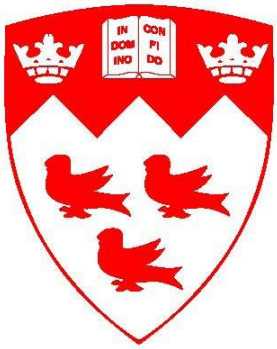
# Does “Lie to Me” Lie to You? An Evaluation of Facial Clues to High-stakes Deception

Lin Su

Degree of Master of Engineering

Department of Electrical and Computer Engineering

McGill University, Montreal




July 2013

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Engineering

# Abstract

During a forensic interview, high-stakes deception is very prevalent notwithstanding the heavy consequences that may result. Studies have shown that most untrained people cannot perform well in discerning liars and truth-tellers. Thus it has become common to adopt various technical aids to compensate for this poor judgment. Examples are polygraphs, functional Magnetic Resonance Imaging (fMRI) and linguistic analysis. However, the deception indicators used in these cases are not reliable.

In the popular TV program *Lie to Me*, micro-expressions have been used for detecting deceit during the investigation of some criminal cases. A micro-expression is considered to be a rapid and involuntary facial expression which could reveal the concealed emotion. Additionally, some psychological studies have stated that certain facial actions are more difficult to *inhibit* if the associated facial expressions are *genuine*. Similarly, these facial expressions are equally difficult to *fake*. This has cast light on the possibility that deception could be detected by analyzing these facial actions. However, to the best knowledge of the author, there is no computer vision research that has attempted to discriminate high-stakes deception from truth using facial expressions. Therefore, this thesis aims to test the validity of facial clues to deception detection in high-stakes situations using computer vision approaches.

We note that only a limited number of the existing databases have been collected specifically for deception detection studies and none of them were obtained from *real-world* situations. In this thesis we present a video database of actual high-stakes situations, which we have created using YouTube.

We have adopted 2D appearance-based methods as the methodology to characterize the 3D facial features. Instead of building a 3D head model as is the current trend, we have extracted invariant 2D features that are related to the 3D characteristic.

In order to discern deception and honesty, we have identified the following deceptive cues from nine separate facial regions through dynamic facial analysis: eye blink, eyebrow motion, wrinkle occurrence and mouth motion. Then these cues were integrated to form a facial behavior pattern vector. A Random Forest was trained using the collected database and applied to classify the facial patterns into deceptive and truthful categories.

Despite the many uncontrolled factors (illumination, head pose and facial occlusion) in the videos in our database, we have achieved an accuracy of 76.92% when discriminating liars from truth-tellers using both micro-expressions and "normal" facial expressions. The results have shown that using facial clues for automated lie detection is very promising from the point of view of practice. In addition, we also challenge the belief expounded in *Lie to Me* that micro-expressions alone are sufficient for detecting lies.

# Abrége

Une des plus grande faiblesse des entrevues d'enquête provient de la déception de l'accusé, malgré les conséquences majeures. Des études indiquent que la majorité de personnes non formés ne peuvent discerner entre les menteurs et les diseurs de vérité. D'où la nécessité d'utiliser des aides technologiques pour atténuer ce mauvais jugement. Par exemple, les investigateurs peuvent se servir de détecteurs de mensonges, de l'imagerie par résonnance magnétique fonctionnel, et de l'analyse linguistique. Malgré leur utilisation, ces techniques manquent souvent de fiabilité pour indiquer les déceptions.

Dans l'émission de télévision américaine *Lie to Me*, des mico-expressions sont utilisés pour détecter la déception pendant certains investigations criminels. Une micro-expression se définit comme une mimique rapide et non volontaire qui peut révéler une émotion cachée. De plus, des études psychologiques indiquent que certaines actions du visage sont plus difficiles à inhiber si les expressions correspondantes sont sincères. De la même façon, ces expressions sont également difficiles à feindre. Ces principes indiquent la possibilité de détecter la déception en analysant les actions du visage. Par contre, à la connaissance de cet auteur, il n'existe aucune recherche dans le domaine de l'imagerie informatique pour discerner entre une grave déception et la vérité en analysant les mimiques. En conséquence, l'objectif de cette thèse est de tester la validité des mimiques pour détecter la déception dans les situations conséquentes en utilisant des techniques d'analyse des images informatiques.

Il existe peu bases de données conçues spécifiquement pour des études avec l'objectif de détecter la déception, qui présentent tous des scenarios fabriquées. Cette thèse est unique puisqu'elle introduit une base de données de vidéos présentant des situations réelles et conséquentes provenant de YouTube.

Nous avons adopté des méthodes basées sur l'apparence 2D comme méthodologie pour caractériser les traits du visage en 3D. Au lieu de construire un modèle de

tête 3D comme il en est la tendance actuelle, nous avons extrait des fonctionnalités 2D invariantes qui sont liées à la caractéristique 3D.

Pour discerner entre la déception et vérité, nous avons identifier des indices de déception par une analyse dynamique de neufs régions sur le visage: les clins d'œil, les mouvements de sourcils, l'apparition des rides, et les mouvements de la bouche. Ces indices ont été intégrés pour former un vecteur représentant le modèle de comportement du visage. Une forêt fortuite *(Random Forest)* a été formée en utilisant la base de données construite pour classifier les comportements du visage en deux catégories, la déception et la vérité.

Malgré la nature chaotique des vidéos (illumination et orientation variable de la tête, et l'occlusion du visage), nous avons atteints une précision de 76.92% pour détecter les menteurs en analysant leurs micro-expressions et expressions « normales ». Nos résultats démontrent que les indices provenant du visage peuvent être appliquées pour détecter des mensonges en pratique. De plus, nous défions l'idée semée par l'émission *Lie to Me* que les micro-expressions seuls suffisent pour détecter les mensonges.

# Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Martin Levine, for his patience, guidance and encouragement throughout my research, and for his enthusiasm and spirit in regard to research from which I would benefit for life.

I would also like to acknowledge Professor Stephen Porter, without whom the database would not have been possible. Besides, many thanks to numerous professors and advisors who have provided me with persistent guidance and support during my years at McGill University.

Finally, I would like to thank all my friends and colleagues at the Centre for Intelligent Machines, particularly Mehrsan Javan Roshtkhari, for their continuous support and constructive suggestions on my research. Additionally, I really appreciate Matthew Balazsi for his help with my French abstract.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1. Introduction

## 1.1. Overview

Do we have the capacity of "tearing off" the mask of a liar's face and revealing the truth behind it? Obviously Neville Chamberlain did not. Otherwise he would not have trusted Hitler with his oath that he would not invade Czechoslovakia, which has turned out to be one of the most shocking lies ever [1].

Speaking of deception detection is reminiscent of a TV show called *Lie to Me*, which has enjoyed a large popularity in the past few years. In this TV series, Dr. Lightman and his team used their "talent" to assist the police with the investigation of some criminal cases. Their "talent" is that they could visually determine if a suspect was lying by interpreting his micro-expressions[1] during the interrogation.

The deception of concern by *Lie to Me* is termed *high-stakes deception*, because interrogation on a criminal case yields a high-stakes scenario. Different from the lies told in our daily lives, deceptions in high-stakes situations are more likely to result in heavy consequences. Therefore a liar in such a situation might experience heavy cognitive load, being aware of the severe penalty he will receive if his lie were caught. Considering the risk of releasing a guilty suspect or mistaking an innocent person, the detection of high-stakes deception is a necessity for a democratic society.

Do humans have the ability to detect high-stakes lies, just like Dr. Lightman and his colleagues did? As will be seen in this thesis, most untrained people are no better than chance at detecting lies [2], and the subjective decision-making process of humans may thereby bias their decisions [3]. Also, even though the viewers of *Lie to Me* attempted to learn how to detect lies from this TV show,

[1] A micro-expression is a rapid and involuntary facial expression which can seemingly reveal a person's genuine emotion. It will be discussed in Chapter 4.

there is evidence showing that they were more likely to misidentify innocent people as liars [4].

However, regardless of the fact that ordinary people are deficient at deceit detection, is the fundamental theory of detecting lies in *Lie to Me* plausible or not? In other words, is a micro-expression reliable as a clue to deception? This thesis aims to investigate on this question, validating if facial clues could be adopted as indicators of deception in high-stakes situations.

The proposed method consists of three stages: pre-processing, dynamic feature analysis and classification, as shown in Figure 1. In the pre-processing stage, face detection and facial landmark localization are firstly applied to register the face[2]. Then an anthropometric model is used to decompose the face into several facial regions. In the dynamic feature analysis stage, the indicators of deception, which are actually facial expressions, are detected in each facial region and collected into a facial behavior description vector. Finally in the classification stage, a binary Random Forest classifier is trained to discriminate deception and honesty.

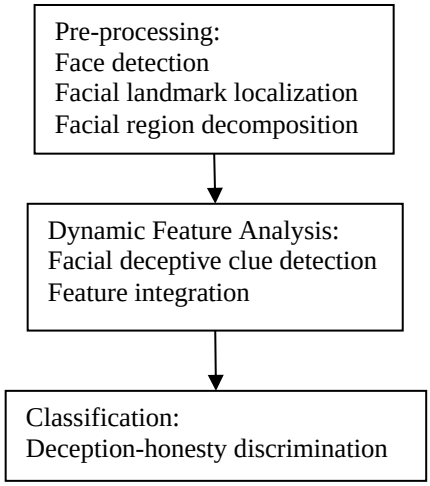



Figure 1. Three stages of the proposed method

[2] A commercial software (PittPatt [90]) was used to locate three major landmarks on the face: left eye, right eye and nose base.

## 1.2. Thesis Contributions

This thesis has several contributions, as discussed as follows.

1) This thesis has set a precedent for future research on high-stakes deception detection using facial clues. To the best knowledge of the authors, there is no computer vision research that has testified the validity of facial expressions as indicators of high-stakes deception. As will be seen in Chapter 6, our results are very promising, implying the research potential of this topic in the future.

2) A database consisting of high-stakes deception videos of real-world situations has been collected from YouTube. In our database, the suspects are either pleading for the safe return of their missing relatives or denying their involvement with the disappearance or death of the victims. All of the criminal cases are real, and approximately half of the suspects were convicted as guilty by overwhelming evidence. As will be seen in Chapter 3, only a limited amount of the existing databases has been collected for deception detection studies, and none of them were obtained from *real-world* situations.

3) The proposed method is at the forefront of analyzing facial expressions in unconstrained environments. Since the videos in our database were collected from YouTube, certain uncontrollable factors add to the difficulty of their analysis. These totally unconstrained and spontaneous videos are subject to temporal variations in illumination, head pose and facial occlusion. However, in the current literature, little research has been conducted to address these issues with regard to facial expression analysis. Instead of seeking a solution to solve them, almost all of the studies to date have excluded certain data that were not ideal for the purpose of analysis. In comparison, the proposed method seeks to address the deception detection problem, but in the presence of exactly these factors.

## 1.3. Chapter Contents

The rest of the thesis is organized as follows.

Chapter 2 is a literature review of the background of the presented research, including an introduction to high-stakes deception, people's performance at

discriminating liars and truth-tellers, current measures and automated methods for detecting lies. In Chapter 3, a review of the current databases for deceit detection studies is presented. Since the proposed method focuses on facial expression analysis, current facial expression databases are also reviewed. Then the database collected by the author is introduced. Chapter 4 presents the theoretical foundation of our deception detection method, and proposes the dynamic feature analysis methods based on these theories. After feature extraction, Chapter 5 explains the experimental procedure of training and testing a binary classifier for discriminating deception and honesty. The results are discussed in Chapter 6. Finally the conclusion and future work are presented in Chapter 7.

# Chapter 2. Literature Review

In this chapter, we present a literature review of deception detection in four sections. In the first section, we will introduce the definition of high-stakes deception and emphasize the necessity that they be detected. Secondly, we will review the performance of humans at discriminating liars and truth-tellers both before and after training. In the third section, several current measures for detecting lies will be presented, and the pros and cons of each measure will be discussed. Finally, the limited amount of research in the computer vision area related to deception detection is reviewed.

## 2.1. High-stakes Deception

People lie twice a day on average [5, 6]. In fact, deception is facilitated by technology, resulting in ubiquitous lies in every kind of interpersonal communications: face-to-face, phone calls and emails [7]. People lie so frequently, because their self-esteem encourages them to hide information that could be harmful to their pride or relationships. Nonetheless, not all lies are bad. An appropriate lie or white lie could sometimes save a relationship crisis, win a business negotiation, or encourage a dying person to live better.

However, despite that deception is ubiquitous in our daily lives, there are certain circumstances where lies are more likely to give rise to heavy consequences to both individual and society [8]. These circumstances are termed *high-stakes situations*, for example, being interrogated by a police officer, defending oneself in the courtroom, or appealing for a parole to a judge.

To detect deceit in such high-stakes scenarios is even more demanding than in daily life, because either failing to catch a liar or wrongly discriminating an innocent person as a liar could lead to disaster. Unfortunately in the latter scenario, even if the truth were uncovered years later, the damage that would be brought to the misjudged innocent person and his family might be invertible and may last forever. As said by Rebecca Sophonow, whose husband was falsely

convicted of murdering a girl in 1981 but later released and compensated $2.6 billion dollars [9], “Compensation doesn’t make you move on with your life and forget” [10]. On the other hand, it is even more dangerous to release a guilty murderer into the society, allowing him to have the opportunity to commit crimes again. Thus, miscarriage of justice is a terrible consequence of failing to detect deceit in high-stakes situations, which not only brings lifelong agony to individuals, but also propagates potential dangers into the society. In light of this, successfully spotting a lying suspect or clarifying an innocent person in a high-stakes situation is of high necessity and importance.

## 2.2. People’s Performance at Lie Detection

People rarely have the capacity of being able to see the truth behind lies, regardless of the fact that lies are occurring on a regular basis.

This fatal problem with human judgment often involves their inability to be objective. Unlike machines, humans are inevitably biased by a myriad of factors, resulting in their “tunnel vision” when they assess the credibility of suspicious people. The first factor is their dreadful self-confidence. People often underestimate the effort they should take to spot liars, but have too much confidence on their judgment [11]. Secondly, people’s judgment largely relies on their naïve intuition. They often believe in emotion-based rather than empirically based clues to deception, such as gaze aversion and fidgeting, which is detrimental to their judgment [12]. The third factor involves their emotions. Interestingly enough, studies have shown that emotionally intelligent people perform worse at deception detection. This is due to their greater sympathetic feelings to others [13]. Fourth, motivation also has a negative influence on deception detection accuracy, which is termed motivational impairment [14, 15]. Finally, people’s subjectivism often leads them to be more skeptical towards exonerating witnesses than incriminating ones [14], consequently degrading their performance at lie detection.

Considering the fact that the judgment of people is easily biased, it is not surprising that their lie detection accuracy is slightly above chance. In [16], Bond et al. synthesized research results from more than 200 documents and found that people could discriminate lie and truth with an accuracy of 54%, with a lie detection rate of 47% and truth detection rate of 61%. Similarly in [17], 192 undergraduate students watched videotaped interviews of mock suspects denying stealing a wallet and obtained an overall accuracy of 55.2%. The detection rate of guilty suspects was 61.5%, which is better than chance; for innocent suspects the percentage is only 49%. Ekman and O'Sullivan have also stated that most ordinary people could only achieve around 50% precision in detecting deceit, while professionals from the Secret Service outperformed them with an accuracy of 70% [18]. Vrij et al. have reported an average accuracy of 72% by thirty-seven police officers after watching videos of high-stakes police interviews, with a lie detection accuracy of 73% and a truth detection accuracy of 70% [19]. Warren et al. [20] have also reported that college students achieved an accuracy of only 50%. Ekman has claimed that the existence of so-called "wizards", who have extraordinary deception detection skills, could achieve higher accuracy than ordinary people [18]. But this discovery has been challenged by Bond [16], giving rise to a long debate between them [21]. To date, there is no definitive conclusion whether individual differences in the capacity of detecting lies actually exist, i.e., whether "wizards" exist. Table 1 is a summary of the reported performance on deceit detection by human observers.

However, no matter whether "wizards" exist or not, psychological studies have shown that people's performance at lie detection could be improved by professional training. 32 Canadian federal parole officers performed below chance, but achieved 76.7% accuracy after participating in an empirically validated training program for two days [22]. In another study, the accuracy of detecting deceit in videotaped narratives increased from 46% to 58% after a two-day abbreviate training program by 26 healthcare professionals [23]. The best precision 80.9% has been achieved in [18], in which 42 legal and mental

professionals were trained in a comprehensive and empirically based training workshop.

Table 1. Performance on deceit detection by human observers. The best performance reported to date is indicated in bold. Note that the number of test samples (where reported) is relatively low.

| ***Reference*** | ***Lie Catchers*** | ***Number of Lie Catchers*** | ***Overall Accuracy*** | ***Liar Spotting Accuracy*** | ***Truth-teller Spotting Accuracy*** |
|---|---|---|---|---|---|
| **[16]** | Diverse (206 documents synthesized) | N/A | 54% | 47% | 61% |
| **[17]** | Undergraduate students | 192 | 55.2% | 61.5% | 49% |
| **[18]** | Police officers, CIA and FBI agents, lawyers, college students, therapists, judges, etc. | N/A | 50% | N/A | N/A |
| **[18]** | Secret Service agents | N/A | 70% | N/A | N/A |
| **[19]** | Police officer | 37 | **72%** | 73% | 70% |
| **[20]** | College students | 20 | 50% | N/A | N/A |

In conclusion, appropriate training programs have effectively improved the performance of humans at lie detection. This has been attributed mainly to the fact that during the learning process, the measure of lying has gradually changed from naïve stereotype to appropriate empirical evidence. From this we observe that, a proper measure of deception is of the highest significance in deceit detection. Therefore in the next section, some of the most important deception detection measures employed in the recent decades will be discussed in detail.

## 2.3. Current Deception Detection Measures

How do people detect lies? Since no one has an extensible nose as Pinocchio does, certain evidence-based measures are obviously required to discriminate liars and truth-tellers. Currently, a variety of measures have been applied to real world applications, including physiological, neuroscience, and psychological linguistic measures.

### 2.3.1. Physiological Measures

A physiological measure is a very straightforward way of distinguishing between guilty and innocent suspects in criminal cases, and has a long history. Centuries ago, people in Arabia used to put hot iron bars on the tongue of a suspect's. If he were burnt, he would be considered as guilty. Similarly in China and England, rice and bread were used, respectively. All of these ancient methods were based on the same obviously naïve theory: guilty suspects secrete less saliva due to their nervousness and anxiety, and will consequently get burnt by the iron bar or choke on the bread.

Later, the physiological measurements have been extended to be more scientific and potentially reliable. For instance, thermal body imaging [24] detects blood flow via special cameras, while voice stress analysis [25] measures the voice signature. A more famous and dominant physiological measurement in the past century is the polygraph [26, 27], which measures multiple physiological signals simultaneously. It is usually combined with a psychological questioning strategy, termed the "Comparison Question Test" [27], resulting in what may be termed as a "psychophysiological technique".

However, the drawback of these physiological approaches is that they only focus on measures emanating from the peripheral rather than central nervous system [27]. In this case, the interpretation of the measurements is not directly related to the emotional states of humans, which are controlled by the central nervous system. The other disadvantage is that they are easily degraded by certain

countermeasures, such as tranquillizers or repeated practice, to control physiological arousal.

Consequently, physiological measures lack reliability for detecting lies, not only because they are based on an untenable hypothesis that physiological symptoms can provide solid indications of deception, but the measurement results might be inaccurate due to countermeasures.

### 2.3.2. Neuroscience Measures

As an alternative method of lie detection, functional Magnetic Resonance Imaging (fMRI) is an objective measurement of mental state [28]. It is typically conducted using an MRI scanner over the head or body. Actually, it has only recently made its debut in the forensic sphere in 2010. When Brian Dugan, serving a pair of life sentences for his murders in 1980s, was charged of another murder, his lawyer proved him to be a psychopath and tried to help him avoid more punishment [28]. On the theoretical side, fMRI is superior to the polygraph since it measures brain activity in the central rather than peripheral nervous system. Studies have shown that deception is associated with the activity in prefrontal brain regions [29], and brain activity seems to be uncontrollable by humans. In this case, it seems that the fMRI approach should be reliable in lie detection tasks.

However, fMRI also has several shortcomings [28, 30-33]. First, most of the fMRI studies have been small and rarely replicated, thereby lacking validity through scientific scrutiny. Second, brain activity varies considerably for each individual, making it unreliable to simply use the average brain activity of a group of healthy and normal people to evaluate the truthfulness of individual high-stakes cases. Third, the theoretical findings of many studies are not consistent with each other, due to the difference in experimental settings and paradigms. Fourth, the accuracy of deception detection of fMRI is vulnerable to covert countermeasures. In addition, both polygraph and fMRI require expensive equipment as well as the cooperation of the suspects.

### 2.3.3. Psychological Linguistic Measures

Psychological linguistic, or verbal measures, is another deceit detection approach which has been widely studied by psychologists in the past few decades.

Response latency has been used as the main verbal indicator of deception in many psychological studies [34]. It is based on the argument that lying seems to require more time for cognitive processing than truth telling. However, frequent lying will make it easier to lie, whereas frequent truth telling will make it more difficult [34]. This finding casts new light on the interrogation strategy. Asking more questions about irrelevant events that are assumed to yield more truthful responses by suspects may actually increase his lying latencies. If the lies are spontaneous, the response should take a longer time than average. If the lies are prepared, the response may take a shorter time than expected [35-37]. Besides asking anticipated and unanticipated question pairs, other questioning strategies have also been proposed and have proved to be effective. Examples are: imposing a cognitive load to elicit more cues to deceit [38-40], adopting drawing or describing spatial contexts as a complementary way of interrogating events in chronological order [41-43], and adding a second supportive interviewer who keeps nodding his head and smiling [37]. In spite of the commonly accepted relationship between response latency and lying, the authors of [44] proposed that the context of a conversation should be taken into consideration rather than blindly defining their correspondence.

In addition to response latency, many other verbal characteristics have also been recorded and examined in many psychological experiments. Speech rate is considered to be one of the effective verbal clues to deception, when it is either faster or slower than the normal rate [35]. Moreover, Porter has argued that some verbal cues, such as word frequency, grammar usage, tentative word frequency, and qualitative details have been widely used and have achieved high accuracy [3]. Also, the statement of guilty suspects will be less consistent with the evidence in comparison with that of innocent suspects [17].

Although verbal cues have been very popular among psychologists, great caution is required when applying them to high-stakes situations because verbal countermeasures are easier to adopt than non-verbal ones [45]. Also, an effective interrogation strategy is a prerequisite for eliciting useful verbal clues.

## 2.4. Current Computer Vision Research on Deception Detection

Very little research has been done on deception detection using automated computer vision approaches. Rather, the focus has been on three types of indicators: body language, gaze aversion, and facial expression.

### 2.4.1. Body Language as a Cue to Deception

There are some research trying to relate deception with agitated and over-controlled behavioral states. The authors of [46-48] were aimed to prove the theory that behavioral states are related to deception discrimination. They have analyzed the position and velocity of face and hand blobs, classifying the suspects into over-controlled, agitated and relaxed states. However, their work was not very persuading, since their experiments were merely based on a very small set of data (18 subjects). The data they used came from the Mock Theft Experiment [49, 50], which was initially collected to analyze the linguistic features [50]. The Mock Theft Experiment simulated a high-stakes situation but the stakes were relatively low.

However, body language as a cue to deception has weak theoretical support. Darwin proposed a theory termed *face>body hypothesis*, which demonstrated that body movements should be easier to conceal than facial expressions. But Ekman has also argued that most people will pay more attention to managing their facial expressions when they are lying and thus will have less control over their body language [51]. Interestingly, although Porter has attempted to relate the departure from the baseline[3] of body movements to deception [3, 52], he has found that most body cues occurred too rarely for statistical analysis [52].

[3] Baseline is considered as the normal behaviors of an individual.

### 2.4.2. Gaze Aversion as a Cue to Deception

Some computer vision studies have used gaze aversion as a clue to deception.

In [53, 54], the researchers have investigated the issue of detecting deceit under simulated high-stakes situations. They have trained a dynamic Bayesian model of eye movements during the baseline session, where the suspect behaved normally before the actual interrogation. Then, if the eye behaviors (gaze direction and blink rate) during the interrogation deviated from the baseline, they would be categorized as deceptive. They have achieved an accuracy of 82.5%. However, this method requires a baseline for every suspect, which is hard to achieve in reality if the suspect intentionally behaved abnormally before the interrogation starts.

Nevertheless, according to DePaulo et al. [55], gaze aversion and fidgeting might have nothing to do with deception. Mann et al. also found that eye contact maintenance has no significant relationship with deception [56, 57]. To date, no authoritative study has made it clear whether gaze aversion is indeed a reliable indicator of deception.

### 2.4.3. Facial Expression as a Cue to Deception

In [58], the authors discriminated genuine with deceptive facial expressions. In their work, genuine expressions are natural or spontaneous, while deceptive expressions are posed or acted. Their approach was based on the theory that genuine expression differs from deceptive one according to the present or absence of one or more Action Units (AUs)[4]. The veracity decision was based merely on simple threshold. Later authors of [59] extended the research in [58] by adopting machine learning methods in the decision-making process. They used CUBRC-CUBS dataset [58] in their experiments. This dataset was a collection of natural (genuine) and posed (deceptive) facial expressions, but has not been made publicly available. Therefore, it is unknown that whether the images in the dataset

[4] Details regarding AUs will be presented in Chapter 4.

were obtained in lab-environment or uncontrolled wild, and how they manually categorized the data into verity and deceit groups. Moreover, their methods were not completely automatic, since the facial points were manually labeled on every image.

### 2.4.4. Multiple Cues to Deception

Some researchers have combined body language and facial micro-expressions to generate more convincing cues to deception [60]. For body language, head and hand movements were measured. For facial analysis, an Active Shape Model (ASM) was used to track the mouth and eyebrow movements. Then, each motion feature was represented by a 5-bin histogram, and fed into the Nearest Neighbor classifier. They have achieved a high accuracy of 81.6%. However, their experimental data were collected in a very low-stakes situation, where each participant told a deceptive opinion and a truthful opinion.

The Silent Talker presented in [61] has used Artificial Neutral Network (ANN) to discriminate deception and honesty based on four cues to deception: eye gaze, eye closure, head movement and blushing/blanching. A classifier was learnt from labeled training data to categorize each cue into one of the defined discrete states. Then ANN was used to integrate different cues to predict the emotional state: deceptive or truthful. They have achieved a classification accuracy of 79% based on their database, which was collected in a low-stakes scenario similar to the Mock Theft Experiment mentioned above.

