CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY IN PRAGUE

MASTER'S THESIS

Emotive Facial Expression Detection

Petr Husák

husakpe1@fel.cvut.cz

CTU–CMP–2017–01

January 9, 2017

Thesis Advisor: Jan Čech

The research was supported by CTU student grant
SGS15/155/OHK3/2T/13.

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Cybernetics

DIPLOMA THESIS ASSIGNMENT

Student:	Bc. Petr Husák
Study programme:	Open Informatics
Specialisation:	Computer Vision and Image Processing
Title of Diploma Thesis:	Emotive Facial Expression Detection

Guidelines:

1. Familiarize yourself with the literature on facial expressions in both:
   a. Psychology, in order to understand types and mechanism of facial expressions. Concentrate on micro-expressions, fast facial expressions appearing involuntarily without a conscious control of a subject.
   b. Computer vision literature, in order to get familiar with various approaches to facial expressions detection/recognition developed in recent years.
2. Collect an annotated dataset for training and testing. (Download published datasets or prepare your own datasets by labeling movies, or other TV content, or acquire a dataset in a controlled laboratory experiment).
3. Propose an algorithm to detect non-neutral facial expressions, preferably the micro-expressions, from videos. Consider to classify instantaneous basic facial expressions types.
4. Evaluate the proposed algorithms and compare quantitatively on a dataset with the ground-truth annotation.

Bibliography/Sources:

[1] Paul Ekman, Wallace F. Friesen. Unmasking The Face: A Guide to Recognize Emotions from Facial Clues. Mallor Books, Los Altos, CA, 2003.

[2] A. Dhall, R. Goecke, S. Lucey and T. Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. IEEE MultiMedia 19 (2012) 34-41.

[3] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao and M. Pietikäinen. Reading Hidden Emotions: Spontaneous Micro-expression Spotting and Recognition, in arXiv 1511.00423, 2015.

[4] Jan Čech, Vojtěch Franc, Michal Uřičář, Jiří Matas. Multi-view facial landmark detection by using a 3D shape model. Image and Vision Computing 47, pp. 60-70, March 2016.

Diploma Thesis Supervisor: Ing. Jan Čech, Ph.D.

Valid until: the end of the winter semester of academic year 2017/2018

L.S.

prof. Dr. Ing. Jan Kybic **Head of Department**	prof. Ing. Pavel Ripka, CSc. **Dean**

Prague, May 27, 2016

Acknowledgement

I would like to express my gratitude to my advisor Ing. Jan Čech, Ph.D. for his valuable advice, guidance and willingness during my study. Furthermore I would like to thank prof. Ing. Jiří Matas, Ph.D. for his advice. My thanks also go to my family for their support.

Author statement

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university thesis.

Prague, ...

signature

Abstract

Facial expressions are important cues to observe human emotions. The thesis studies the involuntary type of facial expressions called micro-expressions. Micro-expressions are quick facial motions, appearing in high stake and stressful situations, typically when a subject tries to hide his or her emotions. Two attributes are present - fast duration and low intensity. A simple detection method is proposed, which determines instants of micro-expressions in a video. The method is based on analyzing image intensity differences over a registered face sequence. The specific pattern is detected by an SVM classifier. The results are evaluated on standard micro-expression datasets SMIC and CASMEII. The proposed method outperformed the competing method in detection accuracy. Further, we collected a new real-world micro-expression dataset consisting mostly of poker game videos downloaded from YouTube. We achieved average cross-validation AUC 0.88 for the SMIC, and 0.81 on the new challenging “in the Wild” database called MEVIEW.

Keywords: face, expressions, micro-expressions, emotions, computer vision, image processing, SVM

Abstrakt

Výrazy v lidské tváři jsou důležitým klíčem k rozpoznání lidských emocí. Tato práce se zabývá typem obličejových výrazů, které vznikají spontánně, bez vědomí člověka. Nazývají se mikro-výrazy a jsou typické svou rychlostí. Objevují se ve vyhrocených a stresových situacích, kdy osoba je nucena své emoce skrývat. Obvyklé rysy mikro-výrazů jsou rychlý průběh a nízká intenzita změny. Navrhujeme metodu, která detekuje mikro-výrazy ve videu. Metoda je založena na analýze rozdílu intenzit v registrovaném obrazu obličeje. Specifický vzor mikro-výrazu je detekován SVM klasifikátorem. Výsledky jsou vyhodnoceny na standardních databázích (CASME2 a SMIC). Navržená metoda je lepší než konkurenční metoda v detekci mikro-výrazů. Dále jsme shromáždili reálné příklady, a to především z pokerových turnajů na YouTube. Dosáhli jsme průměrné cross-validační AUC 0,88 pro databázi SMIC a 0,81 pro naší novou databázi MEVIEW.

Klíčová slova: obličej, výrazy, mikro-výrazy, emoce, počítačové vidění, zpracování digitálního obrazu, SVM

Contents

Abbreviations

AU	Action unit
FACS	Facial action coding system
ME	Micro-expression
MEVIEW	Micro-expressions videos “in the Wild”
ROI	Region of interest
SVM	Support vector machines

1 Introduction

The thesis studies the detection of spontaneous facial expressions appearing in stressful or emotional situations when a subject is attempting to conceal his/her true emotions. Such facial expressions are called micro-expressions. Their appearance features are similar to usual facial expressions. However, micro-expressions are distinct in time periods and quick changes on one or more particular locations in the face. Those micro-expression properties make the detection and recognition problem difficult for inexperienced people.

A motivation of the micro-expression detection in computer vision is an automatic recognition of suspicious behavior. A large variety of disciplines may benefit from revealing the phenomenon, e.g. security services, psychologists, teachers, etc.

We propose a pipeline to detect facial micro-expressions based on intensity difference analysis of consecutive registered face images. Specific facial regions are tracked using facial feature points. The micro-expression phenomenon results in a characteristic change in the difference signal. Two methods are proposed. The baseline method stands on thresholding the highest intensity change in the most emotive frame and the second proposed method uses SVM classifier trained on particular micro-expression patterns from the signal.

All published databases are filmed under stable laboratory conditions. As micro-expressions are difficult to evoke, collecting enough representative data is challenging especially from real-world situations. Therefore, we searched for high-stake or emotional videos on YouTube and annotated micro-expressions in poker games and TV interviews, where the underlying assumptions are satisfied. The occurrence of MEs was annotated in the videos. We call the dataset MEVIEW – Micro-Expressions VIdEos in the Wild.

The thesis is structured as follows. Chapter 1 introduces the facial expressions from a physiological and psychological points of view. Chapter 2 describes micro-expressions. Chapter 3 reviews the micro-expression detection and recognition in literature. Chapter 4 presents the proposed pipeline for the micro-expression detection. Chapter 5 details the implementation. Chapter 6 summarizes the published databases and describes the novel database MEVIEW. Chapter 7 evaluates the performance of the proposed pipeline and finally Chapter 8 concludes the thesis.

1.1 Facial Expressions

The human face is a complex structure deserving a lot of attention. The muscle structure and cooperation makes a face the most emotional part of the human body. It is hard to understand the person's inner state of mind without seeing his/her face. Therefore, facial expressions play an irreplaceable part of non-verbal communication.

The Universality Hypothesis claims that perceiving and emitting facial expressions as emotions are identical regardless the person's origin or cultural background. The original work studying facial expressions and their consequences is written by Charles Darwin [5]. Darwin claimed that facial expressions are innate, i.e. cannot be learned and have an evolutionary meaning for survival. Paul Ekman made an observational study to find out whether humans, worldwide, have a similar appearance of emotions in the face and discovered a certain level of universality [16, 15]. Moreover, he studied an isolated tribe in New Guinea and observed the same signs of facial expressions as reported about civilized people. Additionally, showing them "our" universal expressions, they were able to recognize the emotions as well.

Psychology distinguishes feelings and emotions. While feelings can last for hours, the duration of emotions is limited up to 5s. The emotions are further distinguished to voluntary and spontaneous expressions. Voluntary expressions are all under the subject control. They are easily faked and usually last for longer time. On the other hand, spontaneous expressions are brief.

1.1.1 Physiology of facial expressions

From the physiological point of view, the facial expression is a consequence of facial muscle activity. Those muscles are also called mimetic muscles or muscles of facial expressions. They are part of the group of head muscles, which additionally contain muscles of the scalp, muscles of mastication responsible for moving the jaw, and the tongue. Facial muscles are innervated by the Facial nerve, that branches in the face and its activation causes contractions, which results is various observable movements. The usually visible muscle actions are blocks of skin motion, e.g. eyebrows, lips, cheek, and wrinkles, e.g. on the forehead, between eyebrows, or on the nose.

In more detail, the human face consists of about 20 flat skeletal muscles [1], depicted in Fig. 1.1. The muscles are located under the skin, attached to the skull bone and inserting the facial skin, but not the bones or joints as other muscles responsible for body movements do. The muscles are located near the facial orifices, i.e. mouth, nose, and eyes [18]. However, unlike other facial muscles, they do not move with joints and bones, but mainly with the skin. Consequently, they cause facial surface deformations, which result in a variable facial expression representing emotions [18, 26]. Note that the muscles move in groups rather than alone and control the orifices. According to the location, the taxonomy is partitioned into three groups: oral, nasal and orbital [1].

The oral muscles alter the shape of the oral orifice. This group is responsible for complex mouth motions and allows sophisticated shaping of the mouth, e.g. encircle the mouth, control

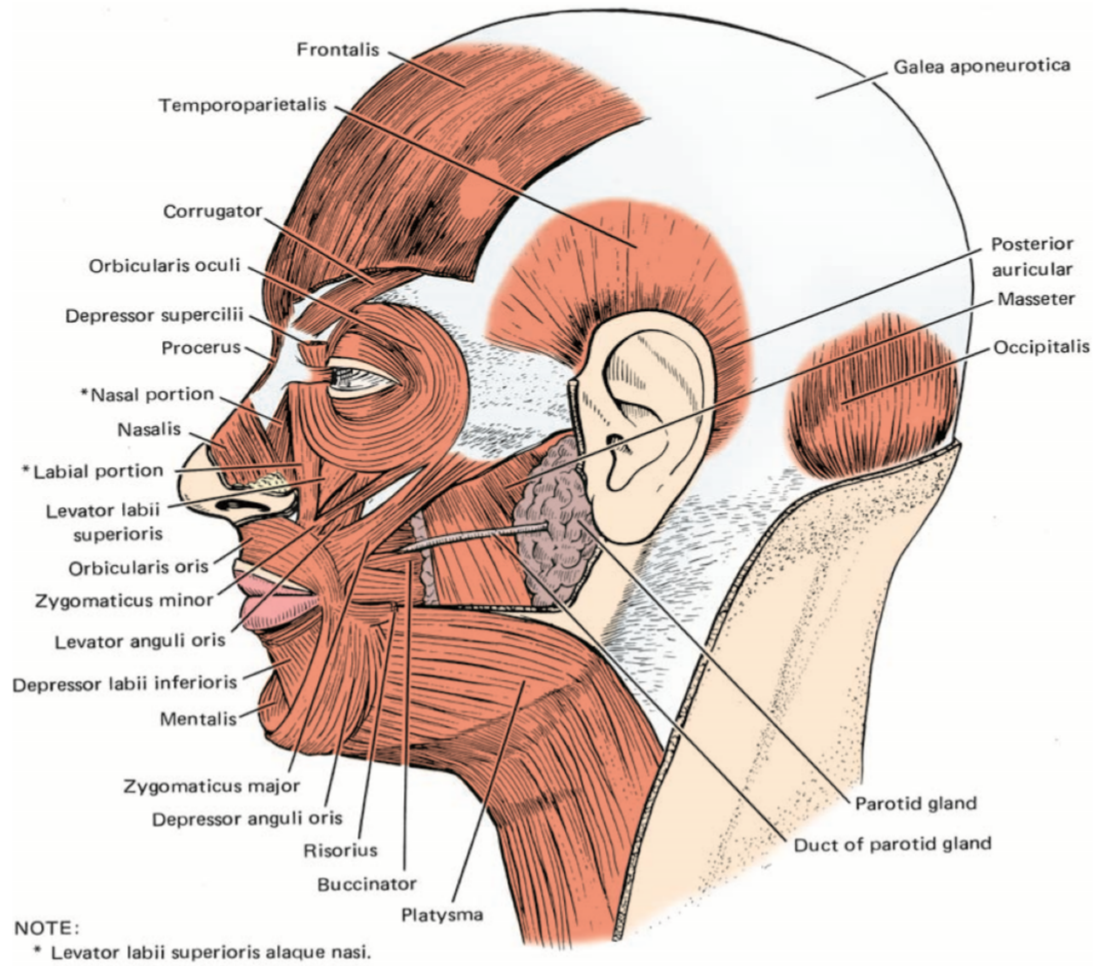

Figure 1.1 Muscles of facial expressions, taken from [18].

angle of the mouth, elevate or depress the lower and upper lip separately or lift and depress the left and right corner or move with cheeks [1, 3, 18].

The nasal group is responsible for the compression and opening of the nostrils. One of the muscles, critical for facial expressions, is located between the eyebrows and pulls the eyebrows downwards and causes wrinkles over the nose [1, 3, 18].

The orbital group of three muscles is primarily responsible for the motion of the eyelid and protecting eyes. All the muscles are inserting the skin around the eyebrows and form vertical wrinkles between the eyebrows [1, 3, 18].

An interesting educational tool showing the particular muscles moves together with emotions can be found in [4]. The tool familiarizes the user with the anatomical and biomechanical foundation of facial expression morphology and offers lots of interactive examples showing the muscle cooperation during the facial expression.

Facial Action Coding System

The Facial Action Coding System (FACS), developed by Ekman et al. [13], is a tool describing every facial muscle activity motion by a set of Action Units (AUs). A particular AU represents

Table 1.1 Description of several AUs together with the muscle name [13].

