Aalto University
School of Science
Master's Programme in Machine Learning and Data Mining

Marcos Luzardo

Eye and Mouth Openness Estimation in Sign Language and News Broadcast Videos

Master's Thesis
Espoo, May 16, 2014

Supervisor: Jorma Laaksonen, D.Sc. (Tech.), Aalto University
Advisor: Ville Viitaniemi, D.Sc. (Tech.)

Aalto University
School of Science
Master's Programme in Machine Learning and Data Mining

ABSTRACT OF
MASTER'S THESIS

Author: Marcos Luzardo

Title:
Eye and Mouth Openness Estimation in Sign Language and News Broadcast Videos

Date: May 16, 2014 **Pages:** 108

Major: Machine Learning and Data Mining **Code:** T-61

Supervisor: Jorma Laaksonen, D.Sc. (Tech.), Aalto University

Advisor: Ville Viitaniemi, D.Sc. (Tech.)

Currently there exists an increasing need of automatic video analysis tools to support sign language studies and the evaluation of the activity of the face in sign language and other videos. Henceforth, research focusing on automatic estimation and annotation of videos and facial gestures is continuously developing. In this work, techniques for the estimation of eye and mouth openness and eyebrow position are studied. Such estimation could prove beneficial for automatic annotation and quantitative evaluation of sign language videos as well as towards more prolific production of sign language material.

The method proposed for the estimation of the eyebrow position, eye openness, and mouth state is based on the construction of a set of facial landmarks that employ different detection techniques designed for each facial element. Furthermore, we compare the presented landmark detection algorithm with a recently published third-party face alignment algorithm. The landmarks are used to compute features which describe the geometric information of the elements of the face. The features constitute the input for the classifiers that can produce quantized openness estimates for the studied facial elements. Finally, the estimation performance of the estimations is evaluated in quantitative and qualitative experiments with sign language and news broadcast videos.

Keywords: facial state recognition, sign language analysis, news broadcast analysis, eye openness, mouth openness, eyebrow position

Language: English

Acknowledgements

To my wife, for her continuing support during the writing of this work. I wish to thank as well the members of the Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL) project for their ideas and contribution to my research, and personal development.

Espoo, May 16, 2014

Marcos Luzardo

Abbreviations and Acronyms

AAM	Active Appareance Model
ASL	American Sign Language
ASM	Active Shape Model
AU	Action Unit
EOG	Electrooculography
FACS	Facial Action Coding System
FinSL	Finnish Sign Language
FPS	Frames per second
HOG	Histogram of Oriented Gradients
IR	Infra Red
LBP	Local Binary Pattern
LED	Light Emitting Diode
MCC	Mathew's Correlation Coefficient
NB	Naive Bayes
NM	Non-manual
RBF	Radial Basis Function
RGB	Red Green Blue (channels in color images)
ROI	Region of Interest
SIFT	Scale-Invariant Feature Transform
SL	Sign Language
SSR	Single Scale Retinex
SURF	Speeded Up Robust Features
Suvi	Suomalaisen viittomakielen verkkosanakirja (The online dictionary of FinSL)
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression

Contents

Chapter 1

Introduction

In the area of sign language (SL) research, studies focusing on the linguistic significance and importance of facial movements while signing is an increasingly active topic. Up to now, research and application of automatic estimation methods for facial movements such as eyebrow position, eye openness, and mouth states in SL is relatively new and there is room for further development. With recent development of several SL corpus projects, there is an increasing need of automatic video analysis and annotation tools to support linguists in their studies. This will enable quantitative evaluation of available video material without expending effort in manual annotation. In the area of news broadcast videos, state estimation of face elements is significant for automatic transcription, information extraction, speaker identification, and genre classification of scenes.

This work studies techniques that can be used to produce automatic annotations of eye and mouth openness to create tagged material for SL video analysis. Classification categories are proposed for the states of all studied facial elements. The proposed openness estimation process consists of a hybrid approach: it combines features extracted from the geometric relations between detected facial landmarks, and different appearance-based algorithms for face landmark detection. An estimation model created from the best performing algorithms is produced and evaluated against linguistic annotations of the tested videos.

A correlation measure and graphical assessment method proposed for evaluation makes this work different from similar studies that employ performance measures such as accuracy percentage as evaluation tools. The videos used in the experiments for training and testing the classifiers are taken from a SL dictionary. A second video data set, containing videos of artificially articulated facial movements, was produced as a part of this work and is used for training the classifiers and assess their applicability. A third video

set taken from news broadcasts is used for qualitative performance analysis of the estimation model.

1.1 Motivation and objectives

The increasing number of SL corpus projects and the appearance of new computer vision methods for video analysis has produced an interest of new, precise methods for automatic annotation of SL video material. This has enabled more opportunities to build tools for estimating the state of facial elements in videos. The main motivation of this work is to develop a method for automatic annotation of eyebrow position, eye openness, and mouth state to create tagged material for SL video analysis. Automatic annotation of SL videos could prove beneficial for deeper understanding of facial movements as well as towards more prolific production of SL material.

The objective of this work is to study techniques that can be used for automatic annotation of the state of facial articulators (eyebrow, eye, and mouth) to help in quantitative and qualitative analysis of non-manual markers in SL, and to evaluate the generalization strength of the proposed model. In this context, this work attempts to answer the following research questions:

1. Can facial landmarks be used effectively for eye and mouth openness estimation?
2. Is it feasible to build a model for the automatic annotation of eye and mouth states in SL videos?
3. Are the annotations produced by the model reliable?
4. Is it possible to automatically estimate the eyebrow position?
5. Can a model for automatic eye and mouth openness annotation be used also for news broadcast video analysis?

Through this work the `SLMotion` software package [47] was used for video processing and face detection. The proposed algorithms, except when indicated otherwise, were developed and tested using the `MATLAB` software package [61].

1.2 Scope of the work

The methodological approach used in this work follows the experimental design principle [30]: the data, methods, and performance measures are selected with the goal of resolving the research objectives in mind. The study

focuses on the estimation of (1) eyebrow position for shift identification, (2) eye openness for blink detection, and (3) mouth state for mouth openness estimation and movement detection in videos. Face detection, as well as expression recognition, is not part of this work. Additionally, since estimations are done frame by frame, automatic pattern identification of actions (blinking and yawning for example) are not evaluated. This also excludes categorization of mouth movements for lip reading, and automatic discrimination between mouthings and mouth gestures.

The estimation of facial elements consists of landmark detection, feature extraction, and classification in per-frame configuration. The performance of the estimations are evaluated in two main experiments: first, SL video annotations; and second, news broadcast video annotations. The experiments employ a small quantity of annotated videos for training and testing. The video material for the experiments is divided according to its origin being either videos from a SL dictionary, or videos produced with intentionally posed facial movements. Quantitative evaluation is performed only for the SL video experiment, while qualitative evaluation is done for the SL and news broadcast experiment.

In this work, the outline to build the model for automatic annotation of the state of facial elements is as follows: the first step is to establish a categorization system for eyebrow position, eye openness and mouth state. The categorization system should describe the facial movements to be estimated. The second step consists of detecting landmarks for eyebrows, eyes, and mouth; in this work this is done by using different algorithms for each element. In the third step, a set of geometric features are computed from the detected landmarks. During the fourth step two types of classifiers are trained using the extracted features. The fifth step consists of the analysis of the classifiers' performance. Finally, the best landmark detection algorithms for each facial element are combined into a model and a qualitative analysis of the annotations produced by the system is performed on randomly selected videos.

1.3 Organization of the thesis

This thesis is arranged as follows: in Chapter 2 the state of recent research in eye and mouth openness estimation is presented. The chapter also introduces the main research done in the domain of this work, considering the eyebrows, eyes, and mouth as non-manual articulators in SL as well as facial expression analysis for news broadcast video.

Chapter 3 describes the proposed categorization system for the eyebrow

position, eye openness, and mouth state. The categorization covers binary as well as multi-class states for each face element. The chapter concludes with a quick review of the video material used in the experiments, annotated manually according to the presented categorization.

Chapter 4 presents the implementation details of the algorithms for landmark detection. Landmark detection algorithms are described for each studied facial element. The chapter also introduces the Supervised Descent Method (SDM), a third-party landmark detection algorithm. The classification results obtained with the proposed landmark detection algorithms are compared to those obtained using the SDM landmarks. In Chapter 5 a set of facial features based on geometrical properties of the face are presented. The features are extracted from the previously detected facial landmarks.

The experiment setup is explained in Chapter 6. The chapter starts by providing further details on the videos used for the experiments. It continues by describing the organization of the experiments according to the landmark detection algorithms and data used for training the classifiers. In the following sections, the chapter addresses the SL video experiment and the news broadcast video experiment. The chapter concludes by describing the methods for statistical learning and construction of the classifiers used in this work.

Chapter 7 describes the methods used for the performance evaluation and the results of the experiments. The chapter starts by establishing the performance measure used, and continues by providing details on the evaluation of the results. Following this, the results of the experiments are presented separately for the sign language and news broadcast experiments. A qualitative evaluation is also provided, taking as reference a linguistic annotation of the same video material. The chapter concludes with discussion of the obtained results.

The conclusions of this work are presented in Chapter 8. Detailed quantitative results, as well as graphical results for each tested video are included in Appendix A and B, respectively.

Chapter 2

Background

The eyebrows, the eyes, the nose, and the mouth (here referred to as *facial elements*) have been a relevant topic in psychological studies where they are used to interpret and understand emotional cues in human communication. This work studies those same elements in the context of language at its most abstract level: the meaningful exchange of information between living beings using complex systems of communication. This complex communication system can be articulated with verbal forms (patterns of sound forming words for example) and non-verbal forms (such as patterns of body movements).

Facial elements are prominent as non-verbal language articulators, specially during face-to-face human communication where verbal and non-verbal articulations co-occur. The different combinations of movement patterns in facial elements, known as *facial expressions*, convey information in connection to the language used.

Linguistic components such as rhythm, intonation, and stress (jointly known as *prosody*) are essential to human communication. In spoken language, prosody is commonly articulated acoustically and supported by facial expressions functioning as emotional modulators. In sign languages the prosodic system involves the articulations from the face, hands, head, and torso [75], which can also have grammatical meaning. Languages (either spoken, signed, or written) present strictly emotional cues that may be consciously or unconsciously expressed, known as *paralinguistic* properties.

This chapter starts by discussing the presence of facial elements in sign language communication as well as the ongoing research in the area. It continues by introducing research of facial elements during communication in news broadcasts. The last section describes the link between facial expression analysis and openness estimation, and also provides information of previous similar research for each facial element studied.

2.1 Facial elements in sign language

Languages have emerged as an instrument used by humans to express themselves. In everyday interactions the vast majority of humans rely on communication through spoken languages. Learning a spoken language is dependent on the ability to hear and repeat acoustic patterns (spoken words), however this is not always possible when the person has some hearing or speaking impediment. In hearing or speaking impaired communities, sign languages (SL) have emerged as a language without the necessity of an acoustic channel of communication, using a visual instead of an acoustic system of codification. Essentially, SLs use hands, body and face to articulate signs and convey meaning.

The study and research of sign languages started in the 1950's, reaching global interest in the 1990's. In recent years the creation of corpus projects for various SLs has allowed the study of SL from new perspectives. Advancement in computer technology and the ability to use digital video have also contributed to the development of SL research.

The European Commission considers SLs as a valuable part of Europe's multilingual diversity, for every spoken language there is at least one corresponding signed language in existence. A number of European countries have taken steps to recognize in their constitution the official status of their sign languages: Finland (1995), Slovak Republic (1995), Portugal (1997), Czech Republic (1998 & 2008), Austria (2005), and Spain (2007). It is estimated that there are about 900,000 SL users in Europe, including the family and friends of the deaf or hearing impaired. The World Federation of the Deaf estimates the number of deaf people in the world to be 72 million [26]. In Finland alone, the number of signers who use Finnish Sign Language (FinSL) is estimated to be 14,000 of which 4,000 to 5,000 are deaf or hard of hearing, and 6,000 to 9,000 are hearing persons. Approximately 300 people use Finland-Swedish Sign Language, of which half are deaf [29].

2.1.1 Non-manual markers

In SL, signs are neither gesture renderings of spoken words nor mime. A sign expresses an action, concept, or thing, and is generally composed of five elements: the hand shape, location of the hand in space, movement, orientation of the palm, and *non-manuals* (NM) [75]. Although hands play a significant role in articulating utterances, SL does not rely on them exclusively. Signers actively use other parts of their body during communication. These movements produced by other parts of the body are referred to as NM.

NM markers in SL are articulations produced by body elements other than hands or arms that convey linguistic meaning. It is acknowledged in the SL community [37, 74] that NM markers can be produced by facial expressions and the following articulators: the upper part of the body, head, mouth, cheeks, eyes, and eyebrows. Manual and NM markers can be expressed simultaneously by signers since the medium through which SL is transmitted is visual, and the brain can process multiple visual stimuli simultaneously.

2.1.2 Function and relevance of non-manual markers

Recent SL research have shown an increased interest in the topic of NM articulators, this is the result of the strong evidence of the significant role of some NM articulations for the SL grammar [37]. Such articulations are used by signers to express prosodic and pragmatic functions, as well as being sentence modifiers. For example, NMs can be employed by syntax in the cases of negation, affirmation and interrogatives. In a similar manner they can be employed by topicalization, conditionals, and agreement. In [74] it is argued that there should be a differentiation between purely emotional facial expressions (and gestures) and linguistically meaningful NM markers.

The prosodic system in SL employs mixtures of manual and non-manual articulations [75]: rhythm is expressed mainly by the hands and arms, while intonation relies on facial movements. Stress also uses manual movements that can be enhanced by body leans. However, the function and significance of NMs are specific to the SL studied.

Intonation in SL commonly employs the upper section of the face (eyebrows, eyes) as markers, including head leans in some situations [75]. Facial articulators can modulate intonation in such a way that they modify the meaning of sentences. In the case of the mouth, in [74] it is suggested that some mouthings (vocalizations, sometimes partial, of spoken words) may be morphological in nature, altering the word to which they are applied as adverbs or adjectives. In some SLs mouthings occur more frequently than in others, and more often with nouns than with verbs. Mouthings are believed to be an expression of contact (code-mixing) with spoken language.

As noted in [74], studies have shown that during SL communication signers do not focus attention on the hands, but instead on the other person's face since that is a valuable source for grammatical and prosodic information. A grammatical function in SL, for example, include the NM expression of head shake to indicate sentence polarity (positive or negative statement). An example of a prosodic function in SL includes head and body leans while signing, in a similar manner to intonation in spoken language.

Not all NM articulations are strictly prosodic, furthermore, the prosodic

system in SL is not solely conveyed by NM articulators. Prosodic NM markers can also happen outside the SL grammar [75]. This supports the notion that face and body movements (outside the NM markers) can also have paralinguistic properties.

In SL it is possible that several NM markers occur simultaneously while signing, this is known as *layering*. Furthermore, blinks and head thrusts are commonly used in some SLs to mark the edges of the prosodic domain. For example in American Sign Language (ASL) in [97] was observed a modulation of eyebrow, eye, and head movements with varying signing rates, concluding that NMs are part of the language planning and production process. Similarly, in [73] evidence was reported that mouthings and mouth gestures have a role in discourse organization, noting the need for a transcription system to annotate mouth movement patterns.

More evidence on the relevance of NM markers can be found in FinSL studies, where it has been argued that upper body, head, and mouth movements express grammatical functions and phonology [42]. In [41], for example, it is shown how NM markers aid to structure sentences in FinSL, where topics are marked prosodically by eyebrow and eye movements.

2.2 Facial elements in news broadcast

Analysis of news broadcast video material has been a common research topic in content-based retrieval of videos. In workshops dedicated to video content analysis like TRECVID [67], news broadcasts in several languages are used as datasets for experiments in automatic segmentation, indexing, tagging, and speaker identification, amongst other applications. Techniques for event analysis, video indexing, and retrieval are of particular interest for news broadcast archival where a large collection of videos is available. While these techniques are not constrained to a specific type of video domain, only applications in news broadcasts are reviewed.

News video analysis includes speech recognition for automatic transcription, text recognition from descriptive tags or text snippets for information extraction, and object and audio recognition for contextual analysis [39]. Moreover, in multimodal analysis, video, audio, and file meta information are mixed for speaker identification, segmentation of scenes, association of related segments, and genre classification of scenes [4].

Identification of facial elements and their states in news broadcasts is mostly employed for face recognition and emotion identification. Applications of face recognition include retrieval of segments with specific participants [68], or recurrently appearing individuals within crowds of related

scenes [7]. Lip reading has been studied as an aid to speech recognition tasks [39] and can be extended to aid in speaker identification and automatic captioning.

Social Signal Processing (SSP) [92] is an emerging research topic in video analysis that focuses on the identification of emotional interactions. In SSP speech analysis, computer vision, and biometry are used to identify social behaviors. SSP experiments have included social network analysis in news videos, where the objective is to detect the behavioral role of the participating subjects. Recognition of roles in conversations or meetings, such as the dominant individual within a group, have mostly used non-verbal and paralinguistic cues as features [91].

An evaluation of the function of emotional expression of news journalists suggests that journalists [71] acknowledge the increasing importance of emotion as a medium to facilitate communication. Emotion can also be used as a tool to route and influence opinions [38]. The objective for facial expression analysis and state estimation of facial elements in news videos is similar as for non-manuals in sign language: automatic annotation of facial movements during communication.

2.3 Automatic estimation of facial states

Besides eye and mouth openness, the automatic estimation of eyebrow movements is also considered as part of this work. *Eyebrow position* refers to the of location of the eyebrows (raised, neutral or lowered for example); *eye openness* refers to the degree of aperture of the eyelids; and *mouth state* refers to the shape formed by the contour of the mouth while being either open, closed or in an intermediate position. Sequences of transitions between states are called *actions*; for example, a blinking action is a transition between eye states from open to closed and to open again.

As previously discussed, facial expression analysis studies the relationship between states of facial elements with the purpose of discovering non-verbal patterns of communication. For this purpose in [25] a classification system and a set of measurements for movements of facial elements named *Facial Action Coding System* (FACS) was introduced. FACS encodes a specific set of actions performed by one or more muscles and defines them as *Action Units* (AU). The AUs enable quantitative analysis of facial expressions and also of facial elements individually. In the SL domain there is no known standard description of mouth morphemes (gestures) or mouthings patterns, thus AU codification of linguistically significant mouth states is not yet available.

Studies in the state identification and tracking of individual facial ele-

ments are strongly related to facial expression analysis [88]. In this context, a project initiative tailored to estimate NM markers using an approach similar to that of facial expression analysis has been reported in [64] for a defined set of linguistically significant facial movements. State and action identification of isolated eyebrow, eye, and mouth elements commonly explore nonverbal communication other than emotion expression. Individual AUs have been used to code states and actions for this purpose. Identification of AUs can employ geometric measurements of facial characteristics or appearance modeling of the studied element.

Motion capture devices have recently been explored as analysis tools for pointing signs [81]. Passive methods of estimation (like non-intrusive video) are typically preferred over active methods (3D motion capture for example) since the former allow for more natural movements during signing. Non-intrusive methods have proven to be qualitatively as reliable as those using 3D motion capture [46], validating and encouraging their development.

Studies of facial elements and the estimation of their states are essential for the future applications of automatic annotations of SL videos. Among these possible applications are sign-to-text and sign-to-speech translation systems. Progress towards a comprehensive translation system that incorporates NM markers estimation has been reported by some projects [13, 24], however the maturity of such systems is still low. Efforts to develop standard procedures to analyze and prepare automatic translation systems, including estimation of NM markers, are still being discussed [85].

2.3.1 Eyebrow position

Research on the estimation of eyebrow position has predominantly been carried out for expression recognition in conjunction with eyes and mouth. Isolated studies for eyebrow estimation are very scarce, however in recent years the SL community's interest for precise estimation of eyebrow position has increased. Early studies in eyebrow movements and their relevance for emotion detection demonstrated that some facial expressions can be identified by eyebrow position alone [10, 25]. Later research has given more attention to studies for precise methods of eyebrow movement detection [23, 56]. In the context of SL, eyebrow position creates emotional intonation and it is also related to syntax, and in some occasions to syntax and intonation at the same time.

A system trained to detect NM markers with eyebrow position, and other facial NM markers in ASL was reported in [56]. An eyebrow position detection system built as a communication interface to replace the clicking action of the mouse peripheral in a computer was described in [34]. In a controlled

environment the system successfully detected eyebrow raises using the distance of the eyebrows from the eyes. Recently, in [23] a study of eyebrow position estimation using images with motion blur, varying illumination, and varying facial expressions was performed. The same work demonstrated the feasibility of a method for biometric recognition and gender identification of individuals using shape-based eyebrow features even under non-ideal conditions.

2.3.2 Eye openness

Estimation of eye openness and blinking actions in video have been of special interest for hypo-vigilance (drowsiness) detection in a wide range of applications [35]. Blinking has also been studied as a non-verbal communication tool, where the pattern and frequency of eye blinks can be related to emotional responses like stress and deceptive behavior [65]. The SL community is interested in voluntary blinking and squinting detection due to the significance of this actions during signing.

Eye-based drowsiness detection from video sources has been studied for applications in the automotive industry, to develop technology to minimize the risk of car drivers falling asleep while driving. For this, *active* approaches where specialized cameras or equipment are mounted inside the car are common. In active approaches cameras are typically placed in front of the driver (for example in the steering wheel or near the sun visor) to acquire video data [35]. Research with stereo cameras and infrared (IR) sensors, high frame rate cameras [77], and special IR cameras [36] have produced good results, even with varying light conditions (day, night, and changes caused by other vehicles' lights).

However, methods relying on active approaches can not in general be used with videos recorded without the mentioned special equipment. Blink detection from video sources recorded with regular cameras has been benchmarked against electrooculography (EOG) methods, and it has been demonstrated that robust results can be achieved [76, 78]. Blink patterns have also been studied besides car driver hypo-vigilance in unconstrained situations [8].

Studies of eye openness and blinking estimation methods using SL videos are scarce. However, blinking estimation methods in general can be employed in SL videos without any special consideration. Recent approaches employ various methods including pixel color similarity [99], descriptors of image features [51], appearance modeling [5], and shape modeling of the eye [95]. Blinking pattern detection in communication has been studied for the development of computer interfaces [50], and to explore eye dynamics during computer use to recognize activities [12, 21].

2.3.3 Mouth state

Research to estimate mouth states has given attention primarily to gesture recognition, lip reading, and hypo-vigilance to a lesser extent [31]. The mouth has a broader range of possible states than other facial elements. Mouth state can be generally characterized by the lips' position (two degrees of movement freedom), the lips' protrusion or contraction, visibility of the teeth, and visibility of the tongue. In facial expression analysis mouth states can describe basic emotions, for example a smile describes happiness, without the need of other visual cues. Considering mouth states produced by speech only, it is possible to understand vocalizations without hearing (lip reading) due to the mouth state patterns produced while speaking.

Estimation methods aimed to aid lip reading typically extract the shape produced by the lips' outer boundaries. Recent methods also follow the lips' inner boundaries to better separate lips from the tongue or teeth in the case they are visible in the image [86]. Lips' boundaries can be estimated with several shape, appearance, and hybrid methods [31]. The shape of the lips can be employed to improve the accuracy in speech recognition tasks [52].

Thresholding methods are a common approach to extract the contours of the lips. The segmentation can be achieved by building statistical models for the color distribution of the lips or by histogram thresholding. Both color segmentation methods [82, 96], as well as histogram thresholding methods [69, 70] are under active research. A common challenge for color segmentation is to isolate lip pixels when the tongue is visible. Tongue identification by color segmentation typically uses more advanced methods.

While color is a strong cue in mouth detection, and in general face detection as well, the used color space determines the suitability of this approach. In addition to RGB (red, green, blue), also other color spaces such as HSV (hue, saturation, intensity) and YCbCr (luminance, blue chroma difference, red chroma difference) are often used for feature detection [55]. Color transformations have also been proved to be a beneficial aid in lip segmentation [28, 57].

The extraction of lip contours using segmentation methods is not always precise. Nevertheless, the contours can be used to define the mouth's region of interest for boosting other methods. Changes in illumination in the mouth region and the presence of tongue or teeth can significantly affect the segmentation results. Illumination normalization can be achieved by the use of specialized filters: for color images via *color constancy* filters [40], or for gray-scale images via *illumination invariant* filters [43].

Chapter 3

Facial state categorization

This chapter details the proposed facial element state categorization and gives an overview of the video material used in the experiments. The primary objectives of this chapter are: to provide a common definition and terminology for openness estimation from imagery for each facial element (eyebrow, eye, mouth), and to introduce the video material that was manually annotated for the experiments according to the presented categorization. Here openness is defined in terms of *states*: the degree of openness in a point in time. Transitions between states are considered *actions*.

The proposed facial state categorization utilizes quantized states for each facial element (eyebrow, eye, or mouth observation) given a video frame. The states can be further divided into *absolute* or *progressive* types: absolute states are binary in nature and can be defined as either open or closed whereas progressive states include intermediate steps between the open and closed states. A summary of the proposed categorizations for each facial element can be seen in Table 3.1. Each category is assigned an identification number for use in both experiment and evaluation phases.

The first section of this chapter discusses the categorization for eyebrow position: the relative position of the eyebrow in relation to its neutral location in the face. The second section studies eye openness: the distance between eyelids and the functional differences between the most common eyelid distances. The third section focuses on mouth state: the location of mouth corners, upper and lower boundaries of the mouth in the face and the shape they describe. The final section of this chapter introduces the source material for the experiments, and describes how the state categorization was employed to produce the ground truth annotation.

	Eyebrow	Eye	V Mouth	H Mouth	Mouth
Absolute	0: neutral 1: shifted	0: closed 1: open	0: closed 1: open	0: neutral 1: shifted	0: closed 1: open
Progressive	0: down 1: neutral 2: raised	0: closed 1: squint 2: open 3: wide	0: closed 1: open 2: wide	0: relaxed 1: narrow 2: wide	0: closed, relaxed 1: closed, narrow 2: closed, wide 3: open, relaxed 4: open, narrow 5: open, wide 6: wide, relaxed 7: wide, narrow 8: wide, wide

Table 3.1: Summary of the categorization values for each facial element. "V" stands for vertical, while "H" stands for horizontal. The number assigned to each state is its corresponding numerical value.

3.1 Eyebrow position

The absolute eyebrow states consists of the positions {*neutral*, *shifted*}. Here the *shifted* position describes any eyebrow movements outside the *neutral* state. The eyebrow is considered shifted when it changes position from a known neutral position, defined by the location of the eyebrow with respect to a detected landmark point in the face. The progressive states {*raised*, *down*, *neutral*} make a more precise distinction between the eyebrow positions. The identified actions for the eyebrow are limited to {*neutral*, *motion*}.

3.2 Eye openness

The degree of *eye openness* can be categorized according to the position of the upper eyelid in relation to the lower eyelid. The proposed categorization system according to eyelid location can be seen in Figure 3.1. In this work the states of the eye are quantized to match the schema of Figure 3.1 with a few simplifications. The absolute states for eye openness are {*open*, *closed*} positions, while the progressive states are {*wide*, *open*, *squint*, *closed*}. In terms of eyelid separation, an occluded or transitional eye position is interpreted as a *squint* state.

The actions identified for the eyes are {*blink*, *squint*, *wide*}. A *blink* is a transition of states between *open* $\rightarrow$ *closed* $\rightarrow$ *open*. The *squint* and *wide* actions are similarly defined as *open* $\rightarrow$ *squint* $\rightarrow$ *open* and *open* $\rightarrow$ *wide* $\rightarrow$ *open* respectively. The duration of the actions does not change the

(a) Wide open (b) Normal (c) Occluded (d) Transitional (e) Closed

Figure 3.1: Transition of states in an eye closure sequence. The eye opening pattern is longer in duration. It can be seen that in state (d) the eyelid covers the pupil (no vision), but it is regarded as partially open in the eyelid distance categorization.

prototypical definition, meaning that one prolonged blink transition is still considered to be a blink.

In the definition of openness based on the eyelid distance, a *closed* state usually implies that the eyelid covers the pupil (no vision), however the pupil can still be covered while the eyelids are not in a complete *closed* state. This scenario extends to the definition of *squint*, where squinting would indicate that the person closes the eyelids near the maximum point while retaining vision, however the eyelid distance to achieve this position can be produced while blocking the pupil. For the sake of simplicity, only the position of the eyelids is considered here in the definition of openness.

3.3 Mouth state

The shape described by the outer boundary of the lips is considered sufficient in this work for the assessment of the mouth states. The outer boundary of the lips is modeled as having two degrees of freedom: horizontal and vertical displacements. In Figure 3.2 the proposed quantization of mouth states is shown. This approach is geometrical, meaning that no prior assumption of mouth movements is made.