## 2.5. Conclusion

In summary, high-stakes deception yields high necessity to be detected, but humans have a deficiency in accomplishing this task and the aforementioned psychological, neuroscience, and linguistic measures are all unreliable. Furthermore, the most prominent issue shared by past psychological and computer vision studies is that researchers rarely employed data obtained in *real* forensic circumstances. To date, only one study has actually employed fMRI scans of the brain of a woman who was convicted of poisoning a child [62]. Other experiments

have all been based on virtual forensic environments, which obviously did not provide real high-stakes situations. In addition, these data were not spontaneous, since most of them were created particularly for experimental studies, thereby creating doubt about their authenticity. We propose an automatic deception detection system capable of providing valid predictions of lying in uncontrollable situations, and base the decision classifier on experimental data captured in high-stakes situations. In the following chapters, the database and the automated methods will be presented.

# Chapter 3. Database for Training Decision Classifiers

A database is a collection of data which provides a platform to test the validity of a theory as well as the robustness of the algorithms for validating it. An ideal database should consist of data collected from the environment where the developed application would be applied. As stated in Chapter 2, we aim to verify the soundness of the theoretical argument that facial clues could be adopted as reliable evidence for detecting high-stakes deception. Therefore, an interrogation interview is our target application context, where a suspect in a forensic case is being questioned by an investigator. The video of the suspect's face captured during the interrogation is our target data. Ideally, the environmental setting for capturing this video should satisfy the following requirements:

*Illumination*: The lighting in the interrogation room should be of constant and moderate brightness, and uniformly illuminated on the suspect's face.

*Camera setup*: The camera should be set at a fixed distance from the suspect, capturing a frontal face without any change in the shooting angle or distance. Furthermore, the suspect's face should be the only face that appears in the camera view.

*Background*: The background of the suspect should be as constant as possible. Normally a wall or a curtain of a solid color is the best choice.

*Suspect*: The suspect is asked to face the camera, without moving dramatically one's head or body so that the camera will neither lose track of the face nor violate the frontal-face assumption. Also, the suspect should be asked to remove any accessories that would occlude the face, including glasses, hat, facial hair and heavy make-up.

To sum up, our ideal database should be a collection of natural facial expressions and micro-expressions on frontal faces without any facial occlusion in a well-illuminated and constant background high-stakes interrogation.

Clearly, in general many of these requirements cannot be attained in a normal situation. Therefore, in this thesis we require that the deception detector be capable of working in much more complex circumstances. In fact, as will be seen, we both train and test our algorithm using video clips obtained from the Internet, mostly using YouTube. This provides us with very natural and complicated setting, perhaps more intricate that a normal environment encountered in police investigations.

Perhaps the most important factor is that the exhibited facial expressions or micro-expressions should be *naturally* elicited by the suspect's internal emotions, rather than artificially acted upon by instructions. This is the role of the interrogator.

In this chapter, we first review the current databases related to deception as well as current databases used for general facial expression analysis. In section 3.3 and 3.4, we will introduce the database that we have collected and explain how we have edited the dataset for later automated analysis.

## 3.1. Review of Current Deceptive Facial Expression Databases

To the best of our knowledge, there is no publicly available database that specifically includes facial expressions of people telling lies in high-stakes situations. In fact, there are only a few datasets that have been used in current deception detection research, but none of these have been made available to the public. Table 2 is a summary of the sample size and emotions collected by these databases, and Table 3 summarizes the confounding factors of them.

Table 2. Sample size and emotions in databases used in deception detection studies

| ***Databases*** | ***Number of Subjects*** | ***Emotions*** | ***Natural/Posed Expressions*** | ***Media Type*** | ***Year Published*** |
|---|---|---|---|---|---|
| Mock Theft **[49, 50]** | 41 | Unknown | Natural | Videos | 2003 |
| RU-FACS **[63]** | 100 | Unconstrained emotions and speech-related mouth movements | Natural | Videos | 2004 |

| | | | | | |
|---|---|---|---|---|---|
| Anonymous **[61]** | 39 | Unconstrained emotions and speech-related mouth movements | Natural | Videos | 2006 |
| CUBRC-CUBS **[58]** | Unknown | Anger, happiness, sadness, fear | Natural/Posed | Images | 2007 |
| Anonymous **[64]** | 41 | Disgust, happiness, sadness, fear, neutral | Natural/Posed | Videos | 2008 |
| Anonymous **[65]** | 27 offenders, 38 students | Unknown | Natural/Posed | Videos | 2008 |
| YorkDDT **[20]** | 20 | Unknown | Natural/Posed | Videos | 2009 |
| Anonymous **[60]** | 220 | Mouth and eyebrow movement | Natural | Videos | 2010 |
| Anonymous **[53]** | 132 | Unknown | Natural | Videos | 2011 |
| Anonymous **[66]** | 100 | Anger, happiness, sadness, disgust, neutral | Natural/Posed | Videos | 2011 |
| Anonymous **[67]** | 60 | Unknown | Natural/Posed | Videos | 2011 |
| Anonymous **[52]** | 78 | Unknown | Natural/Posed | Videos | 2011 |
| Anonymous **[2]** | 59 | Sadness, disgust, fear, happiness, neutral | Natural/Posed | Videos | 2012 |
| Anonymous **[68]** | 52 | Unknown | Natural/Posed | Videos | 2012 |

Table 3. Confounding factors of databases in deception detection studies

| ***Databases*** | ***Illumination*** | ***Background*** | ***Accessories**** | ***Head Pose*** |
|---|---|---|---|---|
| Mock Theft **[49, 50]** | Constant | Constant | Unknown | Unknown |
| RU-FACS **[63]** | Constant | Constant | Unconstrained | Unconstrained |
| Anonymous **[61]** | Unconstrained | Unknown | Unconstrained | Unconstrained |
| CUBRC-CUBS **[58]** | Unconstrained | Unconstrained | Unknown | Unconstrained |
| Anonymous **[64]** | Constant | Unknown | Unknown | Near-frontal pose |
| Anonymous **[65]** | Constant | Unknown | Unknown | Near-frontal pose |
| YorkDDT **[20]** | Constant | Unknown | Unknown | Frontal pose |
| Anonymous **[60]** | Constant | Constant | Unknown | Near-frontal pose |
| Anonymous **[53]** | Unconstrained | Unknown | Glasses | Unconstrained |
| Anonymous **[66]** | Constant | Unknown | Unknown | Near-frontal pose |
| Anonymous **[67]** | Constant | Unknown | Unknown | Near-frontal pose |
| Anonymous **[52]** | Unconstrained | Unconstrained | Unconstrained | Unconstrained |
| Anonymous **[2]** | Constant | Unknown | Unknown | Near-frontal pose |
| Anonymous **[68]** | Unconstrained | Unconstrained | Unconstrained | Unconstrained |

* Accessories include such items as glasses, hats, makeup, facial hair, etc.

In the tables above, RU-FACS [63] is a database specifically collected for deceptive behavior analysis. Participants were told to either lie or tell the truth about their opinion regarding a political or social issue, during an interrogation by retired FBI agents, police, or a county sheriff. The liars were told that they would receive cash rewards if they succeeded in fooling the interrogators. Otherwise

they would have to fill out a boring questionnaire as punishment. The experiment was designed in this way to *mimic* a high-stakes situation for the participants to lie under pressure. Despite this, the stakes in this experiment were relatively low compared to an actual forensic situation. This database was the closest we could find that met our criteria for an ideal database. The emotions elicited in this experiment covered a wide range [69]. The mouth movements caused by utterances were also captured, which is common in an actual interrogation but is ordinarily absent from most existing facial expression datasets. Unfortunately, the RU-FACS database has not been made publicly available, and has been used in only a few research papers [63, 70].

In [60], a similar laboratory experiment was conducted in which 220 participants were questioned by trained interviewers and told to answer questions truthfully or deceptively. This paper did not mention if any reward or punishment was involved in order to raise the stakes for lying.

The so-called Mock Theft Experiment [49, 50] was originally designed to analyze the linguistic features of the audios recorded in a simulated high-stakes scenario. In this experiment, students were told to "steal" a wallet and lie about it during the interrogation in order to obtain money rewards. Similarly in [67], participants were told to steal movie tickets. Compared to [63], these experiments have raised the stakes since it simulated a forensic scenario. But the stakes were still relatively low, because failing to sell one's lie in this case was not at all life-threatening. The databases employed in [53, 61] are similar to the Mock Theft Experiment described in [49, 50, 67].

In [65], an experiment was conducted to analyze verbal and non-verbal deceptive patterns. In this study, both offenders and students were videotaped when telling real and fabricated autobiographical stories. This database has a small sample size and was collected in low-stakes situations.

CUBRC-CUBS [58] is a database consisting of genuine and deceptive facial expressions *in static images*. The premise behind it is if a facial expression

occurred spontaneously or naturally, it was considered to be genuine; otherwise it was taken as a deceptive expression. However, this database is not documented, and is not publicly available. Thus, it is unclear how the images were collected and genuine and deceptive expressions were discerned.

In [2, 64, 66], the authors also collected genuine and deceptive expressions. Different from [58], the expressions were captured *in dynamic videos*. Furthermore, the deceptive expressions are categorized into three types: simulated (express an emotion when feeling no emotion), masked (express an opposite emotion to cover the felt emotion), and neutralized (express no emotion when an emotion is felt) expression. However, the stakes in this experiment were very low.

The YorkDDT dataset [20] recorded two kinds of lies: emotional and unemotional. An emotional lie was obtained when a participant described an unemotional scene while watching an emotional video clip; an unemotional lie was the opposite. This dataset is very small (20 participants) and the lies were also not high-stakes.

The database collected and analyzed in [52, 68] consists of suspects in real forensic cases. The emotions presented by the suspects were various and were captured in high-stakes situations. Partial of this database will be employed in this thesis, and detailed information regarding this will be presented in section 3.3.

In summary, current databases in deception detection studies are very rare, not publicly available, and few of them was obtained in a high-stakes situation. In light of the fact that our method for detection of deception is based on determining expressions, we will also briefly review in the next section the current publicly available datasets specifically created for facial expression analysis.

## 3.2. Review of Current Facial Expression Databases

To date, many databases have been created specifically for facial expression studies and widely used by researchers. A summary and comparison of some

commonly used facial expression databases could be found in Table 4, and Table 5 summarizes the confounding factors of them.

Table 4. Emotions in current facial expression databases

| *Databases* | *Emotions* | *Natural/Posed Expressions* | *Media Type* | *Year Published* |
|---|---|---|---|---|
| KDEF **[71]** | Basic six* + neutral | Posed | Images | 1998 |
| JAFFE **[72]** | Basic six + neutral | Posed | Images | 1998 |
| Multi-PIE **[73]** | Neutral, smiling, squinting, surprised, disgusted and screaming | Posed | Images | 2002 |
| DMFP **[74]** | Basic six + puzzlement, laughter, boredom, disbelief | Posed | Images & Videos | 2005 |
| MMI **[75, 76]** | Basic six + neutral | Posed and natural disgust, surprise, happiness | Images & Videos | 2005 |
| FEED **[77]** | Basic six + neutral | Natural | Videos | 2006 |
| PICS-pain **[78]** | Basic six + neutral + painful | Posed | Images | 2008 |
| CAS-PEAL **[79]** | Neutral, smiling, frowning, surprised, eye closed, mouth open | Posed | Images | 2008 |
| CK+ **[80]** | Basic six + contempt | Posed + Natural smile | Videos | 2010 |
| SEMAINE **[81]** | Anger, disgust, amusement, happiness, sadness, contempt | Natural | Videos | 2010 |
| RaFD **[82]** | Basic six + neutral + contempt | Posed | Images | 2010 |
| SFEW **[83]** | Basic six + neutral | Posed | Images | 2011 |
| AFEW **[84]** | Basic six + neutral | Posed | Videos | 2011 |
| SMIC **[85]** | Micro-expressions: happy, sad, surprise, angry, disgust | Natural | Videos | 2011 |
| USF-HD **[86]** | Smile, surprise, anger, sadness, and micro-expressions | Posed | Videos | 2011 |
| CASME **[87]** | Micro-expressions: Amusement, sadness, disgust, surprise, contempt, fear, repression, tense | Natural | Videos | 2013 |

* Basic six: the basic six facial expressions [88]: anger, happiness, sadness, fear, surprise, and disgust.

Table 5. Confounding factors of current facial expression databases

| *Databases* | *Illumination* | *Background* | *Accessories** | *Head Pose* |
|---|---|---|---|---|
| KDEF **[71]** | Constant | Constant | None | Five yaw angles |
| JAFFE **[72]** | Constant | Constant | None | Frontal pose |
| Multi-PIE **[73]** | 43 different illuminations | Unconstrained laboratory background | None | 9 yaw angles, 3 pitch angles, and two arbitrary angles |
| DMFP **[74]** | Constant | Constant | None | Nine yaw angles |
| MMI **[75, 76]** | Natural vs. controlled | Cluttered vs. solid color | Glasses vs. no glasses | Frontal pose vs. profile |
| FEED **[77]** | Constant | Constant | None | Frontal pose |
| PICS-pain **[78]** | Constant | Constant | None | Frontal pose |
| CAS-PEAL **[79]** | 135 different illuminations | Five different unicolor backgrounds | Glasses, hats | Nine yaw angles, three pitch angles |
| CK+ **[80]** | Constant | Constant | None | Frontal and 30° deviation from the facial midline |
| SEMAINE **[81]** | Constant | Constant | None | Frontal pose |
| RaFD **[82]** | Constant | Constant | None | Five yaw angles |

| | | | | |
|---|---|---|---|---|
| SFEW **[83]** | Unconstrained | Unconstrained | Unconstrained | Unconstrained |
| AFEW **[84]** | Unconstrained | Unconstrained | Unconstrained | Unconstrained |
| SMIC **[85]** | Constant | Constant | Glasses | Frontal pose |
| USF-HD **[86]** | Constant | Constant | Unknown | Frontal pose |
| CASME **[87]** | Constant | Constant | Glasses | Frontal pose |

* Accessories include such items as glasses, hats, makeup, facial hair, etc.

Table 6. Comparison of current facial expression databases

| | ***Posed*** | ***Posed & Natural*** | ***Natural*** |
|---|---|---|---|
| Basic emotions | SFEW, AFEW, KDEF, JAFFE | MMI | FEED |
| Basic & beyond basic emotions | DMFP, RaFD, PICS-pain | CK+ | SEMAINE |
| Beyond basic emotions | USF-HD, Multi-PIE, CAS-PEAL | none | CAMSE, SMIC |

All of these datasets are publicly available, but the major problem regarding most of them is that the collected facial expressions are not natural. A comparison of the existing databases is shown in Table 6. In this table, the databases are categorized according to the degree of naturalness of the expressions as well as the number of emotions included. Basic emotions involve the six basic emotions defined by Ekman [88], while beyond basic emotions involve more complex expressions and micro-expressions. Posed facial expressions were collected by asking the participants to act out certain expressions upon request. However, not everyone is an accomplished actor! Also, as argued in Chapter 2, not all facial muscles can be intentionally contracted by humans. Thus a simulated facial expression might be different from the genuine one.

In order to improve upon these databases, the new trend is to collect videos of natural expressions which are elicited by emotional content. For instance, FEED [77], CASME [87] and SMIC [85] all collected expressions while the participants watched pre-recorded emotion-inducing sequences. The FEED was restricted to the six basic facial expressions, while CASME and SMIC focused on spontaneous

micro-expressions. Even though cash rewards and simple punishments were used when creating the SMIC dataset, the stakes were still too low to be included in our study.

Considering that there is no existing dataset that is ideal for our research, we have collected our own database, which is introduced in the next section.

## 3.3. Introduction to Our Raw Database

Our raw database consists of videos we obtained directly from the Internet. Courtesy of Professor Stephen Porter from the University of British Columbia-Okanagan in Kelowna, Canada, we have obtained a list[5] of emotional pleaders who were asking for help to find their missing relatives or the murderers that killed or dismembered them. Approximately half of the pleaders were later convicted of murdering the missing or dead person, based on conclusive evidence, including blood or DNA matching, security videos, witness testimony, confession, etc. [52].

After an exhaustive search on the Internet, we were able to find about half of the videos involving the forensic cases listed by Porter. These might not be exactly the same as those originally collected by Porter, but the contents were most likely similar.

In addition, we gathered some videos absent from Porter's list, but containing circumstances that were similarly high-stakes. We further validated the guilt or innocence of the suspects, ensuring that all of the criminal cases in our database were closed. In other words, we did not include pending litigations. The criminal cases we have included in our database are listed in Appendix I.

In summary, we have collected Internet videos of a total length of 3.2 hours of 69 suspects. In most of these videos, the suspect is at a press conference appealing to the media for help to find either the missing person or the murderer. Sometimes,

[5] Professor Porter was unable to legally provide these videos but consented to send us a list of so-called pleaders.

the pleader is denying his involvement in the disappearance or death of the victim during a television interview. These situations yield high-stakes circumstances if the suspects attempted to lie, and obviously the guilty ones were all liars. Consequently the deception involved in these situations is what we are interested in. Detailed statistics considering our dataset is shown in Figure 2.

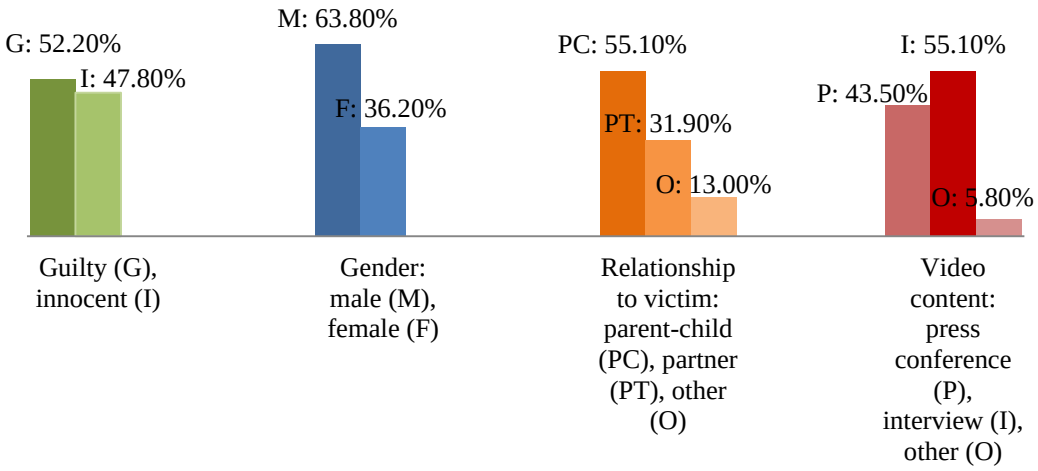



Figure 2. Distribution of data in our raw database (each bin is in percentage of suspects)

Different from most of the existing databases, our database is *completely uncontrived*.

First, the illumination conditions are extremely variable, depending on the video capture environment. Most of the videos were taken indoors, from arbitrary lighting directions and luminance levels, while some of them were taken outdoors under strong sunlight. Consequently, the shading on a face is not necessarily always uniform, and shadows might be formed on some parts of the face.

Second, the head pose is uncontrolled. During the plea or the interview, some of the suspects look down while making the plea. Some of them move their head arbitrarily when engaged in conversation with the interviewer. Thus the head pose

is not fixed at the frontal face position for each individual, resulting in large variations in all the three degrees of freedom[6].

Third, there is no constraint on the appearance of the suspects. There are several circumstances where the suspect is wearing a hat or glasses, or has a mustache or hair covering the forehead.

Fourth, the video quality varies greatly. The resolution could be as low as 320*240 and as high as 1280*720, and the video format is also diverse.

Fifth, the facial expressions are uncontrived. Because of the high-stakes situations the suspects are placed in, most of them speak emotionally. Their emotions are completely natural, rather than act on the basis of instructions. Moreover, most of the expressions occur while speaking, thus the expressions are affected by the mouth movements caused by utterance. In addition, there is no ground truth specifying when and where the expression related to deception occurs. In Porter's work [52, 68], a trained coder examined the videos, frame by frame, and labeled the movement of the Facial Action Units (AUs)[7] on a suspect's face. After that a second trained coder examined some of the videos to assess the veracity. Unfortunately, we do not have their annotations for any videos. Therefore, for our database, the only ground truth is whether a suspect is guilty or innocent, that is, lying or not. This limitation will probably introduce noise into our decision classifier, but it is an inevitable problem for us. The details of training a decision classifier will be elaborated in Chapter 5.

Some sample frames from the videos in our database are shown in Figure 3, from where we can see the confounding factors in our database.

Considering the many uncontrolled factors, our database is far from being the ideal interrogation dataset we specified in section 3.1. Nevertheless, to the best of our knowledge, it is the only one that contains completely natural expressions in

[6] The degrees of freedom of human head will be discussed in the next chapter.

[7] This will be discussed in detail later in the thesis. At this point, the reader can assume that this refers to the action of the facial muscles while emoting.

high-stakes situations. Therefore, our database is currently the most suitable dataset for studying and searching for a solution to real high-stakes deception detection problem. We also note that these uncontrolled factors have rarely been dealt with in the existing literature involving facial expression analysis. Traditional methods proposed for emotion detection have been mostly based on the prerequisite that all of the environmental factors discussed above can be controlled. Obviously, our database is significantly more difficult.

In the next chapter, we will present certain compensation methods that significantly eliminated the negative influence of the factors discussed above. Clearly, the feature descriptors of the face were also selected to be independent of these factors as much as possible. But before analyzing the videos, it is necessary to edit them to be more in accordance with our ideal situations.

## 3.4. Database Editing

Our system is intended to be used in the context of an interrogation, where the suspect is being questioned by an investigator. We assume that for this scenario, the suspect is being recorded by a fixed camera in a single session. Specifically, the suspect should be the only person appearing in view of the camera at a fixed distance from the suspect and at a camera angle in a near-frontal view. In addition, the interrogation environment should be unchanging and well illuminated. Also, the video should be shot at one time without intermissions, in order to avoid inconsistencies between the last frame in the previous shot and the first frame in the next shot.

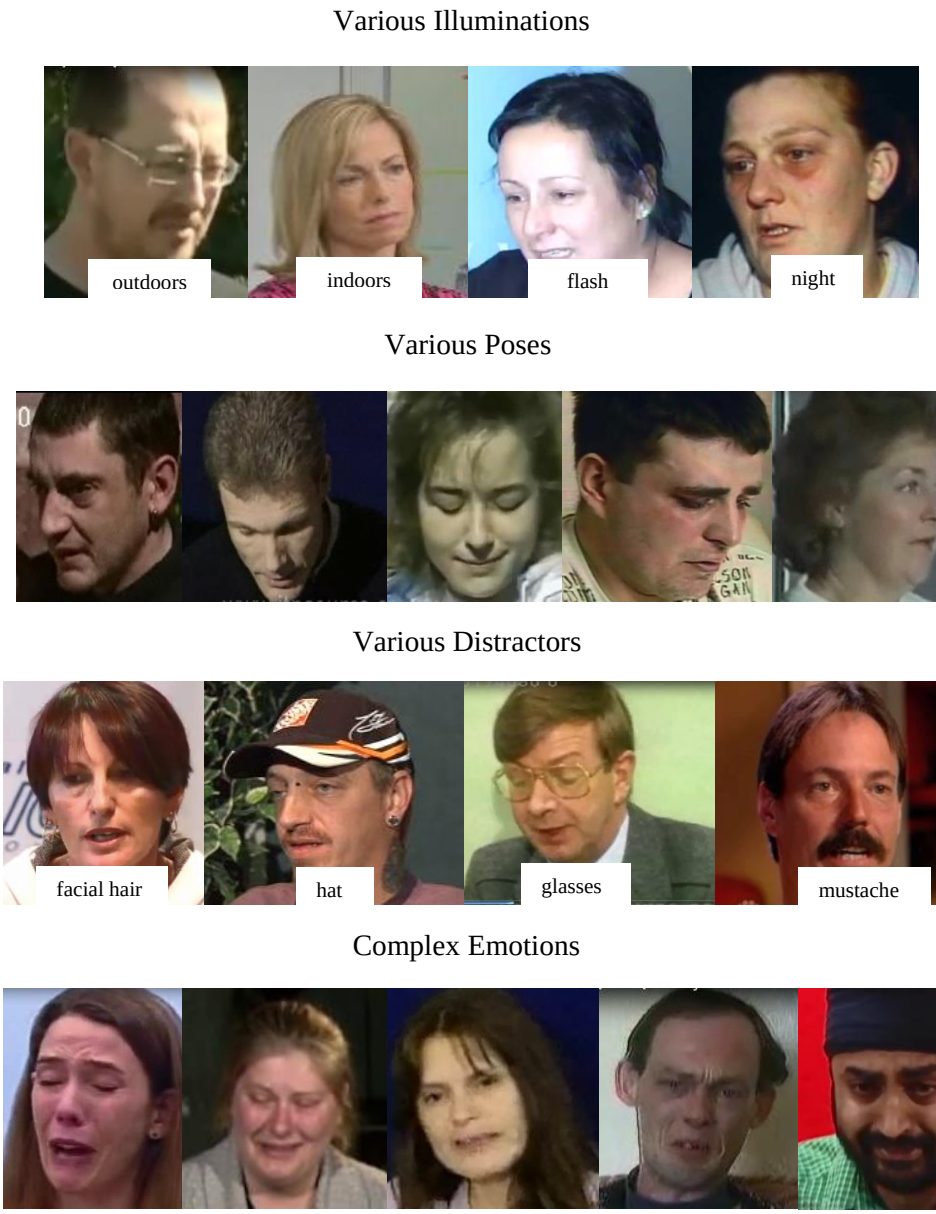



Figure 3. Sample frames from our database.

Note that all the videos in our database are obtained directly from the Internet and therefore did not satisfy the previous conditions. The video could be a news clip on TV, a short interview, or a press conference. The primary content of the video should be a person making an emotional plea to the public appealing for the safe

return of his missing relative or asking for help to find the murderer who cruelly killed or dismembered his dead family member. Often the person denies involvement in the disappearance or death of the victim in an interview. These principal speakers in the videos are all considered as "suspects", and they are either telling the truth or lying to conceal it. In addition, all the tragedies discussed are closely related to murder, disappearance, or sexual assault, yielding high-stakes situations for the suspects if they attempt to lie.