AU	Description	Facial muscle
1	Inner Brow Raiser	Frontalis
2	Outer Brow Raiser	Frontalis
6	Cheek Raiser	Orbicularis oculi
7	Lid Tightener	Orbicularis oculi
9	Nose Wrinkler	Levator labii superioris
10	Upper Lip Raiser	Levator labii superioris
11	Nasolabial Deepener	Zygomaticus minor
12	Lip Corner Puller	Zygomaticus major
13	Cheek Puffer	Levator anguli oris
14	Dimpler	Buccinator
15	Lip Corner Depressor	Depressor anguli oris
16	Lower Lip Depressor	Depressor labii inferioris
17	Chin Raiser	Mentalis
20	Lip stretcher	Risorius w/ platysma

Table 1.2 Basic AUs connected with emotions [21].

Emotion	AU
Anger	4,5,7
Disgust	9,15,16
Fear	1,2,7,20
Happiness	6,12
Sadness	1,15
Surprise	1,2

a certain component of facial muscles movement, see Tab. 1.1. Each emotion in the face can be described by a set of AUs, see Fig. 1.2 .

1.2 Psychology of emotions - basic categories

Ekman describes the six basic emotions [14]. The following section reviews the emotions and concludes their important features. The basic AUs for a specific emotion are listed in Tab. 1.2. Mixing basic expressions creates more different expressions, e.g. angrily surprised.

1.2.1 Anger

Anger is a strong emotional reaction and can also be a dangerous emotion as it might provoke violence. The source of anger has many reasons. Considering the frustration, one can feel anger against an obstacle on the way to success. A different source of anger is a physical threat. When someone wants to hurt us, we naturally feel anger. As well as physical violence, verbal threats or claims also cause the emotion of anger. Other examples are false accusations in uncommitted crimes or someone breaking our personal values.

Anger has a substantial impact on the whole body. The increase of the blood pressure, red face, and tension in the muscles are usually reflected. The physiological response induces the following signals.

- The eyebrows are lowered and squeezing together. There are vertical wrinkles between the eyebrows.
- The eyelids are tight and straight.
- The eyes are tight, focusing on the source of anger. Pupils are narrowed, focusing on the source of anger.
- The lips are either closed tight or gently opened (preparing for yelling).

Together, those characteristics apparently prepare the body for a possible physical or verbal attack.

1.2.2 Disgust

Disgust is a negative emotion usually evoked by smell, taste or vision. Unlike the other emotions, objects evoking disgust are not universal, but cultural or personal, e.g. food. The extreme physiological response is vomiting. The most significant features of the face are in the mouth and nose area.

- The upper lip is lifted.
- There are wrinkles on the nose.
- The cheeks are lifted.
- The eyelids are lifted but are not tight. There are wrinkles under the eyes.
- The eyebrows are pulled down.

1.2.3 Fear

Fear is induced by dangerous or stressful situations. One can feel fear from future events, e.g. fear of loosing money. The reaction to an outbreak of violence may be fear, which can stay or turn in anger. While experiencing fear the person's body is preparing for an escape or a defense against any possible attack, i.e. the heart rate and blood pressure increases, the eyes are open, and the pupils are wide, so the eyes can absorb the maximum amount of the light. In extreme situations fear might induce the muscle function loss, i.e. paralysis.

- The eyebrows are lifted and pulled inward.
- There are wrinkles on the forehead.
- The upper eyelids are lifted.
- The mouth is open, and the lips are tight according to the intensity of the emotion.

1.2.4 Happiness

Happiness or joy is a positive emotion often associated with a smile on the face. The happiness emotion appears when achieving goals. The typical characteristics on the face are:

- The lips corners are pulled back and up.
- The mouth can be open, and the teeth might be visible.

- The cheeks can be raised.
- The wrinkles under the lower eyelid might appear.
- The wrinkles appear outside the eye corners.

1.2.5 Sadness

Sadness appears when a person suffer. The origin of sadness is typically a loss of something. This emotion is calm, not impulsive and often accompanied by tears. During the emotion, the facial muscles lose the tension which may result in typical physiological features

- The inner parts of the eyebrows are pulled down.
- Lips corners pull down, and lips shake.

1.2.6 Surprise

Surprise is a sudden emotion. It comes without thinking and only lasts for a short time. The beginning is an unexpected or wrongly expected situation. Therefore, surprising emotion is either positive or negative. Surprise cannot be anticipated. If there is time to think about the situation, the subsequent reaction will not be the surprise. Surprise often proceeds into another emotion, usually happiness or sadness.

The typical features of surprise are lifted eyebrows, which additionally cause wrinkles on the forehead, widely open eyes, and a dropped jaw.

- The eyebrows are lifted and pulled inward.
- Horizontal wrinkles appear on the forehead.
- The eyes are open wide.
- The jaw is dropped. The mouth is opened, and the lips are tight.

2 Micro-expressions

Micro-expressions are defined as very brief, involuntary facial expressions, see Fig. 2.2. MEs tend to occur in high-stake and stressful situations, especially when something valuable can be gained or lost and the emotions are either deliberately or unconsciously concealed [12, 8, 11, 17]. Therefore, the MEs are a promising cue in catching liars and their detection could be important for police inquiring [32], airport security [20, 41] or psychological examinations. An observer can benefit from the fact, that very subtle facial expressions often leak, even if the subject attempts to conceal unwanted emotions [8, 19]. Compared to a polygraph, revealing a liar using a camera is not invasive and could be applied even without awareness of the subject.

Various reasons exist to hide emotions [30], e.g. cultural conventions, social behavior or even deception. People begin to learn controlling their emotions in their childhood, e.g. to smile when it is appropriate, not to express anger openly, to fake an emotion to deceive an opponent [14]. Ekman studied MEs in context of lies and showed MEs promising in revealing deception [11]. However, MEs do not provide any proof of lying and the context is always important, since MEs only reflect the current emotional state, e.g. an innocent person naturally feels anger in case of false accusation.

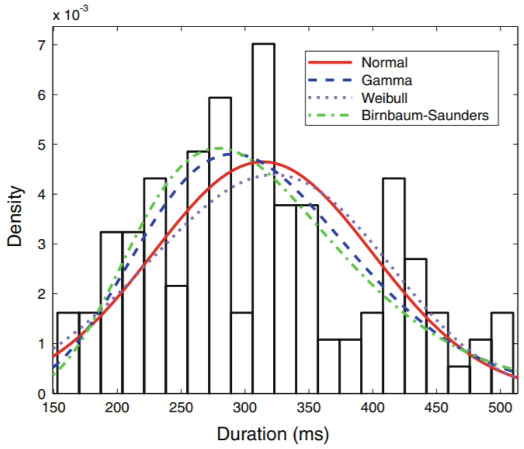

Figure 2.1 Histogram for the ME duration obtained from 109 elicited MEs and four different distributions fit the data. Taken from [46].

Recognizing ME can be helpful in many disciplines. Teachers could teach their students more efficiently. Business courses are promoted to improve the communication skills by reading facial expressions and thus increasing the emotional intelligence.

The definition of ME, however, is not very strict. The time duration is usually the key feature to distinguish MEs from a normal facial expressions [37]. An experimental study [46] discussing the duration of MEs considered the maximum time limit of MEs less than 500ms and showed the mean value 314ms, see Fig. 2.1. According to psychologists, the ME lasts between 1/25s and 1/3s [11, 29]. Due to a short duration and subtle changes on the face surface, noticing MEs is challenging even for people. Haggard and Isaacs, the first ME discoverers, claimed that spotting MEs for people in real time is not even possible [25]. Later Ekman published a study [8] showing the ability can be trained and improved by practicing. The techniques is developed in [9, 29].

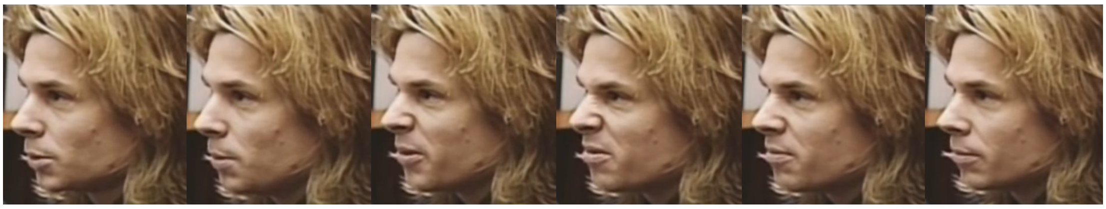

Figure 2.2 Example of a micro-expression in a real situation, source YouTube[1].

The MEs, similarly as common facial expressions, can be divided into six basics categories [10] - anger, disgust, fear, happiness, sadness and surprise. Each of them has a specific appearance in the face which can be described using Facial Action Coding System (FACS) [13], see Tab. 1.2.

[1] https://www.youtube.com/watch?v=HY8f8Ipkskg

3 Related work

Most of the related work is focused on the recognition problem; not many papers deal with the detection although a fast detection is necessary in real situations due to a rare occurrence of MEs. Especially in real videos, it is challenging to distinguish brief facial motions from neutral expressions and to avoid false alarms caused by normal expressions, global facial movements, speaking, occlusions, etc.

3.1 Micro-expression detection

ME detection or spotting refers to a problem of automatically searching a temporal location of ME occurrences. While this problem is common in other topics like eye-blinking detection, the ME detection problem is still largely undiscovered.

Polikovski et al. [35, 36] first described the problem in the computer vision area investigating both the detection and the recognition problem using 3D histogram of gradients. The obtained descriptors were clustered using k-means algorithm to obtain a particular AU. The results were presented only on a small database which is not available online and the MEs were not real but posed by non-professional volunteers.

Wu et al. [42] extracted the features using Gabor filters and detected MEs using GentleSVM. It is a combination of feature selection by GentleBoost and the best n features were used to train the SVM classier. The work deals with both the spotting and the recognition problem. The method was evaluated on METT training examples [9] determined for training people. See Sec. 6.1.4 for description of the METT examples.

Shreve et al. [38] made another research on the ME detection problem using optical flow and estimated the spatio-temporal strains. The results were evaluated on an unpublished database.

Moilanen et al. [31] proposed a method based on feature difference analysis. Features were obtained dividing the face into a 6×6 grid and each box described by LBP. They reported results on SMIC [28] and CASME database [47]. Further, Li et al. [27] integrated works of [31] and [34, 28] and proposed a system for an automatic ME analysis consisting of both spotting and recognition (MESR). They evaluated the results on CASMEII [45] and SMIC_E databases as longer clips. Read more details about the databases in Chap. 6.

Davidson et al. [7] proposed a method based on Histogram oriented Gradients features, the chi-squared dissimilarity measure and evaluated on their in-house dataset. This dataset was recently published as SAMM [6].

Patel et al. [33] used optical flow motion vectors and focused on detection onset, apex and offset frame of a micro-expression.

3.2 Micro-expression recognition

ME recognition problem refers to the classification of a pre-segmented video sequences into two or more classes, i.e. the isolated recognition problem. The pre-segmented sequence might be given by the detection algorithm. The classes can be either emotional states (anger, happiness, disgust, etc.), Action Units or ME/non-ME classes. The problem of detection ME/non-ME classes is possible to apply on the spotting problem together with sliding window. To our knowledge, nobody tried it on longer video sequences.

Polikovski et al. [35, 36] used the 3D gradient descriptor for the classification of AUs. For detection, see Sec. 3.1.

Phister et al. [34] proposed a Temporal Interpolation Model to handle the different video length, the spatio-temporal descriptor LBP-TOP to handle dynamic features and as a classifiers SVM, Multiple Kernel Learning, Random Forests were used.

Guo et al. [24] proposed a combination of LBP-TOP feature extraction and nearest neighbor classifier. The recognition problem was performed on the SMIC database.

Wang et al. [40] proposed a Tensor Independent Color Space (TICS) model and extracted LBP-TOP features and compared the novel TICS, RGB and gray color space. In extended version [39] more color spaces were compared (CIELab and CIELuv) and showed that the performance of TICS, CIELab and CIELuv are better then RGB or gray. Further, in another paper [40] the facial movements were described by Robust PCA and used for recognition.

Xu et al. [44] estimated a movement between consecutive frames using the optical flow estimation, located the local motion, i.e. filtered the global motion, and the principal direction of the residual motion was described by facial dynamic maps. SVM classifier was trained to recognize both the ME/non-ME and the emotion label.

4 Proposed method

A method which spots facial MEs is proposed in this chapter. We expect the input is a video acquired by a standard camera, i.e. at a frame rate 25fps. The source of the video is realistic, non-laboratory and the subject do not have to cooperate, but he/she moves freely. The viewpoint of the camera is not always frontal or stable. The illumination of the scene may change.

Our algorithm is based on the assumption that (usually a small) muscle contraction during the ME results in a sudden measurable change of intensity in a particular region inside the face image. Two effects are possible: (1) texture changes, e.g. wrinkles may appear, or (2) surface normal changes, when a larger textureless region is moved. Either of the effects or both of them are present. Since the magnitude is small, and the phenomenon can easily be confused with a global head/camera motion, speaking, eye-blinks, or presenting a normal (controlled) expression, we proposed an algorithm that first register face to undo the global motion, and then use a classifier to detect typical patterns of the ME from registered intensity difference signal over time.

The pipeline is depicted in Fig. 4.1. The face is first found and landmarks are detected in the video sequence. Then the face image is warped into a canonical coordinate system and split into ROIs describing the facial parts, where a motion caused by MEs is expected. Each ROI measures image intensity changes within a temporal sliding window. The SVM classifier was trained to distinguish the ME from false intensity changes caused by other motions and illumination changes.

4.1 Face rectification

To compensate the global motion, a transformation into a canonical coordinate system (of the generic frontal landmark configuration) is estimated, and the face image is warped accordingly. Facial regions are thus registered. Pixels are as much as possible corresponding over the time.

