are calculated must faster than the LBP based features with sufficient performance, allowing us to test extensions and enhancements to this base system. We first validate it on the CK+ dataset and then we extend this base system with a multi-label discriminant Laplacian embedding (DLE) scheme, to incorporate the AU co-occurrences in the training phase. This multi-label framework is validated on the FERA2015 challenge for AU detection (Valstar *et al.*, 2015). The proposed system is the winner of the AU occurrence detection sub-challenge, successfully detecting AUs in two databases of spontaneous expressions recorded during natural interactions. This chapter is particularly important, since it is one of the few works in the literature that use the joint-AU information in the AU detection learning phase. Another contribution is the first-time use of the DLE framework in the context of facial action recognition.

In Part IV, we present an application type of contribution that uses the AU information, obtained via the system presented in the preceding chapter, to detect cognitive distraction of drivers. Cognitive distraction is a complex mental state that does not necessarily manifest itself with a universal facial expression, as claimed for certain emotions. Yet, it is an important state that can alter one's coordination and decision making abilities and cause danger during driving. With this work we aim to examine the feasibility of its detection using novel feature extraction methods related to facial appearance changes. This part consists of two complementary chapters. In **Chapter 7** we present a new database, called the EPV-DIST, of 46 subjects recorded while driving a simulator and being induced visual and cognitive distraction at the same time via secondary tasks. Publicly available databases on driving are very rare and so are those annotated for distraction. With the EPV-DIST database, we aim to provide the research community a database that is recorded with a realistic setup that represents an in-vehicle configura-

tion and that contains segments annotated for the two types of distractive conditions and driving performance measures.

Then in **Chapter 8** we present our framework to detect sequences of cognitive distraction sequences in comparison to the baseline, where the drivers are not induced any kind of distraction. Our framework is based on dynamic low level descriptors of continuous AU values and the cross-correlations of AUs. The cross-correlations give information about the AU-synchronization characteristics at different time delays and as the experiments demonstrate are helpful in the discovery of facial behaviour during cognitive distraction. The AUs are detected on virtual frames that are generated through three camera views to represent a frontal one. We also propose methods to deal with the subject-biased data distribution and thus present a complete framework that can be integrated in cars for detection of cognitive distraction in real driving situations.

Finally, in **Chapter 9** we give a summary of the learnings from each contribution in these parts, discuss the strong and weak points and also give an outlook on how they can be improved, extended and used in real-world applications. The related publications for each chapter are provided in the corresponding overview sections at the beginning of each part.

Part I

Background

Introduction on Emotions and Facial Expressions

2

As individuals and social beings, we constantly receive internal and external stimuli that affect our decisions, determine our way of communication, cause physiological changes, influence our perception of the world and in the long term transform and build our personal character and social norms. All these transformations, in return, alter how we perceive and appraise future stimuli. Emotions are in the center of this constant loop as the driving force. They continually form the reason and consequence of our actions. They have influenced art, science, philosophy, sociology and other areas ever since they existed. Facial expressions, which constitute the core of this thesis, are one of the most significant means of communicating emotions, may it be voluntarily or involuntarily. In this introductory chapter, we move slightly away from the technical point of view and review the various definitions and categorizations of emotions, the link to facial expressions and describe the Facial Action Units, which are the building blocks of these expressions.

2.1 Emotions and emotion models

How to define what an emotion is and what it covers has been of substantial interest since the great ancient philosophers and still is a matter of discussion in philosophy, psychology, psychiatry and cognitive science among many other disciplines. Quoting from Izard (1969): *"The area of emotional experience and behavior is one of the most confused and ill-defined in psychology."* The word emotion is believed to have originated from the Latin verb *emovere*, which can be described as to move out, stir, or agitate. As a *safe* definition (although one could still debate this), emotion is a psycho-physiological mechanism that occurs as a consequence of internal and / or external stimuli and usually expresses itself in terms of physiological or biomechanical changes (motor expressions) in the body.

Many theories and models have been proposed to understand and define the meaning and nature of emotions. One of the most powerful and influential one is the Darwinian approach, which states that the same state of mind is expressed throughout the world with remarkable uniformity (Darwin, 1872). According to Darwin, emotions are the re-

sult of an evolutionary process and directly related to survival purposes, and so are the corresponding expressions. Following Darwin, researchers like Paul Ekman (Ekman & Keltner (1970)) and Caroll Izard (Izard (1969)) based their emotion models on a categorical approach that states that there are discrete emotions that serve unique purposes, are expressed in a similar fashion universally and are universally understood, or decoded. A less strictly distinct, yet still categorical model is the one of Plutchik, who built a model where every emotion can be generated by mixing eight primary ones (Plutchik (1980)).

Another approach is the dimensional one, which describes and categorizes emotions in terms of multiple continuous dimensions. The most commonly adopted one is Russell's circumplex model (Russell, 1980), which claims that emotions lies on a two dimensional space representing its positive or negative valence and arousal effect. Later, other dimensions have been added, such as control (or dominance). The third most important approach is the appraisal based models. The main and common theory of appraisal models is that emotions occur as a result of how one *appraises* a received stimulus, and is affected by many internal and external factors. The appraisal based approach may be said to focus more on the cognitive basis of the nature of emotions, rather than giving names or assigning dimensions to them. The main proposition of appraisal theories of emotion is that they are elicited and differentiated by the subjectvie interpretation of the personal significance of events (Sander & Scherer, 2009). In the rest of this section we detail the emotion models that have influenced this dissertation work, namely the Basic Emotions Model by Ekman 2.1.1, the 2-D representation by Russell 2.1.2 and the component process model of emotion based on appraisals by Scherer 2.1.3, also discussing the pros and cons of each one, particularly from an automatic facial expression recognition and affective computing point of view.

2.1.1 The Categorical Approach - Ekman's Basic Emotions Model

The categorical approaches, as mentioned earlier, aim at classifying emotions or their neurophysiological or anatomical productions (such as facial expressions) into a known number of categories that can exist independently of one another. These categories are called primary, fundamental or as more commonly used *basic* emotions. The main issues across the basic emotion theories are how many of them there are, which ones they are and which criteria make them basic. Indeed, there exist theories that claim there are only two (e.g. Happiness and Sadness in Weiner & Graham (1984), or Pain and Pleasure in Mowrer (1960)) while some identify 18 of them (Frijda (1986)). One could argue that the choice of words is critical in the choice of these emotions. For example, in many cultures the words anger, annoyance and rage would mean the same thing, yet different theorists include different wording for similar emotions in their categorization. Ortony & Turner (1990) provides an extensive review of basic emotion theories with the contradicting points between all these theories.

A criterion embraced by some theorists is that the basic emotions are psychologically primitive, in the sense that they cannot be decomposed into other emotions. For instance, *distress*, although it is included in the basic emotions according to some, can be explained as a state of anxiety and expectation, sometimes combined with shame and even rage, depending on the context. In addition, this is definitely not the sole descrip-

tion one could find for distress. While joy (or happiness) is a state of positive valence, experienced due to a factor giving pleasure. Describing it using words representing other emotions would be redundant at the best case. We notice, also that it is important to make the distinction between emotions and cognitive states, as also pointed out by Ortony & Turner (1990). Fear could be considered an emotion, for instance, while distraction or interest would be better called cognitive states.

The other criterion adopted by many (Izard (1969), Plutchik (1980)), Ekman *et al.* (1972)), and the one which we will be focusing on more in this section, is that for each basic emotion there exists biological evidence, that is generally assumed to be related to an evolutionary process. In particular, Paul Ekman proposed that an emotion has to have a distinctive facial expression that is universal, to be one of the basic ones (Ekman, 1984). Universality of a facial expression indicates that it is presented in a particular fashion, in the same way among everyone regardless of the age, gender, ethnicity or cultural background; also that it is genetically inherited due to human evolution and not learned, thus, once again, not related to one's social experiences, family or the norms acquired from the society. To find proof of his claims Ekman even conducted a study on a culturally isolated society in New Guinea and has concluded that they could recognize which emotions are represented by portrayals of six facial expressions Ekman (1980). According to Ekman there are six of those emotions: *Happiness*, *Surprise*, *Anger*, *Sadness*, *Fear* and *Disgust*. The facial expressions corresponding to these emotions can be seen in Fig. 2.1, which includes posed expressions that we have acquired by instructing the subjects. Later, Ekman also added *contempt* to the list and increased the number to seven.

For reasons of coherence, this model is revisited and the basic facial expressions are discussed in detail in Section 2.2, in relation with the AUs and the proposed Darwinian explanations. In this section, only a description of the emotions are provided:

- Happiness - Emotion associated to pleasurable, positive stimuli. It can be expressed with various other words, generally depending on intensity, such as joy, contentment, amusement or euphoria.
- Disgust - A negative emotion related to repulsion due to an unpleasurable stimulus. It is one of the easier emotions to naturally induce, for instance using visual or odor stimuli.
- Anger - Another negative emotion that involves high levels of arousal, or excitement. It is also associated with a tendency of aggression towards the source of the emotion, and could be classified in two as cold anger and hot anger (rage).
- Sadness - The emotion of despair, agony and grief due to an unpleasant and generally uncontrollable event.
- Fear - It is the emotion where one also feels a loss of control on the possible outcomes of a stimulus, the reason being either a learned unpleasant consequential event or not knowing the effects of an unfamiliar source.
- Surprise - Emotion felt due to an unexpected, novel event. The *primitiveness* of surprise is the one that is most open to discussion as it can take many forms and

(a) Happiness (b) Disgust (c) Anger

(d) Sadness (e) Fear (f) Surprise

Figure 2.1: Facial expressions corresponding the six basic emotions

is commonly accompanied by one of the other basic emotions. For example, the emotion felt after hearing good news that were unanticipated or winning a prize results in a state of happy-surprise, while unpleasant news lead to sad-surprise or unexpected .

One of the reasons that Ekman's model is so commonly used in affective computing is that it provides a basis with distinctive classes with well-defined rules, so it is suitable for human ratings, as well as automatic classification by machine learning. It is yet open to discussion, how sufficient these six (or seven) words are in terms of coverage and specificity. The model lacks many cognitive states that are quite important to affective computing applications, such as interest, engagement, shame, guilt, stress, distraction or pain. In addition, even if it is assumed that facial expressions are universal, words are not and in fact can have different meanings even to individuals possessing the same mother tongue. Ekman's model does not take into account variations and intensities of the basic emotions, which can indeed be expressed differently. Also Ortony & Turner (1990) argue that emotions should be affectively valenced, positively or negatively, which is not necessarily the case for surprise. He also argues that it may be true that facial expres-

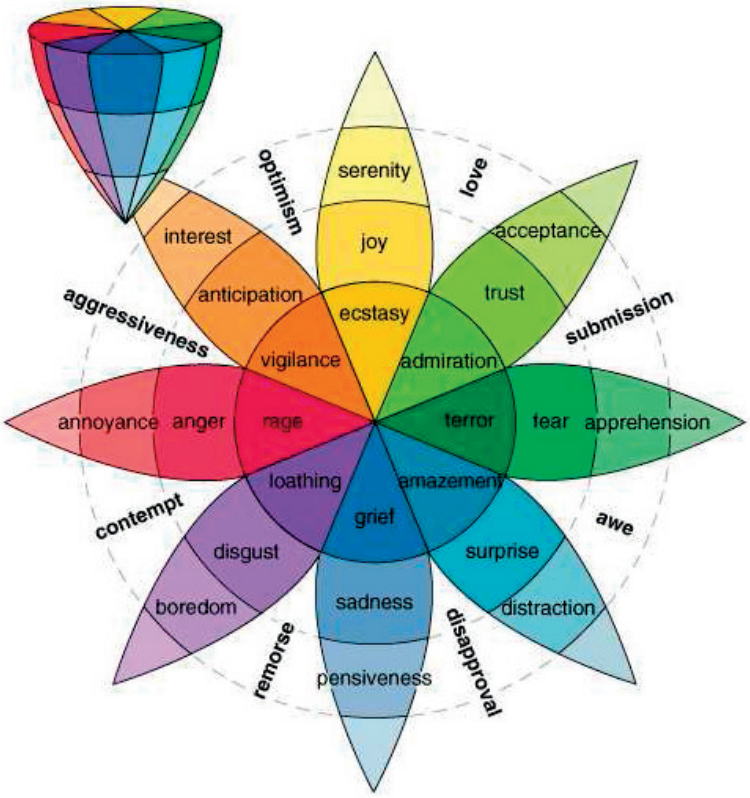


Figure 2.2: Plutchik's wheel of emotions

sions are *hardwired* (and genetically inherited) but not necessarily the corresponding emotions. In summary, although Ekman's model is *practical* in terms of classifying emotions, how adequate and efficient it is to restrict emotions (and the corresponding facial expressions) into six (or seven) categories is very much open to debate.

Another categorical model that contains the intensities and more classes is Plutchik's wheel of emotions (Plutchik, 1980). Plutchik suggests that basic emotions should have an evolutionary background related to survival, and thus proposes eight basic emotions. These emotions are acceptance, anger, anticipation, disgust, fear, joy, sadness as can be seen in Fig. 2.2. According to Plutchik these emotions can be explained in terms of survival purposes, as in fear being related to the need for protection or sadness to the need to maintain possession of a pleasurable object (Plutchik (1980), Ortony & Turner (1990)). Other emotions are generated from these basic ones, as represented in Fig. 2.2, where each of the eight primary colors represents a basic emotion and the intensities change with the color tone, representing other emotions.

In Plutchik's model, thanks to the conical shape and the color palette, the emotions that are closer to each other are placed closer and more emotions can be generated by their mixture, as in fear and trust leading to submission. Also, less intensified emotions are placed further apart and represented by lighter colors, implying they are less significant and harder to be detected, for example interest and serenity contrary to their

stronger counterparts vigilance and ecstasy, placed in the center.

Categorical models are useful in terms of practical applications, for instance when there is a discrete number of virtual emotional avatars that will be activated according to a user's emotion, or when a system is trained to detect for instance when a user is disgusted or happy about a product. However, they are quite limited especially because of language limitations. The direct correspondence with universal facial expressions also make them attractive, but the universality of those expressions is also questionable, as argued for instance in Jack *et al.* (2012), and on which we elaborate further in Sec. 2.2.

2.1.2 The Dimensional Approach

Another approach in defining emotions is to map them on several continuous dimensions that represent certain physiological, cognitive or characteristic projections of the subject. This continuous dimensions approach does not carry the key limitation of categorical models, that depend on a list of names of emotions of a finite number.

The most popular of these mappings is the one by Russell, where every emotion is assumed to exist in a 2-D space defined by the valence and arousal components (Russell (1980)). The valence dimension defines how positive or negative an emotion or an emotional stimulus is, while arousal defines the activation or excitement level. For example, anger of the basic emotions is highly negative in valence and also of high arousal, while a similar affect, impatience, is less negative and less activated. Fig. 2.3 shows the two-dimensional model with various emotions projected, with additional two dimensions that will be discussed further within the appraisal model in Sec. 2.1.3.

More dimensions have been proposed in addition to valence and arousal, as reviewed in Russell (1980). Dominance (or control) is a commonly adopted one. The dominance dimension defines the amount of power that the subject has. It is useful to differentiate affective states that are similar in both valence and arousal. For example, anger and fear are both emotions with high arousal and highly negative valence, but fear involves less control than anger.

The valence arousal model (and variations) are very useful, both in defining affective states and also in terms of affective computing. It intrinsically contains the intensity of the emotion, in addition to the liberty of not being obliged to giving an exact name to it. It is efficient when subjects are asked to evaluate how they feel or when they are asked how someone is feeling by observations, for example using the Self Assessment Mannekin (SAM) (Morris, 1995) or an annotation system using a knob that is continuous over time and affective dimensions. The automatic recognition of these dimensions, therefore, usually requires regression instead of classification. The drawback of the (two or three) dimensional model is that, it lacks in explaining the contextual factors that give rise to an emotion, which may commonly be varying among individuals. Also, in terms of automatic facial expression analysis, the arousal dimension especially is very difficult to be recognized since there is no facial expression that is common in highly aroused and lowly aroused affective states.

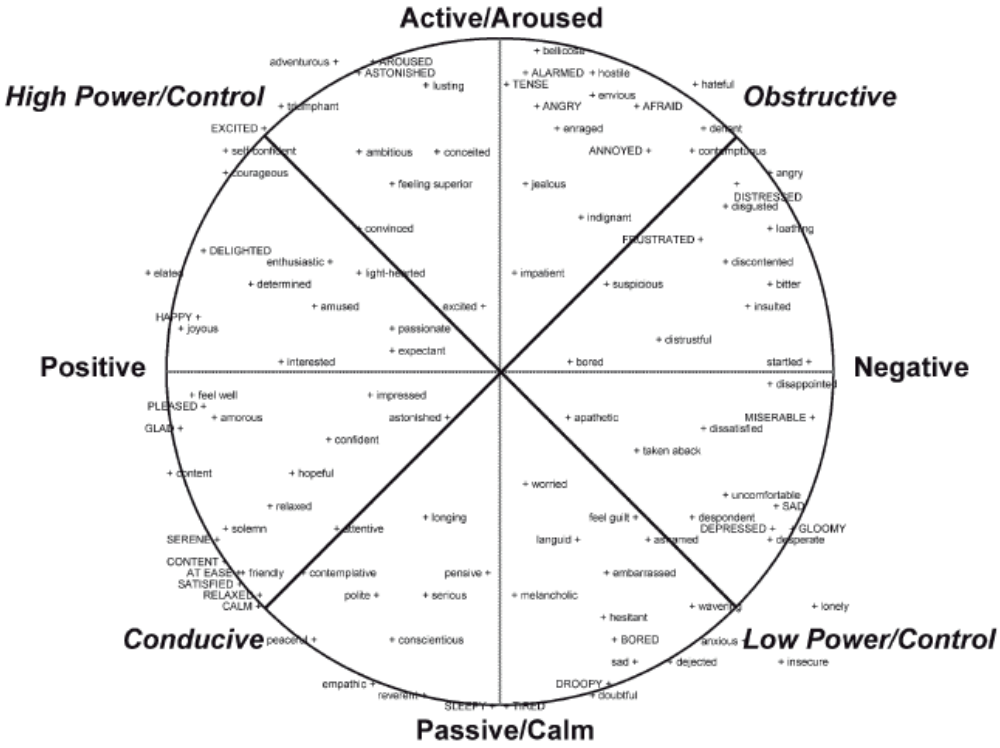


Figure 2.3: A dimensional representation of the combination of the three kinds of emotion models (Scherer (2005))

2.1.3 Appraisal Based Approaches and the Component Process Model of Emotion

Appraisal related theories of emotion aim at defining it in terms of how one appraises, or evaluates an event depending on several internal and external factors. It can be said that they are more interested in the cognitive processes taking place during and after these evaluations and not necessarily giving names or classifying the whole emotional experience.

Magda Arnold and Richard Lazarus were two of the first theorists to support a model based on an evaluation of events (Arnold (1960), Lazarus (1991)). Following their lead Ortony, Close and Collins proposed the OCC model, where emotions are defined as valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is construed (Ortony *et al.* (1990)). The three main components founding an emotion are therefore the type of the stimulus, the affected agent (usually self) and the valence component (positive or negative).

Klaus Scherer adopted a similar yet more detailed approach and proposed the Component Process Model (CPM) of emotions which is a dynamic model that involves a sequence of evaluations and the consequential responses in several organismic subsystems (Scherer, 1987). The CPM defines emotion as an episode of interrelated, synchronized changes in the states of the organismic subsystems in response to the evaluation

Table 2.1: Components of emotions in the CPM with the corresponding functions and organismic subsystems

Emotion Function	Emotion Component	Organismic Subsystem
Evaluation of objects & events	Cognitive	Information Processing (CNS)
System regulation	Peripheral efference	Support (CNS,NES,ANS)
Preparation and direction of action	Motivational	Executive (CNS)
Communication of reaction and and behabioral intention	Motor expression	Action (SNS)
Monitoring of internal state and organisim-environment interaction	Subjective feeling	Monitor (CNS)

of an external or internal stimulus event. The synchronization of changes in multiple physical and cognitive components is an important proposition of the CPM. According to Scherer there are five components of an emotion corresponding to five distinctive functions (Scherer, 2001). Table 2.1, adapted from Scherer (2001) shows these components along with the corresponding organismic subsystems that subserve them: Central Nervous System (CNS), Neuro-Endochrine System (NES), Autonomic Nervous System (ANS) and Somatic Nervous System (SNS).

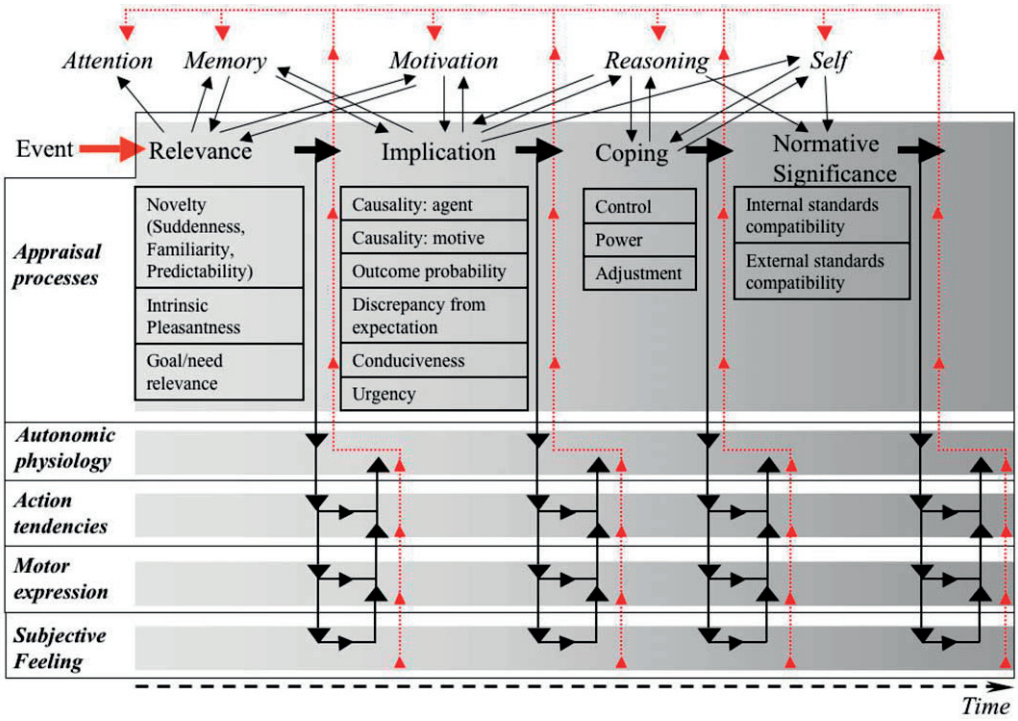


Figure 2.4: An illustration of the component process model of emotion (Sander *et al.* (2005))

Scherer suggests, as an important component of his theory, a set of criteria called the Stimulus Evaluation Check (SEC) (Scherer, 2001). According to the CPM, as an

event unfolds, the individual concerned would evaluate its significance based on these criteria, in a sequential (or causal) manner Scherer (1999). The SEC are grouped in five according to their function as seen in Fig. 2.4, and the definitions of each one are as follows (taking as reference Scherer (2001) and Sander *et al.* (2005)):

- Relevance detection checks: Evaluation of whether an object or event is relevant according to:
  - Novelty Check: (i) Suddenness or abruptness, produces orientation response (ii) Familiarity; based on schema matching (iii) Predictability; based on past observations;
  - Intrinsic pleasantness: Likelihood to result in pleasure or pain; leads to approach or withdrawal
  - Goal-Need relevance: The importance for one's goals or needs
- Implication assessment: A check of causes and possible outcomes of an event:
  - Causal attribution check: Who and why
  - Outcome probability check: Likelihood of consequences
  - Discrepancy from experience check: Consistent or discrepant with individual's expectations
  - Goal / Need conduciveness check: Conduciveness or obstructiveness, leading to positive or negative emotions
  - Urgency check: Depends on goals/needs and temporal consequences, leads to action readiness
- Coping potential determination: Assessment of one's abilities and power related to:
  - Control: Check whether an event or its outcomes can be influenced
  - Power : Check ability to change contingencies
  - Potential for adjustment : Ability to live with and accommodate to the effects
- Normative significance evaluation: Evaluating the event in terms of learned internal and external norms
  - Internal standards check: Self-ideal and internal modal code
  - External standards check : Perceived norms or demands of a reference group

The CPM can be thought of mapping the affective states onto dimensions as well, but this time a larger number that involves the reasons of the stimulating event, the possible outcomes and individual differences in appraising it. It is also inline with previous theories of emotions, such as the valence-arousal-dominance dimensional model. The valence dimension is directly related to the goal / need conduciveness check, arousal to urgency and control to coping potential (Scherer, 2001). Some of the components are

associated with the basic emotions of Ekman. For example, a high novelty would yield the same expressions of surprise or a goal obstructive evaluation is a part of sadness. The theories on related facial expressions for some SEC are reviewed and discussed in the following section (Sec. 2.2). Fig. 2.3 is an illustration of such a mapping scheme.

A very important drawback, which will hopefully be tackled with more experiments in the future, is the lack of a well-defined specific facial expression for each component. A facial expression, or action, can imply many of the appraisals, and it is very difficult to manipulate a specific appraisal using special stimuli. There are on-going studies using facial expressions (and electromyography (EMG) measurements), physiological signals and brain imaging in order to identify what happens when each specific appraisal is activated. One of the few recent studies, for example, has used EMG signals to measure facial activity during manipulations in a gambling task and obtained evidence for the difference facial actions occurring during goal conduciveness and power related appraisals, in addition to their causality and accumulation (Gentsch *et al.*, 2015). Despite these difficulties in experimentation, CPM is arguably the most complete emotional model, that also involves a causal reasoning of how and why an affective state is present.

2.2 Facial Display of Emotions and Facial Action Units

Facial expressions are the most powerful non-verbal way of communicating emotions and contains a lot of information on one's affective, mental and cognitive state. Fig. 2.5 shows examples of facial expression depictions in artworks from the early ages of human civilization and during and after the renaissance. Facial expressions have always been important to humans, yet the first scientific studies were not until the 19th century. Duchenne was the first to analyze the mechanism behind facial expressions using electrical stimulation of the facial muscles on corpses and living bodies of humans (Duchenne, 1876). The most intriguing work was undoubtedly by Charles Darwin in his progressive book *"The expression of emotions in man and animals"* (Darwin, 1872). It is considered as his most revolutionary work after the publication of *"On the Origin of Species by Means of Natural Selection"* in 1859 and *"The Descent of Men, and Selection in Relation to Sex"* in 1871. What made Darwin's work special was that he was the first one to ask the question "Why?" in addition to what and how. As in "Why is that specific muscle contracted when people feel happy or sad or afraid?". The book is also distinguished since it was one of the first scientific works ever to use photographs, and is seen as a milestone in terms of the usage of visuals in publications.

Darwin observed many expressions that humans perform, find counterparts that are observed in various animal species and proposed evolutionary explanations of these expression mechanisms. We will refer to his theories for each kind of expression in Sec. 2.2.2 also in relation to the emotion models described in 2.1, but first we will introduce the Facial Action Coding System (FACS), which defines the Action Unit (AU), which are at the core of this dissertation.

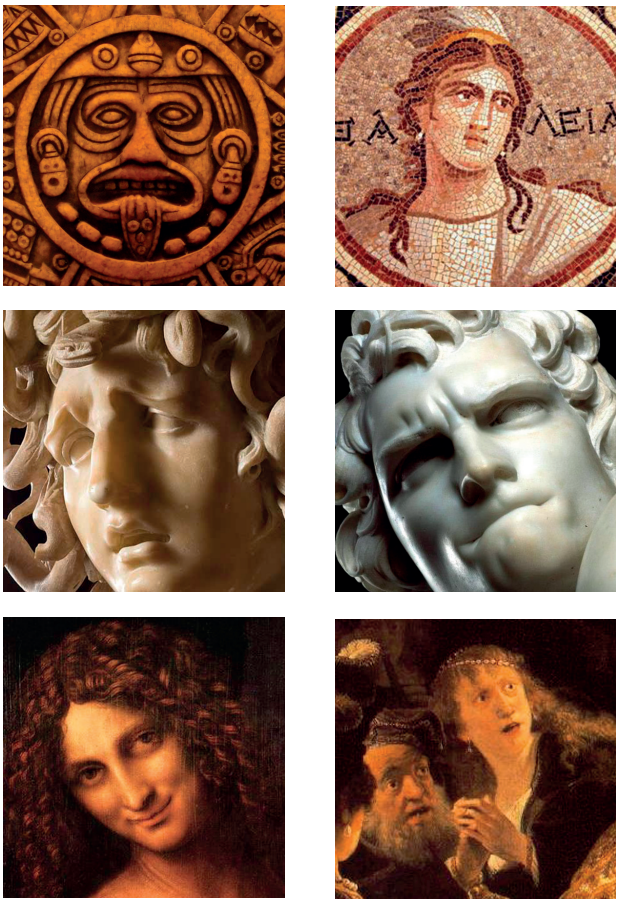


Figure 2.5: Depictions of facial expressions in various forms of art

2.2.1 Facial Action Coding System

The Facial Action Coding System (FACS) was developed by Ekman and Friesen (Ekman & Friesen, 1977) and revised later in 2002 to achieve its final form (Ekman *et al.*, 2002). Previous attempts trying to define and classify facial actions were based on observer's judgments and limited to the taxonomy that was available (see Sec. 2.1). FACS, on the other hand, aims to objectively qualify and quantify every possible facial movement in an objective way as it depends solely on the appearance outcome of muscle movements (their consequent effects on the visible layers of the face) and does not involve any mapping to classes or dimensions related to emotions or other phenomena.

On the human face it is estimated to exist 43 muscles (depending on what is counted as a separate muscle and where the face starts and ends); these are also called mimetic muscles and most of them are directly or indirectly responsible for the various types of facial expressions that we observe. Fig. 2.6 show illustrations of these muscles from a frontal and side view of the face. The individual or grouped contraction or relaxation of these muscles form what is called Action Units (AUs). AUs are the measurement units of FACS. There are 46 AUs which describe the facial movements. Some of these have been

removed later due to redundancy or classified as Action Descriptors (ADs) due to non-existence of a muscular basis. There exist 14 additional AUs which describe the gaze directions and head movements in the three dimensions (yaw, pitch and roll). These AUs are different from the 46 facial ones in terms of how they are defined and they describe actions in a coarser manner. Table 2.2 shows the facial AUs with the corresponding definitions and related muscle movements and Table 2.3 shows the 14 AUs for defining the gaze and head movements. The ADs, such as AD19 (tongue out) or AD 30 (jaw sideways), are not shown in Table 2.2 since they are less precisely defined in terms of muscle movements, as in . The full-list may be found in the FACS manual (Ekman *et al.*, 2002). In the rest of the dissertation what we refer to as simply AUs are the facial AUs, excluding the 14 for head and gaze directions and the ADs.

It can be seen from Table 2.2 that while a single muscle can be responsible for multiple AUs (as in AUs 22, 23, 24 and 28 by the Orbicularis Oris muscle), the contraction / relaxation of multiple muscles can be necessary for a single AU (e.g. AUs 9 and 10), as well. It should be noted that there is no one-to-one correspondence between AUs and facial muscles, and that the AUs are not unitary muscle movements. The eye brow raising movement, for example, is divided in two as inner (AU1) and outer (AU2) yet are controlled by different parts of the same muscle, frontalis. Fig. 2.7 shows an example for each AU that is mentioned in this thesis and included in at least one of the AU detection systems proposed, in total 21 AUs. While most of the AU combinations are additive (in terms of the change in appearance) there are some cases where the combination of AUs result in a different appearance than the addition of the involved AUs. The AUs account for any possible facial expression which may practically reach a few hundred combinations, and there are few AUs that cannot occur simultaneously. These AUs are called antagonistic AUs, as in the case of AUs that involve mouth opening vs. lips pressed (Ekman & Friesen, 1976).

FACS also includes intensity annotations of the AUs. The intensities are defined as A - Trace, B - Slight, C - Marked, D - Severe and E - Maximum, and were introduced in the revised version (Ekman *et al.*, 2002). For example a face image annotated as 1C - 2A -6B - 12C means the inner eye brows are visibly raised with a trace of raise in the outer eye brows accompanied with a pronounced lip pull motion (what might be called a smile) and a slight raise of the cheek. AUs provide a means of objectively defining what is present on the face depending only on the appearance changes and independently of the underlying reason or emotion causing the facial changes.

FACS annotation requires an intense training to obtain a reliable annotation and in most cases it might be very difficult to differentiate which actions are present and at what level. In practice a change of appearance might resemble multiple actions. The same issue is also present in automatic AU detection systems. For example, it is quite difficult to differentiate between the actions AU 25, 26 and 27 using automatic detection methods. All three are related to a mouth opening action, and the difference is sometimes even invisible to the human eye. For this reason in the revised version of FACS (Ekman *et al.*, 2002) the annotation for AUs 26 and 27 have been revised and are always accompanied by AU25, if the lips part. Similarly, AUs 41 (lid droop), 42 (eye slit) and 44 (eye squint) are re-assigned to intensities of AU43 (eyes closed).

A similar problem exists between AUs 12 and 13. Although different muscles are

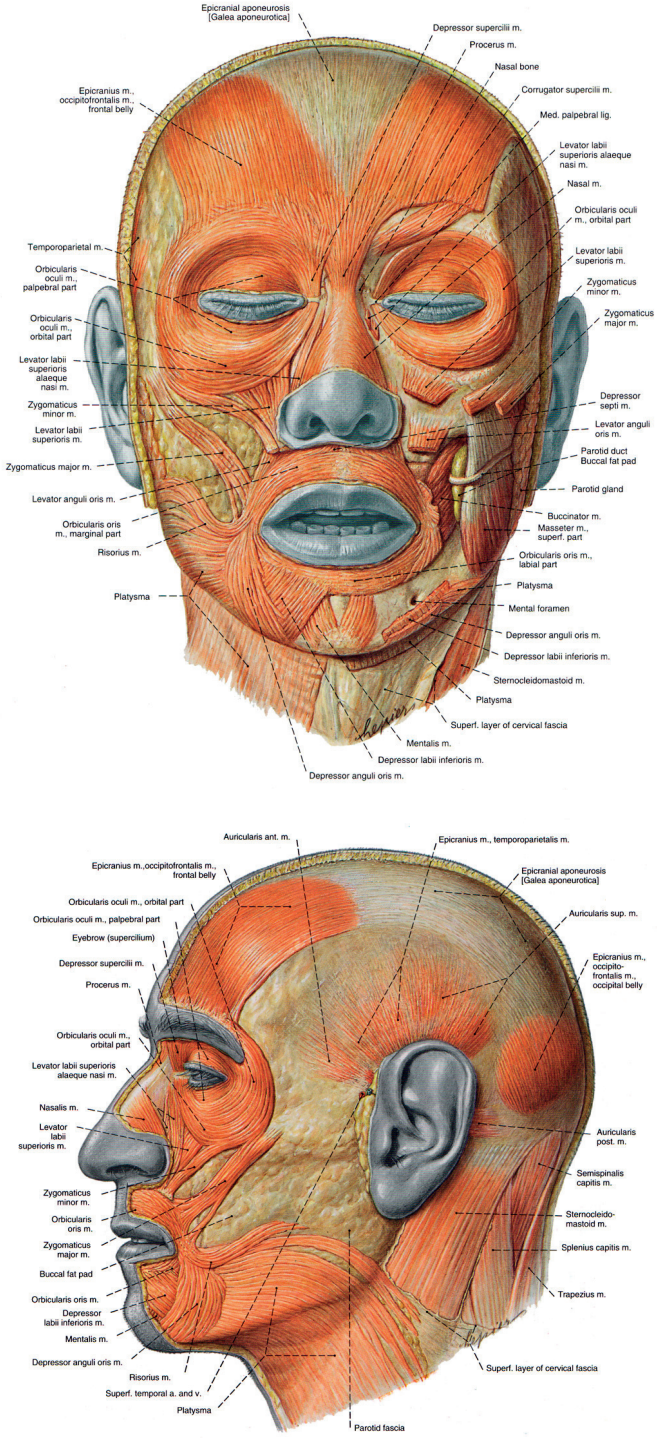


Figure 2.6: Drawing of the facial muscles responsible for facial actions (Sobotta & Figge (1974))

Table 2.2: AUs, FACS definitions and the facial muscles responsible for their activity

AU	Definition	Related Muscles
1	Inner Brow Raiser	Frontalis (Medial)
2	Outer Brow Raiser	Frontalis (Lateral)
4	Brow Lowerer	Depressor Glabellae; Depressor Supercilli; Corrugator
5	Upper Lid Raiser	Levator Palpebrae Superioris
6	Cheek Raiser	Orbicularis Oculi (Orbital)
7	Lid Tightener	Orbicularis Oculi (Palebral)
8	Lips Toward Each Other	Orbicularis Oris
9	Nose Wrinkler	Levator Labii Superioris, Alaeque Nasi
10	Upper Lip Raiser	Levator Labii Superioris, Caput Infraorbitalis
11	Nasolabial Furrow Deepener	Zygomatic Minor
12	Lip Corner Puller	Zygomatic Major
13	Cheek Puffer	Caninus
14	Dimpler	Buccinnator
15	Lip Corner Depressor	Triangularis
16	Lower Lip Depressor Depressor	Labii
17	Chin Raiser	Mentalis
18	Lip Puckerer	Incisivii Labii Superioris & Inferioris
20	Lip Stretcher	Risorius
22	Lip Funneler	Orbicularis Oris
23	Lip Tightener	Orbicularis Oris
24	Lip Pressor	Orbicularis Oris
25	Lips Part	Depressor Labii, or Mentalis or Orbicularis Oris Relaxation
26	Jaw Drop	Masetter; Temporal and Internal Pterygoid Relaxed
27	Mouth Stretch	Pterygoids; Digastric
28	Lip Suck	Orbicularis Oris
38	Nostril Dilator	Nasalis, Pars Alaris
39	Nostril Compressor	Nasalis, Pars Transversa and Depressor Septi Nasi
41	Lid Droop	Levator Palpebrae Superioris Relaxation
42	Slit	Orbicularis Oculi
43	Eyes Closed	Levator Palpebrae Superioris Relaxation
44	Squint	Orbicularis Oculi, Pars Palpebralis
45	Blink	Levator Palpebrae Relax.; Orbicularis Oculi, Pars Palpebralis
46	Wink	Orbicularis Oculi

contracted during the Lip Corner Puller action (AU12 - zygomaticus major) and Cheek Puffer (AU13 - caninus) the difference in facial appearance is not very significant, and therefore AU13 is usually annotated and / or detected as a high intensity AU12. Also, sometimes in order to accurately (manually or automatically) annotate a face one might require the neutral face of the same person, that is with no AUs present. This is because of the person-dependent appearance differences, for example due to naturally higher eye-brows resembling AU1 or AU2 or a downward shaped mouth, resembling AU15. FACS is therefore more efficient in annotating sequences of images rather than single frames. These issues are discussed further in the following sections while reviewing the weaknesses and robustness of the detection systems presented.

Another issue is what is called *competitive* AUs. These AUs might occur at the same time and the appearance change might be dominated by one of the AUs. For instance, a dimpler (AU14) and a cheek puller (AU12) frequently occur simultaneously, yet the AU14 might only be visible from a different view than the frontal one, so it is annotated

Table 2.3: Head and gaze movement related AUs

AU	Definition
51	Head turn left
52	Head turn right
53	Head up
54	Head down
55	Head tilt left
56	Head tilt right
57	Head forward
58	Head back
61	Eyes turn left
62	Eyes turn right
63	Eyes up
64	Eyes down
65	Walleye
66	Cross-eye

as only AU12. The same is true for some mouth actions and Ekman and Friesen admit themselves also that the lower face related actions may not be fully comprehensive as there probably exists an infinite number of possible actions (Ekman & Friesen, 1976). Despite these minor issues, there is no doubt that FACS is the most exhaustive measurement system of facial actions, allows very little subjective interpretation and is very useful for practical applications as well as psychology and neuroscience research (see Sec. 3.5 for example applications).

Temporals of Facial Actions

The evolution of facial expressions and AUs is not instantaneous but a rather dynamic process. The dynamic properties of facial actions are important in discriminating facial configurations with similar appearance but different psychological and cognitive meaning. The temporal features such as the duration, speed of the phases of AUs can be used in applications in recognizing rather complex cognitive states such as pain (Ekman & Rosenberg (1997), Bartlett *et al.* (2014) or distinguishing between posed and spontaneous expressions (Hess & Kleck (1997), Cohn & Schmidt (2004), Valstar *et al.* (2007)).

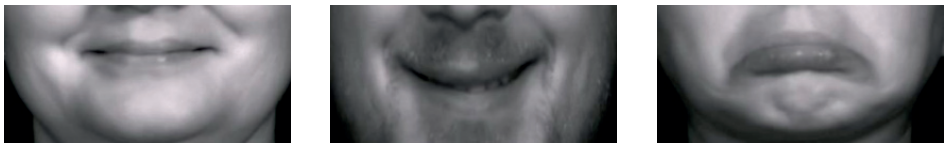
There are four main phases of a facial action. The *onset* is when the related muscles begin contracting (or relaxing depending on the type of action), *apex* is when the action reaches its peak intensity, *offset* is when the muscles begin relaxing and the face begins to return to its original state and finally the *neutral* phase is when there is no more evidence for the particular action. Fig 2.8 shows an example sequence of AU12 activity in time. The sequence is from the MMI database of facial expressions (Valstar & Pantic, 2010), which has been annotated for the existence of AUs and the beginnings of their onset, apex and offset phases.