Most of the videos in our database contain multiple persons: the suspect, the interviewer, missing or dead relatives, the news anchors, etc. A person can appear individually in the camera scene, whereas in most cases, multiple people appear simultaneously. In light of our assumptions regarding the actual application being dealt with in this thesis, it is obvious that the videos need to be edited to mimic the real world situation. Figure 4 illustrates the typical simulated[8] video content in our database and the consequent elimination of the redundant data.

The video editing process is illustrated in Figure 4 and Figure 5, and the strategy is described as follows:

1) First, in order to eliminate the temporal diversity of different videos from different sources, they are temporally normalized to the same frame rate. Before temporal normalization, the original frame rate ranged from 16 frames per second (fps) to 60 fps; this becomes 30fps after normalization. When interpolating a lower frame rate video to the standard frame rate (30fps), intermediate motion between two frames is estimated and interpolated, instead of simply replicating the succeeding frames [89].

[8] In the sense that the "video" shown is constructed from several video clips.

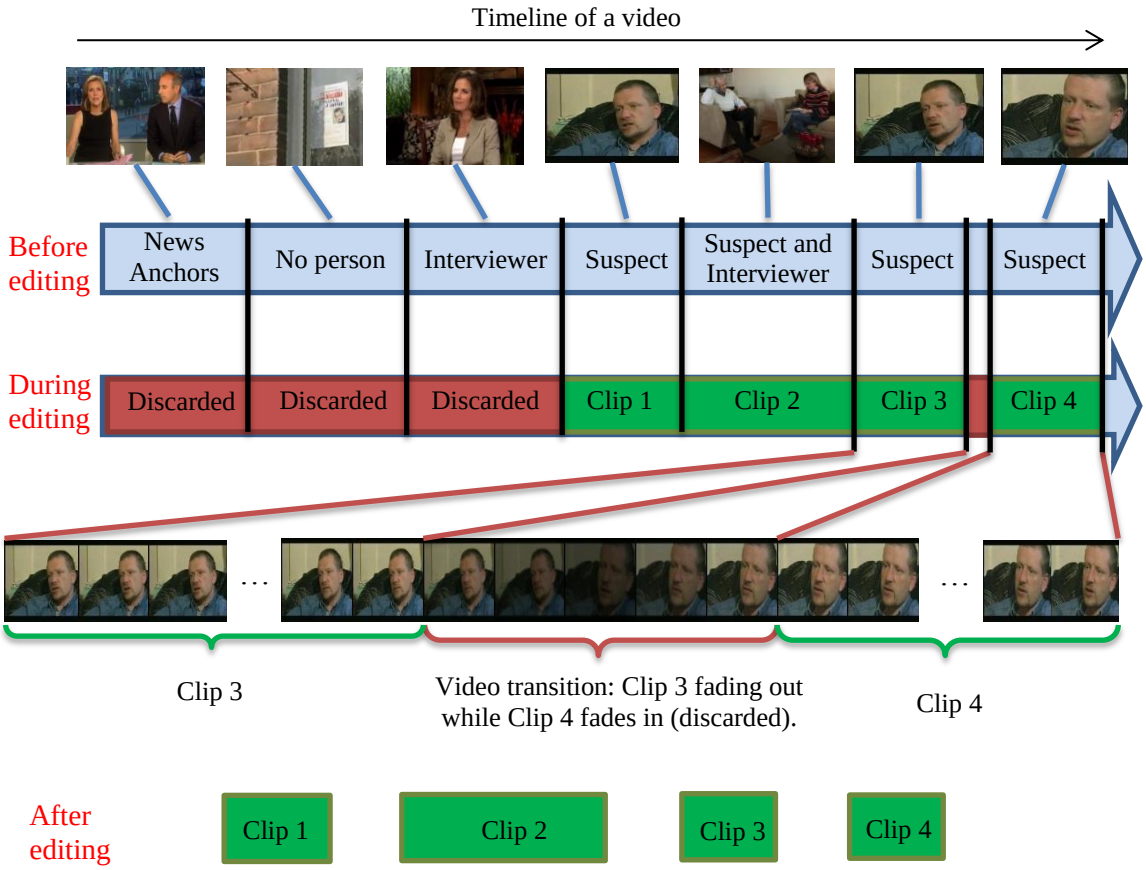



Figure 4. A sample edited video and the editing process. Note that this is for demonstration purposes only and the frames were not actually taken from one video.

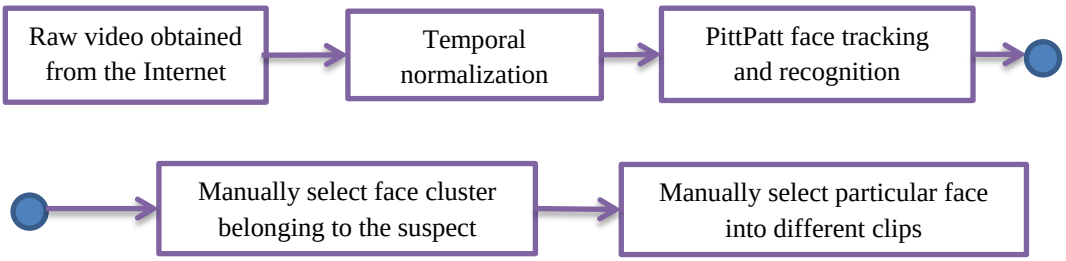



Figure 5. Video editing process

2) Select the frames that contain the suspect's face and locate it in each frame. This step was done using face tracking and recognition provided by PittPatt commercial software [90][9]. PittPatt tracks all faces in a video and then clusters the tracks into groups containing data from a single person. After that, we manually selected the face cluster belonging to the suspect we are interested in.

3) Despite the fact that the face of the suspect has been located, the video may consist of separate different clips. We define the latter as a single shot obtained by the same camera at the same fixed distance and angle. Therefore concatenating different clips into one video might result in inconsistencies of head position, head pose, and facial expressions. Considering this, it is necessary to separate the different clips manually by finding the start and end frame of each clip. However, many videos have additional transitional effects between two clips, which are added during the post-production process before presentation on TV or the Internet. The most common transitional effect is "fade in/out", which occurs when the video frame gradually turns into black or when a frame gradually appears from a black screen. Figure 4 gives an example of this transition effect between clip 3 and clip 4. In this case, the transition frames are usually darker and less reliable for facial expression detection. Therefore, if a clip fades out, the fading frames at the end of the clip need to be discarded. Similarly, for clip fade in.

To sum up, the final video is created by discarding the frames where the suspect does not appear and those affected by transition effects. The result after editing is a set of several independent clips from a video, as shown in Figure 4. Each clip is a sequence of the suspect in a relatively constant environment, which we used to compute the spatial-temporal features to be discussed later. After database editing, we obtained 324 video clips: 51.23% of the clips contain guilty suspects, while 48.77% are innocent. The average length of the video clips is 20 seconds.

The reader should be aware that we have used video clips instead of subjects for training and testing. In other words, if one suspect is guilty (deceptive), the clips

[9] This software is no longer available. The company has been bought out by Google.

belong to this person are all labeled as guilty (deceptive). It is similar for innocent (honest) suspects. Then all the video clips of the same suspect will be treated independently from each other. The experimental procedure will be presented in Chapter 5. In the next chapter, the dynamic features that are used for distinguishing deception and honesty are presented.

# Chapter 4. Dynamic Feature Analysis

Different from the deceit detection methods discussed in Chapter 2, in this thesis we present an automated method for detecting deception in high-stakes situations based on facial expressions. It will be seen in section 4.1 that honest and dishonest suspects present different facial expressions which are uncontrollable no matter what a person's intentions are. This is supported by several psychological theories that will be reviewed in section 4.1. From these theories we are able to determine the specific facial expressions used for distinguishing truth-tellers and liars, and consequently deduce the features for detecting them. Appearance-based methods are used to extract invariant 2D features that are related to the 3D characteristics.

In the rest of this chapter, we will present the dynamic feature analysis of a single video clip. In section 4.2, facial alignment is introduced. The following two sections will discuss facial region localization and feature extraction, respectively. Finally, we will illustrate how the dynamic features in different facial regions are organized and integrated into a concrete representation of a video clip.

## 4.1. Psychological Theories

Since none of the aforementioned measures provide a reliable solution for deceit detection, this thesis aims to test and verify if facial expressions can be used to reliably measure deceit in high-stakes settings. In fact, there have been several psychological theories published in the support of facial expressions being used as a reliable clue to deception detection. In this section, some of these will be reviewed in chronological order. We will then emphasize specific facial expressions that are potential indicators for discriminating deception and honesty in high-stakes situations.

Back in 1862, French neurologist Duchenne, first discovered the difference between a genuine (Duchenne smile) and a fake smile (non-Duchenne smile) [91]. He conducted this experiment using electrical stimulation of the *zygomatic* major muscle, whose contraction pulls the mouth corners up and forms a smile.

However, when Duchenne's experimental subject was amused by a joke, his *orbicularis oculi* muscle contracted simultaneously, thereby pulling up his cheeks and creating "crow's feet" around his eyes, as shown in Figure 6.

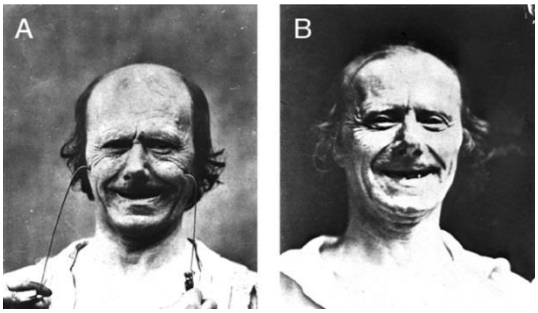



Figure 6. Duchenne's experiment on smiles. (a) non-Duchenne smile (fake smile). The zygomatic major muscle was stimulated electrically. (b) Duchenne smile (genuine smile). [91]

Later Darwin included Duchenne's experiment in his book, *The Expression of Emotions in Man and Animals*, and proposed the *inhibition hypothesis*. He argued that some facial actions that are the most difficult to create voluntarily are also the hardest to be inhibited [88]. Yet this theory had not been empirically tested until recently when Stephen Porter and his team conducted several related experiments [2, 52, 64, 66, 68, 92, 93]. This breakthrough work testified to the validity of Darwin's theory, as well as the possibility that the concealed emotions could be distinguished from the inconsistencies among normal facial expressions. The presence of concealed emotions is referred to as *emotional leakage*.

As early as 1966, Haggard and Isaacs claimed that they had observed some "micro-momentary" expressions when scanning some psychotherapy films [94]. Later Ekman also reported the existence of what he called "micro-expressions", which lasted between 1/25th to 1/5th of a second [1, 95]. Although Porter and ten Brinke [52, 64] have validated the existence of such micro-expressions, they also noted that they occurred very rarely, and only a partial face was involved. In [2, 52, 64, 66, 68, 92], Porter and his team indicated that the emotional leakage they

observed lasted too long to be classified as micro-expressions, but were so ubiquitous that they could possibly be utilized to detect concealed emotions. Recently, Hurley and Frank [67] have also shown that emotional leakage happens everywhere on the face and that facial countermeasures are very rare.

What about normal facial expressions? During the 70's, Ekman et al. [96] proposed the Facial Action Coding System (FACS) to describe facial movements during these expressions. FACS uses Action Units (AUs) to describe the contraction of certain muscles on human faces. For example, Table 7 gives the so-called seven universal facial expressions and their corresponding AUs. In 1980, Ekman et al. [97] reported a list of AUs which were the most difficult to produce deliberately. Based on Darwin's inhibition hypothesis, these AUs should be the hardest to inhibit. Therefore, their detection would indicate facial locations where emotional leakage would most likely happen. For example, according to this list [97], sadness should be hard to fake in the sense that AU1+4 is difficult to be produced intentionally. Therefore to conceal *genuine* sadness is equally difficult as presenting *fake* sadness.

Table 7. Universal expressions coded by Action Units [98]

| ***Emotion*** | ***Action Units*** |
|---|---|
| **Happiness** | AU 6+12 |
| **Sadness** | AU 1+4+15 |
| **Surprise** | AU 1+2+5B+26 |
| **Fear** | AU 1+2+4+5+20+26 |
| **Anger** | AU 4+5+7+23 |
| **Disgust** | AU 9+15+16 |
| **Contempt** | AU R12A+R14A |

In fact, this has been supported by recent studies. Porter et al. [2, 66, 68] have implied that guilty suspects would produce *fake* sadness. Since AU1+4 is hard to produce deliberately, the *fake* sadness would appear like surprise (AU1+2). Moreover, guilty suspects show involuntary leakage of *genuine* happiness (AU6+12) or smirks (AU12) to cover their embarrassment when telling lies. In contrast, innocent suspects would express *genuine* sadness (AU1+4, AU15),

which is hard to inhibit if they were feeling sad. Since happiness only occurs on the face of a guilty suspect, we will exclude AU6 in our feature analysis since we are not distinguishing genuine from fake happiness. The AUs associated with innocent and guilty suspects are summarized in Table 8.

Table 8. AUs of innocent and guilty suspects

| ***Emotion*** | ***Innocent Suspects*** | ***Guilty Suspects*** |
|---|---|---|
| Sadness | AU1+4, AU15 (Genuine) | AU1 or AU2 or AU1+2 (Fake) |
| Happiness | NA | AU6+12 (Genuine)<br>AU12 (Fake) |

In addition to the emotional leakage mentioned above, blinking could also be considered as a clue to deception. Mann et al. [57] have stated that the suspects will blink less frequently when telling lies in high-stakes situations. Leal and Vrij [99, 100] have found that the blinking pattern of liars and truth-tellers differ: liars show a decreased number of eye blinks when they are lying, followed by an increase. In [52], ten Brinke and Porter have also reported a higher blink rate observed in deceptive suspects. Therefore, it appears that blinking activity (AU45) could also be added as a cue to discern deception and honesty.

To sum up, the following AUs are potential indicators for distinguishing liars from truth-tellers in high-stakes situations: AU1, AU2, AU4, AU12, AU15 and AU45. Each AU is related to the movement of a single facial muscle and can result in motion of a facial part or appearance changes in a facial region. Also, multiple AUs can occur simultaneously.

Table 9 summarizes these potential deception indicators (AUs), their associated facial movements and corresponding facial regions.

Table 9. Potential indicators of deception

| *Action Unit* | *FACS Name* | *Facial Movement* | *Facial Region* |
|---|---|---|---|
| **AU1** | Inner Brow Raiser | Horizontal wrinkles occur in the center of the forehead; inner eyebrows move up, forming an oblique shape | Center of forehead; eyebrows |
| **AU2** | Outer Brow Raiser | Short horizontal wrinkles occur above the lateral portions of the eyebrows; outer eyebrows move up, forming an arched shape | Left and right forehead; eyebrows |
| **AU4** | Brow Lowerer | Vertical wrinkles occur between the eyebrows; distance between the eyebrows decreases; partial or entire eyebrows are lowered | *Glabella* area[10]; eyebrows |
| **AU12** | Lip Corner Puller | Lip corners move up obliquely; may create or deepen *nasolabial* furrows | Mouth |
| **AU15** | Lip Corner Depressor | Lip corners move down obliquely; may create or deepen *nasolabial* furrows | Mouth |
| **AU45** | Blink | Eyelids close and open rapidly | Eyes |

Therefore, based on the psychological theories presented above, the proposed method aims to detect the AUs listed in Table 9 and use them to discern deceptive and honest suspects in high-stakes situations. In order to detect and analyze the AUs, the face should be decomposed into several associated physical areas, as listed in Table 9. Before localizing facial regions, it is necessary to first align all faces of the same person in a video clip. This topic is presented in the next section**.**

## 4.2. Facial Alignment

### 4.2.1. Image Enhancement

Illumination conditions in the video clips we use in this study vary significantly since they were collected from the internet (examples can be seen in Figure 8(a)). However, we note that under realistic scenarios, interrogations would take place

[10] The *glabellar* area is the region between the two eyebrows.

indoors, under constant and uniform illumination on the suspect's face. For this reason, as well as to be able to deal with the analysis of a large range of image illuminations, it is necessary to compensate for the unconstrained lighting.

A plethora of illumination compensation methods have been published as a preprocessing step before image or video analysis, especially for face detection or recognition [101, 102]. One of the most conventional and effective methods of illumination compensation is histogram equalization (HE), which adjusts the overall distribution of the image intensity to achieve a higher contrast image. Similarly, adaptive histogram equalization (AHE) [103] adjusts the intensity distribution in each local contextual region, in order to further enhance image details. However, AHE has a disadvantage of excessively amplifying noise in relatively homogeneous areas of an image. Thus, we have adopted Contrast Limited Adaptive Histogram Equalization (CLAHE) [104] to avoid this issue. CLAHE modifies the intensity histogram by setting a clipping limit and uniformly redistributing the exceeded parts into histogram bins (shown in Figure 7). In this way, the hidden details in the image could be brought out, without amplifying too much noise. Figure 8(b) shows some example images after applying CLAHE.

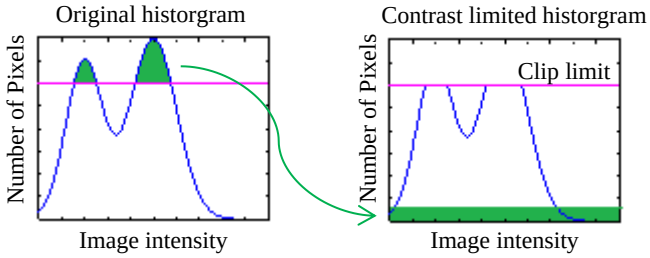



Figure 7. CLAHE method: the part of the histogram which exceeds the clip limit is redistributed uniformly into the histogram bins [104].

(a)

(b)


Figure 8. Image enhancement (a) Raw frames taken from videos in our database (b) Illumination compensated images by applying the CLAHE method to the images in (a).

### 4.2.2. Facial Landmark Detection

The human head has three degrees of freedom (DOFs): pitch, yaw, and roll, as shown in Figure 9. PittPatt[11] [90] locates different numbers of landmarks on a face according to its yaw angle. When the yaw angle is within the range of $[-36°, 36°]$, PittPatt detects at most three landmarks: left eye, right eye, and nose base. Otherwise, it detects at most five landmarks: left/right eye, nose base, eye nose, left/right lower cheek, left/right upper cheek. The definitions of these landmarks and example images in both cases are shown in Figure 10. We considered only faces whose yaw angles fell within $[-36°, 36°]$, thereby ensuring the visibility of both eyes.

[11] PittPatt developed software (which we licensed) for detecting and tracking faces in videos. It was later bought out by Google and the software is no longer available.

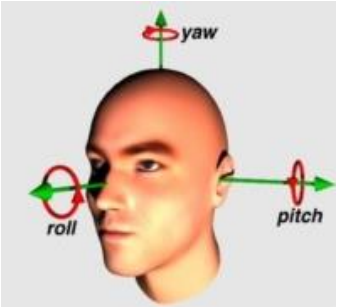



Figure 9. Three DOFs of human head: pitch, yaw, and roll. [90]

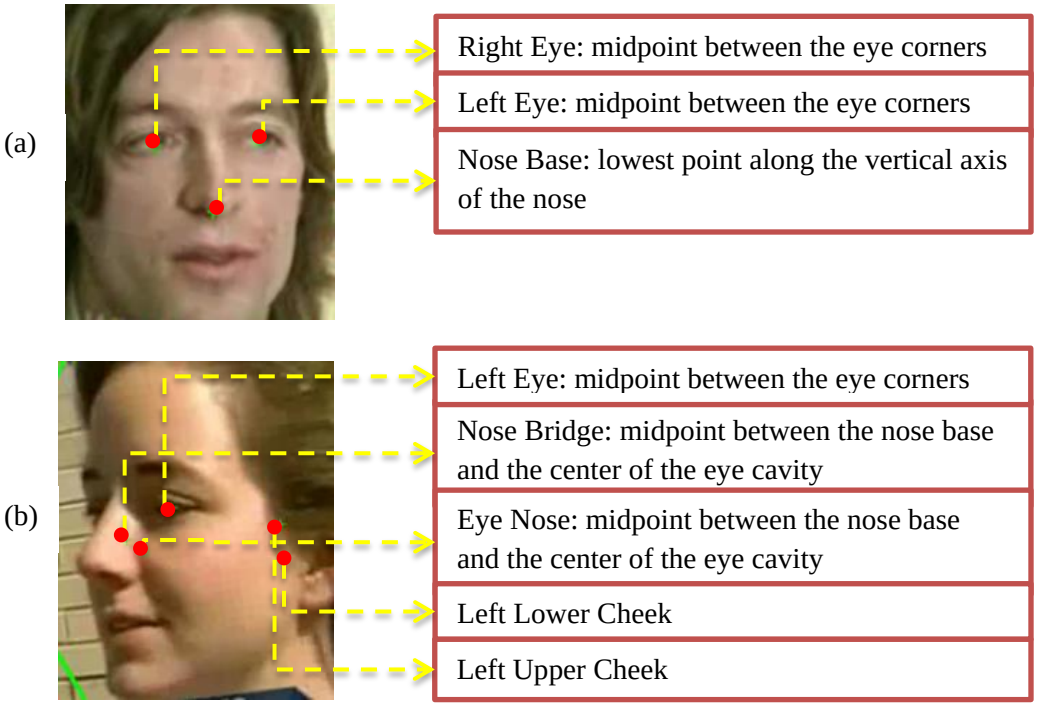



Figure 10. Facial landmarks detected by PittPatt at different yaw angles. (a) yaw angle $\in [-36°, 36°]$, PittPatt detects at most three landmarks: left eye, right eye, and nose base. (b) yaw angle $\in [-180°, 36°) \cup (36°, 180°]$; PittPatt detects at most five landmarks: left/right eye, nose base, eye nose, left/right lower cheek, left/right upper cheek. [90]

### 4.2.3. Nose Landmark Correction

Although the facial landmark detection using PittPatt is relatively fast and accurate, it locates landmarks frame by frame and does not adopt any tracking techniques. Consequently, the obtained landmarks have not been smoothed temporally, resulting in noisy landmark trajectories, which will later degrade the face registration results. To smooth the landmark trajectories, nose base landmark correction, followed by Kalman filtering, is applied.

The nose base is defined by PittPatt [90] as the lowest point along the vertical axis of the nose. Compared to eyes which are rather deformable regarding to their appearance, the area surrounding the nose base has relatively less appearance variations, allowing us to adopt a simple algorithm to correct its location, as described below.

As illustrated in Figure 11 for frame $B$ at time $t$, its previous frame $A$ at time $t-1$ is used as a reference frame for nose landmark correction. A small patch surrounding the nose base landmark in frame $A$ is the "reference patch" for locating the best "matching patch" in frame $B$. In frame $B$, we search for the "matching patch" in a larger search area surrounding its original nose base landmark by minimizing the sum-squared-difference (SSD) between the two patches. To speed up the computation, we employ the Fast Fourier Transform for calculating the SSD metric [105]. Starting from the first frame of a video clip, the nose landmark correction process is applied iteratively. The SSD correction result will be shown in Figure 12.

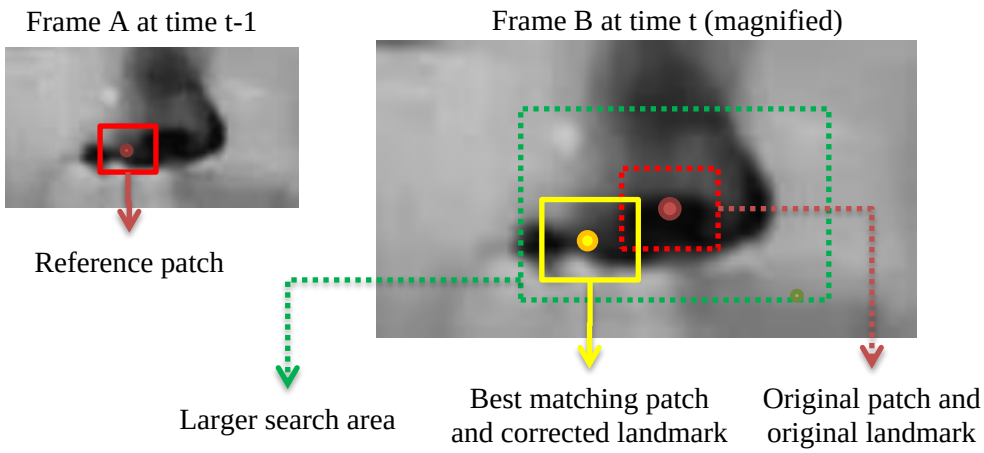



Figure 11. SSD correction of nose base landmark

### 4.2.4. Landmark Trajectory Smoothing

In addition, a Kalman filter [106, 107] is employed to further smooth the landmark trajectories. The Kalman filter is a recursive method for estimating the true state of an object, based on noisy measurements observed at the previous time. In our case, the objective is to track eye/nose motion on the basis of noisy measurements. After applying the Kalman filter, the eye/nose landmark trajectory will be smoother.

For the nose landmark, it is firstly corrected by SSD, and then further smoothed by the Kalman filter, resulting in a smoother trajectory. An example of the trajectory smoothing process of the nose landmark is shown in Figure 12. For the eye landmarks, they are merely smoothed by the Kalman filter.

The smoothness of landmark trajectories is essential to the feature extraction algorithms to be described in section 4.4.

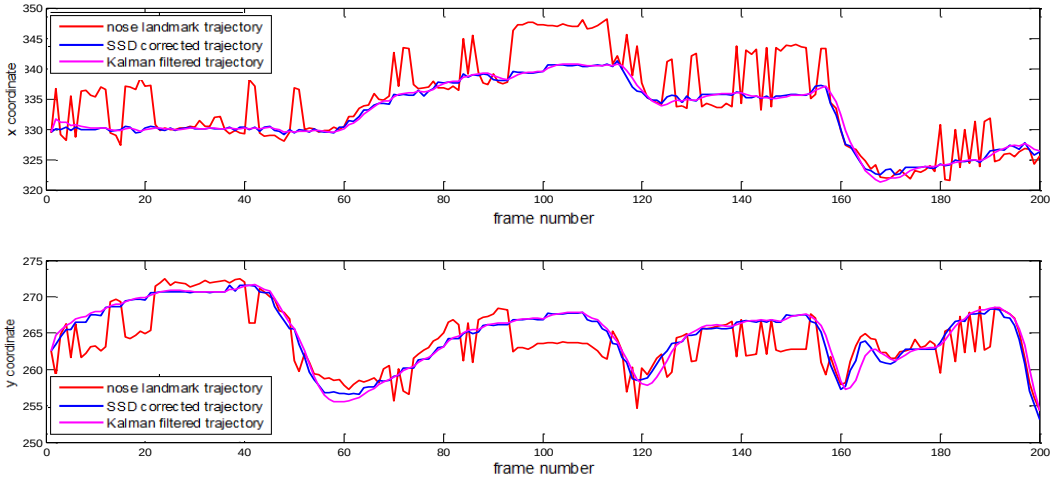



Figure 12. Example of nose landmark trajectory smoothing. The upper and bottom figures are the traces of x, y coordinates of nose landmark as time changes, respectively. In both figures, the red curve is the original trajectory detected by PittPatt. The blue curve is the trajectory after SSD correction. Magenta curve results from the blue curve further smoothed by the Kalman filter.