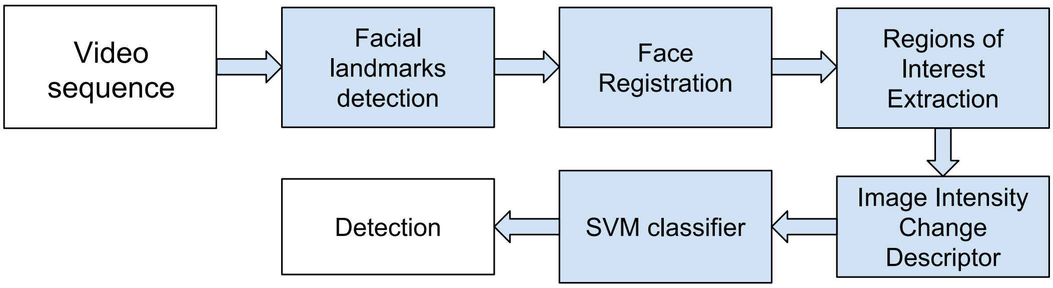
Video sequence
Facial landmarks detection
Face Registration
Regions of Interest Extraction
Image Intensity Change Descriptor
SVM classifier
Detection

Figure 4.1 A flowchart of the proposed automatic ME detection algorithm.

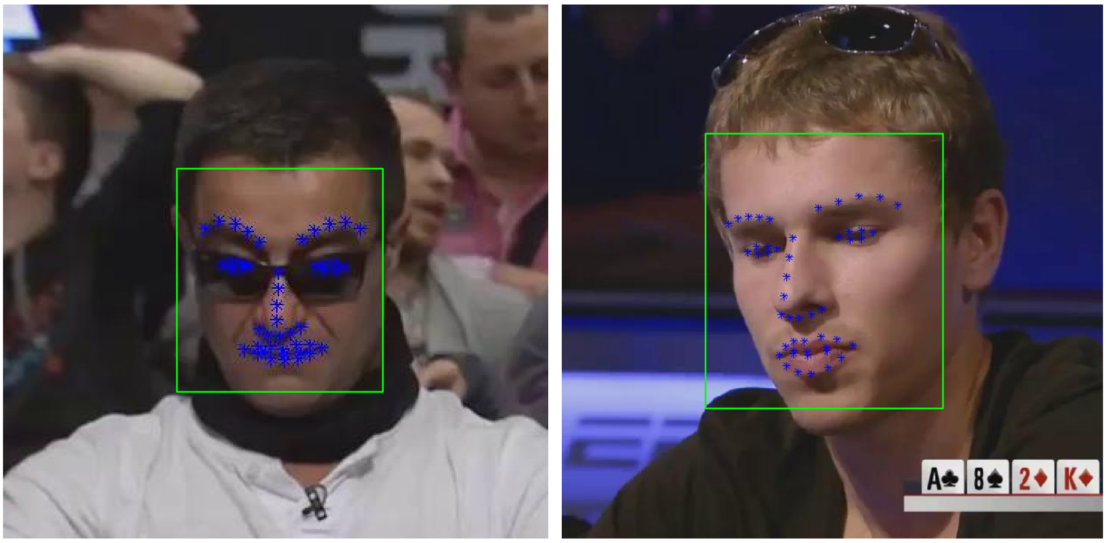

Figure 4.2 Examples of facial landmarks detected by the Intraface detector [43].

4.1.1 Landmarks

Two algorithms for the facial landmarks detection were tested - Chehra [2] and Intraface [43]. Both methods are state-of-the-art implementation of a cascade of linear regressors. The difference is that Chehra fits a 3D model of a cannonical facial shape and Intraface 2D model. The detectors output a set of 49 facial feature points in every video sequence frame t

$$\mathbf{x}^t = [(x_1^t, y_1^t), (x_2^t, y_2^t), \dots (x_{49}^t, y_{49}^t)]. \tag{4.1}$$

This is a standard set of landmarks defined on a contour of the eyes, eyebrows, nose and mouth, see Fig. 4.2.

4.1.2 Transformation

Similarity

Every face image in time t is transformed by similarity into the generic canonical shape model

$$\bar{\mathbf{x}} = [(\bar{x}_1, \bar{y}_1), (\bar{x}_2, \bar{y}_2), \dots (\bar{x}_{49}, \bar{y}_{49})], \tag{4.2}$$

which is an average shape of the landmarks. The generic shape is distributed with Intraface detector.

The similarity transformation is given by scale s, rotation angle φ and translation vector $\mathbf{x}_0$ by

$$\mathcal{S}(\mathbf{x}; s, \varphi, \mathbf{x}_0) = \begin{pmatrix} s\cos\varphi & -\sin\varphi \\ \sin\varphi & s\cos\varphi \end{pmatrix} \mathbf{x} + \mathbf{x}_0. \tag{4.3}$$

It is assumed the landmarks in the image and in the canonical shape are related by similarity.

The estimate is formulated in the least squares sense over all landmark points

$$\mathcal{S}^* = \arg \min_{s,\varphi,\mathbf{x}_0} \sum_{i=1}^{49} ||\mathcal{S}(\mathbf{x}_i; s, \varphi, \mathbf{x}_0) - \bar{\mathbf{x}}_i||_2^2. \tag{4.4}$$

Minimizing the objective is solved by Procrustes analysis [22]. Finally, the image is transformed into grayscale, smoothed by Gaussian filter with $\sigma = 1$ (to remove high-frequency noise of the camera), and warped by the estimated transformation $\mathcal{S}^*$ and cropped into 300×300 pixels image.

Note that the generic mean shape model may differ from the person specific landmark configuration. The shapes cannot be fully registered by the similarity transformation that captures only in-the-plane head/camera rotation, translation motion, and camera zooming. Nevertheless, in theory, this is not a major problem, since due to a fast nature of the ME, we compare intensities between nearby frames. Their landmark configurations are not very different. Although the rectified image is not perfectly frontal for off-the-plane rotations, nearby frames are transformed similarly. More complex transformation, e.g. using a piece-wise planar 3D model, or elastic registration might result in artifacts caused by landmark fluctuations. However, small landmark estimation errors are filtered out by the simple transformation of four parameters.

Rotation & translation

In situations, when the camera is not zooming, it is redundant to estimate the scale. Removing a single degree of freedom of the similarity from equation 4.3, the transformation is

$$\mathcal{R}(\mathbf{x}; \varphi, \mathbf{x}_0) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \mathbf{x} + \mathbf{x}_0. \tag{4.5}$$

The estimation problem is formulated in the least square sense as well as the similarity estimation in equation 4.4 and solved by Procrustes analysis [22].

4.2 Regions of Interest

The structure of facial muscles produces various facial expressions. It is observed that certain emotions always activate a subset of muscles and these motions are described by facial action units [13]. Polikovski et al. [35] came up with the idea dividing the face into regions to capture the specific facial muscles. As the occurrence of MEs is local, i.e. region-dependent, and manifests only in a small part of the face.

Following [35], twelve Regions of Interests (ROIs) were defined from the landmark positions on the face, see Fig. 4.3. Each ROI wraps a group of muscles, their contraction causes a specific facial motion (AU) and thus a change of expression. Additionally, each ROI can detect specific AUs, see Tab. 4.1.

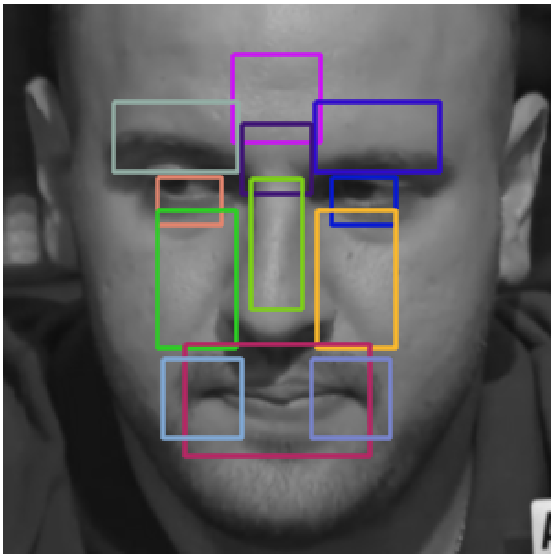

Figure 4.3 Rectified face with twelve regions of interests depicted. Regions of Interests are designed to wrap important facial muscles to determine the location of the ME.

Table 4.1 ROIs and corresponding AUs. [35]

ROI	AUs
Forehead	1,2,4
Left eyebrow	1,2
Right eyebrow	1,2
Between eyebrows	4
Left eye	7
Right eye	7
Nose	9
Left cheek	6,10
Right cheek	6,10
Mouth	16,17,20
Left mouth corner	11,12,13,14,15
Right mouth corner	11,12,13,14,15

4.2.1 Image stabilization by phase correlation

Due to a fluctuation of the landmarks, the displacement between a small residual gitter is observable in a registered video sequence. We model the instability by translation. The best displacement between subsequent ROIs having the maxima correlation measured between raw image intensities.

Having the location of ROI $\mathbf{R}_t$ at time t, the Pearson similarity statistics [23] was calculate to find the best displacement in extended ROI $\hat{\mathbf{R}}_{t+1}$ at time $t+1$. The extended ROI $\hat{\mathbf{R}}_{t+1}$ was chosen to be 10% larger than $\mathbf{R}_t$ to allow only limited displacement. To speed up the computation of the correlation coefficients, the advantage of the Fast Fourier Transform (FFT) was used. Let $\mathbf{F}$ and $\mathbf{G}$ be the Fourier image of $\mathbf{R}_t$ and $\hat{\mathbf{R}}_{t+1}$ respectively. Note that the extended $\hat{\mathbf{R}}_{t+1}$ had to be padded with zeros to the same size as $\mathbf{R}_t$. The phase correlation is obtained by

$$\mathbf{H}(x,y) = \mathscr{F}^{-1}\left(\frac{\mathbf{F}(u,v)\cdot\mathbf{G}^*(u,v)}{|\mathbf{F}(u,v)\cdot\mathbf{G}^*(u,v)|}\right), \tag{4.6}$$

where $*$ denotes complex conjugate number, and the optional displacement by

$$(x^*, y^*) = \arg\max_{x,y} \mathbf{H}(x,y). \tag{4.7}$$

The location of maxima of $\mathbf{H}$ gives the optimal displacement.

4.3 Face description

Facial ROI $k \in \{1,\ldots,12\}$ with an image frame at time t and a sub-window of size $m \times n$ were vectorized

$$I_t^k \in \mathbb{R}^{mn}. \tag{4.8}$$

As the regions are processed independently, the region index k is dropped in the subsequent text.

For each frame, every region I_t is photometrically normalized to suppress possible global illumination changes, e.g. intensity blinking due to fluorescent illumination that interferes with camera frame rate The normalization is done such that the mean is mapped to 0 and standard deviation to 1. The result is divided by total number of pixels in the ROI to balance the different size of ROIs

$$\hat{I}_t = \frac{1}{mn}\frac{I_t - \mu\mathbf{1}}{\sigma}. \tag{4.9}$$

The squared Euclidean difference between image regions at frame t_1 and frame t_2 is used as dissimilarity measure and is given by

$$d_{t_1,t_2} = ||\hat{I}_{t_1} - \hat{I}_{t_2}||_2^2 \tag{4.10}$$

We define the intensity descriptor for face region in frame t by collecting differences d_{t,t_1} over

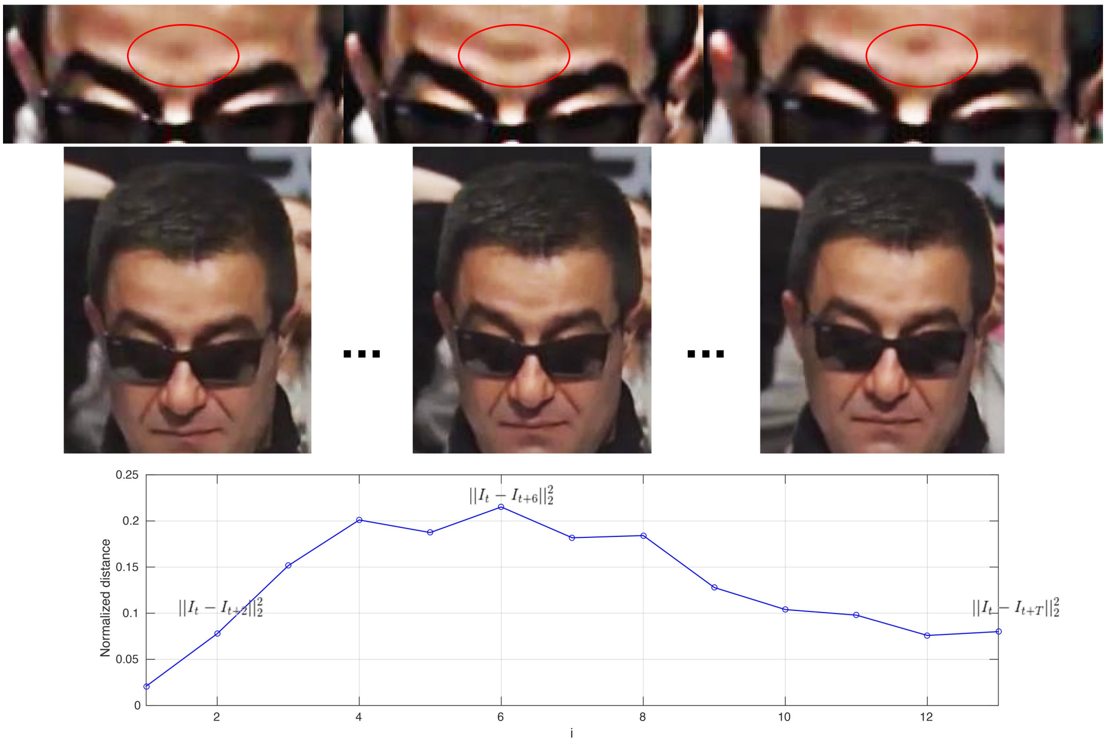

Figure 4.4 The micro-expression (a surprise emotion) after the poker gamer uncovered his cards. The ME starts with the neutral expression and returns after (12 frames) back to the neutral expression. In the middle frame, a raised eyebrow and wrinkles on the forehead can be noticed. The plot shows elements of the descriptor Φ_t for the forehead ROI.

the sliding window $t_1 \in \{1, \ldots, T\}$

$$\Phi_t = [d_{t,t+1}, d_{t,t+2}, \ldots, d_{t,t+T}]. \tag{4.11}$$

The ME is observed for a short period. According to the maximum considered ME duration, the sliding window of length $T = 0.5$s is used. The temporal descriptor measures the changes of image intensities within the sliding window for every ROI, see Fig. 4.4.