The absolute states for the mouth are {*open*, *closed*}. In the case of the progressive states a total of 3 vertical and 3 horizontal positions are defined: {*closed*, *open*, *wide*} in the vertical direction, and {*relaxed*, *narrow*, *wide*} in the horizontal direction. The combined mouth positions total in 9 possible states. The presence of tongue or teeth and their position is neither studied nor evaluated. Mouth actions are described by the non-linguistic elements {*opening*, *closing*, *vocalize*}. Here *vocalize* is manifested as repeated transitions between horizontal and vertical progressive mouth states.

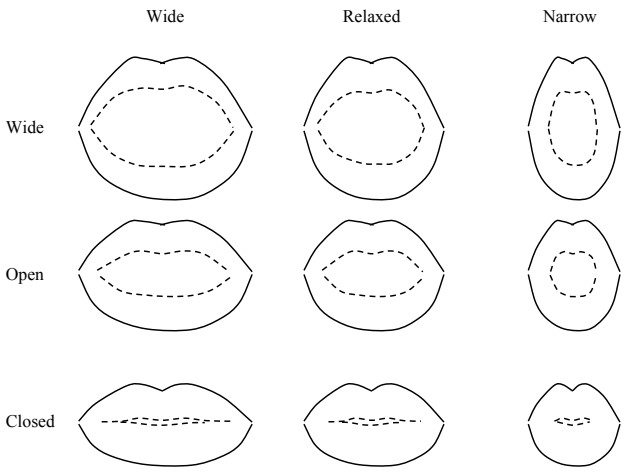


Figure 3.2: Mouth states defined for this work. The states consider the outer border of the lips with horizontal and vertical displacements.

Chapter 4

Landmark detection

This chapter focuses on the algorithms for facial landmark detection, used later to compute the features described in Chapter 5. The landmark detection methods take as input a face region and produce a set of point locations for eyebrows, eye corners, eyelids, mouth corners, and upper and lower lip boundaries.

The presented algorithms rely on accurate face detection. The face detector may report false positives, and to account for this, in `SLMotion` [47] a pruning and interpolation step is performed as postprocessing. In pruning, skin detection is executed: if the amount of skin-like pixels is below a threshold, the face is discarded. The interpolation step fills in undetected faces based on the location of the closest previously detected face in the video frames. Face detection employs `OpenCV`'s [11] implementation of the Viola-Jones rapid object detector [93]. The detected face regions are considered to be accurate, even though it is known that this is not always the case.

The facial landmark detection methods presented in this chapter can be grouped by their use of *segmentation*. The segmentation step divides the face into regions of interest (RoIs) in order to detect landmarks more accurately. Furthermore, different estimation methods are used for each type of face RoIs. This work explores two face region segmentation methods for the proposed landmark detection algorithms.

The content of this chapter is as follows: Section 4.1 introduces the face region segmentation schemes that produce image regions for each facial element. Section 4.2 continues by detailing the preprocessing algorithms employed for landmark detection. The preprocessing algorithms are used to perform either gray-scale photometric normalization, color segmentation, or color normalization. Section 4.3 describes the landmark detection algorithms for eyebrow, eye, and mouth regions; finalizing with a description of a recent third-party method of landmark detection used for comparison.

4.1 Face region segmentation

The detected rectangular face region $\boldsymbol{I}_{\mathrm{F}}$ is the starting point for landmark detection. It is delineated by two coordinate corner points, $(x_0^f, y_0^f), (x_1^f, y_1^f)$. The region $\boldsymbol{I}_{\mathrm{F}}$ is always a square shaped with size $N \times N$ pixels, where N can vary across frames. The landmark coordinates are normalized into the range of $(x, y) \in [0, 1] \times [0, 1]$ with respect to the face region, to account for differences in the face location and for size variation across frames.

Two segmentation methods are studied; each one divides the face region into eyebrow, eye, and mouth RoIs. The first segmentation method (**G1**) is based on *face proportion constraints* [33], while the second method (**G2**) employs the precomputed landmarks from the `flandmark` package [90], a facial feature detection software based on *Deformable Part Models*. Both segmentation methods produce the same type of output: a set of image RoIs to be used as input for the proposed landmark detection methods. The RoIs do not have to be precise. However, the extracted RoIs should leave interfering elements such as hair, nose, or clothing outside the region.

4.1.1 Segmentation by face proportions

The first face segmentation method, referred here as **G1**, employs the cues of *facial symmetry* and *face proportions* to approximate the location of facial RoIs. The face proportion principle states that features in the human face lie within small variations of standard locations, showing patterns according to race and gender [80]. Facial symmetry and face proportion constraints have been researched mainly in the medical and behavioral fields where proof of a strong relationship between face symmetry and perceived aesthetics and healthiness has been demonstrated. However, facial symmetry and proportion have not been very prominent for other applications.

In Figure 4.1 the outline in red dots represents the face region, while the green solid line outlines are the face RoIs. Here the red dots are the estimated facial landmarks and the green dots are the geometric center of the landmarks that they describe. In the segmentation method **G1**, the face RoIs are estimated as:

$$\boldsymbol{I}^{u,r} = \left\{ (x, y) \in \mathbb{N}^2 \,|\, x_0^f \leq x \leq x_0^f \tfrac{N}{2},\ y_0^f \leq y \leq y_0^f + \tfrac{2N}{5} \right\} \tag{4.1}$$

$$\boldsymbol{I}^{u,l} = \left\{ (x, y) \in \mathbb{N}^2 \,|\, x_0^f + \tfrac{N}{2} \leq x \leq x_0^f + N,\ y_0^f \leq y \leq y_0^f + \tfrac{2N}{5} \right\}. \tag{4.2}$$

Here $\boldsymbol{I}^{u,r}$ represents the upper right image area containing the right eyebrow and eye, with (x_0^f, x_0^f) the face region top-left corner point; likewise for

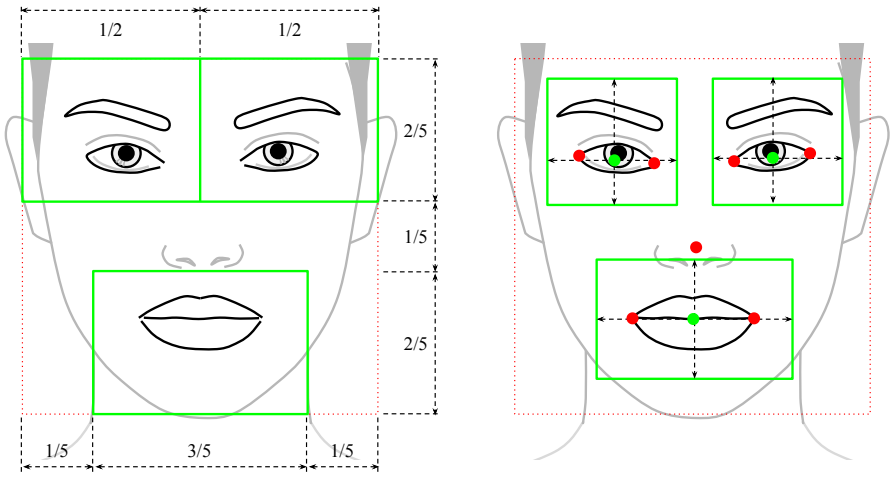


Figure 4.1: Facial RoI extraction methods in comparison. Left: Example application of face proportion and face symmetry method **G1**. Right: Example application of `flandmark`-based area extraction **G2**.

$\boldsymbol{I}^{u,l}$ the upper left image area. The mouth RoI is estimated as:

$$\boldsymbol{I}^{v} = \left\{(x,y) \in \mathbb{N}^2 \,|\, x_0^f + \tfrac{1N}{5} \leq x \leq x_0^f + \tfrac{3N}{5},\, y_0^f + \tfrac{4N}{5} \leq y \leq y_0^f + N\right\}. \tag{4.3}$$

The limits of the RoIs were obtained from the inner face proportions suggested in [33] and adjusted according to visual inspection on random samples of the training data.

4.1.2 Segmentation by face landmarks

The second face segmentation method, referred to as **G2**, employs the landmarks extracted with the `flandmark` software for coarse estimation of individual face elements' RoIs. The eight (x, y) landmark coordinate points from `flandmark` are not directly used due to the presence of noise in the location of the features, observed in previous research [60]. The output of **G2** is similar to that of **G1**, but the area sizes differ.

To segment the right eye, **G2** computes the center $\boldsymbol{c}^r = (c_x^r, c_y^r)$ of the eye corner landmark locations. The same approach is employed for the left eye and mouth landmarks. In Figure 4.1, an example of the extraction is shown, here the face landmarks are red dots, while the computed centers appear as green dots. Using the computed centers, a region is extracted for each facial element with an estimated displacement factor.

The implemented displacement factors were chosen by sampling random frames and computing the average value of coordinate displacement required to cover each facial element. The displacement factor $\boldsymbol{\delta}^u = (\delta_x^u, \delta_y^u)$ for the eye regions is $(1.1, 1.5)$, and the displacement factor $\boldsymbol{\delta}^v = (\delta_x^v, \delta_y^v)$ for the mouth region is $(1.6, 1.3)$. The RoIs are estimated as:

$$\boldsymbol{I}^{u,r} = \left\{(x,y) \in \mathbb{N}^2 \,|\, c_x^r - \delta_x^u \leq x \leq c_x^r + \delta_x^u,\, c_y^r - \delta_y^u \leq y \leq c_y^r + \delta_y^u\right\} \quad (4.4)$$

$$\boldsymbol{I}^{u,l} = \left\{(x,y) \in \mathbb{N}^2 \,|\, c_x^l - \delta_x^u \leq x \leq c_x^l + \delta_x^u,\, c_y^l - \delta_y^u \leq y \leq c_y^l + \delta_y^u\right\} \quad (4.5)$$

$$\boldsymbol{I}^{v} = \left\{(x,y) \in \mathbb{N}^2 \,|\, c_x^v - \delta_x^v \leq x \leq c_x^v + \delta_x^v,\, c_y^v - \delta_y^v \leq y \leq c_y^v + \delta_y^v\right\}. \quad (4.6)$$

4.2 Preprocessing algorithms

The landmark detection algorithms share some preprocessing procedures that are described in this section for later reference (when used). The preprocessing algorithms are used to enhance the face image $\boldsymbol{I}_F$ in order to aid in the estimation of skin and reduce the influence of shadows. Three preprocessing algorithms are used across the landmark detection methods for this:

1. Photometric normalization using *single scale retinex* (SSR) [43].
2. Skin color mask estimation using *color segmentation rules* [49].
3. Color normalization based on *gray world assumption* [40].

The SSR illumination compensation algorithm performs photometric normalization by minimizing the effect of varying illumination in an image. The SSR is a retinex-based method, it combines the working principles of the retina and cortex into a single theory to explain color constancy [53]. The SSR algorithm extracts the *illumination invariant* reflectance component L from the gray-scale image I_G by:

$$L(x,y) = \alpha \log\left(\frac{I_G(x,y)}{I_G(x,y)F(x,y)}\right) - \beta \quad (4.7)$$

$$F(x,y) = k \exp\left(\frac{-(x^2+y^2)}{c^2}\right), \quad (4.8)$$

where α is a gain factor and β an offset parameter; k in the Gaussian kernel $F(\cdot)$ is a normalization factor, and c controls the area covered by the kernel. The reflectance image L minimizes the influence of shadows in the images. The SSR filter used in the experiments is implemented in the `INToolbox` library [94].

The second preprocessing algorithm utilizes a simple skin mask estimation method based on *color segmentation rules*. The segmentation rules mark each

image pixel as being skin or not, thus forming a binary mask. The skin color mask is estimated in the RGB color space by the combination of two rule sets. The first rule set describes skin color in uniform daylight illumination:

$$\begin{aligned}&(R > 95) \wedge (G > 40) \wedge (B > 20) \wedge \\ &(\max\{R, G, B\} - \min\{R, G, B\} > 15) \wedge \\ &(|R - G| > 15) \wedge (R > G) \wedge (R > B).\end{aligned} \tag{4.9}$$

The second rule covers flashlight or daylight lateral illumination:

$$\begin{aligned}&(R > 220) \wedge (G > 210) \wedge (B > 170) \wedge \\ &(|R - G| \leq 15) \wedge (R > B) \wedge (G > B).\end{aligned} \tag{4.10}$$

The complete skin color segmentation creates the output mask by joining (logical OR) the two sets of rules in Equation (4.9) and Equation (4.10). Finally, since darker areas of the face, such as eyebrows, eyes and mouth cavity pixels, are marked as non-skin by the rules, morphological filling is used to regain areas within the face.

The third preprocessing algorithm is related to color normalization in images to achieve *color constancy*. Color constancy means the possibility of constant color identification of an object even under varying illumination conditions. It can be achieved using the *gray world* algorithm; it is based on the assumption that colors in an image average to a neutral gray. This assumption holds only if the image has a pertinent color distribution. Even though the assumption of tendency of colors to sum up to gray may not always hold, the approach is beneficial for the SL video data due to the gray background in the used video data.

The algorithm estimates the average color of the image and compares it with a neutral gray to determine the color of the original object. The implemented method of normalization follows the same principle by computing the mean of each color channel and scaling the channels according the gray average. The scale factors α, β, γ of each of the RGB channels are computed as:

$$(\alpha, \beta, \gamma) = \left(\frac{\frac{1}{3}\sum_i \mu_i}{\mu_R}, \frac{\frac{1}{3}\sum_i \mu_i}{\mu_G}, \frac{\frac{1}{3}\sum_i \mu_i}{\mu_B} \right) \tag{4.11}$$

where μ_i is the average intensity in the channel $i \in \{R, G, B\}$. The scaling is given by:

$$(R', G', B') = (\alpha R, \beta G, \gamma B), \tag{4.12}$$

with (R', G', B') being the new normalized color image channels according to the gray world algorithm.

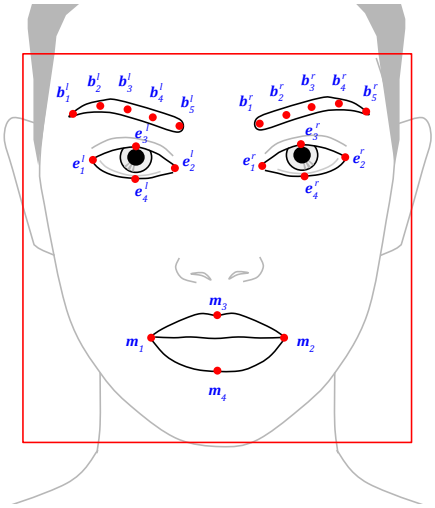


Figure 4.2: Red square: Estimated face region $\boldsymbol{I}_{\mathrm{F}}$. Red dots: Target facial landmarks. The landmarks can be computed using different algorithms.

Segmentation	Element	Landmark detection algorithm
Face Proportions (**G1**) Face Landmarks (**G2**)	Eyebrow	Projection (**Pr**)
	Eye	Radial symmetry (**Cr**) Projection (**Pr**)
	Mouth	Pseudo hue Gradient (**Gr**) Pseudo hue Binary (**Bi**)
Full face image	Eyebrow	Supervised Descent (**Sd**)
	Eye	Supervised Descent (**Sd**)
	Mouth	Supervised Descent (**Sd**)

Table 4.1: Summary of the implemented algorithms for landmark detection.

4.3 Facial landmark estimation

The proposed method for estimating eyebrow position, eye openness, and mouth state is based on the construction of a set of facial landmark points as shown in Figure 4.2. These estimate the position and shape of eyebrows, eyelids (or iris height), eye corners, mouth corners, and upper and lower lip boundaries. Several landmark detection techniques are explored for each facial element as summarized in Table 4.1: oriented projections and pixel similarity for eyebrows, oriented projections and radial transform for eyes, and pseudo-hue masks for lips. For comparison, the landmarks detected using the Supervised Descent Method from the IntraFace package [98] are also considered.

The complete set of estimated landmarks consists of 22 points. In Fig-

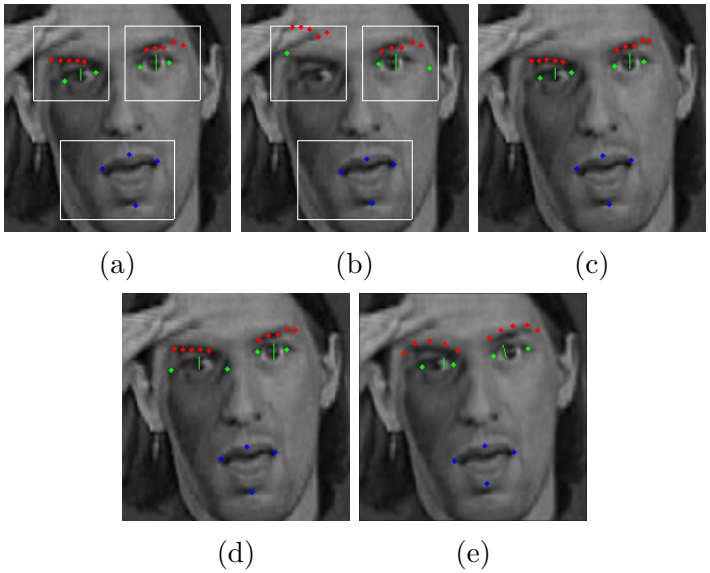


Figure 4.3: Differences in landmark locations with different detection methods. (a) **G1**, **Pr**, **Cr**, **Gr**. (b) **G1**, **Pr**, **Pr**, **Bi**. (c) **G2**, **Pr**, **Cr**, **Gr**. (d) **G2**, **Pr**, **Pr**, **Bi**. (e) **Sd**.

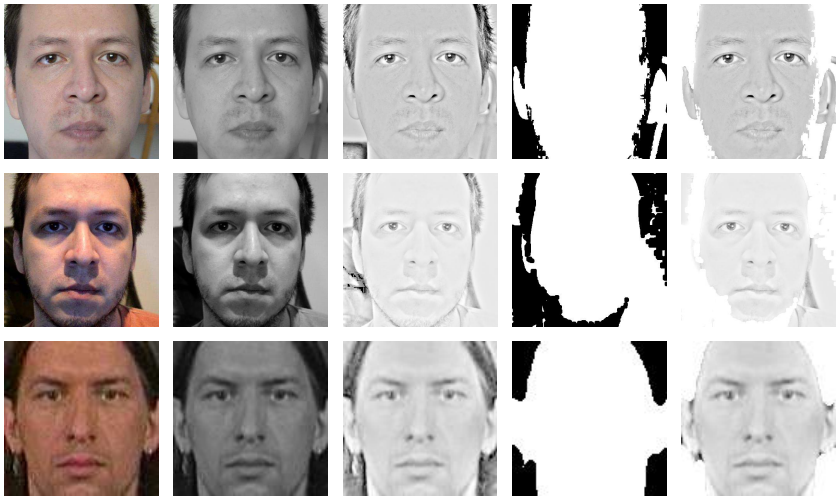
ure 4.3, the differences in landmark locations can be seen for the same video frame using the different landmark detection algorithms. The eyebrow is described with five landmarks, the eye with four, and the mouth also with four landmarks. These are used to calculate a set of geometric features which are detailed in Chapter 5. This section studies the landmark detection algorithms for each type of facial element.

4.3.1 Intensity projection

Intensity Projection (**Pr**) is an algorithm aimed at eyebrow and eye landmark estimation. The algorithm first divides the eye RoI $\boldsymbol{I}^u$ into eyebrow and eye RoIs by estimating a discriminant line. Second, landmark locations are determined for each RoI type using oriented projections. The projection-based algorithm **Pr** determines that the eyebrow and eye locations can be individually determined as local minima in the vertical one-dimensional projection of the image region $\boldsymbol{I}^u$.

To determine the discriminant line between the eyebrow and eye, the **Pr** algorithm performs the following steps:

1. Build a gray-scale image $\boldsymbol{I}_G$ from $\boldsymbol{I}_{\mathrm{F}}$.
2. Build an illumination invariant version $\boldsymbol{I}_L$ of the gray-scale RoI.
3. Extract a skin mask from the RoI and filter out the non-skin from $\boldsymbol{I}_L$.



(a) (b) (c) (d) (e)

Figure 4.4: Example of the preprocessing steps in some images. (a) Original color image. (b) Gray-scale version. (c) SSR processed image. (d) Skin mask. (e) SSR processed image with whitened non-skin.

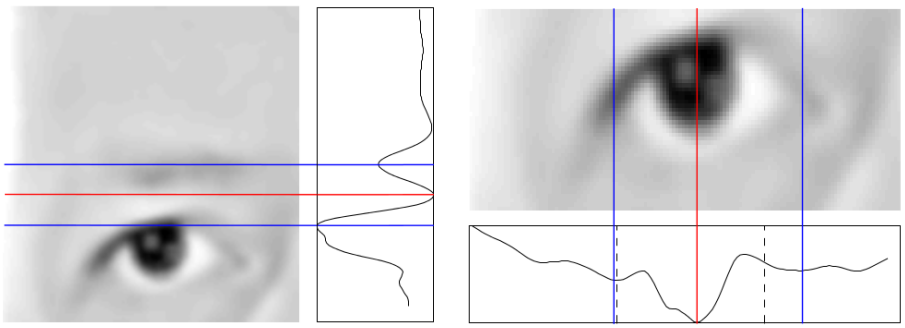
4. Extract the vertical projection of $\boldsymbol{I}^u$ from $\boldsymbol{I}_L$.
5. Estimate the discriminant line.

$\boldsymbol{I}_G$ is processed with the SSR illumination invariant filter to form $\boldsymbol{I}_L$. This is done to reduce the influence of shadows in the image. To eliminate darker non-skin areas, such as hair or background, a skin mask is also computed. The mask consists of tonal segmentation of skin-like color pixels in the image. The skin mask is used to rule out the background and hair pixels in $\boldsymbol{I}_L$. An illustration of the preprocessing steps can be seen in Figure 4.4.

The vertical projection of $\boldsymbol{I}^u$ is used to find the eyebrow and eye discriminant line. The y-coordinates of the eyebrow and eye are estimated to be the two smallest local minima as shown in Figure 4.5. The discriminant line is the global maximum between the identified local minimums. It divides $\boldsymbol{I}^u$ in the eyebrow RoI $\boldsymbol{I}^{u'}$ and the eye RoI $\boldsymbol{I}^{u''}$.

The local minima of the projections are identified using a local filter. The filter consists of a gray-scale dilation of the projection vector $\boldsymbol{x}$, with a flat linear structuring element $\boldsymbol{b}$ with 0 as its origin. The size of $\boldsymbol{b}$ determines the number of neighboring values used in the dilation. Denoting the projection vector as $\boldsymbol{x}(x)$ and the structuring function as $\boldsymbol{b}(x)$, the filter operation is defined as:

$$(\boldsymbol{x} \oplus \boldsymbol{b}) = \max_{s \in \boldsymbol{b}} \{\boldsymbol{x}(x - s)\}. \tag{4.13}$$



(a) (b)

Figure 4.5: Example of a processed eyebrow and eye RoI using intensity projection. (a) vertical projection of $\boldsymbol{I}^{u}$, in blue the eyebrow and eye estimated locations, in red the discriminant line between the eyebrow and eye. (b) horizontal projection of $\boldsymbol{I}^{u''}$, the blue lines are eye corners, the red line marks the pupil, the black dashed line delimits the search windows.

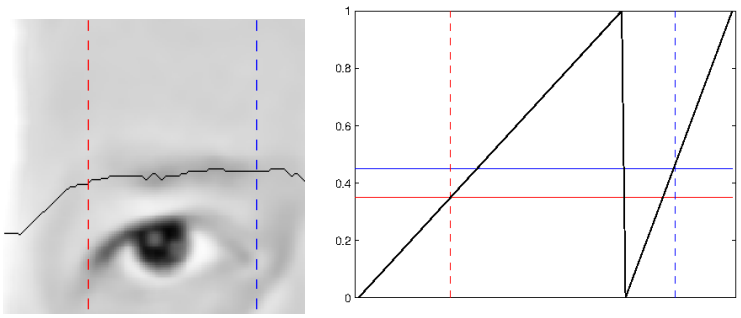
4.3.1.1 Eyebrow landmark estimation

A x-coordinate seed location (the darkest eyebrow pixel) is obtained from the global minimum of the horizontal projection of the eyebrow RoI $\boldsymbol{I}^{u'}$. From the estimated eyebrow seed location, a 1×3 search window is used to form a path towards the left and right ends of the image, consisting of pixels with the lowest intensity difference. Using the intensity values in the estimated eyebrow path, a cumulative sum is computed towards both ends and scaled to the $[0, 1]$ range. Based on a sample of the available training data, it was estimated that the innermost landmark point of the eyebrow resides where the cumulative sum exceeds 0.35. The outermost eyebrow landmark point is estimated to be at the cumulative sum value of 0.45 as seen in Figure 4.6.

It was observed that with larger cumulative sum threshold values the overall shape of the eyebrow becomes more sensitive to lighter pixels. The eyebrow shape for persons with short or dim eyebrows is better estimated with higher threshold values. With the estimated landmarks, the eyebrow path is divided in equal parts to extract the three intermediate locations of the eyebrow landmarks.

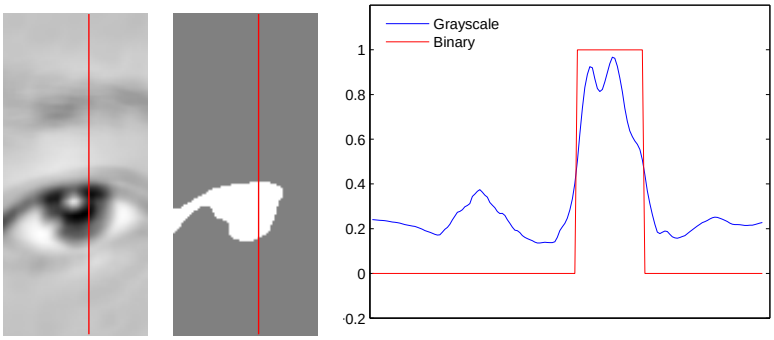
4.3.1.2 Eye landmark estimation

To estimate the locations of the eye corners, the horizontal projection of the eye RoI $\boldsymbol{I}^{u''}$ is segmented into three sections of equal size. The horizontal corner location is determined as the global minimum of the projection in the first and last sections. In **Pr** the landmarks of the iris vertical limits (iris



(a) (b)

Figure 4.6: Eyebrow path estimation example. The threshold limits $[0.35, 0.45]$ are the (red, blue) dashed lines. (a) The estimated eyebrow path appears as a black line across the image. (b) The normalized cumulative sums of the pixel intensities along the estimated eyebrow path.



(a) (b) (c)

Figure 4.7: Example of binarization and eye limit estimation. (a) $\boldsymbol{I}^u$ region demarcated by eye corners, red line shows the x-coordinate location of the estimated pupil. (b) Thresholded image after morphological filling, red line as in previous. (c) Inversed intensity values along the image column at the estimated pupil locations shown in red in (a) and (b).

height) are used instead of the eyelid landmarks. The iris x-coordinate is the global minimum of the complete horizontal projection.

The y-coordinate of each eye corner is determined as the global minimum of the image column vector at the estimated horizontal location of the eye corner. To eliminate the influence of shadows near the eye in $\boldsymbol{I}^{u''}$, pixels that are outside a radial area centered at the mean center of the eye corners are eliminated.

The eye image region $\boldsymbol{I}^u$ is processed with Otsu's global thresholding method [66] to obtain the iris limits. The binary image is morphologically filled to account for reflecting light inside the iris area. The limits of the visible iris are then determined as the image region containing the largest

number of dark pixels in the column corresponding to the estimated pupil location as shown in Figure 4.7.

4.3.2 Eye landmarks from radial symmetry

In face imagery *radial symmetry* is of special interest to locate the iris and estimate the gaze direction, since the iris and pupil show strong radial symmetry cues. Radial symmetry transforms are designed to estimate centers of areas that display some degree of circular appearance around them. This type of transforms have been studied and used to estimate the iris location with some success [59, 100]. The most common transformation employs gradients.

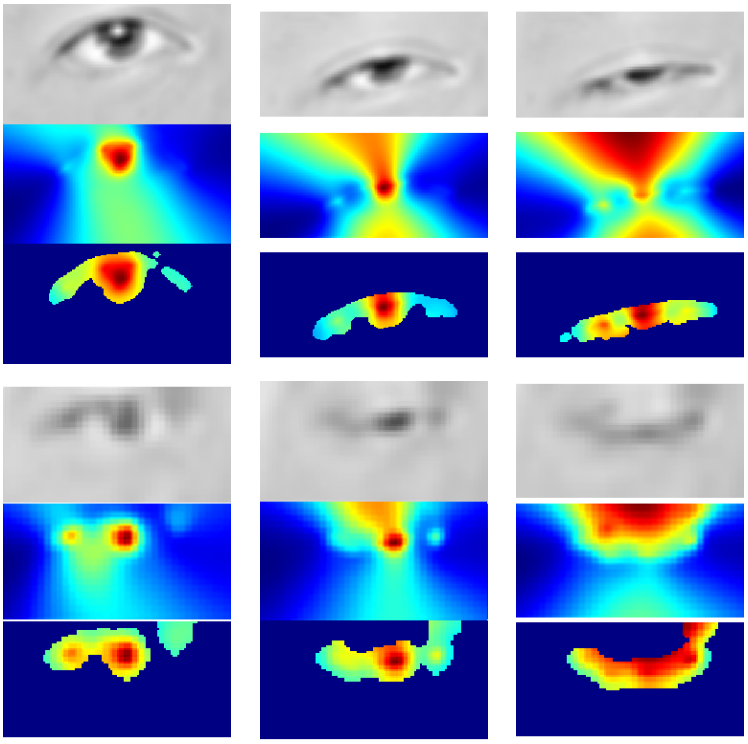
The radial symmetry transform takes an image as input, computes the vertical gradients, and evaluates all pixels as potential centers of radial shapes. The output of the transform consists of a matrix of values indicating how likely each pixel location is surrounded by a radial pattern. The absolute orientations of pixel gradients between the center of a radial shape and the pixels forming the shape should be the same, therefore any pixel location that satisfies this condition receives a high value in the output matrix.