## 4.3. Facial Region Localization

Considering that only three major facial landmarks are obtained so far using PittPatt. However, our facial analysis ranges across the whole face from forehead to mouth, making it necessary to locate more facial points.

Active Shape Model (ASM) [108] is a popular statistical model which fits a new face to the learned model by iteratively searching for the best positions of a set of pre-defined markers. Active Appearance Model (AAM) [109] resembles ASM a lot, except that AAM utilizes texture information in addition to shape constraints. One of the disadvantages of ASM and AAM is that they both need a well-trained model to fit a new face, which needs to be learned from a large diversity of annotated images to ensure generalizability. Another disadvantage is that the performance of the model fitting process is also sensitive to the number of iterations.

Another popular approach is training a cascaded classifier. Usually, AdaBoost or GentleBoost is employed to learn the classifiers using such features as Haar templates, Gabor filter responses, or other local texture features. These have all proved to be fast and effective [102, 110, 111]. However, this method is also based on training, thus the detection performance largely relies on the diversity of the training samples.

As opposed to appearance- or shape-based methods, geometry-based facial landmark localization approaches do not require a training phase. They rely on knowledge of the anthropometric structure of the face. Studies have shown that, despite the variety of facial structures resulting from racial and individual differences, the human face is self-constrained. In other words, the distances between each pair of facial landmarks have inherent geometric relationships with each other, which provides the possibility of building a face model statistically. In [112], hundreds of frontal face images acquired from 150 people from various geographical locations have been carefully measured, and a face model was built.

The anthropometric face model presented in [112] is represented by seven landmarks $P_1 \sim P_7$ and the anthropometric measurements $D_1 \sim D_5$, as illustrated in Figure 13. The distance $D_1$ between the left and right eye is used as the principal measurement, and other distances $D_2 \sim D_5$ are proportional to $D_1$ (Table 10). According to the anthropometric face model, once the eye landmarks $P_1, P_2$ and the distance $D_1$ between them have been obtained, five other landmarks (left eyebrow center, right eyebrow center, midpoint of eyes, nose tip, mouth center) could be geometrically located by calculating $D_2 \sim D_5$ according to Table 10.

However, this model is limited to frontal face images, in which the head has no in-plane or out-plane rotation. As stated in section 3.2, faces in our database are subject to rotation variation. Therefore, head rotation along three orthogonal axes (pitch, yaw and roll) will need to be compensated for if a frontal face model were to be used.

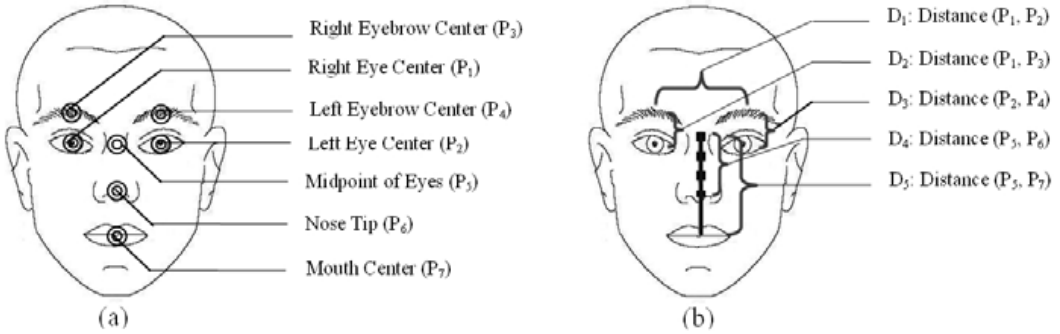



Figure 13. Anthropometric face model. (a) The landmarks in the face model. (b) The distances measured by the anthropometric model. [112]

Table 10. The proportion of the distances (D2, D3, D4, and D5) to D1 measured from subjects coming from different geographical locations. [112]

| ***Proportion*** | ***Description*** | ***Constant*** |
|---|---|---|
| $D_2/ D_1$ | Proportion of the distance between right eye center and right eyebrow center to the distance between eye centers | ≈ 0.33 |
| $D_3/ D_1$ | Proportion of the distance between left eye center and left eyebrow center to the distance between eye centers | ≈ 0.33 |
| $D_4/ D_1$ | Proportion of the distance between midpoint of eye centers and nose tip to the distance between eye centers | ≈ 0.60 |
| $D_5/ D_1$ | Proportion of the distance between midpoint of eye centers and mouth center to the distance between eye centers | ≈ 1.10 |

Suppose that the three landmarks in the original frame are denoted as follows: $P_1'$ (right eye), $P_2'$ (left eye), $P_8'$ (nose base), as shown in Figure 14(a), and the Euclidean distances are defined as:

$$D_1' = EuclDist(P_1', P_2') \quad (1)$$

$$D_4' = EuclDist(P_8', \overrightarrow{P_1'P_2'}) \quad (2)$$

Note that the nose tip landmark is included in the anthropometric face model, but there is no particular landmark for the nose base. Here we use the detected nose base landmark as a substitute for the nose tip landmark in the anthropometric model. They are very close in the frontal face and our landmark localization only serves as a preliminary step to roughly locate the different facial regions.

We will first compensate for head rotation, and then locate more facial landmarks and facial regions according to the anthropometric model.

In-plane rotation only involves the roll angle $\varphi$ as determined by

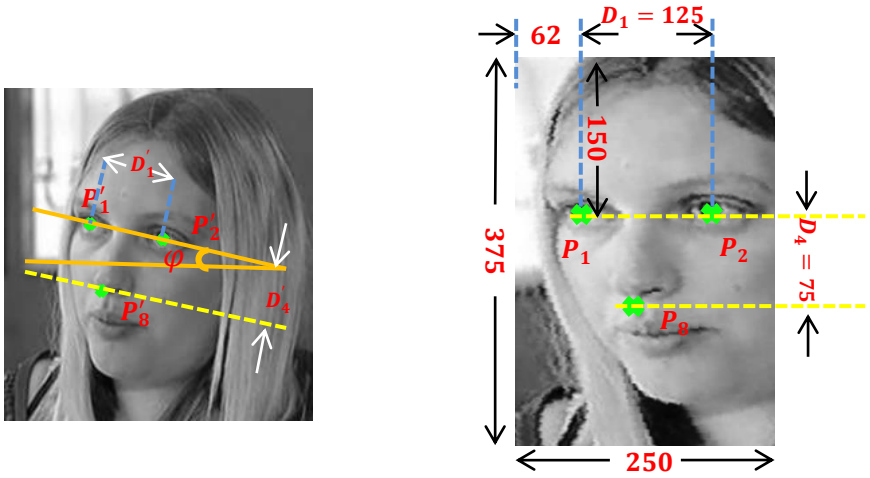
$$\varphi = tan^{-1}\left(\frac{P_{1y} - P_{2y}}{P_{1x} - P_{2x}}\right) \tag{3}$$

Intuitively, it is straightforward to compensate for it by rotating the image by $-\varphi$.

Out-of-plane rotation is more complicated than in-plane rotation, because it is related to 3D and not just 2D. We note that rotation about the yaw axis only affects horizontal distances ($D_1$) in the frontal face model, while rotation along the pitch axis only affects vertical distances ($D_2 \sim D_5$). The simplest approach for compensation for 3D rotation is to stretch $D_1'$ to $D_1$ and $D_4'$ to $D_4$ in the image, such that $D_1$ and $D_4$ satisfy the proportionality factors in Table 10:

$$D_1 = k_1 D_1',\ D_4 = k_2 D_4', and\ \frac{D_4}{D_1} = 0.60 \tag{4}$$

Therefore, the image was resized in the horizontal and vertical directions according to the factors $(k_1, k_2)$, respectively. In our case, we set $D_1 = 125$ and $D_2 = 75$. Finally the face (375×250) is cropped according to Figure 14(b).




(a) (b)

Figure 14. Spatial alignment. (a) Original frame, with three landmarks located. The original distance between the two eyes is $D_1'$, and the distance between the nose base to the line connecting two eyes is $D_4'$. The roll angle of the head is $\varphi$. (b) Aligned face, with the line connecting two eyes horizontal, and the horizontal distance between two eyes 125, the vertical distance between eye and nose 75. The cropped face has a dimension of 375×250, and the right eye is located at (62, 150).

After facial alignment, the frontal face model can be employed to locate other facial landmarks related to the left eye and right eye locations by computing $D_2 \sim D_5$, as shown in Figure 15(a). From these landmarks, nine facial regions can be located: left eye, right eye, left eyebrow, right eyebrow, *glabella* area, left forehead, right forehead, middle forehead and mouth. (shown in Figure 15(b))

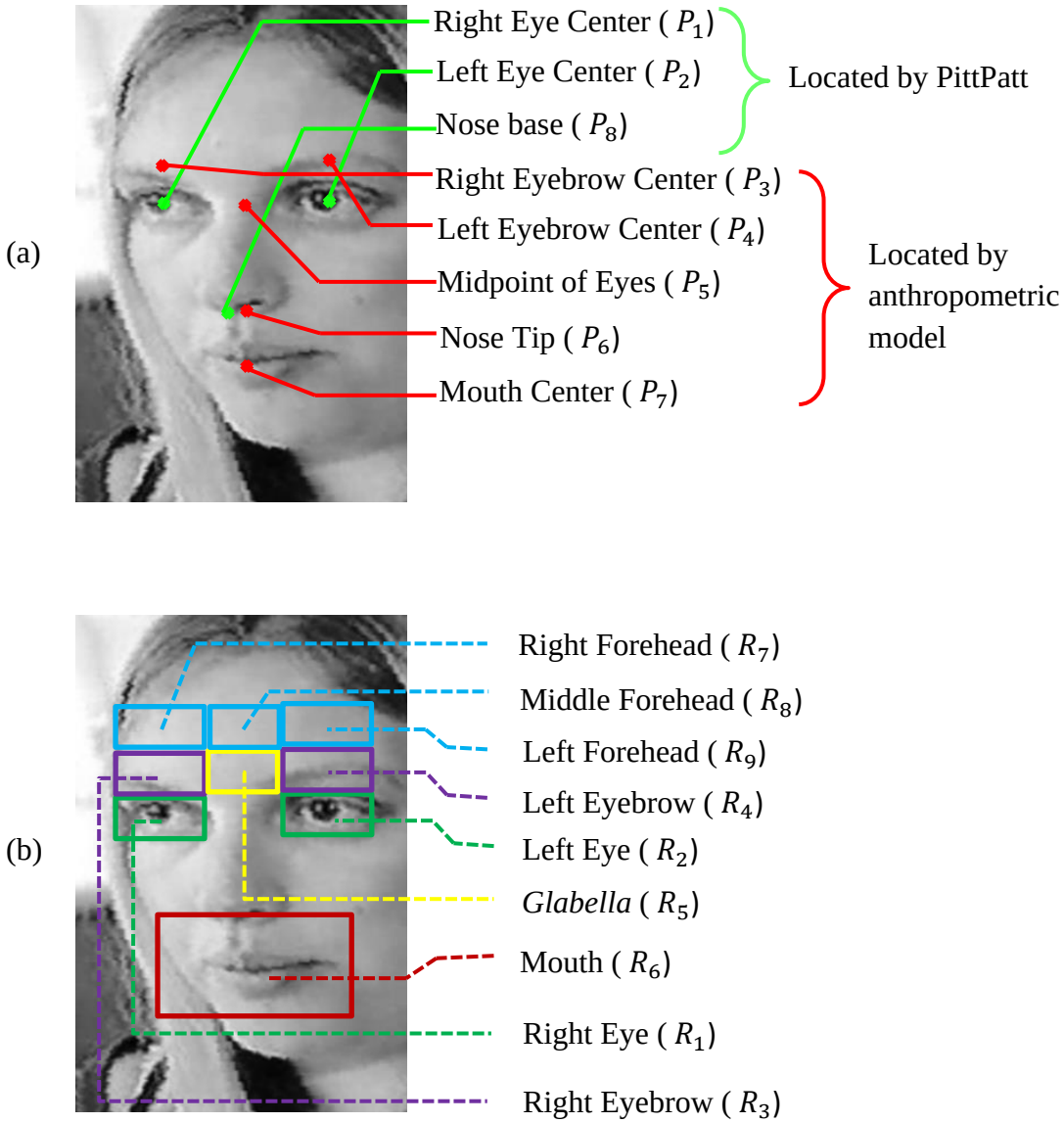



Figure 15. Facial landmark and region localization. (a) Green landmarks are the three primary landmarks located by PittPatt. Red landmarks are additional ones based on the anthropometric model. (b) Nine facial regions are located based on the landmarks in (a).

These facial regions are all regions of interest (ROI). On the basis of Table 8 discussed in section 4.1, each facial region is associated with one or more AUs. Each AU is related to a facial feature, which is called an "event". These events related to deception in the corresponding facial ROI in Figure 15(b) are summarized below in Table 11. The algorithm for detecting each event is presented in section 4.4.

Table 11. Events related to deception in each facial region

| ***Facial Regions*** | ***Events Related to Deception*** |
|---|---|
| $\boldsymbol{R_1, R_2}$ | Blink |
| $\boldsymbol{R_3, R_4}$ | Eyebrow moving upwards or downwards |
| $\boldsymbol{R_5, R_7, R_8, R_9}$ | Wrinkle |
| $\boldsymbol{R_6}$ | Mouth corners moving upwards or downwards |

## 4.4. Feature Extraction

In this section, feature analysis is conducted in each facial region. As stated in the previous section, four kinds of facial events that are related to deception need to be detected. These facial events are also considered as facial actions, and many algorithms have been proposed for solving the *local* facial action detection problem in the literature.

A general process of facial action detection involves local feature extraction and facial action discrimination. The common approaches used in the current research in feature extraction and decision making will be reviewed in below.

Geometric- and texture-based features for describing facial regions have both been very popular. Geometric-based feature involves distance, velocity, and displacement of a facial part. This kind of descriptor is more suitable for describing eyebrow or mouth movements [113-116]. For texture-based feature, the Gabor filter is commonly used to describe spatial compositions of eyebrow or transient wrinkles [113, 115]. Besides, other descriptors like Motion History Image (MHI) and free-form deformation (FFD) have also been applied to describe

the temporal changes in a video [115, 117]. However, there are some common prerequisites for these methods in the current literature, as discussed below.

*Neutral frame assumption:* It is assumed that there is at least one frame in the video clip in which the subject is neutral or expressionless [114, 115, 117, 118]. Normally the first frame or few frames at the beginning of a video were used to define a "baseline". Later the facial action analysis compared the target face with this neutral face. This assumption is reasonable since it eliminates the individual differences by measuring each person's facial action based on one's own baseline.

*Accurate facial landmark assumption:* A common technique for describing facial movements is to use facial landmarks. In this case, the facial movement within a facial region is considered to be represented by the movement of a few landmarks in it. In order to achieve high accuracy at facial action detection, facial landmarks should be located as accurate and many as possible. Active Shape Model (ASM) has been employed in locating 159 and 53 facial points respectively in [115, 119]. Another popular method to detect landmarks is to train boosted classifiers for each facial region, as used in [117, 120, 121]. There are other non-mainstream landmark localization approaches, such as manually locating facial points in the first frame [114] or employing landmark detection software [116]. If this assumption was satisfied, face registration and facial action description would be both easy and straightforward. However, it is very hard to train an ASM or boosted classifier that could be generalized to unconstrained situations with varying illumination, rapid head movement and uncontrolled facial occlusions. Therefore, this assumption is very hard to satisfy in terms of our dataset.

*Head pose assumption:* For accurate facial action detection, some papers have restricted the head pose of the subjects to be exactly or nearly frontal view [120, 122]. Although others did not have this restriction, the head pose in their case did not suffer from large or rapid variations and could be easily registered to compensate for head motions [115, 117]. However, the head rotation in our database is completely uncontrolled, making it impossible to meet this prerequisite.

For the decision-making part, many machine learning techniques have been employed to classify facial actions. Examples are: Gentle Boost in [115, 117], Support Vector Machine (SVM) in [118, 119], Bayes Classifier in [113], combination of Gentle Boost and SVM in [114] and a Hidden Markov Model (HMM) in [116]. The major difficulty of training a classifier is the selection of the training data. The best classification performance would be achieved if the training data had enough variations across all conditions. In our case, a good classifier should not be sensitive to human races, lighting changes, head pose changes and facial occlusions. It is already very hard to collect enough annotated data to train a facial action classifier, not to say training a classifier for each of the facial actions.

In summary, the traditional approaches for facial action detection all depend on certain assumptions, which are obviously not suitable for being employed in our study. Therefore, in the following subsections, we will introduce our methods for detecting the events listed in Table 11.

### 4.4.1. Eye Blink Detection

Eye blink is a dynamic process with the eyelid closing and opening rapidly. Researchers have extensively studied the blink detection problem, due to its various applications in human-computer interaction [123-125], driver fatigue monitoring [126-128] and liveliness detection [129]. Existing methods have either treated blink detection as an open-closed eye classification problem or put emphasis on the change of temporal patterns when blink occurs. Methods used in the literature include matching an open or closed eye template [123, 124, 130], measuring the vertical motion using optical flow [128, 131-133], computing the eye openness by miscellaneous methods [125, 126, 134-137], classifying the eye state using statistical approaches [138-141], and other methods [127, 129, 142]. These methods were proposed specifically for blink detection, based on empirical observations of the blink event.

The blink detection method used in this thesis differs from the traditional methods mentioned above. Instead of focusing on the *blink event* itself, we treat it as an anomaly which occurs within a period of time, while non-blinking is the normal behavior. The *blink event* involves the continuous process of eye closure, closed eye, and eye opening.

The anomaly detection method proposed in [143] has been adopted in this thesis. This method is an on-line real-time approach which detects suspicious behaviors (anomalies) occuring with a low probability in a video. The video is firstly represented by spatio-temporal volumes using dense sampling. Then the volumes are coded by temporal derivative descriptors and a codebook is constructed. Each volume in the video is assigned to all codewords with a degree of similarity. Then, the descriptors and arrangement of multiple volumes inside a larger contextual ensemble are modeled statistically to describe the spatio-temporal composition inside this ensemble. Finally according to the statistical model, the compositions that have lower likelihood of being normal are determined as anomalies. The algorithm is briefly summarized below:

*Algorithm for detecting anomalies:*

1) Construct spatio-temporal volumes $\{STV_1, STV_2, \dots, STV_T\}$ from the video by dense sampling.
2) Compute spatio-temporal derivatives in each volume as descriptors.
3) Construct a codebook consisting of $\Gamma$ codewords: $\{cw_1, cw_2, \dots, cw_\Gamma\}$ from the volume descriptors. Each volume $STV_i$ is assigned to all codeword $cw_j$ with a degree of similarity $w_{i,j}$.
4) Construct ensembles containing multiple spatio-temporal volumes. Each ensemble is represented by the volume descriptors and the relative arrangement of the volumes.
5) For each ensemble, compute its likelihood of being normal by measuring its similarity to the compositions learnt from the previously seen ensembles, yielding a likelihood map for each frame.
6) Threshold the likelihood maps, obtaining binary maps with the anomalous area detected.

In summary, the anomaly detection method detects the anomalous regions in a video based on its probability of occurrence.

For describing each spatio-temporal volume, we have utilized the Histogram of Oriented Gradients [144] based on a Sobel gradient operator kernel to emphasize on both spatial and temporal changes.

By applying the anomaly detection to the eye region in a video, we are able to obtain a likelihood map for each frame, as shown in Figure 16(b). Higher probability of being an anomaly corresponds to a blink event in our case.

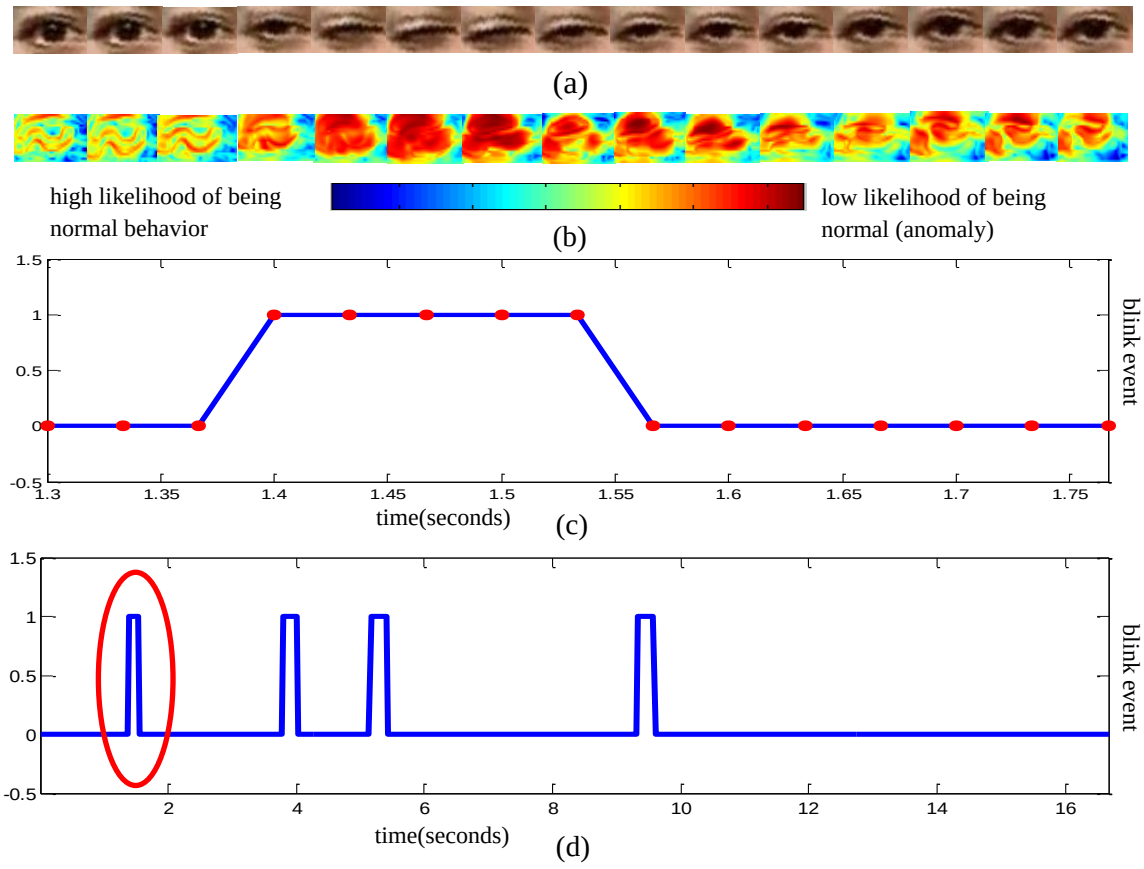



Figure 16. Blink detection. (a) A sequence of the eye ROI where blink occurs. (b) A sequence of the corresponding likelihood maps obtained from the anomaly detection algorithm. (c) A binary sequence where ones are the frames involving a blink event and zeros are 'non-blink' frames. Each red dot on the curve corresponds to a frame in (b). (d) A binary blink event sequence in a 17-second video with four blink events detected. The first blink event (circled in red) from 1.3s to 1.77s is blew up in (c).

Since the threshold for binarizing the likelihood maps was chosen experimentally for each dataset in [143], the thresholds used in [143] would not be suitable for our case since we are not using the same dataset. Therefore, we have adopted an automatic unimodal thresholding method based on the histogram of the likelihoods to determine a data-dependent threshold for each video clip.

Considering that anomalies occur less frequently than normal behaviors, the pixels that belong to an anomaly should only form a minority of all the pixels in a video. Moreover, these pixels should have lower likelihood of being normal. Consider the plot of the histogram of likelihood values for *all* of the likelihood maps in a sample video, shown as the blue curve in Figure 17. The dominant peak of the blue curve is formed by the pixels belonging to normal regions, while the smaller peak to the left of the dominant one is formed by the pixels belonging to anomalous regions. Sometimes the smaller peak is implicitly located under the side lobe of the dominant one. Therefore, we have employed the valley-emphasis thresholding method [145] to find an appropriate threshold between the two peaks. For example, the magenta line in Figure 17 is the threshold determined from the blue curve.

To determine if a frame is a ‘blink’ frame, the histogram of its likelihood map is used. If the corresponding likelihood value of its dominant peak is below the threshold found above, this frame is classified as a ‘blink’ frame; otherwise it is taken as a ‘non-blink’ frame. Figure 17 gives an example of the determination of a ‘blink’ and a ‘non-blink’ frame.

Based on this approach, a binary sequence is obtained from the likelihood sequence in Figure 16(b). The blink event is plotted as an on-off curve, as shown in Figure 16(c). Note that since the average length of blinking is between 0.1s and 0.4s [146], the events detected longer than 12 frames or shorter than 3 frames are eliminated.

As a result, each sequence of eye region is represented by a binary sequence, indicating the blink events detected. This binary sequence is the feature descriptor

of the eye region in a video, and will be combined with other features from other facial regions, and contribute to the final decision. This will be demonstrated in detail in section 4.5.