4.3.1 Feature smoothing

Components of the features Φ_t, further consider as signals, are differences between registered ROIs and the ROIs are not always stable. Thus signals suffer from noise, see Fig. 4.5. To reduce The noise signals are smoothed.

Moving average

The signal $\mathbf{\Phi}$ is smoothed using moving average with a uniform filter of size w equal to one-quarter of the dimension of the dimension of $\mathbf{\Phi}$.

$$\hat{\mathbf{\Phi}} = \mathbf{\Phi} * [\overbrace{\frac{1}{w}, ..., \frac{1}{w}}^{w-times}] \tag{4.12}$$

Convolution of the signal with the uniform filter gives the new descriptor vector.

Polynomial function

The polynomial function of degree four is chosen to fit the signal using linear regression. The solution of the problem leads to an overdetermined system of linear equations. The solution is in the least square sense. The signals replaced by the approximated function.

4.4 Micro-expression detection

Two methods are proposed. The first one, described in Sec. 4.4.1, does not involve any training and is based only on the fact that the highest image intensity changes are in the apex frame of the ME Sec. 4.4.1. The other one is supervised and is trained to capture the pattern of MEs Sec. 4.4.2.

4.4.1 Baseline method

Consider an example of the ME in Fig. 4.4 and the corresponding shape of Φ_t. In the most emotional ME frame (the apex frame), there is supposed to be the largest difference between the apex frame and the preceding frame. Therefore, a score is obtained by aggregating the differences

$$b(t) = \sum_{j=t-T}^{t-1} ||\hat{I}_t - \hat{I}_j||_2^2. \tag{4.13}$$

Note that the signal $b(t)$ is possible to obtain by summing the Φ at the time.

The signal $b(t)$ is thresholded. The maxima usually coincide with the ME apex frames. The idea behind is that in the case of the ME, Φ_t signals over reference frames t tend to be coherent and produce strong response $b(t)$ as demonstrated in Fig. 4.5.

However, the drawback is that other local maxima are generated by other events causing rapid changes in the image, e.g. a global motion or blinking. This weakness is partially mitigated by training a classifier as described in the following section.

4.4.2 SVM classifier

The classifier is trained to distinguish MEs from other false intensity changes. The SVM classifier with RBF kernel was trained on the SMIC_E_HS database. Videos in SMIC have 100fps frame rate. Therefore the feature vectors Φ_t were sub-sampled by taking every 4-th element since our test videos are only 25fps.

The positive samples were collected from the training sequences. Let t_i be all onset frames of the annotated MEs for every region, $i = 1, \ldots, N$. Then positive sample set is defined as

$$P = \bigcup_{i=1}^{N} \{\Phi_{t_i-15}, \Phi_{t_i-14}, \ldots, \Phi_{t_i+5}\}.$$

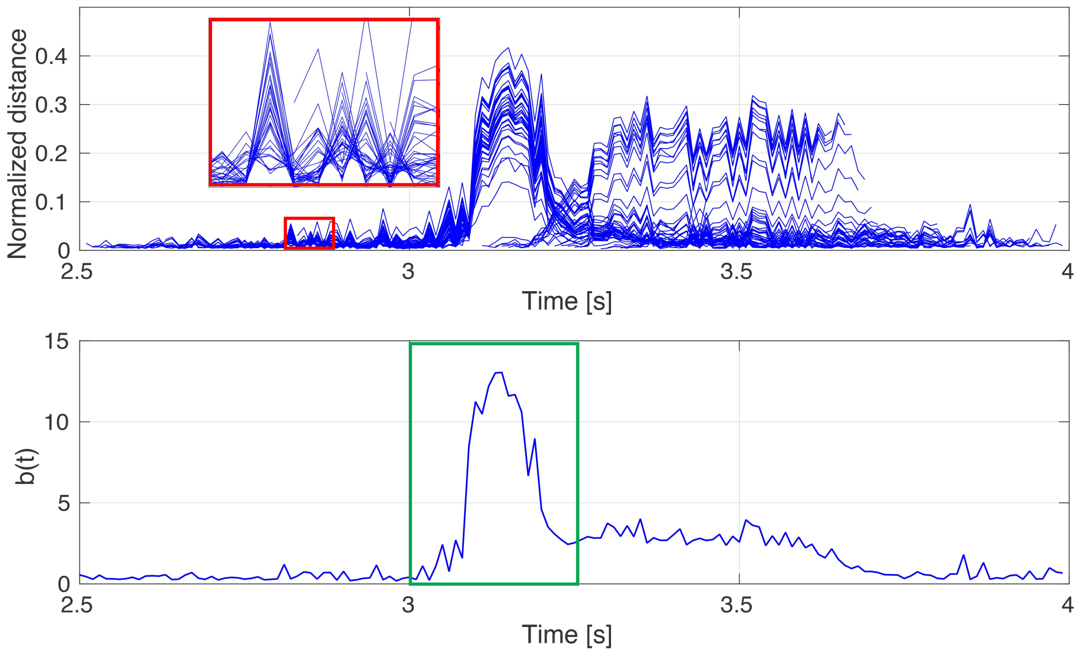

Figure 4.5 The upper plot shows the expanded intensity descriptors Φ_t over reference frames t. The coherence of the signals and the largest difference in the apex frame results in a peak of the aggregated response $b(t)$.

The data augmentation, where 15 frames before and 5 frames after the onset frame of the original 100fps video, ensures that the apex frame of the ME is always present within the feature vectors. An equal number of negative samples were collected from the rest of the video sequences randomly.

Note that a specific SVM classifier was trained for every region of interest $k = 1, \ldots, 12$. The reason of independent processing is that the intensity pattern may be different due to e.g. amount of texture within the region. Finally, SVMs were trained from 800–2000 samples depending on the occurrence of the ME in a particular ROI in the training data.

5 Implementation details

A few tricks to improve the detection were implemented. The regions of eyes were finally removed from the set of detectors as eye blinking is a rapid movement of approximately same duration as MEs. However, filtering the eye blinks is not possible since the spontaneous blinking is often partially overlapping with the MEs.

For “in the Wild” videos, the participants were not always recorded from the frontal pose. Then a subset of ROIs became fully or partially occluded due to a head yaw, see Fig. 5.1. Therefore, we detected a non-frontal head pose and the features from the occluded parts of the face were not considered in further processing. The head pose was estimated simply by computing the natural logarithm of the ratio between the inner eye landmarks and the top nose landmark distances. In the frontal pose, the log-ratio was approximately 0, while the non-frontal pose was detected when the magnitude of the log-ratio exceeds empirically set threshold 0.5.

Finally, the detector response was modified by non-maximal suppression with 20 frames window to avoid multiple responses for the same event in the data.

The entire pipeline was implemented in MATLAB and was not optimized. However, we believe a real-time performance for standard 25fps videos would be easily achieved, since the landmark detection algorithm is real-time. The transformation estimation and the face area warping are cheap operations. Moreover, the differences among the registered ROIs are computed independently, therefore, can be computed and evaluated parallel.

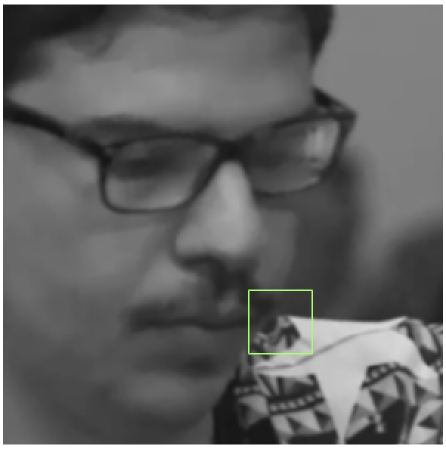

Figure 5.1 The occlusion caused by a head yaw. The ROI does not envelope the facial region and is excluded from the classification.

5.1 Annotation tool

A simple annotation tool was developed to improve the effectiveness of catching MEs. The tool allows playing videos using an external player (VLC player based mex) and controlling the player from MATLAB. It offers simple commands, i.e. play/pause, normal speed or modified speed playing and passing the video frame by frame forward and backward. Additionally, a certain clip of the video can be played repeatedly.

The annotation tool is designed to add multiple labels to events in the video sequence and allows creating video sections, which can be used to distinguish different TV show or poker game parts, e.g. camera switching cards and faces.

The interface, see Fig. 5.2, consists of four windows. The initial window let the user choose the video file. In the other window, the player is controlled. It contains buttons PLAY/PAUSE, frame FORWARD, BACKWARD, play a 1s LOOP and a slider to control the video player speed. The third window is meant for annotations. After starting a section, a user chooses a number of annotation tracks. The tracks are used for different people or different motions in the face. In the current clip, a user clicks the button START MICROEXPRESSION to add the initial label or END MICROEXPRESSION to finish the interval. The section is finalized by clicking the End Section. The last window is the video player. After closing the program, the results are saved in MATLAB data file *mat*.

The annotation tool was useful to spot MEs effectively. The MEVIEW database, described in Chap. 6.2, was annotated using the tool.

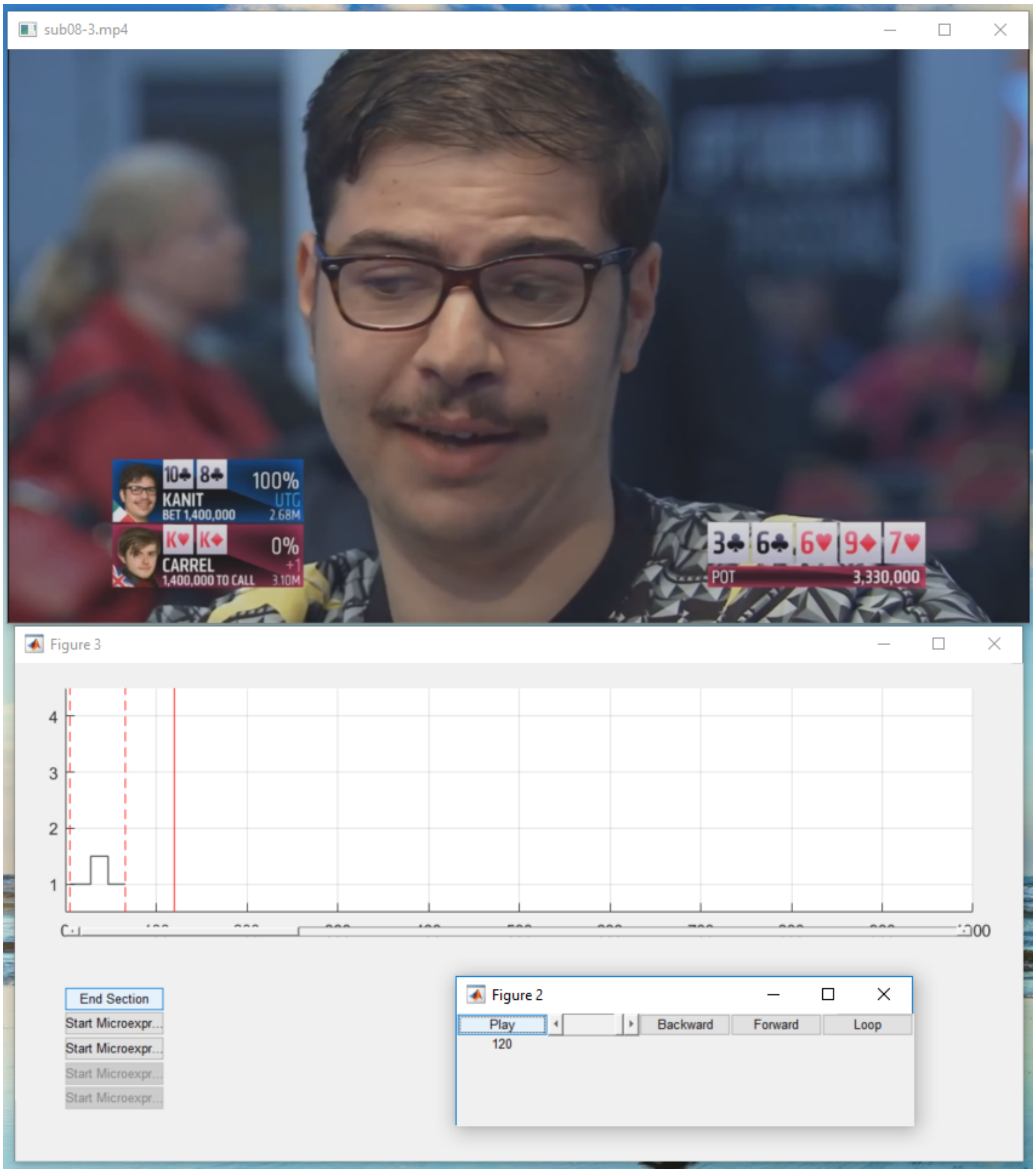

Figure 5.2 Screenshot of the annotation tool developed for the micro-expression spotting. Three main windows are showed in the figure - the video control panel, the annotation panel and the player.

6 Datasets

Obtaining a good database of MEs is highly challenging due to the non-trivial process of evoking MEs. Especially in laboratory conditions, high-stake situations can be hardly evoked as well as to force participants to hide the true emotions. An example for good ME occurrence situations might be criminal interrogations. When the suspect pretends to be innocent and tries to conceal his/her true emotions, the leaked MEs on his/her face would reveal the lie. However, obtaining such database would involve close cooperation with police and either way those confidential data would be surely impossible to publish.

Therefore, the current databases are made either by posing faked MEs, i.e. the participants tried to simulate normal expressions very fast, or, the other approach, which is used in all recent published databases, was inducing MEs by watching emotive videos and trying to suppress any emotions, i.e. the normal expressions.

Games are an another idea based on the stress factor and the need to hide emotions. We collected videos of poker games from YouTube and annotated MEs. The details are described in Sec. 6.2.

In literature, there are databases of spontaneous MEs collected by asking participants to watch an emotive video in the laboratory environment, in front of a high-speed camera. Then participants' task was to conceal any emotions. As a motivation, they either got a money reward CASMEII [47], SAMM [6] or were punished by filling a long form in SMIC [28] in the case of apparent failure. Those databases are described in Sec. 6.1.