The pupil localization algorithm presented in this section is based on the work in [89]. However, the transform did not produced suitable results for eye corner detection during initial experiments of this work. Eye corners usually do not show radial patterns. Instead, the eye corner's radial cues in the image can be attributed to compression distortion or to the use of smoothing filters. Henceforth, oriented projections are used in this work for eye corner detection.

The following steps are performed by the radial-symmetry-based **Cr** algorithm in order to detect the iris landmarks:

1. Compensate illumination in $\boldsymbol{I}^{u''}$ using the SSR filter.
2. Smooth the image region with an average filter.
3. Rule out pixels above an intensity value threshold.
4. Compute the image gradients $\boldsymbol{g}_i$, for all i pixels.
5. Estimate the radial symmetry transform matrix using $\boldsymbol{g}_i$.
6. Select the maximum of the transform matrix as the iris center location.
7. Detect eye corners via oriented projections.

The first step in the **Cr** algorithm processes the eye region $\boldsymbol{I}^{u''}$ with the SSR filter to reduce the influence of shadows (Section 4.2). The resulting



(a) (b) (c)

Figure 4.8: Processed images with **Cr** at different states. Areas with major radial symmetry appear in red. (a) Open eye. (b) Squint. (c) Closed eye.

image $\boldsymbol{I}_L$ is scaled to the $[0, 1]$ range. This step in combination with step three reduces the number of potential iris centers in the image.

For the second step, [89] suggested a Gaussian filter to reduce pixel value variations of small significance. In the other hand, this work proposes the use of an average filter to smooth the image. Gaussian filters remove less details than average filters. However, the average filter is computationally faster making it a good alternative choice. In the average filter operation, each pixel intensity value of the image is averaged within a neighborhood defined by a kernel. The implemented filter uses a default 5×5 square kernel, however, a larger 7×7 square kernel is used for face images larger than 400×400 pixels.

The third step is introduced in order to reduce the computation time of **Cr** by ruling out potential centers in the smoothed and illumination compensated image. Iris and pupil pixels appear darker and show narrower intensity value distribution than skin pixels. The interest is then only on intensity changes from low to high to focus on dark radial patterns. To take advantage of this, the search space of the algorithm is thresholded to the lowest

10% pixel intensities of the image. The threshold then limits the potential optimal center $\boldsymbol{c}^*$ to low intensity pixels only.

For the fourth step, the transform in **Cr** uses the vector field of image gradients $\boldsymbol{g}_i$ for each image location i. As suggested by [89], gradients of small magnitude are ruled out from the optimal center candidates as they typically represent monotonous regions such as skin or sclera. The image gradient is computed by the operation:

$$\boldsymbol{g}_i = \left(\frac{\partial I^{u''}(x_i, y_i)}{\partial x}, \frac{\partial I^{u''}(x_i, y_i)}{\partial y} \right). \tag{4.14}$$

The radial symmetry transform matrix (fifth step) is estimated by first scaling the gradient vector magnitudes such that $||\boldsymbol{g}_i|| = 1$. Then, for every possible center $\boldsymbol{c}_k$ of a circular element, a displacement vector $\boldsymbol{d}_i$ is computed in reference to a pixel position $\boldsymbol{x}_i, i \neq k$. $\boldsymbol{d}_i$ is also normalized to unit length to maintain equal weight for all $\boldsymbol{x}_i$. The displacement is given by

$$\boldsymbol{d}_i = \frac{\boldsymbol{x}_i - \boldsymbol{c}_k}{||\boldsymbol{x}_i - \boldsymbol{c}_k||}. \tag{4.15}$$

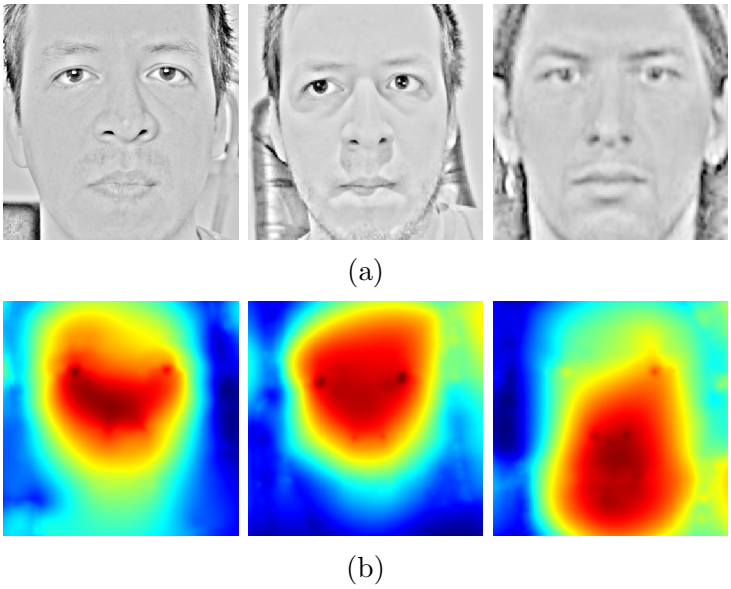
From Equation (4.15) it can be observed that for a given potential radial center $\boldsymbol{c}_k$, the displacement $\boldsymbol{d}_i$ is maximal if it has the same absolute orientation as $\boldsymbol{g}_i$ evaluated at the same position $\boldsymbol{x}_i$. Using this observation the output of the transform at each pixel location forms a value matrix where each value is given by:

$$c_k = \frac{1}{N} \sum_{i=1}^{N} w_{\boldsymbol{c}_k} (\boldsymbol{d}_i^T \boldsymbol{g}_i)^2, \tag{4.16}$$

i.e. the sum of squared dot products between the displacement and gradient vectors for all potential centers k. Here $w_{\boldsymbol{c}_k}$ is a weight factor that incorporates the prior knowledge of the eye intensity values to reduce dominance of local maximums. Since pupils usually appear darker than other regions of the eye, the weight factor is chosen so that more importance is given to dark centers than to bright centers. The optimal iris center $\boldsymbol{c}^*$ in the sixth step is selected as the global maximum $\boldsymbol{c}_k$ of the transform matrix.

The last step computes the eye corners following the estimated location of the iris using oriented projections. The eye RoI is divided in two subregions delimited by the horizontal location of the iris. Within each of the subregions, the eye corner is estimated as the global minimum of the resulting oriented projection. This follows the same principle as in **Pr**.

Examples of transform matrices are given in Figure 4.8 for three eye states. Limiting the search space does not change the locations of the optimal



(a)

(b)

Figure 4.9: Example of processed images in **Cr** with complete faces. The radial symmetry transform is influenced by the presence of nose orifices and other artifacts of light and shadow. (a) Input images. (b) Results of the radial symmetry transform.

centers. For closed eyes, using the threshold preprocessing improves the result making a closer approximation to the eyelid area. The strongest radial pattern during a squint or closed eye position may shift from the iris to the eye corners, or in some cases, eyelid shadows that show some degree of circular appearance. Light reflections in the iris also have a negative effect in the center estimation and weight factors due to gradient variations within the iris area. The influence of more than one circular pattern in the same image can be observed in Figure 4.9, where, due to eyelids or light reflectance, radial characteristics of other facial features are stronger than those of the eye.

4.3.3 Mouth landmarks from color segmentation

To estimate the landmark coordinates of the mouth, the mouth region RoI $\boldsymbol{I}^v$ is used as input. The mouth corners, and upper and lower lip landmarks are employed to extract the state of the mouth. The mouth landmarks are estimated using an approach that utilizes color normalization, transformation, thresholding and projection.

Examples of tested lip segmentation algorithms based on color features can be seen in Figure 4.10. The examples use images extracted from FinSL videos. In the videos, changes of illumination pose difficulties in segmenting upper from lower lip and make extraction of mouth landmarks unstable. The

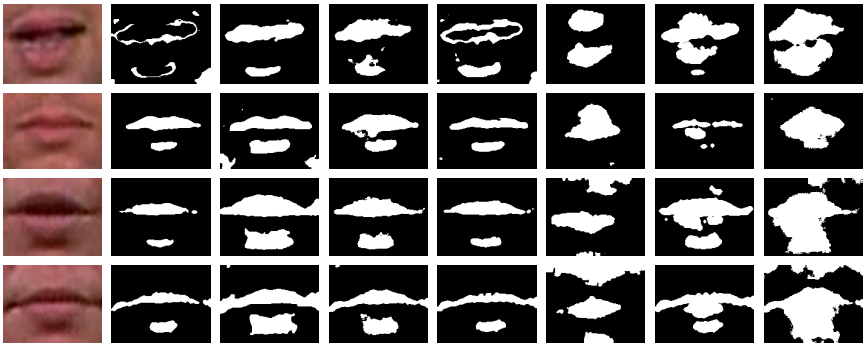


Figure 4.10: Sample lip mask estimation for different color segmentation algorithms using mouth RoIs from the FinSL videos. (a) Original image; (b) [82]; (c) [48]; (d) [69]; (e) [57]; (f) [1]; (g) [9]; (h) [14].

tested color segmentation methods use the RGB color space [48, 69, 82], the YCbCr color space [1, 57], a combination of both [9], and $I3$ from [14].

The algorithms proposed in this section are based on a color transformation by means of *pseudo hue* variations. The transformation have been studied in [27, 28] where its formulation uses only the information of the red and green channels. This allows the use of color normalization techniques since no channel value is used directly. In previous studies binary masks derived from color segmentation methods have been used for pre-labeling lip pixels; in this work the mask itself is considered strong enough to estimate the mouth landmarks. In the proposed algorithm the mouth mask is extracted from the transformed image by Otsu's thresholding method [66].

4.3.3.1 Mouth mask extraction

Two image components are used for mouth mask extraction: the *pseudo hue* component H, and the luminance component L. The mouth mask is extracted by performing these steps on the mouth RoI:

1. Normalize color using the gray world algorithm.
2. Estimate pseudo-hue and luminance components.
3. Compute vertical gradients from three combinations of pseudo-hue and luminance components.
4. Compute the composite matrix R_N using the vertical gradients.
5. Threshold the composite gradient component.

The first step proposes to account for changes in illumination in the mouth RoI in order to obtain a better segmentation. To achieve this color normalization is performed using the gray world algorithm as in Equation (4.12). The normalization is performed in all color channels simultaneously.

In the second step, color transformation is applied to enhance the contrast between skin and lip colors. The *pseudo hue* component H takes advantage of the red and green pixel value difference between lip and skin. This difference is higher for lips than for skin pixels. The pseudo-hue component H is computed by an approximation of the component U from the LUX color space [55] such that $H \approx U$:

$$H = \begin{cases} \dfrac{G}{R} & \text{if } R > G, \\ 1 & \text{otherwise.} \end{cases} \tag{4.17}$$

The luminance component L from the LUX color space is used in order to take advantage of the shadows produced by the mouth. The luminance of the oral cavity is significantly lower than luminance in the lips. This difference is useful to reduce the number of false lip regions by incorporating this information in H. The relative luminance can be computed from the RGB channels with pixel-wise operations as:

$$L = (R+1)^{0.6}(G+1)^{0.3}(B+1)^{0.1} - 1. \tag{4.18}$$

The third step employs the gradient matrix components from [28, 86]. The vertical gradients of the pseudo hue H and the luminance channel L are combined to produce three new image matrices R_{top}, R_{mid} and R_{low}. The vertical gradient operator is denoted here by ∇_y. The new images enhance the contrasting difference between the lip and skin into a set of image edges as follows:

$$R_{\text{top}} = \nabla_y \left(H_N - L_N\right) \tag{4.19}$$

$$R_{\text{mid}} = (\nabla_y H_N) L_N \tag{4.20}$$

$$R_{\text{low}} = \nabla_y H_N, \tag{4.21}$$

where L_N and H_N represents the components scaled to the $[0, 1]$ range. In R_{top} the upper lip edge becomes more apparent, while in R_{mid} the inner border gradients of the lips are the lowest, and finally in R_{low} the lower border of the lips shows stronger negative gradients. In the mid and low image gradients, values greater than zero are ruled out since they represent changes from darker to lighter that are not relevant. The result of this thresholding operation is represented by R^*.

The fourth step constructs a composite edge image R from the set of gradients previously computed. The combined edge image is given by:

$$R = R_{\text{top}} - R^*_{\text{mid}} - R^*_{\text{low}}. \tag{4.22}$$

The final step involves computing two binary masks from the gradient matrices. The first lip mask **Bi** is computed from H_N, while the second mask **Gr** is computed from R_N, the $[0, 1]$ scaled edge image R. The thresholding is performed using Otsu's algorithm. Morphological postprocessing is also performed in order to remove noise in the masks. Both lip masks **Bi** and **Gr** are post-processed in the same way.

4.3.3.2 Mask postprocessing

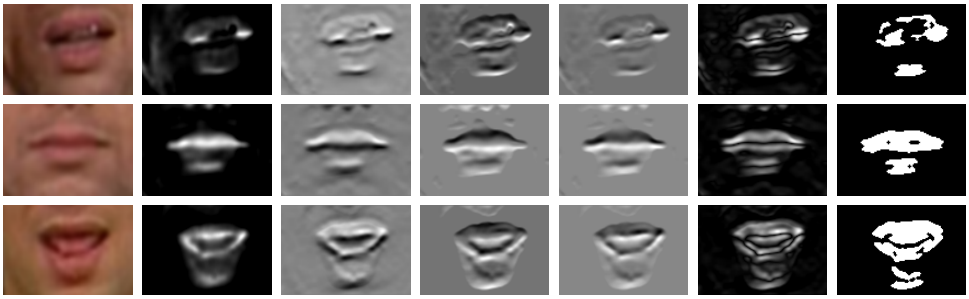
Morphological postprocessing is applied to the binary lip masks to account for falsely marked lip pixels. For this, four postprocessing steps are applied:

1. Morphological closing.
2. Oval mask filtering.
3. Removal of small components.
4. Removal of bordering components.

Morphological closing fills small gaps between the lip mask and connects marginally separated regions; the used structuring element is a disk of size 3. In *oval mask filtering* an oval mask is created and centered at the mouth RoI image $\boldsymbol{I}^v$ with its axes aligned with the image edges; lip mask pixels falling outside the oval mask are eliminated as lip pixel candidates. In the *removal of small components* step, connected components (clusters of pixels) with an individual size less than 10% of the total number of lip candidate pixels are ruled out. Finally, any pixel components connected to the image border are eliminated. In Figure 4.11 an example set of the whole process can be seen where the final lip mask is also shown.

4.3.3.3 Landmark detection

After lip mask postprocessing, the mouth corner landmarks $\{\boldsymbol{m}_1, \boldsymbol{m}_2\}$ are estimated using a horizontal projection of the lip mask. In the resulting projection the x-coordinates of the mouth corners correspond to the leftmost and rightmost occurrence of a value greater than a threshold $\tau = 2$ pixels in the projection vector. The y-coordinate for both landmarks is estimated as the median of the lip mask pixel y-coordinate locations in the column vector of the corresponding mouth corners.



(a) (b) (c) (d) (e) (f) (g)

Figure 4.11: Image examples of the intermediate steps in lip mask estimation for landmark extraction. (a) Original image; (b) Pseudo hue after color normalization; (c) R_{top}; (d) Inverted R_{mid}; (e) Inverted R_{low}; (f) R_N; (g) Lip mask **Gr** after post-processing.

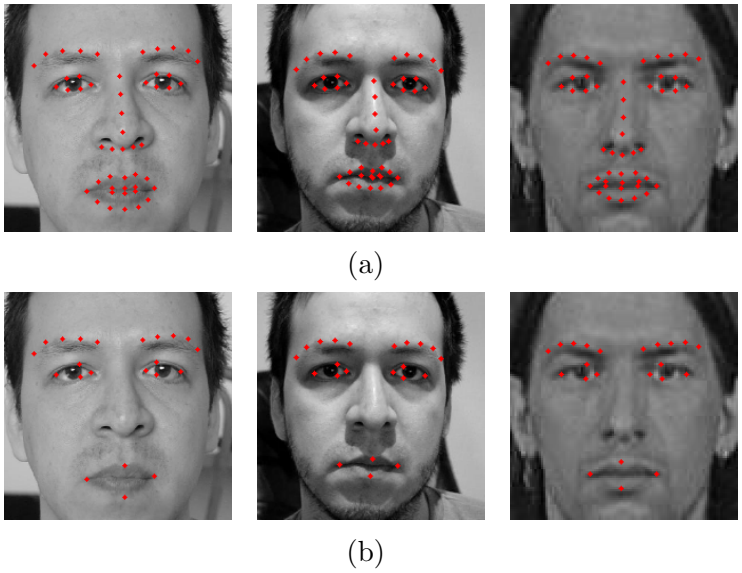
The upper and lower lip landmarks $\{\boldsymbol{m}_3, \boldsymbol{m}_4\}$ are selected based on the vertical projections of the horizontally aligned lip mask image. The alignment is done with respect to the angle formed between $\{\boldsymbol{m}_1, \boldsymbol{m}_2\}$ and the horizon. The estimation of the landmark y-coordinates follows the same principle as for the mouth corners, with threshold $\tau = 2$ pixels. The x-coordinate of $\{\boldsymbol{m}_3, \boldsymbol{m}_4\}$ is the mean of the mouth corners' x-coordinates.

4.3.4 Appearance based method

The appearance based method used in this work is the *Supervised Descent Method* (SDM) [98], a face alignment algorithm inspired by the *Active Appearance Model* (AAM) [18] and the *Constrained Local Models* (CLM) [20]. The SDM algorithm learns during training a sequence of optimal descent directions with a supervised approach. The optimal descent directions are computed using features extracted from known landmark locations at sampled images. The algorithm performs minimization of a non-linear least squares problem without the need to compute the Hessian or Jacobian using the learned descent directions. This approach is a reinterpretation of the cascade regression procedure (also known as Cascaded Pose Regression [22]) using gradient descents.

SDM computes *Scale-Invariant Feature Transform* (SIFT) [58] features around each landmark position. SIFT also provides illumination compensation. The SDM algorithm is a non-parametric method and it does not require an appearance model for fitting. An additional advantage is that the fitting process is fast. However, the algorithm requires large amounts of training data to produce a suitable model.

The SDM algorithm has been used for face recognition, face tracking,



(a)

(b)

Figure 4.12: Facial landmark detection using **Sd** algorithm implementation from `IntraFace`. (a) Full set of estimated landmarks; (b) Set of landmarks used in this work.

facial expression analysis, and facial expression transfer. This work uses only the feature detection capabilities of the algorithm. The implemented SDM algorithm is denoted **Sd** to keep the same naming convention as for the other landmark detection algorithms.

Implementation of the algorithm is provided by the `IntraFace` software package [98]. This package is used in conjunction with the face detection method previously introduced. An example facial landmark detection result can be seen in Figure 4.12, the used landmarks are only a subset of the landmarks available in the `IntraFace` implementation.

Chapter 5

Feature extraction

This chapter details the proposed feature extraction method from which geometric information of the face state is obtained using the detected facial landmarks. Since the face proportions like mouth width and height varies from person to person, a set of subject independent features is constructed. Feature extraction is performed in the same manner without regard of the detection algorithm that produced the facial landmarks. The features are collected in a *feature vector* and are used to produce categorization values of the studied facial elements. All detected facial landmarks are normalized in the range $[0, 1]$ with respect to the face region boundaries prior feature extraction. The eyebrows are described with eleven features, the eyes with two, and the mouth with five for a total of eighteen facial features composing the feature vector. The feature vector consists of $\boldsymbol{o}_k = \{o_k^{E1}, o_k^{E2}, o_k^{B0}, \ldots, o_k^{B10}, o_k^{Mw}, o_k^{M1}, \ldots, o_k^{M4}\}$, computed for each $k = \{1, \ldots, N\}$ video frames.

The features for each face element are post-processed with Principal Component Analysis (PCA) [72] to eliminate noise. The post-processed feature vector is later employed as input for the statistical learning methods in order to build the classifiers for estimating the state of the facial elements. This chapter details the feature extraction scheme for each facial element. In the first section the computation of the eyebrow features is introduced, followed by the eye features, and continuing with the mouth. The last section provides a visual inspection of the extracted features before they are postprocessed for a selected video.

5.1 Eyebrow features

Eyebrow position estimation uses the facial landmarks depicted in Figure 5.1. The left eyebrow is described with the set of landmarks $\{\boldsymbol{b}_1^l, \ldots, \boldsymbol{b}_5^l\}$ where

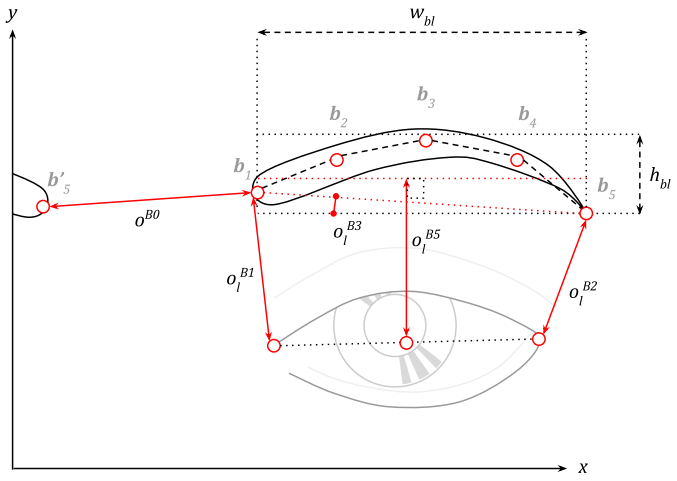


Figure 5.1: Facial landmarks for eyebrow position. The order of the landmarks for right and left eyebrows follows the same pattern; the first landmark is the beginning of the eyebrow from face center towards the ears.

$\boldsymbol{b}_i^l = \{b_{ix}^l, b_{iy}^l\}$, $i = \{1, \ldots, 5\}$ and similarly for the right eyebrow with upper index r.

The eyebrow features o^{B0} to o^{B4} are computed as in [3]. The features measure the distance between eyebrows, distance between the eyebrow corners and eye corners, eyebrow slope, and area of the eyebrow region. In addition, this work proposes the eyebrow feature o^{B5} that uses the eye center as a reference point. Features o^{B1} to o^{B5} are computed for both left and right eyebrows independently, while o^{B0} is computed only once, leading to a total of 11 eyebrow features. With w_b^l (w_b^r) the width and h_b^l (h_b^r) the height of the left (right) eyebrow, the features are computed for the left eyebrow as:

$$o^{B0} = ||\boldsymbol{b}_5^r - \boldsymbol{b}_1^l|| \tag{5.1}$$

$$o^{B1} = ||\boldsymbol{b}_1^l - \boldsymbol{e}_1^l|| \tag{5.2}$$

$$o^{B2} = ||\boldsymbol{b}_5^l - \boldsymbol{e}_2^l|| \tag{5.3}$$

$$o^{B3} = \frac{b_{5y}^l - b_{1y}^l}{b_{5x}^l - b_{1x}^l} \tag{5.4}$$

$$o^{B4} = w_b^l h_b^l \tag{5.5}$$

$$o^{B5} = \frac{\|e_{\mu y}^l - \rho e_{\mu x}^l - (b_{\mu y}^l - \rho b_{\mu x}^l)\|}{\sqrt{\rho^2 + 1}}. \tag{5.6}$$

Here ρ is the slope of the face with respect to the horizon, points $\boldsymbol{e}_\mu^l$ and $\boldsymbol{b}_\mu^l$ are the mean of the landmark coordinates of the left eye corners and

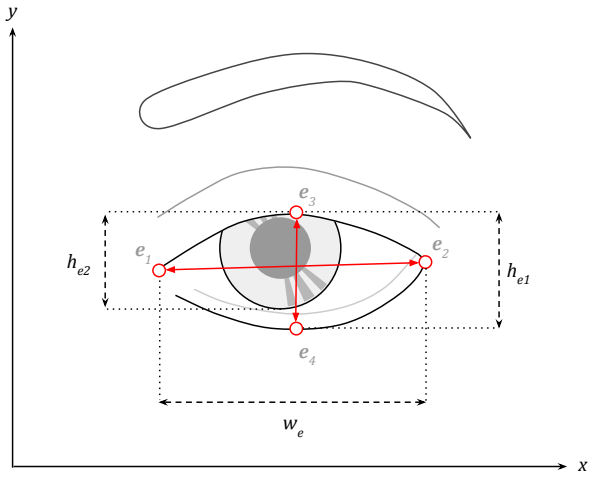


Figure 5.2: Facial landmarks for eye feature extraction. Here h_{e1} is the distance from the upper eyelid to the lower eyelid, while h_{e2} is the distance from the visible upper iris corner to the visible lower iris corner.

eyebrow, respectively. The slope ρ can be estimated from the face landmarks, in this study the slope is estimated using the mouth corners. Features o^{B1}, o^{B2}, and o^{B5} are scaled according to the average feature value of the first five video frames.

5.2 Eye features

Eye openness is determined from the distance between the upper and lower eyelids. The distance between eyelids can vary for the same state as the image scale changes. To account for this variation, the eyelid distance has to be normalized with respect to an invariant facial feature. This approach makes it unnecessary to register and track scale variations of the eye. Using the eye landmarks shown in Figure 5.2, the left eye is described with the landmarks $\{\boldsymbol{e}_1^l, \ldots, \boldsymbol{e}_4^l\}$ where $\boldsymbol{e}_i^l = \{e_{ix}^l, e_{iy}^l\}$ and similarly for the right eye.

The degree of eye openness is estimated from the features o^{E1} and o^{E2}, the left and right eyelid distance respectively. They are computed by using the equation:

$$o^{E1} = \frac{h_{e1}^l}{w_e^l} \tag{5.7}$$

$$o^{E2} = \frac{h_{e1}^r}{w_e^r}, \tag{5.8}$$

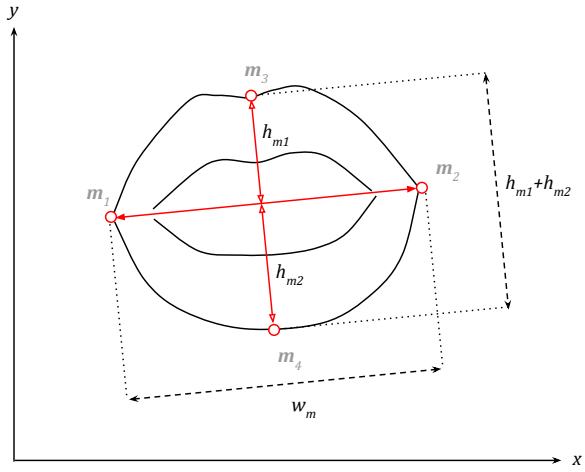


Figure 5.3: Facial landmarks for mouth state feature extraction. The landmarks are positioned at the boundaries of the lips, specifically at the mouth corners and the border of the upper and lower lip.

where $h^l_{e1} = ||\boldsymbol{e}^l_4 - \boldsymbol{e}^l_3||$ and $w^l_e = ||\boldsymbol{e}^l_2 - \boldsymbol{e}^l_1||$. $||\cdot||$ stands for the Euclidean length. Equation (5.8) produces an openness value in the range of $[0, 1]$.

5.3 Mouth features

The set of mouth features accounts for the two degrees of movement freedom during vocalization of words. The mouth landmarks are $\{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_4\}$ with $\boldsymbol{m}_i = \{m_{ix}, m_{iy}\}$ as can be seen in Figure 5.3. Distances $w_m = ||\boldsymbol{m}_2 - \boldsymbol{m}_1||$, $h_{m1} = ||\boldsymbol{m}_3 - \boldsymbol{\mu}_{w_m}||$ and $h_{m1} = ||\boldsymbol{m}_4 - \boldsymbol{\mu}_{w_m}||$ are computed in order to extract the features:

$$o^{Mw} = \frac{w_m}{w_{m0}} \tag{5.9}$$

$$o^{M1} = \frac{h_{m1}}{w_m} \tag{5.10}$$

$$o^{M2} = \frac{h_{m2}}{w_m}. \tag{5.11}$$

Where $\boldsymbol{\mu}_{w_m}$ is the geometric center of the two landmarks describing the mouth corners. Here w_{m0} represents the average w_m of the first five video frames.