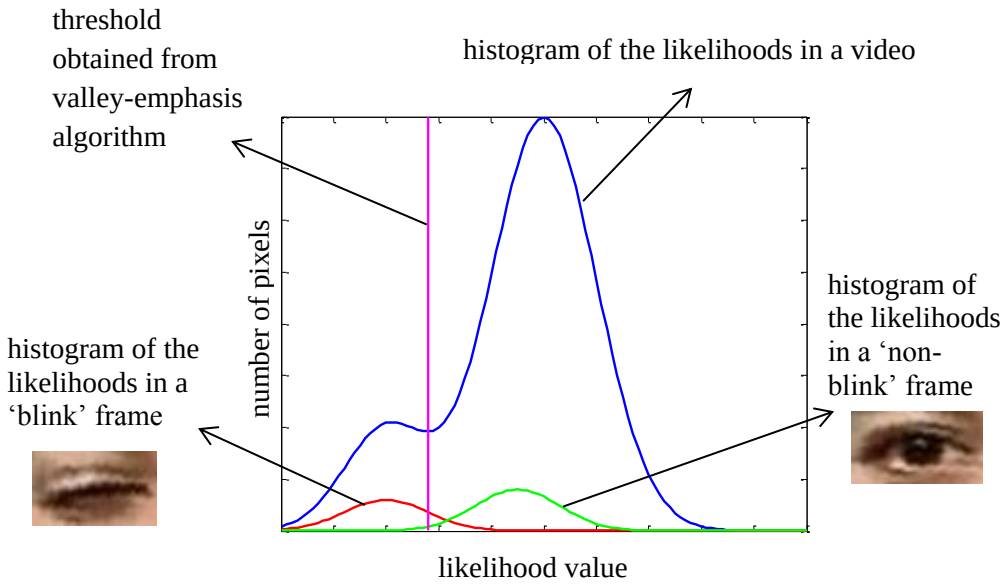



Figure 17. Demonstration of the determination of a 'blink' frame and a 'non-blink' frame. The blue curve is the histogram of the likelihood values in all likelihood maps of a video. The vertical magenta line is the threshold determined by the valley-emphasis thresholding method. The green curve is the histogram of the likelihood values in a frame. This frame is determined as a 'non-blink' frame because the corresponding likelihood value of the dominant peak of its histogram is larger than the threshold. Similarly, the associated frame of the red curve is determined as a 'blink' frame.

### 4.4.2. Detecting Eyebrow Motion

Eyebrow motion is closely related to different facial expressions. As argued in chapter 2, eyebrow movements are involved in both sadness and surprise expressions. People will pull inner eyebrows up when they feel sad. A surprise expression involves upward movement of the entire eyebrows, and this is also an expression of fake sadness. In the literature, eyebrow motion detection has been mostly used for emotional state estimation [60, 113, 114, 120] and avatar

animation [122]. Since the eyebrow movement involves motion instead of appearance change, geometric features are commonly used for characterizing the eyebrow displacement [60, 113, 114, 120].

Due to the various head poses in our database, the outer corners of the eyebrows are often missing from the videos. Therefore we merely focus on the upward and downward motion of the whole eyebrow, instead of treating the inner corner and outer corner separately.

Our eyebrow motion analysis algorithm involves two steps: eyebrow segmentation and eyebrow motion detection. In the first step, the eyebrow is segmented from the eyebrow ROI through the algorithm described below.

*Algorithm for eyebrow segmentation:*

1) Smooth the eyebrow ROI (Figure 18(a)) using a Gaussian filter (Figure 18(b)).
2) Transform the eyebrow ROI from RGB color space to $L^*a^*b^*$ color space. The $L$ component is shown in Figure 18(c).
3) Threshold the $L$ component using Otsu's method [147] to obtain a binarized eyebrow image. (Figure 18(d))
4) Select the blob in the binary image which has the maximum number of pixels along the central vertical line as the eyebrow blob. (Figure 18(e,f))
5) Obtain the convex hull of the eyebrow blob (Figure 18(g)) to achieve smoother outline of the eyebrow.
6) The upper contour of the blob in the previous step is taken as the upper contour of the eyebrow (Figure 18(h)). Also, the midpoint of the upper contour is located as the intersection point of the upper contour and the middle vertical line of the eyebrow ROI (Figure 18(h)).

Then in the second step, the eyebrow motion is detected by tracking the vertical motion of the midpoint of the eyebrow upper contour, as described in the algorithm below.

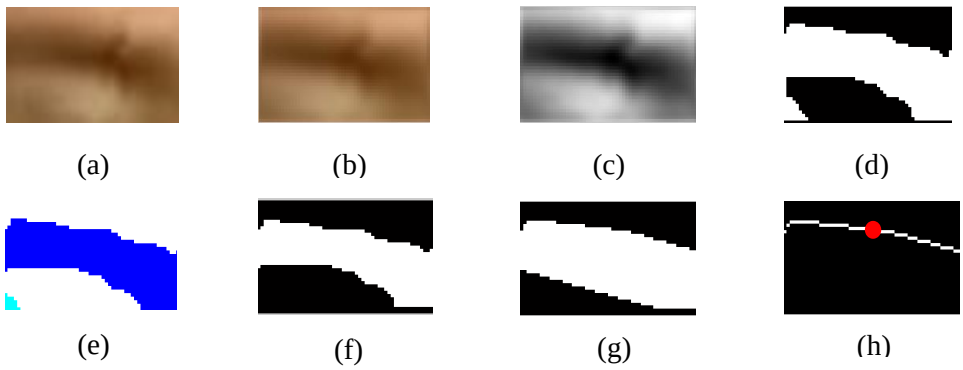



Figure 18. Eyebrow segmentation process. (a) Eyebrow region cropped from the anthropometric model. (b) Eyebrow ROI smoothed by a Gaussian filter. (c) L component in $L^*a^*b^*$ color space. (d) Binary eyebrow image after thresholding (c). (e) Blobs in (d) are labeled with different colors. (f) Selected eyebrow blob. (g) The convex hull of the blob in (f). (h) Eyebrow upper contour and its midpoint (red dot).

*Algorithm for detecting eyebrow motion:*

1) Measure the displacement of the midpoint of the upper contour from the horizontal middle line as the eyebrow displacement value. (Figure 19)
2) Track the displacement along time, obtaining a displacement curve, shown in Figure 20(a).
3) Apply a median filter to the displacement curve to eliminate outliers. (Figure 20(b) blue curve)
4) Apply a moving average filter to the curved obtained in the previous step to get a relative baseline of the eyebrow. (Figure 20(b) magenta curve)
5) Compute the difference between the curves obtained from step 3) and 4) to get the *relative displacement* of the eyebrow. (Figure 20(c))
6) From the curve obtained in the previous step, an *eyebrow raise event* is located if the height of the peak is higher than a threshold $\theta_{eyebrow}$. Similarly, an *eyebrow lower event* is located if the valley is lower than $-\theta_{eyebrow}$. (Figure 20(d))

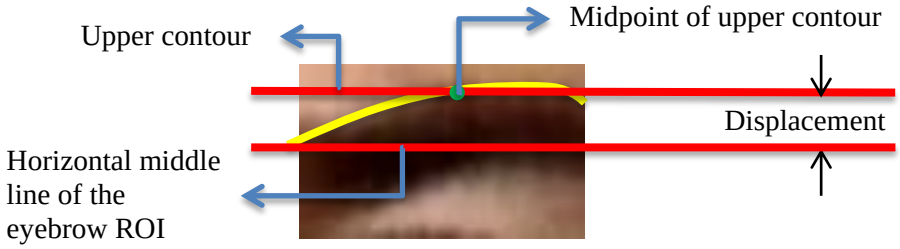



Figure 19. Eyebrow displacement

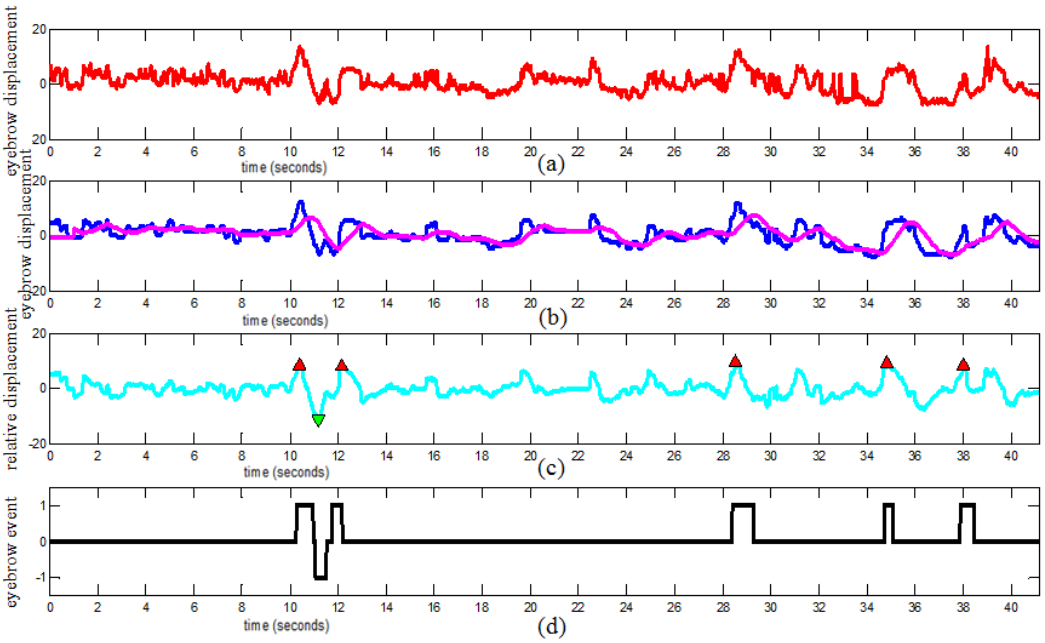



Figure 20. Eyebrow motion detection. (a) Eyebrow displacement varies along time in a video. (b) The blue curve is the median filtered curve of (a) and the magenta one is the moving averaged curve of the blue one. (c) Relative displacement curve. Red arrows are the peaks higher than the threshold, while green arrows are the valleys below it. (d) Eyebrow event curve. 1s indicate where the eyebrow raises, while -1s indicate where it lowers.

In summary, two kinds of events are detected in the eyebrow ROI: *eyebrow raise event* and *eyebrow lower event*. Both of the two events are obtained by thresholding the relative displacement curve mentioned above. The threshold $\theta_{eyebrow}$ is chosen experimentally during the training process, as will be

discussed in the next chapter. The detected eyebrow motion events will be integrated with other facial events in section 4.5, forming a comprehensive representation of the facial movements.

### 4.4.3. Detecting Wrinkles

Since facial wrinkles are generated by muscular movements, their direction is approximately perpendicular to the direction of motion of the corresponding muscles. Wrinkles can either be transient or permanent, both capable of yielding a reliable indicator for facial expression analysis [148, 149] and human age estimation [150-157]. In the literature, various automated features have been applied to represent the appearance of wrinkles, such as Gabor filter [149, 150, 158], Sobel filter [152-154], Hough transform [155], Active Contour [156] and Canny operator [148]. Also, there are other methods for detecting wrinkle segments using a watershed algorithm [151], Markov Chain Monte Carlo sampling [159], line sieving and morphological region growing [157].

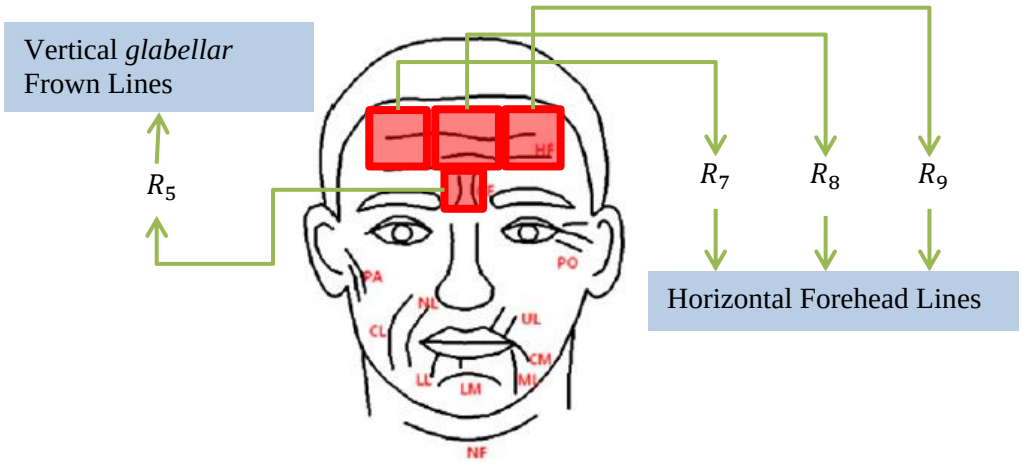



Figure 21. Facial wrinkles on a human face [160]. The horizontal forehead lines and *glabellar* frown lines in the red rectangles need to be detected.

Figure 21 indicates several primary facial wrinkles. Based on our anthropometric model, we focus on horizontal forehead lines in three ROIs ($R_7, R_8, R_9$) and vertical *glabellar* frown lines in $R_5$. The forehead lines in $R_7$ and $R_9$ are produced

by AU2 (outer eyebrow raiser), while those in $R_8$ are produced by AU1 (inner eyebrow raiser). The vertical *glabella* frown lines are generated by AU4 (brow lowerer).

Combinations of AUs can also produce wrinkles. The combination of AU1+2 will result in horizontal lines across the entire forehead ($R_7, R_8, R_9$). The simultaneous appearance of wrinkles in $R_8$ and $R_5$ is an indicator of AU1+4.

Table 12 is a summary of the above mentioned wrinkles in different facial regions and their associated Action Unit.

Table 12. Wrinkle ROI and its associated wrinkle direction and Action Unit

| ***Facial region*** | ***Wrinkle direction*** | ***Associated Action Unit*** |
|---|---|---|
| $\boldsymbol{R_7}$ | horizontal | AU2 |
| $\boldsymbol{R_8}$ | horizontal | AU1 |
| $\boldsymbol{R_9}$ | horizontal | AU2 |
| $\boldsymbol{R_5, R_8}$ | vertical, horizontal | AU1+4 |
| $\boldsymbol{R_7, R_8, R_9}$ | horizontal | AU1+2 |

Since wrinkles in a predominant direction (either horizontal or vertical) need to be detected, oriented Gabor filters will be sufficient to characterize the directional texture.

The Gabor filter is defined by equation:

$$gabor(x, y) = e^{-\frac{x\prime^2+\gamma^2 y\prime^2}{2\sigma^2}} e^{2\pi\frac{x\prime}{\lambda}+\varphi} \tag{5}$$

in which $x' = xcos\theta + ysin\theta, y' = -xsin\theta + ycos\theta$ [161].

Eight orientations of the Gabor filter are often used for detecting edges in different orientations, as shown below in Figure 22.

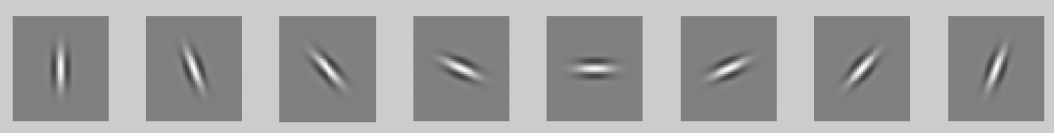



Figure 22. Oriented Gabor filters in eight different orientations by varying $\varphi$.

Though the face orientation has been spatially aligned, the wrinkles might not be exactly horizontal or vertical. Thus each ROI is convolved with a set of Gabor filters in more than one direction, and the combined response is used for detecting wrinkles. For horizontal forehead lines, three orientations were selected: $\varphi_i \in \left\{\frac{3}{8}pi, \frac{4}{8}pi, \frac{5}{8}pi\right\}$. For the vertical *glabellar* lines, three orientations were also selected: $\varphi_i \in \left\{0, \frac{1}{8}pi, \frac{7}{8}pi\right\}$. Three frequencies $\lambda_j \in \{8,12,16\}$ were computed in each direction, forming horizontal and vertical Gabor filter banks. Each directional bank contains 9 filters.

Suppose the Gabor response in a certain direction and scale is denoted as $gabor(\varphi_i, \lambda_j)$. Then the total response of $A$ directions and $B$ scales is defined as:

$$gabor_{response} = \frac{\sqrt{\sum_{j=1}^{B}\sum_{i=1}^{A} gabor(\varphi_i, \lambda_j)^2}}{A * B} \quad (6)$$

Figure 23 (a) shows the horizontal Gabor filter bank, and (b) is a sample *glabella* ROI with vertical frown lines. (c) is the associated Gabor responses by convolving (b) with each filter in (a), and (d) is the total response. Similarly, Figure 24 shows the vertical Gabor filter bank, a sample forehead ROI with horizontal wrinkle lines and its associated Gabor responses. From these two examples we could see that the directional textures are well captured by the oriented Gabor filters.

After computing the Gabor response, each ROI is represented by a Gabor-filtered response image. This image is transformed into an entropy value indicating the overall strength of the edges:

$$Entropy = -\frac{\sum gabor_{reponse} * \log(gabor_{reponse})}{number\ of\ pixels\ in\ the\ Gabor\ response\ image} \quad (7)$$

Then we consider a sequence of ROIs in a video with each frame having been characterized by a single entropy value. The *wrinkle event* could be detected by the algorithm described below.

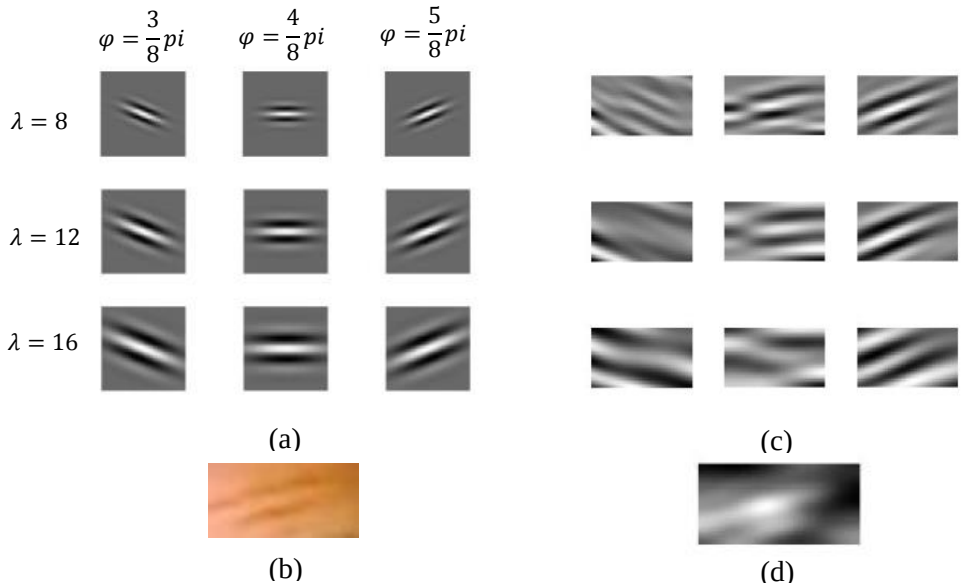



Figure 23. Horizontal Gabor filter bank and responses. (a) Horizontal Gabor filter bank consists of Gabor filters in three directions and three scales. (b) A sample image of forehead ROI with horizontal wrinkle lines. (c) Horizontal filter bank responses obtained by convolving (a) and (b). (d) Total Gabor responses.

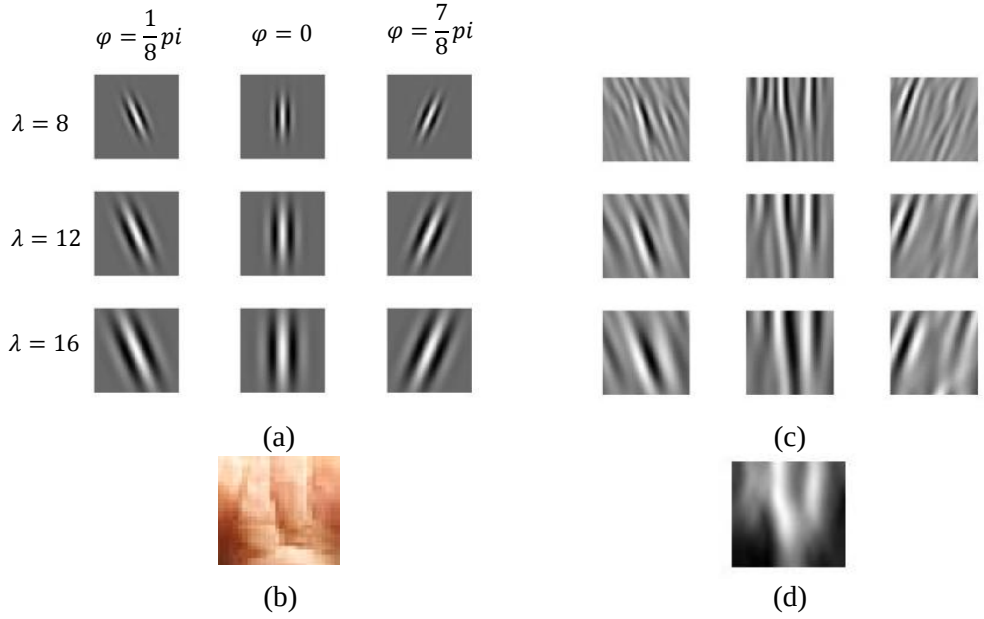



Figure 24. Vertical Gabor filter bank and responses. (a) Vertical Gabor filter bank consists of Gabor filters in three directions and three scales. (b) A sample image of *glabella* ROI with vertical frown lines. (c) Vertical filter bank responses obtained by convolving (a) and (b). (d) Total Gabor responses.

*Algorithm for detecting wrinkles:*

1) A sequence of wrinkle ROI (Figure 25(a)) is filtered by a horizontal or vertical Gabor filter bank, resulting a sequence of Gabor responses (Figure 25(b)).
2) The Gabor responses obtained in the previous step are transformed into entropy values, resulting in an entropy curve, shown in Figure 25(c).
3) The entropy curve is filtered by a median filter to remove outliers, shown in Figure 25(d).
4) The filtered entropy curve is thresholded by $\theta_{wrinkle}$, resulting in a binary sequence in Figure 25(e). The 1s indicate where the wrinkles occur, while 0s indicate where there are no wrinkles. A sequence of 1s is called a *wrinkle event*, representing that the wrinkles are occurring at one time and last for a period of time.

Therefore, for a video clip, wrinkle ROI $R_i \in \{R_5, R_7, R_8, R_9\}$ will have a binary sequence indicating where wrinkle events have occurred. Similar to the *eyebrow raise/lower event*, the *wrinkle event* will be integrated with other events from other facial regions in section 4.5 to achieve a compact representation of the facial movement.

Note that we were unable to arrange for automatically selecting an appropriate threshold $\theta_{wrinkle}$ for detecting a wrinkle event. Therefore the threshold was chosen during our video training process in which cross-validation was used to choose the best threshold. This will be discussed in detail in section 5.2.

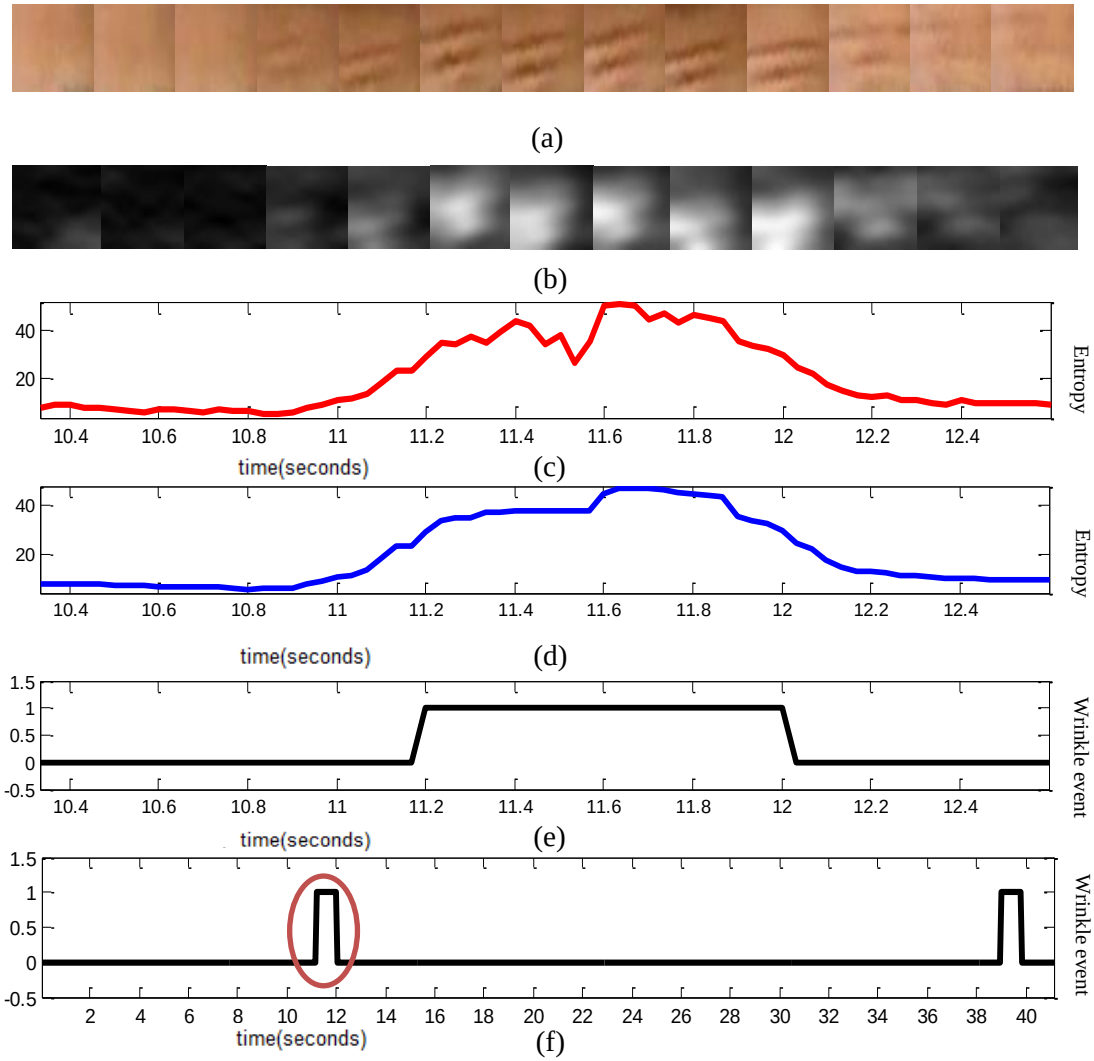



Figure 25. Wrinkle detection. (a) A sequence of wrinkle ROI taken from a video. (b) The corresponding Gabor responses. (c) Entropy curve of the frames in (b). (d) Entropy curve filtered by a median filter. (e) Binary sequence indicating where the wrinkle event occurs by thresholding the entropy curve in (d). (f) Two wrinkle events are detected in a video of 41 seconds. The first event (circled in red) was blown up in (e).

### 4.4.4. Detecting Mouth Motion

Amongst all the facial parts, the mouth is the most deformable one and has attracted considerable research interest. Research on automatically detecting the motion of the mouth can be categorized into three levels, depending on the degree of fineness. At the coarsest level, the mouth position is used for facilitating face detection task. When several face candidates are detected in an image, the mouth is located to verify which face candidate is a real face [162]. At the middle level, the mouth is classified into a few discrete states: open or closed [163, 164], lip corner pulled up or down [116], smile or not [165, 166]. At the finest level, the mouth is accurately tracked and characterized for visual speech recognition [167-169].