Besides the spontaneous databases, other databases have been collected. The problem was that most of them were not published or often contained faked MEs, which is in contradiction with Ekman theory that MEs cannot be controlled [12]. They are discussed in Sec. 6.1.4. All known datasets are summarized in Tab. 6.1.

6.1 Published datasets

In this section, the most important standard databases are described.

Table 6.1 Micro-expressions databases summary

	Polikovski	USF-HS	SMIC	CASME	CASMEII	SAMM	MEVIEW
Micro-expressions	42	100	164	195	255	159	31
Participants	10	Unknown	16	19	26	32	16
Resolution	640 × 480	720 × 1280	640 × 480	640 × 480/720 × 1280	640 × 480	2040 × 1188	720 × 1280
Frame rate	200	29.7	100	60	200	200	25
Spontaneous/posed	Posed	Posed	Spontaneous	Spontaneous	Spontaneous	Spontaneous	Spontaneous
Environment	Laboratory	Laboratory	Laboratory	Laboratory	Laboratory	Laboratory	Real
Main purpose	Recognition	Detection	Recognition	Recognition	Recognition	Recognition	Detection

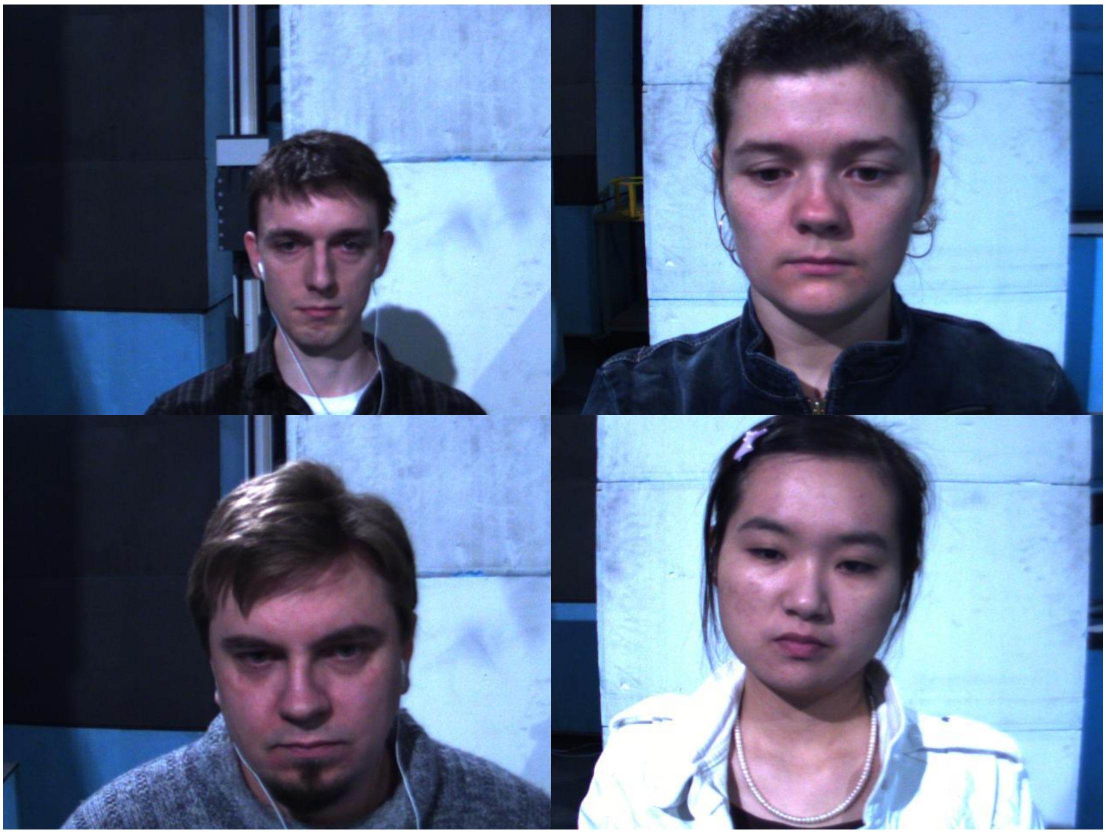

Figure 6.1 Four examples from the SMIC database [28]. The participants were recorded from the frontal poses, while watching an emotive video clip.

6.1.1 SMIC

The original SMIC database [28], see Fig. 6.1, consists of 164 MEs with 20 participants. The resolution is 640 × 480 px and frame rate 100 Hz. The average facial resolution is 190 × 230. The participants were recorded in a laboratory environment with controlled illumination, ensured by four lights in the room corners. Participants sat in stable positions in front of the high-speed camera. The emotional videos were played. After each video, they filled a questionnaire about their feelings from the video clip.

The idea of obtaining MEs was based on two conditions. (1) Stimuli strong enough to elicit the emotion on the neutral face, (2) motivation to conceal any emotion. The first requirement was satisfied by watching emotive videos. When the participant failed in keeping the neutral face, e.g. the participant burst out laughing, he/she had to fill a long, annoying questionnaire containing more than 500 questions.

The videos were clipped only to the duration of MEs with no margin before and after the ME. Therefore, it is suitable for the isolated recognition. The detection problem was performed in published paper clipping neutral parts and classifying ME/non-ME. ME clips were annotated by three categories - positive, negative and surprise. The expressions were not FACS coded.

Later few extended version were published. 10 participants were recorded by a standard video camera (VIS) and a near-infrared camera (NIR) with 25 fps. Especially for the detection

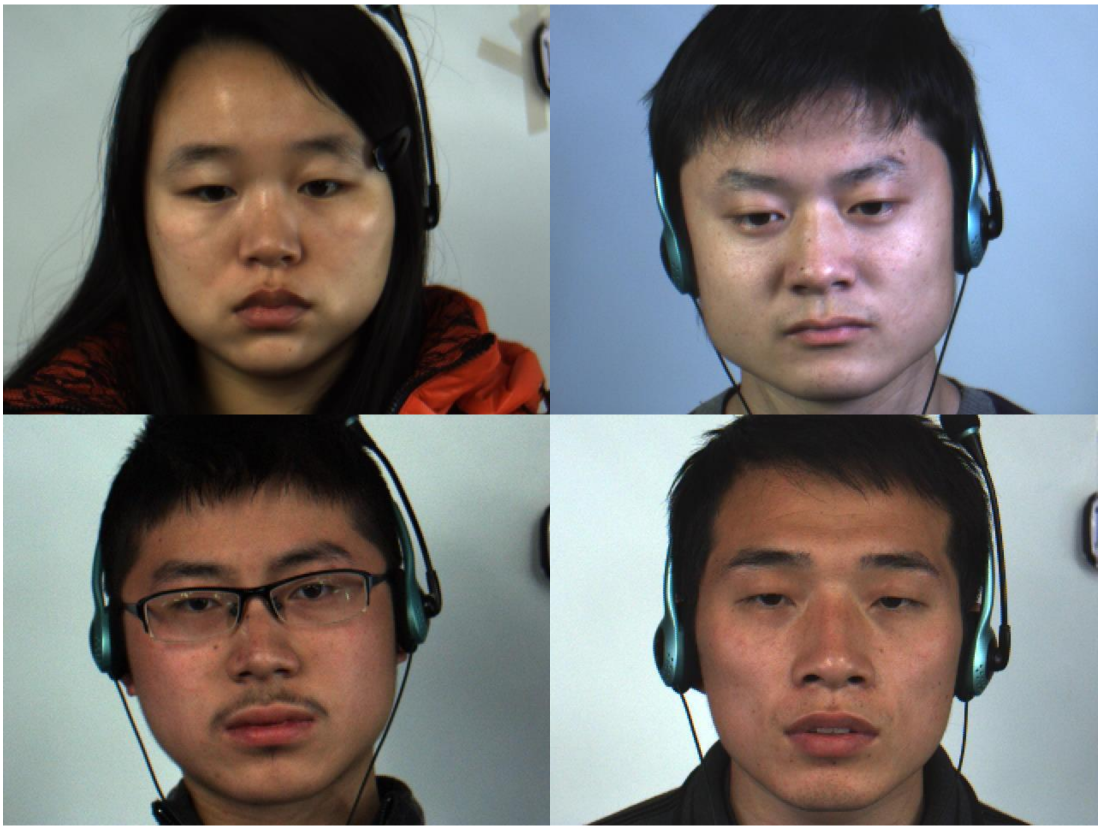

Figure 6.2 The participants from CASME2 database [45] recorded in frontal pose.

problem, longer videos were published (SMIC_E). Time margin before and after the ME was preserved in video clips and onset and offset frames were labeled. The average length of the videos is 5.9s.

6.1.2 CASME

The CASME database [47] consists of 195 elicited MEs from 19 participants captured in 60 Hz frame rate. The CASME data were recorded by two cameras. The first with resolution 1280×720 (CASME-A) and the other one with 640×480 (CASME-B). The cameras also differ in the environment. CASME-A was recorded under a natural light and CASME-B in a room equipped with two LED lights.

The environment was controlled and participants were asked to keep a neutral face and do not move to prevent noisy output. The emotional stimuli were induced by watching very emotional videos, see in Tab. 6.2. Psychological researchers manually detected MEs. The moving facial parts were annotated by AUs and the emotional labels were added to the spotted ME.

However, recorded videos were clipped such that only a small neutral face video interval was preserved around the ME. Therefore, this database is primarily meant for the isolated recognition problem.

Later the improved database CASMEII [45], see Fig. 6.2, was released as an extension of the original dataset. The database contains 255 videos with 26 subjects. CASMEII was recorded

Table 6.2 Examples of stimuli used to induce emotions in CASMEII database [45].

Video Stimuli Description	Emotion Link
Jokes on names	Happiness
Tooth extraction	Disgust
A girl killed by car	Sadness
Dog torturing	Anger

by a high-speed camera with 200fps and resolution 640 × 480. Four LED lamps balanced the problem that lights flickering appeared in high-speed videos.

The protocol was the same as for the CASME database. The onset, apex, and offset frames were annotated in the clipped videos; FACS coded, and emotion type annotated. Few examples of the video stimuli is in Tab. 6.2

Videos were clipped, and the average length of video clips is 1.25s. The ME was usually in the middle. The database is available for research purposes.

6.1.3 SAMM

A new database called SAMM [6] focused primarily on the resolution of MEs. The participants were recorded with 200fps, video resolution 2040 × 1088 and 400 × 400 facial resolution. Together 159 micro-movements from 32 university participants were collected. The experiment was in fully controlled laboratory conditions. Each participant sat in front of the camera in the laboratory room. The light conditions were ensured by two lights with an array of LEDs to balance the aliasing. Aliasing makes a problem for high-speed camera recordings as the common frequency of the lights is 50Hz and causes flickering on the captured images.

SAMM used the similar experiment protocol as the SMIC and CASME. Participants were asked to stay in the same position and keep as much neutral expression as possible with a motivation of winning £50. The emotions were induced by a video stimuli downloaded from the Internet, see Tab. 6.3. The full list of used videos can be found in the original database paper [6]. The source of the videos was not mentioned.

The contribution of the collection was to capture details of micro-movements in the face with a high-resolution camera and to annotate the facial motion by FACS coding system, see Tab. 1.1. SAMM is not focusing on emotional labels.

The database is available for research purposes, and clipped micro-movements can be downloaded directly from the Internet. The full version, which would be useful for detection, is also available, however, because of a large amount of data it is not stored online, but can be sent on a disk.

6.1.4 Others

Polikovski et al. [35] created a collection of posed ME, i.e. the MEs were not spontaneous but performed artificially. The participants were ten students, and they were asked to perform seven basic emotion very fast. The camera resolution was 640 × 480 and 200fps.

Table 6.3 Examples of stimuli used to induce emotions in SAMM database [6].

Video Stimuli Description	Emotion Link
Snake attacks camera	Fear
A dog being kicked	Anger/Sadness
Twin towers collapsing	Sadness
Baby laughing	Happiness

USF-HD [38] is a collection of 100 posed MEs. The participants saw an example of a particular ME and tried to mimic it as quick as possible. The collection consists of 100 MEs with the resolution of 720 × 1280 and with 29.7fps. The database is not available to download.

Micro-Expression Training Tool (METT) [9] is a training tool created by Paul Ekman to teach people recognizing MEs. However, the dataset is not suitable for machine learning as the sequence contains the same neutral expression picture before and after. One frame with a facial expression (often not ME, but the macro-expression) is located in the middle. The principle of teaching people is to train the ability recognize the particular emotion extremely fast. METT is mentioned here as Wu et al. [42] performed experiments on this dataset.

6.2 Collection of “In the Wild” videos

We collected new MEVIEW dataset (Micro-Expressions VIdEos in the Wild), see Fig. 6.3. The dataset consists of videos from poker games, and TV interviews downloaded from the Internet. The advantage of poker games is the stress environment and the necessity to hide emotions. Opponents try to conceal or fake their true emotions, which is a scenario where MEs are likely to appear. The MEs are still rare since the TV show post production often cut the most valuable moments such as the detail of a player’s face while the cards are being uncovered or someone raises, calls or folds his/her cards.

MEs were manually explored in videos following the Ekman advice. First, watch the scene frame by frame and gradually increase the speed after. Special attention was paid around key moments of the game in case of poker videos, or while a person was listening to a hard question in the case of the TV interviews. These suspicious events were checked carefully. The annotator participated in the Ekman’s online course[1] and he was able to detect several MEs in the videos with reasonable confidence. The onset and offset frames of the ME were marked in long videos, and FACS coded, and the emotion types were annotated. In summary, 31 video clips with 16 individuals were collected. The video resolution is 720 × 1280 and frame rate 25. In the experiments, the videos were clipped such that each video clip contains one camera shot with the face which was being analyzed. Therefore the average duration of the videos is 3s. The facial area differs as the camera is often zooming, changing the angle and the scene.