In addition, the features used in [87] for expression recognition are also included in the mouth feature set. Features o^{M3} and o^{M4} are scale independent, meaning that lips at a determined position produce the same output

values for a larger or smaller identical shape. The added features are defined as:

$$o^{M3} = \frac{w_m}{h_{m1} + h_{m2}} \tag{5.12}$$

$$o^{M4} = \frac{h_{m1}}{h_{m2}}. \tag{5.13}$$

5.4 Feature vector arrangement

The extracted features are collected in a feature vector for each studied facial element. Three feature vectors are produced for each face input. The feature vector for the eyebrow contains features o^{B0} to o^{B10}, for the eye it contains features o^{E1} and o^{E2}, while the mouth feature vector contains features o^{M1} to o^{M4} and o^{Mw}. The features for the eyebrow and eye are postprocessed with a temporal average filter of window size three in order to remove noise, the same applies for the mouth features using a windows size four instead. Example values from the filtered feature vectors for a selected video can be seen in Figure 5.4, the shown feature vectors have not yet been postprocessed with PCA.

It can be noted by simple observation that the eye features range between $[0.1, 0.6]$ in the studied videos. The *open* eye state has values around 0.4, while the *closed* state typically falls under the 0.2 value. The state *wide* shows values greater than 0.5, but it can also be observed for values above 0.45, while the squint can range around 0.25.

The eyebrow features range between $[0, 1]$ in the studied videos, however they are scaled for visualization purposes. Taking as example o^{B1} in Figure 5.4, values fluctuating above or below 1 could be interpreted as an event, being a *raised* state if the value is greater than 1.1 and a *down* state if less than 0.9.

The mouth features can not be numerically evaluated by simple observation. However, mouth movements can be identified by fluctuations in each of the extracted features. While it is difficult to identify a specific state from the raw features, the *closed,relaxed* state is noticeable as the period without major fluctuations. For all the mouth features, the maximum value observed was below 4.

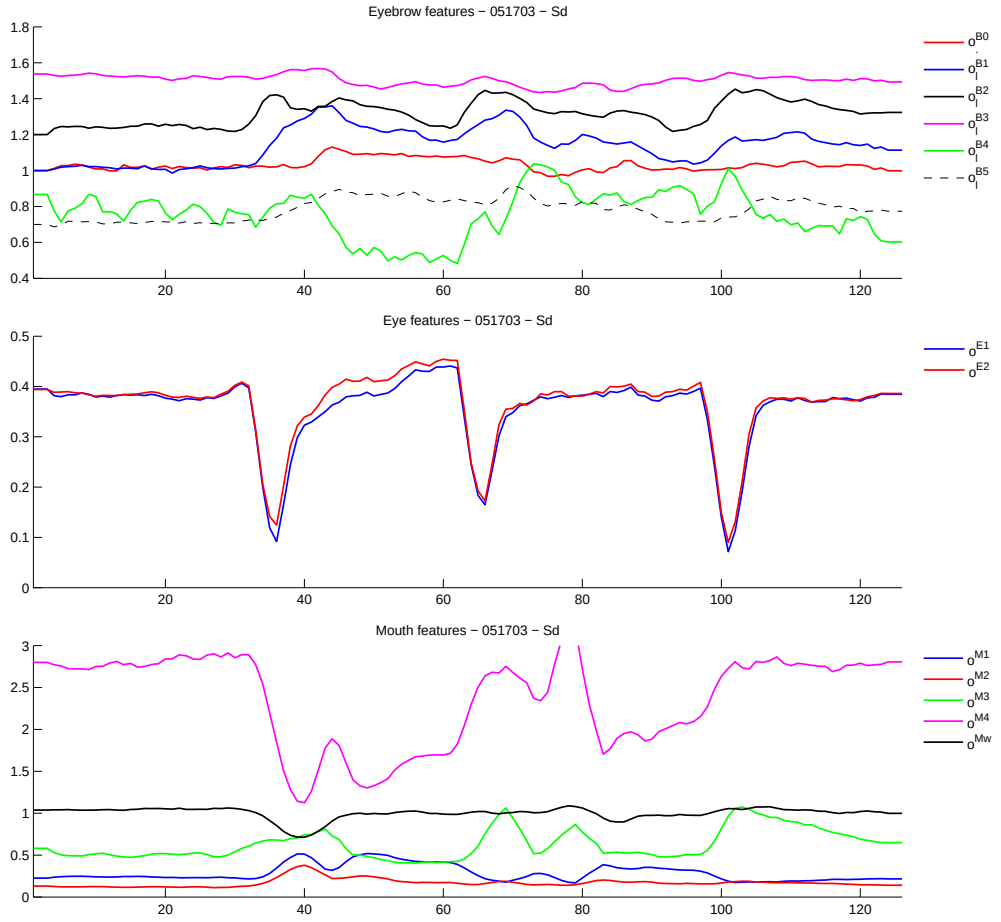


Figure 5.4: Scaled features extracted for a FinSL video using the **Sd** landmark detection algorithm evaluated as time signals.

Chapter 6

Experiments

Two experiments each containing a set of evaluations are designed in order to assess the usefulness of the facial element features extracted after different face segmentation and landmark detection algorithms. The first experiment, the FinSL experiment, employs videos from the Suvi *(Suomen viittomakielten verkkosanakirja)* dictionary for training and testing the classifiers. The second experiment, News Broadcast experiment, uses the best algorithms from the first experiment to generate annotations in news broadcast videos. The tests are designed to enable answering the research questions related to the viability of a facial state estimation method for use in automatic annotation of SL and news broadcast videos.

The tests in the first experiment are organized by the video data used: data for testing and data for training the classifiers. Two classification algorithms are incorporated in the tests: *Naive Bayes* [2] and *Support Vector Machine* [19]. The testing data is divided into videos without facial occlusions and videos with occlusions (where the subject uses glasses or hair style that constantly hides some facial features). The training data is also divided into two main groups: a subset of the Suvi videos and the Posed video set. Both training sets contain only videos without facial occlusions. The tests also consider the number of classes employed for classification: either *absolute* or *progressive* states, as described in Section 4.3. A summary of the organization of tests for the first experiment can be seen in Table 6.1.

The rest of this chapter is organized as follows: the first section provides details on the video material used for the experiments; in the second section, the FinSL experiment and the tests designed to evaluate the performance of the automatic annotations are described; in the third section, the News Broadcast experiment is defined. The last section describes the classification task for the experiments, and the formulation of the employed Naive Bayes and Support Vector Machine classifiers.

Group Type	Training Data	Classifier	Categorization
Occlusion yes/no	Suvi Set	Naive Bayes	Absolute
			Progressive
		Support Vector Machine	Absolute
			Progressive
	Posed Set	Naive Bayes	Absolute
			Progressive
		Support Vector Machine	Absolute
			Progressive

Table 6.1: Summary of the organization of the first experiment.

6.1 Source material for experiments

The video material used in the experiments consists of video sets from three scenarios: selected videos from Suvi a FinSL video dictionary, posed video segments, and Finnish news broadcast segments. All employed video material from the Suvi and posed video segments were manually annotated following the scheme presented in Chapter 3.

The set of posed video segments has been produced as part of this work. The formats (duration, quality, resolution) differ between the Posed, Suvi (FinSL) and news broadcast videos. Audio channels have not been included for any of the video material. The following subsections provide details on the video material for each video type. The Suvi and Posed video set are used primarily for training, test and quantitative evaluation. The news broadcast videos are evaluated qualitatively. Tables 6.2 and 6.3 show the distribution of samples in each data set and facial element according to the proposed categorization.

6.1.1 Finnish sign language sentences

The FinSL video dictionary videos are from the Suvi database. The subset of videos is the same that a previous SL study [41] used, the videos were chosen to take advantage of the available linguistic analysis of NM markers in the study. There are no specifications associated with the videos regarding clothes, hair or glasses; however they can be grouped according to face occlusions.

All selected Suvi videos have a resolution of 720×576 pixels at 25 fps with a 4:3 aspect ratio. The video compression has been performed with the Motion JPEG (MJPG) codec. The recording format has the subjects filmed from the head to the waist in front of a gray background.

	Eyebrow			Eye				V Mouth			H Mouth		
	0	1	2	0	1	2	3	0	1	2	0	1	2
Suvi	39	258	41	26	50	229	33	228	77	33	240	14	84
Posed	176	567	148	54	173	613	51	744	100	47	759	65	67
Test M	42	1275	365	135	280	1079	188	1034	487	161	1191	137	273
Test F	0	347	123	30	161	271	8	282	161	27	360	84	26

Table 6.2: Distribution of annotated video frames for eyebrow, eye, vertical and horizontal mouth states. See Table 3.1 for the explanation of the states.

	Mouth								
	0	1	2	3	4	5	6	7	8
Suvi	163	3	62	48	7	22	29	4	0
Posed	687	28	29	39	23	38	33	14	0
Test M	840	50	79	289	68	113	69	15	63
Test F	255	22	5	88	55	18	17	7	3

Table 6.3: Distribution of annotated video frames for mouth states. See Table 3.1 for the explanation of the states.

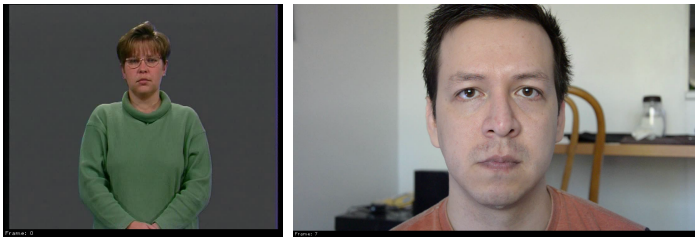
There are no specifications regarding clothes, hair or glasses; however the videos can be grouped according to subjects with or without visible facial occlusions. In this study the videos with occlusions (glasses and hair style) are from the same female subject, henceforth the set is named FEMALE and in the case of the set without occlusions MALE because they are all from the same male subject.

Each sequence has a duration of 4 to 6 seconds. A total of 2490 frames are used from 19 Suvi videos: 338 for training, 1682 for the testing MALE subset, and 470 for the testing FEMALE subset. Examples of the Suvi video frames can be seen in Figure 6.1a.

6.1.2 Posed video segments

A set of video recordings was produced with a subject performing artificial eyebrow, eye and mouth movements. This set is called Posed since it does not follow any natural communicational pattern. The Posed video segments have been employed as training material to test the performance of the classifiers. In the Suvi and Posed video segments only one person is filmed per segment, with the person facing towards the camera.

The Posed set was recorded with a Nikon D5100 photographic camera at 25 fps with a resolution of 1280×720 pixels and 16:9 aspect ratio. The video compression was performed by the Motion JPEG (MJPG) codec. The videos were filmed with frontal daylight illumination to reduce shadows in the face. All recordings frame the subject from the head to the lower portion



(a) (b)

(c)

Figure 6.1: Example frames from the different video sources used in this work. (a) Subject from the Suvi video dictionary. (b) Posed video produced as part of the experiments. (c) News broadcast video showing two faces.

of the neck and include objects in the background. The posed set contains 3 videos with an average duration of 13 seconds each and totaling 891 frames.

6.1.3 News broadcast videos

A different set of videos comes from television recordings of Finnish news broadcasts from the *Yleisradio* (YLE) broadcasting company. In these videos the number of persons varies from zero to many as the scene changes. The news video set is employed to evaluate the applicability of the automatic annotation model for videos outside the Suvi test videos. Possible applications for annotations in news videos include, for example, speaker identification and alignment when more than one person is seen in a sequence.

The news broadcast videos have a resolution of 720×576 pixels, have a rate of 25 fps, and 16:9 aspect ratio. The compression is performed using the H264 (MPEG4 AVC) method. There is no specific format for all sequences, but they share certain common elements when the news anchor is shown: background, body and head position, and range of movement. In a sequence, the news anchor can disappear from the view to introduce sections that do not show faces. Additionally, more than one face can appear during any sequence as shown in Figure 6.1c. The set of news videos contains sequences of various lengths lasting less than a minute, including those with and without faces.

6.2 FinSL video experiment

The performance of the automatic estimation of the state of facial elements is evaluated in quantitative and qualitative experiments. In the quantitative FinSL video experiment the state of each facial element was manually annotated in videos taken from the Suvi dictionary. The annotations were performed frame-by-frame on basis of the visual appearance of the isolated frame, without regard to linguistic significance. The manual annotations follow the proposed categorization scheme introduced in Chapter 3. For the qualitative set of experiments the reference annotations were taken from the work of [41]. The annotations for the qualitative experiment were prepared for a subset of the Suvi material from the point of view of linguistic significance. In the qualitative evaluation the facial state estimations are visually inspected against manual annotations and/or video frames to determine the accuracy of the estimations.

The faces obtained from the object detector are upscaled two times their original size to gain spatial resolution for the landmark detection algorithms, and to remove noise by smoothing the image. For each facial element five test sets were created by combining landmark detection algorithms with the different face segmentation approaches. The first set combines the segmentation method **G1** and the landmark algorithms **Pr**, **Cr**, and **Gr** for eyebrows, eyes, and mouth, respectively. The second set uses the same segmentation method with landmark detection algorithms **Pr**, **Pr**, and **Bi**. The third set changes the segmentation method to **G2** and combines the landmark detection algorithms **Pr**, **Cr**, and **Gr**. The fourth set is similar to the second with segmentation method **G2**. The fifth set uses the landmark detection algorithm in **Sd**.

Each of the five test sets produce a feature set for the statistical learning phase. The features obtained from all available videos are divided in training and testing subsets. The training subset divides into two sections: data from Suvi and data from the Posed videos. The testing subset contains also two sections: MALE and FEMALE test subsets. Naive Bayes and Support Vector Machine classifiers are built separately for each training subset.

6.3 News broadcast experiment

The set of news broadcast videos is not labeled with any particular information regarding the state of the studied facial elements. Henceforth, the evaluation of the experiment is qualitative. The automatic annotations were visually inspected and compared with the news broadcast video frames. A

visual representation of the annotations is produced for each encountered face/subject in order to facilitate the evaluation. The used approach for the qualitative evaluation is similar to the one used for FinSL videos, with the exception that there is no linguistic annotations available for these videos.

The news broadcast videos are not limited to a particular number of faces in a frame. Detected faces have to be aligned to keep track of each individual subject in order to avoid mixing up time vector subjects every time a face re-enters the frame. A simple face tracking algorithm is used for this purpose based on the idea that a face with a location close to a previous known location belongs to the same subject. The skin detection algorithm described in Section 4.2 is used to discard false faces.

6.4 Statistical learning

This section introduces methods for matching the feature vectors constructed from the facial landmarks with a *state* according to the proposed facial state categorization. Two supervised classification methods are employed to examine the discriminative capability of the proposed features. The first classification algorithm employed is the *Naive Bayes* (NB) [72] probabilistic classifier. As its name indicates, it computes a class membership probability given the provided observation by means of statistical inference. The output of the classifier in this case is the class with the highest probability value.

The second classification algorithm employed is the *Support Vector Machine* (SVM) [19] classifier. It utilizes a non-linear kernel function operating in a different space from the original feature space to separate the classes and to assign a score to each of the classes. The score is computed as the dot product between the feature vector and a vector of weights. The output of the classifier is the class with the highest score. The algorithm can use different weight estimation methods and kernel functions that can be chosen to suit a given task.

Classifier performance depends on the type of data and on the classification task. A comparative analysis between several types of classifiers using a common dataset is studied in [15, 16]. The NB classifier's performance provides a reference when model complexity and model implementation is taken into account. Nevertheless, several other classification methods have shown improved performance in real-world scenarios. The SVM classifiers have also shown good performance in different applications, and they provide some flexibility in their parameter selection. SVM is used to evaluate the performance of a different, more complex, classifier to see the potential improvement in the studied task.

6.4.1 Classification task

The classification task consists of training and testing a classifier in order to produce estimations of the states of facial elements. For training and testing purposes, all video frames have a labeled state per face element according to the categorization presented in Chapter 3. The states were manually labeled and are referred to as the ground truth labels. The classification task starts by training a distinct classifier for each facial element. The obtained classifiers are employed to categorize the previously unseen video frames, known as the test set.

Each facial state represents a single class in the classification task. The number of classes in a classification problem affects the formulation of the classifier. When the number of classes is two the problem is named *binary classification*. If there are more than two classes the problem is known as *multiclass classification*. This work considers exclusively cases where an observation can only be a member of exactly one class.

6.4.2 Naive Bayes classifier

The NB classifier is a parametric probabilistic algorithm that is based on the application of Bayes' theorem, with the assumption of conditional independence of features given the class labels (the so called *naive* assumption). The NB classifier is very popular for its simplicity in implementation and high accuracy for small datasets [2]. The NB classifier is known to be fast to train, fast to compute classifications with, and not affected by irrelevant features. Despite the simplicity of the model and that the term "naive" indicates otherwise, it is generally considered a good reference baseline classifier.

The NB classifier can be described in terms of a conditional probability model; given a number of data samples the objective is to know the probability of affinity to a class from a known set of possible classes. The probability model is constructed from the theorem:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

which, given the observed sample $\boldsymbol{x} = \{x_k, \ldots, x_N\}$, and class variable C, can be formulated as:

$$P(C|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C)P(C)}{p(\boldsymbol{x})}. \tag{6.1}$$

The prior probability $P(C = c)$ represents the probability that C takes the value of c regardless of the value of the sample $\boldsymbol{x}$. In other words, the

prior probability describes how frequent is the class C before looking at the samples.

The likelihood $p(\boldsymbol{x}|C)$ represents the conditional probability that an instance from class c has the observed sample value $\boldsymbol{x}$. Similarly, the evidence $p(\boldsymbol{x})$ is the marginal probability of instance $\boldsymbol{x}$ occurring, without considering class association. The evidence is the same for all classes, hence it is also called *normalizing constant*, and, for classification purposes, can be ignored since it scales the posteriors equally. The evidence in the two class case is written as:

$$p(\boldsymbol{x}) = \sum_{C} p(\boldsymbol{x}, C) = p(\boldsymbol{x}|C = 1)P(C = 1) + p(\boldsymbol{x}|C = 0)P(C = 0). \quad (6.2)$$

Combining the elements of the probabilistic model, for K mutually exclusive and exhaustive number of classes $C_i, i = \{1, \ldots, K\}$ and priors subject to $\sum_j P(C_j) = 1$, the posterior can be calculated as follows:

$$\begin{aligned} P(C_i|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|C_i)P(C_i)}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x}|C_i)P(C_i)}{\sum_{j=1} p(\boldsymbol{x}|C_j)P(C_j)}. \end{aligned} \quad (6.3)$$

From Equation (6.3) by assuming the independence of feature components x, the posterior can be written as:

$$P(C_i|\boldsymbol{x}) \propto P(C_i) \prod_{k=1}^{N} p(x_k|C_i). \quad (6.4)$$

The output C'^* of the NB classifier is the class with the highest estimated posterior probability:

$$C'^* = \underset{j}{\operatorname{argmax}}\ P(C_j|\boldsymbol{x}). \quad (6.5)$$

The NB classifier used in this work estimates the likelihood from a sample assuming a normal density function. In this function $\boldsymbol{\mu}_i$ is the vector of mean values computed from the samples in the observed class i, and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the samples. With d the size of the sample vector ($\boldsymbol{x} \in \mathbb{R}^d$), the formulation of the likelihood follows:

$$p(\boldsymbol{x}|C_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right]. \quad (6.6)$$

6.4.3 Support vector machine classifier

SVM classifiers [19] are supervised learning models used for binary classification and pattern recognition. SVM relies on the concept of *separating hyperplanes*, decision boundaries that separate members of one class from the other. Assuming that the data is linearly separable, a SVM classifier's separating hyperplane is the line that separates all members of both classes. The distance from the closest samples of each class to the separating hyperplane is called the *margin*. While finding any hyperplane may be enough to divide the data, by maximizing the margin a better generalization can be achieved. The separating hyperplane that best maximizes the margin is known as the *optimal separating hyperplane*, this is the reason for which SVMs are described as a *maximum-margin* methods.

Not all data is completely linearly separable. If there is no hyperplane that can linearly separate the data, the hyperplane that best separates and maximizes the margin of its nearest samples is the next best option. This is known as the *soft margin* technique, it allows misclassification of points by introducing non-negative slack variables. The slack variables relax the constraints of the SVM and measure the misclassification degree in the samples.

The hyperplanes can be described by $\boldsymbol{w} \cdot \boldsymbol{x}_i + b = 0$, where $\boldsymbol{w}$ is normal to the hyperplane and $\frac{b}{||\boldsymbol{w}||}$ is the perpendicular distance from the hyperplane to the origin. The *support vectors* are the training data points located at the separation margins of the classes. The SVM objective is to find a hyperplane that maximizes the equidistance from the support vectors. For binary classification, the standard formulation of SVM is given by [2]:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \forall_i. \end{aligned} \tag{6.7}$$

Here $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\}$ are training samples with $\boldsymbol{x}_i$ feature vectors and $y_i \in \{-1, 1\}$ target outputs. The soft margin extension introduces C, a regularization parameter or trade-off, between the accuracy and the amount of deviations larger than the non-negative slack variables ξ_i that are tolerated. The standard SVM formulation leads to a quadratic optimization problem that can be solved by introducing Lagrange multipliers. The Lagrangian formulation is given by [2]:

$$\mathcal{L}_P = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{l}\beta_i\xi_i. \tag{6.8}$$

Here α_i and β_i are the Lagrange multipliers. After differentiating with respect to $\boldsymbol{w}$, b and ξ_i and substituting back, the dual form of the original optimization problem is obtained. In this formulation the maximum-margin hyperplane is a function of the support vectors only. The dual form is given by:

$$\begin{aligned} \text{subject to} \quad & \sum_{i=1}^{l} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C, \forall_i. \\ \max_{\alpha} \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \end{aligned} \tag{6.9}$$

The original SVM was aimed at linear classification, a non-linear formulation is also possible for cases were the data is not linearly separable at all. The non-linear adaptation of SVM is similar to the linear version except that the dot product in Equation (6.9) is replaced by a non-linear kernel function. The kernel functions compute generalized dot products between two vectors. It can be observed for example that for the linear classifier $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \boldsymbol{x}_i \cdot \boldsymbol{x}_j$ is a linear kernel. A non-linear mapping function $\boldsymbol{x} \rightarrow \phi(\boldsymbol{x})$ can be used to transform the data from its original space into a higher dimensionality feature space where the data obtains linear structure. The high-dimensional mapping is related to the kernel such that $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j)$. This allows us to find a separating hyperplane. The explicit computation of the mapping function can be avoided since only the dot products of the mapped data need to be determined, this is referred to as the *kernel trick*.

The SVM classifiers can use specialized kernels that best suit different types of data. Under this generalization they became also known as *Kernel Machines* [2]. In this work, a Radial Basis Function (RBF) is used as kernel. The RBF kernel is given by:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2), \tag{6.10}$$

where the parameter γ determines the extent of the kernel in the feature space. $\boldsymbol{w}$ in the feature space equals to $\sum_i y_i \alpha_i \phi(\boldsymbol{x}_i)$. The output of the classifier for a data point $\boldsymbol{x}'$ is determined by:

$$y' = \text{sgn}(\boldsymbol{w} \cdot \phi(\boldsymbol{x}') + b). \tag{6.11}$$

SVM is aimed at binary classification, however, for $K > 2$ classes it can be extended to multiclass classification by the decomposition of the classification task into binary problems. A common approach is to construct K

two-class classifiers, each classifier separates a class from all other samples that do not belong to it. This approach is referred to as *one-vs-all*. The experiments in this work employ an alternative approach referred to as *one-vs-one*. In this approach $K(K-1)/2$ pairwise classifiers are constructed, and the discriminant uses majority voting to select the output of the classifier. Each classification is considered a vote, the final output is the class with the largest number of votes. The SVM implementation used in this work is provided by the `LIBSVM` [17] package.

Chapter 7

Evaluation

In this chapter the method of performance assessment is detailed. The complete number of results to evaluate given the experiment setup described in Chapter 6 is considerable: for each of the four facial elements to evaluate, five combinations of algorithms is tested, each test is evaluated using two different classifiers, two datasets are used for statistical learning producing two different classifiers for each classifier type, additionally each classifier is used for multiclass and binary classifications. This process is performed for the MALE and FEMALE subsets. To evaluate the performance of the selected features and classifiers for each facial element a comprehensive tool is used in order to address the number of dimensions of the experiments, and the characteristics of the data.

Section 7.1 reviews the performance measure employed in the experiments; the studied measure combines the information of positive and negative results in a single performance value. The performance evaluation method is also discussed; a graphical tool is proposed for statistical analysis of the results given the number of tests and videos used in each test. Sections 7.2 and 7.3 are dedicated to the performance review of the quantitative FinSL experiment; first with the MALE and then with the FEMALE subset of videos. The algorithms and classifiers with better quantitative performance are summarized in Chapter 7.4. In Section 7.5 a qualitative analysis is performed using linguistic annotations. In Section 7.6 the results of the qualitative News Broadcast experiment are evaluated; the analysis is performed with the same estimation methods selected from the qualitative FinSL experiment. Discussion of the results for all experiments is presented in Section 7.7.

7.1 Performance measurement

Previous research [83, 84] has explored in detail the invariance of measures in the context of classification performance analysis, showing that the properties of a measure can greatly influence the interpretation of results. Measures should be chosen so that they match the data characteristics of the experiment performed, this way the interpretation of results maintains consistency with the studied task. In this work the selected measures aim at applicability and interpretation for both two class and multiclass classification, as well as being invariant to highly imbalanced classes.

The number of facial states in the test videos are highly imbalanced in all categories. For a given video frame the classification algorithm assigns a class to each eyebrow, eye, and mouth position. The assigned class output takes discrete values for each facial category as described in Table 3.1. Considering binary classification, a full eye *blink* occurs approximately within 5 frames. In a video of 5 seconds containing 2 blinks for example, the class *closed* will represent $\frac{5 \cdot 2}{5 \cdot 30} = 0.07 = 7\%$ of the processed frames. This situation is more evident in multiclass classification since the number of samples in each class can vary significantly: the number of frames that show a *squint* or *wide* eye state is significantly smaller compared to the number of frames in neutral *open* position. Performance measures that account for only positive or negative classifications can yield misleading interpretations if the classifier favors only the dominant class of the test [79]. The performance measure interpretation must be considered when the class distribution is known to be highly biased towards a specific class.

To evaluate the performance of a test for highly imbalanced classes it is preferable to use separate measures for each class. In binary classification this can be achieved with *sensitivity* and *specificity*. Sensitivity indicates how effectively the classifier can identify positive results (the probability that a state is correctly identified as *open* given that the state really is *open*). Specificity indicates how effectively a classifier can identify negative results (the probability that a state is identified as *closed* given that the state is *closed*). By using the confusion matrix in Table 7.1, sensitivity can be formulated as $\frac{TP}{TP+FN}$ and specificity as $\frac{TN}{FP+TN}$. The output value range of sensitivity and specificity falls within the limits $[0, 1]$. According to the specification of the measures, a classification task is considered better as it gets closer to 1 in both values.

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 7.1: Confusion matrix for a binary classification scenario; TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

7.1.1 Mathew's correlation coefficient

Sensitivity and specificity can be represented by a single correlation coefficient measure even with imbalanced classification results [6, 79]. This has the advantage of employing a single measure for preliminary evaluation, allowing the use of sensitivity and specificity for more in-depth analysis if necessary. The correlation performance measure used is an application of Pearson's product-moment correlation coefficient (PCC) [54] to a confusion matrix known as the Matthew's Correlation Coefficient (MCC) [62]. It is worth noting that MCC is strongly related to Pearson's Phi coefficient (and consequently with PCC), where the PCC of two binary variables yields the Phi coefficient, and the Phi coefficient applied in terms of a confusion matrix equals the MCC [6].

The MCC produces output values in the range of $[-1, 1]$. A value of -1 represents a total disagreement between classification and true class values, a value of 1 represents a perfect classification, and a value of 0 represents a classification no better than random assignment. Even with highly imbalanced class membership, MCC provides a baseline for performance evaluation when all observations are assigned to a single class. Given a confusion matrix as presented in Table 7.1, MCC is given by [6]:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{7.1}$$

MCC is also used to measure the performance of classifications when there are more than two classes, by extending the PCC for K classes, as suggested by [32]. The multiclass PCC performance measure R_K, renamed "multiclass MCC" due to its use of the confusion matrix, is explored in [44, 45]. The measure yields consistent results with the two class version of MCC, the output value interpretation is also the same as with the two class version.