In our case, mouth is used for detecting happiness and sadness. This problem has been addressed in classical facial expression analysis. The process usually involves detecting features (local, global, or both) and then training a facial expression classifier [165]. Due to the variations in head pose and utterances, traditional methods involving training a classifier on static images will not perform very well. For example, the authors of [165] have argued that to develop a smile detector for practical application requires an enormous number of training samples with enough variations in head pose and illumination.

Instead we base our detector on a simple model of the expression behavior. A smile is indicated by the mouth corners pulling obliquely upwards (AU12), while sadness involves corners pulling obliquely downwards (AU15). As shown in Figure 26, a smile (AU12) is associated with an increase in the mouth angle $\emptyset$, whereas for sadness (AU15) it is the opposite. For both AU12 and AU15, there is an increase in the width of the mouth. Therefore, we measure the change of angle $\emptyset$ as well as the width $W$ simultaneously to determine AU12/AU15.

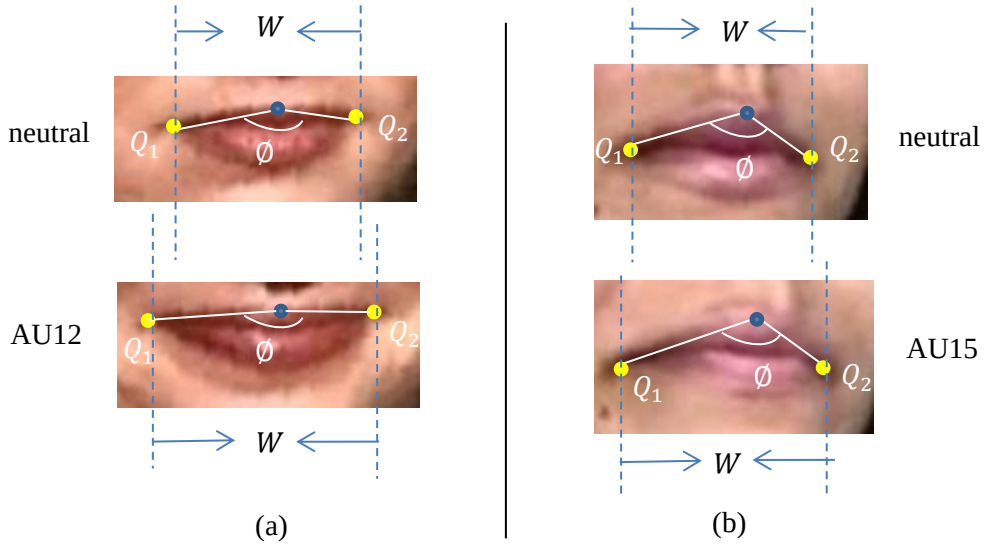



Figure 26. Change of mouth angle and width due to mouth movement. (a) AU12 is associated with an increase in both angle and width. (b) AU15 is associated with an increase in width and a decrease in angle.

Our algorithm for detecting the mouth motion involves two steps: mouth segmentation and mouth motion detection.

In the first step, the mouth is segmented from the mouth ROI by thresholding the "pseudo-hue" [170] of the image, and the feature points characterizing the mouth shape are located. The mouth segmentation algorithm is described as follows.

*Algorithm for mouth segmentation:*

1) The mouth ROI (Figure 27(a)) is smoothed by a Gaussian filter. (Figure 27(b))
2) The so-called "pseudo-hue" [170] of the mouth ROI is computed as

$$PseudoHue = \frac{R}{R+G} \tag{8}$$

shown in Figure 27(c).

3) A binary image is obtained by thresholding the pseudo-hue image, shown in Figure 27(d). The threshold is chosen using the Otsu's method [147].
4) Among all the mouth candidate blobs (Figure 27(e)), we select the largest one. (Figure 27(f))

5) The convex hull of the selected blob is computed and taken as the final mouth blob. (Figure 27(g))
6) Locate the left and right extreme points of the mouth blob as the left and right mouth corners $Q_1$ and $Q_2$. (Figure 27(h))
7) Note that the nose base point has been located by PittPatt. Therefore a vertical line could be drawn from the nose base point to the mouth blob. The highest and lowest intersection points of this line and the mouth blob are denoted as $P_1$ and $P_2$. They are taken as the upper and lower lip feature points. (Figure 27(h))
8) Compute the mouth angle $\emptyset$ which is the sum of angle $< \overrightarrow{P_1Q_1}, \overrightarrow{P_1P_2} >$ and $< \overrightarrow{P_1Q_2}, \overrightarrow{P_1P_2} >$, as shown in Figure 27(h). Width $W$ is measured as the horizontal distance between $Q_1$ and $Q_2$.

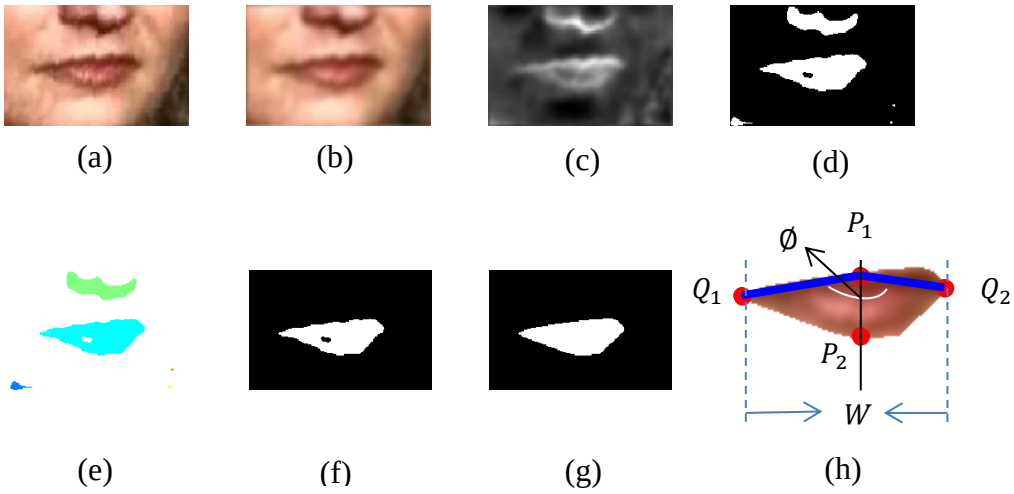



Figure 27. Mouth segmentation process. (a) The original mouth ROI. (b) Gaussian-smoothed mouth ROI. (c) Pseudo-hue image. (d) Binary image after thresholding the pseudo-hue image in (c). (e) Mouth blob candidates labeled in different colors. (f) Selected mouth blob. (g) The convex hull of the mouth blob. (h) Segmented mouth with mouth feature points located and measurements computed.

After segmenting the mouth and computing its angle $\emptyset$ and width $W$, the mouth motion could be detected by tracking the change of $\emptyset$ and $W$. A smile is taken as a happiness event $hp\_event$, and it is detected if both the angle and width increased significantly. Similarly, a sadness event $sd\_event$ is detected if the angle decreased while the width increased.

The algorithm for detecting the happiness events in a video is described as follows.

*Algorithm for detecting happiness events in a video:*

1) Compute the angle $\emptyset$ and width $W$ of the mouth for each frame in a video, resulting in an angle curve $\emptyset(t)$ characterizing the change of the mouth angle as time $t$ varies (Figure 28(a)) and a width curve $W(t)$ representing the change of the width (Figure 28(b)).
2) Since a happiness event involves an increase in width, $W(t)$ is thresholded into a binary sequence $hp_W = W(t) > W\_hp$ (Figure 28(c)). The 1s in this sequence are the candidates for happiness events.
3) Since a happiness event also involves an increase in angle, $\emptyset(t)$ is thresholded into a binary sequence $hp_\emptyset = \emptyset(t) > \emptyset\_hp$. Also, the 1s in this sequence are the candidates for happiness events.
4) The happiness events could be determined by synthesizing the candidates in the previous two steps: $hp\_event = hp_\emptyset \wedge hp_W$.

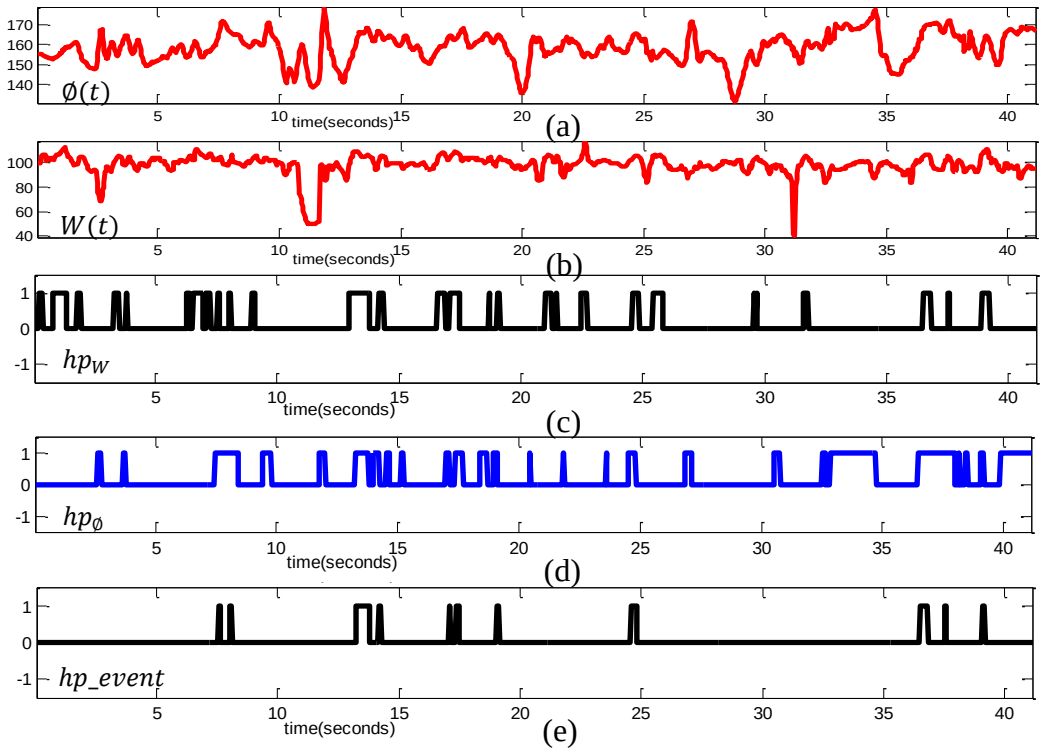



Figure 28. Happiness event detection in a video. (a) Mouth angle curve $\emptyset(t)$. (b) Mouth width curve $W(t)$. (c) Happiness event candidates $hp_W$ are detected where the width is larger than $W\_hp$. (d) Happiness event candidates $hp_\emptyset$ are detected where the angle is larger than $\emptyset\_hp$. (e) Happiness event $hp\_event$ is determined by synthesizing $hp_W$ and $hp_\emptyset$.

Similarly, the sadness event could be determined by $sd\_event = sd_{\emptyset} \wedge sd_W$, where $sd_{\emptyset} = \emptyset(t) < \emptyset\_sd$ and $sd_W = W(t) > W\_sd$.

Considering the individual differences in mouth shape and the influence of utterance and head pose on mouth scale and appearance, the thresholds $\emptyset\_hp, W\_hp, \emptyset\_sd, W\_sd$ cannot be chosen as absolute values. Therefore a *baseline* for each individual is estimated as the average value of the measurements over the whole video. We compute these thresholds as functions of a weighting factor $\rho$:

$$\begin{cases} \emptyset\_hp = mean(\emptyset(t)) + \rho * var(\emptyset(t)) \\ \emptyset\_sd = mean(\emptyset(t)) - \rho * var(\emptyset(t)) \\ W\_hp = mean\big(W(t)\big) + \rho * var\big(W(t)\big) \\ W\_sd = mean\big(W(t)\big) + \rho * var(W(t)) \end{cases} \tag{9}$$

By choosing different $\rho$, the thresholds are changed accordingly, representing how much the angle and width deviate from the individual baseline. Denote $\theta_{mouth} = \rho$, and this threshold could not be determined universally or by algorithmic steps. Therefore, it is also selected based on experimentation with the training set, as will be discussed in section 5.2.

Just like the other events obtained from other facial regions, the happiness and sadness events will be integrated with others and contribute together to the final decision. The integration strategy is presented in section 4.5.

### 4.4.5. Feature Extraction Summary

In summary, there are four kinds of features extracted from the face: eye blink, eyebrow motion, wrinkle occurrence, and mouth motion. Except for the eye blink feature, each of the other features is a function of a threshold $\theta_e \in [\theta_{e\_min}, \theta_{e\_max}]$, where $e \in \{eyebrow, wrinkle, mouth\}$. The continuous interval $[\theta_{e\_min}, \theta_{e\_max}]$ is discretized into $\gamma_e$ discrete values as the threshold candidates. Therefore, the threshold set $\Theta = [\theta_{eyebrow}, \theta_{wrinkle}, \theta_{mouth}]$ has $\gamma_{eyebrow} \times \gamma_{wrinkle} \times \gamma_{mouth}$ possible candidate values. The best threshold set

$\Theta_{best}$ was determined by experimentation using the training data. This process will be elaborated in the next chapter. In the following section, the integration strategy of all the detected events is presented.

## 4.5. Feature Integration

### 4.5.1. Primary and Secondary Features

As stated in Chapter 3, our database consists of 324 video clips. The database is denoted as $V = \{v_1, v_2, \dots, v_N\}$ in which $N = 324$. Each video clip $v_\alpha$ is further decomposed into nine sequences of specific regions on the face, $\mathcal{R}_\alpha = \{R_{\alpha,1}, R_{\alpha,2}, \dots, R_{\alpha,9}\}$. For each region $R_{\alpha,\beta}$, one or two binary feature vectors (events) are computed using the methods discussed in previous sections in this chapter. In total, 12 feature vectors have been computed from the nine facial regions. Each feature corresponds to a facial Action Unit. Since the 12 feature vectors are computed directly from the facial regions and each involves only one region, we refer to them as *primary features*. The primary feature set of video clip $v_\alpha$ is denoted as $P_\alpha = \{p_{\alpha,1}, p_{\alpha,2}, \dots, p_{\alpha,12}\}$. If the number of frames in clip $v_\alpha$ is $n_\alpha$, each primary feature vector $p_{\alpha,i}$ is of dimension $n_\alpha \times 1$.

The 12 primary features and their corresponding facial regions, Action Units and feature vectors are listed in Table 13.

Table 13. Primary features for video clip $v_\alpha$

| ***Action Unit*** | ***AU45*** | ***AU45*** | ***AU1/2*** | ***AU1/2*** | ***AU4*** | ***AU4*** |
|---|---|---|---|---|---|---|
| ***Event*** | right eye blinking | left eye blinking | right eyebrow raising | left eyebrow raising | right eyebrow lowering | left eyebrow lowering |
| ***Facial Region*** | $R_{\alpha,1}$ | $R_{\alpha,2}$ | $R_{\alpha,3}$ | $R_{\alpha,4}$ | $R_{\alpha,3}$ | $R_{\alpha,4}$ |
| ***Feature Vector*** | $p_{\alpha,1}$ | $p_{\alpha,2}$ | $p_{\alpha,3}$ | $p_{\alpha,4}$ | $p_{\alpha,5}$ | $p_{\alpha,6}$ |
| | | | | | | |
| ***Action Unit*** | ***AU12*** | ***AU15*** | ***AU4*** | ***AU2*** | ***AU1*** | ***AU2*** |

| ***Event*** | mouth corners moving up | mouth corners moving down | wrinkle in *glabella* area | wrinkle in right forehead | wrinkle in mid-forehead | wrinkle in left forehead |
|---|---|---|---|---|---|---|
| ***Facial Region*** | $R_{\alpha,6}$ | $R_{\alpha,6}$ | $R_{\alpha,5}$ | $R_{\alpha,7}$ | $R_{\alpha,8}$ | $R_{\alpha,9}$ |
| ***Feature Vector*** | $p_{\alpha,7}$ | $p_{\alpha,8}$ | $p_{\alpha,9}$ | $p_{\alpha,10}$ | $p_{\alpha,11}$ | $p_{\alpha,12}$ |

Considering the fact that some Action Units involve more than one facial region and multiple Action Units are likely to occur simultaneously, *secondary features* are also computed. For example, the left eyebrow raise involves both upward motion in the left eyebrow region ($p_{\alpha,3}$) and appearance changes in the left forehead region ($p_{\alpha,12}$). Therefore a secondary feature indicating the congruence of these two events can be generated by computing the logical conjunction of $p_{\alpha,3}$ and $p_{\alpha,12}$. In total, nine secondary features are computed, resulting in a secondary feature set $S_\alpha = \{s_{\alpha,1}, s_{\alpha,2}, \ldots, s_{\alpha,9}\}$, as shown in Table 14.

Table 14. Secondary features for video clip $v_\alpha$

| ***Action Unit*** | ***AU45*** | ***AU1/2*** | ***AU4*** | ***AU1+2*** | ***AU1+2*** |
|---|---|---|---|---|---|
| ***Event*** | eyes blink | eyebrows raising | eyebrows lowering | right eyebrow raising + wrinkle in right forehead | left eyebrow raising + wrinkle in left forehead |
| ***Facial Region*** | $R_{\alpha,1}$+$R_{\alpha,2}$ | $R_{\alpha,3}$+$R_{\alpha,4}$ | $R_{\alpha,3}$+$R_{\alpha,4}$ | $R_{\alpha,3}$+$R_{\alpha,7}$ | $R_{\alpha,4}$+$R_{\alpha,9}$ |
| ***Feature Vector*** | $s_{\alpha,1} = p_{\alpha,1} \wedge p_{\alpha,2}$ | $s_{\alpha,2} = p_{\alpha,3} \wedge p_{\alpha,5}$ | $s_{\alpha,3} = p_{\alpha,4} \wedge p_{\alpha,6}$ | $s_{\alpha,4} = p_{\alpha,3} \wedge p_{\alpha,8}$ | $s_{\alpha,5} = p_{\alpha,5} \wedge p_{\alpha,10}$ |

| ***Action Unit*** | ***AU1+4*** | ***AU1+4*** | ***AU4*** | ***AU4*** |
|---|---|---|---|---|
| ***Event*** | left eyebrow raising + wrinkle in mid-forehead | right eyebrow raising + wrinkle in mid-forehead | right eyebrow lowering + wrinkle in *glabella* area | left eyebrow lowering + wrinkle in *glabella* area |
| ***Facial Region*** | $R_{\alpha,3}$+$R_{\alpha,8}$ | $R_{\alpha,4}$+$R_{\alpha,8}$ | $R_{\alpha,3}$+$R_{\alpha,5}$ | $R_{\alpha,4}$+$R_{\alpha,5}$ |
| ***Feature Vector*** | $s_{\alpha,6} = p_{\alpha,3} \wedge p_{\alpha,9}$ | $s_{\alpha,7} = p_{\alpha,5} \wedge p_{\alpha,9}$ | $s_{\alpha,8} = p_{\alpha,3} \wedge p_{\alpha,7}$ | $s_{\alpha,9} = p_{\alpha,5} \wedge p_{\alpha,7}$ |

After concatenating the 12 primary features and 9 secondary features, video clip $v_\alpha$ will be described by a feature matrix denoted by:

$$\Omega_\alpha = [p_{\alpha,1}, p_{\alpha,2}, \dots, p_{\alpha,12}, s_{\alpha,1}, s_{\alpha,2}, \dots, s_{\alpha,9}] \tag{10}$$

which has a dimension of $n_\alpha \times 21$. In this matrix, each column defines a feature vector for a specific event as time progresses and each row is a $1 \times 21$ feature vector for a single video frame.

### 4.5.2. Feature Temporal Volumes

In order to create a compact representation for each frame, its temporal context, including the preceding and succeeding frames, should also be taken into consideration. As noted at the end of the previous section, each frame is represented by a $1 \times 21$ binary feature vector. Considering a context of $T$ consecutive frames centered at frame $t$, a Feature Temporal Volume (denoted as $FTV_{\alpha,t}$) can be obtained by summing up the $T$ feature vectors (Figure 29).

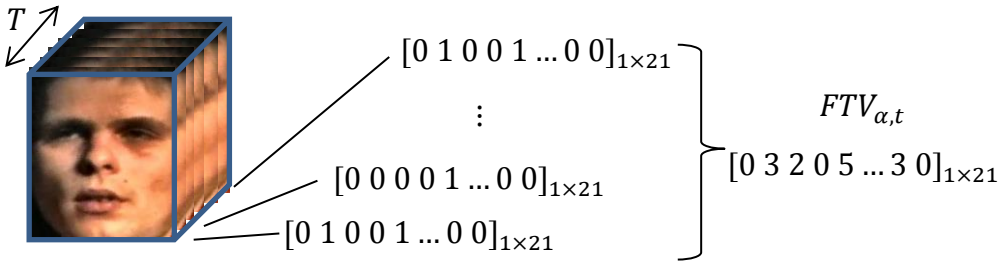



Figure 29. Construction of a feature temporal volume $FTV_{\alpha,t}$

Therefore, video clip $v_\alpha$ can be represented by a "bag of FTVs" (BOF). Each FTV in the bag is a $1 \times 21$ vector, and there are $m_\alpha$ FTVs in the bag, where $m_\alpha = n_\alpha - T + 1$.

In the next chapter, we will explain how the FTVs are used to train the classifier as well as how a prediction is made through a voting process.

# Chapter 5. Methodology

As stated in Chapter 3, our database consists of $N$ video clips, forming a pool of clips $V = \{v_1, v_2, \dots, v_N\}$. Furthermore, each video clip $v_\alpha$ is represented by a "bag of FTVs" (BOF) $\{FTV_{\alpha,1}, FTV_{\alpha,2}, \dots, FTV_{\alpha,m_\alpha}\}$ using the feature analysis methods presented in Chapter 4. Figure 30 is a demonstration of the global structure of the database in terms of BOFs.

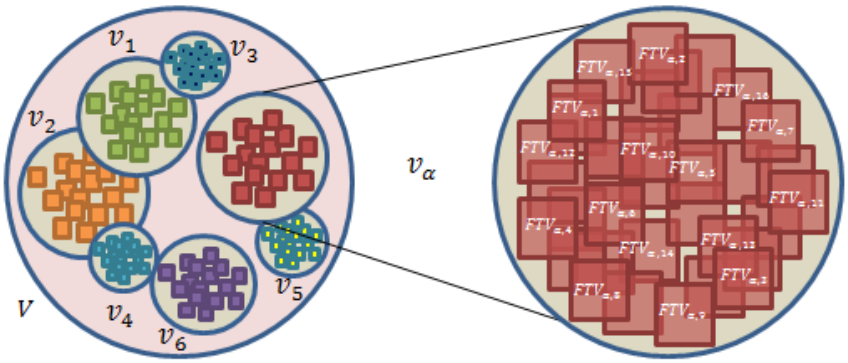



Figure 30. Global database structure. Our database is a pool of video clips (left). Each video clip is described by a BOF (right).

This chapter will focus on the decision-making step which relates the BOFs to the intrinsic nature of a video clip, that is, whether it contains deception or not.

To determine whether a suspect is deceptive or honest is a binary classification problem. A binary classifier can be constructed through a supervised learning process. In this process, the data are divided into two parts, one for training the classifier and the other for testing. During training, the classifier is learnt based upon labeled training samples, while the testing session predicts the class labels of the unseen data. As discussed in the previous chapter, there are three undetermined feature thresholds $\theta_{eyebrow}, \theta_{wrinkle}$ and $\theta_{mouth}$. Since the features of the video clips depend on these thresholds, it is necessary to select the best values of the thresholds which maximize the classifier performance through a cross-validation [171] process using the training data.

As presented in Chapter 3, the guilty suspects in our training database are liars. Therefore their video clips are labeled as positive samples with a label of 1. Similarly, the clips of innocent suspects are labeled by 0. To discriminate them, Random Forests [172] have been adopted as the binary classifier.

In this chapter, the theory of the Random Forests is briefly reviewed in section 5.1. Then the experimental procedure and the details of the training and testing processes are elaborated in section 5.2.

## 5.1. Random Forests

In this thesis, the binary classifier used for discriminating guilty and innocent suspects is the Random Forest [172]. A Random Forest (RF) is an ensemble algorithm designed for bagging decision trees. Due to its computational efficiency and simplicity of implementation, it has been used for classification, regression and clustering. The advantage of using an RF is that the variance is known to be reduced when bagging all decision trees and the generalization error will always converge, avoiding the occurrence of overfitting [172]. In this section, the methodology for constructing and using an RF for prediction is briefly reviewed.

### 5.1.1. Introduction to the RF Structure

A Random Forest is an ensemble of trees $\{tree_i\}_1^k$, as shown in Figure 31. Each tree in the forest consists of three kinds of nodes: root node (black circle in Figure 31), internal node (red circle in Figure 31) and leaf node (green circle in Figure 31). The root node and each internal node are associated with a nodal rule $r$ in the form of an inequality. For an instance, given $F$ features $\{\mathcal{F}_i\}_1^F$, the inequality usually involves comparing the $i-th$ feature $\mathcal{F}_i$ to a value $c$, that is, $\mathcal{F}_i \leq c$. For example, in Figure 31, the *root node* has a nodal rule of $\mathcal{F}_2 \leq 0.5$ which is equivalent to the question "Is the value of the second feature ($\mathcal{F}_2$) of the instance no greater than 0.5?" Different answers lead to different paths to the next node. Any path from the root to the leaf represents a classification rule, and the class label at the leaf node is the classification result.

The construction of a tree involves designing the tree structure and creating the nodal rules. To predict the class of a new instance using a tree involves finding a path from the root to the leaf following the nodal rules. The class label at the leaf node is the predicted label of the instance. Detailed algorithms regarding construction and prediction will be presented in the next two sections.

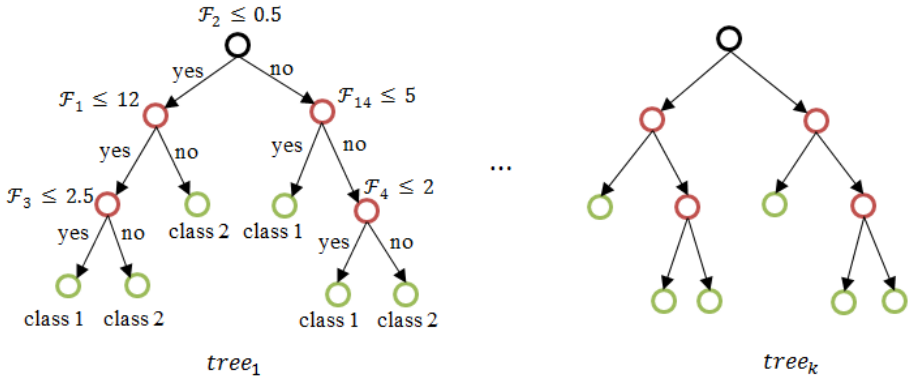



Figure 31. A sample Random Forest. The Random Forest is an ensemble of trees. Each tree has internal nodes (in red) and leaf nodes (in green). In $tree_1$, sample nodal rules at the internal nodes are presented. $\mathcal{F}_i$ represents the $i-th$ feature of an instance.