The intensity plot in Fig. 2.8 is, of course, only representative and the dynamics of

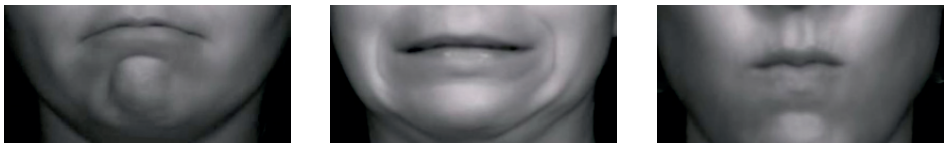
(a) AU1 - Inner Brow Raiser (b) AU2 - Outer Brow Raiser (c) AU4 - Brow Lowerer

(d) AU5 - Upper Lid Raiser (e) AU6 - Cheek Raiser (f) AU7 - Lid Tightener

(g) AU9 - Nose Wrinkler (h) AU10 - Upper Lip Raiser (i) AU11 - Nasolabial Deepener

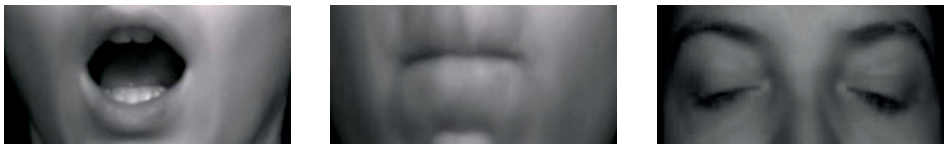


(j) AU12 - Lip Corner Puller (k) AU14 - Dimpler (l) AU15 - Lip Corner Depressor



(m) AU17 - Chin Raiser (n) AU20 - Lip Stretcher (o) AU23 - Lip Tightener

(p) AU24 - Lip Pressor (q) AU25 - Lips Part (r) AU26 - Jaw Drop



(s) AU27 - Mouth Stretch (t) AU28 - Lip Suck (u) AU45 - Blink

Figure 2.7: AUs according to the FACS

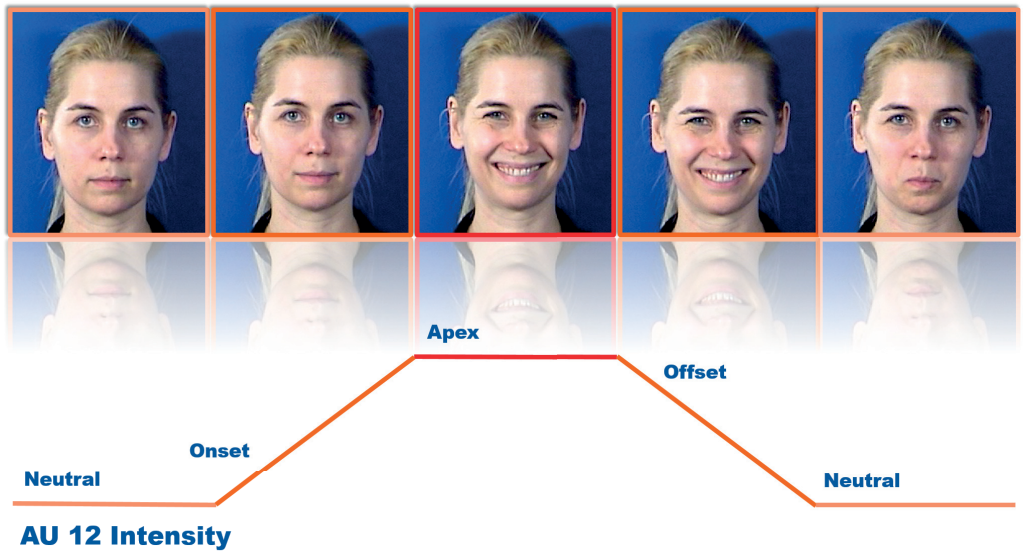


Figure 2.8: The dynamics of an example AU12 sequence from the MMI database. The sequence starts with a neutral phase, the action starts with the **onset**, reaches its peak at the **apex**, starts decreasing in intensity at the **offset**, and again reaches the neutral state.

facial actions are rarely that symmetric. In fact, it might happen that multiple apices are present along a sequence, that is a second peak of intensity with a precedent offset phase that never reaches the neutral. It is also important to notice that annotating the temporal phase of an action based on a single frame is almost impossible based on a single frame, and the preceding and following frames are also required. As stated earlier, these phases and the transitions between them contain important information. For instance, Cohn & Schmidt (2004) have shown that spontaneous smiles have a longer onset phase compared to posed smiles and can contain multiple AU12 apices; or as demonstrated by Valstar & Pantic (2006), spontaneous and deliberately displayed eye brow actions (AUs 1,2 and 4) differ in the duration, speed of onset and offset phases and in the order and timing of the actions.

2.2.2 Expressions of emotion models

Certain facial expressions are assumed to belong to certain emotions, and similarly certain emotions are thought to evoke specific expressions by certain theorists, including Darwin himself (Darwin (1872), Ekman (1989)). Although there are debates and questions on the validity of both correspondences, we list in this section those expression assumptions in terms of AUs in addition to those corresponding to certain components in the appraisal based model explained in Sec. 2.1.

Table 2.4 shows a list of AUs associated to the six basic emotions and some components of the CPM, with some of these AU combinations shown in Fig. 2.9 on images from the CK+ database (Kanade *et al.*, 2000). The AU - appraisal correspondences are mainly based on Kaiser & Wehrle (2001), where the authors review works that have stud-

Table 2.4: AUs related to the six basic expressions and some appraisal components

Expression	Related AUs
Happiness	6, 12, 25
Anger	4, 7, 10, 17, 22, 23, 24, 25
Sadness	1, 4, 6, 11, 15, 17
Disgust	9, 10, 16, 17, 25, 26
Surprise	1, 2, 5, 25, 26, 27
Fear	1, 2, 4, 5, 20, 25, 26, 27
Appraisal	
Novelty	
(High) Suddenness	1, 2, 5, 25
(Low) Familiarity	4,7
(High) Intrinsic pleasantness	6 , 12
Goal conduciveness	
Obstructive	4, 7, 17, 23
Conducive	6, 12
(Low) control	1, 15, 41
(Low) Power	20, 26
(High) Power	17, 24
(Low) External norms	10

ied evidences of facial expressions of appraisals, and Scherer (1993). In CPM, contrary to Ekman's approach, facial expressions, as well as other modalities (e.g. physiological or vocal changes), follow a causal pattern with response to an emotional stimulus that depends on the individual's evaluation. According to Darwin's theory some expressions might be explained as serviceable, as in they actually serve for a momentary or long-term purpose, while some arise from his principle of *antithesis* as opposite in nature to a serviceable action that serves no purpose and some carry only the purpose of a nervous discharge (Darwin (1872)). For example, the AU5 action (eyes wide open, see Fig. 2.9a) is thought to be related to action tendency, as in an orientation towards an unexpected or novel stimulus (Sander & Scherer, 2009). Similarly, the eye-brow raising actions aim at increasing the field of vision, and the ability to move the eyeballs easily in any direction (Darwin (1872), Hess & Thibault (2009)). However, shrugging shoulders, for example, has no explicable purpose and is theorized as an antithesis action.

In particular, the expression associated to the happiness emotion, or high intrinsic pleasantness and goal conducive appraisals of an event, or positively valenced states is a smile, generally accompanied by a cheek-puller action and the appearance of crow-feet wrinkles around the eyes (Fig. 2.9b). This is also called a *Duchenne smile* and is considered a sign of genuine enjoyment. Smiles are not particular to a positiveley valenced emotion though, the underlying reasons may differ considering the social context, e.g. they are commonly expressed as a sign of dominance (or control). Genuine expressions of happiness or amusement on the other hand are more reflexive and do not carry a purpose in the short-term, in an evolutionary point of view.

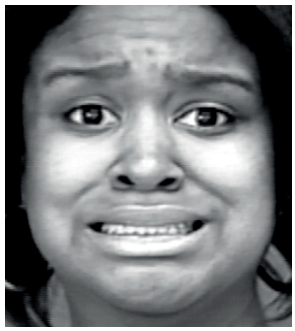
The expression of anger involves the contraction of the corrugator muscle (AU4, or a frown) and is what resembles a distressed state (Fig. 2.9c). A similar expression is

(a) 1 + 2 + 5 + 25 + 27 *

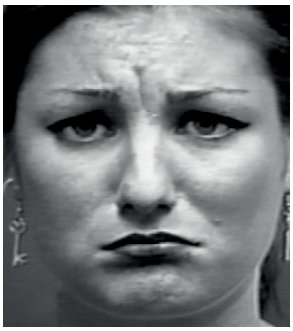
(b) 6 + 12 + 25 *

(c) 4 + 7 + 17 + 23 + 24 *

(d) 4 + 7 + 10 + 16 + 25



(e) 1 + 4 + 5 + 12 + 20 + 25 *



(f) 1 + 4 + 15 +17 *

(g) 4 + 9 + 10 + 25 *

Figure 2.9: AU combinations related to the basic emotions and appraisal components. (* Images are from the CK+ database)

hypothesized for the low familiarity and a goal obstructive evaluation of an event in the CPM, which also causes a negatively valenced and possibly high arousal state, as also confirmed in a study by Van Reekum (2000). An example is how people lower their eye-brows also when they do not understand or they are concentrated on something, which presents a cognitive difficulty. Also tightened lips displaying the teeth is a very common anger expression (Fig. 2.9d), which can be interpreted as a sign of aggression, thus action tendency and an attack response. Tightened closed lips, on the other hand,

is also related to an assessment of high control over the emotional event in addition to anger.

A particular expression that is common to fear and sadness is the combined action of AU1 + AU4, which is very difficult to perform deliberately (to pose), therefore it is a strong and reliable indicator of sadness or fear (or the goal obstructive SEC in CPM). The lower face actions during fear (AU20 & AU25, see Fig. 2.9e) are indicators of a withdrawal action in response to a threat. Sadness, on the other hand, is typically expressed with lowered mouth corners (AU15) and a raised chin (AU17) (Fig. 2.9f), also a sign of low control. In a Darwinian perspective, sadness expression can be interpreted as a socially adaptive behavior, signaling others of the discontent, and to elicit sympathetic responses in observers. Finally, the disgust expression involves wrinkling of the nose (AU9) and raising the upper lip (AU10) (Fig. 2.9g), this expression is encountered when a situation or event does not comply with one's self or social norms as explained by CPM. Darwin explains the disgust expression as part of a vomiting response.

We have listed, in this section, the facial actions that are related to some categorical emotions, with explanations in terms of emotional dimensions and with correspondence to the SECs in Scherer's CPM. As can already be seen, the different emotional models involve variations also in their correspondences to facial expressions. However, they all have in common the concept of a resulting facial muscle movement in common, which leads us once again to AUs and their added value in terms of emotion theories. In this thesis, we mainly focus on AU detection methods with certain references to how they apply within these different emotion models.

2.3 Conclusion

In this chapter we have provided an introduction to the concepts that are used or addressed in the dissertation. In particular, in Sec. 2.1 we have reviewed the various definitions for emotions and the models that are used to classify, differentiate and conceptualize emotions. We detailed Ekman's basic emotions model, the valence-arousal dimensional model and Scherer's component process model of emotions. We have also discussed their strong and weak points, also in terms of an affective computing perspective.

In the second part (Sec. 2.2) we have moved on to the facial display of emotions, summarized why they are important and explained the FACS, which defines the primary units of facial motions, AUs, the automatic detection of which this dissertation focuses on. We explain each AU in detail with the corresponding anatomical basis, then describe the temporal components of facial actions. Finally, we reviewed the link between facial expressions and the two emotion models; we have explained each expression corresponding to the six basic emotions as well as certain appraisals of the CPM. We have also provided the theories on the evolutionary basis of each. In general, we have tried to demonstrate the importance of the concepts and issues that lie at the heart of the rest of the thesis.

Tools for Facial Action Unit Detection and Expression Recognition

3

The concept of *Affective Computing* has entered our lives not before the 1990's, and since then it has constituted an important research and application area as a bridge between computer science, signal processing, wearable device technology, psychology and neuroscience among many others. It is defined by Rosalind Picard as computing that relates to, arises from, or influences emotions in her famous work, that is considered to have introduced the term (Picard, 1997). Affective computing can be performed using three main modalities: audio, physiological or visual signals, that are also commonly used in combination with each other, in a multimodal fashion. Audio cues, that are used in affective computing, can be verbal (the actual words) as well as non-verbal, as in features of the voice such as the intonation, or the speed and duration of the speech and its segments. Some physiological signals related to emotions are the electroencephalography (EEG), electrocardiography (ECG), Galvanic Skin Response (GSR), electrodermal activity (EDA), heart rate (HR), breathing rate or even brain imaging techniques such as the functional Magnetic Resonance Imaging (fMRI). Visual cues, on the other hand, may involve the pose, displacement and gestures of the various body parts (or the whole body) and sometimes the identity of the person involved.

Among those, facial cues are arguably the richest and most effective sources of information in terms of affective computing. In terms of real-world applications, facial information is more continuously accessible compared to audio cues, for instance, which are absent when the person is not speaking or producing a sound. It is non-invasive, as opposed to physiological signals, which require special equipment attached to the body and whose performance is easily affected by noise. Finally, it is less subject-dependent (nearly all humans perform the same set of AUs, for example) and contains more affect related material compared to gestures of other body parts. Automatic Facial Expression Analysis (AFEA), therefore, is an essential branch, or component, of affective computing as well as the more recent areas of Socially-Aware Computing (Pentland (2005)) and Social Signal Processing (SSP) (Pantic *et al.* (2011)), which aim at understanding and modeling social interactions and making use of this ability in Human-Computer Interaction (HCI) systems.

In this chapter we review the existing methods, tools, performance measures involved in and databases and applications related to AFEA systems. Note that, facial analysis systems (including face recognition) can be performed on 2-dimensional (2D) or 3-dimensional (3D) data; 2D being the grey-level or RGB images and 3D meaning additional depth information. Since the contributions in this dissertation are applied on 2D data only, we mainly include the state-of-the-art in 2D face analysis systems. An AFEA system consists of three main parts: Face acquisition (detection and/or tracking), facial feature extraction and classification according to the task (Figure 3.1). The term *facial expression analysis* comprises the differentiation of a facial image according to all categorizations or dimensional projections previously listed in Chapter 2, i.e. basic emotions, valence-arousal dimensions, SEC and AU presence. They involve the same steps in terms of acquisition and feature extraction and differ mainly in the final classification stage.

The face acquisition step refers to the automatic localization of either the whole face or certain landmarks on it. It can further be divided in two as detection and tracking; detection referring to frame-wise independent localization and tracking to making use of the position inferred on one or more precedent frames. The next step is extracting features that are relevant to the facial expression, which can be of two types: geometric or appearance (texture) based. The final step is obtaining a decision (e.g. on which AUs are present) using these features and by means of classification, clustering or regression if the desired output is of continuous form. An optional step is dimension reduction or feature selection, whose explanation is also included in this final phase within this thesis as they are frequently directly linked to the classification scheme and / or use similar techniques.

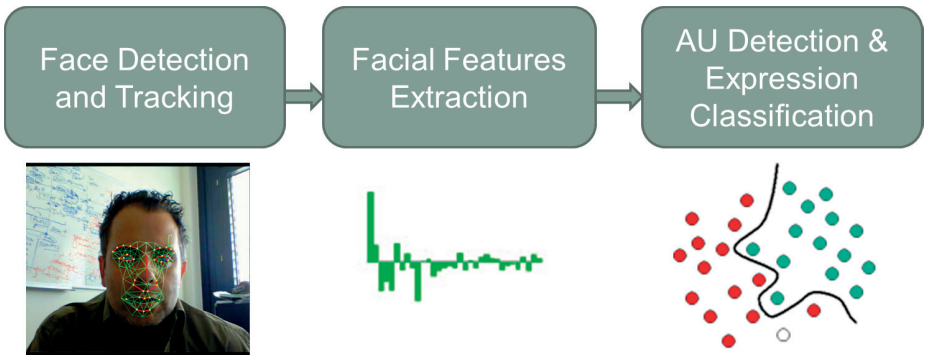


Figure 3.1: The three steps of an automatic FER or AU detection system.

All of these components involve a training phase (to learn the parameters or to learn the class separation), for which a (great) number of data with ground-truth is needed. The ground-truth information is obtained by manual or semi-automatic methods, as for example in the FACS annotations explained in Sec. 2.2.1. Thanks to the research effort by the facial analysis community, many databases have been published with annotations available for either the basic emotions, two (or more) continuous emotion dimensions or AU presence, which can be used for training, validation and testing new AFEA systems. In the rest of the chapter we review important examples of all these components

mentioned, the existing databases and proposed applications, with particular emphasis on those utilized, adopted or used for comparison in this dissertation.

3.1 Face and Facial Points Detection Methods

Face localization and detection methods can be sub-categorized in two as holistic and local (or part-based) approaches. Holistic approaches aim to model the whole face as a single object and look for the complete face in an image. Local approaches, on the other hand, are able to localize parts or landmarks on the face and from there estimate the location of the face. They can, in turn, be said to be better handling partial or occluded views of faces. The term face detection refers to detecting the presence and location of all faces in an image, while localization refers to finding the location of a single one. We will use the terms interchangeably from here on, since the focus of the dissertation is the analysis of individuals' faces. Facial points refer to salient landmarks on the face that can be consistently identified across different faces. Facial points detection can be classified as shape, appearance and regression based methods. In this section we give a short review and explain three important facial point detection methods, which are used and referred to in latter chapters of the dissertation, in detail: Active Appearance Models (AAM), Constrained Local Model (CLM) and Supervised Descent Method (SDM). For methods other than those provided here, Hjelmås & Low (2001), Li *et al.* (2004), Zhang & Zhang (2010), Wang *et al.* (2014) provide extensive reviews from earlier attempts to more up-to-date ones.

Earlier approaches to face detection used information such as skin color or motion (e.g. Wang & Chang (1997), Chai & Ngan (1999)). Modern techniques rely on learning methods and classifiers that utilize appearance-based descriptors of the face or its landmarks. The most commonly known and used face detection system is the one by Viola & Jones (2004). It has drawn a lot of attention when it was published, as it is the first real-time face-detection method and since then has been integrated in many facial analysis systems, including on mobile platforms, and is still commonly employed as the initialization method for newer and more precise methods. It is based on learning a cascade of weak classifiers using Adaboost Schapire & Singer (1999). The weak classifiers use Haar-like features, which can simply be explained as horizontal and vertical pixel differences at different scales. The face detector performs an exhaustive search in the image and outputs the square with the highest score to be the closest to a face. The downside of the Viola-Jones algorithm is that, its training is tedious and requires thousands of face and non-face images, it usually outputs multiple detections for a single face, and consequentially the output is not precise enough for further facial analysis methods (e.g. expression recognition or AU detection). In addition, the method is not very robust against occlusions or variations of head-pose, especially if relevant samples had not been included in the training.

3.1.1 Active Appearance Models

The AAM, proposed by Cootes *et al.* (2001), aims to represent, or explain, an object (the face in this case) in terms of a set of model parameters, which are obtained by

constraining solutions to be valid instances of a model. Active appearance models are statistical models of deformable objects which contain both the shape and texture variation among a set of training images of the object. It is more suitable for applications in facial analysis, since it provides positions of the facial features such as eyes, brows, nose, mouth etc. as well as the strict boundaries of the face. They can be seen as an extension to the Active Shape Models (ASM), which proposed such a statistical modeling of the shape variation for the first time (Cootes *et al.*, 1995).

The images in the training set are first aligned for the shape (generally using Procrustes analysis) and then texture normalized to reduce the effect of the change in lighting conditions. Given a set of shapes s represented as a vector of $2 \times n_{landmarks}$ and g the matrix of grey-level pixel intensities within s, the training process of AAMs consists first of obtaining statistical shape and texture models separately by applying Principal Component Analysis (PCA):

$$\mathbf{s} = \overline{s} + \mathbf{\Phi}_s \mathbf{b}_s \tag{3.1}$$

$$\mathbf{g} = \overline{g} + \mathbf{\Phi}_t \mathbf{b}_t \tag{3.2}$$

where $\overline{s}$ and $\mathbf{\Phi}_s$ represent the mean and eigenvectors of the covariance matrix of the shape, and $\overline{g}$ and $\mathbf{\Phi}_t$ represent those of the texture. In order to obtain a complete model of appearance the model parameter vectors $\mathbf{b}_s$ and $\mathbf{b}_t$ are concatenated and a third PCA is applied to this concatenated vector:

$$\mathbf{s} = \overline{s} + \mathbf{Q}_s \mathbf{c} \tag{3.3}$$

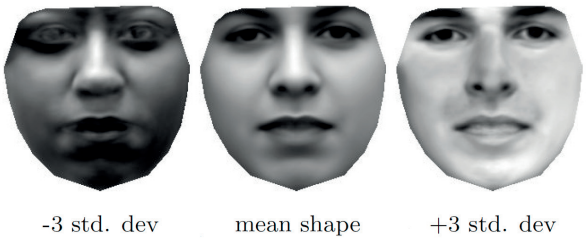
$$\mathbf{g} = \overline{g} + \mathbf{Q}_t \mathbf{c} \tag{3.4}$$

where $\mathbf{c}$ is the complete appearance model parameters vector, and $\mathbf{Q}_s$ and $\mathbf{Q}_t$ are the principal modes of the combined variation. Using this model a new face instance can be generated by alternating the model parameters $\mathbf{c}$, which consists of non-rigid parameters for the shape and rigid parameters for the global transformation, defining the overall Point Distribution Model (PDM).

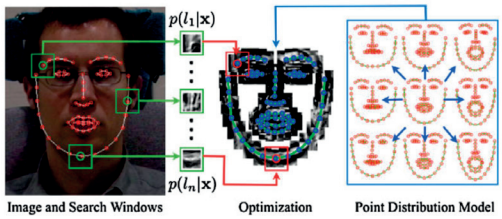
The idea of the AAM search algorithm is then to synthesize a new example by the adjustment of these model parameters (Fig. 3.2a), and it is generally treated as a minimization problem of the difference between the synthesized image and the original unseen image, so that the two are as close as possible.

The main limitation of the original AAM is that the fitting is constrained by the variation among the training set. Therefore, it is more accurate in person-specific applications (Gross *et al.*, 2005). The original AAM was proposed as a 2D model trained for faces with near-frontal poses. This, of course is not sufficient for real applications. However, if many pose configurations are included in training, then the accuracy of landmark detection decreases for all. In addition, some pose variations result in self-occlusion of certain landmarks and AAM is not flexible enough to handle this issue.

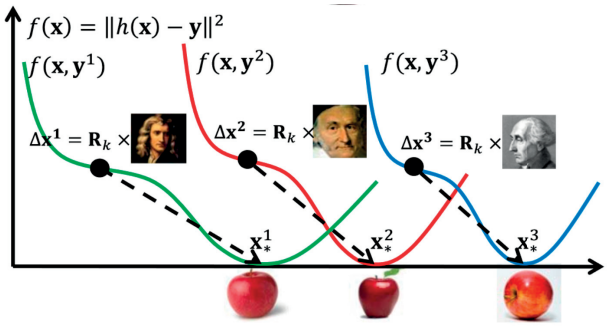
To overcome this problem many methods have been proposed, namely 3D (Ramnath *et al.* (2008)), or 2D+3D model (Xiao *et al.*, 2004) extensions,the multi-view nonlinear active shape model by Romdhani *et al.* (1999) which employs the Kernel PCA, manifold approaches as in (Osadchy *et al.*, 2007) or the use of multiple AAM and adequate model switching according to the pose (Li *et al.* (2005),Yüce *et al.* (2011)).



(a) Variation of appearance in the AAM as moving away from the mean



(b) Illustration of the CLM



(c) Illustration of the SDM

Figure 3.2: Representation of the AAM, CLM and SDM methods. Images from (a) Sorci (2009), (b) Saragih *et al.* (2011), (c) Xiong & De la Torre (2013)

3.1.2 Constrained Local Models

CLM is the name given to the ensemble of methods that aim at localizing a set of points on a given image, constrained by an overall statistical shape. The first CLM was proposed by Cristinacce & Cootes (2006), where again a PDM is generated, but this time local appearance patches around the shape points are tried to match in the image using feature templates, instead of the actual pixel values of the whole image, compared to the AAM, which naturally leads to a better generalization.

The original method can be summarized as replacing the texture vector $\mathbf{g}$ in Eq. 3.2 with a vector that is the concatenation of patches extracted around each feature and normalised to have zero mean and unit variance. Each of these patches output a *response image* that defines a cost term. The total cost is optimized by manipulating the shape parameters, explaining the constraint coming from the overall shape model (Fig. 3.2b).

These response images can also be created using other appearance descriptors, such as LBP or Histogram of Oriented Gradients (HoG).

A notable adoption of the technique is by Saragih et al. (Saragih *et al.*, 2011), where they have proposed a non-parametric distribution to approximate the response image and the shape fitting is reduced to a regularized mean-shift. This method is very efficient and is robust against partial occlusions and a larger variety of head-poses, and is therefore frequently employed in facial analysis applications[1].

3.1.3 Regression based methods - Supervised Descent Method

Another efficient approach to facial point detection and tracking is regression-based methods. Regression-based methods learn a mapping from local image patches to a probability over the parameter space, the 2D position of the facial points on the next frame, in this case. Due to their efficient computation and robustness against variability of head-pose and image resolution they have been effectively used recently for facial point detection. In a relatively early attempt Cristinacce & Cootes (2007) have extended the ASM with a GentleBoost regression scheme. In Valstar *et al.* (2010) SVM regression is combined with conditional Markov Random Fields (MRF), in Dantone *et al.* (2012) the authors use Conditional Regression Forests that are conditional to the global face properties, such as the head-pose. Cao *et al.* (2014b) propose a two-level boosted regression with explicit shape correction, which was further extended by Richter *et al.* (2014) to allow for different feature channels and include head pose information to improve detection performance.

Xiong & De la Torre (2013) proposed to use the Supervised Descent Method (SDM) for the minimization of non-linear least-squares problems and applied it successfully to the problem of facial point tracking. The real-time performance and the publicly available implementation makes it the most commonly used state-of-the-art face and facial-point tracking system[2]. It is an extension of Newton's gradient descent method which aims at minimizing a function by sequential updates or cascaded regression (Fig. 3.2c). SDM carries out this function in a supervised manner, i.e. for a training set of known facial point locations and corresponding templates or appearance features (e.g. SIFT) SDM learns a series of parameter updates and generic descent directions and for an unseen image the extracted templates are projected onto the learned descent direction to obtain the displacement update of the facial features.

Common to all regression-based methods, for tracking the points in a sequence it requires an initial estimate of the positions, which is typically chosen as the mean shape of the training set, scaled and translated using a face detector. For the rest of the frames, previous locations of the facial points are used to regress from.

3.2 Geometric and Appearance Based Feature Extraction

Feature extraction is the process of obtaining representations from images, or sequences of images, that are ideally relevant to the discrimination task and that can be

[1] Code available at: https://github.com/kylemcdonald/FaceTracker

[2] Available at: http://www.humansensing.cs.cmu.edu/intraface

formulated in a fixed-size vector form to be used in a classifier or regressor. For facial analysis systems the features used in the literature can be categorized in two: Geometric features and Appearance based features. Geometric features are the ones that are calculated through the locations of certain points (landmarks) on the image and do not include any pixel intensity information (except for being used to locate these landmarks). Appearance based features, on the other hand, rely on these pixel intensities either directly or via image transformations on a global level or extracted locally. Geometric and appearance based features can also be used in combination (in feature-level or classifier-level fusion) and have even shown better performance in certain cases compared to their individual usage. In this section we review and explain commonly used features of both types, with a discussion on their weak and strong points.

3.2.1 Geometric Features

The geometric features of the face consist of the ones that involve the actual location of facial landmarks. This location information is then converted in a feature representation either via normalized direct coordinates of the image or as a function of the distance between multiple points. The geometric features can be calculated in a single frame or can be calculated over two or more frames as a difference or trajectory function.

Features of this type that have been used in various works can be listed as follows: (a) Locations of facial landmarks (e.g. eyebrows, mouth contours etc.) (b) Distance between landmark-pairs (e.g. distance between two mouth corners or uppermost and lowermost points of the eye contours) (c) Angle of the lines joining landmark-pairs (d) Angle between edges of polygons joining 2+ landmarks (e) Difference of these features in the current frame and a reference frame (f) Trajectory of these features along a sequence frame, represented in a fixed size feature vector, e.g. via coefficients of a polynomial fit (g) Coefficients of a shape model fitted on the specific image.

The location-related information has to be normalized with respect to the overall size of the face so as to remove the effects of the distance to the camera and individual differences in face size. This can be performed using the distance for example between the most extreme points of the chin on the horizontal axis, if these points are provided by the facial landmark detector, or the distance between the innermost eye points which is unaffected by facial actions. In addition, a precise face registration needs to be applied prior to extracting the features, as all three of the yaw, pitch and roll rotations of the head directly influence the feature values. Another solution is to use a representation that is intrinsically invariant to these transformation.

Geometric features are also highly affected by subjective differences in facial landmark configurations. This problem cannot be solved through a simple normalization of the face size etc. as individuals also differ in terms of the relative positions of facial features such as the eyes, nose, mouth, even in the neutral (expressionless) case, i.e. when no facial action is present. To avoid the effect of these individual factors some systems use a neutral face of the subject as a reference frame and the features extracted from other frames are calculated relative to the values obtained from this reference frame (e.g. Lien *et al.* (1998), Senechal *et al.* (2011), Yüce *et al.* (2013a)). This requires, of course, the presence of a neutral frame and the knowledge that it is indeed expressionless, which

is attainable with tests performed on databases with human annotations yet not always feasible in real applications. To overcome this problem Senechal *et al.* (2011) have proposed to generate a neutral face of the subjects by PCA reconstruction using a basis created only with expressionless faces.

Since the efficacy of geometric features is directly related to the precision of the facial landmark detection system, geometric features are not very suitable to detect subtle facial expressions. The feature vectors contain the accumulated noise introduced by the facetracker, for example when using the evolution of locations in two or more frames. When the landmark detector is accurate, however, geometric features are very effective in detecting AUs (especially some AUs that are marked relatively more by movement of salient facial points) and even the temporal phases of AUs (e.g. Pantic & Patras (2006), Valstar & Pantic (2012)), particularly on data where the head-pose does not vary significantly over sequences.

3.2.2 Appearance Features

Appearance features are based on the texture information in an image, that is the pixel intensities. These pixel intensities can be used as they are directly as features for facial action recognition. However, this requires an accurate registration of the faces, as well as intensity normalization for illumination and individual skin color differences. This direct appearance information has been used for example within an AAM framework in Mahoor *et al.* (2009). Feature transformations are in general less influenced by these aforementioned factors and they are able to represent effectively additional information on the face, such as edges, corners, frequency etc. This information is more *meaningful* and discriminative in terms of facial actions, therefore it is common to apply a transformation on a face image and form a feature vector through this transformation. Chew *et al.* (2012) have investigated the benefits of using feature transformations compared to using raw pixel information for the task of AU detection and they have concluded that feature transformations are useful in cases of alignment errors and illumination variations, but not so much when these conditions are *perfect*.

The most commonly used appearance descriptors in the literature are the HoG (Dalal & Triggs, 2005), Discrete Cosine Transform (DCT) (Ahmed *et al.*, 1974), LBP and its variants in 3D and the frequency domain, filter banks (Gabor wavelets in particular) and SIFT features. The construction of the HoG features are similar to that of the SIFT and an example of usage in the facial action context can be found in Chew *et al.* (2012). The DCT features provide a direct representation of the texture frequency and has been used for facial expression recognition in Gao *et al.* (2014) and AU detection in Gehrig & Ekenel (2011). LBP, Gabor wavelets and SIFT features are explained in detail in the subsections below, since they are used extensively in the following chapters of the thesis.

Similar to the shape (geometric) features, explained in Section 3.2.1, appearance based features can also be constructed frame-wise or using multiple frames of a sequence. The dynamic feature extraction is performed either by using a distance function between feature vectors of different frames or by extending the feature transformation techniques to 3D, the third dimension being the time.

Local Binary Patterns

Local Binary Patterns (LBP) have been introduced by Ojala *et al.* (1996) as a method for general usage in texture description and soon after was discovered to be a perfect tool for facial image analysis. The original LBP is based on representing pixels in terms of their comparison with neighboring pixels and many variants have been proposed continously for a better representative power, including its frequency counter-part named Local Phase Quantization (LPQ), which provides additional robustness against image blurs (Ojansivu & Heikkilä, 2008). In Huang *et al.* (2011) and Shan *et al.* (2009) are reviews on its multiple variants with a focus on the uses in facial image analysis problems, including face detection, face recognition and facial expression analysis. In this section we give the definition and formulation of the basic LBP and some example variants from the literature so as to supply a sufficient background for the following chapters.

For a pixel located at coordinates p_c and intensity $I(p_c)$, the original LBP transform is calculated using the intensities of neighboring pixels p_k in a P neighborhood as :

$$LBP(p_c) = \sum_{k=0}^{P-1} t(I(p_k) - I(p_c)).2^k \tag{3.5}$$

where the function t is a binary thresholding function of the form

$$t(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases} \tag{3.6}$$

The resulting value is a P-bit binary value, or a non-negative integer $< 2^P - 1$. Since it depends on pixel differences, LBP can efficiently represent the edge information in the image in a way that is robust against illumination changes. The choice of neighborhood is an important factor when extracting LBP features. The original LBP was proposed within a rectangular neighborhood, while it is also very common to use a circular neighborhood, resulting in what is called the *circular LBP*. A feature vector is then created using histograms over the whole image or over smaller predefined windows, although it is safe to say that LBP is more effective for modeling local texture information. See Chapter 4 for an example implementation with overlapping windows of various sizes.

The resolution of the LBP can also be varied using the distance of neighboring pixels to the center pixel, as well as the number of neighboring pixels taken into calculation in this neighborhood. Fig. 3.3 shows some examples of representations at different resolutions; $LBP_{8,1}$ for example denotes the LBP that is calculated in an 8-neighborhood with 1 pixel radius, or $LBP_{16,2}$ is the one that uses 16 pixels within a 2 pixel radius circle and so on. This is a parameter that needs to be chosen depending on the type of application and images. A lower resolution (larger neighborhood), for instance, may be more suitable when the images are noisy. Multiple resolutions of LBP can also combined via concatenation of feature vectors or a decision-level fusion to create multi-resolution LBP (Ojala *et al.*, 2002).

Two commonly used extensions of the LBP are the uniform-LBP and rotation invariant LBP (Ojala *et al.*, 2002). They both decrease the histogram size significantly by grouping the possible patterns. The uniform-LBP keep 58 of the original circular

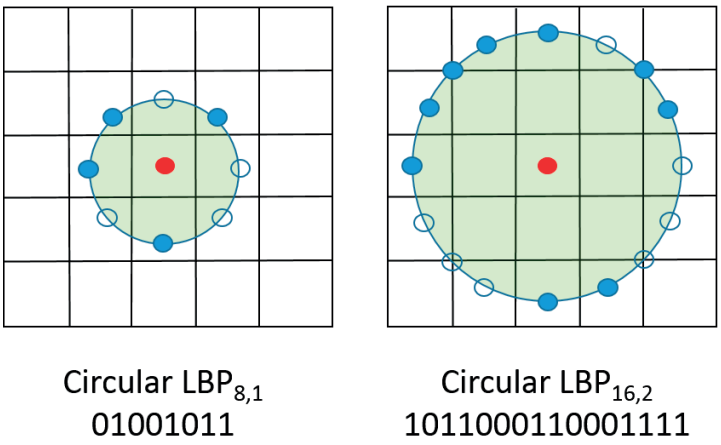


Figure 3.3: Illustration of two LBP resolutions. Pixel interpolations with larger values than the center pixel (in red) are shown in filled circles.

$LBP_{8,1}$, which contain at most two bitwise transitions between 0 and 1 and which have been shown to be the most representative, and groups the rest in a single bin resulting in a 59-bin histogram. The rotation invariant LBP groups the patterns that result in the same binary vector when an in-plane rotation is applied on the neighborhood. For example the vectors 00110000 and 00001100 are placed in the same bin, and the same for all the patterns that differ only in terms of the *roll* type of rotation. The rotation invariant LBP allows for directly using it without the need for correction for in-plane pose variations.

The LBP can also be extended to represent dynamic patterns. The LBP computed on Three Orthogonal Planes (TOP) is such an extension, and has been used extensively for facial expression recognition and AU detection (Zhao & Pietikainen, 2007). LBP-TOP allows obtaining patterns for both the appearance and motion. In addition to the horizontal and vertical spatial dimensions X and Y the third dimension T represents the time and allows computing LBP on the three planes XY, XT and YT independently of each other.

Gabor Wavelets

Wavelet transforms allow representing the texture in images in different frequency bands and Gabor wavelets in particular have been proposed (Daugman, 1985) as a model of the simple cells in the visual cortex and they possess the desirable characteristics of capturing salient visual properties, such as spatial localization, orientation selectivity, and spatial frequency. 2D Gabor kernels are spatial bandpass filters that achieve the theoretical limit for conjoint resolution of information in the 2D spatial and 2D Fourier domains, i.e. an optimal localization in both domains (Lee, 1996). They have been recognized as one of the most successful feature extraction methods for texture classification and in particular for face representation. They form a well-established image decomposition because of their spatial locality and orientation selectivity characteristics and have been used successfully for face and facial expression recognition and analysis

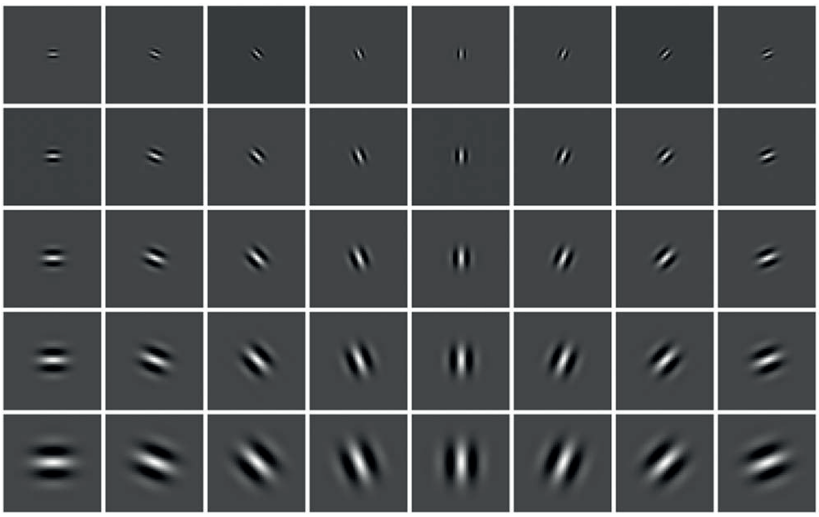


Figure 3.4: Real-part of Gabor wavelet kernels with 5 different scales (vertical axis) and 8 different orientations

(Lyons *et al.* (1998), Shen & Bai (2006)).

The 2D Gabor wavelet is defined as follows:

$$\psi(\vec{x};\nu,\mu) = \frac{k_{\nu,\mu}^2}{\sigma^2} e^{(-\frac{k_{\nu,\mu}^2 \|\vec{x}\|^2}{2\sigma^2})} [e^{(ik_{\nu,\mu}\vec{x})} - e^{(-\frac{\sigma^2}{2})}] \tag{3.7}$$

with $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} xcos\phi + ysin\phi \\ -xsin\phi + ycos\phi \end{pmatrix}$ and $e^{(ik_{\nu,\mu}\vec{x})}$ is the oscillatory wave function, whose real part and imaginary parts are respectively the cosine and sine functions. μ controls the orientation of the filters while ν scales the center of the filter in the frequency domain Daugman (1985). Fig. 3.4 shows the real-part of different Gabor kernels produced by variation of the μ and ν parameters.

Gabor wavelets have been extensively used for facial action and expression recognition in the literature due to these properties mentioned above. A common practice is to use Gabor filters in combination with the LBP transform, constructing what is called Local Gabor Binary Patters (LGBP) (Zhang *et al.*, 2005) and by projection on the different orthogonal planes as a natural extension LGBP-TOP (Almaev & Valstar, 2013) . Gabor and LGBP transform are revisited in Chapter 5, where we propose to use LGBPs with the additional curvature properties for facial AU detection.

Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) has been proposed by Lowe (2004) as a means to represent, detect, match and track objects in images and image sequences. What makes the SIFT special compared to similar descriptors such as HoG or Harris corner detectors, is its invariance to changes in scale, i.e. an object (or keypoint) is represented with the same feature vector regardless of its size in an image allowing for efficient object matching in multiple images and tracking in videos.

What gives SIFT its scale-invariance is the use of Difference of Gaussians (DoG) obtained by Gaussian filtering with two different σ values on different scales, creating a *Gaussian pyramid*. Afterwards the local extrema are detected by comparing pixels with its neighbors over scale and space, resulting in the detection of key-points in image.

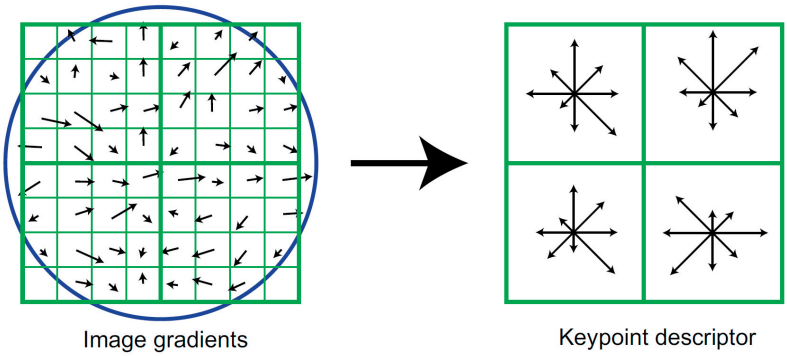


Figure 3.5: Illustration of the SIFT keypoint descriptor construction (Image from Lowe (2004))

Each keypoint is then assigned an orientation to achieve invariance to rotation. The gradient magnitude and direction is calculated in a neighborhood around the keypoint resulting in a histogram of orientation of 36 bins to cover the 360 degrees of direction. The feature representation is achieved through keypoint descriptors, which are either predefined or detected using the keypoint detection algorithm. The keypoint descriptor construction is illustrated in Fig. 3.5. The 16 by 16 neighborhood around the keypoint is divided into 16 sub-blocks of 4 by 4 , for each of which an 8 bin orientation histogram is created. This results in a total of 128 bin values and thus a vector of features of the same size. The choice of neighborhood and number of bins are parameters to be tuned, but it is very common to use the original configuration proposed in Lowe (2004).