MCC provides a good single performance metric where using other performance measures would have required per-class analysis of each experiment. For this reason MCC is a reasonable tool to quickly assess multiclass per-

formance in balanced and imbalanced classifications. Consider two $N \times K$ classification matrices: Y as the class memberships assigned by a classifier, and Y' the true class memberships. Generally, the sample PCC is defined as [32]:

$$R_K = \frac{\operatorname{cov}(Y, Y')}{\sqrt{\operatorname{cov}(Y, Y)\operatorname{cov}(Y', Y')}} \tag{7.2}$$

with

$$\begin{aligned} \operatorname{cov}(Y, Y') &= \sum_{k=1}^{K} \frac{1}{K} \operatorname{cov}(Y_k, Y'_k) \\ &= \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} (Y_{nk} - \bar{Y_k})(Y'_{nk} - \bar{Y'_k}) \end{aligned} \tag{7.3}$$

where $\bar{Y_k} = \frac{1}{N}\sum_{n=1}^{N} Y_{nk}$ and $\bar{Y'_k} = \frac{1}{N}\sum_{n=1}^{N} Y'_{nk}$ are the means of the class (column) k. In this formulation, PCC supports multiple classes by means of the inner product of the corresponding columns of each matrix and subtracting their means.

The sample PCC can be applied in terms of a confusion matrix by including a $K \times K$ confusion matrix C, where C_{kk} represents the number of correctly classified frames y_{ik} in class k, and C_{kl} the frames classified as class k that belong to class l, $l \neq k$. With the binary variables $\boldsymbol{y}_k = y_{1k}, ..., y_{nk}$, $y_{ik} \in \{0, 1\}$ for the discrete case. The final formulation of MCC then follows the same construct as the sample PCC, and using Equation (7.3) it can be seen that:

$$\text{MCC} = \frac{\sum_{klm} \left(C_{kk} C_{lm} - C_{kl} C_{mk} \right)}{\sqrt{\sum_{k} \left[\left(\sum_{l} C_{kl} \right) \left(\sum_{l',k' \neq k} C_{k'l'} \right) \right]} \sqrt{\sum_{k} \left[\left(\sum_{l} C_{lk} \right) \left(\sum_{l',k' \neq k} C_{l'k'} \right) \right]}}, \tag{7.4}$$

which can be considered as a K-dimensional description of the two class MCC. The multiclass MCC has the same value range of $[-1, 1]$ as in the two class version.

In the context of the video data, a MCC value of 0.25 indicates that the estimations only follow partially the movement of the facial element and are not very reliable. More explicitly, that the estimations have a major tendency to follow the true pattern (a few continuous frames or several sporadic frames) than to not following it. A MCC value of 0.5 indicates that the estimations may miss some frames or events, but otherwise follow greatly the

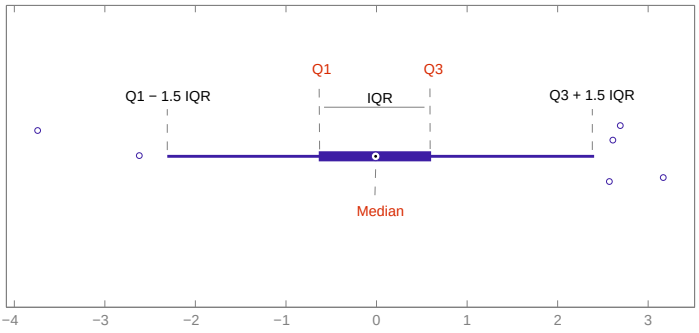


Figure 7.1: Horizontal representation of a Box Plot as used in this work. $Q1$ and $Q3$ represent the upper and lower quartiles of data respectively, with the interquartile range ($IQR = Q3 - Q1$) as their difference. The outliers are represented with bubbles outside the range of the whiskers.

true patterns of movement. A classification model reaching this performance value is considered to be good but noisy. A MCC value of 0.75 indicates that the true movement patterns are following very closely, with a few misalignment errors or ambiguous intermediate states. At a value of 0.75 the classification model is considered to be reliable. Any higher MCC value adds to the model reliability, however, reaching 0.85 is considered to be a near perfect estimation.

7.1.2 Graphical evaluation

For numerical evaluation the general performance of a classifier across all available videos in a test is considered. This task would typically employ the average MCC produced for all videos. However, the average of the correlation coefficients can be biased in the presence of outliers. For this reason, the distribution of the MCCs is used instead in order to perform a more comprehensive analysis.

The *box-and-whisker* diagrams [63] are used to analyze the distribution of the MCC measures obtained from the assessment of the experiments. As shown in Figure 7.1, the box-and-whisker diagram allows statistical graphic evaluation: the diagram setup in this work uses the median, 25th and 75th percentiles and 99.3% boundaries. This graphical tool helps to achieve fast performance comparison when the number of test samples is large.

From the graphical evaluation, the stability of the classification can be identified by the length of the box and whiskers, marking the potential MCC outliers. The box plot analysis is not intended to replace the average and variance obtained from the MCC measure, but to complement the evaluation.

It is worth noting, however, that the number of samples in each test may also affect the interpretation of the box plot results.

7.2 Quantitative FinSL MALE experiment

The discriminative properties of the extracted features were first tested with the MALE subset. The properties of the MALE subset are described in Chapter 6. The classifiers were trained using two different data sets: Suvi videos and Posed videos. The subset training videos from Suvi are 0006-01, 0058-02, 0058-04 (in article-sentence format). The test dataset is given by the MALE subset 0350-01, 0466-03, 0473-01, 0517-03, 0655-03, 0687-01, 0777-04, 0823-04, 0972-03, 1138-01, 1207-01, 1216-01.

The performance evaluation employs the box-plot graphical method of MCC coefficients. The tests are grouped by classifier type and by data used for training the classifier. An identical second plot is produced from the binary classification results. This approach is employed for each facial element. Detailed performance values can be found in Appendix A.

7.2.1 Eyebrow position

For eyebrows, the features o^{B5} and o^{B10} were extracted using the mouth corners as alignment horizon. This makes the eyebrow results biased towards the precision of the mouth landmarks. The 11 eyebrow features were post-processed to reduce the observed noise using *principal component analysis* (PCA). The 4 strongest PCA coefficients were selected based on classification performance sampling, and thus the eyebrow feature vector dimensionality is reduced to 4. The number of coefficients was selected by taking the highest median MCC value of each possible PCA projection.

The resulting correlation values of eyebrow classifications are below the 0.25 mark in the majority of classifiers. The strongest results were obtained using **Sd** with the Suvi training set and the NB classifier, reaching a median correlation of 0.3. A median of 0.24 was obtained with **Gr/G2** with small variations, but narrower performance result distribution. The performance improved in binary classifications: the median correlation of **Sd** with the Suvi training and NB classifier reached value 0.41, with the **Gr/G2** algorithm results reached a median value of 0.39 using the same classifier type and training data.

The classifiers trained with the Posed set showed little or no improvement over the Suvi-trained classifiers. It is worth mentioning that with this

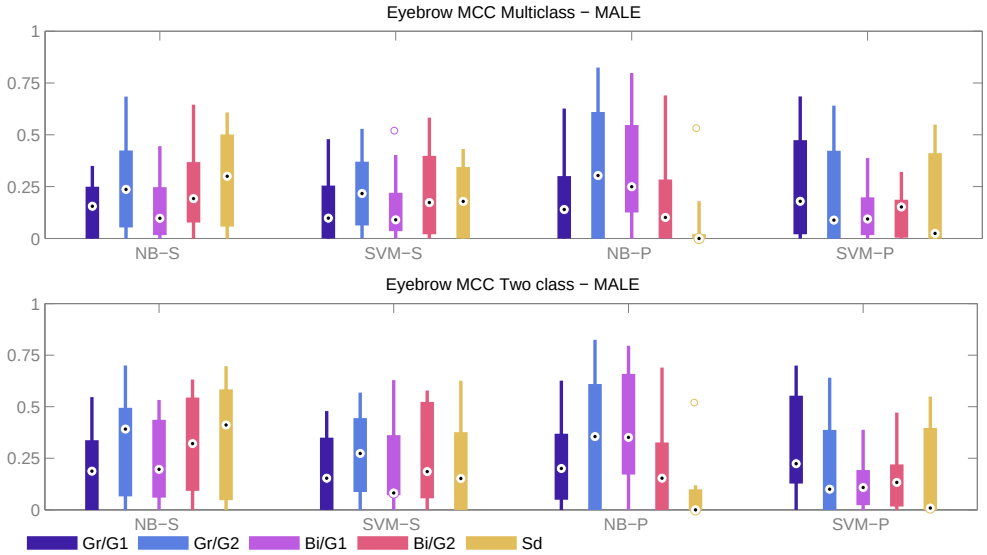


Figure 7.2: MCC distribution of the classification labels of the eyebrow estimations for the MALE test set. The figure shows classifier performance using Naive Bayes (NB) and Support Vector Machines (SVM) with either Suvi (S) or Posed (P) training data.

classifier, the **Gr/G2** algorithm outperformed the other options with median correlation of 0.31. However, the correlation distribution shows that the **Gr/G2** performance is unreliable with Posed training. The results can be seen in Figure 7.2.

7.2.2 Eye openness

The eye openness estimation shows the highest correlation medians of all facial elements as can be seen in Figure 7.3. The eye feature vector was postprocessed with PCA to remove noise, preserving the original dimensionality. The numbers of features used was 2. Eye openness is determined using features from both eyes at the same time and produce a single state estimation.

The classifiers trained with the subset from Suvi showed more robust performance than those trained from the Posed set. In multiclass classification the strongest results were obtained using the **Sd** algorithm with SVM classifier, the median correlation was 0.69. The **Pr/G2** algorithm with SVM and Suvi training got the highest median of the proposed algorithms with a MCC value of 0.53. In binary classification, the strongest correlations were obtained with the Suvi training and SVM classifier. A median correlation of 0.73 was obtained using the **Sd** algorithm, while with the **Pr/G2** algorithm the median correlation reached 0.72.

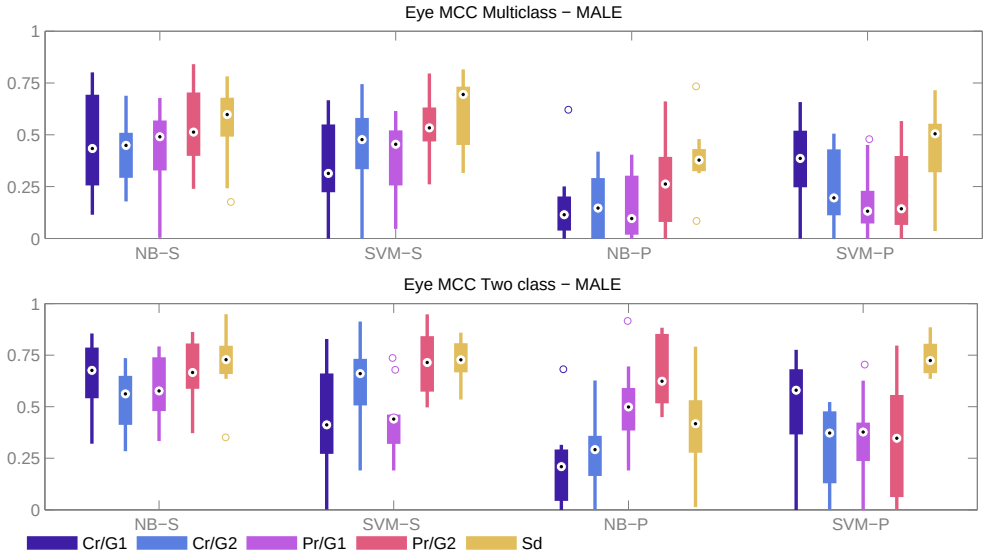


Figure 7.3: MCC distribution of the classification labels of the eye estimations for the MALE test set.

The classification performance using the classifiers trained with the Posed data was consistently lower than the performance of the classifiers that used the Suvi training data. The **Sd** algorithm had the highest median correlation amongst the Posed training set, with comparable results in binary classification using the SVM classifier. The proposed algorithm **Pr/G2** performed comparably with the NB classifier, while the **Cr/G1** algorithm performed noticeably better using the SVM classifier.

7.2.3 Mouth state

The correlation results for the mouth estimations favored the classifiers with Posed data training as shown in Figure 7.4. The feature vector for the mouth state estimation is postprocessed with PCA, retaining the strongest 2 PCA coefficients. The number of categories (classes) in the mouth state estimation is 9.

The results from the SVM classifier and Posed training dataset had good correlations for multiclass classification, having the strongest results with the **Sd** algorithm at a median MCC of 0.45. The binary classification favored the NB classifier, in this case with Suvi training and **Sd** algorithm at a median MCC of 0.74. The **Gr/G2** algorithm showed better results than **Sd** in multiclass classification with Suvi training, however in binary classification the **Sd** algorithm had a better performance.

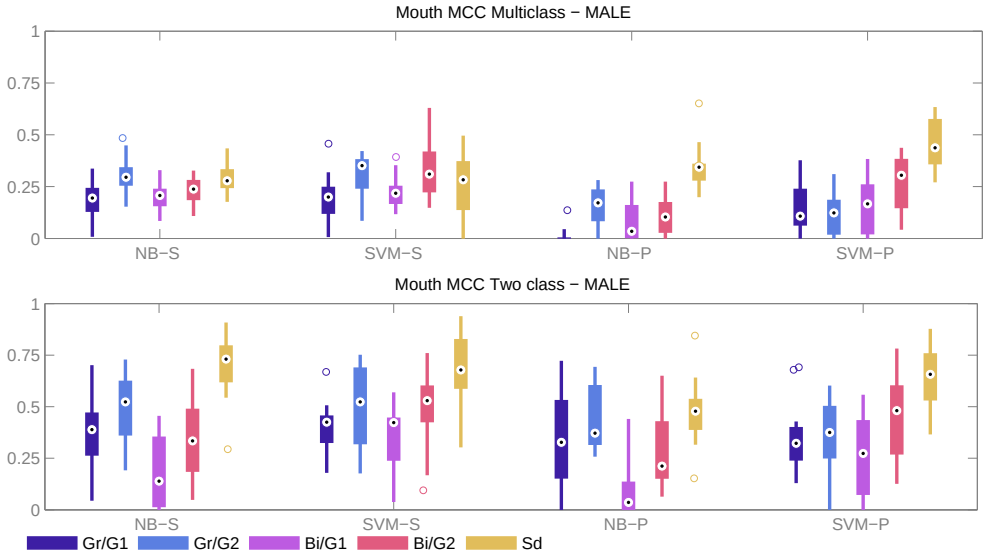


Figure 7.4: MCC distribution of the classification labels of the mouth estimations for the MALE test set.

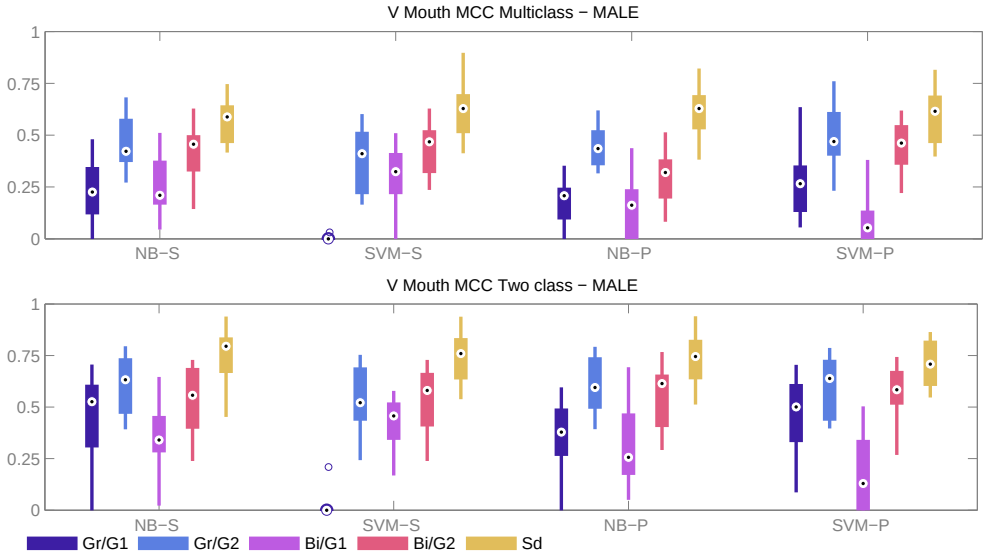


Figure 7.5: MCC distribution of the classification labels of the vertical mouth estimations for the MALE test set.

The vertical mouth feature vector was obtained after postprocessing the mouth feature vector with PCA and retaining the strongest PCA coefficient, thus reducing the feature vector dimensionality to 1. As shown in Figure 7.5, the NB classifier with Posed training data and **Sd** algorithm yielded a correlation value of 0.63. The Posed training data also benefited the proposed

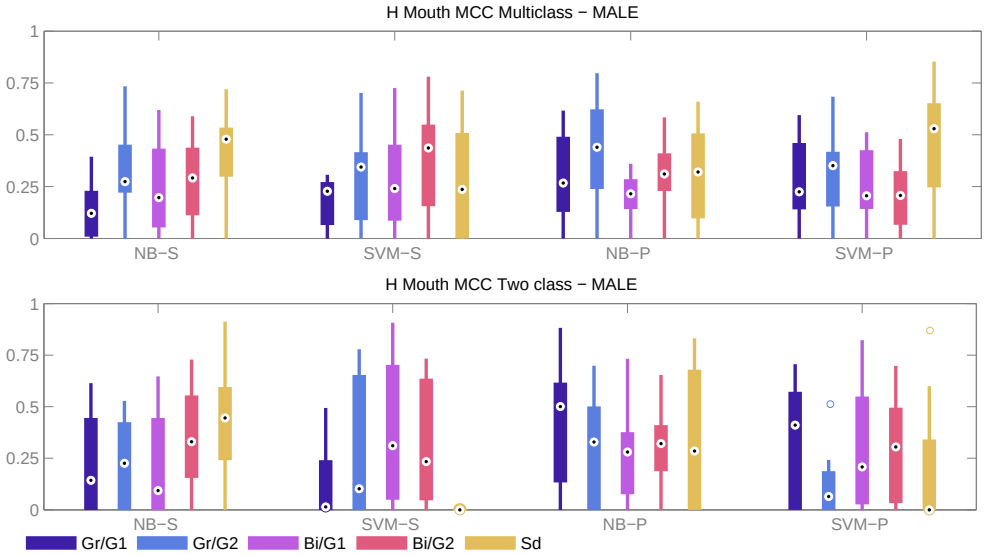


Figure 7.6: MCC distribution of the classification labels of the horizontal mouth estimations for the MALE test set.

Gr/G2 algorithm, which produced a median correlation value of 0.47 using the SVM classifier. In binary classification the algorithms **Sd** and **Gr/G2** got the leading results with the same classifier/training data combination as in multiclass classification, the median correlations achieved were 0.75 and 0.64, respectively.

The horizontal mouth feature vector was also postprocessed with PCA, in this case retaining the strongest 3 PCA coefficients. The SVM trained with the Posed data had the best performance with a median correlation value of 0.53. Amongst the classifiers trained with the Posed data, the **Bi/G2** option was preferred from the proposed algorithms since it showed higher performance in binary classification than the **Gr/G2** algorithm. The best median MCC of the **Bi/G2** algorithm was obtained using the NB classifier, producing a correlation value of 0.31. In binary classification, the proposed algorithm retained a median correlation value of 0.32 with the same classifier/training data combination, as can be seen in Figure 7.6.

7.3 Quantitative FinSL FEMALE experiment

The FEMALE test dataset is formed by the Suvi video subset 0346-01, 0807-04, 0966-03, and 1029-02. The training video sets are the 0006-01, 0058-02, and 0058-04, the same as with the MALE test set. The Posed dataset is also used for training as with the MALE test. The FEMALE test set poses

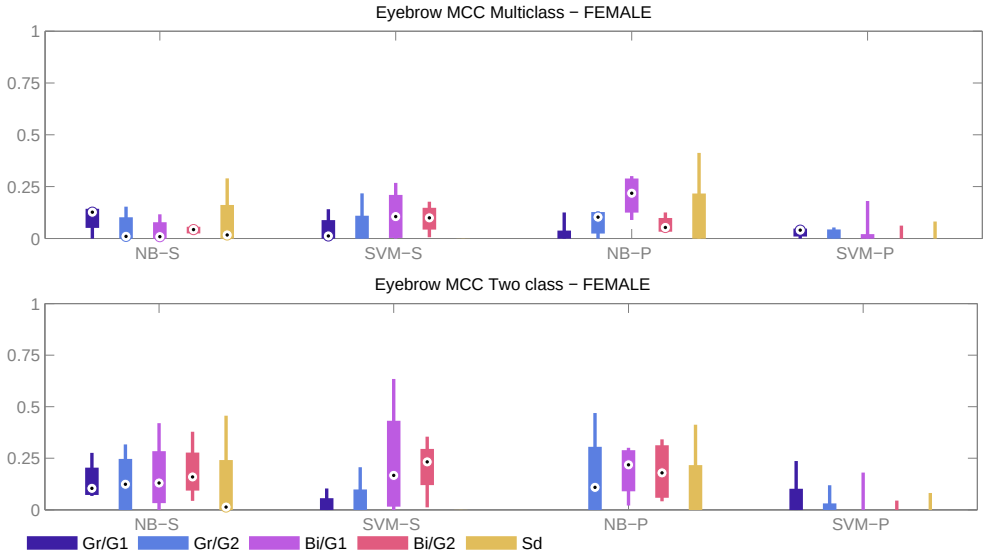


Figure 7.7: MCC distribution of the classification labels of the eyebrow estimations for the FEMALE test set.

difficulties for automatic estimation of states of facial elements due to the presence of artifacts partially occluding the face of the subject. For this reason, this data set is considered as a different group within the Suvi test data.

All feature vectors of each facial element were post-processed with PCA as with the MALE set. The feature vector dimensionality for eyebrows is 4, for eyes 2, for mouth 2, for vertical mouth 1, and for horizontal mouth 3. The quantitative evaluation is performed following the same procedure as with the MALE test.

7.3.1 Eyebrow position

In Figure 7.7 the eyebrow MCC distribution can be seen for the tested classifiers. The eyebrow position estimation results are significantly lower than for the MALE test set. The resulting median correlation value of classifications for all the tested landmark algorithms was below 0.25, with the higher median value at 0.23 using the **Bi/G1** algorithm with the NB classifier and Posed training data for the multiclass scenario.

The eyebrow position binary classification results improve slightly overall. The best combination is again the same as in multiclass classification, reaching a median MCC value of 0.25. It should be noted that in this test set the **Bi/G2** algorithm with Suvi training data and SVM classifier also

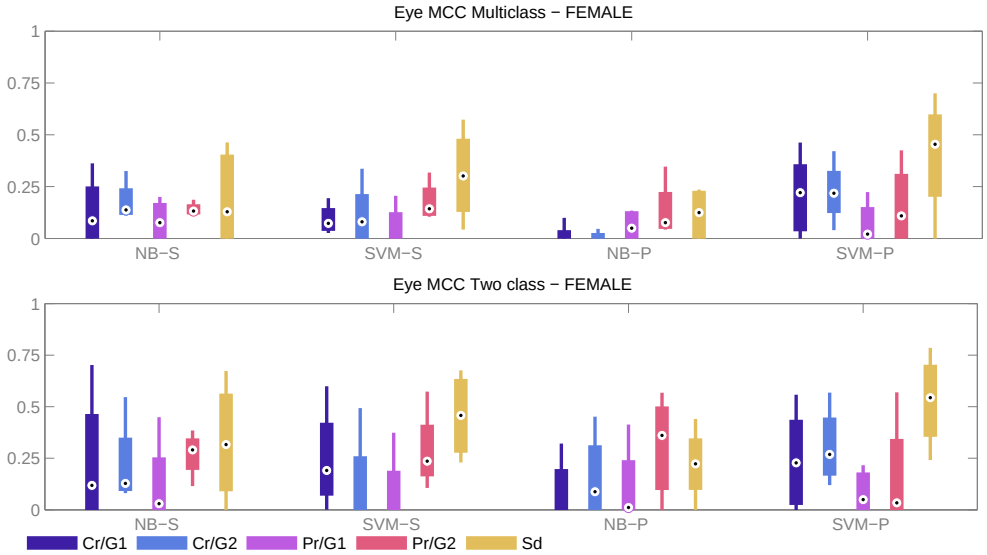


Figure 7.8: MCC distribution of the classification labels of the eye estimations for the FEMALE test set.

produced a median correlation of 0.25 in binary classification.

7.3.2 Eye openness

The eye openness estimation results in the multiclass scenario favored again the classifiers trained with the Posed data and the **Sd** landmark detection algorithm. The SVM classifier trained with Posed data yielded the best median classification correlation value of 0.48. The **Cr/G2** landmark detection algorithm reached a median correlation value of 0.24 using the Posed training data and SVM classifier. The **Cr/G1** algorithm reached a similar result, however its MCC value distribution showed a more unstable performance.

In binary classification the best results were obtained using the Posed training data. The **Sd** algorithm reached a median MCC value of 0.56 using the SVM classifier. From the proposed landmark detection algorithms the best median MCC value reached the 0.35 value using the **Pr/G2** algorithm. Results from the **Cr** algorithm using classifiers trained with the Posed data set showed little improvement in the binary classification scenario. The resulting MCC box plots can be seen in Figure 7.8.

7.3.3 Mouth state

The mouth state estimation results showed that, again, the classifiers trained with the Posed data produced the highest median MCC results. The SVM

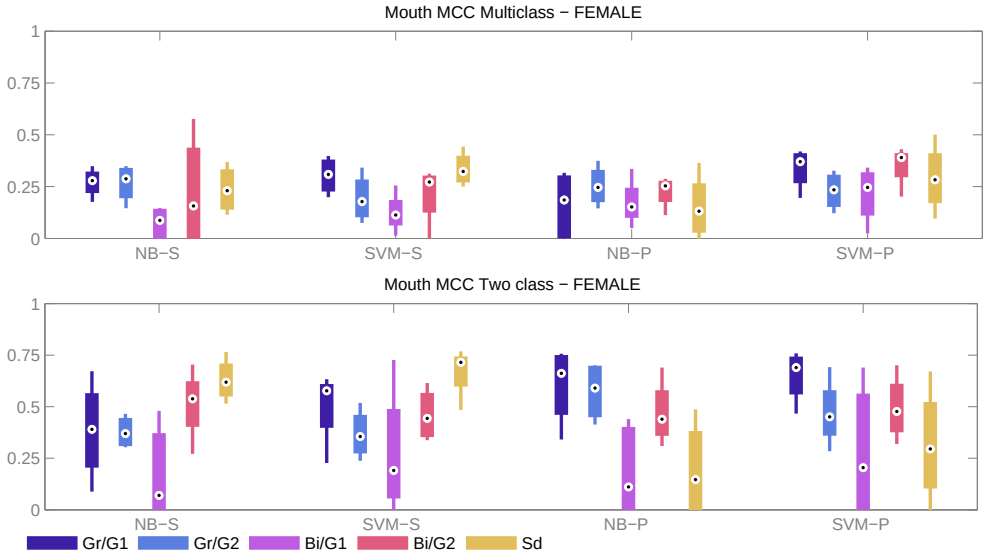


Figure 7.9: MCC distribution of the classification labels of the mouth estimations for the FEMALE test set.

classifier trained with features extracted from the detected landmarks of the **Bi/G2** algorithm had the best performance at 0.37 median correlation value. The **Gr/G1** landmark algorithm obtained similar results in the same setting with a median correlation value of 0.36. In binary classification the results favored the **Sd** algorithm and classifiers trained with the Suvi data. In this group the SVM classifier showed the best results at 0.74 median MCC value using the **Sd** landmark algorithm. From the proposed landmark algorithms, the **Gr/G1** held a median correlation value of 0.73, being close to the highest result in binary classification. The summary can be seen in Figure 7.9.

The vertical mouth estimation results in Figure 7.10 show that the **Sd** landmark algorithm consistently produced the highest median MCC value in all test groups. The SVM classifier trained with Suvi data had the highest median MCC value 0.67. The **Bi/G2** algorithm is the closest to the highest using Posed data training and the NB classifier with a 0.51 median correlation value. The same results hold for the performance in binary classification. The highest median MCC reached the 0.75 value by using the Posed training data, the NB classifier and the **Sd** landmark detection algorithm. The **Gr/G1** algorithm is slightly behind at 0.74 median correlation value with Posed data training and the SVM classifier.