[1] http://www.paulekman.com/micro-expressions-training-tools/

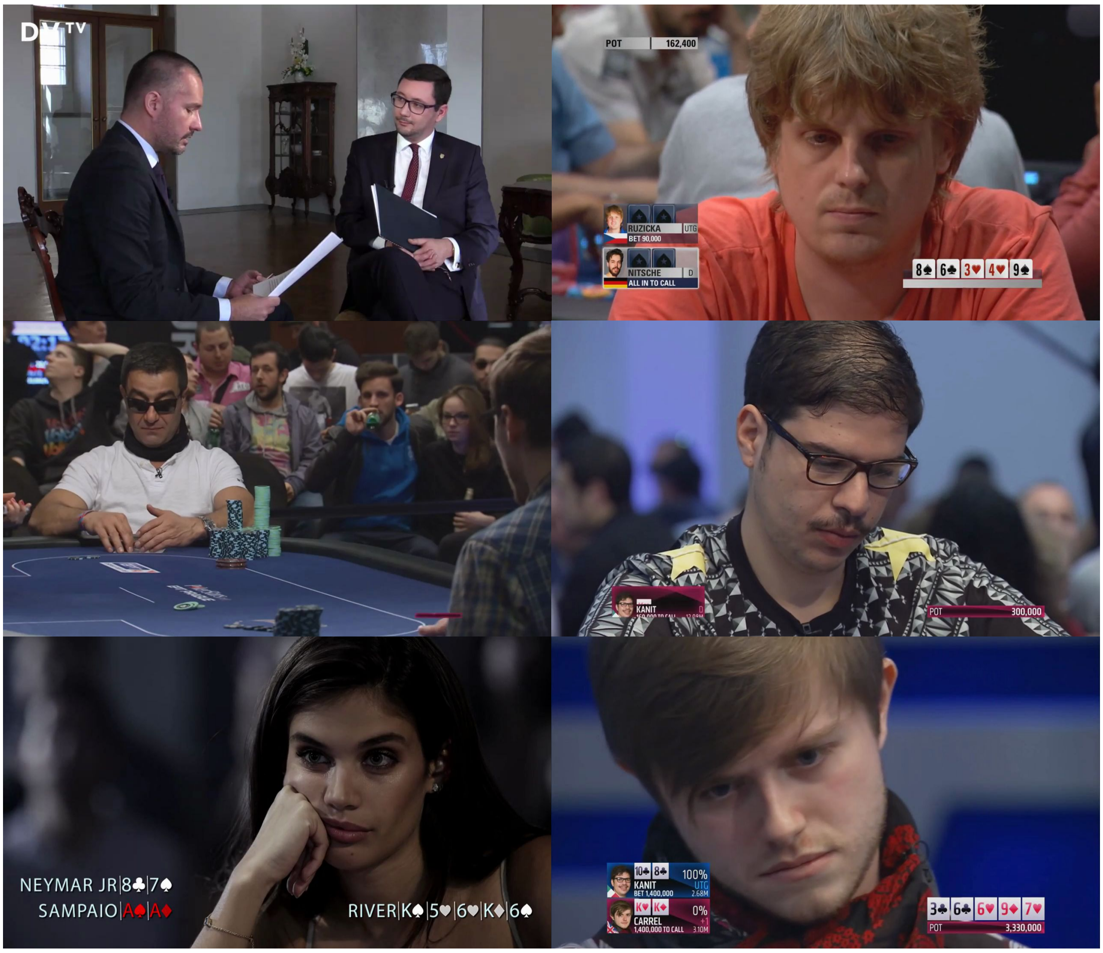

Figure 6.3 Examples from the database MEVIEW contain many different phenomena besides the frontal head pose, e.g. small facial resolution, sunglasses, occlusion, in-the-plane and off-the-plane head rotation.

7 Experiments

The proposed method was tested on both standard SMIC [28] and CASME II [45] datasets, and on our MEVIEW dataset. Another suitable database would be SAMM. However, due to the recent release, we did not get the full version. We performed the cross-validation and cross-dataset experiments with testing the proposed methods in the detection pipeline and evaluated the baseline method with multiple design choices in each block of the pipeline, see Fig. 7.6. The algorithm was tested against competing method of Li et al. [27] with favorable results.

7.1 Evaluation protocol

For evaluation, we strictly follow the evaluation protocol of [27]. Having annotated onset and offset frames of the ME, then a frame is considered to be correctly detected if it is in range [onset-($N/4$), offset+($N/4$)], where N is the maximal considered length of a ME, $N = 64$ for CASMEII (200 Hz), $N = 32$ for SMIC (100 Hz), and $N = 8$ for MEVIEW dataset (25 Hz). It means the interval is expanded by a small margin to tolerate a low uncertainty of the annotation in the precise ME interval. All such correctly detected frames are counted as True Positives (TP). If a detection is out of the range, then all N frames are counted as False Positives (FP). The True Positive Rate (TPR) is TP divided by a number of annotated positive frames, and the False Positive Rate (FPR) is FP divided by the total number of frames in a sequence without the number of annotated positive frames. The performance is evaluated by receiver operating characteristic (ROC) curve and the area under the curve (AUC).

For our method, at each frame, we have twelve detectors that are spotting MEs. The detection of the ME is considered if at least one detector fires.

7.2 Cross-validation on SMIC dataset

In the SMIC_E_HS database, there are 20 participants. Therefore a 20-fold cross-validation, always leaving all videos of one participant out, was performed to train the SVM classifier and to estimate the detection accuracy.

The videos were manually checked, and only ROIs with an observable muscle motion in the annotated ME interval was considered as a valid sample in the training set. Cross-validation ROC curves are shown in Fig. 7.1, together with the mean ROC curve. It is seen, that ROC curves have a relatively high variance for cross-validation runs. The reason is probably that MEs of certain participants are much easier/more difficult detectable than for others. Certain subjects seem almost like having a wax face, spotting their MEs is even difficult for a manual

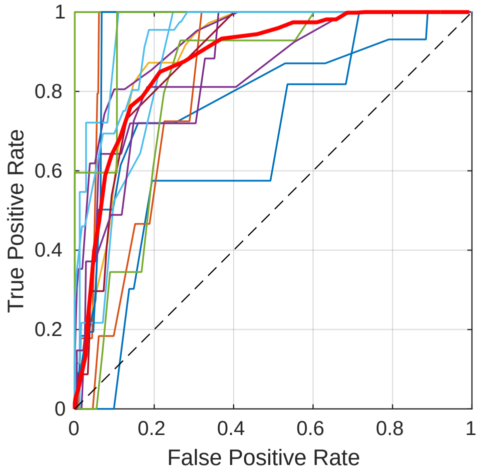

Figure 7.1 The 20-fold cross-validation on SMIC_E_HS dataset. The mean ROC curve, highlighted in red, has AUC = 0.88.

inspection of the video played slowly and repeatedly. On the hand, we noticed movements that might be micro-expressions that were not annotated in the dataset.

The mean ROC curve is compared to the baseline method, Sec. 4.4.1, and the result of [27] in Fig. 7.3. The baseline methods contain choices - Intraface landmarks, similarity transformation and none smoothing of the features. The mean ROC curve outperforms both the baseline thresholding and the method [27]. Nevertheless, average result of the cross-validation is shown.

7.3 Cross-database experiment

The SVM detectors were trained on the entire SMIC_E_HS database and the resulting classifiers were evaluated on CASMEII, see Fig. 7.4 and on the MEVIEW database, see Fig. 7.5.

The CASMEII contains videos in 200 Hz frame rate. Therefore the videos were sub-sampled taking every 8th frame. The proposed SVM detector outperformed the other methods. For CASMEII, we can see the SVM significantly outperformed the baseline thresholding, and it is slightly more accurate than [27].

Results on our MEVIEW dataset are inferior to the results on CASMEII and SMIC. The reason is that MEVIEW dataset is much more challenging and includes many “in-the-Wild” phenomena that cause false positives. Nevertheless, the strongest detection in both the SVM and baseline methods belongs to true MEs. An example of detection scores of all twelve ROIs for several frames of the poker game video, around an event shown in Fig. 4.4, is presented in Fig. 7.2. We can see the true ME have the highest score, while several top scoring events false events appear, probably due to improperly compensated motion.

The proposed algorithm produces some false positives. They are often caused by eye blinks. Eye blink duration is similar to MEs. Excluding the eye regions helped to reduce the false alarms. However, we observed that the eye blinks causes a shift of the entire set of landmarks and thus influence the transformation, which may result in false detection in other regions than

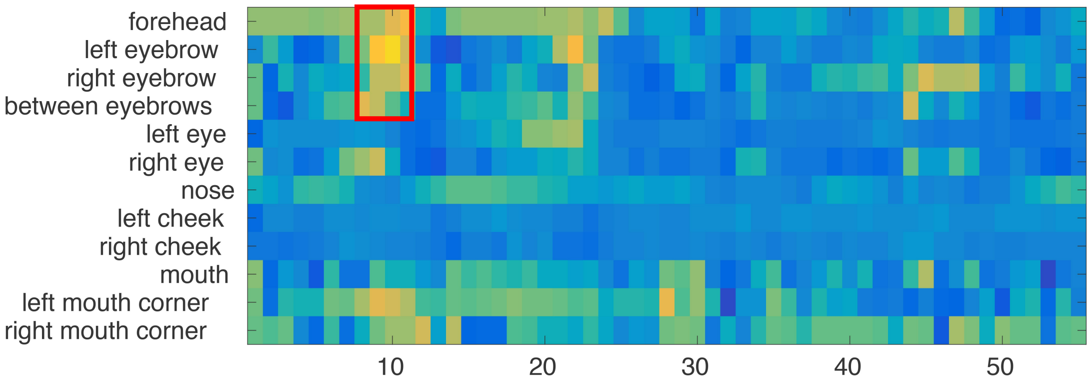

Figure 7.2 The output of the SVM detectors for twelve ROIs. Yellow color refers to higher score of the SVM. The ME is marked by a red rectangle.

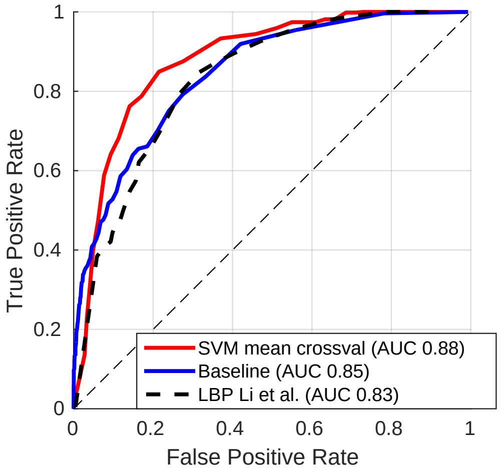

Figure 7.3 Comparison among the mean SVM cross-validation ROC, the baseline method the LBP-based method by Li et al. [27] on SMIC_E_HS.

around the eyes. Completely filtering the eye blink instances would on the other decrease the true positive rate, since the eye blinks often co-occur with MEs or MEs often follow shortly after the eye-blinks.

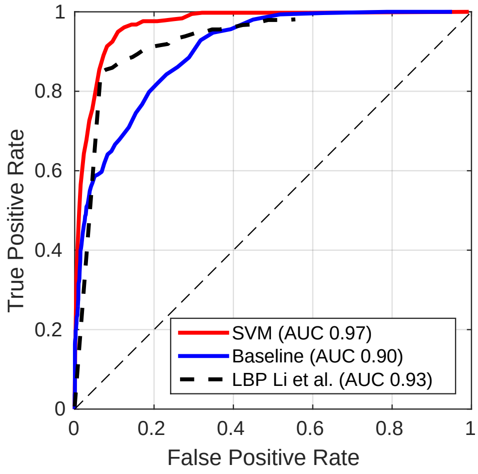

Figure 7.4 The ROC curves of the cross-dataset SVM, baseline method and LBP by Li et al. [27] on CASMEII.

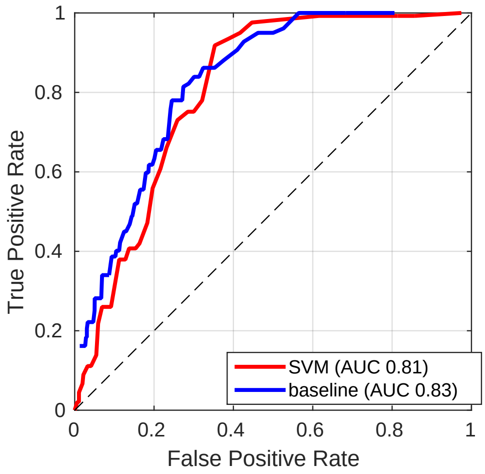

Figure 7.5 Comparison of the ROC between the cross-dataset SVM and baseline method on MEVIEW database.

7.4 Introspection experiments

We measured the impact of particular choices in the pipeline, see Fig. 7.6. All the methods are described in Chap. 4. Additionally, an option that the face registration was skipped was tested. The ROIs were defined from landmarks independently and were placed based on the location of their landmarks. The size of the regions was derived from the inter-ocular distance and their orientation was defined by the vector connecting the eye centers. This method is further called *withoutRegistration*.

A summary of results all tested options is present in Tab. 7.1, 7.2 and 7.3. It is observable that the Chehra landmarks [2] are superior to the Intraface landmarks [43] at first sight. We can see that the signal smoothing matters, as the non-smoothed features are inferior. The more detailed introspection follows.

7.4.1 Landmark selection

First, the impact of landmark selection was measured on databases CASME2, see Fig. 7.7 and SMIC_E_HS, see Fig. 7.8. The curves were obtained by fixing either Chehra or Intraface algorithm, and all combination of the rest of the pipeline were depicted on the x-axis. The AUC is compared.

Chehra algorithm outperformed the Intraface algorithm on both databases in almost every case of the rest of the pipeline combinations. The differences are not too significant, especially in situations when the faces were not rectified.

Intraface provides a robust landmark detection for challenging examples, e.g. non-frontal poses, expressions, low facial resolution. Intraface regresses 2D landmark positions directly. While Chehra fits a 3D model doing the estimation of the 3D pose and the output landmarks are the projection of the model. It seems this strategy is more stable for easier frontal poses with neutral expressions. Nevertheless, the differences are rather minor.