It is relatively recent that SIFT is utilized for facial analysis applications (e.g. Ding *et al.* (2013), Ringeval *et al.* (2014)), where the keypoints are provided by the face-tracker or facial landmark detector. The keypoint descriptors are calculated around these landmarks to be fed into the classifier for action detection or face recognition. SIFT descriptors are invariant to illumination changes in addition to scale and rotation differences. Their calculation is very fast allowing for their usage in real-time applications, as we present in an example in Chapter 6. An extension of the original SIFT is the 3D-SIFT that can be applied to 3D data as well as image sequences which are considered as volumes, in order to represent motion of objects and recognize actions (Scovanner *et al.*, 2007).

3.3 Machine Learning Methods for Feature Selection and Classification

After extracting features from the facial image, the final step is to use that information to make a decision on the image, be it detecting the existing AUs, recognizing one

of the basic expressions or mapping it via regression to the continuous emotional dimensions (see Section 2.1 for various categorization methods). This process can involve a feature selection or dimension reduction step, which serves to decrease the number of dimensions by removing the redundancy between features or directly eliminating the irrelevant ones. In this section we review the machine learning tools that are used in the various chapters of this dissertation and give descriptions of the performance metrics that are used to evaluate the contributions.

3.3.1 Principal Component Analysis

PCA, also called the Karhunen-Loeve transform, is a linear orthogonal transformation technique for multivariate data aiming at mapping the data on a lower dimensional space in a way that keeps an important portion of the variation among the data points. It is useful when applied prior to classification in cases where the data contains a lot of redundancy or noisy features, which is information that is not useful for discrimination. It is based on eigen-analysis of the covariance matrix of the data.

Given a data matrix $\mathbf{X}$ of n data points of dimension d each and its mean vector $\overline{\mathbf{X}}$ of size d, the $d \times d$ covariance matrix $\mathbf{C}_x$ and its eigencomposition are calculated as:

$$\mathbf{C}_x = (\mathbf{X} - \overline{\mathbf{X}})^T (\mathbf{X} - \overline{\mathbf{X}}) \tag{3.8}$$

$$\mathbf{C}_x \mathbf{\Phi} = \mathbf{\Phi} \mathbf{\Lambda} \tag{3.9}$$

$\mathbf{\Phi}$ being the eigenvectors and $\mathbf{\Lambda}$ the diagonal matrix containing the corresponding eigenvalues $\lambda_1 \cdots \lambda_d$. The value of the eigenvalues represent the amount of variance explained by the corresponding eigenvector or *principal component*. For dimension reduction applications these values can be used to determine the number of dimensions wished to keep, for example if a certain percentage of the total variance is wished to be retained then the sum of the eigenvalues of the principal components kept should not exceed the same percentage of the sum of all eigenvalues. Unseen data can then simply be projected on the new space created by multiplying it with the corresponding eigenvectors, being a rotation in the original space.

Turk & Pentland (1991) have proposed a simple trick to overcome the computational problem of calculating the huge covariance matrix and its Singular Value Decomposition (SVD) when the number of dimensions is much larger than the number of datapoints. It is simply to compute the covariance matrix on the transposed data matrix $\mathbf{X}$ and multiplying the calculated *pseudo-eigenvectors* by the covariance matrix in order to obtain the same first n eigenvectors for a much lower complexity. Note that, projections of images on PCA-bases can also be used directly as features for classification as a simplified representation, as in the case of the Eigenfaces method (Turk & Pentland, 1991).

3.3.2 Linear Discriminant Analysis

A linear discriminant function is the set of linear relations that provides the best separation between classes. In the binary class case it can be notated as:

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \tag{3.10}$$

where $\mathbf{w}$ is the weight vector, w_0 is the bias factor and the datapoint $\mathbf{x}$ is said to belong to class C_0 if $y(\mathbf{x}) < 0$ and to C_1 if $y(\mathbf{x}) > 0$. The case where $y(\mathbf{x}) = 0$ is called the $d-1$ dimensional separating hyperplane, where the dimension of the data is d. In cases where the number of classes k is greater than two, one can take two approaches. The first one is the *one-versus-the-rest* approach where $k-1$ classifiers of this kind are learned, one for each class. The second approach is the *one-versus-one* where $k(k-1)/2$ classifiers are learned to separate each class from one another, and the final decision can be assigned by majority voting. From now on, we will be referring to the binary-class case ($k = 2$) in the explanation and notations, which has mainly been adopted from Bishop (2006).

In linear discriminant analysis or *Fischer's linear discriminant* the linear classification model is considered as a dimensionality reduction problem and for a d-dimensional data vector define $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ as a projection to a single dimension. The classification provided by Eq. 3.10 can still be applied as if $y(\mathbf{x}) > w_0$ $\mathbf{x}$ belongs to class C_1 and vice versa. This projection may cause a loss of the discrimination present in the original space if the choice of $\mathbf{w}$ is not made to maximize the separation between classes. Taking $\mathbf{m}_0$ and $\mathbf{m}_1$ as the mean vectors of the data belonging to the two classes, the choice can be made to maximize the separation between the class means in the projection space, i.e. maximizing

$$m_1 - m_0 = \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_0) \tag{3.11}$$

with

$$m_k = \mathbf{w}^T\mathbf{m}_k \tag{3.12}$$

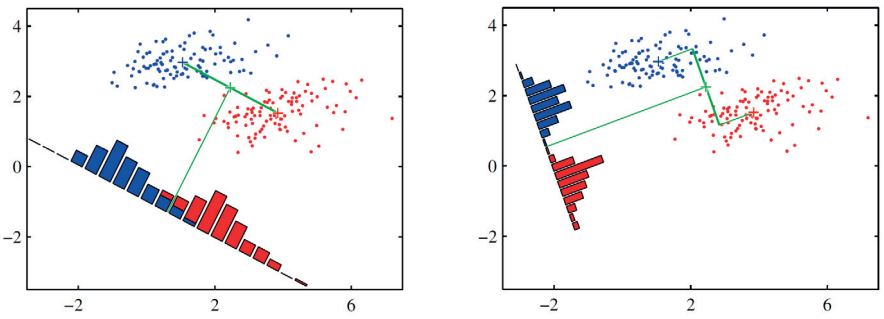


Figure 3.6: Illustration of the LDA. (a) shows the separation obtained by projecting on the line connecting the means of the classes (b) shows the advantage of Fischer's discriminant. Images are from Bishop (2006).

Constraining $\mathbf{w}$ to have unit length we obtain $\mathbf{w} \propto \mathbf{m}_1 - \mathbf{m}_0$, which provides a fair classification by projecting the data on the line joining the two means with the separation being in the middle. However, as shown in Fig. 3.6a, this does not necessarily lead to the best discrimination as we do not take into account the inter and intra class variations. Fischer proposed to solve this problem by maximizing the inter-class separation while minimizing the intra-class variation so as to minimize class overlap on the projected

space. Fischer's criterion is defined as the ratio of between-class variation to the within-class variation:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{3.13}$$

where the between-class covariance matrix $\mathbf{S}_B$ and within-class covariance matrix $\mathbf{S}_W$ are defined as:

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T \tag{3.14}$$

$$\mathbf{S}_W = \sum_{n \in C_0} (\mathbf{x_n} - \mathbf{m_0})(\mathbf{x_n} - \mathbf{m_0})^T + \sum_{n \in C_1} (\mathbf{x_n} - \mathbf{m_1})(\mathbf{x_n} - \mathbf{m_1})^T \tag{3.15}$$

To maximize $J(\mathbf{w})$ we differentiate Eq. 3.13 by $\mathbf{w}$ and obtain:

$$(\mathbf{w}^T \mathbf{S}_B \mathbf{w}) \mathbf{S}_W \mathbf{w} = (\mathbf{w}^T \mathbf{S}_W \mathbf{w}) \mathbf{S}_B \mathbf{w} \tag{3.16}$$

The magnitude of the projection is not important, but only its direction. Also we know that $\mathbf{S_B}$ is always in the direction of $(\mathbf{m}_1 - \mathbf{m}_0)$ .Therefore, dropping the scalar factors we obtain Fischer's linear discriminant as:

$$\mathbf{w} \propto {\mathbf{S}_W}^{-1} (\mathbf{m}_1 - \mathbf{m}_0) \tag{3.17}$$

Fischer's discriminant provides a projection that can be used as a dimension reduction technique and also a classifier by using appropriate thresholding. LDA is able to take into account the differences in the within-class variations, but a drawback is that it reduces the classification problem to $k - 1$ dimensions, which may not be sufficient for a complex classification.

3.3.3 Support Vector Machines

The Support Vector Machine (SVM) is a maximum margin binary classifier, which tries to separate two classes by a margin whose width is maximized so as to decrease the generalization error over all training instances. SVM is a very successful machine learning tool that has been effectively used for a variety of classification and regression problems (Bishop, 2006).

Given a set of N training instances and corresponding labels $(\mathbf{x}_i, l_i)$, with $x_i \in \mathbb{R}^d$ and $l_i \in -1, 1$, and coming back to the linear separation problem:

$$y(\mathbf{x}) = \mathbf{w}^T \phi(x) + b \tag{3.18}$$

where $\phi(x)$ is an optional feature transform, b is a bias parameter and the labels l_i take values -1 or $+1$ according to the sign of $y(x_i)$. A margin is defined as the distance that between the closest instance and the decision hyperplane. SVM tries to maximize this margin by solving the optimization problem:

$$\begin{aligned} &\underset{w,b,\xi}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C \sum_i \xi_i \\ \text{subject to} \quad & l_i(\mathbf{w}^T \phi(x_i) + b) \geq 1 - \xi_i \\ \text{and} \quad & \xi_i \geq 0, \quad i = 1, \cdots N \end{aligned} \tag{3.19}$$

ξ_i are called the slack variables, introduced to cope with potential *non-separable* instances lying in the decision margin region providing a so called "soft-margin" and making the optimization problem solvable (Smola & Schölkopf, 2004). $C > 0$ is the cost (penalty) parameter that indirectly defines how tight the margin of the classifier will be. A larger C means a larger penalty, thus a smaller margin. Although a large C ensures less classification error in the training data, it can also lead to *overfitting* and thus a worse performance for unseen data, particularly when noise was present in the training data.

The Kernel Trick

Although SVM is by nature a linear classifier, it also allows for non-linear classification using what is called the Kernel Method or the kernel trick. A kernel is defined as a dot-product in the feature space and this feature space can be created using a mapping function ϕ.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \forall \quad \mathbf{x}_i, \mathbf{x}_j \tag{3.20}$$

where $\langle \cdot, \cdot \rangle$ defines a dot-product and the kernel matrix K therefore defines a pairwise relation between samples. The kernel trick is the ability to use this pairwise relation instead of explicitly defining the feature mapping ϕ. In addition, since it is defined as a dot-product, K is positive definite symmetric (PSD) kernel, also called *Mercer kernel*, and can be used within machine learning tools to replace the original feature space.

Many kernels have been proposed and used, mostly depending on the particular application type. A review of these could be found in Schölkopf & Smola (2002). Some commonly used kernels are:

- Linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.21}$$

  It is the kernel corresponding to an identity mapping $\phi(x) = x$, the pairwise similarity measure is thus just a dot-product.

- RBF kernels:

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2) \tag{3.22}$$

  Also called the Gaussian kernel, the Radial Basis Function (RBF) kernels impose a similarity related to the Euclidean distance regularized by the parameter $\gamma > 0$, called the bandwidth parameter, defining the width of the kernel.

- Polynomial kernels:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + b)^p \tag{3.23}$$

  They have two parameters to tune: $p \geq 1$, the degree of the polynomial, and b, the offset. Note that, the linear kernel is a polynomial kernel of degree 1.

The parameters of the kernels (if applies) and the cost parameter C of the SVM are generally optimized using a grid-search and $n - fold$ cross-validation on the training set. That is, each time $1/n$ of the training data is left out for validation, while the rest of the data is used to train a classifier with all possible discrete sets of parameters to be tuned and this is repeated n times. The best parameters are then the set that gives the

best average result over the n-folds and a final classifier is trained with these parameters using all training data, to be used with the new *unseen* data. This ensures the separation of training and validation data and thus overfitting of the parameters to the training instances.

3.3.4 Boosting Methods

Boosting is the term that is used to define the ensemble of methods that aim at combining multiple classifiers to produce a *committee* of decisions, that is used as the final classifier and whose performance is ideally better than any of the *"base"* classifiers. These base classifiers are generally chosen from *weak learners*, which are very simple classifiers with only better than random performance, and still can create very powerful classifiers when combined in a boosting scheme.

$$f(\mathbf{x}) = sign\left(\sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})\right) \tag{3.24}$$

In 3.24 each $f_m(\mathbf{x})$ is a weak learner (a decision function based on the data $\mathbf{x}$) and $\alpha_1, \alpha 2, \cdots, \alpha_M$ are the corresponding weights that are learned using the boosting algorithm.

The most well-known and commonly used boosting algorithm is AdaBoost or adaptive boosting, that was developed by Freund *et al.* (1996). It is the method that is also used in the Viola & Jones (2004) face detection method. The main idea is to give emphasis by weighting the instances that are misclassified at each step of the classifier, i.e. adapt the classifier to better handle problematic instances. The AdaBoost.M1 algorithm, also known as the *Discrete AdaBoost* since the base classifiers return a discrete label, is outlined in Algorithm 1.

Algorithm 1 Discrete AdaBoost (AdaBoost.M1)

1: Initialize the weights $w_i = 1/N$, $i = 1, \cdots, N$, with N = number of training instances
2: **for** $m = 1$ to M **do**
3: Fit a classifier $f_m(\mathbf{x})$ on training data using weights w_i
4: Compute

$$err_m = \frac{\sum_{i=1}^{N} w_i I(l_i \neq f_m(\mathbf{x}_i))}{\sum_{i=1}^{N} w_i} \tag{3.25}$$

5: Compute $\alpha_m = log((1 - err_m)/err_m)$
6: Set $w_i \leftarrow w_i exp[\alpha_m \cdot I(l_i \neq f_m(\mathbf{x}_i))]$, $i = 1, \cdots, N$
7: Renormalize s.t. $\sum_i w_i = 1$
8: **end for**
9: Output $f(\mathbf{x}) = sign\left[\sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})\right]$

Note that, the type of weak learner as well as the overall error function in Eq. 3.25 can be modified based on the application. GentleBoost, proposed by Friedman *et al.* (2000), for example chooses the f_m at each iteration that minimizes the weighted least-squares

term $\sum_{i=1}^{N} w_i(l_i - f_m(\mathbf{x}_i))^2$. While the original AdaBoost favors the highest possible α values, GentleBoost is more *conservative* algorithm that uses Newton updates for the weights. It has been shown to converge faster and outperform AdaBoost and its variants (Friedman *et al.*, 2000).

In the case where each of the weak-learners use a single feature to make a decision, boosting algorithms weight the feature of each performance, or in other words learn the best separating feature at each step. This information can be used as a method for feature selection. GentleSVM, which means selecting the most discriminative features with GentleBoost and using them in an SVM classification framework, has been successfully employed for classification problems and is also used in the following chapters of this dissertation.

3.3.5 Random Forests

Random Forest (RF) is another powerful classification method that can be defined as an ensemble of trees that are each trained using randomly selected training instances. This selection of samples from the training data is called *bagging* or *bootstrap aggregation*. Bootstrap methods allow for the non-parametric estimation of the data distribution and therefore can be used as a way of estimating the classifier or regressor accuracy for a given parameter set.

Given a training set with data-label pairs $\{x_i, l_i\}$, bagging averages the prediction over a collection of bootstrap samples, drawn from the training data with replacement and reduces the variance of accuracy by averaging these noisy yet unbiased models. This bagging estimate is defined simply as:

$$f_{bag}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(\mathbf{x}) \tag{3.26}$$

B being the number of bags, this expression is a Monte Carlo estimate of the true bagging estimate, approaching it as $B \to \infty$ (Hastie *et al.*, 2005).

Random Forests combine the idea of bagging with trees, which can be defined as a series of splits. The trees are grown using a random selection of features (variables) for each split. The trees are grown a certain depth, which is the minimum node size, being a parameter to tune. In random forests the classifier is learned on out-of-bag estimates to estimate the performance of a tree only on instances that were not included in the corresponding *bag*. This technique removes the need for an additional independent test set (Breiman, 2001). The generalization error computed on this out-of-bag estimates converges as the number of trees (thus the number of bags) increases. The final classification decision is then made by the voting of all these trees as illustrated in Fig. 3.7.

Random forests are very effective classifiers that are robust against overfitting and can better handle class imbalance, due to their bagging-based nature. Their main drawback is the large number of parameters that one needs to tune in order to get the optimal results, such as when to stop splitting the nodes, the maximum number of random features chosen at each split, the number of instances in the bags and the total number of trees to train. Since they also provide a ranking of the features, random forests can also

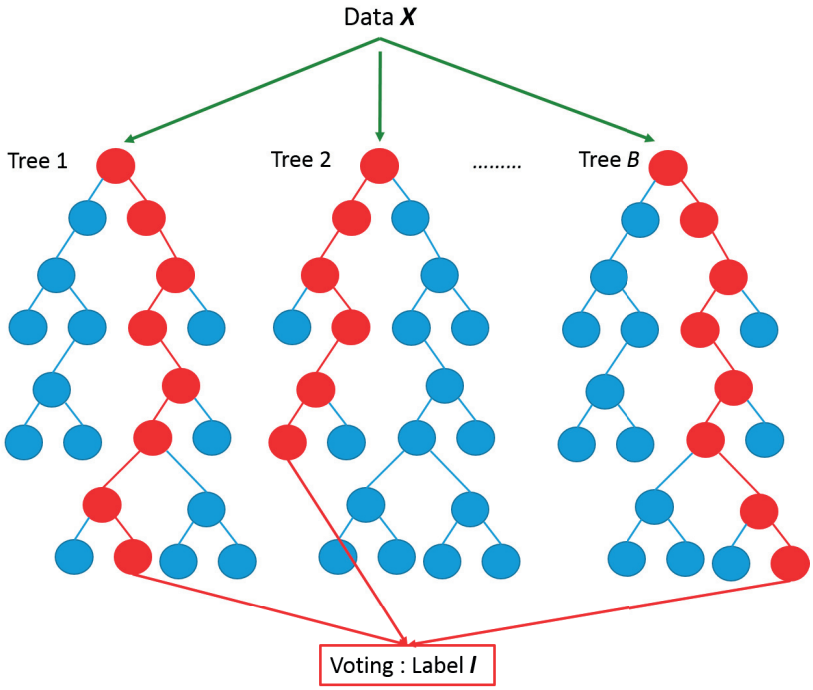


Figure 3.7: Illustration of the Random Forest Technique

be used as a feature selection tool so as to prune the features which show low overall performance among the splits that they have been selected in.

3.3.6 Performance Metrics

There exist various metrics used to quantify and compare the performance of classification (or detection) and regression systems, each of which measure and give emphasis to a different quality. The classification problem can be one-class, two-class, multi-class or multi-label. One-class problems refer to cases of novelty or outlier detection problems, for example. The difference between a one-class classification and the two-class, or binary-class, one is that in the two-class problem both the negative and positive data points are included in learning the discrimination, while one-class classifiers are trained using only the class that is *of interest* or available at training time.

Multi-class problems have more than two mutually independent classes, where each sample can only belong to one and only one of the 3+ classes. Face recognition systems that match a given face to an identity from a set of identities are examples of this type. The multi-label case, on the other hand, does not have the mutual independence assumption, i.e. data points can possess multiple labels at the same time. Multi-label problems can be binary or multi-class, for instance as in the case of the identity, sex, age and profession of a person. Due to the nature of the problems attacked in this thesis, we will mainly refer to the metrics that are relevant to binary-class classification or regression problems, that output continuous values. However, most of these metrics also apply to

the multi-class case, and some to the multi-label one.

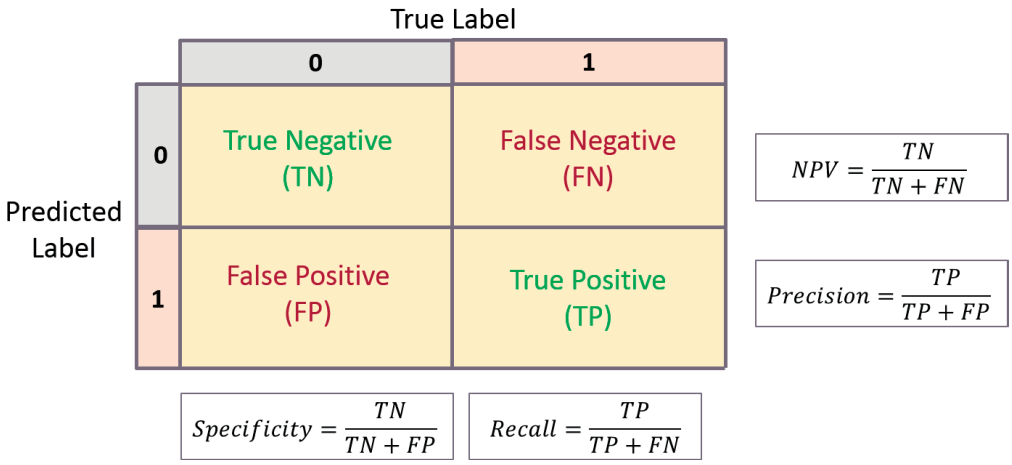


Figure 3.8: The confusion table of a two-class (0 for the negative case or 1 for the positive) classification problem and the definitions of four performance metrics: Precision, Recall, Specificity and Negative Predictive Value (NPV)

- **Confusion Table:** Also called a *Contingency Table*, it shows the correspondence between each true-class and the predicted output. It of size n_{class} (number of classes) by n_{class}. In the binary-class case the cells represent the number of the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), where one of the classes is treated as the one being of interest to detect (Fig.3.8).

- **Overall Accuracy (OA):** It is the ratio of correctly classified instances out of all instances. It is a useful measure taking into account equally all classes, but can be misleading in case of imbalanced classes.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.27}$$

- **Precision:** As the name implies, it is a measure of how precise a system is and defined as the ratio of correctly labeled positive instances over all instances that are labeled as positive. In other words, it explains how many of the cases detected as positive are *indeed* of the class of interest. It is also named the Positive Predictive Value (PPV).

$$Precision = \frac{TP}{TP + FP} \tag{3.28}$$

- **Recall:** It is a metric of the *sensitivity* of the system, and explains how many of the positively labeled instances a system is able to detect. It is also called the *True Positive Rate (TPR)*.

$$Recall = \frac{TP}{TP + FN} \tag{3.29}$$

- **False Positive Rate (FPR):** It explains how many of the negative instances were classified incorrectly as positive, and is complementary to the term *specificity*,. It is different from the *False discovery rate (FDR)*, which defines the ratio of instances incorrectly labeled as positive overall instances labeled as positive.

$$FPR = \frac{FP}{FP + TN} = 1 - specificity \tag{3.30}$$

- **F1-Score:** The balanced F-Score (or F_1 score) is a measure that takes into account both the *precision* and *recall* of a system, calculated as the harmonic mean of the two. It is a balanced metric that favors the sensitivity and the precision at the same time. It is particularly meaningful in cases of class imbalance. For instance, in a classification problem with 90% of the data carrying the negative label, even if the system classifies all instances as negative it will reach a 90% OA, but the F-Score will be too low, or indefinite. It ranges between 0 and 1, 1 being the perfect classification case.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.31}$$

- **Receiver Operator Characteristics (ROC) curve:** The ROC curve shows the relationship between the FPR and TPR values, i.e. it shows how many false positives have to allowed to reach a certain amount of true positives. In case the classifier employed outputs a continuous confidence measure (e.g. the distance to hyper-plane in SVM, or the probability output of the RF), it can be generated by tuning the decision threshold from the minimum to the maximum value. It is very useful to show the balance between the two measures, and can be used to obtain a threshold that is suitable for the particular application. For instance, in some medical diagnosis problems it is important to detect all pathological cases no matter the number of false positives, as a precaution. While in user state monitoring systems that provide feed-back or use the information for system adaptation, it is important for the acceptability of the system not to output too many unnecessary detections, so in this latter case the FPR is more important. In the best case the curve would be where $TPR = 1$ for all $FPR \geq 0$ and in the random classification case $TPR = FPR$ for all values. The Area Under Curve (AUC) is a measure used to compare ROC curves, which is 1 in the perfect case, 0.5 for random, and ≤ 0.5 for worse than random (a concave curve) (Fig. 3.9).

- **Root Mean Squared Error (RMSE):** It is used when the decision output is not binary or it is of continuous type, thus applies to regression problems. It is defined as the root of the expected value of the squared differences between the true labels and the predicted outputs.

- **Intraclass Correlation Coefficient (ICC):** It is also used as a metric to compare non-binary output systems. It is calculated as the difference of within-target mean squares and residual sum of squares, normalized by the within-target mean squares. In addition to regression problems, it is also very commonly used to assess the quality of human ratings when there are multiple raters.

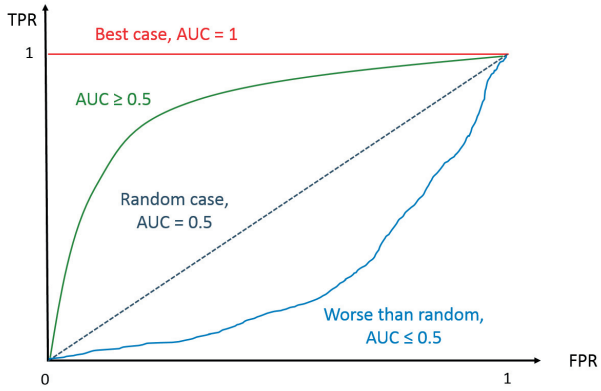


Figure 3.9: Receiver Operator Characteristics (ROC) curve and the possible outcome cases, with the area under curve (AUC) shown for each case

- **Concordance Correlation Coefficient (CCC):** It is a measure that combines the Pearson's correlation coefficient (ρ) and the mean square error in a single metric:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.32}$$

  where σ_x^2 and σ_y^2 are the variances and μ_x and μ_y are the means of the continuous valued vectors x and y that are in comparison.

3.4 Existing Databases

Facial expression databases are crucial tools for facial analysis research, both for training and testing. Publicly available image and video databases help advancements in the field and allow for objective comparison of system performances. Creating a database of facial expressions is a tedious task mainly due to providing the ground-truth. Databases currently available to the community come with annotations for the six (or seven) basic expressions, the emotional dimension (valence, arousal, dominance etc.), AU existence or intensity, or a specific condition or state, for instance pain, interest, engagement, distraction.

In the earlier years of automatic face analysis research the databases mostly consisted of posed expressions, i.e. the subjects were given explicit instructions to perform a certain facial action combination. Although this provides convenience in terms of data annotation (since the sequence-level ground-truth labels come automatically during data acquisition), the data obtained is not natural and quite different from what one observes in real-world applications. The later trend, therefore, is to obtain expressions in a spontaneous manner, that is either by emotion elicitation or by having human-raters annotate the data through visual observation. Human annotations, of course, bring along the problem of subjectivity, even in the case of FACS annotations, which is the most objective and well-defined system to date. For this purpose some databases use multiple human

raters that annotate the same data and a measure of reliability is provided along with the annotations.

Table 3.1 gives a list of some commonly used publicly available databases, the type of annotations (ground-truth) provided, the existence of spontaneous expressions and 3D data. Posed expression databases such as the CK (Kanade *et al.*, 2000) and its extension CK+ (Lucey *et al.*, 2010), the MMI (Valstar & Pantic, 2010) and the Bosphorus (Savran *et al.*, 2008) have been widely used in the field and provide a well-established performance baseline for AU and expression recognition systems. The existence of both kind of labels also allow for training and testing systems that use AUs to recognize the basic expressions. The GEMEP-FERA database is between the posed and spontaneous category (Valstar *et al.*, 2011). It contains actors uttering meaningless sentences while trying to perform in a certain emotional way. The emotions intended therefore are not real, but the expressions are natural. It also allows for multimodal analyses since it contains videos of the whole upper body and speech.

Databases containing spontaneous expressions use either emotion elicitation using visual stimuli such as movies (e.g. NVIE - Wang *et al.* (2010b), DISFA - Mavadati *et al.* (2013)) or they are recorded during natural interactions between people (RECOLA - Ringeval *et al.* (2013)) or people interacting with a software (AVEC databases - Valstar *et al.* (2013) and Valstar *et al.* (2014)) or an avatar controlled by another person, as in SEMAINE (McKeown *et al.*, 2012) where the avatars represent and act in a certain emotional state (e.g. joyful or angry). The AVEC 2013 and 2014 databases are particular in the sense that it contains labels for the Beck's Depression Index (BDI) score, which is a questionnaire measuring a person's depression level. The database can therefore be used to investigate visual and audio cues of different levels of depression.

In this section we have provided an overview of some of the existing databases, elicitation and annotation methods. In the following chapters we explain in more detail the corresponding databases that are actually used for training or test purposes.

Table 3.1: List and Comparison of Publicly Available Databases of Facial Expressions and AUs. Basic Exp.: annotated for the 6 (or 7) basic expressions, AUs: annotated for the AUs, Val. - Arou.: annotated for the valence and arousal dimensions (might also be annotated for other dimensions), Dynamic: involves sequences of images, 3D: involves 3D data, Spontaneous: involves spontaneous (non-posed) expressions, Partially: only in a portion of the data, or partially fulfilling the condition

Database	Basic Exp.	AUs	Val. - Arou.	Dynamic	3D	Spontaneous	Other
AR-FACE (Martinez, 1998)	✓	✗	✗	✗	✗	✗	Occlusion
JAFFA (Lyons *et al.*, 1998)	✓	✗	✗	✗	✗	✗	-
KDEF (Lundqvist *et al.*, 1998)	✓	✗	✗	✗	✗	✗	Multiple views
Multi-PIE (Gross *et al.*, 2007)	✓	✗	✗	✗	✗	✗	Multiple views
SAL (Douglas-Cowie *et al.*, 2007)	✗	✗	✓	✓	✗	✓	Multimodal
BOSPHORUS (Savran *et al.*, 2008)	✓	✓	✗	✗	✓	✗	Occlusion & Pose
CK+ (Lucey *et al.*, 2010)	✓	✓	✗	✓	✗	Partially	-
MMI (Valstar & Pantic, 2010)	✓	✓	✗	✓	✗	Partially	Side-view & AU temporals
NVIE (Wang *et al.*, 2010b)	✓	✗	✗	✗	✗	✓	Near-Infrared Lighting
UNBC-McMaster (Lucey *et al.*, 2011)	✗	✓	✗	✓	✗	✓	Pain Scores
GEMEP-FERA (Valstar *et al.*, 2011)	✓	✓	✗	✓	✗	Partially	Multimodal
SEMAINE (McKeown *et al.*, 2012)	✗	✓	✓	✓	✗	✓	Dyadic interaction
RECOLA (Ringeval *et al.*, 2013)	✗	✗	✓	✓	✗	✓	Multimodal
DISFA (Mavadati *et al.*, 2013)	✗	✓	✗	✓	✗	✓	AU intensities
AVEC (Valstar *et al.*, 2013)	✗	✗	✓	✓	✗	✓	Depression Scores
BP4D (Zhang *et al.*, 2014)	✗	✓	✗	✓	✓	✓	AU intensities (partial)

3.5 Applications

In Chapter 2 we have already discussed why facial expressions are important and how they provide information on one's affective or cognitive state. In this section we provide a short review of some of the application areas in the literature, that use automatic detection of AUs or facial expressions as a source of information. Note that this review does not include applications of face recognition (e.g. in forensics or face verification); for the corresponding applications the reader is referred to Jain & Li (2005). Automatic facial expression analysis is firstly an essential component of HCI (which can also be called Human-Machine Interfaces (HMI) in this case). Systems that are able to understand affective and cognitive states of users use this information to mediate and adapt their behaviour for a more user-friendly and efficient system. HCI applications range from gaming to medical and learning assistance.

One of the most commonly encountered use of automatic analysis of facial expressions is in the field of marketing. Analyzing people's facial reactions when they watch advertisements or browse a product allow measuring their liking and intent to purchase the product. Using this information producers, retailers or online shopping sites can build marketing strategies, reorganize product placements and infer about their target population (Texeira *et al.* (2012), McDuff *et al.* (2014)). This idea does not only apply to direct purchasing behaviour, but also to assessment of liking multimedia content in general. For example, as a very interesting recent application, a comedy club in Barcelona, Spain has placed cameras in front of the audiences' faces, detects every time you smile, and charges you according to your number of smiles / laughs during the performance.

Another field of application is in the health-care area. Certain psychopathologies have been shown to have as symptoms flat or abnormal affect. Abnormal affect is defined as not feeling or expressing a feeling in an expected way, for example getting extremely raged as response to an amusing stimulus with no apparent side reasons. Flat affect, on the other hand, is the deficiency to lack to feel or express a feeling when you are expected to. Flat and abnormal affect is encountered in patients with schizophrenia, depressive disorder, manic-depressive disorder and certain types of autism spectrum disorder (e.g. Asperger's syndrome). These disorders are also characterized by not being able to recognize others facial expressions (Sander & Scherer (2009), Kring & Stuart (2005)). Automatic facial expression recognition tools can be used for the diagnosis (Kächele *et al.* (2014), Valstar *et al.* (2014), Cohn *et al.* (2009)) and treatment (Gordon *et al.* (2014), Picard (2009)) of those psychopathologies and evaluation of psychiatric intervention. Another application in the healthcare area are systems that measure the pain level, for instance after surgeries or during regular monitoring of the elderly, and help take appropriate measures (Ashraf *et al.* (2009), Sikka *et al.* (2015)).

A similar type of approach is also used to model, understand and provide feed-back to learning and teaching systems, the main idea being that understanding and being able to interpret the student's feelings and cognitive states (e.g. distracted, confused) one can adapt their behaviour for a more efficient interaction. In the case of online-learning systems or teaching using robots facial analysis software become even more essential, as real-time system adaptation is feasible (e.g. Whitehill *et al.* (2008), Cooper *et al.* (2010), Grafsgaard *et al.* (2013)).

Another application field that is on the rise is driver and pilot monitoring systems via facial analysis. These systems detect particular states of the drivers (e.g. fatigue, distraction, rage) that can be hazardous for driving and allow for taking appropriate measures. An extensive review on existing systems is given in Chapter 8 along with our contribution on detecting cognitive distraction of drivers using AUs.

Facial analysis tools can also be useful in better understanding human behaviour and the underlying neurological processes. Researchers are working towards an accurate AU detection system (as in the focus of this dissertation) in order to be able to fully-automate the FACS coding process. FACS codes are then used by theoretical psychologists and cognitive neuroscientists to study the mechanisms that give way to facial expressions and also certain deficiencies in the brain that cause impairments.

3.6 Conclusion

In this section we have provided a review and explanation of the main tools that are needed for an automatic facial expression recognition system: Face and facial landmark detection, geometric and appearance based feature extraction and machine learning for feature selection and classification. We have presented an overview of existing systems, databases and related application fields. The review covers both AU detection and Facial Expression Recognition (FER) systems. We have given emphasis to methods and resources that are utilized in the following chapters of the thesis and the proposed contributions. With this section we conclude the introductory part of the thesis.

Part II

Individual and Multi-Label Action Unit Detection

Overview

Automatic facial action unit (AU) detection in videos is the key ingredient to most systems that use a subject face for either interaction or analysis purposes. From new generation game consoles to market research or software used for the treatment of psychopathologies, many applications and devices nowadays make use of facial analysis of users, consumers or patients, as previously reviewed in Section 3.5. Automated facial action detection and classification therefore continues to be an important research issue in the computer vision area. With the ever growing range of possible applications, achieving a high accuracy in the simplest possible manner gains even more importance. In this part of the thesis we propose two different methods to increase frame-level AU detection accuracy, both of which utilize LBP based features with different extensions. LBP and its variants have already been proved to be very effective descriptors of the facial texture (see Sec. 3.2.2); with our two contributions we present two novel extensions to achieve state-of-the-art AU detection performance.

In Chapter 4 we aim to reinforce LBPs using certain image processing techniques, that have been used in the domain of automatic facial analysis for the first time. These are the bilateral filter, morphological opening by reconstruction and morphological top-hat transform by reconstruction. These image transformations are applied on the face image prior to the LBP transform in order to enhance different appearance-based properties of the face, e.g. via smoothing of certain regions. The proposed system combines these texture based features with additional features based on facial point geometrical relations between the neutral and expressive frames and we show that it achieves detection rates higher than methods previously proposed, using a small number of features and basic support vector machine classification, by using some fundamental image processing tools.

Our second contribution, presented in Chapter 5, we explore the use of curvature Gabor wavelets together with LBP once again for the problem of AU detection in videos. Gabor wavelets, as reviewed in Section 3.2.2, are efficient texture descriptors with the ability to represent appearance in various scales and orientations. Curvature Gabor wavelets extend this representation to different curvature degrees. In this contribution we investigate their use in facial action recognition for the first time in the literature, in addition to the advantages of using different kernel sizes for the wavelets, making what is called Multiple Local Curvature Gabor Binary Patterns. The proposed framework proves very efficient and we show with experiments that the inclusion of various degrees of curvature and kernel sizes substantially increase detection accuracy of AUs, obtaining the state-of-the-art performance on the CK+ dataset (Lucey *et al.*, 2010), composed mostly

of posed expressions. We also show the application of the system in a cross-database manner and discuss the strengths and drawbacks of the methodology.

To summarize, in this part of the dissertation we investigate two novel methods so as to show how to more efficiently use LBP based appearance features for the problem of AU detection in video frames. The proposed methods have also appeared in the publications Yüce *et al.* (2013a) and Yüce *et al.* (2013b).

Improving LBP based AU detection using morphological and bilateral filters

4

4.1 Introduction

In this chapter we propose a novel extension to LBP based AU detection that uses certain fundamental image processing tools as a preprocessing step so as to increase the efficiency of the output appearance features, as our main contribution. With additional novelties, extensions and feature combinations we build a framework that significantly increases AU detection accuracy.

Although the LBP and its many variants have been extensively investigated for AU detection and expression recognition purposes (see Shan *et al.* (2009) and Section 3.2.2 for a review), too few of the works have gone further than extracting histograms on a fixed grid in 2D or 3D (the third dimension being time). In Senechal *et al.* (2010) and Senechal *et al.* (2011), the authors have successfully used the difference of the LGBP histograms between the neutral image and the peak expression. We adopt a similar approach, however we compute the LBP histograms obtained from overlapping windows and compute a single feature per window, which is the χ^2 distance between the histograms, resulting in a smaller number of features which search more extensively throughout the image.

In addition, we apply three different filters (using morphology by reconstruction and bilateral filters) separately before applying the LBP transform on the image. This lets us obtain three different LBP transforms which define more clearly the edges than directly applying the LBP transform, and we show with experimental results that indeed the new features proposed achieve a better accuracy. We also show that by combining these texture features with certain shape features we can achieve detection performances higher than other methods that have reported results on the same database that we use for our tests.

The majority of the work included in this chapter has already been published in the Proceedings of the FG'13 conference (Yüce *et al.*, 2013a). In the rest of the chapter we first describe the shape features, preprocessing methods and texture feature extraction

procedure along with the feature selection and classification method that is used. In Section 4.3 we present the results obtained using texture features by themselves and in conjunction with shape features and compare these results to other methods. Finally, in Section 4.4 we conclude the chapter with a discussion.

4.2 Proposed Method for Improving LBP based AU Detection

In this section we explain in detail the method proposed for the AU detection system. Since our main contribution is in features extraction, the emphasis is also given to this component of the system.

4.2.1 Shape Features

To obtain the shape features we need to localize the face and certain points on it, either by manual human annotations or with the help of a face tracking system. In order to avoid any noise possibly introduced by automatic face tracking and to better observe the improvement provided by the proposed texture based features (explained in Section 4.2.2) we use manual annotations of 68 facial points for the tests presented in this work.

The face is divided into three regions and only a certain group of the facial points are used corresponding to each region. The reason for doing this is that none of the action units causes a substantial change in the whole face or all of the 68 points defined, but only a specific portion. So, we can reduce the computational burden and noise caused by the feature extraction and selection processes. More precisely, we use 29 points and the texture contained inside and around for each of the upper face (AUs 1,2,4,5 and 7), middle face (AUs 6 and 9) and lower face (AUs 12,15,17,20,23,24,25 and 27) action units. The selected points for each type can be seen in Fig.4.1.

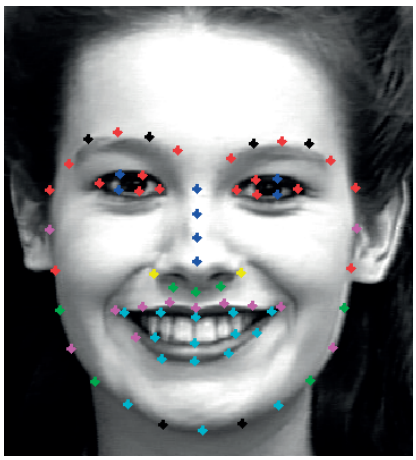


Figure 4.1: Points used in feature extraction; Upper face points are shown in red, blue or yellow; Middle face points in green, blue, magenta or yellow; Lower face points in cyan, magenta or yellow; Black points are not taken into account for any AU

The shape features are then obtained using the initial frame (containing no expression) and peak expression frame (referred to as peak frame throughout the rest of the chapter) of each video sequence containing an expression of N frames, similarly to Valstar & Pantic (2006) with the difference of using only 2 frames rather than the whole sequence. All of the shapes (68 points) were aligned to a single shape to exclude the effect of translation, rotation and scale. The first features obtained is the position change in horizontal and vertical directions of the 29 points defined, which is called set S. Thus, we compute for each point i in S

$$F_1(i) = x_{i,N} - x_{i,1} \tag{4.1}$$
$$F_2(i) = y_{i,N} - y_{i,1} \tag{4.2}$$
$$F_3(i) = \sqrt{(x_{i,N} - x_{i,1})^2 + (y_{i,N} - y_{i,1})^2} \tag{4.3}$$

where $x_{i,N}$ denotes the position in x coordinate of point i in frame number N, or the peak frame, and similarly $x_{i,1}$ that in the first, or neutral, frame.