### 5.1.2. Algorithms for Constructing an RF

Constructing an RF involves growing $k$ trees using a training dataset $D$. A bootstrap replicate $D_i$ for constructing $tree_i$ is created by randomly selecting $\omega$ samples with replacement from $D$. Then $tree_i$ is grown from the root node and recursively split into two sub-nodes. At each node, the best splitting rule $r^*$ is chosen as the one that maximizes the "goodness-of-split" function $\mathbb{G}(r, D_i)$. The "goodness-of-split" function reflects the decrease of the "impurity" [173] of the data at the leaf nodes when a split is performed. $\mathbb{G}(r, D_i)$ has different forms based on different splitting criteria, and the Gini criterion [173] is used in this thesis.

The algorithm for constructing an RF is described as follows. The detailed pseudo-code can be found in Appendix II.

*Algorithm of constructing a Random Forest:*

1) For every $i \in \{1,2,\dots,k\}$, do the following:

   (a) Randomly select dataset $D_i$ of size $\omega$ from $D$ with replacement.

   (b) Grow $tree_i$ using $D_i$ by conducting the following steps recursively, until the number of instances at each leaf node of the tree is smaller than $d_{min}$.

      (i) Randomly select $f$ features $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_f\}$ without replacement from the $F$ features.

      (ii) Pick the best splitting rule $r^*$ which maximizes $\mathbb{G}(r, D_i)$.

      (iii) Partition $D_i$ according to rule $r^*$ into two subsets, and create a new node for each of them.

2) The ensemble of trees $\{tree_i\}_1^k$ is the constructed Random Forest.

### 5.1.3. Algorithm for Using an RF for Prediction

To predict the class of a new instance $\xi$, all the trees in the RF are independently applied to predict its class. As stated in section 5.1.1, the prediction process of each tree involves finding a path from the root to the leaf following the nodal rules. Then the *class label* $Class_i(\xi)$ at the leaf node is taken as the prediction by $tree_i$. Finally, the class that has the most votes from all of the trees is taken as the final prediction from the RF. The algorithm is described as follows.

*Algorithm for using the Random Forest to predict the class of a new instance $\xi$:*

1) For every $i \in \{1,2,\dots,k\}$, predict the class of $\xi$ using $tree_i$. Denote the resultant prediction as $Class_i(\xi)$.
2) The predicted class $Class^*(\xi)$ is the majority class of $\{Class_i(\xi)\}_1^k$.

## 5.2. Experimental Training Procedure

In this section, the experimental procedure for training the RF and using it for prediction is elaborated. Also, the strategy of choosing the best thresholds through cross-validation is also presented.

Suppose there are $N$ video clips in total in our database (Figure 32(a)). The N video clips are randomly partitioned into two sets: a training set of $N_{train}$ clips and a test set of $N_{test}$ clips, where $N_{train} = \frac{4N}{5}$ and $N_{test} = \frac{N}{5}$ (Figure 32(b)). The training set is used for training the RF classifier and the test set is used for testing the performance of the RF.

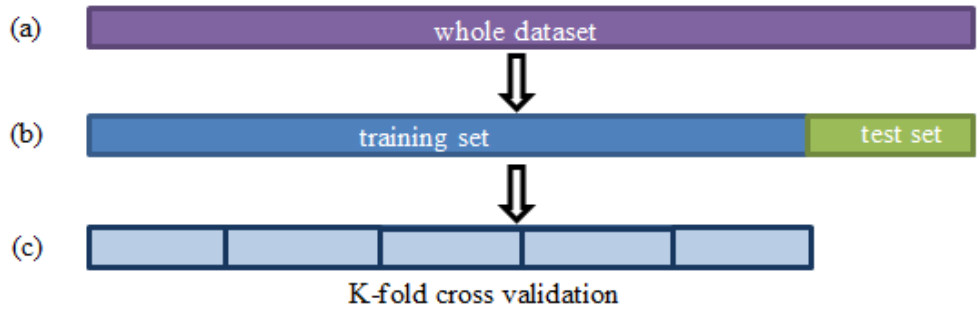



Figure 32. The general experimental procedure. (a) The whole dataset. (b) The whole dataset is partitioned into a training set and a test set. (c) K-fold cross-validation is performed on the training set to choose the best values for the three thresholds which maximizes the performance of the classifier.

As mentioned in the previous chapter, there are three undetermined thresholds for the eyebrow motion, wrinkle detection and mouth motion features. This gives a threshold set $\Theta = [\theta_{eyebrow}, \theta_{wrinkle}, \theta_{mouth}]$, which has $\gamma_{eyebrow} \times \gamma_{wrinkle} \times \gamma_{mouth}$ possible candidate values. Therefore the training set is firstly used for selecting the best threshold set $\Theta_{best}$ which maximizes the classifier performance using K-fold cross-validation (Figure 32(c)). After the best threshold set $\Theta_{best}$ has been chosen, its corresponding features are considered as the finalized features. Then the whole training set with the finalized features is used for training the final classifier. The cross-validation process is discussed in section 5.2.1.

### 5.2.1. K-fold Cross-validation

K-fold cross-validation is applied to the training set to pick the best thresholds. During the cross-validation, training set is both used for training and testing (self-validating). Specifically, for each threshold set $\Theta$, the training set of $N_{train}$

samples are randomly partitioned into $K$ subsets. The RF classifier is trained $K$ times, each time leaving out one subset for testing and using the complementary $K-1$ subsets for training. Then the test results of the $K$ experiments are averaged. After all the $\gamma_{eyebrow} \times \gamma_{wrinkle} \times \gamma_{mouth}$ possible threshold sets have been cross-validated, the threshold set $\Theta_{best}$ which maximizes the performance of the classifier[12] is chosen.

However, the choice of $K$ is still an open problem. Larger $k$ reduces the estimation bias by using more training data, but results in a smaller test set which may not give precise test results. Generally, a $K$ as small as 3 will be adequate for large datasets, while small datasets often require a larger $K$. Also, the computational time is often taken into consideration when choosing $K$ because it increases linearly as $K$ grows [174]. In this thesis we chose $K = 5$ as a compromise.

After the best threshold set $\Theta_{best}$ has been chosen, the features of the training data are frozen as indicated in Chapter 4. This provides us with the final RF classifier[13]. Finally the RF is tested on the $N_{test}$ video clips, which have not been used during training. The test results will be reported in the next chapter.

### 5.2.2. Details of Training and Testing

The previous section has introduced the experimental procedure in a general manner. In this section we will discuss the training and testing processes in detail.

When we refer to *training*, this implies performing training to learn a classifier. It involves both the training within the cross-validation and the training for learning the final classifier after the best thresholds have been chosen. When we say *testing* here, we mean every time the classifier is used to predict the label of a new sample. It involves both the validation step within the cross-validation and the testing after the final classifier has been trained.

[12] The criterion for evaluating the classifier performance will be presented in section 5.2.3.
[13] Referred to as a forest.

The details of training and testing are discussed below, respectively.

**Training:** Suppose there are $C$ video clips that are used for training. Each clip $v_\alpha$ has a class label (either 1 or 0) and an associated bag of FTVs (BOF) $\{FTV_{\alpha,1}, FTV_{\alpha,2}, \dots, FTV_{\alpha,m_\alpha}\}$. Each FTV is assigned the same class label as the clip and treated independently from the others. All of these FTVs are assembled to form a single training set, called the Big BOF, which has $\sum_{\alpha=1}^{C} m_\alpha$ FTV samples. For each case, the FTVs are assumed to be independently and identically distributed in the Big BOF, regardless of their originating video clips. Then each tree of the RF is trained by randomly sampling a subset of the FTVs inside the Big BOF. This process is demonstrated in Figure 33.

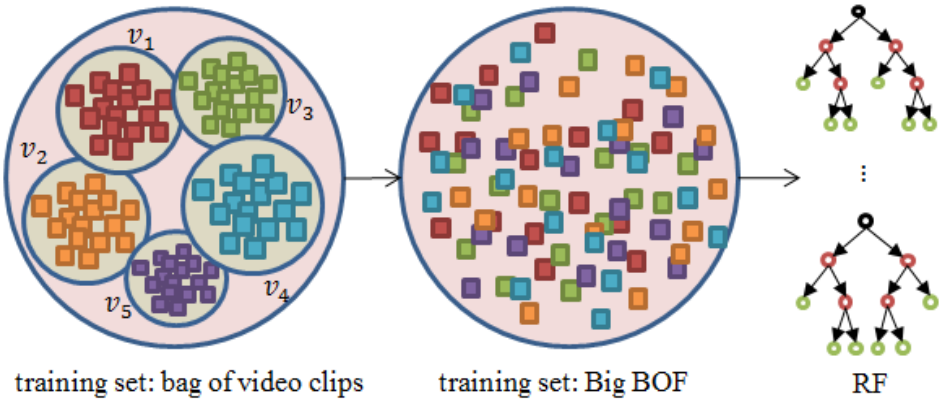



Figure 33. Details of the training process. The training set has $C$ video clips, each one associated with a BOF (left). Nevertheless, the FTVs are treated independently from each other as training samples and form a Big BOF (middle). Then the RF is constructed using the Big BOF by randomly sampling the FTVs and constructing the decision trees (right).

**Testing:** The goal of the testing process is to predict the class labels of the test video clips. Each clip $v_\alpha$ has $m_\alpha$ FTVs, and all the FTVs are supplied to the RF. Each FTV has a class label predicted by the RF, forming a bag of candidate class labels for this clip.

If the number of positive labels (1s) among the candidate labels is $pl_\alpha$, the label of the clip is voted by the $m_\alpha$ candidate labels according to a voting rule:

$$label(v_\alpha) = \begin{cases} 1, & if\ pl_\alpha > \mu m_\alpha \\ 0, & otherwise \end{cases} \quad (11)$$

where $\mu \in [0,1]$ indicates the percentage of FTVs who have voted for 1. The parameter $\mu$ is added to the threshold set $\Theta$ and chosen experimentally during the cross-validation process, by searching for a combination of $\Theta$ and $\mu$ that could give the best validation result. The testing process is demonstrated in Figure 34.

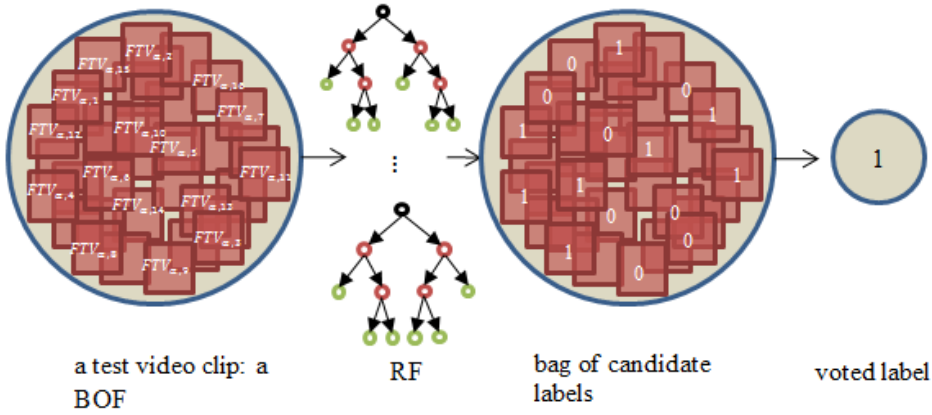



Figure 34. Details of the testing process. A test video clip can be represented by a BOF (left). Each FTV belonging to this clip is input into the RF ("the forest"), thereby producing a predicted class label. These labels are placed into a bag of candidate labels (middle). All candidate labels then vote for the final label of the clip (right).

### 5.2.3. Criterion for Evaluating the Classifier Performance

In the previous sections, we have mentioned that the best thresholds are chosen when the best performance of the classifier is achieved during the cross-validation process. We have adopted two criteria for defining the best performance. These two are applied independently and the results of both will be discussed in the next chapter.

Among the $N_{test}$ video clips in the test set, $n_p$ of them are clips containing deception (label is 1). For clips without deception, the number is $n_n$ (label is 0). After the prediction and voting step, each clip has a predicted label. As shown in

Figure 35, $n_{tp}$ of the $n_p$ clips are correctly predicted as 1 (true positive) while $n_{fn}$ of them are falsely predicted as 0 (false negative). Similarly, $n_{tn}$ is the number of clips correctly predicted as 0 (true negative) and $n_{fp}$ is for falsely classifying 0 as 1 (false positive).

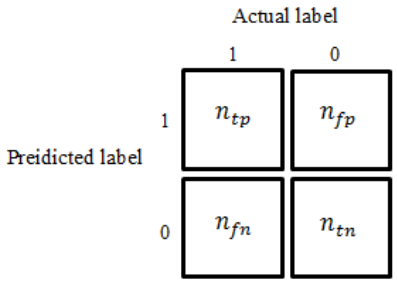



Figure 35. The test samples are categorized into true positive, false positive, true negative and false negative based on their actual and predicted labels. The number of samples belonging to each category is shown in the figure.

Thus the two criteria can be defined as follows.

The first criterion is the *accuracy* of spotting liar or truth-teller. It is defined as:

$$Accuracy = \frac{n_{tp} + n_{tn}}{N_{test}} \tag{12}$$

This measure indicates the percentage of clips whose class labels are correctly predicted. When using this criterion, the best threshold set, $\Theta_{best}$, is chosen as the one that gives the highest accuracy.

The second criterion is the *area under the receiver operating characteristic* (ROC) *curve*, termed AUC. The ROC curve captures the effect on the true positive rate (TPR) and the false positive rate (FPR) when the parameters vary. The TPR and FPR are computed as follows:

$$TPR = \frac{n_{tp}}{n_p}, \qquad FPR = \frac{n_{fp}}{n_n} \tag{13}$$

A sample ROC curve is shown in Figure 36. The area of the shaded part under the ROC curve is computed as the AUC. Based on this criterion, the best classifier performance is achieved when the AUC value is the largest.

In our case, the ROC curve is obtained by varying $\mu$ from 0 to 1 when the thresholds are fixed. During the cross-validation process, an AUC value could be computed for each threshold set $\Theta$, and $\Theta_{best}$ is chosen as whichever gives the largest AUC value.

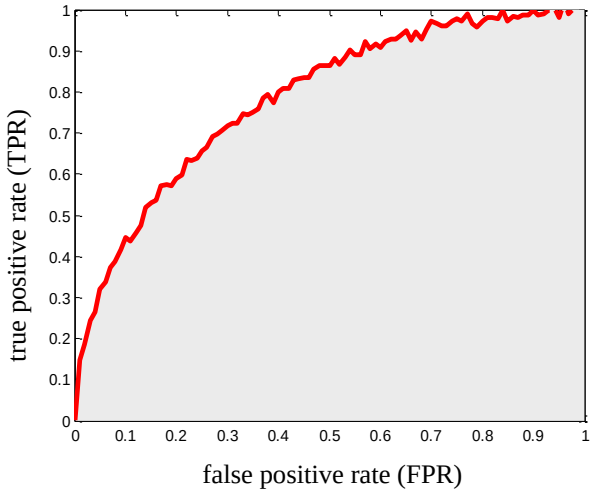



Figure 36. A sample ROC curve. The area of the shaded part is the AUC.

The results produced by the two criteria are both reported in the next chapter. Both of them could be used in actual applications, based on different requirements of the system. The accuracy criterion emphasizes more the precision of spotting liars and truth-tellers, while the AUC criterion gives an overall evaluation of the system. The ROC curve permits the user to be aware of the tradeoff between detecting a liar and detecting a truth-teller.

Specifically, if the risk of mis-detecting a liar is considered to be higher than that of mis-detecting a truth-teller, the user might need a higher TPR. In this case the ROC curve is a good reference to search for a balance between a high TPR and a tolerable FPR.

The next chapter will discuss the test results based on the two criteria.

# Chapter 6. Results and Discussion

In this chapter, the experimental results are discussed. Specifically, in section 6.1, the main test results are discussed. As introduced in Chapter 3, our database has various challenges: uncontrolled illumination, various head poses and facial occlusion. Although the first two have been largely compensated for during the feature analysis process in Chapter 4, the last one remains an unsolved problem in our deception detection task. Therefore in section 6.2, the effect of different facial occlusions on the test result is investigated. Finally, in section 6.3 we discuss whether micro-expressions are reliable evidence for spotting liars as the TV series *Lie to Me* implies, as opposed to what are normally called just expressions.

## 6.1. Main Results

As mentioned in the previous chapter, there are two criteria for evaluating the performance of the classifier: accuracy and AUC. The test results based on both criteria will be discussed here. But before that, the sensitivity of the test results to the FTV size is discussed. Then the best results will be presented at the end of this section.

In Chapter 4, we introduced the Feature Temporal Volumes (FTVs) as the elemental descriptor of a video clip. An FTV is computed from $T$ frames, as shown in Figure 29. When computing an FTV, the *temporal context* around each frame is taken into consideration in order to achieve the most compact and superior representation of the facial movements. The temporal size $T$ of an FTV implies the number of frames that should be considered as the duration of a "compact" expression.

In the next two subsections, the effect of $T$ on the test results is investigated based on the two criteria mentioned above.

### 6.1.1. Test Results Based on Accuracy

As presented in Chapter 5, the accuracy of the test results reflects the percentage of video clips that are correctly classified into deceptive and honest categories. Besides the accuracy, we have also computed the true positive rate (TPR) and true negative rate (TNR). Since the liars are considered as positive samples, the TPR reflects the precision of spotting them whereas the TNR reflects the precision of spotting truth-tellers.

Figure 37 shows the variations in accuracy as well as the TPR and the TNR as functions of $T$. (Detailed results are listed in the table in Appendix III)

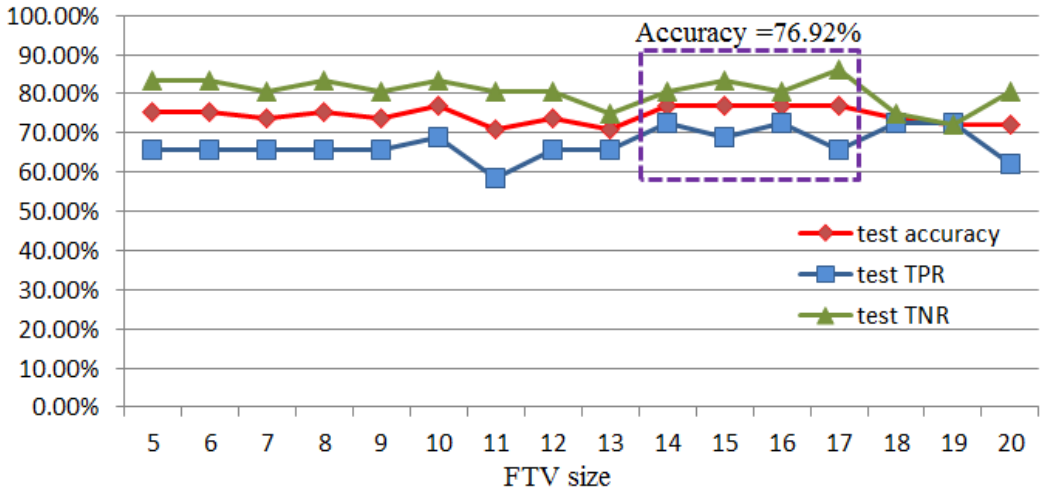



Figure 37. The change of accuracy, TPR and TNR as the FTV size varies. The results inside the dashed rectangle are the best results: accuracy=76.92%.

Several conclusions could be drawn from this figure:

1) The accuracy has a mean of 74.52% with a variance of 0.04%, implying that the FTV size does not significantly influence the test accuracy.
2) The best accuracy 76.92% is achieved when $T = 14{\sim}17$. Note that the frame rate of our videos is 30 frames per second. Therefore the ideal $T$ is approximately half a second, which implies a reasonable duration of a

*compact* facial expression. We can infer that most of the discriminating facial cues to deception are within half a second.

3) In general, the TNR is higher than the TPR, which means our system does a better job of spotting truth-tellers than catching liars. This phenomenon might be attributed to the process for assigning ground truth labels to each video clip. In Chapter 5, when a suspect has multiple video clips, every clip was assigned to the same label as the suspect. In other words, every clip of a guilty suspect was labeled as 1 while that of an innocent suspect was labeled as 0. Indeed, if a suspect is honest, all of his clips are definitely deception-free. But this is not true for deceptive suspects. If one suspect is deceptive, we could only say that at least one clip belonging to him is deceptive. It is very likely that some of his clips might not contain deception. Therefore there might be some deception-free samples, which are mislabeled as deception samples, resulting in a lower TPR.

It might seem that we could determine the *frame-by-frame* facial action codes of all the video clips to obtain the ground truth of when a deceptive expression has occurred. However, this would be contrary to the model we have used for classifying the facial expressions, are based on time intervals (using the FTVs) .

### 6.1.2. Test Results Based on AUC

The test AUC value is also a function of $T$, as shown in Figure 38.

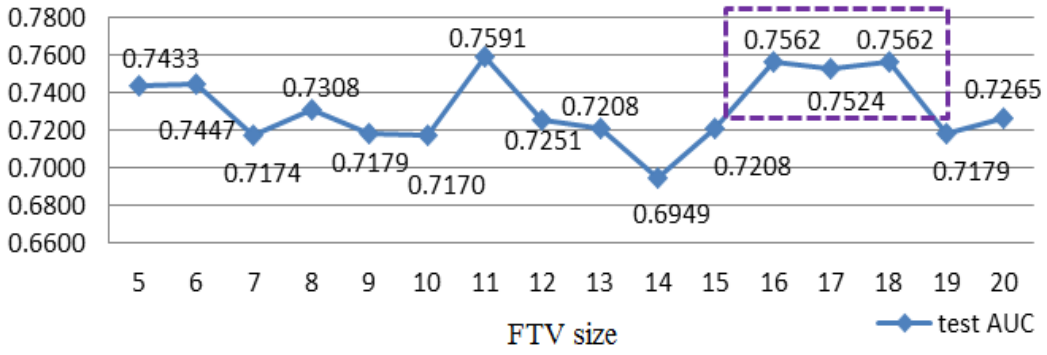



Figure 38. The change of AUC as the FTV size varies. The AUC values in the dashed rectangle are the best AUC values: 0.7562.

Similar conclusions as in section 6.1.1 could be drawn from the figure:

1) The mean AUC value is 0.73, and the variance is $3.2 \times 10^{-4}$. The conclusion is similar to the first conclusion in the previous section; the AUC value is not very sensitive to the change of $T$.
2) The best AUC=0.7562 is obtained when $T = 16, 18$, which is also approximately half a second. This reconfirms that $\frac{1}{2}s$ is a reasonable duration of a *compact* facial expression.

### 6.1.3. Discussion of the Best Results

From section 6.1.1 we can see that the best accuracy is 76.92%. This outperforms all of the results obtained by human observers, as discussed in section 2.2. The only exception that has been stated in the literature was obtained by mental professionals after a comprehensive training workshop [18]. This study reported an accuracy of 80.9%. Nonetheless, we note that since our test result was obtained under very challenging conditions, it is reasonable to believe that our system would perform better in an actual interrogation situation in a more controlled environmental setting.

However, it is feasible for a user of the system reported in this thesis to customize the tradeoff between the precision of spotting a truth-teller and that of catching a liar by using the ROC curve in section 6.1.2. In section 6.1.2, the highest AUC value is 0.7562, and the corresponding ROC curve is shown in Figure 39. The ROC curve could be used for searching for a balance between the TPR and the FPR. For example, if the user of the system demands an 80% liar-spotting precision, the risk of wrongly classifying an honest suspect into a deceptive one is also as high as 60%. Therefore, if the user wants to achieve a desirable deception detection performance, he can integrate the result of our system with that of other lie detectors.

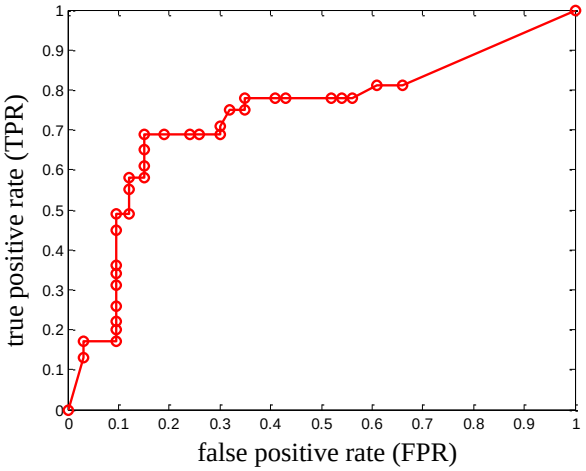



Figure 39. ROC curve when highest AUC=0.7562 is achieved in the previous section. Note that the points at (0,0) and (1,1) on the curve were not obtained from the experiment. They were added in order to complete the ROC and compute the AUC value.

## 6.2. Effect of Facial Occlusions on Deception Detection Results

As discussed in Chapter 3, most of the existing facial expression databases have excluded facial occlusions during the data collection process, in order to provide a relatively easy scenario for automated expression analysis. Therefore in the current literature, there is only a limited amount of research that has discussed the effect of facial occlusions on facial expression analysis. In [175], the researchers have investigated on how the facial expression classification was affected by partial occlusions on different facial regions. In [176], the authors have shown that their facial expression recognition methods were robust to facial occlusions.

Figure 3 has shown that facial occlusion is very common in our database. In fact, facial occlusion should be considered an inevitable factor in an actual interrogation scenario. The interviewee might feel uncomfortable or irritated when instructed to remove any facial occlusions, leading to unnatural facial expressions. Therefore, the effect of facial occlusions on the deception detection task will be examined in this section.

The major facial occlusions in our database are: glasses, mustache and facial hair. In this thesis, facial hair is considered to be the fringe covering the forehead. Examples with respect to these three facial occlusions can be found in Figure 40. To investigate the effect of facial occlusions on our algorithms, we have designed five scenarios shown in Table 15. These are all based on the same experimental procedure described in Chapter 4, but used different data in the database.