The horizontal mouth estimation results show that the performance of the **Gr/G1** algorithm was the highest in the multiclass scenario. The highest median correlation value was 0.59, achieved by the NB classifier with

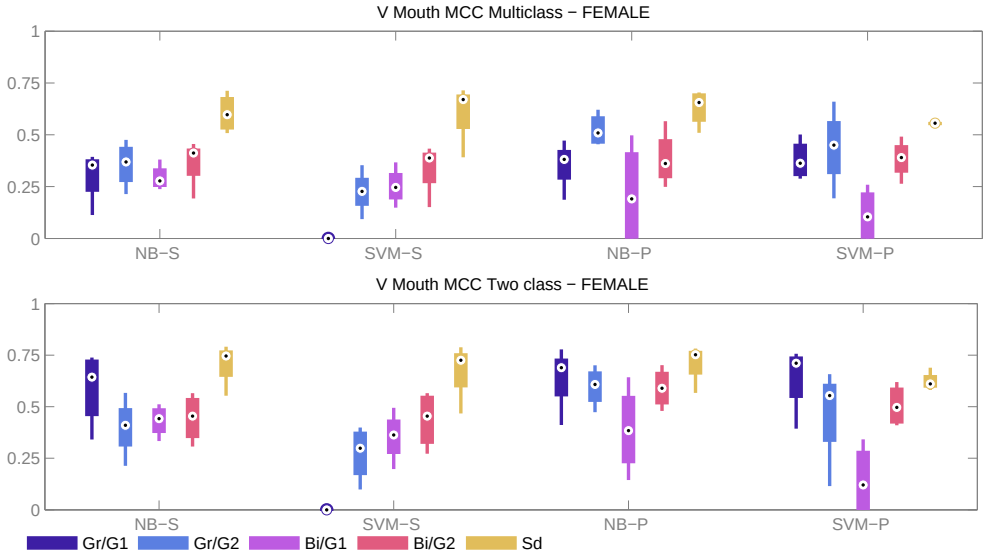


Figure 7.10: MCC distribution of the classification labels of the vertical mouth estimations for the FEMALE test set.

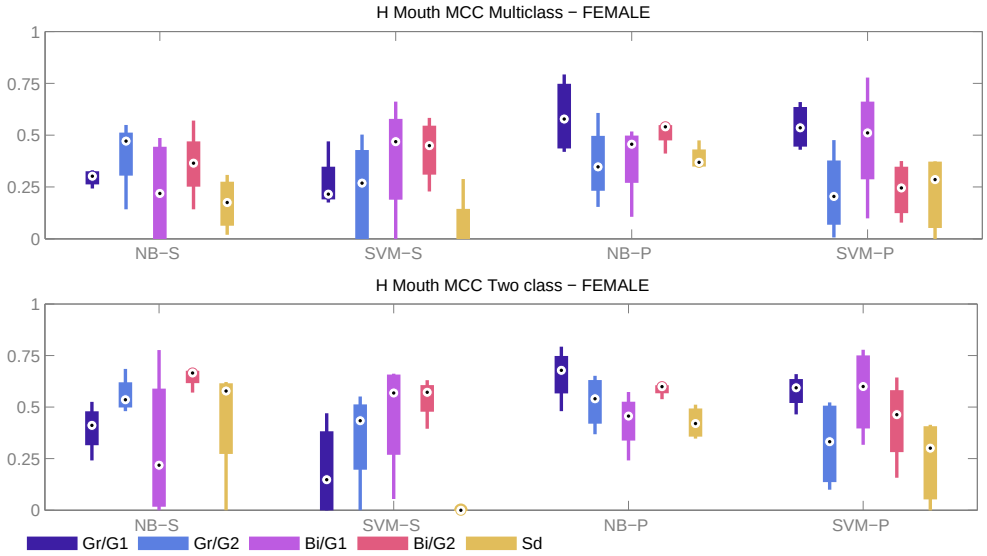


Figure 7.11: MCC distribution of the classification labels of the horizontal mouth estimations for the FEMALE test set.

Posed training data. In binary classification **Gr/G1** got the highest median correlation value again with the same combination of classifier/training data as in the multiclass scenario. The binary classifiers in the Posed training data group favored instead the NB classifier using features extracted from the **Bi/G2** landmarks, which also showed the narrower MCC distribution

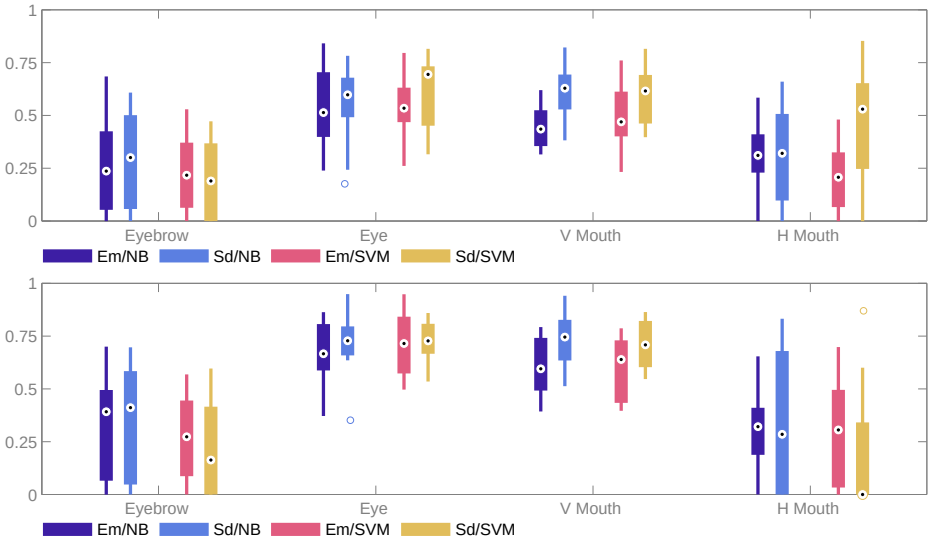


Figure 7.12: MCC distribution of **Em** and **Sd** for each facial feature with MALE test set. Top: Multiclass classification. Bottom: Binary classification.

in this test setting. This was the only facial element experiment where the **Sd** algorithm underperformed in most classification settings. The complete results can be seen in Figure 7.11.

7.4 Quantitative experiments summary

To summarize the quantitative experiments, the best performing landmark detection algorithms proposed for each facial element are grouped together to form the Landmark Ensemble Method (LEM). The LEM method is here denoted **Em** to maintain the naming nomenclature in the results. The selected algorithms in **Em** use **G2** for facial feature segmentation, **Gr** for eyebrow position (PCA processed using the strongest four components), **Pr** for eye openness (PCA processed using two components), **Gr** for vertical mouth state (PCA processed using the strongest component), and **Bi** for horizontal mouth state (PCA processed using the strongest three components). After PCA processing the feature vector dimensionality of all facial elements is reduced from 18 to 10 elements.

The performance of the **Em** algorithm for each facial element is compared once more against the performance the **Sd** algorithm in order to select a landmark detection algorithm for the qualitative evaluation. The results in Figure 7.12 show that the MCC distributions of the **Em** algorithm are close to those of the **Sd** algorithm differing at most by 0.1 in the median correlation

	Eyebrow	Eye	V Mouth	H Mouth
Red	down	closed		
Yellow		squint	wide	narrow
White	neutral	open	closed	relaxed
Green	raised	wide	open	wide

Table 7.2: Color coding of quantized face states.

value. The test data used was the MALE subset of Suvi.

The facial element estimations from the best classifiers and **Sd** landmarks were selected for qualitative evaluation. The selected classifiers were: Suvi NB for eyebrow, SVM for eye, Posed NB for vertical mouth, and Posed SVM for horizontal mouth.

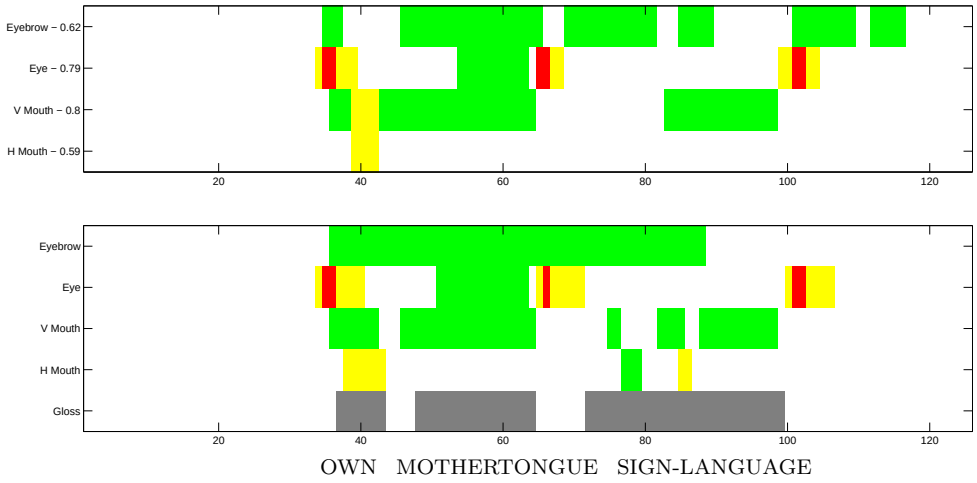
7.5 Qualitative FinSL experiment

For the qualitative FinSL experiment the eyebrow position, eye openness, and mouth state automatic estimations were compared against linguistic annotations of the same video data. Moreover, the results obtained from the previously selected facial state estimation model were codified into a set of color symbols and super-imposed in the videos for frame-level evaluation. A timeline representation of the results was also produced to enable evaluation against the manually annotated ground truth data.

The timeline representation shows a track for each facial element, with the mouth element divided into horizontal and vertical movements. The color coding of the tracks (also shown in Table 7.2) is {white, yellow, green, red} where white corresponds to the *neutral*, *open*, *relaxed* or *closed* states for eyebrow, eye or mouth, respectively. The yellow corresponds intermediate states like *squint*, *wide*, or *narrow* for eye, and vertical and horizontal mouth, respectively. The green color code corresponds the states *raised*, or *wide* for eyebrow and eye, *open* vertical mouth or *wide* horizontal mouth. The red color code is reserved to indicate the states *down*, and *closed* for eyebrow and eye, respectively. A set of colored symbols were super-imposed in the video to indicate the identified state of each facial element. They use the same color coding system as the timeline representation.

In Figure 7.13a an example video frame used for qualitative evaluation can be seen. Similarly, in Figure 7.13b the timeline plot of the estimated states of facial elements for the example video is shown, the MCC value appears next to the label of each timeline track. The MCC values for each element in the example video are thus: 0.62 for eyebrow, 0.79 for eyes, 0.8 for vertical mouth, and 0.59 for horizontal mouth state estimation. A median filter of 5

(a)



(b)

Figure 7.13: Suvi video 051703. (a) Frame 51 with super-imposed symbols representing the estimated states. (b) Top: timeline representation of estimations. Middle: ground truth reference annotations. Color coding for eyebrow as in Table 7.2, gray = sign gloss. Bottom: gloss transcript of the signed sentence 'My mother tongue is sign language' in Finnish SL.

frames of length has been applied to remove noisy detections. Timeline plots for the remaining tested videos can be found in Appendix B.

The qualitative set of experiments employed annotations from the point of

view of linguistic significance prepared for a subset of the Suvi material [41]. The linguistic annotations include events such as head tilts, nods, raised eyebrows, lowered eyebrows, blinks, and squints. No linguistic annotation has been made to mouth movements, nevertheless the identifiable mouthings were matched to spoken words. The annotations also include a translation by articulation and a its equivalent sentence. The glosses (articulation occurrences and equivalent translation) for all studies Suvi videos were included for the qualitative analysis.

In the included example (Figure 7.13), the eyebrow estimations coincided with the linguistic annotations except in the non-linguistic visual changes or perspective illusions (head tilting) visible in frames 103–117. The fading-out phase of the raised eyebrows in frames 72–88 is not deemed linguistically significant, but were still detected. For the eye openness estimations, the blinks were correctly detected around frames 38, 64 and 102, and the same holds for the widening of the eyes in frames 56–62. The mouth MCC is high in the vertical movements, activity was detected from frame 37 to 63, but the section in frames 83–98 showing open lips with closed jaw was only partially detected. The horizontal mouth movement estimation detected activity in frames 39–43, however, the latter frames were not detected.

7.6 Qualitative news broadcast experiment

The performance evaluation of the news broadcast experiment was qualitative since there is no available ground truth data (labeled facial states) of the subjects' facial gestures. The best classifiers and the landmarks from the **Sd** algorithm from the FinSL experiment were employed to generate automatic estimations for each detected face in a video frame. The estimations were again super-imposed into the videos for visual inspection. The super-imposed estimations consists of a set of symbols depicting the estimated eyebrow position, eye openness and mouth states. Timeline plots depicting the estimations were produced for each detected face.

In the example frame shown in Figure B.20, two subjects were detected, one of them a static face. The eyebrow movements of the news reader are clearly present during frame sequences 170–175, 200–300, and 335–426. The automatic estimations are noisy, however the first and last eyebrow raise sequences are identified. The eyebrow estimations fail during the second sequence, this is also the sequence with most neutral estimations. Regarding the eyes, the estimation correctly identified 5 blinks at frames 72, 133, 216, 267, and 334. By adjusting the intensity of the median filter the eye blinks become more noticeable. At frames 1–10 the head of the news presenter is

(a)

(b)

Figure 7.14: YLE sequence test. The color coding is the same as with previous experiments. (a) Example frame of NM estimation with super-imposed symbols representing states with more than one face appearing in the same frame. (b) timeline representation of the estimated states for each facial element for the two subjects.

tilted down, producing the illusion of closed eyes shown in the timeline plot. The vertical mouth estimations show opening and closing mouth estimations during most of the sequence. The estimations fail only during sudden short changes probably due to the postprocessing median filtering. The horizontal mouth estimations frequently had difficulties with the *narrow* state, for example in frames 300–312 a *wide* state is identified as *narrow* state. Due to the fast mouth movements, the observed imprecisions are possibly caused by

the median filtering.

In the timeline representation, the moment when the second face appears in the video can be observed. The plot of the second subject (bottom in Figure B.20b) shows that the estimations' color coding starts approximately at frame 26. In the same way, both subjects disappear from the video at frame 425 and the estimations become unavailable (shown in gray in the timeline representation).

The estimated face states of the static subject are constant: eyebrow lowered, squinted eyes, closed vertical mouth, wide horizontal mouth. This is a strong indicator that the face is not live. It is worth noticing that eye openness and horizontal mouth state estimations are consistently correct trough the sequence. The eyebrow position and vertical mouth state are not correctly estimated, possibly due to the non-frontal pose, nevertheless they do not fluctuate. Additional results from the news broadcast video experiment is shown in Appendix B.

7.7 Discussion

As the proposed features primarily rely on geometric characteristics of the face, the landmark locations should be accurate to obtain good results. The results of the different landmark detection algorithms show that **Sd** is the most precise single method. Nevertheless, the results obtained from the proposed **Em** algorithm have shown to be comparable with those of the **Sd** algorithm. It can also be observed that the basic patterns of movement for the eyebrow, eye, and mouth are identifiable with relative accuracy from the features prior PCA, even without a classification algorithm. The obtained models were precise enough detecting blinks and vertical mouth openings for unsupervised use, while the eyebrow and horizontal mouth movements showed mixed results requiring further work for reliable unsupervised applications.

In this work, the estimation of the eyebrows' position is dependent on the distance of the eyebrow landmarks from other facial elements and on the correct alignment of those elements. This may add noise to the estimated eyebrow positions since changing the head pose can alter the way the locations of the facial elements are perceived, adding uncertainty to the proposed approach. Making use of auxiliary head pose information could improve the results of the classification task in these circumstances.

Regarding eye openness, the assumption of symmetry makes it possible to identify eye states in situations where one eye is occluded by determining only the observable eye's activity. In all test sets the eye openness estimation

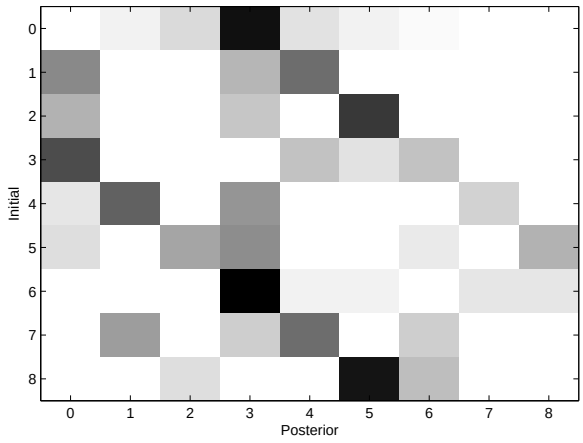


Figure 7.15: Mouth state transition sequences. The mouth state categories 0-8 are explained in Table 3.1, page 21. Each row indicates the most probable posterior state given the starting mouth state is the one represented by the row. Transitions to the same state are not shown.

was the most reliable, and even more in the binary classification. Taking into consideration the SVM classifier, the imbalanced dataset used for training, and the good results obtained, specialized kernels could be studied to further improve the results.

The MCC results for the full range (horizontal and vertical) mouth movements are relatively weak. It is possible to identify certain mouth positions as least occurring (*wide*,*narrow* for example), these could be ruled out or merged into a similar positions to increase the performance of the classifier. The mouth positions could also reveal patterns of movements or transitions between states. An example of this can be seen in Figure 7.15 where state transition pairs have been extracted from the manual annotations of the Suvi videos. The graph shows how often a specific transition occurs in all the studied Suvi material.

The vertical mouth estimations showed strong results, comparable to those achieved for eye openness. The performance difference between the proposed ensemble algorithm **Em** and **Sd** is narrow, the correlation result distributions overlap considerably. The horizontal mouth estimations were the most unstable amongst the evaluated face elements. The graphical evaluation shows the unreliability of the estimations as large distribution tails. This may be due to the fact that changes in the mouth width are small in most cases during vocalization, and can be mistakenly identified as landmark location noise. The vertical and horizontal mouth state classifiers were the most benefited from the Posed training data, possibly because this video set

included a more even distribution of samples in each class.

The use of MCC and box plots for evaluation of the experiments allowed visual analysis of the performance results for all test groups and the comparison of the different landmark estimation methods in each test. MCC provided a measure capable of representing both specificity and sensibility at the same time, shortening evaluation time. However, the interpretation of MCC may not be as straightforward as is the case of more traditional measures.

The facial state estimations obtained from the news broadcast videos evidenced a bias towards the learned parameters from the Suvi video training material. This may be because in the employed classification model the training was performed with the same subject's facial proportions. While the classifiers are able construct a good model, categories such as *wide* and *squint* have shown to be more difficult to generalize. This effect can be seen in Figure B.20 where the second subject's timeline shows that the still image is in a recurrent *down* eyebrow state even when it is clear she is not. This may, however, be due to the low scale of the image and head pose. The extracted features were calculated from subjects with observed small (squinted appearance) eyes, ultimately influencing the estimations. Adding new labeled training data to the classifiers can increase the applicability of the presented methods for unconstrained videos.

Chapter 8

Conclusions

Finding patterns in the manual and non-manual articulation of signs is a key in the study of the prosodic system (rhythm, intonation, and stress) in all sign languages. Automatic estimation and tracking of the activity of articulators can help linguists in the study of sign composition and in the gathering of quantitative data to support the ongoing research in the area. Whereas evidence of non-manuals as grammatical and prosody devices has been available from very early in the field of sign language studies, there is still much to learn on their organization and interaction with the sign language syntax.

Facial elements in sign language have often been associated as non-manual prosodic articulators. Facial expressions (such as eyebrow movements and eye blinks) can mark intonation and modify the meaning of the signed sentence. This type of intonational device has been observed in some sign languages functioning as a modulator of hand actions, providing further evidence that it is a part of the prosodic system. The emergence of mouth articulations and their consistency in sign language is still being studied: at the phonology (observed in mouthings) and the morphology (observed in mouth gestures) levels: Their occurrence is widespread in sign languages.

This work has studied techniques for eye and mouth openness estimation in sign language and news broadcast videos. The research aimed at these two applications by employing specific video material for each one: FinSL videos to aid linguists in the automatic estimation of non-manual facial articulators, and news broadcast videos to test the estimation tool and explore its use as an early speaker identifier. The study presented here has proposed a categorization for the states of each facial element, studied different landmark detection algorithms, proposed a feature extraction scheme for each facial element, and proposed an evaluation framework to determine the performance of the trained classifiers. Using the feature extraction scheme it was pos-

sible for the classifiers to capture eye openness and mouth states precisely enough to enable quantitative studies of phonetics, using a feature vector of 10 elements, which provides an answer to the first research question: "Can facial landmarks be used effectively for eye and mouth openness estimation?". This indicates by observing the final experiment results that it is possible to extract efficient features for openness estimation. The feasibility of using the estimation tool for eye and mouth openness was established using just three video sequences as training data. The news broadcast experiment also provided good results, successfully answering the second research question of this work: "Is it feasible to build a model for the automatic annotation of eye and mouth states in SL videos?".

The obtained classification performance using the **Sd** landmark detection algorithm indicated that this algorithm was more precise at landmark detection than the proposed ensemble algorithm **Em**. Nevertheless, the difference is small enough to consider the overall results still comparable. The proposed landmark detection algorithms can be further improved using different noise removal methods, or by using different feature dimensionality reduction algorithms on the proposed feature vectors. The segmentation method **G1** showed good results with frontal faces; admittedly, it will perform badly if the face is not well aligned.

The third research question "Are the annotations produced by the model reliable?" considered the practical application of the obtained model. Eye openness and vertical mouth estimations proved to be the most reliable. Moreover, the **Cr** algorithm can be used be used in the ensemble **Em** algorithm (or individually for the **Sd** algorithm) for the estimation of gaze. For all the presented methods the assumption of eye symmetry made it possible to identify any stage in situations where only one eye is visible by determining the activity of the observable eye. The eyebrow position and horizontal mouth estimations showed noisy results and low reliability during the evaluation. The Naive Bayes and SVM classifiers were found to produce very similar accuracies. With the used multiclass MCC performance measure, the variance of the results between the test set videos was found to be large particularly for the eyebrow and horizontal mouth estimations.

The categorization scheme for mouth states proved to be useful. From the manual annotations using the proposed categorizations, it can be seen that some mouth states are infrequent in the studied Suvi videos. The frequency of transitions between these states shows patterns of movements that could be statistically studied further. This would facilitate to study mouth state estimation as an aid towards automatic discrimination between mouthings and mouth gestures. The application of automatic mouth state estimation to other unconstrained videos (different signed and spoken languages) could fur-

ther reveal the feasibility of a generalized mouth state categorization scheme.

To answer the fourth research question "Is it possible to automatically estimate the eyebrow position?", the proposed recognition method for the eyebrow position did not provide conclusive results over its practical use in automatic detection of eyebrow shifts. Further work would be needed to develop more reliable methods for the recognition of eyebrow states. For example, a different dimensionality reduction technique or noise removal filter could add stability to the eyebrow position estimations.

The news broadcast video material was used to explore eye and mouth openness for liveness detection, aiming at further application to speaker identification. The results obtained in the news broadcast experiment using the estimation model from the FinSL experiment provided evidence of the applicability of the estimation model in more unconstrained videos than the FinSL material used in training. Even with the simple face alignment algorithm employed, estimations can be obtained as well as identify liveness by observing the timeline plots, this provides an answer to the fifth research question: "Can a model for automatic eye and mouth openness annotation be used also for news broadcast video analysis".

Future work can focus on regression analysis instead of classification to obtain a more precise scale of estimations in each possible facial element. This could prove beneficial, for example, if the exact degree of eye openness is of interest. Another research path could explore different classification algorithms, or specialized SVM kernels for highly imbalanced data. Lastly, the information of previous frames could be incorporated to boost new estimations and lead to the detection of actions instead of per-frame states.

Bibliography

[1] Abtahi, S., Hariri, B., and Shirmohammadi, S. Driver drowsiness monitoring based on yawning detection. In *Instrumentation and Measurement Technology Conference (I2MTC), 2011 IEEE* (2011), IEEE, pp. 1–4.

[2] Alpaydin, E. *Introduction to machine learning*, second ed. MIT press, 2010.

[3] Araujo, R., Miao, Y.-Q., Kamel, M. S., and Cheriet, M. A fast and robust feature set for cross individual facial expression recognition. In *Computer Vision and Graphics*. Springer, 2012, pp. 272–279.

[4] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems 16*, 6 (2010), 345–379.

[5] Bacivarov, I., Ionita, M., and Corcoran, P. Statistical models of appearance for eye tracking and eye-blink detection and measurement. *Consumer Electronics, IEEE Transactions on 54*, 3 (2008), 1312–1320.

[6] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*, 5 (2000), 412–424.

[7] Barr, J. R., Bowyer, K. W., and Flynn, P. J. Detecting questionable observers using face track clustering. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on* (2011), IEEE, pp. 182–189.

[8] Benoit, A., and Caplier, A. Hypovigilence analysis: open or closed eye or mouth? blinking or yawning frequency? In *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on* (2005), IEEE, pp. 207–212.

[9] bin Abdul Rahman, N. A., Wei, K. C., and See, J. RGB-H-CbCr skin colour model for human face detection. *Faculty of Information Technology, Multimedia University* (2006).

[10] Black, M. J., and Yacoob, Y. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Computer Vision, 1995. Proceedings., Fifth International Conference on* (1995), IEEE, pp. 374–381.

[11] Bradski, G. *Dr. Dobb's Journal of Software Tools*.

[12] Bulling, A., Ward, J. A., Gellersen, H., and Troster, G. Eye movement analysis for activity recognition using electrooculography. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33*, 4 (2011), 741–753.

[13] Campr, P., Dikici, E., Hruz, M., Kindiroglu, A., Krnoul, Z., Ronzhin, A., Sak, H., Schorno, D., Akarun, L., Aran, O., et al. Automatic fingersign to speech translator. *Proceedings of eNTERFACE* (2010).

[14] Canzler, U., and Dziurzyk, T. Extraction of non manual features for videobased sign language recognition. In *IAPR workshop on machine vision applications* (2002), pp. 318–321.

[15] Caruana, R., Karampatziakis, N., and Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (2008), ACM, pp. 96–103.

[16] Caruana, R., and Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 161–168.

[17] Chang, C.-C., and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3 (2011), 27.

[18] Cootes, T. F., Edwards, G. J., and Taylor, C. J. Active appearance models. In *Computer Vision-ECCV'98*. Springer, 1998, pp. 484–498.

[19] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[20] Cristinacce, D., and Cootes, T. F. Feature detection and tracking with constrained local models. In *BMVC* (2006), vol. 17, pp. 929–938.

[21] Divjak, M., and Bischof, H. Real-time video-based eye blink analysis for detection of low blink-rate during computer use. In *First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS 2008)* (2008), pp. 99–107.

[22] Dollár, P., Welinder, P., and Perona, P. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 1078–1085.

[23] Dong, Y., and Woodard, D. L. Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study. In *Biometrics (IJCB), 2011 International Joint Conference on* (2011), IEEE, pp. 1–8.

[24] Dreuw, P., Forster, J., Gweth, Y., Stein, D., Ney, H., Martinez, G., Llahi, J. V., Crasborn, O., Ormel, E., Du, W., et al. Signspeak–understanding, recognition, and translation of sign languages. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (2010), pp. 22–23.

[25] Ekman, P., and Friesen, W. V. Facial action coding system: A technique for the measurement of facial movement.

[26] European Commission. Europe: a continent of many sign languages. `http://ec.europa.eu/languages/languages-of-europe/sign-languages_en.htm`, August 2013.

[27] Eveno, N., Caplier, A., and Coulon, P.-Y. New color transformation for lips segmentation. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on* (2001), IEEE, pp. 3–8.

[28] Eveno, N., Caplier, A., and Coulon, P.-Y. Accurate and quasi-automatic lip tracking. *Circuits and Systems for Video technology, IEEE Transactions on 14*, 5 (2004), 706–715.

[29] Finnish Association for the Deaf. Sign language users. `http://www.kl-deaf.fi/`, August 2013.

[30] Fisher, R. A. The design of experiments.

[31] Gómez-Mendoza, J.-B. *A contribution to mouth structure segmentation in images aimed towards automatic mouth gesture recognition.* PhD thesis, L'institut national des sciences appliquées de Lyon, 2012.

[32] Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry 28*, 5 (2004), 367–374.

[33] Grammer, K., and Thornhill, R. Human (homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of comparative psychology 108*, 3 (1994), 233.

[34] Grauman, K., Betke, M., Lombardi, J., Gips, J., and Bradski, G. R. Communication via eye blinks and eyebrow raises: Video-based human-computer interfaces. *Universal Access in the Information Society 2*, 4 (2003), 359–373.

[35] Hansen, D. W., and Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 32*, 3 (2010), 478–500.

[36] Haro, A., Flickner, M., and Essa, I. Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on* (2000), vol. 1, IEEE, pp. 163–168.

[37] Herrmann, A., and Steinbach, M. *Nonmanuals in Sign Language.* Benjamins Current Topics. John Benjamins Publishing Company, 2013.

[38] Ho, S. S., and McLeod, D. M. Social-psychological influences on opinion expression in face-to-face and computer-mediated communication. *Communication Research 35*, 2 (2008), 190–207.

[39] Hu, W., Xie, N., Li, L., Zeng, X., and Maybank, S. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41*, 6 (2011), 797–819.

[40] Hunt, R. W. G. *The reproduction of colour.* Wiley, 2005.