Then, we also take as features the change in position of all points with respect to each other in the peak and initial frames, i.e.

$$F_4(i,j) = \sqrt{(x_{i,N} - x_{j,N})^2 + (y_{i,N} - y_{j,N})^2} - \tag{4.4}$$
$$\sqrt{(x_{i,1} - x_{j,1})^2 + (y_{i,1} - y_{j,1})^2} \tag{4.5}$$
$$F_5(i,j) = atan\frac{|y_{i,N} - y_{j,N}|}{|x_{i,N} - x_{j,N}|} - atan\frac{|y_{i,1} - y_{j,1}|}{|x_{i,1} - x_{j,1}|} \tag{4.6}$$
$$\tag{4.7}$$

for all points $i \neq j$ in S. Obtaining in the end the feature set $F_s = [F_1, F_2, F_3, F_4, F_5]$ of 899 shape features.

4.2.2 Texture Features

The texture related features that we propose to use are based on LBP histograms obtained from overlapping windows of various sizes. The LBP transform is applied on three images obtained by three different filters and the final features are the histogram differences between initial and peak frames. These filters are the bilateral filter, opening by reconstruction filter and black top-hat by reconstruction filter. We explain in the following subsections how each of them works and why they are relevant to our task, in addition to a brief description of the LBP transform and the feature extraction procedure. More details on LBP and its variants can be found in Sec. 3.2.2 of the dissertation.

4.2.2.1 Bilateral Filter

The first preprocessing method we perform in order to eliminate irrelevant facial deformations or noise present in the image is the bilateral filter. The bilateral filter is a non-linear filter introduced by Tomasi & Manduchi (1998) and has been vastly used mainly

for the purposes of image denoising and for creating special effects in photographs. Its main advantage compared to linear filters is that it smooths an image while preserving the edges with the help of two different kernels called the domain and range filter. The equation of the bilateral filter is given as

$$\hat{I}(p_c) = w_c^{-1} \sum_{k \in Q} e^{-\frac{\|p_c - p_k\|^2}{2\sigma_d^2}} e^{-\frac{(I(p_c) - I(p_k))^2}{2\sigma_r^2}} I(p_c) \tag{4.8}$$

where Q is the particular neighborhood taken around the pixel located at p_c and I denotes the corresponding gray-level intensity. The normalization factor w_c is simply the summation of the weights over the neighborhood Q.

$$w_c = \sum_{k \in Q} e^{-\frac{\|p_c - p_k\|^2}{2\sigma_d^2}} e^{-\frac{(I(p_c) - I(p_k))^2}{2\sigma_r^2}} \tag{4.9}$$

The first kernel in Eq. 4.8 is the simple Gaussian smoothing filter, called the domain filter in this case. The second one, called range filter, is where the non-linearity appears and it smoothes the image in the intensity domain. This means that, the neighboring pixels with intensity values close to the center pixel are assigned a smaller weight than the pixels that have a larger intensity difference. Thus, the areas which contain edges (high intensity changes) are less affected by the smoothing performed by the domain filter.

The bilateral filter is suitable for our case, since our main source of information is contained on the edges created by the facial actions, and we want to smooth out the regions that contain other irrelevant deformations. The main issue with bilateral filters is the choice of the 3 parameters σ_d, σ_r and the neighborhood size, which affect directly the amount of smoothing and edge preserving. No optimization of these parameters exists in the literature and the optimal parameters depend highly on the application, so, in this work, we choose empirically as parameters $\sigma_d = 3$, $\sigma_r = 50$ and a square neighborhood of size 11, which provides a reasonable smoothing. An example result of the bilateral filter and the LBP transform applied on it can be seen in Figures 4.2d and 4.2l. As expected the bilateral filter - LBP transform combination results in smoother regions, so that the main patterns explaining the facial features are better viewed and, of course, identified.

4.2.2.2 Morphological Operations by Reconstruction

The second type of preprocessing that we use is based on mathematical morphology. Opening and closing are two of the most commonly used morphological operations. Morphological opening serves to identify or isolate structures (or connected components) that are brighter than their environment. Morphological closing isolates and flattens image structures that are darker than their surroundings and that have a smaller support than the structuring element (SE), which is used for the consecutive dilation and erosion operations. Depending on the structuring element, the way that the image behaves under these filters thus provides information on structural features of the objects present in the image. They have been frequently used to obtain feature sets using

varying sizes of structural elements in tasks like image classification and segmentation, especially in remote sensing applications (Dalla Mura *et al.*, 2011).

Based on this ability of defining bright and dark structures in images, we adopt the idea of using the morphological filters as a preprocessing method applied before the LBP transform. The standard opening and closing operations, however, result in the deformation of important geometrical structures as well. To prevent this severe effect, a shape preserving method called morphological filtering by reconstruction was proposed (Crespo *et al.*, 1995), with the idea of avoiding deformation of structures larger than the structuring element.

Opening and closing by reconstruction are performed in two steps. In the case of opening, first a marker image I_M is obtained by applying erosion (represented by ϵ) on the original image I, using the structural element B.

$$I_E = \epsilon_B(I) \tag{4.10}$$

The second phase is iteratively performing a geodesic dilation starting with the marker image I_E until no further change in the image pixels is obtained. The geodesic dilation on an image is defined simply as the pixel-wise minimum ($\wedge$) of the elementary dilation (dilation with the smallest structuring element, represented as δ_1) on the image and a mask image, which is in our case the original image, I (Dalla Mura *et al.*, 2011). After n iterations we obtain the opening by reconstruction, I_{OR}, in the form

$$I_{OR} = \delta^n_{1,I}(I_E) = \delta_{1,I}(\delta_{1,I} \ldots (\delta_{1,I}(I_E))) \tag{4.11}$$

with

$$\delta_{1,I}(I_E) = \wedge\{\delta_1(I_E), I\} \tag{4.12}$$

and

$$\delta^{n+1}_{1,I}(I_E) = \delta^n_{1,I}(I_E) \tag{4.13}$$

Closing by reconstruction (I_{CR}) is obtained, similarly, by iteratively applying the geodesic erosion operation on the marker image obtained by dilating the original image with a structural element B, until the resulting image is identical to the one in the previous iteration. The geodesic erosion is defined as the pixel-wise maximum ($\vee$) of the elementary erosion of the marker image and the mask image, which is once again our original image I.

$$I_{CR} = \epsilon^n_{1,I}(I_D) = \epsilon_{1,I}(\epsilon_{1,I} \ldots (\epsilon_{1,I}(I_D))) \tag{4.14}$$

with

$$\epsilon_{1,I}(I_D) = \vee\{\epsilon_1(I_D), I\} \tag{4.15}$$

We use as our morphological preprocessing methods the opening by reconstruction and the black top-hat by reconstruction method. The black top-hat transform (also called the closing by top-hat or top-bottom transform) is the residual of a closing image when compared to the original image:

$$I_{BTR} = I_{CR} - I \tag{4.16}$$

(a) Original neutral image

(b) LBP of the original neutral image

(c) Neutral image with bilateral filter

(d) LBP of neutral image with bilateral filter

(e) Neutral image with morphological opening

(f) LBP of neutral image with morphological opening

(g) Neutral image with top-hat transform

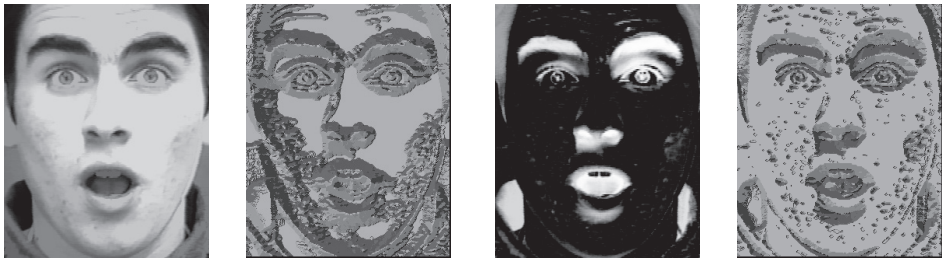
(h) LBP of neutral image with top-hat transform

(i) Original peak expression image

(j) LBP of the original peak expression image

(k) Peak expression image with bilateral filter

(l) LBP of peak expression image with bilateral filter



(m) Peak expression image with morphological opening

(n) LBP of peak expression image with morphological opening

(o) Peak expression image with top-hat transform

(p) LBP of peak expression image with top-hat transform

Figure 4.2: Examples of the preprocessed images and their LBP transforms for the neutral (no AU present) and the peak of expression cases

Example results of the opening by reconstruction, black top-hat transform and the LBP transform applied on top can be seen in Figures 4.2f & 4.2n and 4.2h & 4.2p respectively. As we can see the opening performed serves to flatten the bright areas on the face, emphasizing the important intensity changes caused by the facial features, and to help the LBP transform obtain clearer structures. The black top-hat transform, on the other hand, identifies the dark regions on the face (such as the mouth opening and eyebrows) and therefore cause the LBP to have more significant boundaries around these regions. As the structuring element we use a disk shape of size 30 by 30 pixels, for images of size 640 by 490. All filter parameters were chosen based on visual observations for this work.

4.2.2.3 Feature Extraction by Uniform Local Binary Pattern Histogram Differences

LBP is an efficient gray-scale texture descriptor proposed by Ojala *et al.* (1996) and has been used widely in various texture description and classification problems, including expression recognition and AU detection, along with its many variants. Its main advantage is that it is invariant to illumination changes since it is defined by the relationship of a pixel with its neighbors, thus can identify successfully the microstructures in an image. More details about different LBP types and their usage in the literature can be found in Section 3.2.2.

The basic LBP is defined for a pixel p_c as

$$LBP(p_c) = \sum_{k=0}^{P-1} l(I(p_k) - I(p_c)).2^k \tag{4.17}$$

where $I(p)$ denotes the intensity of a pixel p, and P is the total number of pixels in the chosen neighborhood of the center pixel p_c. The function l is a simple thresholding function in the form

$$l(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases} \tag{4.18}$$

In the end we obtain a binary pattern of P bits for each pixel. By varying this number P and the radius of the circular neighborhood one can obtain LBP at different resolutions. In this work we use the uniform LBP on a neighborhood of radius=1 and $P = 8$. Uniform LBP (Ojala *et al.*, 2002) is an extension of the standard LBP, where the binary patterns are grouped according to the number of 0/1 transitions that they contain, and the patterns containing more than 2 transitions (non-uniform patterns) are assigned the same identity, since it was shown that they occur much less frequently than the others, namely the 58 uniform patters. So, for each pixel in a region of interest we assign a value from 0 to 58, and obtain a 59 bin histogram for that region. Figures 4.2b and 4.2j show the uniform LBP transformation of an example face with and without expression from the CK+ database (Lucey *et al.*, 2010).

In our experiments we scale each face region (upper, middle or lower as explained in Section 4.2.1) in the initial and peak frames to a standard size of 240 to 120 pixels. Then

we obtain the 59 bin uniform LBP histograms of 324 overlapping windows of different sizes, the smallest window size being 40 by 40 while the largest one is 240 by 120 containing the whole region of interest. Figure 4.3 shows an illustration of the windows with the smallest size along with the first two slid versions; the overlap size is $M_1 - 20$ by $M_2 - 20$ for each window of size M_1 by M_2. Most of the works to date using LBP histograms for action unit detection have used standard size non-overlapping windows. However, for each AU the most important information may be contained in windows of different sizes and positioned in various locations. For instance, for AU2 (outer brow raise) the large window containing both of the eye brows is intuitionally more important than the smaller window containing only the inner brows, while for AU1 (inner brow raise) it is not the case. Therefore, we prefer not to discard any of these overlapping regions, and let the feature selection step choose the most relevant ones.

Figure 4.3: Illustration showing the smallest window size used for LBP histogram extraction and the first two overlapping translated versions

Once we have obtained the histograms for each of the windows on each of the initial and peak frames, we compute the histogram variation between the two frames, the reason being, using the change in the LBP profiles rather than the profiles directly in the peak frame eliminates the variations due to identity and provides a stronger feature set (Senechal *et al.*, 2010). Instead of the direct difference of 2 histograms and using every bin as separate features as done by Senechal *et al.* (2010), we use the χ^2 distance, D_{χ^2}, which is defined as

$$D_{\chi^2}(H_N, H_1) = \sum_{b \in B} \frac{(H_N(b) - H_1(b))^2}{(H_N(b) + H_1(b))/2} \tag{4.19}$$

where $H_N(b)$ denotes the value at bin b of the histogram for the Nth frame, and B denotes the set of all the bins. The texture features for the region of concern is thus these distance measures for each of the 324 windows.

Applying the LBP transform and obtaining these texture features explained, for all three of the preprocessed images (bilateral filter, opening by reconstruction, black top-hat by reconstruction) we have our final set of 972 texture related features. The three different filtering methods, combined with the local binary pattern transform, allow us

to obtain an extended set of features explaining the facial structure and as presented in the next section provide a much better AU detection accuracy compared to the LBP used alone, both in combination with the shape features and by themselves.

4.2.3 Feature Selection and Classification

Once the full set of features (shape + texture) is obtained, we perform feature selection using the GentleBoost algorithm (Friedman *et al.* (2000), see Section 3.3.4 for details) to choose the most relevant features for each of the AUs. We therefore perform this process 15 times independently, for the action units 1,2,4,5,6,7,9,12,15,17,20,23,24,25 and 27. Feature selection is a crucial step in the AU detection process, since it discards the irrelevant and redundant features which constitutes a huge portion of the total number of features extracted, due to the large number of LBP windows and inter-point relations we use for building our features set. For each action unit 200 features are extracted in total as result of the GentleBoost, then the optimal number of features is chosen by performing leave-one-subject-out tests (explained in detail in Section 4.3) with 30,50,100,150 and 200 features for each AU separately.

For the detection of action units using these selected features, we train 15 Support Vector Machine (SVM), once again for each AU. The SVM are binary, the classes being if the specific AU is present in the image sequence or not. As kernels we use Gaussian RBF, and optimize the classifier parameters σ and C using a 5-fold cross validation on the training set. The cross-validation tests and parameter optimization are explained in more detail in Section 4.3.

4.3 Experimental Results

For all the experiments that we performed we have used the Extended Cohn-Kanade (CK+) database (Lucey *et al.*, 2010), which consists of a total of 593 image sequences of 123 different subjects posing in various facial expressions and contains different numbers of examples of many action units. The action units present on the peak frame of each sequence were identified by human coders for each sequence. We have applied our methods to detect 15 action units which have a reasonable number of occurrences in the database. We take, for each AU, as positive examples all the sequences that it is present in the peak frame, regardless of the intensity of the action. For the LBP and the three proposed filters we have used our own implementation on C++, for the GentleBoost we have used the method provided within OpenCV[1] and finally for the SVM classification we use the publicly available LibSVM library (Chang & Lin, 2011)[2].

For each of the tests presented, we have performed a leave-one-subject-out (LOO) cross-validation; meaning, all sequences of a specific subject were excluded in the set used to train the classifier, then the classifier was tested on the excluded sequences and the overall accuracy was calculated by adding the number of correctly classified sequences for each subject. The best parameters set $\{\sigma, C\}$ of the SVM (corresponding to

[1] http://opencv.org/

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

the highest classification rate) were chosen out of 25 possible combinations, using a 5-fold cross validation on the training set for each subject. The LOO tests were performed for each AU using 30,50,100,150 and 200 features and the one giving the highest overall accuracy was chosen as the final result.

We group the results we obtained in two parts: The first one is the AU detection performance using only texture features in the feature selection and classification, and compares the two results obtained by the preprocessing methods, explained in Section 4.2.2, applied before the LBP transform and by the LBP transform applied directly on the original image. The second part presents the detection results obtained by using these texture features in conjunction with the geometric features detailed in Section 4.2.1, and compares these results to other methods in the literature that have reported results on the same database.

4.3.1 Experiments with only texture features

First, we train our feature selector and classifiers using only the texture features, not including yet the geometric features, in order to observe the advantage of applying the preprocessing methods proposed over using LBP transform directly on the image by itself. Table 4.1 presents the number of features used, overall accuracy and area under the ROC curves, which are presented in Fig.4.4, for each of the 15 action units and for both methods. The overall accuracy (OA) stands for the correct classification rate for both the positive and negative examples for each AU.

We can see from these results the significant increase in accuracy when we use the extended set of texture features, i.e. with the preprocessing applied. For all AUs we obtain a higher accuracy and AUC with the feature extraction method using the filters, resulting in an average increase of 2.34% in the OA, 4.57% in the AUC, which is more meaningful than the OA due to the unbalanced number of positive and negative examples. The number of features giving the highest accuracy in each case is particularly interesting, since for certain AUs this number is higher for the method using only LBP, although the total number of features before feature selection is only one third of the other method (324 vs. 972). This fact serves to show us that the increase in accuracy is not at all dependent on the number of features extracted but rather on their ability to describe the facial actions.

These tests show not only the advantage of the preprocessing methods proposed, but also the potential of the system when it is completely automated. The texture features are mostly independent from the facial point annotations, for which we used manual annotations at this step, except for obtaining the relevant region part of the face. This can be easily and efficiently performed using existing facial landmark detection methods in the literature (as performed in the following chapters) and we see, as explained in the following section and presented in Table 4.2 that we achieve accuracy measures competitive with other state-of-the-art methods even using only texture features.

Table 4.1: AU Detection Results for the preprocessing + LBP texture features (Pre+LBP) and for only LBP texture features (LBP). NP: Number of positive examples for the AU in the database, nFts: Number of features used, OA(%): Percentage overall accuracy, AUC(%): Area under ROC curve

AU	NP	nFts		OA		AUC	
		Pre+LBP	LBP	Pre+LBP	LBP	Pre+LBP	LBP
1	177	200	150	**0.909**	0.862	**0.959**	0.889
2	117	200	150	**0.941**	0.933	**0.977**	0.929
4	194	100	150	**0.879**	0.821	**0.942**	0.882
5	102	30	100	**0.919**	0.890	**0.936**	0.871
6	123	100	150	**0.890**	0.869	**0.929**	0.885
7	121	100	150	**0.836**	0.836	**0.867**	0.831
9	75	50	150	**0.975**	0.955	**0.992**	0.958
12	131	150	50	**0.934**	0.909	**0.954**	0.933
15	94	50	150	**0.922**	0.880	**0.936**	0.894
17	202	150	50	**0.895**	0.865	**0.940**	0.911
20	79	100	100	**0.926**	0.921	**0.949**	0.916
23	60	100	150	**0.922**	0.894	**0.882**	0.779
24	58	50	150	**0.926**	0.914	**0.896**	0.843
25	324	150	100	**0.880**	0.858	**0.937**	0.925
27	81	100	150	**0.961**	0.959	**0.987**	0.950
Avg.				**0.914**	0.891	**0.939**	0.893

4.3.2 Experiments with shape and texture features combined

The second group of experiments we perform is using the shape features (explained in Section 4.2.1) in combination with the texture features explained in Section 4.2.2. Once again we conduct the experiments using the LBP on top of three preprocessing methods, and using LBP directly on the image separately. In the first case the feature selection algorithm is fed 1871 features in total, while in the second this number is 1231. In this study aiming to test the efficiency of the proposed texture features we use only manual annotations of the facial points in order to eliminate the bias factor introduced by the possible noise from facetracking. Due to the high accuracy of these features and the ratio of the shape vs. texture features, the feature selection tends to select shape features more frequently in the LBP features without preprocessing case, as expected. Therefore, the difference in accuracies obtained by the two different methods is less significant than that presented in Section 4.3.1. With the preprocessed features we obtain 94.74% overall accuracy and 96.97% AUC, while with only the LBP features we obtain 94.13% accuracy and 96.01% AUC as average over the 15 AUs tested.

The preprocessed features achieve higher accuracy and AUC for 12 AUs, the exceptions being AU 23 and 24 for only the overall accuracy, which is rather meaningless since they have very few positive examples, and AU25 (jaw drop) for both accuracy and AUC, which has proven by the performance difference between using shape+texture features and only texture features (shown in Table 4.2), to be very dependent on the features

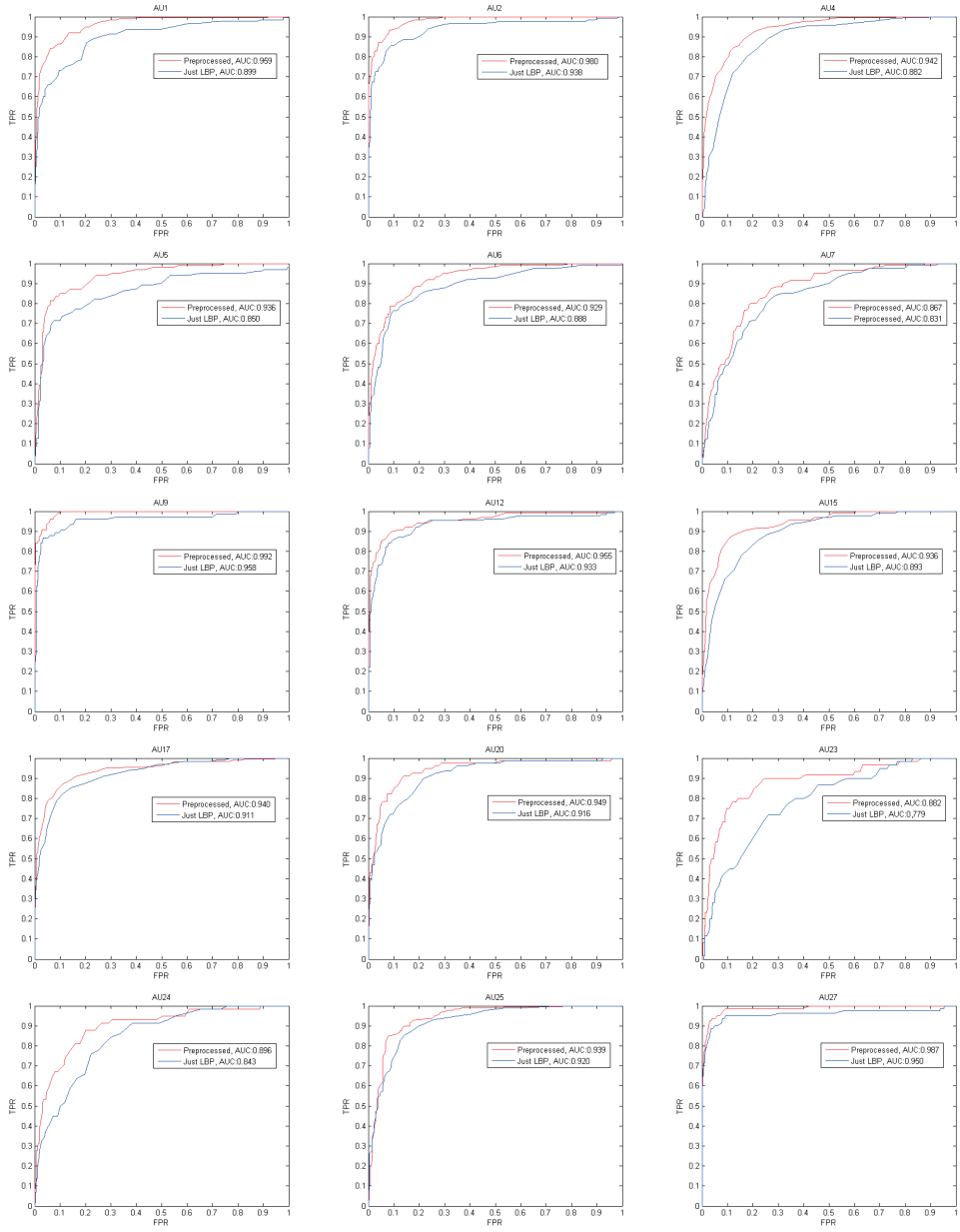


Figure 4.4: Receiver Operator Characteristics curves for each of the Action Units included in the experiments. Red curves are the ones obtained using preprocessing and LBP texture features, while blue curves are the ones obtained using only LBP texture features

provided by the geometry of the facial points rather than the texture. Comparing these two performances (shape+texture vs. texture) we see that while shape features bring about a higher accuracy in all AUs, for some of them this change is more substantial, like AU1 (inner brow raise) in addition to AU25. This tells us that for these AUs, change

of location of facial points contains more important information than the change in texture contained in or around. It makes complete sense in the case of AU1 and AU25, for example, where we do not see a significant texture variation on the area related to these actions but an obvious position change of certain facial points.

We also compare our results with three different methods that have reported results on the same database. The first one is the method by Senechal et al. (Senechal *et al.*, 2010) in which they use as features the histogram differences of Local Gabor Binary Patterns (LGBP) in non-overlapping fixed size windows, and build a special kernel using this difference for the classifier. Since separate AU performances were not reported and the lower AUs are not the same ones tested in this work, we can only compare the mean upper AU detection performance. The best results that they achieve is with the special kernel which is 97.3% AUC, while for us this measure is 96.8%. With, the Gaussian RBF kernel, however, they achieve 96.2%, from which we can deduct that with a much lower number of features selected efficiently, higher performances can be achieved.

Table 4.2: AU Detection Results comparison using our method with shape + texture features (SHTXT), our method with texture features only (TXT), the method proposed by Valstar *et al.* (2012) (referred to as Valstar) and the method proposed by Bartlett *et al.* (2006) (referred to as Bartlett). OA: Overall accuracy, F1: F1 measure, AUC: Area under ROC curve

AU	OA				F1			AUC		
	SHTXT	TXT	Valstar	Bartlett	SHTXT	TXT	Valstar	SHTXT	TXT	Bartlett
1	0.965	0.909	0.918	0.92	0.938	0.841	0.826	0.983	0.959	0.95
2	0.976	0.941	0.939	0.88	0.939	0.836	0.833	0.991	0.977	0.92
4	0.911	0.879	0.870	0.89	0.862	0.809	0.630	0.968	0.942	0.91
5	0.944	0.919	0.904	0.92	0.829	0.745	0.596	0.976	0.936	0.96
6	0.911	0.890	0.930	0.93	0.778	0.716	0.811	0.946	0.929	0.96
7	0.882	0.836	0.870	0.88	0.688	0.531	0.290	0.917	0.867	0.95
9	0.992	0.975	0.928	1	0.966	0.895	0.573	0.998	0.992	1
12	0.944	0.934	0.930	0.95	0.865	0.838	0.836	0.974	0.954	0.98
15	0.953	0.922	0.969	0.85	0.839	0.726	0.361	0.956	0.936	0.91
20	0.963	0.926	0.908	0.92	0.849	0.690	0.517	0.973	0.949	0.84
24	0.946	0.926	0.935	0.92	0.682	0.511	0.497	0.945	0.896	0.88
25	0.959	0.880	0.851	0.89	0.963	0.889	0.748	0.984	0.937	0.93
27	0.985	0.961	0.964	0.99	0.945	0.855	0.854	0.996	0.987	1
Avg.	**0.949**	0.915	0.916	0.909	**0.857**	0.760	0.638	**0.969**	0.943	0.926

The comparison with the other two methods can be seen in Table 4.2 for the 13 common AUs that were tested in all three works. The first method (Valstar *et al.*, 2012) proposes using as features only the position change of facial points throughout the whole sequence and does not report the AUC measure so we compare the F1 measure instead, noting that we tune our parameters to give the highest classification accuracy and not the highest F1. The second method (Bartlett *et al.*, 2006) uses only Gabor features with an Adaboost classifier. We achieve in average, and for most of the action units, superior performance compared to the 2 methods, both when we use shape and texture features together and when we use only the texture features. Once again, the shape features we use depend highly on the accuracy of the facial points, for which we have only used human annotations at this stage, but the promising accuracy measures obtained for both types of features already show the strength of the proposed features in detecting action

units.

4.4 Conclusion

In this chapter we have presented a simple, novel and efficient method for extracting features for AU detection in videos that is based on LBP applied separately to images processed by three different filtering methods, namely the bilateral filter, opening by reconstruction and black top-hat by reconstruction. The results obtained show that this method provides a significant increase in the accuracy measures for all 15 action units tested compared to using LBP by itself.

We have also used the extracted texture related features along with certain transient geometric features, and demonstrated that we achieve performances superior to existing approaches tested on the same database. The presented results show the performance by using human annotations for calculating the geometric features, the reason being to show the efficiency of the proposed appearance-based feature amelioration and to avoid possible bias introduced by face tracking error. Our experiments using only texture features, which are mainly independent of the tracked points, result in very high performances already, proving the strength of the features proposed in detecting facial actions. Our contribution in this chapter shows that using simple, yet competent, image processing methods the performance of LBP based appearance features for AU detection can be increased substantially.

Multiple LCGBPs for Facial Action Unit Recognition

5

5.1 Introduction

As already reviewed in previous chapters the FACS is the most objective means of describing and quantifying facial actions (see Section 2.2.1) and automatic detection of the AUs have proven very useful in many facial analysis applications especially for HCI. Curvature Gabor features have recently been shown to be powerful facial texture descriptors with applications on face recognition (see Section 3.5 for a short review). In this chapter we introduce their use in facial action unit (AU) detection within a novel framework that combines multiple Local Curvature Gabor Binary Patterns (LCGBP) on different filter sizes and curvature degrees. The proposed framework proves to provide very accurate detection results, which is an important quality of automatic facial analysis systems.

In this work we propose as features the variation among frames of a combination of LCGBP as descriptors of facial action. LCGBP is an extension to the LGBP which have been used extensively for face recognition and AU detection (e.g. Zhang *et al.* (2005), Senechal *et al.* (2010)), since they have proven to be quite robust against variations of conditions such as illumination. By adding the affect of curved formations, which are common in the facial texture, the curvature Gabors provide a much more efficient way of representing the facial components (Hwang *et al.*, 2011) and have already been shown to be successful in recognizing facial identity (Arar *et al.*, 2012). Here, we apply this idea by using the change in LCGBP histograms between neutral and expressive images for detecting the AUs. It has been shown that using this variation of histograms between frames is more efficient than using the histograms themselves directly (Senechal *et al.* (2010), Yüce *et al.* (2013a) - our contribution presented in Chapter 4). The main contribution of this work is introducing a unique way of extracting Gabor features, which includes the curvature information and proves to be very powerful descriptors for facial actions by the very high accuracy results.

The majority of the work included in this chapter has already been presented in Yüce *et al.* (2013b). The rest of the chapter is formed as follows: First, we explain the formulation of LCGBP in Section 5.2, then in Section 5.3 we describe the framework that we

propose for AU detection and detail the parameter selection and test settings. Section 5.4 presents the results obtained by several experiments on two databases and comparisons with other types of features and some existing methods in the literature. Finally, we report our conclusions, discuss the advantages and weak points of the proposed system and possible future directions for further improving the system in Section 5.5.

5.2 Local Curvature Gabor Binary Patterns

5.2.1 Curvature Gabor (CG) Wavelets

Gabor wavelets have been recognized as one of the most successful feature extraction methods for face representation. They form a well-established image decomposition because of their spatial locality and orientation selectivity characteristics. Therefore, they are optimally localized in the space and frequency domains, and can be used successfully in facial image processing for face and facial expression recognition and analysis.

The conventional Gabor wavelet definition is as follows:

$$\psi(\vec{x};\nu,\mu) = \frac{k_{\nu,\mu}^2}{\sigma^2} e^{(-\frac{k_{\nu,\mu}^2 \|\vec{x}\|^2}{2\sigma^2})} [e^{(ik_{\nu,\mu}\vec{x})} - e^{(-\frac{\sigma^2}{2})}] \tag{5.1}$$

where $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} xcos\phi + ysin\phi \\ -xsin\phi + ycos\phi \end{pmatrix}$ and $e^{(ik_{\nu,\mu}\vec{x})}$ is the oscillatory wave function, whose real part and imaginary parts are respectively the cosine and sine functions. μ controls the orientation of the filters while ν scales the center of the filter in the frequency domain (Daugman, 1985).

A typical neutral face image contains curve-like features because it contains permanent facial components such as eyes, nose, cheeks, lips, and eyebrows as well as straight features. Since facial expressions are generated by the movement of groups of muscles in any orientation and transient features like wrinkles and furrows, images with expressions contain even more curvature characteristics than straight ones. Therefore, as one way of modeling curve-like features, we include CG wavelets for face representation in addition to the conventional Gabor wavelets.

Peters *et al.* (1997) obtained CG wavelets by adding a curvature parameter to the conventional Gabor formulation as follows:

$$\psi(\vec{x};\nu,\mu) = \frac{k_{\nu,\mu}^2}{\sigma^2} e^{(-\frac{k_{\nu,\mu}^2 \|\vec{x}\|^2}{2\sigma^2})} [e^{(ik_{\nu,\mu}\acute{x})} - e^{(-\frac{\sigma^2}{2})}] \tag{5.2}$$

$$\vec{x} = \begin{pmatrix} \acute{x} \\ \acute{y} \end{pmatrix} = \begin{pmatrix} xcos\phi + ysin\phi + c(-xsin\phi + ycos\phi)^2) \\ -xsin\phi + ycos\phi \end{pmatrix} \tag{5.3}$$

where c corresponds to the curvature ratio.

CG wavelets do not have the orientation symmetry as in conventional Gabor wavelet as shown in Fig. 5.1 (Arar *et al.*, 2012). For the conventional Gabor wavelet setting, it is usually sufficient to have 8 orientations. However, this number should be increased to 16 to obtain the same orientation utilization in case of CG wavelets.

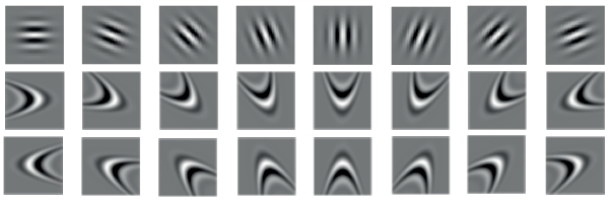


Figure 5.1: Illustration of orientation asymmetry in CG wavelets with $c = 0.1$ (middle and bottom row) in comparison with the conventional Gabor wavelet (top row).

In CG wavelets, one can use different curvature degrees, i.e., $c = \{0.05, 0.1, 0.2\}$, and Gaussian sizes, i.e., $\sigma \in \{0.5\pi, \pi, 2\pi\}$, for multi-curvature utilization as well as scale space utilization. In this way, a much stronger representation power of modeling facial structures is obtained by extracting both fine and coarse features with straight and curved filters as can be seen applied on an example image from the MMI database (Valstar & Pantic, 2010) in Figure 5.2.

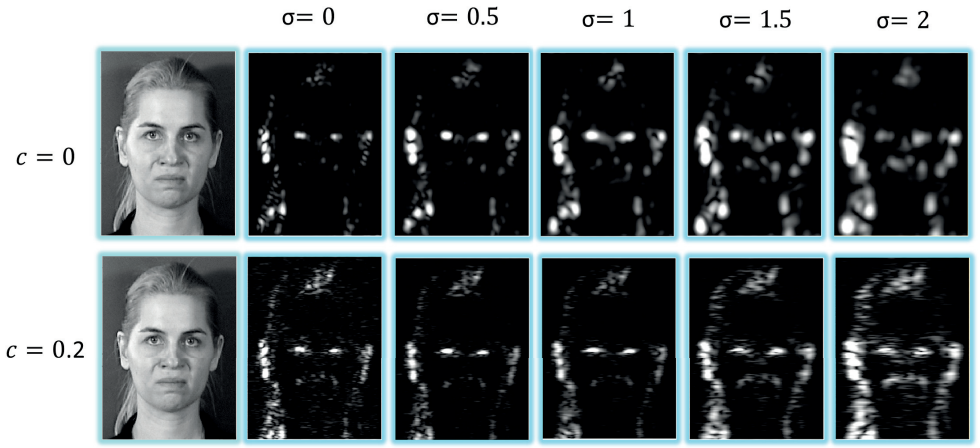


Figure 5.2: Gabor wavelets of different Gaussian sizes (σ) and curvature degrees (c) applied on an image

5.2.2 Local Binary Patterns

The LBP transformation has been proposed as a texture description method (Ojala *et al.*, 1996) and has proven to be very effective in representing facial texture and been widely used for both face and facial action recognition as explained in the previous chapter and in Section 3.2.2. It maps the texture variation around each pixel to a binary pattern and the histogram of these patterns in a local window can be used directly as a descriptor for that certain region of interest. The computation of the pattern for a pixel at position x of an image I is as follows:

$$LBP_P(x) = \sum_{p=0}^{P-1} t(I(x_p) - I(x)).2^p \tag{5.4}$$

In this representation, each $I(x_p)$ is a neighboring pixel of the center pixel $I(x)$ on a neighborhood defined by the number of pixels P as well as the shape (rectangular or circular) and the distance to the central pixel which determines the resolution of the transformation. The function $t(x)$ is the simple thresholding function which returns 1 if the input pixel difference is positive and 0 if it is negative. In this way we obtain a $P-bit$ binary value, or an integer between 0 and $2^P - 1$ to represent each pixel.

In this work we use an 8-pixel circular neighborhood with the radius 1, giving 256 possible patterns. It has been shown, however, that only 58 of these patterns, called the uniform patterns, contain the important part of the texture information (Ojala *et al.*, 2002). So, we can reduce the size of the descriptor to 59 bins by assigning all the non-uniform patterns into a single bin.

Applying the LBP on top of Gabor magnitude images with various scales and orientations results in obtaining a richer representation and finer description of the facial texture (Zhang *et al.*, 2005). In our work we extend this variation of descriptors by also including multiple curvature degrees and Gaussian sizes, obtaining the Local Curvature Gabor Binary Patterns (LCGBP) representation as can be seen in Figure 5.3. Of course, this extension substantially increases the number of features obtained, in addition introduces more redundancy between features and possible noise for the final classification task. Therefore, whether using directly the LCGBP histogram bins as features or, as we perform in this particular work, using a dissimilarity measure for the histograms between frames, a feature selection or dimension reduction technique is essential to be able to perform a meaningful classification using these features. The details on how we compute the histogram dissimilarity as well as the feature selection technique and the types of selected features are explained in more detail in the following sections.

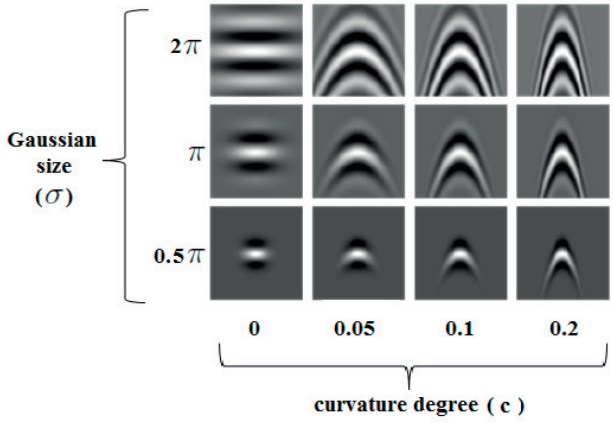


Figure 5.3: Visualisation of Gabor wavelets with various Gaussian sizes and curvature degrees

5.3 Facial Action Recognition Framework

This section describes in detail each step in our automated facial action unit detection system using LCGBP as seen in Fig. 5.4.

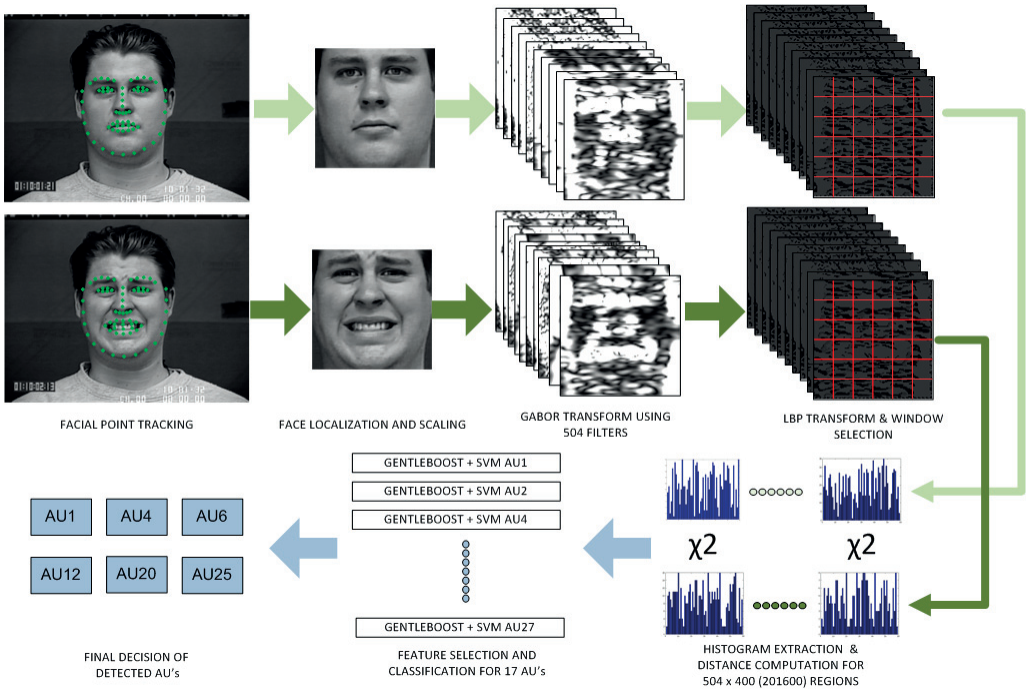


Figure 5.4: Complete flowchart of the proposed framework for an input video

5.3.1 Face Localization

To be able to perform an effective feature extraction among all images in the dataset, we first need to locate our region of interest, which is the face, as accurately and consistently as possible. Existing face detection systems, which output a rectangular region around the face, are generally not reliable enough to extract appearance features since the performance varies across subjects, expressions and head poses. Therefore, we choose to use a facial point tracking system instead, which provides more stable boundaries for the face region.

In this work, we localize 66 facial landmarks as seen in Fig. 5.4, using a publicly available automatic face tracking system proposed by Saragih *et al.* (2011). The face tracker is based on constrained local models (CLM) (Cristinacce & Cootes, 2006) with regularized landmark mean-shift as the fitting strategy. The CLM, similar to the Active Appearance Model (AAM) (Cootes *et al.*, 2001), uses a combined model for the shape and texture, but the model in CLM consists of templates of appearance around each facial landmark point, as explained in more detail in Section 3.1.2. This local nature of the CLM combined with the fitting algorithm proposed in Saragih *et al.* (2011) allows accu-

rately tracking facial points even under extreme head poses, intensive facial expressions and presence of occlusions.