7.4.2 Impact of facial rectification

Then, the impact of facial rectification was measured on all three databases, see Fig. 7.9, Fig. 7.10 and Fig. 7.11. Four registration methods were fixed and all combinations of landmarks and feature smoothing methods combined. The AUC was compared.

In the laboratory conditions on databases CASME2 and SMIC_E_HS, the rotation and translation registration method outperformed all other methods, see Fig. 7.7 and 7.8, which confirms

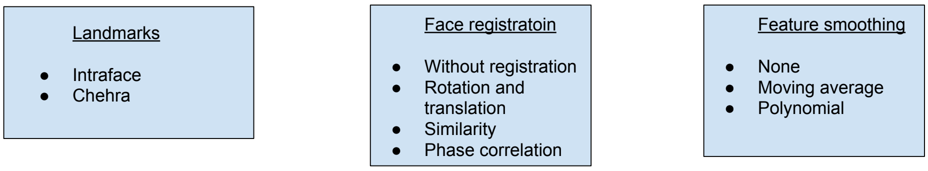

Figure 7.6 Flowchart of different method choices in each block of the pipeline.

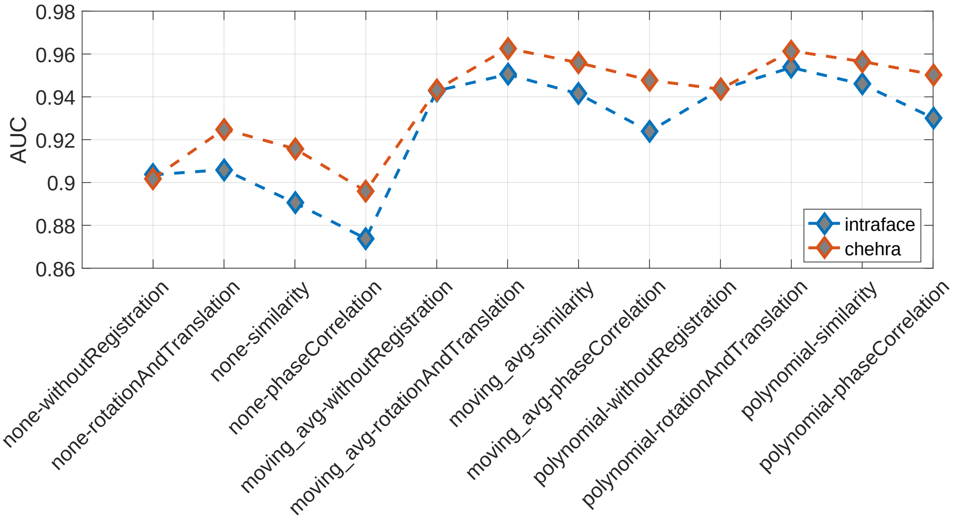

Figure 7.7 Comparison of the area under the ROC curve (AUC) between different landmarks for combinations of registration methods and feature smoothing methods in the pipeline for the database CASME2.

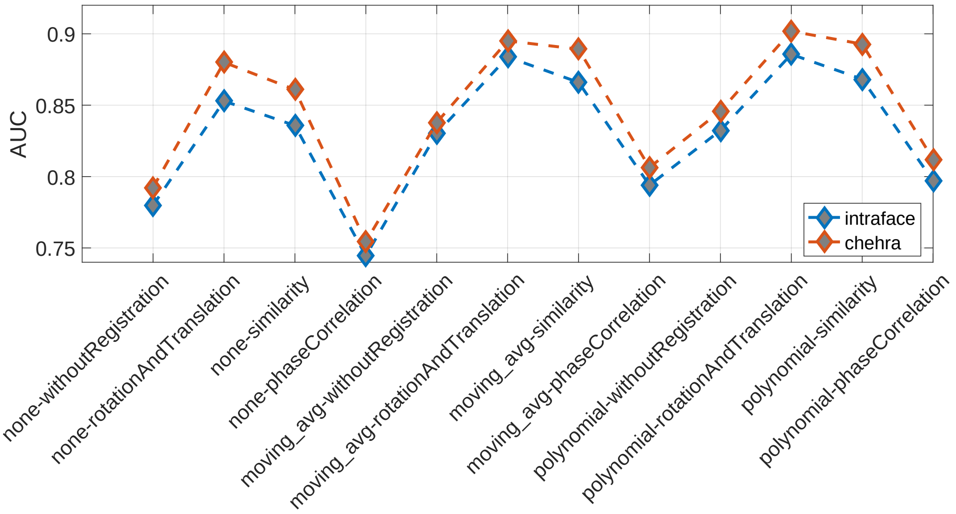

Figure 7.8 Comparison of the area under the ROC curve (AUC) between different landmarks for combinations of registration methods and feature smoothing methods in the pipeline for the database SMIC_E_HS.

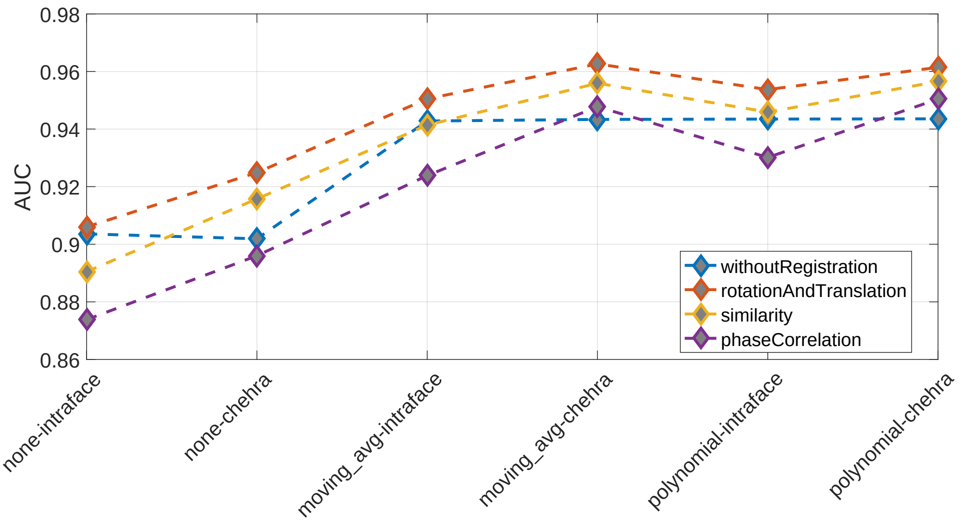

Figure 7.9 Comparison of the area under the ROC curve (AUC) among different face registration methods for combinations of smoothing methods and landmark algorithms in the pipeline for the database CASME2.

our expectations. The participants were in stable position. Therefore the scaling factor is redundant.

Results of similarity are comparable to rotation and translation. The results on CASME2 and SMIC_E_HS are slightly worse then rotation and translation. However, on the MEVIEW, similarity outperformed the other two methods more significantly.

The option *withoutRegistration* surprisingly outperformed almost all methods on the MEVIEW database, Fig. 7.11. The reason might be that participants often change the pose thus the estimated transformation (both similarity and rotation and translation) warps a non-frontal facial pose. Placing the landmarks directly to the face can suppress the effect despite the landmark fluctuation. Results of *withoutRegistration* option on CASME2 database are comparable to similarity described before. The CASME2 videos have the best facial resolution, i.e. the landmarks are supposed to be accurate. Rotation and translation reduce the noise even more. On the other hand, SMIC_E_HS has a larger margin between the similarity and the registration option *withoutRegistration*. The facial resolution is lower and also light conditions are worse than the CASME2 database.

The phase transformation has the worst results on all three datasets. The aim was to compare the template, i.e. each ROI with larger search area in consecutive frames and find the best match within. The expectation was that finding the most correlated match within the ROI nearby area increases the robustness of the method and results in more accurate descriptors. However, the human skin does not contain significant texture. Especially ROIs as forehead or cheeks have the almost featureless surface, and the correlation is not informative.

To provide a more accurate estimate, an improvement could be using a prior displacement that would force the uncertain solution, caused by ambiguity due to the weak texture, to stay close to zero.

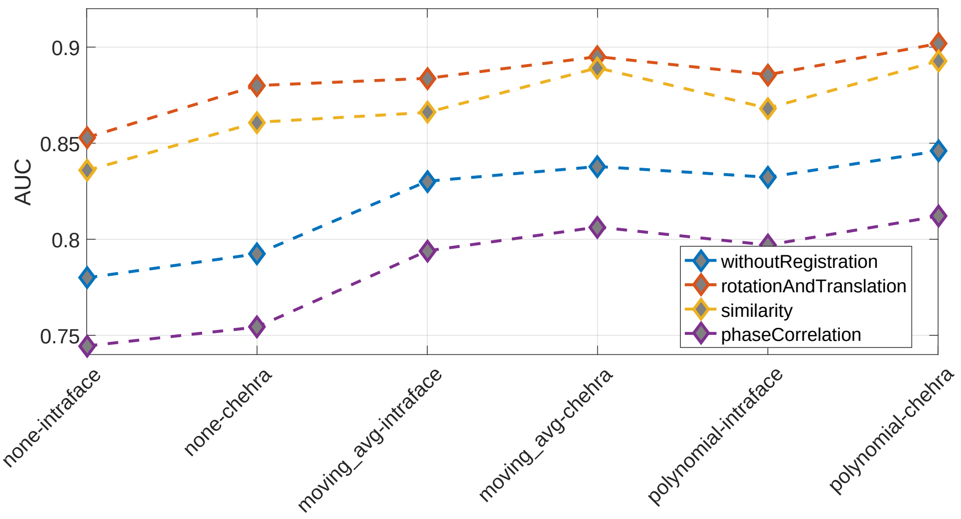

Figure 7.10 Comparison of the area under the ROC curve (AUC) among different face registration methods for combinations of smoothing methods and landmark algorithms in the pipeline for the database SMIC_E_HS.

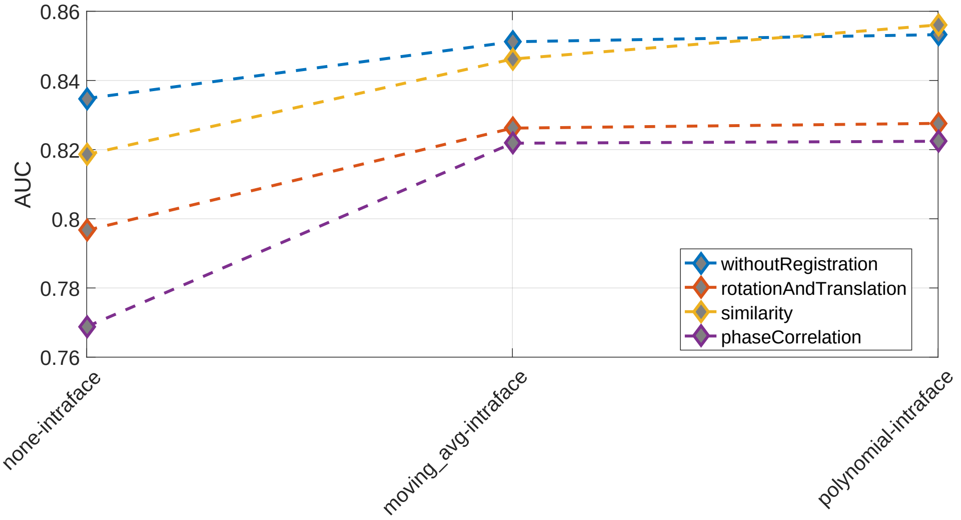

Figure 7.11 Comparison of the area under the ROC curve (AUC) among registration methods for different feature smoothing methods in the pipeline for the database MEVIEW.

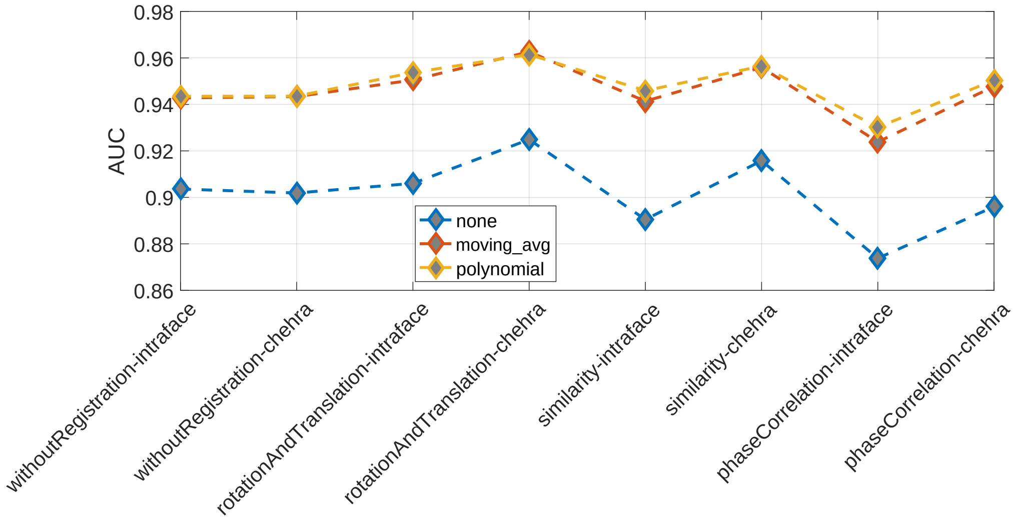

Figure 7.12 Comparison of the area under the ROC curve (AUC) among different feature smoothing methods for combinations of registration methods and landmark algorithms in the pipeline for the database CASME2.

7.4.3 Feature smoothing

In the last experiment, the smoothing methods were compared among each other. The smoothing method was fixed and the landmark algorithms and registration method were combined.

The aim was to reduce the noise of the descriptors that might be caused by e.g. inaccurate facial rectification.

On all three plots Fig. 7.12, 7.13 and 7.14 can be seen that the feature smoothing has positive impact on the accuracy. Moving average and polynomial function fitting have almost the same influence on the AUC measure on all three databases.