[41] Jantunen, T. The equative sentence in Finnish Sign Language. *Sign Language & Linguistics 10*, 2 (2007), 113–143.

[42] Jantunen, T., and Takkinen, R. Syllable structure in sign language phonology, 2010.

[43] Jobson, D. J., Rahman, Z.-u., and Woodell, G. A. Properties and performance of a center/surround retinex. *Image Processing, IEEE Transactions on 6*, 3 (1997), 451–462.

[44] Jurman, G., and Furlanello, C. A unifying view for performance measures in multi-class prediction. *arXiv preprint arXiv:1008.2908* (2010).

[45] Jurman, G., Riccadonna, S., and Furlanello, C. A comparison of MCC and CEN error measures in multi-class prediction. *PloS one 7*, 8 (2012), e41882.

[46] Karppa, M., Jantunen, T., Viitaniemi, V., Laaksonen, J., Burger, B., and Weerdt, D. D. Comparing computer vision analysis of signed language video with motion capture recordings. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (Istanbul, Turkey, may 2012), European Language Resources Association (ELRA).

[47] Karppa, M., Viitaniemi, V., Luzardo, M., Laaksonen, J., and Jantunen, T. SLMotion: An extensible sign language oriented video analysis tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'14)* (2014), European Language Resources Association (ELRA).

[48] Kim, J. Y., Na, S. Y., and Cole, R. Lip detection using confidence-based adaptive thresholding. In *Advances in Visual Computing*. Springer, 2006, pp. 731–740.

[49] Kovac, J., Peer, P., and Solina, F. *Human skin color clustering for face detection*, vol. 2. IEEE, 2003.

[50] Królak, A., and Strumiłło, P. Eye-blink detection system for human–computer interaction. *Universal Access in the Information Society 11*, 4 (2012), 409–419.

[51] Lalonde, M., Byrns, D., Gagnon, L., Teasdale, N., and Laurendeau, D. Real-time eye blink detection with GPU-based SIFT tracking. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on* (2007), IEEE, pp. 481–487.

[52] Lan, Y., Theobald, B.-J., Harvey, R., Ong, E.-J., and Bowden, R. Improving visual features for lip-reading. In *Proceedings of*

International Conference on Auditory-Visual Speech Processing (2010), vol. 201.

[53] Land, E. H., McCann, J. J., et al. Lightness and retinex theory. *Journal of the Optical society of America 61*, 1 (1971), 1–11.

[54] Lee Rodgers, J., and Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician 42*, 1 (1988), 59–66.

[55] Liévin, M., and Luthon, F. Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video. *Image Processing, IEEE Transactions on 13*, 1 (2004), 63–71.

[56] Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., and Neidle, C. Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL. In *IEEE International Conference on Automatic Face and Gesture Recognition* (2013).

[57] Liu, Z., and Liu, C. Fusion of color, local spatial and global frequency information for face recognition. *Pattern Recognition 43*, 8 (2010), 2882–2890.

[58] Lowe, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (1999), vol. 2, Ieee, pp. 1150–1157.

[59] Loy, G., and Zelinsky, A. Fast radial symmetry for detecting points of interest. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 25*, 8 (2003), 959–973.

[60] Luzardo, M., Karppa, M., Laaksonen, J., and Jantunen, T. Head pose estimation for sign language video. In *Image Analysis, 2013. SCIA 2013. 18th Scandinavian Conference on* (2013), Springer Berlin Heidelberg, pp. 349–360.

[61] MATLAB. *version 8.0 (R2012b)*. The MathWorks Inc., Natick, Massachusetts, 2012.

[62] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure 405*, 2 (1975), 442–451.

[63] McGill, R., Tukey, J. W., and Larsen, W. A. Variations of box plots. *The American Statistician 32*, 1 (1978), 12–16.

[64] Metaxas, D., Liu, B., Yang, F., Yang, P., Michael, N., and Neidle, C. Recognition of nonmanual markers in american sign language (ASL) using non-parametric adaptive 2D-3D face tracking. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (Istanbul, Turkey, may 2012), European Language Resources Association (ELRA).

[65] Minkov, K., Zafeiriou, S., and Pantic, M. A comparison of different features for automatic eye blinking detection with an application to analysis of deceptive behavior. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on* (2012), IEEE, pp. 1–4.

[66] Otsu, N. A threshold selection method from gray-level histograms. *Automatica 11*, 285-296 (1975), 23–27.

[67] Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A. F., and Queenot, G. TRECVID 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2013* (2013), NIST, USA.

[68] Ozkan, D., and Duygulu, P. Interesting faces: A graph-based approach for finding people in news. *Pattern Recognition 43*, 5 (2010), 1717–1735.

[69] Pan, J., Guan, Y., and Wang, S. A new color transformation based fast outer lip contour extraction. *Journal of Information & Computational Science 9* (2012), 2505–2514.

[70] Panning, A., Niese, R., Al-Hamadi, A., Michaelis, B., and Inroduction, I. A new adaptive approach for histogram based mouth segmentation. *Proceedings of the World Academy of Science, Engineering and Technology 56* (2009), 779–784.

[71] Pantti, M. The value of emotion: An examination of television journalists' notions on emotionality. *European Journal of Communication 25*, 2 (2010), 168–181.

[72] Pearson, K. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*, 11 (1901), 559–572.

[73] Petitta, G., Sallandre, M.-A., and Rossini, P. Mouth gestures and mouthing in two sign languages (LIS and LSF). *Theoretical Issues in Sign Language Research (TISLR XI)* (2011).

[74] Pfau, R., and Quer, J. Nonmanuals: their grammatical and prosodic roles. *Sign languages (Cambridge Language Surveys)* (2010), 381–402.

[75] Pfau, R., Steinbach, M., and Woll, B. *Sign language: an international handbook*, vol. 37. Walter de Gruyter, 2012.

[76] Picot, A., Caplier, A., and Charbonnier, S. Comparison between EOG and high frame rate camera for drowsiness detection. In *Applications of Computer Vision (WACV), 2009 Workshop on* (2009), IEEE, pp. 1–6.

[77] Picot, A., Charbonnier, S., and Caplier, A. Drowsiness detection based on visual signs: blinking analysis based on high frame rate video. In *Instrumentation and Measurement Technology Conference (I2MTC), 2010 IEEE* (2010), IEEE, pp. 801–804.

[78] Picot, A., Charbonnier, S., Caplier, A., and Vu, N.-S. Using retina modelling to characterize blinking: comparison between EOG and video analysis. *Machine Vision and Applications 23*, 6 (2012), 1195–1208.

[79] Powers, D. Evaluation: From precision, recall and F-measure to ROC., informedness, markedness & correlation. *Journal of Machine Learning Technologies 2*, 1 (2011), 37–63.

[80] Prendergast, P. M. Facial proportions. In *Advanced Surgical Facial Rejuvenation*. Springer, 2012, pp. 15–22.

[81] Rinfret, J., Parisot, A.-M., Szymoniak, K., and Villeneuve, S. Methodological issues in automatic recognition of pointing signs in Langue des signes Québécoise using a 3D motion capture system. *Theoretical Issues in Sign Language Research (TISLR XI)* (2011).

[82] Soetedjo, A., Yamada, K., and Limpraptono, F. Y. Lip detection based-on normalized RGB chromaticity diagram. In *6th International Conference on Information & Communication Technology and Systems* (2010), pp. 63–67.

[83] Sokolova, M., Japkowicz, N., and Szpakowicz, S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence*. Springer, 2006, pp. 1015–1021.

[84] Sokolova, M., and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management 45*, 4 (2009), 427–437.

[85] Stein, D., Schmidt, C., and Ney, H. Analysis, preparation, and optimization of statistical sign language machine translation. *Machine translation 26*, 4 (2012), 325–357.

[86] Stillittano, S., Girondel, V., and Caplier, A. Lip contour segmentation and tracking compliant with lip-reading application constraints. *Machine Vision and Applications 24*, 1 (2013), 1–18.

[87] Tang, F., and Deng, B. Facial expression recognition using AAM and local facial features. In *Natural Computation, 2007. ICNC 2007. Third International Conference on* (2007), vol. 2, IEEE, pp. 632–635.

[88] Tian, Y., Kanade, T., and Cohn, J. F. Facial expression recognition. In *Handbook of face recognition*, second ed. springer, 2011, ch. 19, pp. 487–519.

[89] Timm, F., and Barth, E. Accurate eye centre localisation by means of gradients. In *VISAPP* (2011), pp. 125–130.

[90] Uřičář, M., Franc, V., and Hlaváč, V. Detector of facial landmarks learned by the structured output SVM. In *VISAPP* (2012), SciTePress — Science and Technology Publications, pp. 547–556.

[91] Vinciarelli, A., and Mohammadi, G. Towards a technology of nonverbal communication: Vocal behavior in social and affective phenomena. *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives* (2011), 133–156.

[92] Vinciarelli, A., Pantic, M., and Bourlard, H. Social signal processing: Survey of an emerging domain. *Image and Vision Computing 27*, 12 (2009), 1743–1759.

[93] Viola, P., and Jones, M. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (2001), vol. 1, IEEE, pp. I–511.

[94] Štruc, V., and Others. Inface: A toolbox for illumination invariant face recognition.

[95] Wang, L., Ding, X., Fang, C., Liu, C., and Wang, K. Eye blink detection based on eye contour extraction. In *Image Processing: Algorithms and Systems* (2009), p. 72450.

[96] Wang, S.-L., Lau, W.-H., Liew, A. W.-C., and Leung, S.-H. Robust lip region segmentation for lip images with complex background. *Pattern Recognition 40*, 12 (2007), 3481–3491.

[97] Wilbur, R. B. Effects of varying rate of signing on ASL manual signs and nonmanual markers. *Language and speech 52*, 2-3 (2009), 245–285.

[98] Xiong, X., and De la Torre, F. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (2013), IEEE, pp. 532–539.

[99] Yang, F., Yu, X., Huang, J., Yang, P., and Metaxas, D. Robust eyelid tracking for fatigue detection. In *Image Processing (ICIP), 2012 19th IEEE International Conference on* (2012), IEEE, pp. 1829–1832.

[100] Zhang, W., Chen, H., Yao, P., Li, B., and Zhuang, Z. Precise eye localization with AdaBoost and fast radial symmetry. In *Computational Intelligence and Security*. Springer, 2007, pp. 1068–1077.

Appendix A

Classification performance

The following tables summarize the performance of the extracted features using each landmark estimation algorithm. The acronyms used here are as described in Chapter 6.4, with the exception of Br (eyebrow), Ey (eye), and Mo (mouth) this being either VMo (vertical mouth) or HMo (horizontal mouth) . The specificity and sensibility for the multiclass classification shows the average over all individual sensibility and specificity values per class.

A.1 Suvi training performance

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.101	.455	.728	.029	.015	.006	.133	.558	.607	.078	.091	.021
	Pr+Gr/G2	.254	.574	.802	.047	.023	.005	.329	.793	.649	.057	.050	.017
	Pr+Bi/G1	.136	.489	.762	.020	.030	.005	.232	.600	.699	.035	.044	.021
	Pr+Bi/G2	.236	.573	.791	.045	.022	.006	.310	.624	.759	.055	.079	.017
	Sd	.293	.611	.806	.048	.017	.005	.346	.620	.795	.067	.063	.008
Ey	Cr/G1	.456	.575	.873	.051	.030	.003	.645	.704	.934	.026	.049	.002
	Cr/G2	.420	.507	.867	.025	.010	.002	.532	.823	.790	.018	.012	.009
	Pr/G1	.430	.566	.867	.037	.020	.003	.589	.698	.865	.025	.031	.029
	Pr/G2	.531	.658	.895	.037	.025	.002	.670	.862	.866	.020	.017	.009
	Sd	.559	.736	.895	.033	.009	.003	.720	.775	.915	.020	.024	.013
VMo	Gr/G1	.221	.423	.771	.028	.012	.005	.452	.735	.734	.049	.016	.044
	Gr/G2	.457	.632	.845	.015	.016	.002	.605	.800	.829	.020	.009	.026
	Bi/G1	.260	.501	.756	.020	.012	.002	.367	.825	.518	.028	.010	.037
	Bi/G2	.415	.597	.815	.018	.021	.004	.530	.861	.665	.028	.008	.040
	Sd	.569	.718	.880	.010	.010	.001	.745	.862	.900	.018	.008	.012
HMo	Gr/G1	.135	.492	.749	.022	.039	.003	.190	.632	.713	.077	.121	.027
	Gr/G2	.323	.642	.812	.051	.030	.005	.202	.577	.796	.059	.115	.015
	Bi/G1	.236	.578	.773	.045	.039	.008	.210	.357	.910	.068	.141	.007
	Bi/G2	.274	.636	.795	.038	.034	.008	.351	.782	.855	.058	.078	.013
	Sd	.426	.733	.844	.037	.016	.002	.425	.790	.896	.073	.050	.008

Table A.1: NB results for the MALE test set trained with Suvi videos.

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.090	.519	.742	.034	.027	.008	.111	.334	.812	.048	.071	.004
	Pr+Gr/G2	.229	.583	.785	.033	.020	.005	.264	.529	.788	.041	.057	.009
	Pr+Bi/G1	.156	.540	.753	.025	.024	.006	.203	.383	.827	.043	.078	.012
	Pr+Bi/G2	.217	.576	.779	.049	.027	.008	.274	.402	.890	.067	.091	.008
	Sd	.182	.598	.772	.038	.049	.014	.201	.254	.956	.051	.060	.002
Ey	Cr/G1	.341	.473	.824	.051	.020	.004	.440	.382	.964	.059	.046	.001
	Cr/G2	.454	.525	.857	.039	.017	.003	.622	.647	.952	.037	.046	.002
	Pr/G1	.393	.478	.825	.029	.008	.002	.428	.336	.964	.023	.028	.005
	Pr/G2	.549	.556	.876	.019	.016	.002	.719	.675	.978	.019	.036	.001
	Sd	.616	.680	.894	.031	.013	.002	.726	.705	.959	.008	.017	.003
VMo	Gr/G1	.003	.389	.667	.000	.006	.000	.017	1.00	.007	.003	.000	.000
	Gr/G2	.383	.567	.801	.022	.016	.004	.533	.886	.612	.027	.006	.050
	Bi/G1	.307	.524	.771	.020	.011	.002	.402	.842	.539	.028	.008	.029
	Bi/G2	.433	.606	.818	.015	.020	.004	.539	.864	.669	.023	.008	.039
	Sd	.612	.705	.883	.018	.016	.002	.737	.878	.863	.015	.008	.022
HMo	Gr/G1	.164	.499	.749	.022	.028	.006	.115	.110	.967	.035	.011	.001
	Gr/G2	.297	.567	.783	.056	.027	.009	.266	.272	.986	.103	.078	.001
	Bi/G1	.285	.608	.787	.051	.030	.009	.366	.549	.928	.109	.119	.005
	Bi/G2	.376	.669	.821	.069	.038	.009	.327	.513	.948	.080	.093	.002
	Sd	.239	.526	.752	.089	.038	.013	.000	.000	1.00	.000	.000	.000

Table A.2: SVM results for the MALE test set trained with Suvi videos.

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.097	.396	.701	.004	.029	.000	.138	.930	.173	.007	.005	.007
	Pr+Gr/G2	.029	.151	.664	.007	.005	.001	.117	.898	.164	.023	.012	.002
	Pr+Bi/G1	.025	.145	.685	.004	.009	.002	.158	.932	.210	.028	.014	.009
	Pr+Bi/G2	.040	.111	.687	.000	.005	.000	.185	.944	.206	.015	.004	.014
	Sd	.041	.490	.674	.026	.004	.002	.068	.179	.873	.059	.023	.016
Ey	Cr/G1	.124	.338	.777	.023	.004	.002	.197	.335	.838	.103	.052	.027
	Cr/G2	.178	.376	.795	.008	.002	.000	.221	.652	.562	.036	.057	.043
	Pr/G1	.079	.352	.762	.009	.002	.001	.048	.724	.335	.074	.029	.115
	Pr/G2	.140	.375	.796	.001	.005	.001	.270	.670	.585	.010	.034	.057
	Sd	.153	.589	.791	.065	.026	.006	.327	.993	.321	.066	.000	.077
VMo	Gr/G1	.303	.547	.790	.012	.007	.001	.591	.806	.793	.025	.007	.006
	Gr/G2	.357	.635	.788	.009	.010	.001	.400	.776	.621	.016	.004	.004
	Bi/G1	.293	.524	.770	.003	.009	.000	.432	.833	.573	.005	.008	.011
	Bi/G2	.368	.529	.780	.011	.009	.001	.445	.904	.493	.011	.001	.012
	Sd	.604	.787	.882	.007	.001	.001	.709	.867	.848	.008	.002	.002
HMo	Gr/G1	.294	.549	.780	.001	.009	.001	.397	.753	.731	.011	.015	.009
	Gr/G2	.408	.634	.810	.025	.022	.005	.559	.615	.907	.006	.039	.009
	Bi/G1	.217	.467	.711	.054	.024	.005	.303	.339	.924	.099	.062	.010
	Bi/G2	.361	.616	.795	.023	.040	.007	.646	.783	.898	.002	.023	.006
	Sd	.169	.504	.708	.013	.007	.002	.444	.350	.987	.067	.043	.001

Table A.3: NB results for the FEMALE test set trained with Suvi videos.

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.077	.250	.691	.007	.020	.001	.137	.795	.321	.020	.021	.028
	Pr+Gr/G2	-.020	.368	.655	.024	.003	.004	-.028	.374	.605	.023	.038	.003
	Pr+Bi/G1	.106	.303	.714	.014	.008	.002	.224	.822	.402	.068	.031	.024
	Pr+Bi/G2	.108	.194	.714	.007	.017	.001	.225	.869	.344	.021	.017	.037
	Sd	-.014	.494	.663	.001	.000	.000	-.028	.000	.988	.002	.000	.000
Ey	Cr/G1	.092	.307	.771	.004	.001	.000	.245	.212	.972	.049	.039	.001
	Cr/G2	.102	.335	.773	.023	.008	.002	.078	.412	.662	.061	.049	.056
	Pr/G1	.015	.297	.755	.016	.001	.001	-.007	.370	.657	.060	.092	.150
	Pr/G2	.177	.406	.796	.007	.011	.001	.288	.546	.723	.030	.059	.013
	Sd	.305	.661	.826	.039	.017	.005	.455	.954	.471	.034	.002	.092
VMo	Gr/G1	.000	.375	.667	.000	.005	.000	.000	1.00	.000	.000	.000	.000
	Gr/G2	.225	.531	.737	.008	.008	.001	.274	.804	.449	.014	.005	.010
	Bi/G1	.252	.470	.751	.006	.007	.001	.354	.829	.495	.011	.008	.011
	Bi/G2	.340	.506	.773	.012	.009	.002	.436	.908	.479	.015	.001	.016
	Sd	.611	.725	.875	.016	.013	.002	.676	.888	.785	.015	.001	.010
HMo	Gr/G1	.269	.541	.764	.014	.019	.001	.180	.197	.967	.045	.047	.001
	Gr/G2	.196	.480	.719	.081	.030	.009	.354	.299	.970	.045	.037	.001
	Bi/G1	.383	.656	.816	.074	.053	.012	.463	.760	.813	.061	.172	.013
	Bi/G2	.428	.616	.799	.018	.015	.002	.542	.547	.946	.008	.009	.001
	Sd	.054	.425	.671	.019	.004	.000	.000	.000	1.00	.000	.000	.000

Table A.4: SVM results for the FEMALE test set trained with Suvi videos.

A.2 Posed training performance

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.190	.554	.769	.050	.033	.006	.245	.590	.712	.044	.024	.018
	Pr+Gr/G2	.274	.632	.797	.109	.044	.015	.289	.387	.910	.115	.058	.015
	Pr+Bi/G1	.340	.642	.810	.069	.034	.007	.382	.651	.778	.074	.020	.033
	Pr+Bi/G2	.156	.487	.735	.050	.021	.006	.193	.519	.679	.051	.041	.021
	Sd	.008	.464	.699	.034	.026	.008	.020	.368	.668	.033	.090	.033
Ey	Cr/G1	.137	.364	.784	.032	.019	.002	.195	.961	.219	.038	.003	.036
	Cr/G2	.137	.421	.787	.032	.018	.002	.263	.954	.326	.036	.008	.028
	Pr/G1	.140	.358	.779	.027	.005	.003	.504	.437	.950	.032	.052	.008
	Pr/G2	.255	.402	.815	.041	.017	.004	.662	.580	.985	.027	.042	.001
	Sd	.386	.579	.859	.020	.015	.001	.401	.762	.683	.048	.019	.025
VMo	Gr/G1	.165	.376	.748	.014	.010	.003	.344	.511	.845	.034	.028	.027
	Gr/G2	.449	.640	.845	.009	.015	.001	.610	.740	.898	.017	.012	.012
	Bi/G1	.146	.444	.734	.024	.012	.003	.305	.407	.888	.036	.036	.009
	Bi/G2	.304	.520	.808	.016	.011	.003	.549	.665	.917	.022	.016	.010
	Sd	.611	.687	.879	.017	.015	.002	.734	.874	.857	.014	.008	.024
HMo	Gr/G1	.297	.595	.795	.042	.027	.008	.405	.730	.881	.101	.094	.010
	Gr/G2	.435	.713	.862	.058	.031	.005	.299	.702	.836	.066	.111	.021
	Bi/G1	.207	.517	.780	.010	.015	.002	.261	.887	.625	.049	.089	.031
	Bi/G2	.321	.608	.844	.021	.014	.003	.293	.984	.654	.033	.001	.020
	Sd	.311	.536	.762	.054	.034	.009	.342	.392	.980	.118	.105	.001

Table A.5: NB results for the MALE test set trained with the Posed videos.

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.217	.553	.783	.059	.036	.007	.290	.573	.774	.055	.045	.016
	Pr+Gr/G2	.224	.612	.793	.071	.033	.010	.234	.678	.671	.063	.060	.017
	Pr+Bi/G1	.114	.541	.736	.018	.015	.004	.118	.471	.680	.017	.020	.009
	Pr+Bi/G2	.049	.473	.704	.010	.005	.002	.076	.398	.672	.015	.031	.014
	Sd	.092	.541	.721	.057	.022	.010	.066	.583	.526	.060	.078	.027
Ey	Cr/G1	.358	.472	.845	.047	.018	.004	.504	.763	.807	.061	.067	.012
	Cr/G2	.230	.448	.805	.036	.020	.003	.305	.878	.486	.034	.019	.037
	Pr/G1	.173	.311	.781	.023	.005	.001	.349	.222	.981	.039	.028	.002
	Pr/G2	.214	.351	.813	.038	.007	.003	.339	.527	.809	.080	.033	.025
	Sd	.442	.572	.868	.032	.011	.003	.737	.668	.983	.006	.015	.001
VMo	Gr/G1	.275	.456	.789	.027	.022	.005	.456	.718	.756	.040	.024	.054
	Gr/G2	.486	.626	.846	.023	.022	.003	.600	.777	.849	.021	.009	.021
	Bi/G1	.024	.359	.687	.038	.018	.005	.133	.385	.753	.059	.034	.023
	Bi/G2	.447	.652	.847	.014	.009	.002	.566	.736	.875	.020	.008	.006
	Sd	.591	.684	.880	.016	.017	.002	.710	.788	.942	.013	.018	.006
HMo	Gr/G1	.278	.539	.770	.037	.032	.009	.337	.469	.931	.080	.062	.008
	Gr/G2	.321	.582	.833	.033	.022	.003	.102	.247	.889	.025	.032	.007
	Bi/G1	.249	.579	.774	.031	.020	.005	.294	.483	.904	.085	.123	.007
	Bi/G2	.203	.519	.742	.026	.015	.006	.289	.412	.921	.055	.043	.010
	Sd	.446	.624	.819	.079	.039	.010	.201	.216	.979	.081	.086	.001

Table A.6: SVM results for the MALE test set trained with the Posed videos.

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	-.031	.154	.636	.009	.001	.002	-.163	.747	.101	.003	.027	.006
	Pr+Gr/G2	.075	.375	.678	.004	.015	.003	.083	.753	.319	.087	.050	.035
	Pr+Bi/G1	.207	.613	.746	.007	.004	.001	.189	.792	.435	.012	.033	.011
	Pr+Bi/G2	.065	.305	.717	.001	.014	.001	.186	.771	.440	.017	.035	.010
	Sd	.031	.502	.668	.055	.011	.005	.027	.230	.774	.056	.017	.052
Ey	Cr/G1	-.016	.314	.748	.005	.002	.000	.053	.939	.103	.029	.002	.008
	Cr/G2	-.018	.238	.749	.003	.005	.000	.138	.949	.150	.041	.005	.014
	Pr/G1	-.011	.195	.751	.028	.008	.002	.016	.315	.700	.079	.047	.144
	Pr/G2	.135	.318	.780	.015	.016	.002	.299	.376	.900	.062	.076	.007
	Sd	.091	.424	.769	.022	.017	.001	.222	1.00	.161	.025	.000	.013
VMo	Gr/G1	.355	.551	.807	.011	.005	.001	.642	.735	.920	.019	.008	.005
	Gr/G2	.523	.755	.849	.005	.004	.001	.597	.798	.790	.007	.007	.041
	Bi/G1	.178	.452	.742	.066	.047	.007	.389	.479	.889	.034	.072	.005
	Bi/G2	.384	.566	.817	.014	.022	.001	.590	.793	.803	.007	.005	.006
	Sd	.631	.748	.890	.006	.011	.001	.713	.831	.896	.007	.001	.005
HMo	Gr/G1	.592	.692	.848	.026	.028	.003	.657	.708	.939	.013	.005	.001
	Gr/G2	.364	.606	.795	.027	.023	.005	.525	.664	.871	.013	.025	.008
	Bi/G1	.384	.646	.815	.027	.037	.006	.432	1.00	.546	.015	.000	.035
	Bi/G2	.512	.713	.874	.003	.023	.001	.586	.975	.761	.001	.002	.002
	Sd	.390	.534	.752	.003	.006	.001	.425	.324	.959	.005	.017	.004

Table A.7: NB results for the FEMALE test set trained with the Posed videos.

Extraction Method		All Classes						Two Classes					
		Averages			Variances			Averages			Variances		
		Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe	Mcc	Sen	Spe
Br	Pr+Gr/G1	.022	.181	.665	.001	.003	.001	.021	.761	.225	.016	.024	.001
	Pr+Gr/G2	.038	.220	.649	.013	.007	.004	.059	.679	.326	.113	.070	.034
	Pr+Bi/G1	-.151	.460	.642	.046	.017	.008	-.173	.183	.737	.046	.094	.004
	Pr+Bi/G2	-.163	.418	.619	.021	.002	.001	-.197	.019	.828	.016	.000	.011
	Sd	.246	.577	.719	.056	.026	.011	.231	.750	.403	.068	.188	.020
Ey	Cr/G1	.196	.394	.798	.044	.013	.003	.230	.413	.804	.057	.034	.016
	Cr/G2	.224	.413	.809	.018	.005	.001	.306	.818	.484	.028	.022	.013
	Pr/G1	-.018	.304	.744	.042	.004	.002	-.014	.246	.735	.052	.066	.163
	Pr/G2	.151	.386	.788	.032	.014	.002	.145	.360	.785	.065	.060	.007
	Sd	.400	.487	.851	.067	.017	.006	.528	.942	.562	.040	.003	.098
VMo	Gr/G1	.379	.491	.821	.007	.006	.001	.643	.821	.829	.021	.001	.023
	Gr/G2	.438	.636	.820	.028	.021	.004	.470	.749	.728	.044	.002	.039
	Bi/G1	.023	.353	.668	.061	.015	.007	.050	.552	.488	.077	.093	.010
	Bi/G2	.384	.553	.802	.007	.010	.001	.505	.844	.642	.008	.006	.023
	Sd	.554	.711	.861	.000	.007	.000	.623	.662	.964	.002	.004	.003
HMo	Gr/G1	.540	.674	.843	.010	.021	.001	.578	.693	.906	.005	.002	.001
	Gr/G2	.223	.531	.762	.031	.021	.008	.321	.345	.927	.035	.052	.003
	Bi/G1	.474	.695	.840	.060	.052	.011	.573	.875	.809	.035	.047	.009
	Bi/G2	.236	.530	.746	.014	.014	.003	.432	.555	.864	.032	.064	.009
	Sd	.212	.483	.714	.037	.001	.003	.230	.208	.949	.043	.052	.004

Table A.8: SVM results for the FEMALE test set trained with the Posed videos.

Appendix B

Timeline plots for tested videos

All timeline plots figures show on top the estimated states for each facial element, and below the manually annotated ground truth reference states. The color coding for the facial elements is as shown in Table B.1. The gloss of the videos (shown in gray color) marks the timing of sign articulation. Transitional movements such as raising or lowering hands at the beginning or end of a sentence are not considered in the gloss.