Once we locate the facial landmarks using the face tracker, we crop the image using the most extreme landmarks on the facial mask obtained, with a certain safety margin (Fig. 5.4). No registration or texture warping is performed, since the databases that we use to train and test our system were recorded in quite constrained situations with respect to head pose and since the types of features we use have proven to be robust against misalignment. Since the databases that we use in this work contain very little pose variation we only scale each detected face region to a fixed size of 120 by 120 pixels, and do not perform any other registration.

5.3.2 Feature Extraction

After locating and scaling the face region we extract the appearance features using a combination of LCGBP transforms, which is the LBP transform applied on top of the image filtered by various curvature Gabor wavelets, as explained previously in Section 5.2 and as represented in Fig. 5.4. For our training and testing purposes we apply this filtering to the frame with a neutral expression and the frame with the peak of the posed expression separately for each sample video, since we utilize the comparison between those frames. For the CK+ database (Lucey *et al.*, 2010) these frames correspond to the first and last frames respectively, while on the MMI database (Valstar & Pantic, 2010) they are once again the first frame and the one marked as the first frame of apex phase of the expression (see Sec. 2.2.1 for a definition), obtained from the annotation provided within the database for certain recordings. At this point the system requires that a frame is marked as neutral expression, then the method can be applied to any other frame of the same subject to detect action units at different intensities. This automatization problem can be effectively solved by projecting the subject face with any expression to the PCA space created by examples of expressionless faces, as proposed in Senechal *et al.* (2010). However, we have not tested this method in the scope of this thesis.

The first step of feature extraction is applying the Gabor transforms to the input images. The classic method for generating Gabor representations of images is to apply wavelets in different scales and orientations with a fixed Gaussian size. In addition to adding the curvature component in various degrees we also include wavelets with different Gaussian sizes, similar to Arar *et al.* (2012). This is expected to result in a richer representation of finer details of facial texture components, which are crucial for high accuracy action recognition, compared to a single Gaussian size, and so is proven with our test results (presented in the following section). To be more precise we use Gabor wavelets of 3 different scales ($\nu \in \{0, 1, 2\}$), 8 (or 16 in case of curvature because of the asymmetry, see Fig. 5.1) orientations ($\mu \in \{0, \cdots, 7\}$), 3 Gaussian sizes ($\sigma \in \{\pi/2, \pi, 2\pi\}$) and 4 curvature degrees ($c \in \{0, 0.05, 0.1, 0.2\}$). This results in a total of 504 separate filters ($1 \times 3 \times 3 \times 8 + 3 \times 3 \times 3 \times 16$).

Next we apply the uniform LBP transform on each of the magnitude images of the outputs of these 504 filters for both the neutral and peak expression frame. Then to obtain the local texture information we calculate the histograms on 400 overlapping windows of sizes 20 by 20, 20 by 40, 40 by 20 and 40 by 40 with an overlap size of 10, all units

in pixels. The conventional tendency in the literature for LBP histogram extraction has been to use non-overlapping windows of a fixed size, but as shown in Chapter 4 of the dissertation, varying the size and performing a more extensive search using overlaps, combined with a powerful feature selection step, results in a more informative feature set. Then we compute for each of these windows the χ^2 distance of corresponding histograms in the neutral and peak expression frames, and obtain our full set of features of size 201600 (400×504). Using these alterations from the neutral face as features not only eliminates the variation caused by identity (Senechal *et al.* (2010), Yüce *et al.* (2013a)) but also allows tracking the relative intensity of the movement between frames.

5.3.3 Relevant Feature Selection and AU detection

The extensive representation and search strategy chosen in the feature extraction technique results in a huge number of features which causes two main problems. The first problem is that most of these features are correlated with each other so using them in combination in a classification task introduces an unnecessary computational burden. Secondly, only a portion of them are relevant to the task, i.e. detecting a specific action unit. The irrelevant features cause only noise and a decrease in accuracy in classification. Therefore we need to use a feature selection method that addresses both of these problems and that is specific to each action unit. Boosting techniques allow both reducing the dimensionality of the feature vector and eliminating the irrelevant features, since they are trained in a manner that maximizes the classification rate.

We adopt in this work the GentleBoost technique, since it has already been shown in the literature to be effective when used in combination with Support Vector Machines (SVM) (Valstar *et al.* (2012),Yüce *et al.* (2013a)), which is the classification method that we utilize. For 17 AUs, which have a reasonable number of examples in our training database (CK+), we select 1000 features out of 201600 using GentleBoost separately, so we obtain the most relevant features in terms of Gabor scale, size, orientation and curvature ratio as well as the location in the 2D space. Then we train, once again for each AU, an SVM , for which the two output classes are whether the AU is present or not. We perform the leave-one-subject-out tests on the CK+ database for each AU using 100, 200, 300, 400, 500, 750 and 1000 features in the SVM and at each case choose the number of features giving the highest overall accuracy rate. Using the publicly available LibSVM implementation (Chang & Lin, 2011) we have performed the tests with both linear-SVMs and RBF kernels (parameters optimized using a 5-fold cross validation). Here, however, we only report results using the RBF kernels, since they result in better accuracy compared to the linear SVM in every AU, but there is no substantial difference when comparing different types of features. These results are presented in the following section.

5.4 Experimental Results

In this section we report the results of our experiments performed on the Extended Cohn-Kanade database of facial expressions (CK+) (Lucey *et al.*, 2010), where each sample video starts with a neutral expression and ends with the peak of the expression.

We train and test our system using only this final frame of each sequence. We have implemented the LBP and Gabor filters on C++, for the GentleBoost we have used the method provided within OpenCV[1] and finally for the SVM classification we use the publicly available LibSVM library (Chang & Lin, 2011)[2]. The code for the CLM based facetracker is also publicly available[3].

All presented results are those obtained by a leave-one-subject-out test, i.e. training the Gentle-SVM classifier on samples of 122 subjects and testing it on the remaining subject. We also present our results obtained by training the system on CK+ and testing on 253 videos of the MMI database (Valstar & Pantic, 2010) to demonstrate the generalizability property of the system. The used videos are those annotated for the temporal phases (onset, apex, offset) of the AUs (see Sec. 2.2.1), and once again we use only the first frame annotated as apex, as the test frame for each video.

5.4.1 Comparing types and combinations of Gabor features

We first compare the test results obtained by various parameter settings for the LCGBP and also using only LBP as a baseline comparison method. All settings are kept the same for this comparison, except that for the LBP the maximum number of features tested in the training phase of SVM is kept at the physical maximum, i.e. 400.

We have tested 14 configurations in addition to the standard LBP features; namely 12 settings for LCGBP with 3 scales and 8 (or 16) orientations and a fixed Gaussian size (σ) chosen from 0.5π, π or 2π and fixed curvature degree (c) from $0, 0.05, 0.1, 0.2$ (0 meaning standard LGBP with 9600 total features, each of the rest yields 19200), one setting combining all proposed σ choices with $c = 0$ (28800 features) and one setting combining all possible σ and c choices (201600 features), which is the setting for the main proposed system. The comparison in three types of accuracy measures (overall accuracy, F1 and AUC of the ROC) averaged over 17 AUs (Upper face AUs 1, 2, 4, 5, 6, 7, 9 and lower face AUs 11, 12, 15, 17, 20, 23, 24, 25, 26, 27) can be seen in Fig. 5.5 and Table 5.1 for a more detailed view. The ROC curve was obtained by alternating the SVM decision threshold.

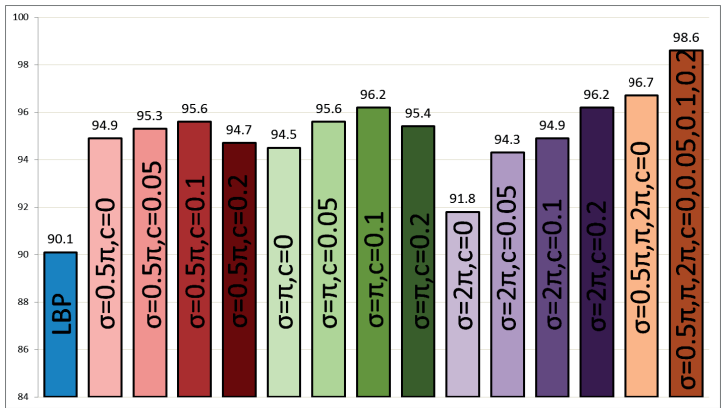
The first observation, other than the definite superiority of LGBP to standard LBP, is that for all fixed σ settings the curvature Gabors perform significantly better than the non-curvature standard Gabor setting, which is the first indication of the effectiveness of curvature features for facial action recognition. Another important comparison is the one between the 4 non-curvature LGBP settings. Using different sizes of Gaussians in the Gabor formulation in combination with each other results in a substantial increase in accuracy with respect to any fixed σ configuration. This indicates the necessity of alternating the Gaussian size along with the scale and orientation in any Gabor setting, which contradicts with the usual tendency in the literature for selecting Gabor wavelets for facial expression or action unit detection.

The proposed setting, which is combining 3 different σ values and 4 different curvature degrees gives the highest classification accuracy for all action units, as expected.

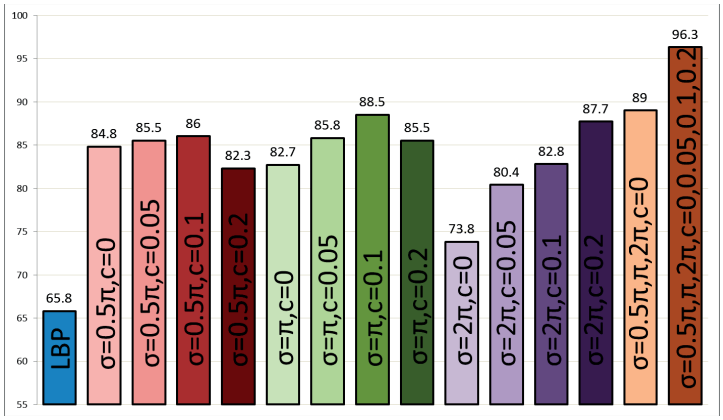
[1] http://opencv.org/

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

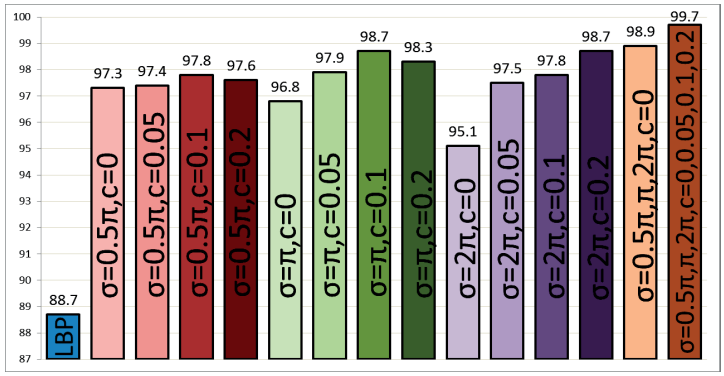
[3] https://github.com/kylemcdonald/FaceTracker



(a) Overall Accuracy (%)



(b) F1 (%)



(c) Area Under ROC Curve (%)

Figure 5.5: Comparison of three different accuracy measures for different LCGBP feature settings & LBP

The results for each AU tested can be seen in Table 5.2 in comparison with the non-curvature case combining different σ's. The superiority is clearly not because of the greater number of features extracted (201600 vs. 28000), but because the various cur-

	OA	F1	AUC
LBP	0.901	0.658	0.887
$\sigma = 0.5\pi, c = 0$	0.949	0.848	0.973
$\sigma = 0.5\pi, c = 0.05$	0.953	0.855	0.974
$\sigma = 0.5\pi, c = 0.1$	0.956	0.860	0.978
$\sigma = 0.5\pi, c = 0.2$	0.947	0.823	0.976
$\sigma = \pi, c = 0$	0.945	0.827	0.968
$\sigma = \pi, c = 0.05$	0.956	0.858	0.979
$\sigma = \pi, c = 0.1$	0.962	0.885	0.987
$\sigma = \pi, c = 0.2$	0.954	0.855	0.983
$\sigma = 2\pi, c = 0$	0.918	0.738	0.951
$\sigma = 2\pi, c = 0.05$	0.943	0.804	0.975
$\sigma = 2\pi, c = 0.1$	0.949	0.828	0.978
$\sigma = 2\pi, c = 0.2$	0.962	0.877	0.987
$\sigma = 0.5\pi, \pi, 2\pi, c = 0$	0.967	0.890	0.989
$\sigma = 0.5\pi, \pi, 2\pi, c = 0, 0.05, 0.1, 0.2.$	0.986	0.963	0.997

Table 5.1: Overall accuracy (OA), F1 and area under ROC curve (AUC) values for different settings (averaged over 17 AU's)

vature degrees and filter sizes allow extracting those that are relevant to each specific action unit. We observe that for some AUs the difference between the two cases is less significant than others, and this can be explained by the variation of amount of curvature that shapes the deviation from the resting state for each action unit. However, observing Table 5.3, which shows the ratio of features chosen by the GentleBoost with respect to σ and c values and the deviation among action units, we can say that none of the types of features show a too powerful dominance over others in none of the AUs, although the non-curvature features are selected significantly less frequently than the rest. This suggests that every type of feature chosen is of similar importance to the detection task in hand and their combination is essential for such a highly accurate classification performance.

5.4.2 Comparison with existing work

We compare our results, as shown in Table 5.4, with three recently conducted works: Valstar *et al.* (2012), Senechal *et al.* (2010) and our previous work in Yüce *et al.* (2013a) (also presented in Chapter 4) which have reported results on the Cohn-Kanade database and have used similar techniques either in the feature extraction or the classification phase. Valstar *et al.* (2012) have used the evolution of certain facial landmarks throughout the video sequence as features and utilized the Gentleboost and SVM as the feature selection and classification methods. In Yüce *et al.* (2013a), we have also used Gentleboost and SVM with a combination of shape features similar to Valstar *et al.* (2012) and LBP features that are improved with the help of three different preprocessing filters (see Chapter 4 for details). The work on Senechal *et al.* (2010), on the other hand, uses as features directly the bins of histogram difference of LGBP magnitude images extracted

AU	No Feat.		OA		F1		AUC	
	Curv.	No curv.	Curv.	No curv.	Curv.	No curv.	Curv.	No curv.
AU1	750	750	0.976	0.958	0.959	0.928	0.995	0.983
AU2	1000	1000	0.992	0.987	0.978	0.965	0.997	0.998
AU4	750	750	0.963	0.935	0.942	0.897	0.994	0.976
AU5	750	1000	0.985	0.965	0.956	0.895	0.997	0.992
AU6	1000	1000	0.985	0.955	0.963	0.884	0.998	0.991
AU7	750	750	0.968	0.936	0.917	0.835	0. 996	0.969
AU9	750	300	1	0.995	1	0.979	0.998	0.994
AU11	1000	300	0.997	0.979	0.969	0.786	0.999	0.984
AU12	1000	1000	0.988	0.968	0.973	0.923	0.998	0.994
AU15	1000	750	0.988	0.969	0.962	0.897	0.999	0.993
AU17	1000	1000	0.975	0.956	0.963	0.935	0.993	0.989
AU20	500	1000	0.983	0.975	0.937	0.905	0.996	0.991
AU23	750	500	0.993	0.971	0.967	0.838	0.999	0.993
AU24	750	1000	0.993	0.965	0.964	0.796	0.999	0.989
AU25	500	1000	0.979	0.966	0.982	0.969	0.994	0.994
AU26	1000	1000	0.989	0.959	0.938	0.721	0.999	0.987
AU27	200	1000	0.998	0.995	0.994	0.981	0.999	0.999
Avg.			**0.986**	0.967	**0.963**	0.89	**0.997**	0.989

Table 5.2: Number of features giving the maximum overall accuracy, Overall accuracy (OA), F1 and area under ROC curve (AUC) values for combinations of LCGBP (Curv.) and LGBP (No curv.) for 17 AUs

from 16 non-overlapping windows with a fixed Gaussian size and no curvature, and as classification adopts SVM with a specially trained kernel.

As seen in Table 5.4, our method certainly outperforms all the other state-of-the-art methods on the CK+ database in AU detection accuracy. The comparison with the two methods (Valstar *et al.* (2012) and Yüce *et al.* (2013a)) using the same type of feature selection and classification, shows the efficiency of the type of features utilized in our system, while the comparison with Senechal *et al.* (2010), which uses a rather complicated classification scheme, proves the utility of using curvature based features and combining different sizes of Gabor wavelets.

5.4.3 Cross database performance

We also test our system trained on the CK+ database on 253 sample videos from the MMI database (Valstar & Pantic, 2010), which were annotated for the temporal phases of the expression, to obtain the peak expression frame. This allows us to reach an idea on the generalizability of the system, since the two databases were recorded independently. The F1 and AUC results obtained, as well as the F1 results from Valstar *et al.* (2012) for the common AUs can be seen in Table 5.5. The comparison with Valstar *et al.* (2012) is not necessarily firm, since the number of training and testing examples in the 2 cases are different; yet it gives the reader a view in the state-of-the-art for cross-database performance.

c \σ	0.5π	π	2π	Total
0	3.7 ± 0.6	4.8 ± 0.9	5.4 ± 1.4	13.9 ± 1.6
0.05	7.6 ± 1	9.2 ± 1.2	10.6 ± 1.4	27.3 ± 2.2
0.1	7.6 ± 1.2	8.8 ± 1	11.3 ± 1.1	27.7 ± 1.1
0.2	8.2 ± 1	9.7 ± 0.8	13.3 ± 2.6	31.1 ± 2.3
Total	27 ± 2.6	32.5 ± 1.8	40.5 ± 2.7	

Table 5.3: Mean and standard deviation of percentage of features chosen from different Gaussian sizes defined by σ and curvature values (c).

Type of acc.	F1(%)		AUC(%)		AUC(%)	
No of AUs	14		16		14	
Method	Valstar *et al.* (2012)	O.W.	Senechal *et al.* (2010)	O.W.	Yüce *et al.* (2013a)	O.W.
	61.86	**96.09**	96.45	**99.69**	96.9	**99.65**

Table 5.4: Accuracy comparison of our method with three other methods; No of AUs represents the number of common AUs taken into accuracy consideration and "O.W." stands for our work

	Valstar *et al.* (2012)	Our work			Valstar *et al.* (2012)	Our work	
AU	F1	F1	AUC	AU	F1	F1	AUC
1	0.255	0.481	0.871	12	0.400	0.455	0.764
2	0.467	0.471	0.872	15	0.229	0.200	0.673
4	0.414	0.493	0.758	20	0.341	0.143	0.563
5	0.149	0.154	0.501	24	0.292	0.154	0.667
6	0.571	0.302	0.673	25	0.746	0.608	0.715
7	0.211	0.174	0.796	26	0.203	0.045	0.527
9	0.286	0.519	0.916	27	0.591	0.524	0.956
Avg.	0.368	0.337	0.732				

Table 5.5: Accuracy results of testing our system on the MMI database with the training on CK+ and F1-measure comparison with Valstar *et al.* (2012) for the same experiment settings.

The significantly lower results compared to the within database results formerly presented may be related to a number of factors. First one is that there are certain differences in the way that the action units are posed and annotated in the two databases, so the training data has no experience on what it will be tested on. For instance, in the MMI database, action units 12 (lip corner puller) and 13 (sharp lip puller) are annotated as separate actions while they have a negligible difference in the appearance change they cause. Also, AU26 (jaw drop) is many times posed in the MMI database by moving the jaw to left or right side, while in the CK database this never occurs.

Another important factor is the difference in intensity of the posed expressions in the two datasets. This factor can be better observed on the AUs where the F1 measure is low

but the AUC measure is acceptably high, suggesting a successful generalization (e.g. AU1, AU2, AU9, AU27). The AUC compared to measures like F1 or true positive rate or false positive rate is more robust against this factor, since it is obtained by alternating the decision threshold of the classifier. Nonetheless, these initial results that we present do not assess a firm low or high generalizability quality on the system, but rather serve as a means of showing its applicability to different types of data. Further tests need to be performed with the system trained on a dataset with more variability to be able to assess the generalization, which will be the next direction we plan to improve the system.

5.5 Conclusion and Discussion

We have presented a novel framework for facial action unit detection in videos. The proposed system consists of extracting a combination of curvature Gabor features at different filter sizes, applying the LBP on top and computing the difference in histograms for neutral and peak frames. Then the obtained features are used in an AU specific feature selection and classification process to detect the present AUs. We achieve 98.6% accuracy, 96.3% F1 and 99.7% AUC scores in average for the leave-one-out test performed on the CK+ database, which is to our knowledge the highest reported to date.

We also report our results for a cross-database test, which are not as highly accurate, yet promising, especially for certain AUs. To assess the generalizability of the system, further tests should be performed with a training set containing a larger variability among expressions. However, the extremely high accuracy presented in this work already shows the representation and discriminative power of the proposed features, which we believe will constitute an important position in future facial action recognition and expression analysis research.

This is the first time that the curvature type Gabor wavelets have been used for facial action detection. They are particularly suitable for the task because of their ability to represent curved structures on the face, such as wrinkles and furrows, which are important indicators of facial muscle movements. The proposed system, however, has some drawbacks. This first one is the low generalizability, that is the very high accuracy obtained on test samples from the same dataset as the training does not apply to the cross-database performance. We do demonstrate, though, the promise for certain AUs especially for the AUC measure, which is a more meaningful metric with tunable systems. A comparably more suitable use of the system would be on subject-dependent systems, as in an HCI system that is custom-trained on a specific users facial actions.

Another drawback is the high computation time for the multiple Gabor wavelets proposed. The high number of the different wavelets lead to a heavy computation (in the order of a few seconds per frame, in the worst case), although they are capable of representing facial features in multiple scales, shapes etc. In the next chapter (Chapter 6) we propose a real-time AU detection system that can be more easily extended to other methods, such as our proposed multi-label AU detection framework.

Part III

Multi-Label Action Unit Detection

Overview

Following the contributions presented in Chapters 4 and 5, in this part of the thesis we present our final contribution on the AU detection problem with a completely different approach. Firstly, we present a real-time subject-independent AU detection framework that is easy to apply to large amounts of data and secondly we present a multi-label solution to the AU detection problem using a manifold embedding extension that incorporates the co-occurrence information between multiple AUs. This is quite a novel approach in the sense that it is the first time in the literature that the inter-relations of AUs is exploited in such a scheme.

This part of the dissertation presents a contribution with two levels. First, we present our real-time AU detection method that is based on extracting SIFT features around facial landmarks detected through an SDM based facetracker. The methods presented in Chapters 4 and 5 have shown state-of-the-art performance on the CK+ database, yet they do not run in real-time and / or require a heavy parameter tuning stage within training. This makes it difficult to study and test extensions and use them in real-world applications or on a large amount of test data. To this end, we propose our real-time AU detection system using SIFT and validate it on the CK+ database, similarly to the previous contributions.

The system that is presented is then extended to a multi-label embedding scheme for the AU detection problem and we also present in this chapter our participation in the Facial Expression Recognition and Analysis (FERA) 2015 sub-challenge for spontaneous action unit occurrence detection. The problem of AU detection is a multi-label classification problem by its nature, which is a fact overseen by most existing work. The correlation information between AUs has the potential of increasing the detection accuracy. We investigate the multi-label AU detection problem by embedding the data on lower dimensional manifolds which prioritize multi-label correlation. For this, we apply the multi-label Discriminant Laplacian Embedding (DLE) method as an extension to our base system. The extended system also uses SIFT features around a set of facial landmarks that is enhanced with the use of additional non-salient points around transient facial features. Both the base system and the DLE extension show better performance than the challenge baseline results for the two databases in the challenge. The proposed extension achieves close to 50% as F1-measure on the testing partition in average, a score that is 9.9% higher than the baseline in the best case and more accurate than other participants, winning the participated sub-challenge.

Our contributions in this part form an easily applicable AU detection system that is validated through multiple datasets and an internationally recognized challenge on the

subject. The work presented here include parts of the publications Gao *et al.* (2014), Ringeval *et al.* (2014) and Yüce *et al.* (2015). In the next and final part of the thesis we present the system being used for driver monitoring, showing an example function in a real-world application.

Multi-Label Action Unit Detection

6

6.1 Introduction

We have reviewed in the previous chapters why automatic AU detection is a useful and important tool and what the main open issues in the field are. Real-world AU detection systems are to be utilized during natural interactions and thus require robustness against low intensities and short durations of facial muscle contractions during spontaneous behavior, uncontrolled scene configurations and subjective appearance variations, primarily among many other factors. In this chapter we address these issues by proposing an AU detection system that is validated on first the CK+ database, as in the previously proposed systems, and secondly through the Facial Expression Recognition and Analysis Challenge 2015 (FERA2015) sub-challenge on AU occurrence detection. The FERA2015 challenge (Valstar *et al.*, 2015) is important since it tries to stimulate research on AU detection in a way that addresses the difficulties listed above within the datasets used. Both SEMAINE (McKeown *et al.*, 2012) and BP4D Zhang *et al.* (2014) databases include spontaneous facial expressions that contain AUs of various intensities and durations that were recorded in a mostly unconstrained manner.

Since the first FERA challenge (FERA 2011 (Valstar *et al.*, 2011),(Valstar *et al.*, 2012)) many advances have been proposed. Variants of Local Binary Patterns (LBP) are still popular for static and dynamic 2D or 3D action unit detection because of their efficiency (e.g. Almaev *et al.* (2013),Bayramoglu *et al.* (2013)). SIFT (Scale Invariant Feature Transform) descriptors have also been used efficiently within various frameworks (Ding *et al.* (2013), Zhu *et al.* (2011)). In this chapter we also propose a base system that uses SIFT features, validate its efficacy on the CK+ (Lucey *et al.*, 2010) and then propose an extension to this system that uses SIFT on an enhanced set of facial landmarks that includes non-salient points around transient facial features and a multi-label manifold embedding in the training phase to integrate co-occurrence relations between different AUs.

The AUs generally occur in combinations during natural behavior and these combinations form a correlation pattern between them. However, there is not a lot of work in the literature on the use of multi-label information for AU detection rather than treating

them as independent labels. This information may prove valuable since AU recognition is actually a multi-label problem, i.e. a data point belongs to multiple labels. A well-known work that uses this information in a temporal manner is the one by Tong *et al.* (2007). Mahoor *et al.* (2009) have combined Laplacian Eigenmaps, which is a locality preserving method for embedding the data on a lower dimensional manifold, with spectral regression to learn separate sub-spaces for AUs to detect their intensity. There have also been other attempts to use manifold learning type of projections for facial expression recognition, e.g. Shan *et al.* (2006). In this work we also propose an extension to our system that uses Discriminant Laplacian Embedding (DLE) (Wang *et al.*, 2010a). DLE is a method that combines the concept of Laplacian Eigenmaps (Belkin & Niyogi, 2003) and a multi-label adaptable variant of Linear Discriminant Analysis (LDA), which constitutes the discriminative part of the system. It has been applied successfully for classification of multi-label data (Wang *et al.*, 2010a), and this is the first time it is applied on the AU detection problem.

The challenge data is composed of three partitions: training, development and test. The test set is the one that is used for ranking and it is not available to the participants. Therefore, in contrast to the development set, the results used for ranking are obtained in a *blind fashion* by sending a software to the organizers without possibility of parameter tuning etc. We show that our proposed system performs successfully on both of the challenging datasets used (BP4D and SEMAINE) and outperforms the challenge baseline for both the development and test sets. In addition, we present the initial analysis on the effect of the proposed DLE extension on both the development and test partitions. We show improvement for certain AUs, while also observing that the performance is quite data dependent.

In the rest of this chapter, we first explain the proposed base system in Sec. 6.2 along with its performance on the CK+ database and then we present the proposed extension of multi-label DLE in Sec. 6.3. Sec. 6.4 demonstrates our results for the FERA 2015 challenge on the development and test partitions of the two challenge datasets in comparison with the baseline results as well as between the two proposed methods. In Sec. 6.5 we present our conclusions on the results and effectiveness of the proposed method and list potential improvement methods. Parts of the work included in this chapter have been adapted from the following publications: Yüce *et al.* (2015), Gao *et al.* (2014) and Ringeval *et al.* (2014).

6.2 Proposed AU Detection System Overview

In this work, we have used an AU detection system that uses SIFT features as the main difference with the systems presented in Chapters 4 and 5. Another difference is the use of an SDM based facetracker instead of CLM, which operates faster and is more robust against large head-pose variations. In this section we provide a brief review of this system, which is then used as a *base system* for the multi-label extension, as explained in Section 6.3.

Common to any automatic facial analysis system, the initial step for AU detection is to locate the face region and facial landmarks in the images, for which we employ a face

tracker based on the SDM (Xiong & De la Torre, 2013). The SDM starts with an initial guess and estimates the shape using a cascade of regression models that are learned at each step using local texture features (SIFT features in this case) extracted from the landmarks estimated in the previous step, as explained in more details in Section 3.1.3. The initialization for the tracker is performed using the commonly used Viola-Jones facetracker (Viola & Jones (2004) and see Section 3.1). The tracker in the end provides 49 landmarks on the face for each frame in the tracked video sequence. Subsequent to the facial landmarks tracking each tracking is aligned to reduce the effects of appearance variations due to the head pose. For this purpose, we have used an affine warping scheme using the detected eye locations. This technique was used in our applications on the CK+ dataset and the FERA2015 challenge (including the multi-label extension), since the data contains small head-pose variations in terms of yaw and pitch, thus affine warping provides a registration that is sufficiently good.

After aligning the face and scaling it to a fixed size of 200 by 200 pixels we extract local appearance features around the 49 landmarks using SIFT (Lowe (2004) and Sec. 3.2.2). The SIFT descriptors are extracted in the 32 by 32 local neighborhood around each landmark, resulting in a feature vector of size $128 \times 49 = 6272$. We reduce this dimension using PCA, that retains 98% of the total variance contained in the features and the extracted feature vectors are used in an L1-regularized linear-SVM classification scheme.

Table 6.1 shows the leave-one-subject-out AU detection accuracy on the CK+ database (Lucey *et al.*, 2010). Note once again that, although the accuracies are lower those presented in Chapters 4 and 5, since the feature extraction scheme is much faster (compared to application of multiple Gabor or reconstruction filters) the method is much faster to train and apply. Therefore, it allows for more easily investigating extensions and improvements. In addition, since it does not involve a rigorous feature selection process it is better generalizable to different datasets recorded in independent conditions. The complete AU detection framework operates at $> 15fps$, while the speed of previous methods presented are in the order of a few seconds per frame. Another application of the *base* AU detection system with SIFT features can be found in Ringeval *et al.* (2014) within the framework of multimodal valence and arousal recognition in natural videos. The work has not been included in this thesis for reasons of consistency.

For the FERA challenge, the training is performed on a custom training set that is the combination of the neutral and peak frames of each sequence in the CK+ database (Lucey *et al.*, 2010), non-speech frames of the training partition of the GEMEP-FERA database (Valstar *et al.*, 2011) and examples from the SEMAINE (McKeown *et al.*, 2012) and BP4D (Zhang *et al.*, 2014) training partitions down-sampled such that there is a certain minimum number of examples of each AU that is present in each sequence. The resulting combination is a set of 6713 examples and each AU retains a positive/negative sample ratio of at least 10%. The C parameters of the linear-SVMs are learned through a 5-fold cross-validation within this training set. Differently from what has been explained above, the final vector dimension after the PCA was chosen separately for each AU, and this dimension is learned through the development set accuracies. For the final system submission we also learn a threshold for the distance to the separating hyper-plane in a way that maximizes the F1-score on the development set. This threshold can be an

Table 6.1: AU Detection Accuracy on the Extended CK database, OA: Overall accuracy, F1: F1-Score AUC: Area under ROC curve

AU	OA	F1	AUC
1 (Inner Brow Raiser)	0.919	0.860	0.948
2 (Outer Brow Raiser)	0.941	0.829	0.939
4 (Brow Lowerer)	0.851	0.767	0.908
5 (Upper Lid Raiser)	0.939	0.810	0.966
6 (Cheek Raiser)	0.894	0.712	0.928
7 (Lid Tightener)	0.847	0.573	0.884
9 (Nose Wrinkler)	0.981	0.923	0.989
12 (Lip Corner Puller)	0.954	0.894	0.974
15 (Lip Corner Depressor)	0.932	0.770	0.950
17 (Chin Raiser)	0.934	0.904	0.977
23 (Lip Tightener)	0.939	0.632	0.912
25 (Lips Part)	0.951	0.955	0.988
27 (Mouth Stretch)	0.971	0.894	0.985

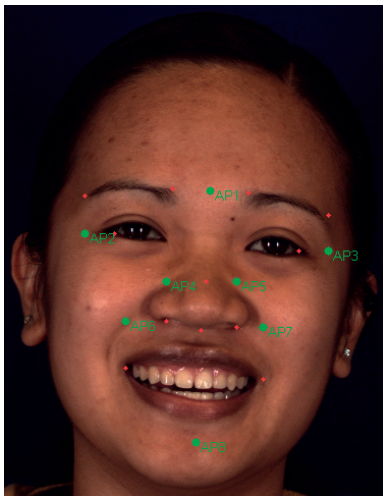
effective biasing parameter between the precision and recall and may depend on factors such as database recording conditions or subjective appearance differences. However, the results reported on the development set still use the default 0 as the decision threshold in order to have a fair comparison with the baseline and between the proposed systems.

6.2.1 Extending the Facial Mask

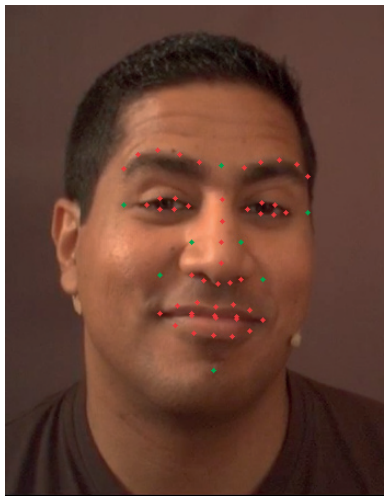
Using the locations of the 49 points provided by the SDM facetracker we estimate the position of eight additional points that contain important local appearance information related to certain AUs. The positions of these points are calculated after an affine warping performed to correct for the in-plane rotation. They mostly mark the non-salient landmarks on the face, which generally appear as result of a muscle contraction, compared to the main set of 49, which locate the salient facial features. The locations of these additional points (AP) are illustrated on an example image from the BP4D database in Fig. 6.1(a) as well as the landmarks that are used in their calculation.

$AP1$ is located at the center of the inner most points of the two eyebrows and locates a critical region mostly for $AU4$ (brow lowerer) but also for $AU1$ (inner brow raise). $AP2$ and $AP3$ are located around the crow-feet wrinkles close to the two eyes and their positions are calculated using the corners of the eye-brows and the center of the eye. These wrinkles are important indicators of $AU6$ and are potentially correlated to $AU7$ (eye-lid tightener). $AP4$ and $AP5$ are located on each side of the nose and mainly added to include appearance information that occurs during AUs such as $AU10$ (lip raiser) or $AU9$ (nose wrinkler, not included in this challenge). The positions of $AP6$ and $AP7$ are calculated as the x position of the corresponding corner of the mouth and the y location of the nose tip. These points mark the nasolabial furrows, whose appearance change with action units like $AU6$ (cheek raiser) and $AU10$. Finally, $AP8$ is the point located on the chin that is obtained such that it is equidistant to the nose tip with $AP1$, assuming a

vertical symmetry on the face. This point is mainly important for $AU17$ (chin raiser) but also contains information for other AUs that cause a shape and appearance change on the chin and the lower lip. These eight points provide more coverage on the face and thus additional appearance information. Note that the locations of these points are invariant to pose change since they are calculated relatively to the 49 points obtained from the SDM face tracker after correction for the in-plane rotation. Fig. 6.1(b) shows the complete mask consisting of 57 points and resulting in a SIFT feature vector of dimension 7296.



(a) 8 Additional facial landmarks (green) and the landmarks (red) used in their computation on a sample image from the BP4D database.



(b) The full 57 points mask on a sample image from the SEMAINE database.

Figure 6.1: Facial landmarks obtained from the face tracker and 8 additional points.

Note that, this facial mask extension is only part of the FERA 2015 application and was not used for the the CK+ database. In the following section we explain the multi-label extension and its application on the FERA2015 challenge.

6.3 Discriminant Multi-Label Manifold Embedding for AU Detection

The problem of AU detection has rarely been treated as a multi-label problem. In this work we apply the multi-label DLE method proposed in Wang *et al.* (2010a) to investigate the advantage of using the mutual information between AUs instead of treating them independently. DLE makes use of the similarities of samples in the training data in terms of both the features and labels, allowing the integration of the correlation between multiple labels.

The method can be summarized as the combination of LDA and Laplacian Embedding (Belkin & Niyogi, 2003) in a multi-label setting in order to utilize the locality information of the data in a supervised manner. Given a data matrix X of n samples and the corresponding label matrix Y of size $n \times P$ (P being the number of different labels,

meaning Y contains a binary-valued vector for each sample indicating the existence of each AU in our case) the embedding is performed by solving the eigenvalue problem:

$$(A^{-\frac{1}{2}} {S_w}^{\frac{1}{2}} S_b {S_w}^{-\frac{1}{2}} A^{-\frac{1}{2}})U = \mathbf{\Lambda} U \tag{6.1}$$

$\mathbf{\Lambda}$ being the set of eigenvalues and U the combination of eigenvectors, that will be used to project the data matrix X. S_w and S_b are the within-class and between-class scatter matrices, defined similarly to those in standard LDA (explained previously in Sec 3.3.2 and revisited later in this section) and $A = XLX^T$, with L being the graph Laplacian (Belkin & Niyogi, 2003). $L = D - W$, D being the diagonal of W and W defined as the "Label Correlation Enhanced Pairwise Similarity" in the work that we have adopted (Wang *et al.*, 2010a) and is formulated as the weighted sum of the *feature similarity matrix*, W_x and the *label similarity matrix*, W_L:

$$W = W_x + \mu W_L \tag{6.2}$$

W_x is the $n \times n$ pairwise similarity matrix, similar to most embedding algorithms, and is calculated through the Gaussian kernel similarity function (aka. heat kernel). The bandwidth of the kernel function σ is fixed as the average of all absolute pairwise differences in the training set.

W_L, on the other hand, is calculated using the pairwise similarities between the label vectors $\mathbf{y}$ of each sample and for two samples i and j is formulated as:

$$W_L(i, j) = \frac{{\mathbf{y}_i}^T C \mathbf{y}_j}{||\mathbf{y}_i|| ||\mathbf{y}_j||} \tag{6.3}$$

C is the $P \times P$ label-correlation matrix calculated from the training data. Embedding C in W_L allows weighting the pairwise label similarities by how correlated two labels are and thus results in placing two samples that have co-existence in highly correlated labels close to each other in the final embedding space. To give an example, if two samples are both labeled as $AU1 = 1$ and $AU2 = 1$ these samples will be close to each other in the final space because of the high correlation between $AU1$ and $AU2$ (c.f. Fig. 6.2), whereas they would have been placed further if we had not incorporated this correlation of labels. Fig. 6.2 shows the correlation between every AU, where the high correlation between certain AUs can be marked with a lighter color. Samples from the BP4D database were excluded in the computation of $AU25$ and $AU45$ correlations, since these were not annotated. In addition to the labels defining whether each of the 14 AUs under question within this challenge exist or not, we add a 15th binary label to include the cases where none of these AUs exist. This additional label, of course is not correlated to any of the 14 AUs (Fig. 6.2).

Finally, μ in Eq. 6.2 is the balance parameter between the pairwise feature and label similarity matrices and was optimized on the development partition of the challenge data separately for each AU.

The second component of the DLE is the multi-label LDA. The standard LDA aims to project the data on a lower dimensional space in which the distance between samples with different labels are maximized and samples with the same labels are densely placed

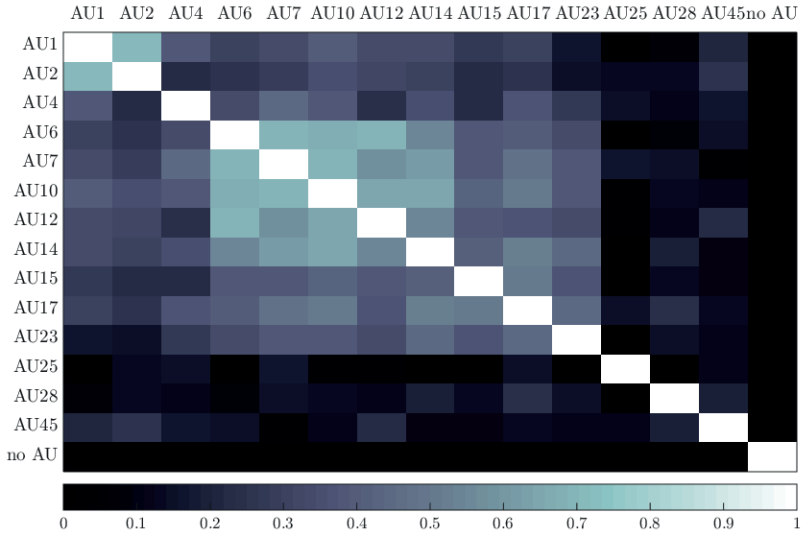


Figure 6.2: Action Unit Correlation Matrix on the Training Set

close to each other. This is performed by maximizing the between-cluster scatter and minimizing the within-class scatter (see Sec. 3.3.2). In the multi-label case the corresponding matrices are defined as the sum of the single-label scatter matrices for each type of label p (Wang *et al.* (2010a)):

$$S_b = \sum_{p=1}^{P} S_b^p, S_b^p = (\sum_{i=1}^{n} Y_{ip})(\mathbf{m}_p - \mathbf{m})(\mathbf{m}_p - \mathbf{m})^T \tag{6.4}$$

$$S_w = \sum_{p=1}^{P} S_w^p, S_w^p = \sum_{i=1}^{n} Y_{ip}(\mathbf{x}_i - \mathbf{m}_p)(\mathbf{x}_i - \mathbf{m}_p)^T \tag{6.5}$$

where $\mathbf{m}_p$ is the mean of all samples belonging to the label p and $\mathbf{m}$ is the multi-label global mean:

$$\mathbf{m} = \frac{\sum_{p=1}^{P}\sum_{i=1}^{n} Y_{ip}\mathbf{x}_i}{\sum_{p=1}^{P}\sum_{i=1}^{n} Y_{ip}} \tag{6.6}$$

These two kinds of projections allow us both to learn a correlation enhanced lower-dimensional manifold and a multi-label discrimination of the data. Projecting the data on the embedding space defined by U, which is learned by solving Eq. (6.1), we obtain a lower dimensional feature vector as input to the classifier. These features are more discriminative while still containing the correlations between the AUs in addition to the locality properties of the original data. In our tests, the data matrix X of size n by d is obtained by projecting the full data matrix on the PCA space. This initial dimension reduction step enables removing most of the redundancy in the data and thus, allows for a more efficient embedding. The final number of eigenvectors to be used is chosen by optimizing the F1-score, which is a balanced compromise between precision and recall, on the development partitions for both datasets and separately for each AU.