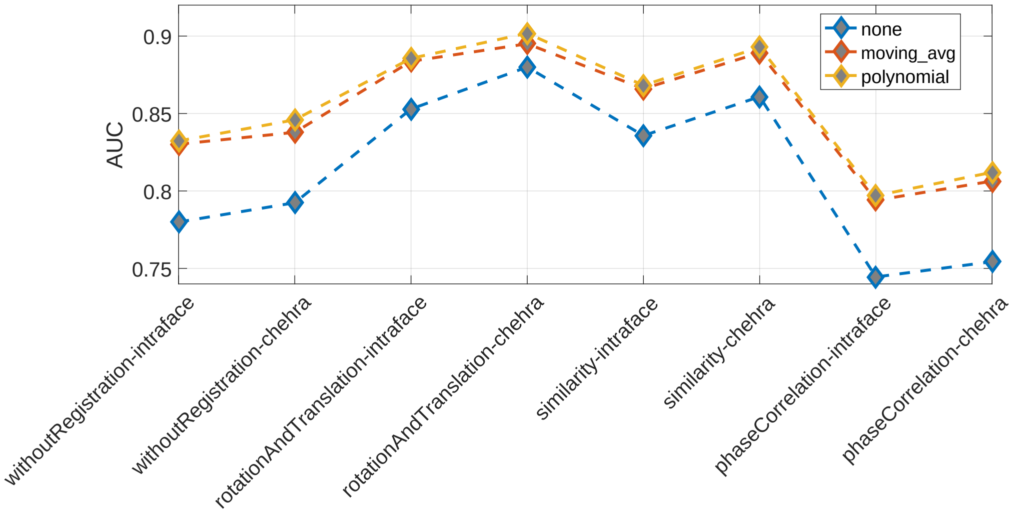

Figure 7.13 Comparison of the area under the ROC curve (AUC) among different feature smoothing methods for combinations of registration methods and landmark algorithms in the pipeline for the database SMIC_E_HS.

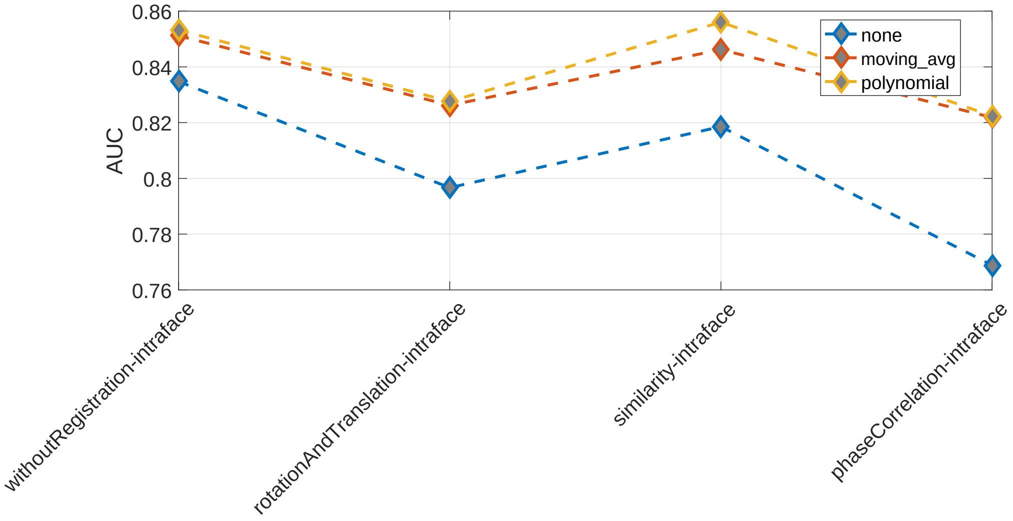

Figure 7.14 Comparison of the area under the ROC curve (AUC) among different feature smoothing methods for different registration methods in the pipeline for the database MEVIEW.

Table 7.1 Results of the pipeline for the CAME2 database of the baseline method.

#	Method	AUC
1	CASME2-moving_avg-rotationandtranslation-chehra	0.96263
2	CASME2-parabolic-rotationandtranslation-chehra	0.96144
3	CASME2-parabolic-similarity-chehra	0.95649
4	CASME2-moving_avg-similarity-chehra	0.95592
5	CASME2-parabolic-rotationandtranslation-intraface	0.95371
6	CASME2-moving_avg-rotationandtranslation-intraface	0.95049
7	CASME2-parabolic-fft-chehra	0.9504
8	CASME2-moving_avg-fft-chehra	0.94777
9	CASME2-parabolic-similarity-intraface	0.9459
10	CASME2-parabolic-withoutRegistration-chehra	0.94356
11	CASME2-parabolic-withoutRegistration-intraface	0.94347
12	CASME2-moving_avg-withoutRegistration-chehra	0.94335
13	CASME2-moving_avg-withoutRegistration-intraface	0.94279
14	CASME2-moving_avg-similarity-intraface	0.94138
15	CASME2-parabolic-fft-intraface	0.93015
16	CASME2-none-rotationandtranslation-chehra	0.9249
17	CASME2-moving_avg-fft-intraface	0.92385
18	CASME2-none-similarity-chehra	0.91576
19	CASME2-none-rotationandtranslation-intraface	0.90603
20	CASME2-none-withoutRegistration-intraface	0.90359
21	CASME2-none-withoutRegistration-chehra	0.90189
22	CASME2-none-fft-chehra	0.896
23	CASME2-none-similarity-intraface	0.89047
24	CASME2-none-fft-intraface	0.87381

Table 7.2 Results of the pipeline for the SMIC_HS database of the baseline method.

#	Method	AUC
1	SMIC_HS-parabolic-rotationandtranslation-chehra	0.90181
2	SMIC_HS-moving_avg-rotationandtranslation-chehra	0.89504
3	SMIC_HS-parabolic-similarity-chehra	0.89279
4	SMIC_HS-moving_avg-similarity-chehra	0.88923
5	SMIC_HS-parabolic-rotationandtranslation-intraface	0.88564
6	SMIC_HS-moving_avg-rotationandtranslation-intraface	0.88374
7	SMIC_HS-none-rotationandtranslation-chehra	0.88001
8	SMIC_HS-parabolic-similarity-intraface	0.86822
9	SMIC_HS-moving_avg-similarity-intraface	0.86601
10	SMIC_HS-none-similarity-chehra	0.86089
11	SMIC_HS-none-rotationandtranslation-intraface	0.85291
12	SMIC_HS-parabolic-withoutRegistration-chehra	0.84594
13	SMIC_HS-moving_avg-withoutRegistration-chehra	0.83787
14	SMIC_HS-none-similarity-intraface	0.83574
15	SMIC_HS-parabolic-withoutRegistration-intraface	0.8323
16	SMIC_HS-moving_avg-withoutRegistration-intraface	0.83021
17	SMIC_HS-parabolic-fft-chehra	0.81203
18	SMIC_HS-moving_avg-fft-chehra	0.80637
19	SMIC_HS-parabolic-fft-intraface	0.797
20	SMIC_HS-moving_avg-fft-intraface	0.79397
21	SMIC_HS-none-withoutRegistration-chehra	0.7924
22	SMIC_HS-none-withoutRegistration-intraface	0.77993
23	SMIC_HS-none-fft-chehra	0.75432
24	SMIC_HS-none-fft-intraface	0.74447

Table 7.3 Results of the pipeline for the MEVIEW database of the baseline method.

#	Method	AUC
1	MEVIEW-parabolic-similarity	0.85604
2	MEVIEW-parabolic-withoutRegistration	0.85324
3	MEVIEW-moving_avg-withoutRegistration	0.85119
4	MEVIEW-moving_avg-similarity	0.84619
5	MEVIEW-none-withoutRegistration	0.83477
6	MEVIEW-parabolic-rotationandtranslation	0.82759
7	MEVIEW-moving_avg-rotationandtranslation	0.82621
8	MEVIEW-parabolic-fft	0.82242
9	MEVIEW-moving_avg-fft	0.82185
10	MEVIEW-none-similarity	0.81858
11	MEVIEW-none-rotationandtranslation	0.79667
12	MEVIEW-none-fft	0.76882

8 Conclusion

In the thesis, we studied a specific type of facial expressions called micro-expressions. The importance of the micro-expression detection is that micro-expressions are involuntary thus reveal the true emotions of a person. Therefore, having a good detection method might be useful for criminal investigation, airport security or psychological examination.

A simple detection method was proposed. The method was designed to spot MEs in non-laboratory conditions, but also in real-world videos. The method is based on measuring the brief local image intensity changes caused by facial muscle contractions, which result in a specific pattern.

Despite few micro-expressions databases were already published, all of them were recorded in laboratory conditions with a similar protocol and are mostly designed for the isolated recognition. It was desired to find another way of obtaining micro-expressions, preferably in the real situations.

We collected the MEVIEW dataset of “in-the-Wild” phenomena, consisting of mostly poker tournament videos downloaded from YouTube. Poker games seem to be rich in micro-expression occurrences as the basic assumptions, e.g. stress, fear of loss, are fulfilled.

The proposed method was evaluated on SMIC and CASMEII databases and our challenging MEVIEW database. The proposed algorithms were compared against a method of [27] with favorable results. An introspection throughout the designed choices was carried out.

As a limitation of the proposed method, we see the number of false alarms. Nevertheless, we observed the correct micro-expressions tend to give very high scores. More examples collected by manual inspection of high scoring events in long videos by experienced human annotators could be important and more data would surely allow designing a more sophisticated classifier with higher detection accuracy.

Bibliography

[1] Muscles of facial expression. https://www.kenhub.com/en/library?sequence=facial-muscles. Accessed: 2016-12-19. 5, 6

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014. 15, 36

[3] Radomír ČIHÁK. *Anatomie 2. 2. vyd. Praha*. Grada Publishing, 2002. 6

[4] Victoria Contreras. Artnatomya. http://www.artnatomia.net/, 2005. 6

[5] Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 5

[6] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2016. 12, 26, 29, 30

[7] Adrian K Davison, Moi Hoon Yap, and Cliff Lansley. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 1864–1869. IEEE, 2015. 12

[8] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003. 10

[9] Paul Ekman. *Micro Expressions Training Tool*. Emotionsrevealed. com, 2003. 10, 12, 30

[10] Paul Ekman. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan, 2007. 11

[11] Paul Ekman. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009. 10

[12] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969. 10, 26

[13] Paul Ekman and Wallace V Friesen. Facial action coding system. 1977. 6, 7, 11, 16

[14] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003. 7, 10

[15] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013. 5

[16] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *California Mental Health Research Digest*, 8(4):151–158, 1970. 5

[17] Paul Ekman, Maureen O'Sullivan, and Mark G Frank. A few can catch a liar. *Psychological science*, 10(3):263–266, 1999. 10

[18] Margaret J Fehrenbach and Susan W Herring. *Illustrated anatomy of the head and neck*. Elsevier Health Sciences, 2015. 5, 6

[19] Mark G Frank and Paul Ekman. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of personality and social psychology*, 72(6):1429, 1997. 10

[20] Mark G Frank, Carl J Maccario, and Venugopal Govindaraju. Behavior and security. *Protecting Airline Passengers in the Age of Terrorism. Greenwood Pub Group, Santa Barbara, California*, pages 86–106, 2009. 10

[21] Wallace V Friesen and Paul Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983. 7

[22] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991. 16

[23] A Ardeshir Goshtasby. *Image registration: Principles, tools and methods*. Springer Science & Business Media, 2012. 18

[24] Yanjun Guo, Yantao Tian, Xu Gao, and Xuange Zhang. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3473–3479. IEEE, 2014. 13

[25] Ernest A Haggard and Kenneth S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*, pages 154–165. Springer, 1966. 10

[26] Carl-Herman Hjortsjö. *Man's face and mimic language*. Studen litteratur, 1969. 5

[27] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Reading hidden emotions: spontaneous micro-expression spotting and recognition. *arXiv preprint arXiv:1511.00423*, 2015. 12, 32, 33, 34, 35, 44

[28] Xiaobai Li, T. Pfister, Xiaohua Huang, Guoying Zhao, and M. Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6, April 2013. 12, 26, 27, 32

[29] David Matsumoto and Hyi Sung Hwang. Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, 35(2):181–191, 2011. 10

[30] David Matsumoto, Seung Hee Yoo, and Sanae Nakagawa. Culture, emotion regulation, and adjustment. *Journal of personality and social psychology*, 94(6):925, 2008. 10

[31] Antti Moilanen, Guoying Zhao, and Matti Pietikäinen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1722–1727. IEEE, 2014. 12

[32] Maureen O'Sullivan, Mark G Frank, Carolyn M Hurley, and Jaspreet Tiwana. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530–538, 2009. 10

[33] Devangini Patel, Guoying Zhao, and Matti Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 369–380. Springer, 2015. 12

[34] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *2011 International Conference on Computer Vision*, pages 1449–1456. IEEE, 2011. 12, 13

[35] S. Polikovsky, Y. Kameda, and Y. Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*, pages 1–6, Dec 2009. 12, 13, 16, 17, 29

[36] Senya Polikovsky and Yoshinari Kameda. Facial micro-expression detection in hi-speed video based on facial action coding system (facs). *IEICE transactions on information and systems*, 96(1):81–92, 2013. 12, 13

[37] Xun-bing Shen, Qi Wu, and Xiao-lan Fu. Effects of the duration of expressions on the recognition of microexpressions. *Journal of Zhejiang University Science B*, 13(3):221–230, 2012. 10

[38] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar. Macro-and micro-expression spotting in long videos using spatio-temporal strain. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 51–56. IEEE, 2011. 12, 30

[39] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. *IEEE Transactions on Image Processing*, 24(12):6034–6047, 2015. 13

[40] Sujing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In *ICPR*, pages 4678–4683. Citeseer, 2014. 13

[41] Sharon Weinberger. Airport security: Intent to deceive? *Nature*, 465(7297):412–415, 2010. 10

[42] Qi Wu, Xunbing Shen, and Xiaolan Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. In *International Conference on Affective Computing and Intelligent Interaction*, pages 152–162. Springer, 2011. 12, 30

[43] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 15, 36

[44] Feng Xu, Junping Zhang, and James Wang. Microexpression identification and categorization using a facial dynamics map. 2016. 13

[45] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 12, 28, 29, 32

[46] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37(4):217–230, 2013. 10

[47] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7, April 2013. 12, 26, 28