	Eyebrow	Eye	V Mouth	H Mouth
Red	down	closed		
Yellow		squint	wide	narrow
White	neutral	open	closed	relaxed
Green	raised	wide	open	wide

Table B.1: Color coding of quantized face states.

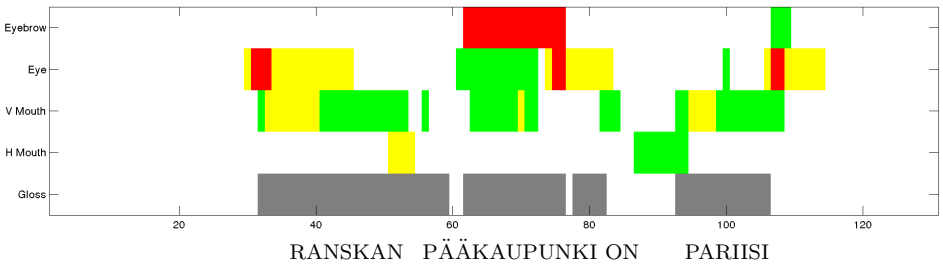


Figure B.1: Timeline of annotations for Suvi video 000601 (Training video).

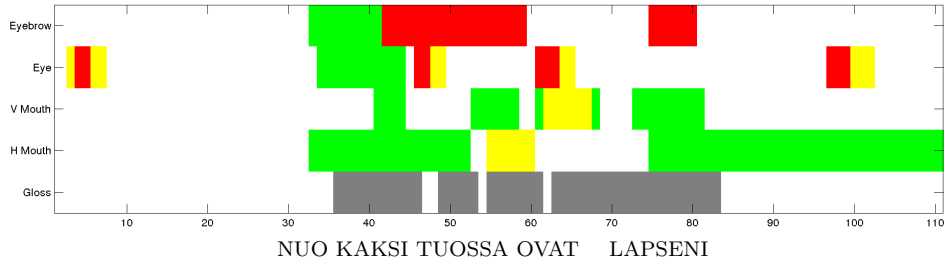


Figure B.2: Timeline of annotations for Suvi video 005802 (Training video).

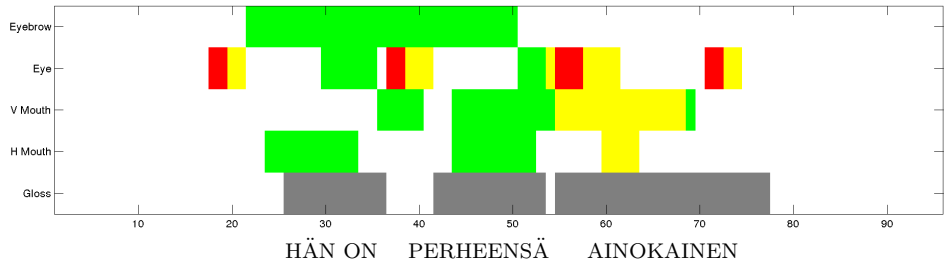


Figure B.3: Timeline of annotations for Suvi video 005804 (Training video).

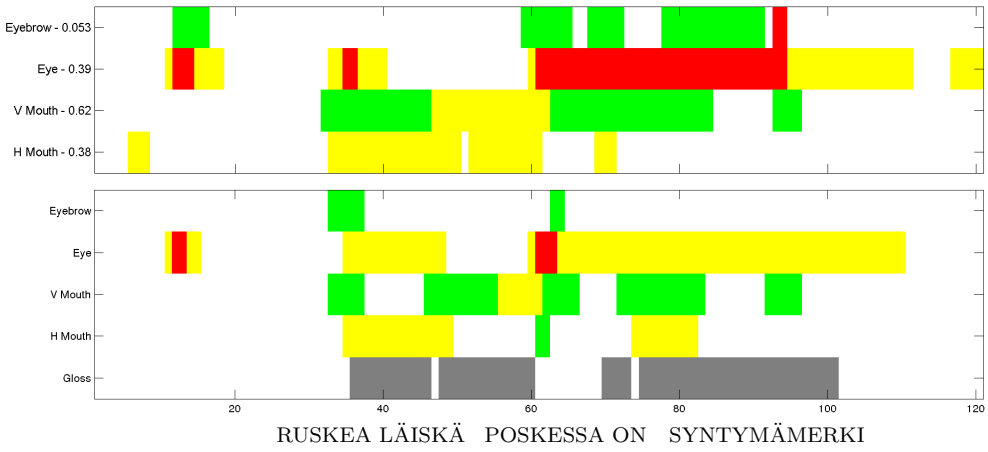


Figure B.4: Timeline of estimations for Suvi video 034601.

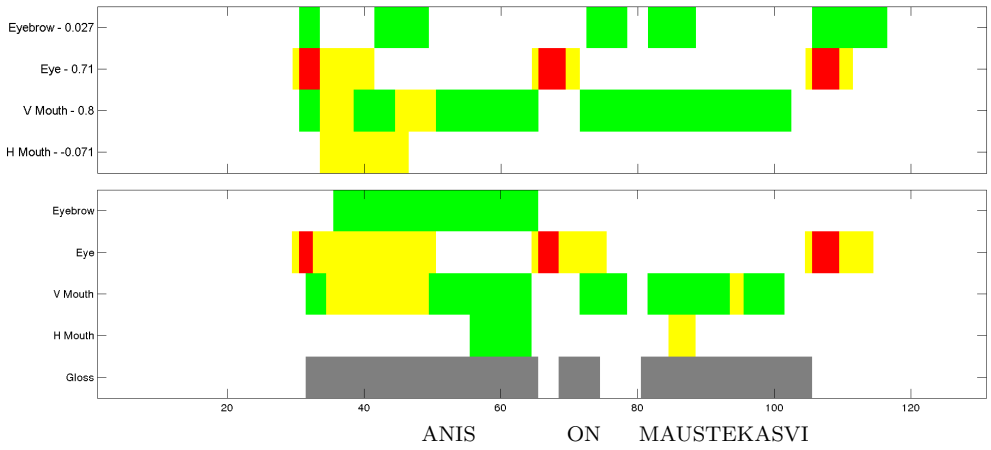


Figure B.5: Timeline of estimations for Suvi video 035001.

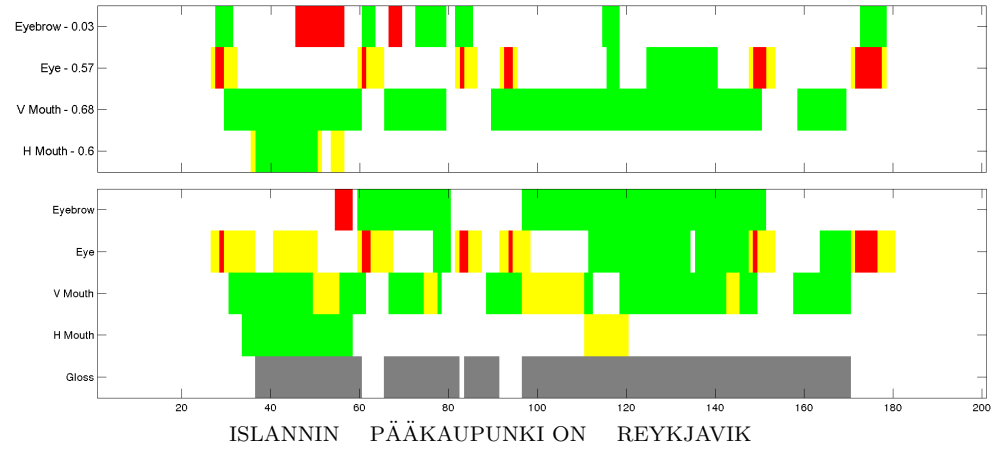


Figure B.6: Timeline of estimations for Suvi video 046603.

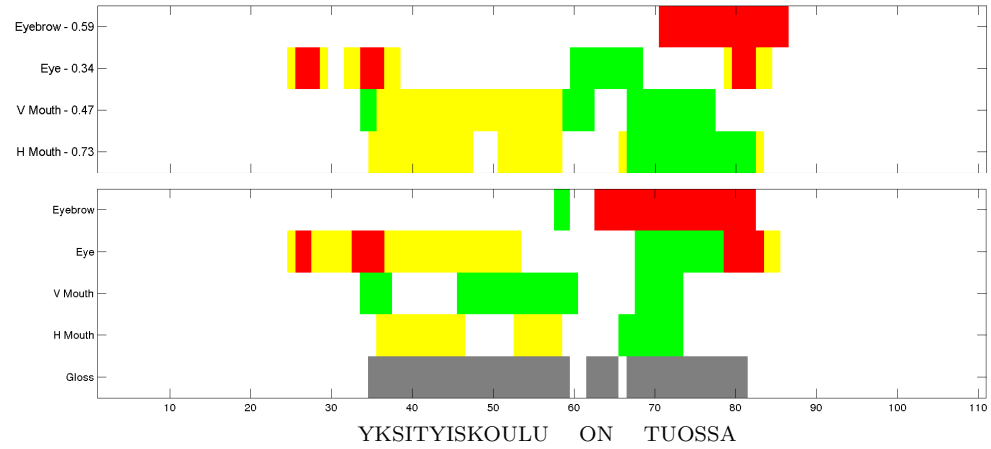


Figure B.7: Timeline of estimations for Suvi video 047301.

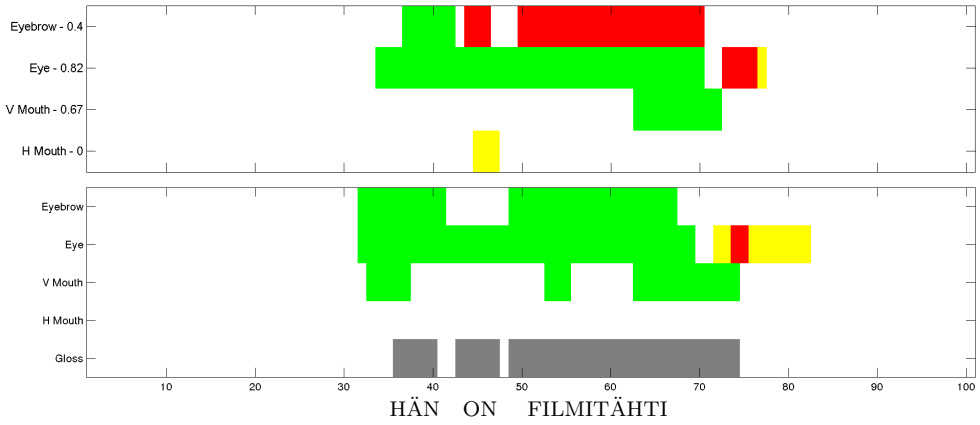


Figure B.8: Timeline of estimations for Suvi video 065503.

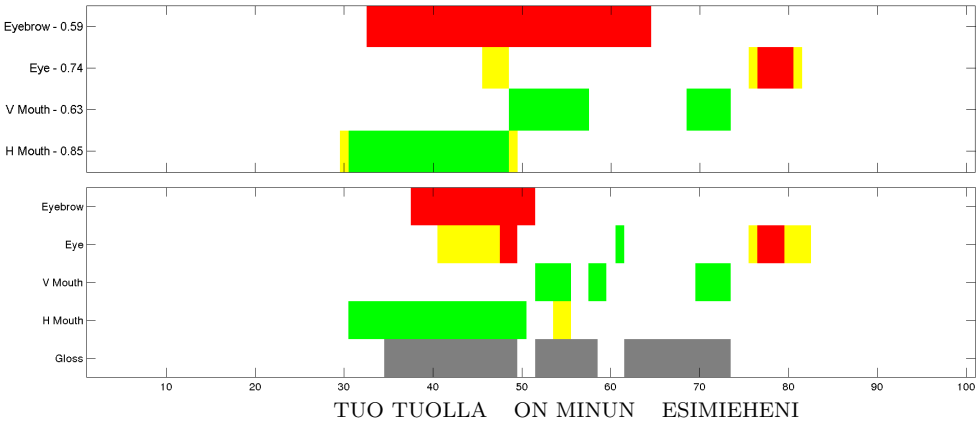


Figure B.9: Timeline of estimations for Suvi video 068701.

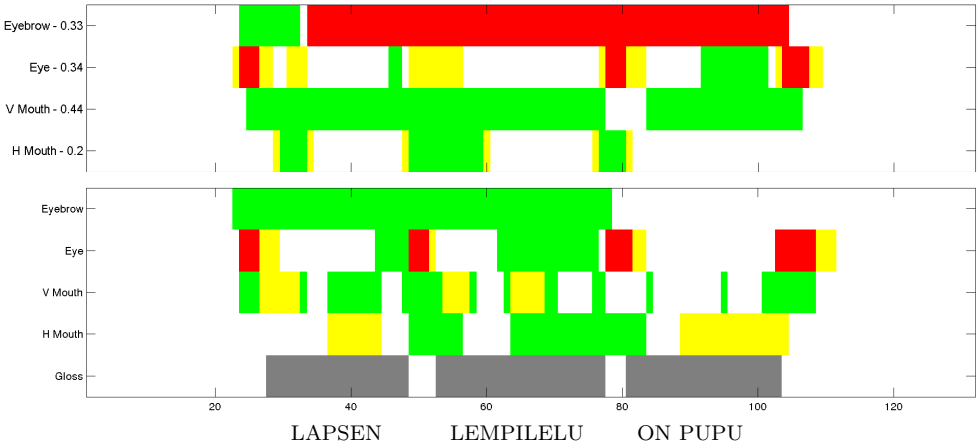


Figure B.10: Timeline of estimations for Suvi video 077704.

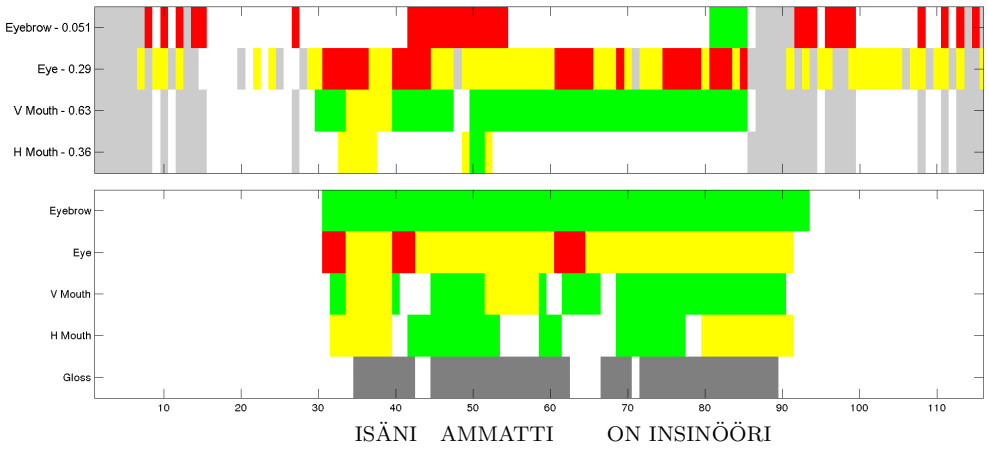


Figure B.11: Timeline of estimations for Suvi video 080704.

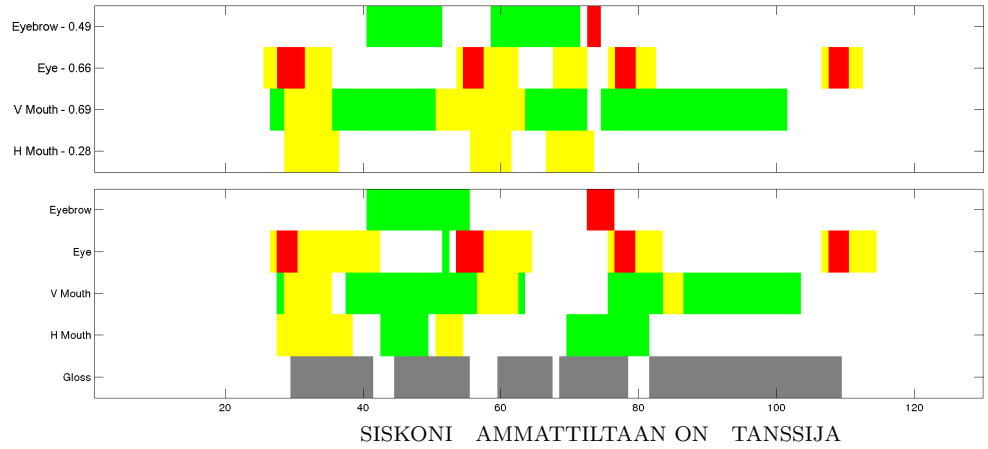


Figure B.12: Timeline of estimations for Suvi video 082304.

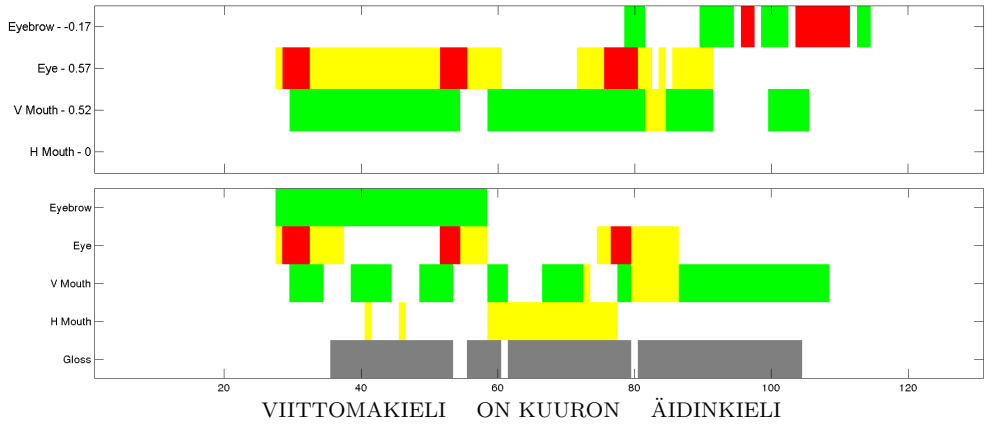


Figure B.13: Timeline of estimations for Suvi video 096603.

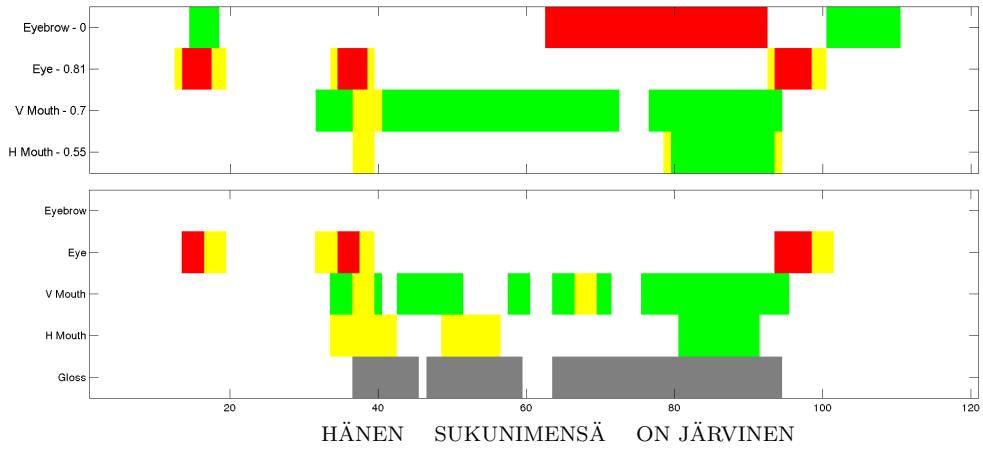


Figure B.14: Timeline of estimations for Suvi video 097203.

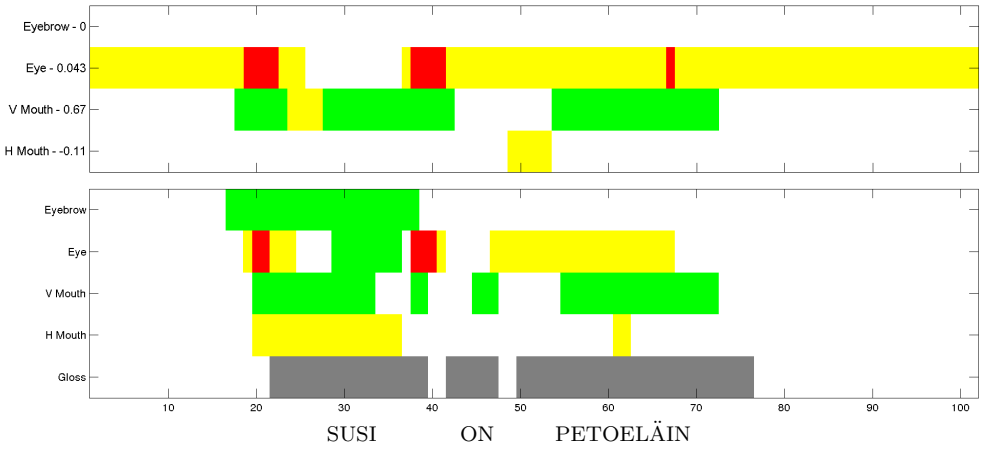


Figure B.15: Timeline of estimations for Suvi video 102902.

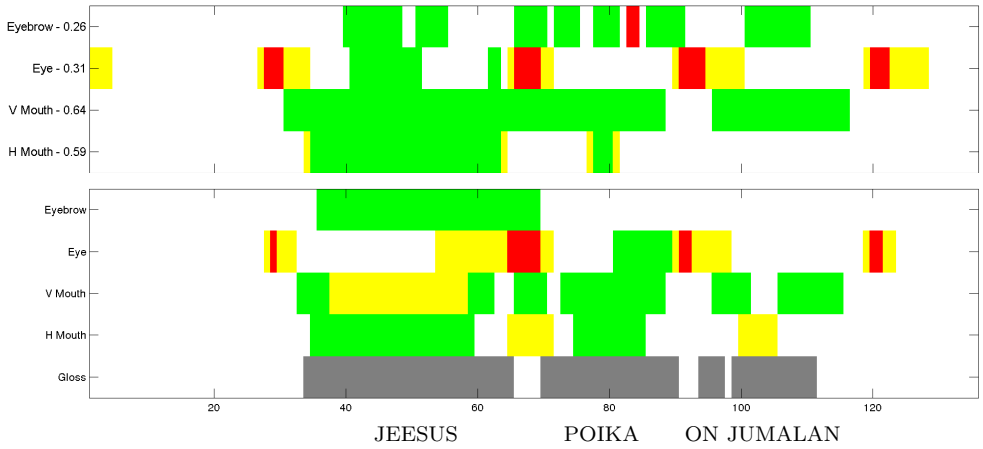


Figure B.16: Timeline of estimations for Suvi video 113801.

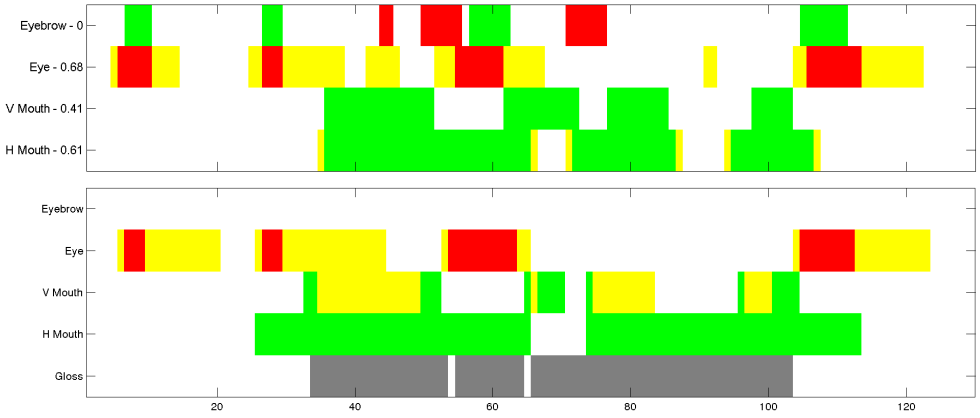


Figure B.17: Timeline of estimations for Suvi video 120701.

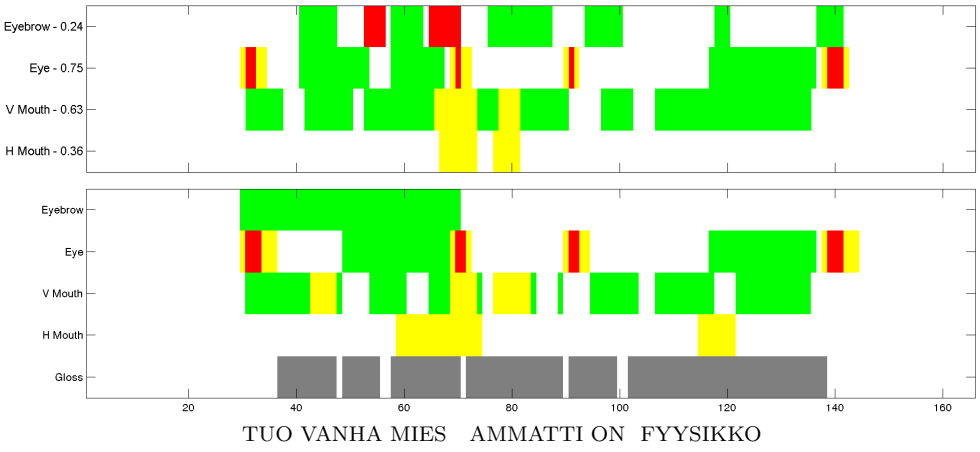
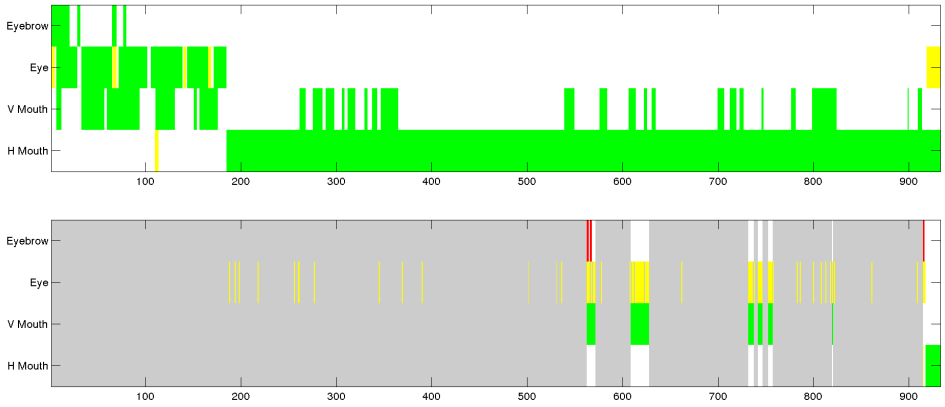


Figure B.18: Timeline of estimations for Suvi video 121601.

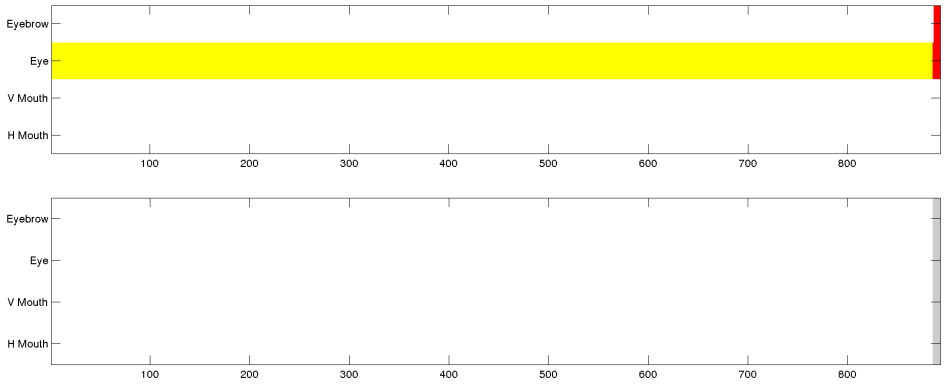
(a)



(b)

Figure B.19: YLE video 2. The color coding is the same as with previous experiments. (a) Example frame of NM estimation with super-imposed symbols representing states. Top: News anchor. Bottom: Two static faces introduced. (b) Timeline representation of the estimated states. Top: Timeline of the news anchor (frames 1–187) and first introduced face "Matti Vanhanen". Bottom: Timeline of second introduced face "Jyrki Katainen". The news anchor does not get its own timeline due to error in the simple face alignment algorithm. The second introduced face is mostly not detected resulting in large blank areas in the timeline plot.

(a)



(b)

Figure B.20: YLE video 3. The color coding is the same as with previous experiments. (a) Example frame of NM estimation with super-imposed symbols representing states. Top: Three static faces. Bottom: news anchor. (b) Timeline representation of the estimated states. Top: Timeline of static face "Eero Lankia" and the news anchor (frames 872–883). Bottom: Timeline of static face "Mikko Alkio". One static face was not detected at all. The lack of liveness of the faces is more clearly identified in this example. Estimation errors in the static faces are due to the poor facial states initialization.