6.4 Results on the FERA 2015 Challenge

The FERA 2015 challenge consists of three sub-challenges: AU occurrence detection, AU intensity recognition with prior occurrence knowledge and full AU occurrence and intensity detection. The challenges are on two spontaneous facial expression databases with AU annotations: SEMAINE database (McKeown *et al.*, 2012) and BP4D database (Zhang *et al.*, 2014). The participants are provided with a training and development set for each database and are asked to run their programs on independent test sets, which they do not see and they have no prior knowledge about, except that they are part of the same database. The test partitions of both datasets contain subjects from the development and training sets as well as unseen subjects. For more details on the databases and partitions the reader is referred to the challenge baseline paper (Valstar *et al.*, 2015).

We have participated in the AU occurrence detection sub-challenge and in this section report our results on the development and test partitions. The fact that these databases are recorded spontaneously makes the task more challenging since the recorded subjects act completely naturally without any instructions about their facial behaviour. This causes more variant AU appearance and occurrences in terms of intensity and combination compared to databases with constraints on the action units or facial expressions performed, e.g. CK+ (Lucey *et al.*, 2010), GEMEP-FERA (Valstar *et al.*, 2011) or MMI (Valstar & Pantic, 2010)).

6.4.1 Results on the Development Set

We first test our proposed system and the Discriminant Laplacian Embedding extension on the development set. No image from the development sets of the two datasets were included in the training of the classifiers or learning the PCA and embedding bases. However, the development set allows us to obtain the number of features that are optimal for each action unit and also to learn the threshold for the SVM decision value to be used on the test partition. This threshold was kept at 0 for the results on the development set in order to obtain a fair comparison with the baseline results and also between the two proposed methods. The DLE was applied to the PCA projected matrix of dimension $d = 1000$.

Table 6.2 shows the average Overall Accuracy (OA) and F1 measures obtained using the two systems (PCA-SIFT and DLE-SIFT) in comparison with the baseline results. Tables 6.3 and 6.4 present our results for each AU on the BP4D and SEMAINE development partitions, respectively in terms of OA, Area Under ROC Curve (AUC) and F1. As can be seen from Table 6.2 both systems achieve significantly better performance than the baseline systems (with geometric and LGBP-TOP features (Almaev *et al.*, 2013)) on the development set. For a detailed per-AU comparison the reader is referred to the baseline paper (Valstar *et al.*, 2015). The increased accuracy compared to the baseline shows the efficacy of the chosen features and also the advantage of the enhanced set of facial landmarks. The enhanced set indeed results in an average F1-score increment of 0.7% and 2.6% on the BP4D and SEMAINE development partitions respectively, compared to the 49 point standard set, tested under the same conditions.

Table 6.2: Comparison with Average Baseline Results on the Development Partitions

	SEMAINE		BP4D	
	OA	F1	OA	F1
Baseline Geometric	0.735	0.351	0.712	0.580
Baseline Appearance	0.680	0.298	0.639	0.539
PCA-SIFT (Prop. 1)	0.793	0.417	**0.735**	0.589
DLE-SIFT (Prop. 2)	**0.802**	**0.435**	**0.735**	**0.591**

Table 6.3: Results on the BP4D Development Partition

	PCA-SIFT			DLE-SIFT		
AU	OA	AUC	F1	OA	AUC	F1
1 (Inner Brow Raiser)	0.717	0.674	0.395	0.694	0.695	**0.41**
2 (Outer Brow Raiser)	0.669	0.563	**0.284**	0.664	0.563	0.262
4 (Brow Lowerer)	0.791	0.786	0.472	0.805	0.78	**0.509**
6 (Cheek Raiser)	0.809	0.888	**0.802**	0.801	0.873	0.783
7 (Lid Tightener)	0.698	0.765	**0.761**	0.691	0.756	0.746
10 (Lip raiser)	0.734	0.795	0.781	0.729	0.789	**0.789**
12 (Lip Corner Puller)	0.859	0.933	**0.877**	0.839	0.914	0.857
14 (Dimpler)	0.606	0.699	0.611	0.598	0.706	**0.616**
15 (Lip Corner Depressor)	0.732	0.779	**0.447**	0.728	0.769	0.43
17 (Chin Raiser)	0.642	0.724	0.573	0.70	0.761	**0.604**
23 (Lip Tightener)	0.831	0.782	0.486	0.838	0.783	**0.487**
Average	0.735	0.762	0.589	0.735	0.763	**0.591**

Table 6.4: Results on the SEMAINE Development Partition

	PCA-SIFT			DLE-SIFT		
AU	OA(%)	AUC(%)	F1(%)	OA(%)	AUC(%)	F1(%)
2 (Outer Brow Raiser)	0.804	0.753	**0.308**	0.822	0.732	0.306
12 (Lip Corner Puller)	0.671	0.677	0.480	0.682	0.731	**0.512**
17 (Chin Raiser)	0.957	0.889	**0.394**	0.957	0.886	0.303
25 (Lips Part)	0.757	0.74	0.482	0.725	0.74	**0.494**
28 (Lip pucker)	0.978	0.906	0.509	0.982	0.947	**0.672**
45 (Blink)	0.591	0.683	**0.329**	0.649	0.668	0.324
Average	0.793	0.775	0.417	0.803	0.784	**0.435**

The advantage of using a DLE with multi-label information over standard PCA, on the other hand, is not that clear. Although for some AUs the method is more efficient, in average the improvement remains marginal. The difference is clearer when tested on the SEMAINE database, which suggests that the success of the method may depend on the data distribution or the similarity of the distribution of data between the training and test sets. On the SEMAINE database the clearest improvement is on $AU28$, while on BP4D

it is on $AU4$ and 17 . The same improvement not appearing on the two databases further suggests that the data distribution is an important factor. More training data will probably provide better variability and thus better generalization of the success of embedding.

6.4.2 Challenge Results for AU Occurrence Detection on the Unseen Test Set

This section presents the results we have obtained on the test partitions, which constitute the main challenge. Tables 6.5 and 6.6 show the F1-scores obtained on the BP4D and SEMAINE test partitions respectively using the two proposed systems and in comparison with the challenge baseline results obtained with the geometric and appearance features. The first observation is that both of the proposed systems clearly outperform the challenge baseline on the test set except for some AUs, namely AUs 2, 12, 14 and 45, for which geometric features are apparently more effective. As weighted average on the two databases, the best F1 score (**0.499**) is obtained with the PCA-SIFT system and is 0.099 higher than the challenge baseline with appearance features (improved by 24.8%) and 0.054 higher than that with geometric features (improved by 12.3%). The proposed system also outperforms all other participating systems: Gudi *et al.* (2015) who proposed a deep learning framework (reported an overall F1-score of 0.458) and Baltrušaitis *et al.* (2015) who have proposed improvements through various person-specific normalization methods (reported an overall F1-score of 0.48).

Table 6.5: F1-Scores on the BP4D Test Partition

	Our Results		Baseline	
AU	PCA-SIFT	DLE-SIFT	Geo.	App.
1	**0.261**	0.226	0.188	0.180
2	0.167	0.149	**0.185**	0.159
4	**0.283**	0.233	0.197	0.225
6	**0.729**	0.697	0.645	0.671
7	0.785	**0.802**	0.799	0.751
10	**0.802**	0.742	0.801	0.799
12	0.779	0.784	**0.801**	0.792
14	0.625	0.599	**0.72**	0.666
15	**0.348**	0.223	0.238	0.139
17	**0.380**	0.325	0.311	0.245
23	**0.441**	0.424	0.320	0.239
Average	**0.508**	0.473	0.473	0.442

For BP4D, the accuracies obtained are much lower compared to the development set. This is expected as the few parameters (number of features, decision thresholds and μ in 6.2) that we have were optimized on the development set. However, on the SEMAINE test partition we obtain better results than the development set, which is possibly an indicator that the SEMAINE test and development partitions are more similar to each other compared to BP4D and that our classifiers are able to generalize well enough to

Table 6.6: F1-Scores on the SEMAINE Test Partition

	Our Results		Baseline	
AU	PCA-SIFT	DLE-SIFT	Geo.	App.
2	0.655	0.663	0.569	**0.755**
12	**0.769**	0.759	0.595	0.517
17	0.215	**0.255**	0.091	0.066
25	**0.623**	0.613	0.445	0.400
28	0.251	**0.262**	0.250	0.009
45	0.325	0.347	**0.396**	0.209
Average	0.481	**0.483**	0.391	0.326

this unseen dataset. As explained in the definitions of the sets (Valstar *et al.*, 2015), the test set of BP4D was indeed recorded at a different time, possibly under different conditions.

Our results show that the DLE system achieves a marginal increase in the mean accuracy on the SEMAINE database, compared to the system with standard PCA. Better results are obtained for AUs 2, 17, 28 and 45. However, this is not the case for the BP4D, with higher F1 measure only for AUs 7 and 12. This probably implies once again that the DLE is highly dependent on the data distribution and that the BP4D development and test partitions contain more variation of AU combinations than that is contained in the training set compared to the SEMAINE database. More tests with more variability and a higher number of training data is needed to reach a conclusion on the benefits of the method, which will be performed next as an extension to this work. Better tuning of the parameters may also greatly increase the accuracy. Another possible cause of the problem is the very low rank of the between-class scatter matrix. Our further work will include using the Laplacian of a dissimilarity matrix instead of the LDA terms in the formulation.

6.5 Conclusion

In this chapter, first we have presented an AU detection system that uses SIFT features extracted around facial landmarks detected through an SDM facetracker and fed into SVM classifiers to detect the existence of AUs on the frame level. We presented the detection results obtained on the CK+ database for validating this base system.

Then, we have presented an extension of this system that uses an enhanced facial landmarks configuration that includes points around transient facial features and a Multi-label Discriminant Laplacian Embedding with integrated correlation between AUs. The system is applied to the FERA 2015 sub-challenge for spontaneous AU occurrence detection on the SEMAINE and BP4D databases. We obtain a significant increase of accuracy compared to the challenge baseline with both the proposed systems and in the best case obtain 49.85% F1 score as average on the test sets of the two databases, outperforming other participants and obtaining the best results on the challenge.

Comparing the efficiency of the proposed extension of DLE, we observe better re-

sults on certain AUs, but only a small increase in the average accuracy is obtained. Using multi-label information in AU detection is a difficult task mainly due to the large number of labels and huge variability in terms of their co-occurrence. In this work, taking into account the different performances on the two test partitions we reach the conclusion that the success of the multi-label DLE might also be data dependent and needs further analysis. The high accuracy of the method on some AUs suggests that the DLE is worthy of more investigation. Our future work will include integrating a dissimilarity matrix instead of the LDA related matrices, which are generally of very low rank. Another extension for further improvement would be to include the temporal adjacency of data points in the embedding or classification scheme.

Part IV

Automatic detection of driver's cognitive distraction

Overview

Visual driver monitoring systems keep gaining attention in the computer vision community, especially with the wide possibility of applications made available by the introduction of semi-automatic and automatic vehicles. The Advanced Driver Assistance Systems (ADAS), integrated within several components of the cars aim at providing increased security and comfort to the drivers while being as least distractive and disturbing as possible. The assistance systems that involve automatic visual detection of driver states (e.g. fatigue, emotional disturbance or distraction) therefore need a high true positive and low false positive detection rate. In addition, this precision needs to be obtained in real-time and in a way that is robust against the varying conditions that occur during driving (e.g. changes in illumination or driver's position).

Driver's distraction is a very important factor that is among the most influential ones to cause human error while driving, thus leading to traffic accidents, injuries and casualties. Efficient detection of distracted states of the driver will allow in-vehicle systems to launch an alarm system that can vary in intensity from a visual message on the console to a beeping sound and even automatic stopping of the car. The methods proposed in this part to detect cognitive distraction may also be applied to other applications, such as detecting the engagement level during learning applications or for monitoring the treatment of certain psychopathologies that cause difficulty of concentration and engagement. Examples of similar applications have been reviewed in Section 3.5 in the introductory part of the dissertation.

In this part of the thesis, we present two main contributions in Chapters 7 and 8, which are contributions that are complementary to each other. First, in Chapter 7 we introduce the EPV-DIST database, which is a multi-camera visual database of 46 people driving a simulator with different distraction conditions induced. The recording setup for our database has been planned to represent a configuration that is feasible to place in a car and work robustly in different light conditions. The experiment consists of three baseline driving conditions and two separate distraction conditions, namely visual and cognitive distraction induced through secondary tasks performed at the same time while driving. The software that we use allows recording the wheel and pedal information continuously, which is then used to assess the driving performance at each segment of the driving experience. Chapter 7 starts by reviewing existing definitions of driver distraction and works that have studied its effects on driving performance and car accidents. Then we explain the acquisition system and distraction induction methods to obtain the EPV-DIST database. Finally, we present statistical analyses of the subjects' driving performances so as to show the effectiveness of the distraction induction.

Secondly, in Chapter 8 we propose a method based on Facial AUs to detect the conditions where the drivers were induced cognitive distraction in the EPV-DIST database. AUs model every unitary muscle movement on the face and are the most objective means of defining and quantifying what is on the face. The system avoids using other categorizations of facial expressions (as reviewed in Section 2.1), since cognitive distraction is a complex mental state that does not have an existing correspondence to a pre-defined facial expression. The system proposed first generates a virtual frontal view from the three frames captured by the multi-camera system, then applies the AU detection we have previously proposed in Chapter 6. Then, we extract features from the dynamic continuous value outputs of the AU detection system, independently for each of the 14 AUs detected, and also from their cross-correlations at different delay points. This second type of features allows exploiting the dynamic inter-relations of AUs and their synchronization behaviour in different conditions, with the hypothesis that it will improve detection of facial expressions that are indicators of this complex state. All features are then fedinto SVM or Random Forests (RF) classifiers to obtain a decision on each sequence that has been labeled as cognitive distraction or not. Our findings show that the corresponding facial behaviour are very subject-dependent and although we achieve an acceptable accuracy in the subject independent tests, it is evident that a subject-based training will help obtaining higher precision. Chapter 8 first reviews existing systems of visual driver monitoring systems, then explains our proposed cognitive distraction detection system, presents the results obtained on the EPV-DIST database, and discusses the subjectivity problem as well as proposals of possible uses of the proposed system.

This final part of the thesis constitutes mainly an application of the previous contributions, with added novelty in terms of the database introduced and the detection system that combines several analysis methods. The majority of the work presented in Chapters 7 and 8 has also been submitted as a journal article and is under review at the time of writing (Yüce *et al.*, Under Review).

A Database for Spontaneous Facial Expressions of Distraction During Driving

7

7.1 Importance of Distraction during Driving

Driver monitoring in real-time is an emerging topic thanks to the availability of faster software and smaller and more powerful hardware that can easily be integrated in consumer vehicles. In addition to systems that record and analyze driving data, e.g. wheel movement, speed and acceleration, or driver's physiological signals, research on visual monitoring systems are also on the rise and such systems will soon be more and more frequently integrated in automobiles on the market. In this work we introduce a visual database of drivers in a simulator setting while being induced different types of distraction. The database has been recorded in realistic conditions with a system that can be integrated in a vehicle and provides temporal annotations for driving conditions and performance.

There has been a long discussion on how to define driver distraction. Pettitt *et al.* (2005), Lee *et al.* (2008) and more recently, Regan *et al.* (2011) have published works on how to define the term and compile the existing definitions. Lee *et al.* (2008) summarizes it as the diversion of attention away from activities critical for safe driving toward a competing activity. In Pettitt *et al.* (2005) there is a more extensive definition and driver distraction is stated as a delay by the driver in the recognition of information necessary to safely maintain the driving task, due to some event, activity, object or person, within or outside the vehicle that compels or tends to induce the driver's shifting attention away from fundamental driving tasks by compromising the driver's auditory, biomechanical, cognitive or visual faculties, or combinations thereof (Regan *et al.* (2011), Stutts *et al.* (2001)). Driver distraction can be in three types depending on its source and demand: visual, manual and cognitive (Strayer *et al.*, 2011). Even though we introduce a database that includes all three types of distraction, the automatic system presented in Chapter 8 focuses on detecting cognitive distraction, which can be defined as a diversion of the driver's attention from the driving task, not necessarily requiring any sharing of visual processing or involving or demanding a biomechanical action. It includes internally

induced distraction, such as mind wandering or daydreaming but excludes cases such as boredom, sleepiness, or driving under the influence of alcohol or drugs and substances that alter the mental state.

Many studies show that driver distraction is one of the most important causes of traffic accidents, along with alcohol use and speeding. A study conducted in France showed that 17% of 453 accidents that resulted in admittance to the emergency room was caused by a high mental distraction of the responsible driver (Galera *et al.*, 2012). More recent studies from the same group have shown that induced distractive thoughts led to less micro-regulation of both speed and lateral position and narrowed visual scanning of the driving scene (Lemercier *et al.*, 2014), mind wandering is the cause of 8% of close to 1000 accidents according to emergency room interviews with the drivers (Bakiri *et al.*, 2013) and that it affects 85.2% of the drivers especially in situations requiring less attention from the driver such as an everyday commute or a monotonous motorway (Berthié *et al.*, 2015).

In Neale *et al.* (2005), the authors have collected and analyzed almost 43000 hours of driving data and shown that 78% of the crashes and 65% percent of near-crash incidents involve driver inattention due to various secondary tasks. A similar study sponsored by the United States department of transportation showed that drivers investigated were engaged in non-driving related tasks in 71% of crashes (Olson *et al.*, 2009). Again in the U.S., it is estimated that around 20% of all police-reported road crashes involve driver distraction as a contributing factor (Victor *et al.*, 2013). In Young & Salmon (2012) the authors provide a large scale examination of the relationship between driver distraction and driver errors, along with a list of existing studies. Even though the numbers differ depending on the type and amount of data analyzed in each study, they all show that internally or externally caused driver distraction is a critical risk factor. Therefore, in this thesis (Chapter 8) we address the problem of automatic detection of cognitive driver distraction using visual monitoring of the driver's face and propose a system that is tested on simulation data and that can easily be integrated in real cars for applications like an early alert system or activation of countermeasures in order to help the driver regain his attention on the driving task.

In this chapter, we present the EPV-DIST database that we have recorded for purposes of training and testing our system for automatic detection of driver distraction, as well as providing the research community a database that is rich in the number of subjects and types of annotations related to driving. The database consists of 46 subjects driving a simulator while performing additional tasks in order to induce visual and cognitive distraction. In the following sections of the chapter, first we describe the recording setup that was used in the experiments in Sec. 7.2 as well as a summary of the demographics of the subjects recorded. Then, we explain the methods used to induce the distraction state aimed at in 7.3 with a statistical analysis of the effects on driving performance and conclude the chapter with a discussion on the usefulness of the database 7.4.

7.2 Description of Recording Configuration

One of the main contributions of this work is the introduction of a new video database with induced distraction during simulated driving. In this section we describe the details of the database, which we name EPV-DIST, which is short for the EPFL-PSA-Valeo Near Infrared (NIR) Multi-Camera Database of Visual and Cognitive Distraction during Driving. The aim of the database is to provide videos of natural behavior of drivers while performing additional visual and cognitive tasks. Another point worth mentioning is that the recording setup is built in a way that can be integrated directly in an actual consumer vehicle (in terms of camera positions), and provides robustness against real-life driving conditions, such as ambient light and head pose variations.

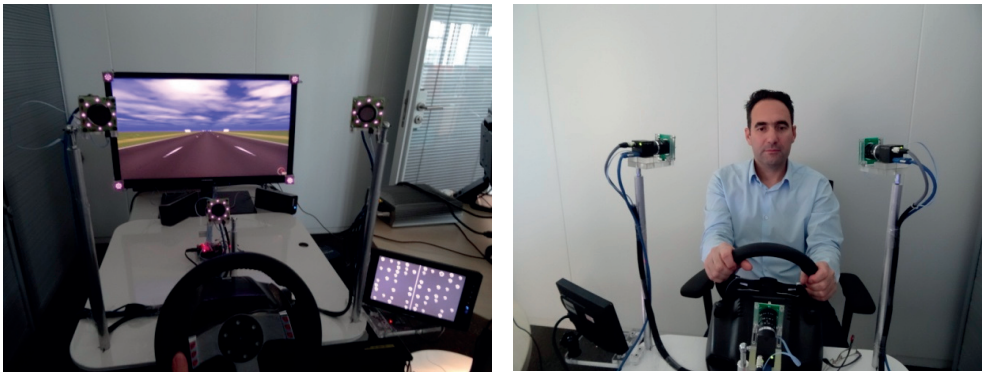
We have recorded 48 subjects, two of whom had to be excluded from the database due to technical problems during the recording. The subjects were recruited from students and research and administrative staff of EPFL and EPFL Innovation Park. As the mental tasks were prepared in French, they were asked to possess a sufficient level of understanding and speaking in French and also have a sight enough to drive without glasses, in order to avoid reflections of the NIR lighting. They did not need to possess a driving license as the simulator is very simple and since we have included a familiarization step prior to the recorded experiments. The subjects' ages are between 19 and 52 with an average of 30. The number of female and male subjects are equal and all subjects have given their consent for the use of their data in research on automatic visual behavior analysis. The total length of the recordings is approximately 25 minutes per subject, making a total of more than 19 hours of recording.

We plan to make the database publicly available solely for research purposes in the future, to help advance the research on facial behavior analysis during driving under various, predefined conditions. In the rest of this section we give details on the data acquisition setup.

7.2.1 Data Acquisition System

The EPV-DIST dataset consists of multi-view videos that are recorded using three NIR cameras and a special lighting equipment per camera with adequate filters, in order to filter out ambient light. Figure 7.1 shows the recording setup and the position of the recorded subject during the experiments.

Both the choice of the recording material and the placement of the cameras (see Fig. 7.1a) are based on a realistic application in a real consumer vehicle. Since the light conditions change very often during driving we have performed the recordings with cameras with a wide wavelength capture range (PointGrey Flea3) and adequate band-pass filtering. We have also built three integrated NIR-Light Emitting Diode (LED) (850 nm) circular lighting circuits that can be placed around the cameras and illuminated in synchronization with the frame-grab of the cameras using a micro-controller. These lighting systems make sure there is constant illumination around the face and the bandwidth filter, at the same wavelength as the LEDs, filters out a substantial amount of the ambient light. This allows for a continuous visibility of the driver's face with close to constant illumination and is suitable for a real application in a car in any light condition, e.g. when



(a) The Three-Camera Acquisition Setup (b) Driver's position during the recording

Figure 7.1: The recording setup of the experiment

there is too much sunlight or while passing through a tunnel. The camera-light pair has already been used in our previous experiment on posed expressions of stress (Gao *et al.*, 2014), where we have shown the feasibility of facial expression recognition with such a system in the single camera case. In addition to the circular LEDs we have placed similar lighting at the four corners of the simulator screen which are illuminated in turns with the lighting around the cameras, to create the glint and dark/bright pupil effect for use in future research for gaze analysis. The gaze related features have not been included in the automatic detection system presented in Chapter 8 in order to avoid possible bias and noise in the experimental results due to unoptimized gaze measurements.

As for the choice of the number and positions of the cameras, there exist two constraints for a realistic setup. The first one is the amount of head-pose coverage using a virtual frontal view generation from all cameras, explained later in Chapter 8. The second one is the feasibility of placement of the cameras in an actual car, without blocking the driver's sight and where there is already support to place the camera. We have chosen to use a three-camera system and placing the cameras as can be seen in Fig. 7.1. The first camera, the semi-frontal one (referred to as the frontal camera from this point on) is placed in the representative position inside the console behind the wheel. The left camera (with respect to the driver) is where would be the highest point of the left pillar in a car. Finally, the right camera is on the representative position of the bottom-left corner of the rear-view mirror.

The three cameras record frames synchronously at a rate of 20 fps as seen in Fig. 7.2a, 7.2b and 7.2c for the left, right and frontal cameras, respectively. We only use every second of these frames, those that correspond to the bright ones. These three images are then used to reconstruct a virtual frontal view of the driver's face as seen in Fig. 7.2d. The details of this reconstruction are given in Subsection 8.3.1. This three camera system allows invariability against head pose changes, which occur quite frequently while driving, and also against occlusions that occur in one or more views. All these properties of the setup provide a realistic sense into our database, as it would be a suitable setup to integrate in an actual vehicle.

(a) Left Camera (b) Frontal Camera

(c) Right Camera (d) Generated Virtual Frontal View

Figure 7.2: Images acquired by the three cameras and the corresponding generated virtual frontal view.

7.3 Experiment Protocol and Methods for Distraction Induction

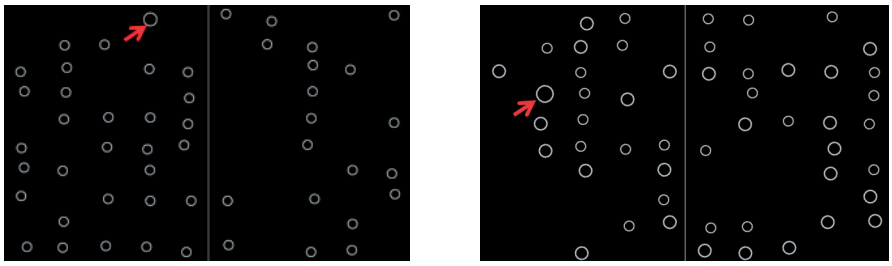
The driving task we used for our experiments is the Lane Change Test (LCT) (Mattes, 2003). LCT is a simple, easy to manipulate driving simulator that has become the standard simulator for testing secondary tasks while driving (ISO, 2010). It has been commonly used in the past for experiments involving such secondary tasks (Harbluk *et al.* (2007), Engström & Markkula (2007)). We have used a Logitech G27 wheel and pedals set for the control and the LCT allows continuous recording of the wheel and pedal motions, which are useful in providing a metric for the driving performance (explained in detail in Sec. 7.3.2). The LCT consists of a series of lane change tasks which are presented as road signs on the simulator screen (see Fig. 7.1a) and the drivers are asked to change their lane according to the sign presented, as soon as they see the sign and as quickly as possible before passing by the sign. We have fixed the maximum speed at 60 km/h and the distance between two signs at 150 meters, which results in Lane Change Sequence (LCS) of approximately 9 seconds, since the drivers were asked

to maintain the maximum speed. The road signs appears ~ 1 sec after the introduction of the distraction (if it applies), and disappear 40 meters later, giving the drivers around 2.4 seconds to perform the appropriate lane change.

7.3.1 Driving Conditions

All subjects were asked to perform the driving task in three conditions: The baseline, solely the driving task without any distractors; visual distraction, a visual secondary task which requires the driver to take the eyes off the road; and cognitive distraction, where the attention is directed to a non-driving related task without the need to take the eyes off the road. All factors other than the distractive agents were kept the same for the three conditions. Each subject has completed a total of five driving tasks, three baselines, one visual distraction and cognitive distraction. The order of the distraction related tasks have been randomized among subjects, such that there is an equal number of subjects who have performed the visual task before the cognitive one and vice versa. This randomization is to decrease the secondary effects of uncontrollable confounding factors, such as fatigue or disengagement.

The baseline condition is to obtain a ground measure for the driving performance and facial behavior without workload of the subject. It is performed three times in total, the beginning (after the familiarization part, which is not recorded), between two distractive conditions and the end. Each one consists of 18 lane changes, equally distributed for the 6 possible types of lane change between the left, right and middle lanes, in order to avoid effects of learning. In the end, we obtain 54 LCS, 9 for each lane change type, per subject.



(a) Easy condition (b) Difficult condition

Figure 7.3: Sample questions from the SURT task

The visual task used to induce visual distraction is the Surrogate Reference Task (SURT). The SURT is a standardized test (ISO, 2012), where the participants are shown two images with many circles and are asked to pick the side of the screen on which the biggest circle is present (Fig. 7.3). The images are shown on an additional screen at the right of the main simulator screen (see Fig. 7.1a), which is a representation of a screen position in the center console in an automobile. SURT is a visual-manual task with a high level of distraction as it requires a relatively long glance at the secondary screen in order to differentiate the biggest circle among many similar objects. The driver then gives his/her response by pressing the left or right pedal behind the wheel, which

constitutes the manual part of the task. The SURT has been applied at two difficulty levels, which differ by the number and sizes of the circles (Fig. 7.3a and 7.3b). All subjects performed the task with both difficulty levels, which were randomized in order, once again for the reasons listed above. The analysis of the visual distraction task videos have not been included in this thesis, since the focus is cognitive distraction, which is a *mental state* that is harder to detect by face - head related visual cues. The visual distraction component of the database will be used in future experiments and have been included in the database to provide the community videos of varying head poses during driving, with a measure of the driving performance.

The final driving condition is the induced cognitive distraction, which forms the main focus of this part of the thesis. This was performed using an auditory version of the Operation SPAN (OSPAN), developed by Turner & Engle (1989) and has been used by the National Highway Transports Safety Administration in US (NHTSA) as a standard task simulating driver cognitive distraction. The OSPAN task makes use of the working memory and attention of the driver and does not require the visual attention as the task is presented in audio and the response is either manual or by speech.

The OSPAN task is composed of two components, the first one is making simple calculations and the other memorizing words. At each LCS the driver is told a simple mathematical calculation statement, e.g. *Two times four plus one is ten*. The driver needs then to press the corresponding hand pedal behind the wheel if they think the statement is true or false. The choice of pedal for the right and wrong answers have been randomized to reduce the effects of the natural tendency to unintentionally think that one side represents the correct one. Right after the statement the subjects also hear a simple word in French, e.g. *maison (house)*, *rouge (red)* or *chemise (shirt)*, which they were asked to memorize and repeat at the end. The LCS that we analyze do not include the part where the drivers repeat the words they had to memorize. In the easy condition the participants hear two mathematical calculations only including addition and subtraction along with two words to memorize, while in the hard condition they hear three calculations that also include multiplication and three accompanying words. All participants receive an equal number of easy and hard tasks following each other and the order of which has been randomized equally among participants. The calculations and words have been recorded prior to the experiment and was repeated from a speaker in the experiment room, synchronized to appear at the same instant for every subject. The OSPAN task creates an additional load to the working memory of the subject and aims to pull the attention of the driver off the road and the driving task. Note that we have not aimed at any positive or negative valence effect of the cognitive distraction, for example as performed in Chan & Singhal (2013) by selecting words related to positive and negative emotions.

In order to put the driver in a multitasking condition each distraction task started a couple of seconds before the appearance of the lane change sign. The participants had not been given any instruction prior to or during the experiment regarding the priority of the driving vs. secondary task. Each participant, therefore, chooses such a priority depending on his/her own workload and sometimes in a varying manner for each task, as observed from their recorded data.

7.3.2 Measuring Driving Performance

The LCT simulation system allows recording the wheel and pedal motion at all times, which we then use in order to calculate a performance measure for each of the LCS. The measure that we calculate is the Mean Deviation from the normative lane change behavior, or Mdev. The Mdev is measured by calculating the area between the expected driving behavior for a specific lane change and the actual one (Fig 7.4). It is a standard way of quantifying the driving performance on a simulator over short distances (ISO, 2010). The area between the two trajectories is sensitive to perception (missing the sign), reaction time, quality of the maneuver and lane keeping (Mattes, 2003). Note that the Mdev could also have been used as an indirect measure of the level of cognitive distraction, and a ground-truth for the classification problem in this thesis (Chapter 8). However, due to the distribution of the Mdev values, this would have caused an imbalanced classification or regression task. Also, the problem we address in this work is to detect when a driver is imposed an additional mental load, or cognitive distraction and not the facial behavior during *unsuccessful* driving but before that happens. Therefore, we use the Mdev values only to show whether the induced distraction has an overall effect on the driving performance.

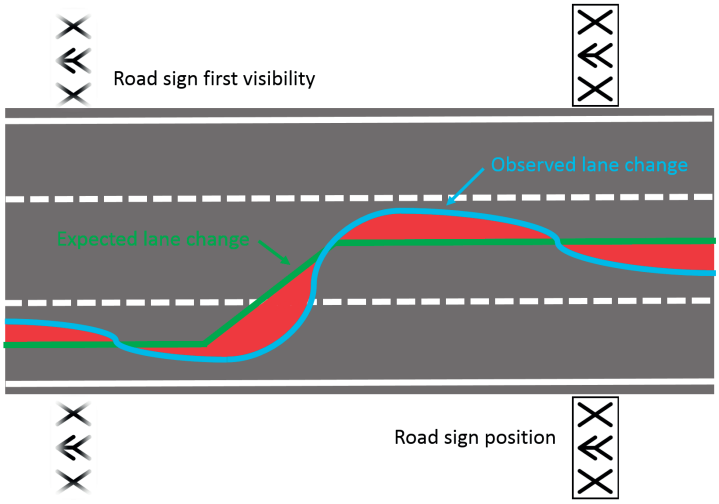
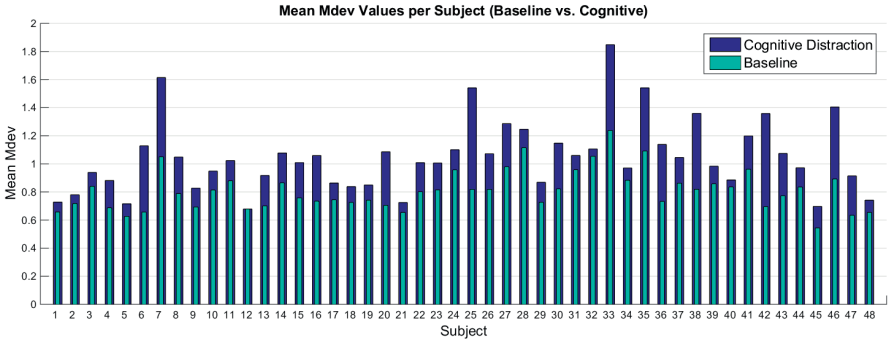
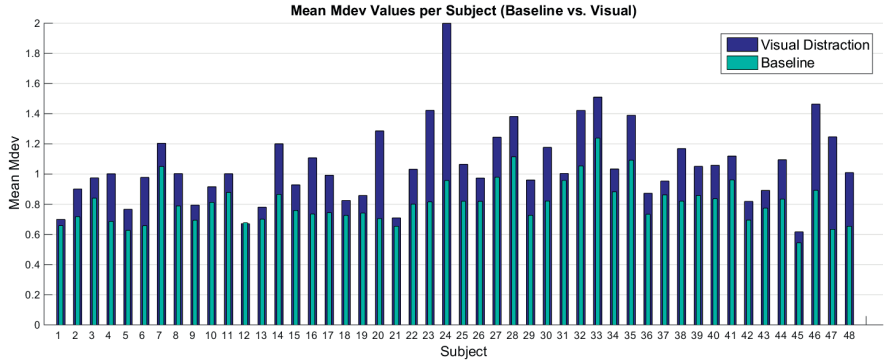


Figure 7.4: Mdev calculation as the area between the expected lane change and the observed driver behavior

In order to show the effectiveness of the visual and cognitive distraction induction that we used, we have performed a statistical analysis of the Mdev performance values, comparing the three tasks. We have calculated the Mdev value for each 8.5 seconds sequence (the first 0.5 seconds were removed to remove noise) and Fig. 7.6 shows the distribution of the Mdev values among all sequences for every subject in the three conditions. The initial observations are the difference between the Baseline (BL) versus the Cognitive Distraction (COG) and Visual Distraction (VIS) in variance (0.098 for BL vs. 0.295 for COG and 0.278 for VIS) and the shift in the mean value (0.809 for BL vs. 1.048 for COG and 1.053 for VIS) median (0.772 for BL vs. 0.944 for COG and 0.941



(a) Baseline vs. Cognitive Distraction



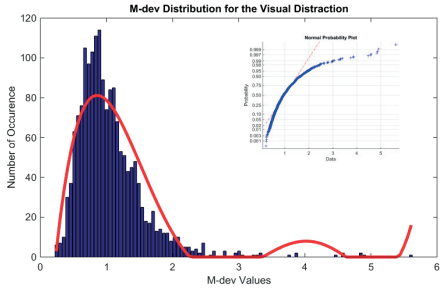
(b) Baseline vs. Visual Distraction

Figure 7.5: Mean Mdev values for each subject in three conditions: Baseline, Cognitive Distraction and Visual Distraction.

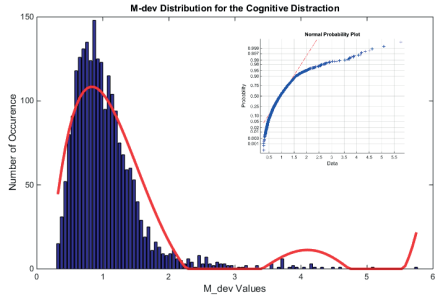
for VIS) and the maximum values (2.481 for BL vs. 5.769 for COG and 5.631 for VIS). In addition, we have performed a Wilcoxon signed-rank statistical test (WSR) test on the mean Mdev values of the subjects in the BL vs. COG and BL vs. VIS conditions. The WSR test is a non-parametric paired difference test used to compare two related samples (Wilcoxon, 1945). It is used to compare ordinal random variables that are non-Gaussian distributed, which fits perfectly our case (Fig. 7.6c, 7.6b and 7.6a). The signed-rank test gave a $p - value < 0.01$, showing that the two distributions (in both comparisons) are significantly different from each other, proving the effectiveness of the manipulation for the visual and cognitive distraction. We also observe a higher mean Mdev value for all 48 subjects in both distractive cases (see Fig. 7.5).

7.4 Conclusions

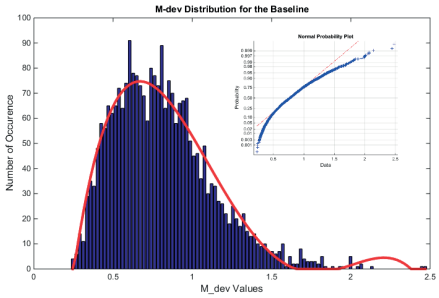
We have presented a database, called EPV-DIST, of 46 people recorded using three cameras while driving a simulator in baseline, visual distraction and cognitive distraction conditions. The recordings have been configured to represent a configuration that could be integrated and work robustly inside an actual car during real driving conditions. To



(a) Distribution of the visual M_dev values



(b) Distribution of the cognitive M_dev values



(c) Distribution of the baseline M_dev values

Figure 7.6: Comparison of the M_dev values for the three driving conditions, with a polynomial fit plotted on top of each distribution. The smaller plots show the deviation from a normal distribution, indicating that the distributions cannot be assumed normal.

the best of our knowledge this is the first database to include such a high number of subjects, in addition to the first time use of the multi-camera setup.

We have also presented some statistics on the driving performance of the participants and showed that the distraction induction methods result in a significant decrease in performance compared to the baseline, using the Mdev value as a performance metric. In the next chapter we present our proposed methods to detect cases of cognitive distraction using AUs, their dynamics and their inter-relations. We believe that the introduction of this database will stimulate further research in the field of driver state monitoring.

Action Unit based Cognitive Distraction Detection

8

8.1 Introduction

Cognitive distraction, which can be caused by mind wandering or cognitive load, is a factor that seriously affects the driving performance as it causes a diversion of attention from the primary task, that is driving. The studies on statistics of the effects of cognitive distraction, as well as other types of distraction, have been reviewed in Section 7.1. In this chapter we present a novel method to automatically detect driver's cognitive distraction using dynamic information of AUs as well as their correlations. For this purpose, we use the EPV-DIST database that has been presented in Chapter 7. The EPV-DIST database is a database of 46 subjects recorded using three NIR camera-light systems while driving a simulator and performing additional tasks at the same time, to induce visual and cognitive distraction separately.

In this chapter we only aim to tackle the problem of visual detection of cognitive distraction, since the focus of the thesis is facial actions and visual distraction is identified with changes in the head-pose and gaze avertion towards the distractive agent. Cognitive distraction is not a mental or affective state that has pre-defined universal expressions, such as those explained in Section 2.2.2. Its automatic detection is therefore a very challenging task, which is why in this work we propose novel feature extraction and data normalization schemes to handle it.

In the rest of the chapter, we first give a review of existing methods for automatic visual driver monitoring systems in Sec. 8.2. Then in Sec. 8.3 we give an overview of the system that we use to extract the visual features related to cognitive distraction, including a summary of the AU detection system that has already been presented in Chapter 6. In Sec. 8.4 we present our results for classifying the LCS as baseline or cognitive distraction, as previously described in Chapter 7. In Sec. 8.5 we give a discussion on the effectiveness of the system, as well as the individuality of the concerned facial actions and possible applications of the detection system. Finally, we conclude the chapter in Sec. 8.6.

8.2 Related Work on Visual Driver Monitoring

This section presents a brief review of existing work on visual driver monitoring for various applications as well as different modalities used for detecting various types of driver distraction. An extensive review is given in Dong *et al.* (2011) and Kang (2013), the interested reader is referred to these publications for more approaches and applications not listed here.

Over the years most of the research on visual driver monitoring systems have focused on fatigue detection, which is another critical factor for human error in driving and can be considered related to cognitive distraction, yet excluded from the definition that we adopt for cognitive distraction (see Sec. 7.1). An approach on fatigue detection, rather close to ours is the work by Vural *et al.* (2007), where the authors use many AUs, including head-pose, and analyzed their relation to fatigue during a three-hour simulator driving experiment after midnight. As expected, the most relevant features were related to eye-blink (AU45) and also outer brow raise (AU2), as the subjects tried to remain awake. In Rongben *et al.* (2004) an automatic mouth movement analysis is performed to detect fatigue related actions, and also speaking, while in Gu & Ji (2004) AUs are used within a Dynamic Bayesian Network (DBN) to detect driver vigilance. The head pose dynamics have also been successfully exploited in a real-time driver awareness detection system (Murphy-Chutorian & Trivedi, 2010). Another commonly used visual cue for fatigue detection is the Percent Eye Closure Measure (PERCLOS), as used for instance in Bergasa *et al.* (2006).

As for automatic detection of distraction, a non-vision based system is presented in Tango *et al.* (2010) where the driving information, such as the speed, position of the pedal and steering wheel have been used to detect visual distraction tested with various machine learning methods. Wöllmer *et al.* (2011) also used the driving information and non-vision based head tracking data to detect cases of visual distraction while performing various tasks. In Liang *et al.* (2007) the authors used eye movements and driving performance data within a Bayesian Network framework to predict $\sim 80\%$ of distraction cases while interacting with an in-vehicle information system (IVIS). A similar study is presented in Jimenez *et al.* (2012) where the gaze angle and fixation data was used once again to recognize distraction induced by the IVIS. In D'Orazio *et al.* (2007), the eye movements are analyzed to predict visual inattention using Neural Networks. The gaze information was used along with head movements and lane position of the vehicle in Kutila *et al.* (2007) to detect induced visual and cognitive distraction using a stereo-vision system integrated in trucks and passenger cars. For cognitive distraction, the authors achieve 68% on the truck experiments and 86% for the passenger car experiments. However, the low number of drivers tested (3 for the passenger car, 12 for the truck) is insufficient to discuss the generalization capability of the system. In Jabon *et al.* (2010) several features related to the coordinates of 22 facial landmarks and driving data were used to predict accidents. In Ragab *et al.* (2014) the arm position, eye closure, eye gaze, facial expressions, and orientation provided by Kinect to detect visual and manual distraction on 6 subjects.

The closest approach to one that is presented in this work is the one by Li & Busso (2015) where the authors use AUs, gaze and head pose information to detect visual

and cognitive distraction. Apart from the difference in the methodology to induce the cognitive distraction, a very important difference in their approach is that they ask human evaluators to extract sequences of distraction based on the facial behaviour of the drivers. This fact makes their work and ours incomparable, as the problem they try to tackle becomes how to detect human perception of expressions of distraction. Nonetheless, it provides us a list of AUs related to this problem, which can be used for comparison with the discoveries in our work. With the experiments performed on 20 subjects, the F-score for detection cognitive distraction is 73.8% and for visual distraction 80.8% (Li & Busso, 2015).

To the best of our knowledge, ours is the first work that presents a completely automatic system, that can be integrated in a vehicle, to detect cognitive distraction with an objective ground-truth, using non-intrusive visual monitoring of the driver's face and tested on such a large variety of subjects.

8.3 System Overview and Proposed Feature Extraction Scheme

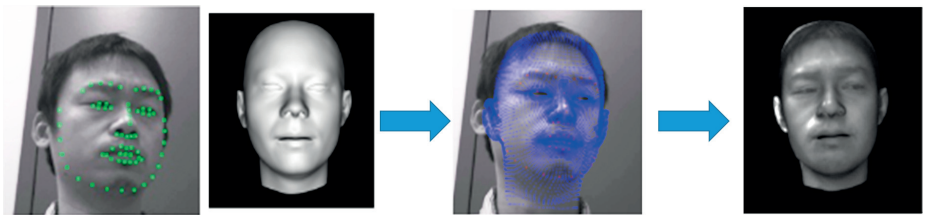


Figure 8.1: Virtual face generation pipeline

This section describes the methods built and adopted in order to detect the presence of cognitive distraction in the sequences of the EPV-DIST database (introduced in Chapter 7) via the driver's facial actions. In our context, this means classifying each recorded LCS as belonging to the baseline or cognitive task, as explained in Section 7.3. The outline of the pipeline is as following: First we generate a virtual frontal view of the driver's face in each frame using a Bilinear 3D face model and texture mapping from a 2D image. Then we detect 14 AUs on the generated virtual frontal view of the face by extracting Scale Invariant Feature Transform (SIFT) features and applying SVM classification for each AU separately. Next, we extract features from the dynamic continuous valued output of the SVMs, also investigating the correlated behavior between the AUs and finally feed these features in an SVM or Random Forest classifier to obtain a binary response for each sequence as distracted or not. The details of each method, as well as their implementation are given in the rest of the section.

8.3.1 Virtual View Generation from Three Cameras

The model based face pose normalization / frontalization has been applied widely in face recognition (Gao *et al.* (2009), Asthana *et al.* (2011), Taigman *et al.* (2014)). It is also known as virtual face frontal view generation. One can fit a 2D deformable mesh

model to a non-frontal face and apply non-linear warping to generate a virtual frontal face (Gao *et al.*, 2009). However, it has been shown that warping with a sparse 2D mesh model is sub-optimal due to artifacts and discontinuity. Instead, we fit a 3D dense mesh model and map texture directly to the mesh vertices. The frontal view face is rendered by applying inversed rigid motion of the 3D face model. Fig. 8.1 shows the concept of our face pose frontalization method.

Fitting a 3D dense face mesh model with texture information is far from efficient for real time application. We adopted a feature based 3D mesh model fitting and which provides a good balance between fitting efficiency and accuracy for real time virtual frontal view generation. In order to recover the expression variations in addition to identity variations of human faces for emotion detection applications, a bilinear 3D morphable model (Cao *et al.*, 2014a) is considered. The model has two sets of parameters, which control expression changes and identity changes separately.

A sparse set of facial landmark features are then selected and the objective of the 3D model fitting is to minimize the projection error of these facial landmark features, with respect to the corresponding 2D features detected on a 2D facial image. In total 68 salient facial features are selected and the 2D salient facial features are detected and tracked using SDM (Xiong & De la Torre, 2013), which is reused in the AU detection step as explained in Sec. 8.3.2.

The feature based 3D face mesh model fitting can be easily extended to multiple camera setup. The coefficients of the bilinear face model are not dependent on the viewpoint because they are characterizing 3D object's shape and not its projection on the image plane. It has been shown in Faggian *et al.* (2008) and Ecabert *et al.* (2015) that fitting a 3D morphable model in a multi-view setup provides more accurate and robust results. Therefore, we reconstruct the mesh based on the tracking from the three cameras positioned as shown in Fig. 7.1a.

To generate a virtual frontal view image, texture information needs to be extracted from a 2D image and the values are mapped on the corresponding vertices of the reconstructed 3D face mesh. The texture information can be extracted from a specified camera view, or the optimal camera view, or an adaptive fusion of multiple camera views. Given a reconstructed 3D face mesh $\mathbf{f}$ and its estimated projection operator $\mathbf{L}$ with respect to an input face image $\mathbf{I}$, the visible vertices in $\mathbf{f}$ are determined by checking the normals and the viewing angle. Those vertices are projected on the 2D image plane with the projection operator, The underlying pixel values T are assigned to their corresponding visible vertices. An example of a rendered frontal face image is show in Fig. 8.1. In this work, we obtain the pixel values from the view with the smallest absolute yaw angle, which is mostly the frontal view due to the nature of the driving and secondary tasks. Fig. 7.2 shows the three views and the reconstructed virtual face on an highly expressive real-case frame from our database.

8.3.2 AU detection from Virtual Frontal View

Once we generate the virtual frontal view from the three cameras, we detect 14 AUs from the generated frame. For this purpose we adopt the system that we have proposed in Chapter 6 and Yüce *et al.* (2015) and that has won the AU occurrence de-

tection sub-challenge of the FERA2015 (Valstar *et al.*, 2015). The FERA2015 challenge was organized to promote advances in research on AU and AU intensity detection. It is composed of two challenging datasets (BP4D (Zhang *et al.*, 2014) and SEMAINE (McKeown *et al.*, 2012)) with spontaneous and natural behavior each annotated frame-wise for the presence and intensities of AUs. The participants were provided with two sets of training and development partitions and asked to send a working program that would be applied on two unseen test partitions, in order to assess the efficiency of the systems in a blind manner, i.e. without the advantage of parameter tuning or usage of prior knowledge on the data. Our framework presented in the scope of FERA2015 allows us to obtain a continuous AU occurrence signal for 14 AUs, which are listed in Table 8.1 along with their definitions. For more information on these and other AUs please see Chapter 2.2.1, where we provide an extensive review on the FACS. An overview of the system is presented here and more details have been provided in Chapter 6.

Table 8.1: Detected AUs and their definitions

Action Unit	Definition
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU6	Cheek Raiser
AU7	Lid Tightener
AU10	Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU23	Lip Tightener
AU25	Lips Part
AU28	Lip Pucker
AU45	Blink

The initial step in the AU detection system is to locate the facial landmarks, around which we will then acquire the relevant appearance based features. For this purpose, we use the state-of-the-art face tracker based on the SDM (Xiong & De la Torre, 2013). The SDM starts with an initial guess and estimates the shape using a cascade of regression models that are learned at each step using local texture features (e.g. SIFT) extracted from the landmarks estimated in the previous step. Note that, since the virtual view generation and AU detection systems are currently implemented as two separate pipelines, we reapply the SDM tracker on the generated virtual view. In the future, these systems will be combined for efficiency reasons. The SDM outputs the locations of 49 landmarks and using this mask we calculate the locations of 8 additional *non-salient* landmarks. The details for the calculation are present in Section 6.2. These additional

points (AP) are generally excluded from face trackers or facial landmark detectors as they mark transient features of the face and their annotation and detection are not as trivial as the non-transient landmarks. However, they contain very important local appearance information related to facial actions as many appearance changes occur around these points during certain muscle contractions. These points can be seen on an example virtual face image from the EPV-DIST database in Fig. 8.2 along with original SDM landmarks and their locations and some of the related AUs are listed as follows:

- AP1 - The center of the eyebrows, relevant to AU4 and AU1
- AP2 and AP3 - Around the crow-feet wrinkles, relevant to AU6 and AU7
- AP4 and AP5 - Sides of the nose, relevant to AU10 and AU9 (nose wrinkler)
- AP6 and AP7 - Nasolabial furrows, relevant to AU6 and AU10
- AP8 - On the chin, relevant to AU17

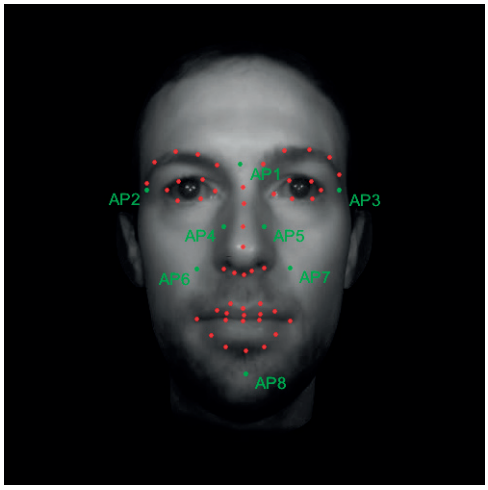


Figure 8.2: The facial mask used to obtain appearance features. The red points show the original SDM landmarks and the green ones are the additionally calculated points.

After obtaining the landmarks from the face tracker, the face is aligned using the eye locations to correct for any possible in-plane rotation still remaining from the virtual view generation. This is performed before the APs are extracted, so that the calculation of their locations is invariant to the head-pose. Later, the face is scaled to a fixed size of 200 by 200 pixels and the SIFT features (Lowe, 2004) are extracted around the 57 landmarks in total. SIFT features have been effectively used in mainly object recognition and tracking (Li & Allinson (2008),Zhou *et al.* (2009)) and successfully applied on the AU detection problem as well (Ding *et al.* (2013), Ringeval *et al.* (2014)). The SIFT descriptors extracted in a 32 by 32 local neighborhood around each landmark result in a feature vector of size 7296, which is then reduced using Principal Component Analysis (PCA), retaining a certain number of final features learned for each AU separately.

These features are used to train a L1-regularized linear-SVM classifier for each AU separately on a custom made training set that includes images from the CK+ (Lucey *et al.*, 2010), GEMEP-FERA (Valstar *et al.*, 2011) databases in addition to the challenge datasets SEMAINE (McKeown *et al.*, 2012) and BP4D (Zhang *et al.*, 2014). The training set consists of 6713 images in total and, in addition to the well-established standard database CK+ of posed expressions, includes many non-posed, or spontaneous, examples of expressions from the other three databases. This fact is particularly useful when the system is applied on real data, as in the case of our application.

The results we obtained on the unseen test-set of the two challenge datasets are revisited in Table 8.2 in comparison with the best challenge baseline results. The presented F1 scores on this challenging AU detection problem shows the efficiency of the system and proves suitable for use in a real application. Note that, although, the original work (Chapter 6 and Yüce *et al.* (2015)) proposes a multi-label manifold embedding scheme to improve AU detection and achieves a better result on one of the two unseen partitions, we have chosen not to adopt this part of the system in order to obtain a better generalization on unseen data.

Table 8.2: F1-Scores on the FERA Challenge (revisited)

Database	BP4D	
AU	Prop. System Yüce *et al.* (2015)	Best Baseline Valstar *et al.* (2015)
1	**0.261**	0.188
2	0.167	**0.185**
4	**0.283**	0.197
6	**0.729**	0.645
7	0.785	0.799
10	**0.802**	0.801
12	0.779	**0.801**
14	0.625	**0.72**
15	**0.348**	0.238
17	**0.380**	0.311
23	**0.441**	0.320
Average	**0.508**	0.473
Database	SEMAINE	
AU	Prop. System Yüce *et al.* (2015)	Best Baseline Valstar *et al.* (2015)
2	**0.655**	0.569
12	**0.769**	0.595
17	**0.215**	0.091
25	**0.623**	0.445
28	**0.251**	0.250
45	0.325	**0.396**
Average	**0.481**	0.391

The SVM classifiers each give a continuous value output, which is the distance to the hyper-plane. It has been long debated in the community whether the output of classifiers trained in a binary manner should be used to quantify the intensity of AUs. For example, recently Girard *et al.* (2014) have shown that the intensity of smiles are better recognized

using classifiers directly trained with annotated intensities . Nonetheless, we use the decision to the hyper-plane of the SVM as a relative intensity measure as it provides enough comparative information when the purpose is not a direct AU intensity detection defined by the FACS (Friesen & Ekman (1978), Sec. 2.2.1.

8.3.3 Feature Construction

For the classification between the sequences belonging to the baseline and cognitive distraction in the EPV-DIST database (Chapter 7), we extract features from the AU signals obtained using the system described in Sec. 8.3.2. The sequences that contain too few frames due to misconduct during recording or face detection failure because of heavy occlusion (e.g. by the hands on top of the steering wheel) have been removed from the analysis resulting in a total of 4520 LCS.

The first set of features, which we will refer to as Feature Set 1 from this point on, comes directly from the continuous individual AU signals. For each of the 14 AU signals (see Table 8.1 for the list) we obtain the mean, variance, maximum and minimum values along the 8.5 second sequences. This process is performed by dividing the sequence in four in time. The reason for splitting the signals in time is to make use of the differences in AU behavior that may occur on different portions (or quarters) of the LCS. For instance, a person might display a facial reaction while listening to the calculation sentence he/she needs to respond to, or during the lane change task which follows the auditory input. Splitting the feature extraction into smaller segments makes it feasible to extract this sort of dynamic information.

The second type of features (Feature Set 2) are derived from the cross-correlations of AUs on different time delay levels. While constructing these features we were inspired by the Appraisal Model of Emotion, as proposed by Scherer (Scherer, 2001), which states that the activation of certain physiological components are coupled, or synchronized, when we are faced with an emotional stimulus. An extensive explanation of this model, also called the CPM may be found in Section 2.1.3. Also following this theory, Kroupi *et al.* (2013) have shown coupling between the phase and amplitude of the EEG and EDA signals while the subjects are watching emotionally stimulating music videos. Another example of a similar analysis is the multiple works by Williamson et al., who have shown the existence of a difference in coordination, movement, and timing of vocal and facial components between patients suffering from Major Depressive Disorder (MDD) vs. control subjects (Williamson *et al.* (2014), Williamson *et al.* (2013)), winning the AVEC 2013 (Valstar *et al.*, 2013) and AVEC 2014 (Valstar *et al.*, 2014) challenges on automatic detection of MDD severity.

Using a similar idea, we calculate the cross-correlation between each of the 14 AUs, within a delay of -80 to $+80$ frames with 2 frames interval. This corresponds to a signal of length 81 for each AU pair and allows modeling the sequential behavior between AUs on a scale of -4 to $+4$ seconds. From those signals we extract, once again, the mean, variation, maximum and minimum values, in addition to the location in time of these maximum and minimum values, and the correlation values at delays corresponding to $-40, -30, -20, -10, 0, 10, 20, 30$ and 40 frames, i.e. at each second in a bi-directional manner. This enables us to obtain an extensive set of features that represent factors like

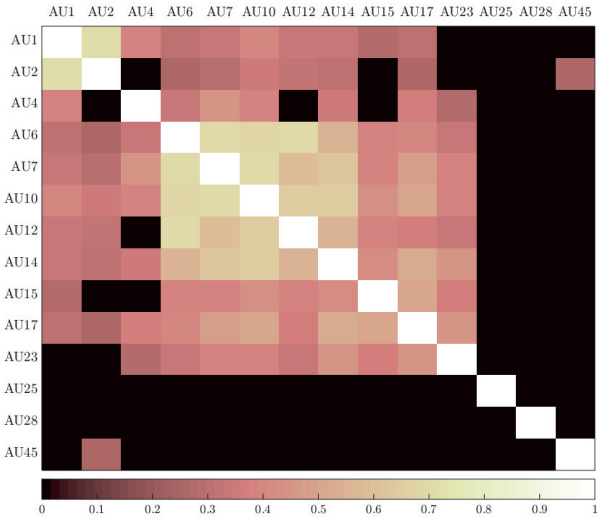


Figure 8.3: Correlation table for the 14 AUs, a higher correlation indicates a high number of co-occurrences between AUs in the training set. The matrix was truncated for values < 0.25 to indicate the cross-correlations that were excluded

the total amount of co-activation and its variation, moments of maximum and minimum synchronization and the level of co-activation at certain levels of delay between AUs located in similar or different parts of the face. Finally, we truncate the feature set according to the correlation priors between AUs. This truncation serves for keeping AU combinations that frequently occur and removing those with little or no correlation. In case such an unusually high correlation is observed, for instance caused by a distortion from the virtual view reconstruction due to heavy head-pose, this process will make sure this noisy observation has no effect on the overall feature set. As correlation priors we use the co-occurrence table of AUs obtained from the AU detection training set, as used in Chapter 6, and use a threshold of 0.25 as shown in Fig. 8.3.

Our hypothesis is that this dynamic co-activation information will help better differentiate the facial behavior of the complex mental state that is cognitive distraction. In Section 8.4, we show that, indeed the cross-correlation based features improve the accuracy on a subject based analysis, yet they are not so helpful for the subject-independent classification task.

8.3.4 Person Specific Normalization for Classification using SVM and Random Forests

The final component of the distraction detection system is the classification part. For this, we use linear SVM for the subject-based tests, where the training and test examples are relatively on similar manifolds compared to between-subject tests. For the subject independent tests, we therefore compare the performance of the SVM with RF classifiers. RF are known to be less effected by over-fitting thanks to their bagging mechanism (Breiman, 2001). They learn the best splitting by multiple features each

time randomly choosing a random subset of samples and features. They are also more suitable for cases with a large number of features.

For the SVM training and testing we make use of the LibSVM library (Chang & Lin, 2011) and for the RF we use the Scikit-learn machine learning library for python (Pedregosa *et al.*, 2011)[1]. The hyper-parameter C for the SVM and the number of trees, maximum number of features and minimum number of samples per split hyper-parameters of the RF were optimized using a 5-fold cross-validation on the training data in a subject independent manner.

While analyzing the data we have discovered that although both type of features are very effective in discriminating the distraction and baseline sequences of individual subjects (see Subsection 8.4.1) the performance on the subject-independent tests are very low. We assume that the reason is that the types of features that we use are discriminative enough to model individual behavior, yet they remain too person-specific. Indeed, visualizing the data, we have seen that most of the subjects are clustered among their own samples. Fig. 8.4a shows an illustration of this phenomenon using the first two principal components of Feature Set 1 after PCA applied on all training data. We can see that even using two dimensions the data points belonging to the test subjects can be easily separated into the labels, yet the same is not true for the training data, for which samples from the two labels do not demonstrate any noticeable pattern and are scattered across the feature space instead. After the subject-based normalization, on the other hand, we observe that the data is much better centralized (Fig. 8.4b). Of course, the projection on 2D is not very meaningful when using complex classification methods; Fig. 8.4a and 8.4b are only for illustrative purposes.

To overcome this problem, instead of the common convention of normalizing the whole training data to zero-mean and unit standard deviation, we propose to perform this operation subject-wise, as:

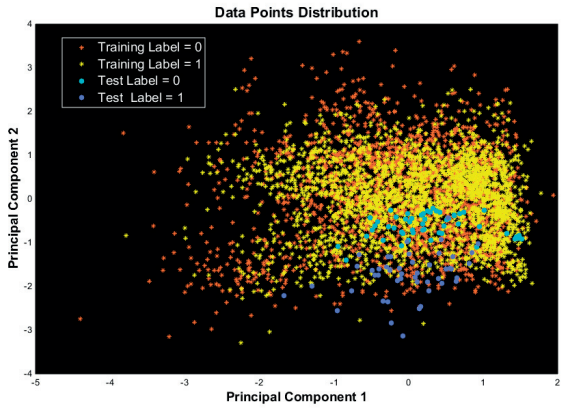
$$\forall x \in F : x = \frac{x - \overline{x}_s}{\sigma_s} \tag{8.1}$$

where x is any point in the features set F belonging to the subject s, $\overline{x}_s$ is the mean of all data points belonging to subject s and σ_s the standard deviation. Even though, this may seem as a factor preventing a real-time application on an unseen subject, the only implication it brings is actually the need for some seconds of frames from the considered subject. In other terms, the person-adaptation is completely unsupervised, does not require any re-training of the classifier (as we only need to change the placement of the test subject) and as seen in Section 8.4.2 increases substantially the subject-independent detection rates.

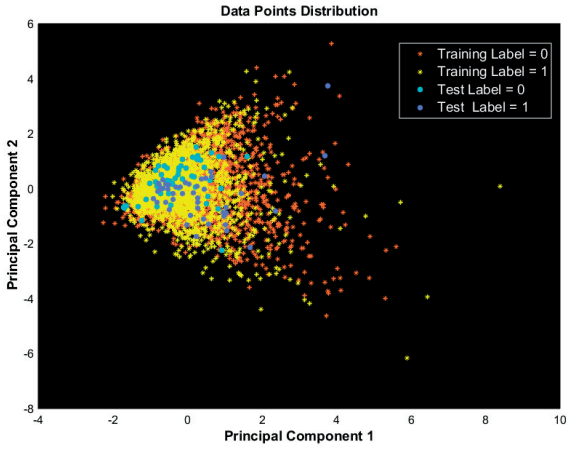
8.4 Classification Results

This section presents our classification result for the baseline vs. cognitive distraction cases. Out of the 4520 LCS in total ($\sim$ 100 per subject), the number of the sequences for the cognitive distraction case is 2156. We present our experimental results for the

[1] http://scikit-learn.org



(a)



(b)

Figure 8.4: The data distribution on the two first principal axes with an example subject chosen as the test case, before (top) and after (bottom) the subject based normalization

classifiers trained per-subject and in a subject independent manner using different feature configurations and classifiers. Table 8.3 presents the accuracies for the best performing systems for the two types of experiments, serving as a summary of the results and the details are presented in the rest of the section.

Table 8.3: Results of Best Performing Systems for Subject Independent and Dependent Cases - OA: Overall Accuracy, F1: F-score, Prec.: Precision, Rec.: Recall

	OA (%)	F1 (%)	Prec.	Rec.
Sub. Dependent	95.51	95.16	96.38	93.97
Sub. Independent	68.10	65.57	67.22	64.00

8.4.1 Subject-Dependent Cognitive Distraction Detection

We first train classifiers independently for each of the 46 subjects in a leave-one-out manner. That is, we learn the classifier hyper-parameters using a 5-fold cross-validation and train the classifier with the best parameters on all sequences for a certain subject points except for one, and test it on the left-out sequence, or data point. The classifier we use for the subject-dependent tests is the linear SVM. In Table 8.4 we show the results obtained using Feature Set 1 alone, Feature Set 2 alone and the two in combination, and compare the accuracies obtained with and without the truncation of correlation features as explained in Sec. 8.3.3.

Table 8.4: Subject Dependent Detection Results - OA: Overall Accuracy, F1: F-score, FT: Feature Truncation on Set 2

Feature type	OA (%)	F1 (%)
Feature Set 1	93.74	93.39
Feature Set 2	93.85	93.47
Feature Set 2 + FT	93.89	93.49
Features Sets 1 + 2	94.88	94.57
Feature Sets 1 + 2 + FT	**95.51**	**95.16**

As shown in Table 8.4, the best results are obtained by combining the features extracted directly from AU signals (Feature Set 1) and those from the cross-correlations (Feature Set 2), supporting our hypothesis that the dynamic inter-relations of AUs are useful in determining individuals' expressions of cognitive distraction. Using Feature Set 2 alone also proves as efficient as using Feature Set 1. The best accuracies obtained are 95.51% with a standard deviation (std.) across subjects of 3.44 for the overall accuracy and 95.16% for the F-score with std. 3.67. These values are calculated over all data points, which corresponds to an average weighted by the number of sequences per subject. The very high accuracy measures, and low variation among subjects, demonstrate the efficiency of the proposed system, when it is trained on labeled data of a specific subject. A side-observation is that, the feature truncation improves accuracy in both of the relevant cases (Feature Set 2 alone and Feature Set 1 and 2 combined), validating the usefulness of exploiting prior AU co-occurrence information. The per-subject accuracies for the best performing system are shown in Fig. 8.5, which will be referred to again in the following subsections.

8.4.2 Subject-Independent Cognitive Distraction Detection

The second set of experiments we have performed is the subject-independent tests, that is carried out in a leave-one-subject-out manner. This time, we also use RF in comparison with SVM, since RF are known to be less affected by overfitting on training data, or subjects in our case. Table 8.5 presents the results obtained using both classifiers, Feature Set 1 and 2 alone and in combination, additional PCA (retaining 98% of the total variance, performed for SVM only since RF internally handle the problem of irrelevant features) and the subject-based normalization as explained in Sec. 8.3.4.

Table 8.5: Subject Independent Detection Results - OA: Overall Accuracy, F1: F-score, FS: Feature Set, SN: Subject-wise data normalization

Cl. type	Feature type	OA (%)	F1 (%)
SVM	FS 1	63.36	59.69
	FS 1 + PCA	63.74	58.42
	FS 1 + SN	**65.5**	**62.32**
	FS 1 + PCA + SN	65.35	61.77
	FS 2	57.85	54.95
	FS 2 + PCA	59.38	55.31
	FS 2 + SN	61.39	58.95
	FS 2 + PCA + SN	62.72	62.14
	FS 1 + 2	61.82	59.19
	FS 1 + 2 + PCA	59.58	54.65
	FS 1 + 2 + SN	62.99	60.92
	FS 1 + 2 + PCA + SN	63.96	61.49
RF	FS 1	63.98	61.93
	FS 1 + SN	**68.10**	**65.79**
	FS 2	57.19	59.31
	FS 2 + SN	63.98	61.49
	FS 1 + 2	58.83	60.61
	FS 1 + 2 + SN	65.29	64.17

The best results are obtained using RF classifier with Feature Set 1 alone when the person-specific normalization is applied with overall accuracy 68.10% (std. = 12.71) and F-score 65.79% (std = 14.02). The person-specific normalization is indeed very effective with all features types, especially with RF. This confirms our rationale explained in Sec. 8.3.4, claiming that the data points of each subject are clustered separately in the feature space. However, it is not effective enough to obtain an accuracy close to the classifiers trained in a subject-based manner (Sec. 8.4.1). As it can be seen in Fig. 8.5, which shows the comparison of subject dependent and independent results for each subject, this effect is more critical in certain subjects (e.g. Subjects 4, 7, 13) and less in others (e.g. Subjects 5, 6). Also, we observe that the correlation related features (Feature Set 2) do not increase the detection efficiency when used in combination with Feature Set 1, and also result in lower results when used alone. These results suggests the individuality of such dynamic multi-AU patterns, i.e. that this kind of information is more meaningful when it is learned on each subject independently. This problem of individuality is discussed further in the rest of the chapter.

8.4.3 A look into the relevant features

In order to see which AUs or AU pairs are the most relevant to our proposed classification task we inspect the correlations of each feature in Feature Set 1 and 2 with the ground-truth labels for baseline and cognitive distraction. Since the subject-dependent

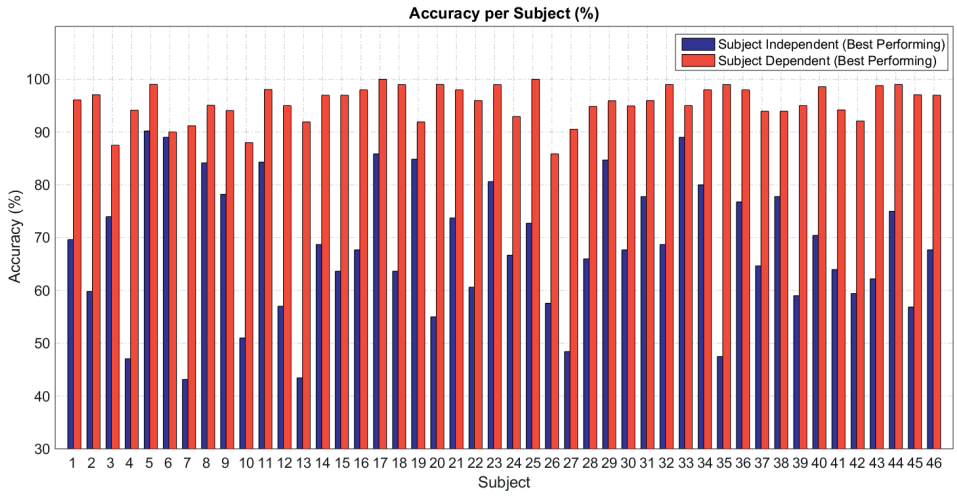


Figure 8.5: Overall Classification Accuracies for each subject, for the best performing methods in subject independent and subject-based training conditions

classification is significantly more efficient compared to the subject-independent one, we find it more meaningful to perform this analysis on a subject level as well.

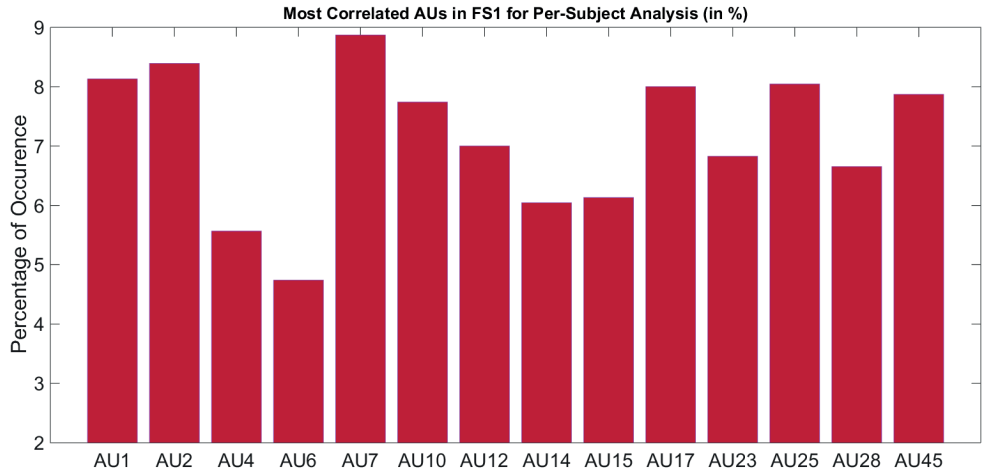


Figure 8.6: Percentage of the most correlated AUs within the top 50 for each subject in Feature Set 1

First, we calculate the correlation of all 224 features from Feature Set 1 with the labels for each of the 46 subjects. Then, for the 50 most correlated features for each subject we look at which AU signal and which temporal segment they belong to. Fig. 8.6 shows the total percentage of each AU among those features, Fig. 8.7 shows the mean,

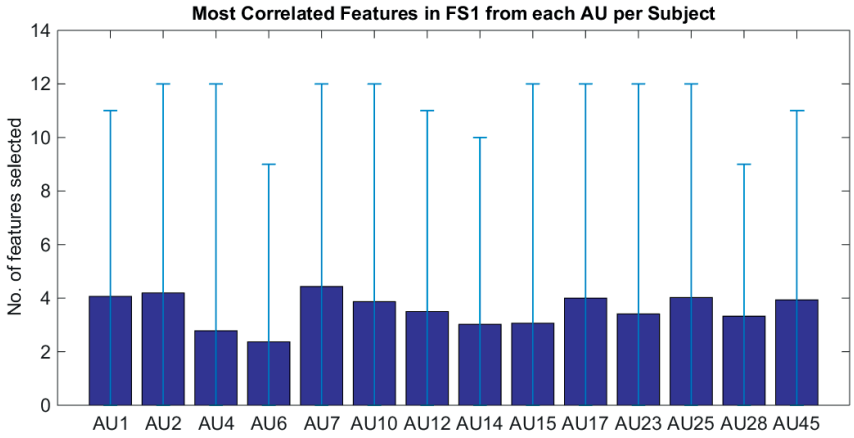


Figure 8.7: Average (per subject) number of features included in the 50 most correlated AUs set, with the min. and max. values shown by the error bar

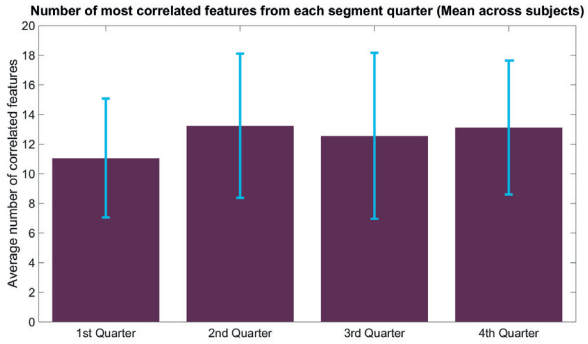


Figure 8.8: Average (and std.) number of features selected from each time segment of the sequences within the 50 most correlated for each subject in Feature Set 1

minimum and maximum number of each AU per subject and Fig. 8.8 shows the mean and standard deviation for each of the four segments (std. removed from Fig. 8.6 for clarity of presentation). We observe that the AU that appears most in the analysis is AU7 (eye-lid tightener), which indeed appears frequently in expressions related to concentration, thinking or focusing. It is followed by the outer and inner eye-brow raise motions AU2 and AU1, lips part AU25 and chin raise AU17, without any clear difference in amount of occurrence. The fact that many AUs occur frequently in the list of correlated features once again shows the large variety of expressions related to cognitive distraction, and helps explaining the difficulty in obtaining a highly-accurate subject independent system. Two of the five most correlated AUs (AU1 and AU17) are also in line with the features found relevant to human perception of cognitive distraction, reported in Li & Busso (2015).

For the temporal segments, none of the segments seem to dominate the others; yet,

the first quarter is observed to appear less (see Fig. 8.8). This is expected, since it corresponds to the first two seconds of the LCS where the secondary task is presented (mental calculation) and the lane change task appears only in the second quarter, forcing the driver to divide his attention and workload between tasks.

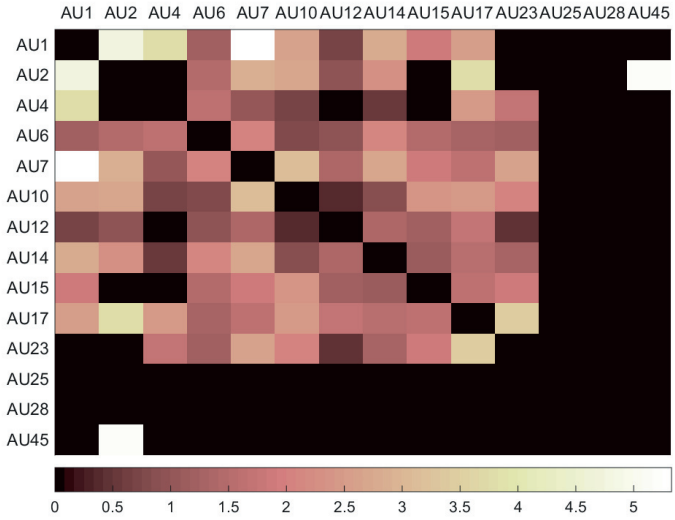


Figure 8.9: Percentage of the most correlated AU pairs within the 100 most correlated for each subject in Feature Set 2

We perform the same procedure for Feature Set 2 and plot the percentage occurrence of features belonging to AU pairs as seen in Fig. 8.9. This time we investigate the 100 most correlated features, as the whole set is larger (of size 750). Some relevant AU pairs worth mentioning are $AU7-AU1$, $AU45-AU2$, $AU1-AU2$, $AU1-AU4$, $AU17-AU23$ and $AU2-AU17$. Although it is harder to interpret the features this time, we observe that most of them are related to eye / eye-brow actions, as in the single AU case. Interestingly, the most correlated upper AU pair includes the AUs identified as related to fatigue in Vural *et al.* (2007), which possibly implies an effort to regain attention, commonly in the two conditions. The lower face combination $AU17-AU23$, on the other hand, can appear in expressions related to pensiveness or assessment of coping potential depending on the simultaneous upper face actions (see Section 2.2.2). Therefore, it makes sense that this combination is relatively meaningful for the differentiation between cognitive distraction and baseline.

8.5 Discussion

The presented results demonstrate that, although a subject based training or adaptation is necessary in order to obtain a highly precise detection, the subject-independent system still achieves an acceptable accuracy in detecting the cognitive distraction sequences. It is a known fact that the cognitive distraction is not one of the basic emo-

tions (or states) that are conveyed similarly by everyone in terms of facial expressions (Section 2.1.1). The subject dependency issue is therefore expected in systems aiming at recognizing such complex expressions. As for the unsupervised subject-based normalization proposed, with an unseen driver the system requires only some couples of seconds of images of the driver's face to increase the detection rate $\sim 4\%$.

As stated earlier, the system is designed so that it can be completely integrated into a passenger vehicle. All components, except for the texture mapping to the generated 3D mesh, work in real-time ($> 15fps$). The texture mapping still needs to be optimized for real-time operation, and integrating the two separately implemented pipelines of virtual view generation and AU detection will also increase speed.

A possible real-world application could be to integrate the system within the human-machine interface of the vehicle, and to activate a visual or audio alert to warn the driver in case a critical level of distraction is detected, as a part of Advanced Driver Assistance Systems (ADAS). With the semi-automatic driven cars slowly entering our lives, such systems gain even more importance, for instance to assess the driver's state when the driver needs to retake the car's control or to decide when it is safe (or suitable) to switch to fully autonomous driving. The current system outputs a decision based on 8.5 seconds of recording due to the definition of the lane change task, but is fully adaptable to shorter or longer durations and to the fusion of multiple sequences. For instance, a moving window that collects distraction info in time can be utilized and the relevant alert system could be activated when the number of segments involving cognitive distraction reaches a certain threshold. According to the detected level of cognitive distraction the severity of the countermeasure can also be adjusted, ranging from a small alerting beep or a message on the console to automatically slowing down or even stopping the car when conditions are suitable.

Further user studies need to be performed in real driving conditions to assess the robustness of the detection, considering the head movements in real conditions, but a general convention in driver feed-back systems is to alert the driver timely, only when really needed and in a way that does not annoy the driver. This requires a good balance between the precision and recall, i.e. false positives and false negatives. The proposed system is suitable to be adjusted for the precision/recall ratio (e.g. by tuning the decision level of the classifiers) and the length of the temporal window and the previously mentioned threshold can also be tuned to obtain the best compromise between the driver comfort and safety.

Distraction does not affect our lives only in the driving context. Knowing that abnormalities in maintaining attention are symptoms of disorders such as Attention deficit hyperactivity disorder (ADHD), Asperger's syndrome or other Autism spectrum disorders, and considering the high accuracy of our subject dependent system, another possible use of the proposed system could be a personalized monitoring system to provide feed-back during treatments, that involve interaction with a human or a machine, of individuals suffering these disorders. A review of works on such interactive technologies can be found in Boucenna *et al.* (2014) and Section 3.5.

8.6 Conclusion

In this chapter we have demonstrated a complete pipeline to discriminate the cognitive distraction segments from the baseline based on AUs, for the sequences in the EPV-DIST database. The proposed system first reconstructs a virtual frontal face image using the input from the three cameras, applies AU detection on the virtual image, then uses features extracted from the dynamic AU signals and cross-correlations of AU pairs to classify segments in the two driving conditions.

Using different configurations and methods we obtain an accuracy of $\sim 95\%$ when the system is trained separately on each subject, and $\sim 68\%$ in the subject-independent case. Based on these results, and further analyses we identify that facial expressions of cognitive distraction vary hugely among subjects and also report the AUs and AU pairs that show relevance most commonly among the subjects. The completely automatic non-intrusive detection system is ready to be accommodated in consumer vehicles for use within applications aiming to prevent, or decrease, human error in accidents. Our further research will include the gaze and head-pose related features and their benefits for detecting the various types of distraction along with AUs. Compared to existing related work, this study is the one performed with the highest number of participants using solely automatic analysis of facial actions.

With this chapter we conclude the presentation of our contributions. This chapter has been an application of most of the concepts and our other contributions presented in the previous chapters. It is important as it shows a *real-life* application of how the presented methods can be beneficial in our daily lives.

9 Conclusions

In this final chapter we summarize our contributions presented in this thesis and discuss the benefits they bring about, as well as their limitations. We also present an outlook for future research in the related areas and propose ideas for advancement.

9.1 Summary and discussion of findings

Our first contribution (in Chapter 4) presents a system that looks at automatic facial expression analysis from an image processing point of view, applying fundamental filtering tools to the problem. All three of these tools (bilateral filtering, morphological opening by reconstruction and morphological top-hat transform by reconstruction) have the property of edge preserving and thus when they are applied to a face image they conserve the information relevant to facial actions and eliminate those that are to be considered noise. In addition, the novel scheme of extracting the LBP from overlapping windows of different sizes also enhances the potential of LBP based texture representation and advances the state-of-the-art in the area.

One potential drawback of the system is the rather long computation times for the iterative reconstruction algorithms, as well as the computational burden brought by adding more windows for extraction. However, with today's systems this is a very manageable issue since the proposed algorithms are very suitable to be parallelized and programmed on the Graphics Processing Unit (GPU).

In Chapter 5 we have proposed an extended feature extraction scheme using curvature Gabor wavelets. It is common to use Gabor filters in combination with LBPs, but this is the first time the curvature Gabor filters have been investigated in terms of facial action recognition, and AU detection in particular. We have presented a detailed analysis of the effects on accuracy of combining different curvature degrees and filter sizes and showed that one can achieve very high detection rates using this enriched representation of appearance. The curvature type of features are of special importance in terms of facial actions since they can better explain the changes in texture of the skin due to the muscle movements (e.g. on the wrinkles accented) but also the curved form of permanent facial features such as the eyes, mouth contours etc.

We have also presented results for a cross-database experiment, and showed the method's potential for generalization, especially for certain AUs. The performance, as expected, is lower than the one for within-database tests and it is true that the very high dimension of the feature space might have caused the system to better predict data that is similar to the training instances. More tests need to be performed with a larger training set to reach a conclusion. The very high accuracy of the single database case, already carries a lot of potential, for example for systems that are custom optimized for a certain subject or a specific image acquisition configuration. The very high precision of the system makes it suitable to replace manual FACS coding with a semi-automatic system, e.g. for neuroscience research that investigates facial behaviour patterns. The drawback of computation time is also valid for this case, but the system is also eligible for speed optimization using GPUs, as each of the filters operates individually and they are not combined until the feature selection step.

In the second part of the thesis contributions (Chapter 6) we have presented a multi-label approach to AU detection by an embedding scheme to incorporate AU correlations in the training. This time, our base system uses SDM based face tracking and SIFT for the appearance features, allowing us to obtain real-time performance. Although the accuracies are somewhat lower than the systems presented in Chapters 4 and 5, the high training and operation speed has allowed us to test various extensions and improvements also with more flexibility in terms of generalization. These extensions include an extended set of facial landmarks, not included in the facetracker mask, to mark transient facial features that are important for facial actions.

The multi-label embedding extension was applied on the FERA 2015 challenge for AU occurence detection and was shown to outperform both the baseline and other contesting systems. Although this extension works better compared to our proposed *base system* for a portion of the tested data, it has not been the case for all AUs in all datasets of the challenge's unseen test partition. The obtained results may suggest that the success of the proposed might be data dependent. Probably, the embedding framework is more effective when the test data has a similar distribution to the training in terms of AU co-occurences. Nonetheless, considering the low number of studies on this topic, we show that the label correlation information is useful for learning to detect AU occurrences. It is very intuitive that the AUs cannot be considered completely independent of each other and a certain amount of prior on what combinations they appear in helps building a more informative basis to embed the data. Our further studies will focus on making the method better generalizable to unseen data obtained in conditions completely independent than those for the training.

The final part of the dissertation focuses on a visual driver monitoring system, that combines the know-how presented in the prior chapters with novel frameworks to achieve a full-pipeline real-world application. In the first half (Chapter 7) we present a new video database of driver behaviour while being induced distraction. This database is important mainly in two senses. Firstly, it is a large source of data that is recorded in conditions representing those that can actually be integrated in an actual car to work in real driving situations. This is obtained through the multi-camera system that acquires NIR videos. This fact makes the setup convenient for use in large head-pose variations and ambient light changes, which are the most critical challenges in terms of computer vision for out-

door and natural conditions. Secondly, the database has been validated for the induction of the two types of distraction and contains per-sequence annotations for the driving performance and the raw driving information. These two qualities make the database very beneficial for future research in affective computing and computer vision.

In Chapter 8 we propose a framework to detect sequences of cognitive distraction in this dataset. The system uses the AU detection system presented in Chapter 6 to obtain a continuous valued output for 14 AUs then extracts dynamic information from these *AU signals* and their cross-correlations. Both inspecting the data and from the results we have observed that the facial manifestation of the cognitive distraction state is highly dependent on the individual. This is an expected outcome as it is a state that does not directly correspond to any of the basic emotions (and thus basic expressions) or a unique combination of SECs in the CPM approach. We believe this is one of the reasons why automatic detection of cognitive distraction through facial expressions has not been studied very often in the literature. To this effect we have also proposed a normalization method, that increases the generalizability across subjects, yet still not achieving the high accuracies obtained in the subject-based case.

When using the additional information from AU correlations we have been inspired by the synchronization theory of the CPM of emotion. AUs can be considered as individual physio-mechanical responses to *emotional* stimuli (although cognitive distraction is not an emotion but more a cognitive state) and thus their synchronization behaviour would alter in the distracted state compared to the neutral one. The experimental results show that, indeed, this correlation information is relevant to the detection task in the subject-based case, but not so much in the subject-independent case. We believe, the main reason is once again the individual differences in how one processes and expresses this cognitive state. The inclusion of head-pose and gaze behaviour could also be beneficial for the detection problem, but we have chosen not to include those in the scope of this dissertation as the main focus is the facial actions of this complex state of mind. We aim to perform this as the next step, as well as detecting the visual distraction sequences in order to fully validate the presented database.

In terms of application, this contribution has great impact potential in the driving industry, but also in other domains. For transport systems it is an important feature to be able to detect when the driver is splitting his/her workload rather than focusing solely on the driving task. For visual distraction, this is an easier task, as it involves simply recognizable head and gaze movements. For cognitive distraction, however, it is a greater challenge and with our system we able to achieve successful performance, especially for a subject-based system. The use of this system in a consumer vehicle would be to activate an alarm when a critical value for cognitive distraction is detected, over data collected within a time window. Safety is, of course, of primary importance but user studies still need to be performed to assess the acceptance of drivers, i.e. to perform this alerting operation without causing annoyance or discomfort. The proposed system also has potential to be used in other domains where knowing a subjects engagement vs. distraction level is important, e.g. HCI systems for learning applications or assessing the progress of treatment of disorders such as ADHD or Asperger's syndrome.

9.2 Outlook and Future Perspectives

It is without any doubt that the field of affect computing and automatic facial analysis will keep growing. Cloud computing and powerful mobile devices make it especially attractive for researchers to advance the techniques and build new applications. We believe the future of affective computing lies in more subject-based or user oriented systems, that can adapt themselves according to the particularities of individuals, their needs and personal usage habits.

In our contributions we have aimed to propose solutions to the subject dependency issue that require as little supervision, or manual intervention, as possible. Further work would be to perform this in a completely unsupervised manner using for example unsupervised domain adaptation methods. Another approach we would like to venture is active (or online) learning methods, which is a semi-supervised scheme that need little data labeling and more importantly does not require re-training of the system. This is an important quality in real-world applications that can achieve the adaptation-to-subject property mentioned earlier.

One approach that we have not had the chance to include in the thesis is exploiting the collective dynamics of AUs for their detection. We have shown how their prior co-occurrence information can improve AU detection accuracy, yet this contribution remained in the static case. The temporal adjacency of frames containing similar multi-label vectors is a factor that can be included in the embedding scheme. The label and temporal adjacency information can also be used directly in the classification model. Recently dynamic models, such as the CRF, are being effectively used for detecting AUs or expressions on a frame-level. We would like to investigate the advantage of a dynamic model that contains *edges* not only between labels in a single frame or between the same label across frames (in the first order sense) but also across labels and across frames with a higher order. This model would allow us to learn the temporal evolution of AUs in relation with each other across sequences.

Also related to this temporal aspect, on the next step we would like to apply space-time graph clustering methods to the facial analysis domain, the AUs forming the nodes in the *space* component. This type of clustering allow identifying temporal patterns that are specific to a test group, e.g. to differentiate between patients with MDD from healthy ones. One possible drawback of the approach is that it is highly effected by noisy observations since the graph is constructed using the binary activation of the nodes (AUs in this case). However, with the systems we have proposed in this thesis we achieve precisions that are good enough to be used to build such graphs. Graph clustering will also allow identifying in a more constructed the way the temporal patterns that are related to the classification performance, similar to what we have performed for cognitive distraction detection. Finally, this differentiation through graphs can provide the context needed to better specialize the dynamic models used for AU detection, in a *feed-back loop* manner.

Bibliography

Iso 26022:2010:: Road vehicles. ergonomic aspects of transport information and control systems. simulated lane change test to assess in-vehicle secondary task demand, 2010.

Iso/ts 14198:2012: Road vehicles – ergonomic aspects of transport information and control systems – calibration tasks for methods which assess driver demand due to the use of in-vehicle systems, 2012.

Ahmed N, Natarajan T and Rao KR. (1974). Discrete cosine transform. *Computers, IEEE Transactions on*, **100**(1), 90–93.

Almaev TR and Valstar MF. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 356–361. IEEE.

Almaev TR, Yüce A, Ghitulescu A and Valstar MF. (2013). Distribution-based iterative pairwise classification of emotions in the wild using lgbp-top. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 535–542. ACM.

Arar NM, Gao H, Ekenel HK and Akarun L. (2012). Selection and combination of local gabor classifiers for robust face verification. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, 297–302. IEEE.

Arnold MB. (1960). *Emotion and personality.* Columbia University Press.

Ashraf AB, Lucey S, Cohn JF, Chen T, Ambadar Z, Prkachin KM and Solomon PE. (2009). The painful face–pain expression recognition using active appearance models. *Image and vision computing*, **27**(12), 1788–1796.

Asthana A, Marks TK, Jones MJ, Tieu KH and Rohith M. (2011). Fully automatic pose-invariant face recognition via 3d pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 937–944. IEEE.

Bakiri S, Galéra C, Lagarde E, Laborey M, Contrand B, Ribéreau-Gayon R, Salmi LR, Gabaude C, Fort A, Maury B and others . (2013). Distraction and driving: Results from a case–control responsibility study of traffic crash injured drivers interviewed at the emergency room. *Accident Analysis & Prevention*, **59**, 588–592.

Baltrušaitis T, Mahmoud M and Robinson P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. *11th IEEE International Conference on Automatic Face and Gesture Recognitions (FG 2015), FERA 2015 Challenge*.

Bartlett MS, Littlewort GC, Frank MG, Lainscsek C, Fasel IR and Movellan JR. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, **1**(6), 22–35.

Bartlett MS, Littlewort GC, Frank MG and Lee K. (2014). Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, **24**(7), 738–743.

Bayramoglu N, Zhao G and Pietikainen M. (2013). Cs-3dlbp and geometry based person independent 3d facial action unit detection. In *Biometrics (ICB), 2013 International Conference on*, 1–6. IEEE.

Belkin M and Niyogi P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, **15**(6), 1373–1396.

Bergasa LM, Nuevo J, Sotelo M, Barea R, Lopez ME and others . (2006). Real-time system for monitoring driver vigilance. *Intelligent Transportation Systems, IEEE Transactions on*, **7**(1), 63–77.

Berthié G, Lemercier C, Paubel PV, Cour M, Fort A, Galéra C, Lagarde E, Gabaude C and Maury B. (2015). The restless mind while driving: drivers' thoughts behind the wheel. *Accident Analysis & Prevention*, **76**, 159–165.

Bishop CM. (2006). *Pattern recognition and machine learning*. Springer.

Boucenna S, Narzisi A, Tilmont E, Muratori F, Pioggia G, Cohen D and Chetouani M. (2014). Interactive technologies for autistic children: a review. *Cognitive Computation*, **6**(4), 722–740.

Breiman L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Cao C, Weng Y, Zhou S, Tong Y and Zhou K. (2014a). Facewarehouse: a 3d facial expression database for visual computing. *Visualization and Computer Graphics, IEEE Transactions on*, **20**(3), 413–425.

Cao X, Wei Y, Wen F and Sun J. (2014b). Face alignment by explicit shape regression. *International Journal of Computer Vision*, **107**(2), 177–190.

Chai D and Ngan KN. (1999). Face segmentation using skin-color map in videophone applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, **9**(4), 551–564.

Chan M and Singhal A. (2013). The emotional side of cognitive distraction: Implications for road safety. *Accident Analysis & Prevention*, **50**, 147–154.

Chang CC and Lin CJ. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3), 27.

Chew SW, Lucey P, Lucey S, Saragih J, Cohn JF, Matthews I and Sridharan S. (2012). In the pursuit of effective affective computing: The relationship between features and registration. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **42**(4), 1006–1016.

Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, Zhou F and La Torre FD. (2009). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 1–7. IEEE.

Cohn JF and Schmidt KL. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, **2**(02), 121–132.

Cooper DG, Muldner K, Arroyo I, Woolf BP and Burleson W. Ranking feature sets for emotion models used in classroom based intelligent tutoring systems. In *User Modeling, Adaptation, and Personalization*, 135–146. Springer. (2010).

Cootes TF, Edwards GJ and Taylor CJ. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 681–685.

Cootes TF, Taylor CJ, Cooper DH and Graham J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, **61**(1), 38–59.

Crespo J, Serra J and Schafer RW. (1995). Theoretical aspects of morphological filters by reconstruction. *Signal processing*, **47**(2), 201–225.

Cristinacce D and Cootes TF. (2006). Feature detection and tracking with constrained local models. In *BMVC*, volume 1, 3. Citeseer.

Cristinacce D and Cootes TF. (2007). Boosted regression active shape models. In *BMVC*, 1–10.

Dalal N and Triggs B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 886–893. IEEE.

Dalla Mura M, Benediktsson JA, Chanussot J and Bruzzone L. The evolution of the morphological profile: From panchromatic to hyperspectral images. In *Optical Remote Sensing*, 123–146. Springer. (2011).

Dantone M, Gall J, Fanelli G and Van Gool L. (2012). Real-time facial feature detection using conditional regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2578–2585. IEEE.

Darwin C. (1872). 1965. the expression of the emotions in man and animals. *London, UK: John Marry*.

Daugman JG. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, **2**(7), 1160–1169.

Ding X, Chu WS, De la Torre F, Cohn JF and Wang Q. (2013). Facial action unit event detection by cascade of tasks. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2400–2407. IEEE.

Dong Y, Hu Z, Uchimura K and Murayama N. (2011). Driver inattention monitoring system for intelligent vehicles: A review. *Intelligent Transportation Systems, IEEE Transactions on*, **12**(2), 596–614.

D'Orazio T, Leo M, Guaragnella C and Distante A. (2007). A visual approach for driver inattention detection. *Pattern Recognition*, **40**(8), 2341–2355.

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry L, McRorie M, Martin LJC, Devillers J, Abrilian A, Batliner S and others . (2007). The humaine database: addressing the needs of the affective computing community. In *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, 488–500.

Duchenne GB. (1876). *Mécanisme de la physionomie humaine: où, Analyse électrophysiologique de l'expression des passions*. J.-B. Baillière.

Ecabert C, Gao H and Thiran JP. (2015). 3d bilinear face model fitting from multiple cameras. Technical report.

Ekman P and Friesen W. (1977). A technique for the measurement of facial movement. *Consulting Psychologists Press, Stanford University, Palo Alto*.

Ekman P, Friesen WV and Hager JE. (2002). *Facial Action Coding System [E-book]*. Salt Lake City, UT: Research Nexus.

Ekman P. (1980). *Face of man: Universal expression in a New Guinea village*. Garland, New York.

Ekman P. (1984). Expression and the nature of emotion. *K. Scherer and P. Ekman (Eds.), Approaches to emotion*, **3**, 319–344.

Ekman P. (1989). The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, 143–164.

Ekman P and Friesen WV. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, **1**(1), 56–75.

Ekman P, Friesen WV and Ellsworth P. (1972). *Emotion in the Human Face: Guidelines for Research and an Integration of Findings: Guidelines for Research and an Integration of Findings*. Pergamon.

Ekman P and Keltner D. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, **8**(4), 151–158.

Ekman P and Rosenberg EL. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Engström J and Markkula G. (2007). Effects of visual and cognitive distraction on lane change test performance. In *Proceedings of the 4th international driving symposium on human factors in driver assessment, training, and vehicle design*, 199–205.

Faggian N, Paplinski A and Sherrah J. (2008). 3d morphable model fitting from multiple views. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, 1–6. IEEE.

Freund Y, Schapire RE and others . (1996). Experiments with a new boosting algorithm. In *ICML*, volume 96, 148–156.

Friedman J, Hastie T, Tibshirani R and others . (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, **28**(2), 337–407.

Friesen E and Ekman P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*.

Frijda NH. (1986). *The emotions.* Cambridge University Press.

Galera C, Orriols L, MBailara K, Laborey M, Contrand B, Ribéreau-Gayon R, Masson F, Bakiri S, Gabaude C, Fort A and others . (2012). Mind wandering and driving: responsibility case-control study. *BMJ*, **345**, e8105.

Gao H, Ekenel HK and Stiefelhagen R. Pose normalization for local appearance-based face recognition. In *Advances in Biometrics*, 32–41. Springer. (2009).

Gao H, Yuce A and Thiran JP. (2014). Detecting emotional stress from facial expressions for driving safety. In *Image Processing (ICIP), 2014 IEEE International Conference on*, 5961–5965. IEEE.

Gehrig T and Ekenel HK. (2011). Facial action unit detection using kernel partial least squares. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2092–2099. IEEE.

Gentsch K, Grandjean D and Scherer KR. (2015). Appraisals generate specific configurations of facial muscle movements in a gambling task: Evidence for the component process model of emotion. *PloS one*, **10**(8), e0135837.

Girard JM, Cohn JF and De la Torre F. (2014). Estimating smile intensity: A better way. *Pattern Recognition Letters*.

Gordon I, Pierce MD, Bartlett MS and Tanaka JW. (2014). Training facial expression production in children on the autism spectrum. *Journal of autism and developmental disorders*, **44**(10), 2486–2498.

Grafsgaard J, Wiggins JB, Boyer KE, Wiebe EN and Lester J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.

Gross R, Matthews I, Cohn J, Kanade T and Baker S. (2007). The cmu multi-pose, illumination, and expression (multi-pie) face database. Technical report, Technical report, Carnegie Mellon University Robotics Institute. TR-07-08.

Gross R, Matthews I and Baker S. (2005). Generic vs. person specific active appearance models. *Image and Vision Computing*, **23**(12), 1080–1093.

Gu H and Ji Q. (2004). Facial event classification with task oriented dynamic bayesian network. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, II–870. IEEE.

Gudi A, Tasli HE, den Uyl TM and Maroulis A. (2015). Deep learning based facs action unit occurrence and intensity estimation. *11th IEEE International Conference on Automatic Face and Gesture Recognitions (FG 2015), FERA 2015 Challenge*.

Harbluk JL, Burns PC, Lochner M and Trbovich PL. (2007). Using the lane-change test (lct) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces. In *Proceedings of the 4th international driving symposium on human factors in driver assessment, training, and vehicle design*, 16–22.

Hastie T, Tibshirani R, Friedman J and Franklin J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, **27** (2), 83–85.

Hess U and Kleck RE. (1997). Differentiating emotion elicited and deliberate emotional facial expressions. *What the face reveals*, 271–285.

Hess U and Thibault P. (2009). Darwin and emotion expression. *American Psychologist*, **64**(2), 120.

Hjelmås E and Low BK. (2001). Face detection: A survey. *Computer vision and image understanding*, **83**(3), 236–274.

Huang D, Shan C, Ardabilian M, Wang Y and Chen L. (2011). Local binary patterns and its application to facial image analysis: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, **41**(6), 765–781.

Hwang W, Huang X, Noh K and Kim J. (2011). Face recognition system using extended curvature gabor classifier bunch for low-resolution face image. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 15–22. IEEE.

Izard CE. (1969). The emotions and emotion constructs in personality and culture research. *Handbook of modern personality theory*, 496–510.

Jabon ME, Bailenson JN, Pontikakis E, Takayama L and Nass C. (2010). Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Computing*, (4), 84–95.

Jack RE, Garrod OG, Yu H, Caldara R and Schyns PG. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, **109**(19), 7241–7244.

Jain AK and Li SZ. (2005). *Handbook of face recognition*, volume 1. Springer.

Jimenez P, Bergasa LM, Nuevo J, Hernandez N and Daza IG. (2012). Gaze fixation system for the evaluation of driver distractions induced by ivis. *Intelligent Transportation Systems, IEEE Transactions on*, **13**(3), 1167–1178.

Kächele M, Schels M and Schwenker F. (2014). Inferring depression and affect from application dependent meta knowledge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 41–48. ACM.

Kaiser S and Wehrle T. (2001). Facial expressions as indicators of appraisal processes. *Appraisal processes in emotion: Theory, methods, research*, 285–300.

Kanade T, Cohn JF and Tian Y. (2000). Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 46–53. IEEE.

Kang HB. (2013). Various approaches for driver and driving behavior monitoring: a review. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, 616–623. IEEE.

Kring A and Stuart BK. (2005). Nonverbal behavior and psychopathology. *The new handbook of methods in nonverbal behavior research*, 313–339.

Kroupi E, Vesin JM and Ebrahimi T. (2013). Phase-amplitude coupling between eeg and eda while experiencing multimedia content. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 865–870. IEEE.

Kutila M, Jokela M, Mäkinen T, Viitanen J, Markkula G and Victor T. (2007). Driver cognitive distraction detection: feature estimation and implementation. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, **221**(9), 1027–1040.

Lazarus R. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Lee JD, Young KL and Regan MA. (2008). Defining driver distraction. *Driver distraction: Theory, effects, and mitigation*, 31–40.

Lee TS. (1996). Image representation using 2d gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **18**(10), 959–971.

Lemercier C, Pêcher C, Berthié G, Valéry B, Vidal V, Paubel PV, Cour M, Fort A, Galéra C, Gabaude C and others . (2014). Inattention behind the wheel: How factual internal thoughts impact attentional control while driving. *Safety science*, **62**, 279–285.

Li J and Allinson NM. (2008). A comprehensive review of current local features for computer vision. *Neurocomputing*, **71**(10), 1771–1787.

Li N and Busso C. (2015). Predicting perceived visual and cognitive distractions of drivers with multimodal features. *Intelligent Transportation Systems, IEEE Transactions on*, **16**(1), 51–65.

Li SZ, Lu X, Hou X, Peng X and Cheng Q. (2005). Learning multiview face subspaces and facial pose estimation using independent component analysis. *Image Processing, IEEE Transactions on*, **14**(6), 705–712.

Li Y, Chen J, Gao W and Yin B. (2004). Face detection: a survey.

Liang Y, Lee J and Reyes M. (2007). Nonintrusive detection of driver cognitive distraction in real time using bayesian networks. *Transportation Research Record: Journal of the Transportation Research Board*, (2018), 1–8.

Lien JJ, Kanade T, Cohn JF and Li CC. (1998). Automated facial expression recognition based on facs action units. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 390–395. IEEE.

Lowe DG. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**(2), 91–110.

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

Lucey P, Cohn JF, Prkachin KM, Solomon PE and Matthews I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 57–64. IEEE.

Lundqvist D, Flykt A and Öhman A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91–630.

Lyons M, Akamatsu S, Kamachi M and Gyoba J. (1998). Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 200–205. IEEE.

Mahoor MH, Cadavid S, Messinger DS and Cohn JF. (2009). A framework for automated measurement of the intensity of non-posed facial action units. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, 74–80. IEEE.

Martinez AM. (1998). The ar face database. *CVC Technical Report*, **24**.

Mattes S. (2003). The lane-change-task as a tool for driver distraction evaluation. *Quality of work and products in enterprises of the future*, 57–60.

Mavadati SM, Mahoor MH, Bartlett K, Trinh P and Cohn JF. (2013). Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, **4** (2), 151–160.

McDuff D, Kaliouby RE, Cohn JF and Picard R. (2014). Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *Affective Computing, IEEE Transactions on*.

McKeown G, Valstar M, Cowie R, Pantic M and Schröder M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, **3** (1), 5–17.

Morris JD. (1995). Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, **35**(6), 63–68.

Mowrer O. (1960). Learning theory and behavior.

Murphy-Chutorian E and Trivedi MM. (2010). Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *Intelligent Transportation Systems, IEEE Transactions on*, **11**(2), 300–311.

Neale VL, Dingus TA, Klauer SG, Sudweeks J and Goodman M. (2005). An overview of the 100-car naturalistic study and findings. *National Highway Traffic Safety Administration, Paper*, (05-0400).

Ojala T, Pietikäinen M and Harwood D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, **29**(1), 51–59.

Ojala T, Pietikäinen M and Mäenpää T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **24**(7), 971–987.

Ojansivu V and Heikkilä J. Blur insensitive texture classification using local phase quantization. In *Image and signal processing*, 236–243. Springer. (2008).

Olson RL, Hanowski RJ, Hickman JS and Bocanegra JL. (2009). Driver distraction in commercial vehicle operations. Technical report.

Ortony A, Clore GL and Collins A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Ortony A and Turner TJ. (1990). What's basic about basic emotions? *Psychological review*, **97**(3), 315.

Osadchy M, Cun YL and Miller ML. (2007). Synergistic face detection and pose estimation with energy-based models. *The Journal of Machine Learning Research*, **8**, 1197–1215.

Pantic M, Cowie R, DErrico F, Heylen D, Mehu M, Pelachaud C, Poggi I, Schroeder M and Vinciarelli A. Social signal processing: the research agenda. In *Visual analysis of humans*, 511–538. Springer. (2011).

Pantic M and Patras I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **36**(2), 433–449.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V and others . (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, **12**, 2825–2830.

Pentland AS. (2005). Socially aware, computation and communication. *Computer*, **38** (3), 33–40.

Peters G, Krüger N and Von Der Malsburg C. (1997). Learning object representations by clustering banana wavelet responses. *Proceedings of the 1st STIPR*, 113–118.

Pettitt M, Burnett GE and Stevens A. (2005). Defining driver distraction. In *12th World Congress on Intelligent Transport Systems*.

Picard RW. (1997). *Affective computing*, volume 252. MIT press Cambridge.

Picard RW. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1535), 3575–3584.

Plutchik R. (1980). *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division.

Ragab A, Craye C, Kamel MS and Karray F. A visual-based driver distraction recognition and detection using random forest. In *Image Analysis and Recognition*, 256–265. Springer. (2014).

Ramnath K, Koterba S, Xiao J, Hu C, Matthews I, Baker S, Cohn J and Kanade T. (2008). Multi-view aam fitting and construction. *International Journal of Computer Vision*, **76**(2), 183–204.

Regan MA, Hallett C and Gordon CP. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis & Prevention*, **43**(5), 1771–1781.

Richter M, Gao H and Ekenel HK. (2014). Extending explicit shape regression with mixed feature channels and pose priors. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 1013–1019. IEEE.

Ringeval F, Eyben F, Kroupi E, Yuce A, Thiran JP, Ebrahimi T, Lalanne D and Schuller B. (2014). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*.

Ringeval F, Sonderegger A, Sauer J and Lalanne D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 1–8. IEEE.

Romdhani S, Gong S, Psarrou A and others . (1999). A multi-view nonlinear active shape model using kernel pca. In *BMVC*, volume 10, 483–492.

Rongben W, Lie G, Bingliang T and Lisheng J. (2004). Monitoring mouth movement for driver fatigue or distraction with one camera. In *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*, 314–319. IEEE.

Russell JA. (1980). A circumplex model of affect. *Journal of personality and social psychology*, **39**(6), 1161.

Sander D, Grandjean D and Scherer KR. (2005). A systems approach to appraisal mechanisms in emotion. *Neural networks*, **18**(4), 317–352.

Sander D and Scherer K. (2009). *Oxford companion to emotion and the affective sciences*. Oxford University Press.

Saragih JM, Lucey S and Cohn JF. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, **91**(2), 200–215.

Savran A, Alyüz N, Dibeklioğlu H, Çeliktutan O, Gökberk B, Sankur B and Akarun L. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, 47–56. Springer. (2008).

Schapire RE and Singer Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, **37**(3), 297–336.

Scherer KR. (1987). Toward a dynamic theory of emotion. *Geneva studies in Emotion*, **1**, 1–96.

Scherer KR. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, **7**(3-4), 325–355.

Scherer KR. (1999). On the sequential nature of appraisal processes: Indirect evidence from a recognition task. *Cognition & Emotion*, **13**(6), 763–793.

Scherer KR. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, **92**, 120.

Scherer KR. (2005). What are emotions? and how can they be measured? *Social science information*, **44**(4), 695–729.

Schölkopf B and Smola AJ. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Scovanner P, Ali S and Shah M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, 357–360. ACM.

Senechal T, Bailly K and Prevost L. (2010). Automatic facial action detection using histogram variation between emotional states. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, 3752–3755. IEEE.

Senechal T, Rapp V, Salam H, Seguier R, Bailly K and Prevost L. (2011). Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 860–865. IEEE.

Shan C, Gong S and McOwan PW. (2006). A comprehensive empirical study on linear subspace methods for facial expression analysis. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, 153–153. IEEE.

Shan C, Gong S and McOwan PW. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, **27**(6), 803–816.

Shen L and Bai L. (2006). A review on gabor wavelets for face recognition. *Pattern analysis and applications*, **9**(2-3), 273–292.

Sikka K, Ahmed AA, Diaz D, Goodwin MS, Craig KD, Bartlett MS and Huang JS. (2015). Automated assessment of childrenś postoperative pain using computer vision. *Pediatrics*, **136**(1), e124–e131.

Smola AJ and Schölkopf B. (2004). A tutorial on support vector regression. *Statistics and computing*, **14**(3), 199–222.

Sobotta J and Figge FHJ. (1974). *Atlas of human anatomy, Vol. 1*. Hafner Press.

Sorci M. (2009). *Automatic face analysis in static and dynamic environments*. PhD thesis, École Polytechnique Fédérale de Lausanne.

Strayer DL, Watson JM and Drews FA. (2011). Cognitive distraction while multitasking in the automobile. *Psychology of Learning and Motivation-Advances in Research and Theory*, **54**, 29.

Stutts JC, Association AA and others . (2001). *The role of driver distraction in traffic crashes*. AAA Foundation for Traffic Safety Washington, DC.

Taigman Y, Yang M, Ranzato M and Wolf L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 1701–1708. IEEE.

Tango F, Botta M, Minin L and Montanari R. (2010). Non-intrusive detection of driver distraction using machine learning algorithms. In *ECAI*, 157–162.

Texeira T, Wedel M and Pieters R. (2012). Emotion-induced engagement in internet video ads. *Journal of Marketing Research*, **49**(2), 144–159.

Tomasi C and Manduchi R. (1998). Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, 839–846. IEEE.

Tong Y, Liao W and Ji Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29**(10), 1683–1699.

Turk M and Pentland A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, **3**(1), 71–86.

Turner ML and Engle RW. (1989). Is working memory capacity task dependent? *Journal of memory and language*, **28**(2), 127–154.

Valstar M, Girard J, Almaev T, McKeown G, Mehu M, Yin L, Pantic M and Cohn J. (2015). Fera 2015-second facial expression recognition and analysis challenge. In *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*.

Valstar M, Martinez B, Binefa X and Pantic M. (2010). Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2729–2736. IEEE.

Valstar M and Pantic M. (2006). Fully automatic facial action unit detection and temporal analysis. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, 149–149. IEEE.

Valstar M and Pantic M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 65.

Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, Cowie R and Pantic M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 3–10. ACM.

Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, Schnieder S, Cowie R and Pantic M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 3–10. ACM.

Valstar MF, Gunes H and Pantic M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, 38–45. ACM.

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

Valstar MF, Mehu M, Jiang B, Pantic M and Scherer K. (2012). Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **42**(4), 966–979.

Valstar MF and Pantic M. (2012). Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **42**(1), 28–43.

Van Reekum CM. (2000). *Levels of processing in appraisal: Evidence from computer game generated emotions*. PhD thesis, University of Geneva.

Victor MTW, Lee JD and Regan MA. (2013). *Driver Distraction and Inattention: Advances in Research and Countermeasures*, volume 1. Ashgate Publishing, Ltd.

Viola P and Jones MJ. (2004). Robust real-time face detection. *International journal of computer vision*, **57**(2), 137–154.

Vural E, Cetin M, Ercil A, Littlewort G, Bartlett M and Movellan J. Drowsy driver detection through facial movement analysis. In *Human–Computer Interaction*, 6–18. Springer. (2007).

Wang H, Huang H and Ding CH. (2010a). Discriminant laplacian embedding. In *AAAI*.

Wang H and Chang SF. (1997). A highly efficient system for automatic face region detection in mpeg video. *Circuits and Systems for Video Technology, IEEE Transactions on*, **7**(4), 615–628.

Wang N, Gao X, Tao D and Li X. (2014). Facial feature point detection: A comprehensive survey. *arXiv preprint arXiv:1410.1037*.

Wang S, Liu Z, Lv S, Lv Y, Wu G, Peng P, Chen F and Wang X. (2010b). A natural visible and infrared facial expression database for expression recognition and emotion inference. *Multimedia, IEEE Transactions on*, **12**(7), 682–691.

Weiner B and Graham S. (1984). An attributional approach to emotional development. *Emotions, cognition, and behavior*, 167–191.

Whitehill J, Bartlett M and Movellan J. (2008). Automatic facial expression recognition for intelligent tutoring systems. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, 1–6. IEEE.

Wilcoxon F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 80–83.

Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G and Mehta DD. (2014). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 65–72. ACM.

Williamson JR, Quatieri TF, Helfer BS, Horwitz R, Yu B and Mehta DD. (2013). Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 41–48. ACM.

Wöllmer M, Blaschke C, Schindl T, Schuller B, Färber B, Mayer S and Trefflich B. (2011). Online driver distraction detection using long short-term memory. *Intelligent Transportation Systems, IEEE Transactions on*, **12**(2), 574–582.

Xiao J, Baker S, Matthews I and Kanade T. (2004). Real-time combined 2d+ 3d active appearance models. In *CVPR (2)*, 535–542.

Xiong X and De la Torre F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 532–539. IEEE.

Young KL and Salmon PM. (2012). Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods. *Safety science*, **50**(2), 165–174.

Yüce A, Sorci M and Thiran JP. (2013a). Improved local binary pattern based action unit detection using morphological and bilateral filters. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 1–7.

Yüce A, Arar NM and Thiran JP. Multiple local curvature gabor binary patterns for facial action recognition. In *Human Behavior Understanding*, 136–147. Springer. (2013b).

Yüce A, Gao H, Cuendet G and Thiran JP. (Under Review). Action units and their cross-correlations for detection of cognitive distraction during driving.

Yüce A, Gao H and Thiran JP. (2015). Discriminant multi-label manifold embedding for facial action unit detection. In *11th IEEE International Conference on Automatic Face and Gesture Recognitions (FG 2015), FERA 2015 Challenge*.

Yüce A, Sorci M and Thiran JP. (2011). Head pose detection using fast robust pca for side active appearance models under occlusion. In *Proceeding of the The 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV 2011)*.

Zhang C and Zhang Z. (2010). A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research.

Zhang W, Shan S, Gao W, Chen X and Zhang H. (2005). Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, 786–791. IEEE.

Zhang X, Yin L, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P and Girard JM. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, **32**(10), 692–706.

Zhao G and Pietikainen M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29**(6), 915–928.

Zhou H, Yuan Y and Shi C. (2009). Object tracking using sift features and mean shift. *Computer vision and image understanding*, **113**(3), 345–352.

Zhu Y, De la Torre F, Cohn JF and Zhang YJ. (2011). Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *Affective Computing, IEEE Transactions on*, **2**(2), 79–91.

ANIL YÜCE

Personal Information

Email: anilyuce@gmail.com
Phone: +41 78 693 38 68
Address: Avenue du Léman 77, CH-1005 Lausanne, Switzerland
Date of Birth: July 10, 1987
Nationality: Turkish

Education

June 2010 - present **Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
PhD in Dynamic Facial Expression Analysis for Emotion Recognition in Video,
Supervisor: Prof. Jean-Philippe Thiran, Co-supervisor: Prof. David Sander(UniGE)
Enrolled in Doctoral School of Electrical Engineering

2010 **Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
MSc in Electrical and Electronics Engineering, Information Technology Track, GPA: 5.74/6.00 (Ranked 1st in department)
MSc Thesis entitled "Combining frontal and side AAMs for robust face tracking"
Supervisors: Prof. Jean-Philippe Thiran, Dr. Matteo Sorci
MSc Semester Project entitled "Bilateral Spline filters and SURE based Bilateral Filters"
Supervisor: Prof. Michael Unser

2008 **Middle East Technical University (METU), Ankara, Turkey**
BSc in Electrical and Electronics Engineering, GPA: 3.60/4.00

Work Experience

June 2010 – present Teaching and Research Assistant at *EPFL Signal Processing Laboratory (LTS5)*

- *Facial Action Unit and Expression Detection in Videos*
  - Research on improving detection accuracy and exploiting dynamics of facial actions
  - Implementation of a state-of-the-art real-time detection system (C++)
  - Analysis of existing databases for various applications including multimodal emotion detection (Funded by SNF project IM2.IP1)
- *Dynamic Facial Expression Analysis for Driver Behaviour Analysis Applications (Funded by joint project with PSA Peugot-Citroën and Valeo)*
  - Recorded multiple databases with posed and spontaneous facial expressions in driving conditions (including hardware assessment and setup)
  - Performed analysis of facial expressions / action units for applications such as stress detection, visual and cognitive distraction, spontaneous emotion detection in realistic driving conditions (Using C++, Matlab and Python with libraries such as OpenCV, NumPy, Scipy)
- *Automatic detection of difficult intubation using facial analysis (side project)*
  - Involved in recording and analyzing images and videos of patients before surgery
  - Built an acquisition device for CHUV, a user interface in Matlab and an interface for handling the database using Html, Php and mySql
- *Teaching Activities*
  - Supervision of 4 semester and MSc diploma projects
  - Teaching Assistant in master level Pattern Recognition and Image Analysis, Image Processing and Advanced Digital Communication courses

July 2013 – Oct. 2013 Research Intern at the Mixed Reality Lab, University of Nottingham, U.K.
Conducted research on analysis of dynamics of facial action units

Feb. 2010 – June 2010 Intern at *EPFL Signal Processing Laboratory (LTS5), Switzerland*
- Resarch on face tracking across pose using multiple active appearance models and a fast-robust PCA (Matlab and C++)

Summer 2007 Intern at *TEPA Inc. , Ankara, Turkey*
- Design of a test protocol for consumer ECG devices

Summer 2006 Intern at *ASELSAN Inc*.- Radar department *, Ankara, Turkey*
User interface design in Matlab for Radar calculations

Awards and Honors

FERA 2015 Challenge, AU Occurrence Detection Sub-challenge Winner, AFGR 2015
IET Anna Barbara Reinhard Prize for Student Excellence, 2010
EPFL Excellence Scholarship, 2008-2010
METU High Honor Student (4 times), Honor Student (4 times) , 2004 - 2008
T.E.D. Prof. Dr. Rüştü Yüce Award for Academic Success, 2004
Turkish University Entrance Exam ranked 186. (among 1.8 million people), 2004

Computer Skills

Programming Languages: C/C++, Matlab, Java, Python, PHP, Html, mySql
Software: Adobe Photoshop, Gimp, ImageJ, VirtualDub, Microsoft Office

Language Skills

English - Fluent (spoken and written - TOEFL score: 116/120)
French - Fluent (spoken and written - B2 level certificate)
Italian - Intermediate
German - Basic
Turkish – Native

List of Publications

Journal Articles

Yüce A, Gao H., Cuendet G.L. and Thiran J.P., "Action units and their cross correlations for detection of cognitive distraction during driving", (Under Review).

Cuendet G.L., Schoettker P., Yüce A., Sorci M., Gao H., Perruchoud C. and Thiran J.P., "Facial Image Analysis for Fully-Automatic Prediction of Difficult Endotracheal Intubation", accepted in IEEE Transactions on Biomedical Engineering (BME), 2015.

Ringeval F., Eyben F., Kroupi E., Yüce A., Thiran J.P., Ebrahimi T., Lalanne D. and Schuller B., "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data", accepted in Pattern Recognition Letters, 2014.

Conference Articles

Yüce A., Gao H. and Thiran J.P., "Discriminant multi-label manifold embedding for facial action unit detection", In Proceedings of 11th IEEE International Conference on Automatic Face and Gesture Recognitions (FG 2015), FERA 2015 Challenge, Ljubljana, Slovenia, 2015.

Gao H., Yüce A. and Thiran J.P., "Detecting emotional stress from facial expressions for driving safety", In Image Processing (ICIP), 2014 IEEE International Conference on, pp 5961–5965, Paris, France, 2014.

Almaev T.R., Yüce A., Ghitulescu A. and Valstar M.F., "Distribution-based iterative pairwise classification of emotions in the wild using lgbp-top", In Proceedings of the 15th ACM International conference on multimodal interaction (ICMI), pp 535–542, Sydney, Australia, 2013.

Yüce A., Arar N.M. and Thiran J.P., "Multiple local curvature gabor binary patterns

for facial action recognition", In Proceedings of 4th International Workshop on Human Behavior Understanding (HBU), in conjunction with ACM Multimedia (ACM MM), pp 136–147, Barcelona, Spain, 2013.

Yüce A., Sorci M. and Thiran J.P., "Improved local binary pattern based action unit detection using morphological and bilateral filters", In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pp 1–7, Shanghai, China, 2013.

Cuendet G.L., Yüce A., Sorci M., Schoettker P., Perruchoud C. and Thiran J.P., "Automatic Mallampati Classification Using Active Appearance Models", International Workshop on Pattern Recognition for Healthcare Analytics, in conjunction with International Conference on Pattern Recognition (ICPR), Tsukuba Science City, Japan, 2012.

Yüce A., Sorci M. and Thiran J.P., "Head pose detection using fast robust pca for side active appearance models under occlusion", In Proceeding of the The 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV 2011), Las Vegas, NV, 2011.

Book Chapters

Thiran J.P., Yüce A., Sorci M., "Facial expression analysis for emotion recognition and perception modeling", in Interactive MultiModal Information Management, EPFL Press, 2014.