(a)

(b)

(c)

Figure 40. Sample frames that have facial occlusions in our database. (a) Glasses (b) Mustache (c) Facial hair

Table 15. Five facial occlusion scenarios

| | ***Scenario*** | ***Experimental data in the scenario*** |
|---|---|---|
| **1** | All data | Whole database including all kinds of facial occlusions |
| **2** | No glasses | Database after removing suspects that wear glasses |
| **3** | No mustache | Database after removing suspects that have a mustache |
| **4** | No facial hair | Database after removing suspects that have facial hair |
| **5** | No occlusion | Database after removing suspects that have at least one of the three following facial occlusions: glasses, mustache, facial hair |

The test results of the five scenarios are shown in Table 16.

Table 16. Test results of five facial occlusion scenarios

| | *Scenario* | *Accuracy* | *TPR* | *TNR* |
|---|---|---|---|---|
| **1** | All data | 76.92% | 72.41% | 80.56% |
| **2** | No glasses | 75.44% | 82.14% | 68.97% |
| **3** | No mustache | 73.08% | 76.67% | 68.18% |
| **4** | No facial hair | 87.04% | 93.55% | 78.26% |
| **5** | No occlusion | 75.00% | 78.26% | 70.59% |

Several conclusions can be drawn from this table:

1) From this table we observe that removing data with facial hair occlusion has increased the test accuracy from 76.92% to 87.04%, indicating that facial hair is the major obstacle to detecting deception using the proposed method. Also, the increase in accuracy largely attributed to the increase in the TPR.

   In our database, facial hair can occlude the forehead and sometimes even the eyebrows, which could affect as many as six facial regions: $R_3, R_4, R_5, R_7, R_8$ and $R_9$, as shown in Figure 41. Based on the feature integration strategy discussed in section 4.5.1, these six facial regions are unfortunately related to 16 out of the 21 features. Moreover, the feature extraction methods applied to eyebrow motion and wrinkle detection, respectively rely on the assumption that the eyebrows and forehead are visible. Since liars are more likely to express *surprise* as discussed in Chapter 4, they will exhibit more appearance changes on the forehead due to the raise of the eyebrows. Therefore it is very likely that the facial hair occlusion has impeded the detection of clues to deception across the forehead of a liar, resulting in a lower TPR.

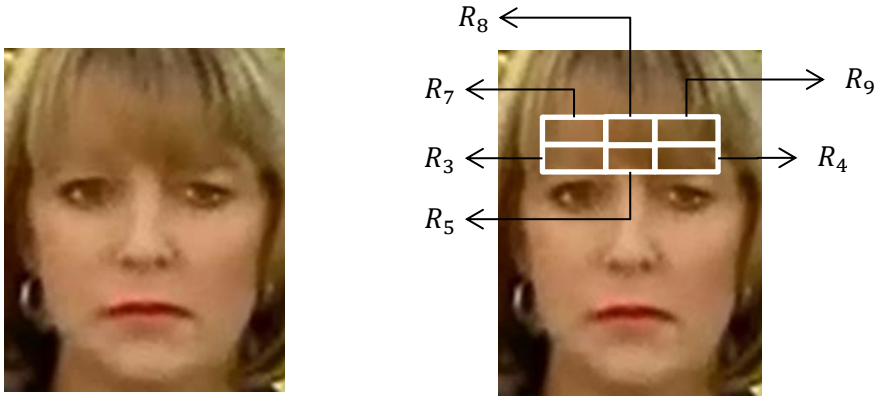



Figure 41. The influence of facial hair on facial regions. (a) A sample frame of a woman who has hair covering her forehead and eyebrows. (b) The six facial regions that are influenced by the hair occlusion.

2) The test results when excluding glasses or mustaches is very close to that of all the data, implying that eye blink and mouth motion detection (introduced in section 4.4.1) are not sensitive to facial occlusions.

In summary, facial hair occlusion needs to be considered when our system is applied to actual interrogation scenarios. Since it might be inappropriate to ask all the interviewees to expose their foreheads, other techniques with respect to solving this problem should be investigated in the future.

## 6.3. Micro-expression vs. Macro-expression as a Facial Clue to Deception

Micro-expressions, a rapid and involuntary facial expression which lasts only 1/25th to 1/5th of a second [95], has been introduced in section 4.1. Macro-expressions usually last longer than micro-expressions. In the TV series *Lie to Me*, micro-expressions play an important role in revealing deception during an interrogation. However, Porter has claimed that most of the observed so-called *emotional leakages* last too long to be classified as micro-expressions [52].

The existing computer vision literature has rarely reported on research on micro-expression detection or classification. Both [85] and [177] have addressed the problem of discriminating micro-expressions from emotionless samples. In [85], micro-expressions were further classified into positive (happiness) and negative (disgust, fear, surprise and sadness) ones. In [86], both macro- and micro-expressions in video sequences were detected. However, to the best knowledge of the author, no computer vision research has been conducted to study the difference between macro- and micro-expressions as indicators of deception. Therefore, this section is focused on investigating whether macro- or micro-expressions can be relied upon for a deception detection task.

In section 4.2, we describe different facial expressions as *events* that are detected in nine facial regions. Each event has a duration, which implies how long a facial expression lasts. If the duration of an event is shorter than 1/5th of a second, we classify it as a micro-expression; otherwise it is a macro-expression.

In this section, three scenarios are considered based on the facial expressions that are included in training and testing, shown in Table 17. Specifically, in the *micro* scenario, only the detected facial events which are shorter than $\frac{1}{5}s$ are used for computing the FTVs. Then the classifier is trained and tested on the resulting FTVs. Similarly, in the *macro* scenario, only the facial events which are longer than $\frac{1}{5}s$ are retained for computing the FTVs. In the *all expressions* scenario, all the *events* have been retained. The test results for the three scenarios are shown in Table 18.

Table 17. Three facial expression scenarios

| | ***Scenario*** | ***Facial expressions included in the scenario*** |
|---|---|---|
| **1** | All expressions | Both macro and micro-expressions are included |
| **2** | Macro | Only macro-expressions are included |
| **3** | Micro | Only micro-expressions are included |

Table 18. Test results of three facial expression scenarios

| | ***Scenario*** | ***Accuracy*** | ***TPR*** | ***TNR*** |
|---|---|---|---|---|
| **1** | All expressions | 76.92% | 68.97% | 83.33% |
| **2** | Macro | 73.85% | 62.07% | 83.33% |
| **3** | Micro | 56.92% | 51.72% | 61.11% |

We observe the followings from Table 18:

1) Neither using macro-expressions nor micro-expressions alone outperformed the results when both of them were used. This implies that the combination of the two improves the ability to discriminate liars and truth-tellers.
2) The accuracy based on *macro*-expressions alone is very close to that of all expressions, and the TNR is the same as that of all expressions. However, the test accuracies, TPR and TNR are all very low when using micro-expressions alone. This implies that macro-expressions are more reliable than micro-expressions when used as cues of deception. This conclusion is consistent with Porter's observation that most of the emotional leakages that are indicators of deception last longer than a micro-expression [52].

In summary, accurate deception detection tasks require both micro- and macro-expressions, and the detection accuracy of using micro-expressions alone is only slightly better than chance. This conclusion is consistent with Porter's argument in [52], but contradicts the popular belief presented by the TV series *Lie to Me*. Nonetheless, the essential idea of *Lie to Me* and the psychological basis of this thesis are the same: deception can be detected through involuntary facial movements. Therefore, does *Lie to Me* lie to you? To some degree it does not, as it has advocated for using facial clues for catching liars. But does it lie about micro-expressions alone being adequate for accomplishing the lie detection task? Our experiments imply that it does!

# Chapter 7. Conclusion

As a first attempt in the literature at proving the validity of facial clues to deception detection in unconstrained environments, the results in this thesis are very promising. We achieved a 76.92% accuracy for spotting liars and truth-tellers in high-stakes situations. Considering that the database we used is rather challenging, it is possible that better results might be obtained from data collected in an actual criminal interrogation scenario, where the environmental setting is much more controlled. Thus, we have successfully proved the fundamental theory behind *Lie to Me*, that is, facial clues are reliable deception indicators in high-stakes situations.

Additionally, we note that the 76.92% accuracy was obtained by detecting both "normal" facial expressions plus micro-expressions. With regard to solely the latter, our results have also challenged the popular belief expounded in the TV show *Lie to Me*, that *micro-expressions alone are sufficient* for lie detection. Our results have shown that the accuracy of using micro-expressions alone for deception detection is only slightly better than chance. The deceit detection system performs best when both micro- and macro-expressions are used.

The investigation of the effect of facial occlusions on the system performance has shown that facial hair is the major obstacle. However, is it possible to ask all suspects to expose their entire forehead? This problem has to be further studied and paid more attention to in the future study. Also, the head pose compensation method used in the thesis is preliminary. Precise face registration technique might be beneficial to more accurate facial behavioral recognition. Furthermore, the size of our database is relatively small. In order to build a more convincing system, more similar criminal cases should be added to our database.

Finally, as suggested by many psychologists [51, 92, 178], there are some pitfalls when designing a deception detection system. First, only the external facial expressions can be detected; we can never know the actual internal emotion that triggers the expression. Second, the fact that concealed emotions are not always

associated with the presence of deception highlights the necessity of taking the context into consideration. Third, the absence of emotional leakage does not mean the suspect is innocent; further questioning needs to be conducted. Finally, Porter also suggested integrating verbal, non-verbal, and facial cues to achieve higher accuracy [3]. These pitfalls and suggestions should always be taken into consideration when designing a deception detection system in the future.

# References


[1] P. Ekman, Telling Lies : Clues to Deceit in the Marketplace, Politics, and Marriage. New York: W.W. Norton, 1992.

[2] S. Porter, L. ten Brinke, and B. Wallace, "Secrets and Lies: Involuntary Leakage in Deceptive Facial Expressions as a Function of Emotional Intensity," Journal of Noverbal Behavior, vol. 36, pp. 23-37, 2012.

[3] S. Porter and L. ten Brinke, "The Truth about Lies: What Works in Detecting High-Stakes Deception?," Legal and Criminological Psychology, vol. 15, pp. 57-75, 2010.

[4] T. R. Levine, K. B. Serota, and H. C. Shulman, "The Impact of Lie to Me on Viewers' Actual Ability to Detect Deception," Communication Research, vol. 37, pp. 847-856, 2010.

[5] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein, "Lying in Everyday Life," Journal of Personality and Social Psychology, vol. 70, pp. 979-995, 1996.

[6] J. T. Hancock, J. Thom-Santelli, and T. Ritchie, "Deception and Design: the Impact of Communication Technologies on Lying Behavior," in Conference on Computer Human Interaction, New York, 2004, pp. 130-136.

[7] J. T. Hancock, "Digital Deception: When, Where and How People Lie Online," in Oxford Handbook of Internet Psychology, K. McKenna, T. Postmes, U. Reips, and A. N. Joinson, Eds., ed Oxford: Oxford University Press, 2007, pp. 287-301.

[8] L. ten Brinke and S. Porter, "Discovering Deceit: Applying Laboratory and Field Research in the Search for Truthful and Deceptive Behaviour," in Applied Issues in Investigative Interviewing, Eyewitness Memory, and Credibility Assessment, B. S. e. a. Cooper, Ed., ed New York: Springer Science and Business Media, 2013.

[9] P. J. Wilson, "Wrongful Conviction: Lessons Learned from the Sophonow Public Inquiry," Canadian Police College, 2003.

[10] (2010). Interview: Rebecca Sophonow. Available: http://www.cbc.ca/fifth/2009-2010/the_wrong_man/rebecca_sophonow.html. [cited July 10, 2013]

[11] A. Vrij, Detecting Lies and Deceit: the Psychology of Lying and the Implications for Professional Practice. Chichester, England: Wiley, 2000.

[12] G. D. R. Team, "A World of Lies," Journal of Cross-cultural Psychology, vol. 37, pp. 60-74, 2006.

[13] A. Baker, L. ten Brinke, and S. Porter, "Will Get Fooled Again: Emotionally Intelligent People Are Easily Duped by High-stakes Deceivers," Legal and Criminological Psychology, 2012.

[14] K. Ask and P. A. Granhag, "Motivational Bias in Criminal Investigators' Judgements of Witness Reliability," Journal of Applied Social Psychology, vol. 37, pp. 561-591, 2007.

[15] S. Porter, S. McCabe, M. Woodworth, and K. A. Peace, ""Genius is 1% inspiration and 99% perspiration"...or is it? An Investigation of the Effects of Motivation and Feedback on Deception Detection," Legal and Criminological Psychology, vol. 12, pp. 297-309, 2007.

[16] C. F. J. Bond and B. M. Depaulo, "Accuracy of Deception Judgments," Personality and Social Psychology Review, vol. 10, pp. 214-234, 2006.

[17] M. Hartwig, P. A. Granhag, L. A. Stromwall, A. G. Wolf, A. Vrij, and E. R. a. Hjelmsater, "Detecting Deception in Suspects: Verbal Cues as a Function of Interview Strategy," Psychology, Crime and Law, vol. 17, pp. 643-656, 2011.

[18] M. O'Sullivan and P. Ekman, "The Wizards of Deception Detection," in Deception Detection in Forensic Contexts, P. A. Granhag and L. A. Stromwall, Eds., ed Cambridge, England: Cambridge University Press, 2004, pp. 269-286.

[19] A. Vrij, S. Mann, E. Robbins, and M. Robinson, "Police Officers Ability to Detect Deception in High Stakes Situations and in Repeated Lie Detection Tests," Applied Cognitive Psychology, vol. 20, pp. 741-755, 2006.

[20] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," Journal of Nonverbal Behavior, vol. 33, pp. 59-69, 2009.

[21] A. Vrij, P. A. Granhag, and S. Porter, "Pitfalls and Opportunities in Noverbal and Verbal Lie Detection," Psychological Science in the Public Interest, vol. 11, pp. 89-121, 2010.

[22] S. Porter, M. Woodworth, and A. R. Birt, "Truth, Lies, and Videotape: An Investigation of the Ability of Federal Parole Officers to Detect Deception," Law and Human Behavior, vol. 24, pp. 643-658, 2000.

[23] S. Porter, M. Juodis, L. ten Brinke, R. Klein, and K. Wilson, "Evaluation of a Brief Deception Detection Training Program," Journal of Forensic Psychiatry and Psychology, vol. 21, pp. 66-76, 2010.

[24] L. Warmelink, A. Vrij, S. Mann, S. Leal, D. Forrester, and R. P. Fisher, "Thermal Imaging as a Lie Detection Tool at Airports," Law and Human Behavior, vol. 35, pp. 40-48, 2011.

[25] M. Gamer, H.-G. Rill, G. Vossel, and H. W. Godert, "Psychophysiological and Vocalmeasures in the Detection of Guilty Knowledge," International Journal of Psychophysiology, vol. 60, pp. 76-87, 2006.

[26] W. G. Iacono, "Effective Policing: Understanding How Polygraph Tests Work and Are Used," Criminal Justice and Behavior, vol. 35, pp. 1295-1309, 2008.

[27] P. C. Stern, The Polygraph and Lie Detection. Report of the National Research Council Committee to Review the Scientific Evidence on the Polygraph. Washington, DC: The National Academies Press, 2004.

[28] V. Hughes, "Science in Court: Head Case," Nature, vol. 464, pp. 340-342, 2010.

[29] S. E. Christ, D. C. Essen, J. M. Watson, L. E. Brubaker, and K. B. McDermott, "The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence from Activation Likelihood Estimate Meta-analysis," Cerebral Cortex, vol. 19, pp. 1557-1566, 2009.

[30] G. Ganis, J. P. Rosenfeld, J. Meixner, R. A. Kievit, and H. E. Schendan, "Lying in the Scanner: Covert Coutermeasures Disrupt Deception Detection by Functional Magnetic Resonance Imaging," NeuroImage, vol. 55, pp. 312-319, 2010.

[31] S. A. Spence, "Playing Devil's Advocate: the Case against fMRI Lie Detection," Legal and Criminological Psychology, vol. 13, pp. 11-25, 2008.

[32] D. D. Langleben, "Detection of Deception with fMRI: Are We There Yet? ," Legal and Criminological Psychology, vol. 13, pp. 1-9, 2008.

[33] S. A. Spence, A. Hope-Urwin, S. T. Lankappa, J. Woodhead, J. C. L. Burgess, and A. V. Mackay, "If Brain Scans Really Detected Deception, Who Would Volunteer to be Scanned?," Journal of Forensic Sciences, vol. 55, pp. 1352-1355, 2010.

[34] B. Verschuere, A. Spruyt, E. H. Meijer, and H. Otgaar, "The Ease of Lying," Consciousness and Cognition, vol. 20, pp. 908-911, 2011.

[35] M. G. Boltz, R. L. Dyer, and A. R. Miller, "Are You Lying to Me? Temporal Cues for Deception," Journal of Language and Social Psychology, vol. 29, pp. 458-466, 2010.

[36] G. L. J. Lancaster, A. Vrij, L. Hope, and B. Waller, "Sorting the Liars from the Truth Tellers: the Benefits of Asking Unanticipated Questions on Lie Detection," Applied Cognitive Psychology, vol. 27, pp. 107-114, 2013.

[37] S. Mann, A. Vrij, D. J. Shaw, S. Leal, S. Ewens, J. Hillman, et al., "Two Heads are Better Than One? How to Effectively Use Two Interviewers to Elicit Cues to Deception," Legal and Criminological Psychology, 2012.

[38] A. Vrij, P. A. Granhag, S. Mann, and S. Leal, "Outsmarting the Liars: Toward a Cognitive Lie Detection Approach," Current Directions in Psychological Science, vol. 20, pp. 28-32, 2011.

[39] A. Vrij, R. P. Fisher, S. Mann, and S. Leal, "A Cognitive Load Approach to Lie Detection," Journal of Investigative Psychology and Offender Profiling, vol. 5, pp. 39-43, 2008.

[40] A. Vrij, S. Mann, S. Leal, and R. P. Fisher, ""Look into My Eyes": Can An Instruction to Maintain Eye Contact Facilitate Lie Detection?," Psychology, Crime and Law, vol. 16, pp. 327-348, 2010.

[41] D. A. Leins, R. P. Fisher, and A. Vrij, "Drawing on Liars' Lack of Cognitive Flexibility: Detecting Deception through Varying Report Modes," Applied Cognitive Psychology, vol. 26, pp. 601-607, 2012.

[42] A. Vrij, S. Mann, S. Leal, and R. P. Fisher, "Is Anyone There? Drawings as a Tool to Detect Deceit in Occupation Interviews," Psychology, Crime and Law, vol. 18, pp. 377-388, 2012.

[43] A. Vrij, S. Mann, S. Jundi, L. Hope, and S. Leal, "Can I Take Your Picture? Undercover Interviewing to Detect Deception," Psychology, Public Policy, and Law, vol. 18, pp. 231-244, 2012.

[44] E. Reynolds and R.-S. Johanna, "Cues to Deception in Context: Response Latency/Gaps in Denials and Blame Shifting," British Journal of Social Psychology, vol. 50, pp. 431-449, 2011.

[45] L. Caso, A. Vrij, S. Mann, and G. D. Leo, "Deceptive Responses: the Impact of Verbal and Non-verbal Countermeasures," Legal and Criminological Psychology, vol. 11, pp. 99-111, 2006.

[46] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, and J. F. Nunamaker, "Automatic Extraction of Deceptive Bahavioral Cues from Video," Integrated Series in Information Systems, vol. 18, pp. 495-516, 2008.

[47] S. Lu, G. Tsechpenakis, D. N. Metaxas, M. L. Jensen, and J. Kruse, "Blob Analysis of the Head and Hands: A Method for Deception Detection," in Hawaii International Conference on System Sciences, Hawaii, 2005.

[48] G. Tsechpenakis, D. N. Metaxas, M. Adkins, J. Kruse, J. K. Burgoon, M. L. Jensen, et al., "HMM-Based Deception Recognition from Visual Cues," in International Conference on Multimedia and Expo, 2005, pp. 824-827.

[49] J. K. Burgoon, J. P. Blair, and E. Moyer, "Effects of Communication Modality on Arousal, Cognitive Complexity, Behavioral Control and Deception Detection During Deceptive Episodes," presented at the Annual Meeting of the National Communication Association, Miami Beach, Florida, 2003.

[50] J. K. Burgoon, J. Blair, T. Qin, and J. F. Nunamaker Jr, "Detecting Deception through Linguistic Analysis," in Intelligence and Security Informatics, ed: Springer, 2003, pp. 91-101.

[51] P. Ekman, "Darwin, Deception, and Facial Expressions," Annals of the New York Academy of Sciences, vol. 1000, pp. 205-221, 2003.

[52] L. ten Brinke and S. Porter, "Cry Me a River: Identifying the Behavioural Consequences of Extremely High-Stakes Interpersonal Deception," Law and Human Behavior, 2011.

[53] N. Bhaskaran, I. Nwogu, M. G. Frank, and V. Govindaraju, "Lie To Me: Deceit Detection via Online Behavioral Learning," presented at the International Conference on Automatic Face and Gesture Recognition and Workshops, 2011.

[54] N. Bhaskaran, I. Nwogu, M. G. Frank, and V. Govindaraju, "Deceit Detection via Online Behavioral Learning," in ACM Symposium on Applied Computing, New York, 2011, pp. 29-30.

[55] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to Deception," Psychological Bulletin, vol. 129, pp. 74-118, 2003.

[56] S. Mann, A. Vrij, E. Nasholm, L. Warmelink, S. Leal, and D. Forrester, "The Direction of Deception: Neuro-linguistic Programming as a Lie Detection Tool," Journal of Police and Criminal Psychology, vol. 27, pp. 160-166, 2012.

[57] S. Mann and R. Bull, "Suspects, Lies, and Videotape: An Analysis of Authentic High-stake Liars," Law and Human Behavior, vol. 26, pp. 365-376, 2004.

[58] T. E. Slowe and V. Govindaraju, "Automatic Deceit Indication through Reliable Facial Expressions," presented at the IEEE Workshop on Automatic Identification Advanced Technologies, 2007.

[59] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju, "Real-time Automatic Deceit Detection from Involuntary Facial Expression," in International Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-6.

[60] N. Micheal, M. Dilsizian, D. N. Metaxas, and J. K. Burgoon, "Motion Profiles for Deception Detection Using Visual Cues," in European Conference on Computer Vision, 2010, pp. 462-475.

[61] J. Rothwell, Z. Bandar, J. O'Shea, and D. McLean, "Silent talker: a new computer‐based system for the analysis of facial cues to deception," Applied cognitive psychology, vol. 20, pp. 757-777, 2006.

[62] S. A. Spence, C. J. Kaylor-Hughes, M. L. Brook, S. T. Lankappa, and W. I. D., "'Munchausen's Syndrome by Proxy' or a 'Miscarriage of Justice'? An Initial Application of functional Neuroimaging to the Question of Guilt versus Innocence.," Eur Psychiatry, vol. 23, pp. 309-314, 2008.

[63] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," Journal of Multimedia, vol. 1, pp. 22-35, 2006.

[64] S. Porter and L. ten Brinke, "Reading Between the Lies: Identifying Concealed and Falsified Emotions in Universal Facial Expressions," Psychological Science, vol. 19, pp. 508-514, 2008.

[65] S. Porter, N. L. Doucette, M. Woodworth, J. Earle, and B. MacNeil, "Halfe the world knowes not how the other halfe lies: Investigation of verbal and non‐verbal signs of deception exhibited by criminal offenders and non‐offenders," Legal and Criminological Psychology, vol. 13, pp. 27-38, 2008.

[66] S. Porter, L. ten Brinke, A. Baker, and B. Wallace, "Would I Lie to You? "Leakage" in Deceptive Facial Expressions Relates to Psychopathy and Emotional Intelligence," Personality and Individual Differences, vol. 51, pp. 133-137, 2011.

[67] C. M. Hurley and M. G. Frank, "Executing Facial Control During Deception Situations," Journal of Noverbal Behavior, vol. 35, pp. 119-131, 2011.

[68] L. ten Brinke, S. Porter, and A. Baker, "Darwin the Detective: Observable Facial Muscle Contractions Reveal Emotional High-Stakes Lies," Evolution and Human Behavior, 2012.

[69] M. G. Frank and P. Ekman, "Appearing Truthful Generalizes Across Different Deception Situations," Journal of Personality and Social Psychology, vol. 86, p. 486, 2004.

[70] S. Lucey, A. B. Ashraf, and J. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," Face Recognition, pp. 275-286, 2007.

[71] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska Directed Emotional Faces-KDEF. CD-ROM from Department of Clinical Neuroscience, Psychology section," ed. Stockholm, Sweden: Karolinska Institutet, 1998.

[72] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, 1998, pp. 200-205.

[73] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, 2002, pp. 46-51.

[74] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, et al., "A video database of moving faces and people," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, pp. 812-816, 2005.

[75] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, 2005, p. 5 pp.

[76] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in Proc. Int'l Conf. Language Resources and Evaluation, W'shop on EMOTION, 2010, pp. 65-70.

[77] F. Wallhoff, "Facial expressions and emotion database," Technische Universität München, 2006.

[78] P. Hancock. (2008). The psychological image collection at stirling (pics). Available: http://pics.psych.stir.ac.uk/ [cited July 10, 2013]

[79] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, et al., "The CAS-PEAL large-scale Chinese face database and baseline evaluations," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 38, pp. 149-161, 2008.

[80] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 94-101.

[81] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in Multimedia and Expo (ICME), 2010 IEEE International Conference on, 2010, pp. 1079-1084.

[82] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the Radboud Faces Database," Cognition and Emotion, vol. 24, pp. 1377-1388, 2010.

[83] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, 2011, pp. 2106-2112.

[84] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted Facial Expressions In The Wild Database," Technical Report TR-CS-11-02, Australian National University2011.

[85] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in Computer Vision (ICCV), 2011 IEEE International Conference on, 2011, pp. 1449-1456.

[86] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, 2011, pp. 51-56.

[87] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME Database: a dataset of spontaneous micro-expressions collected from neutralized faces," in 10th IEEE conference on automatic face and gesture recognition, Shanghai, 2013.

[88] C. Darwin, The Expression of the Emotions in Man and Animals. London,: J. Murray, 1872.

[89] J. Cornet. (2005). Frame Rate Conversion with motion estimation. Available: http://jcornet.free.fr/linux/yuvmotionfps.html [cited July 10, 2013]

[90] "PittPatt SDK," 5.2.2 ed: Pittsburgh Pattern Recognition, 2011.

[91] G. B. Duchenne de Boulogne, The Mechanism of Human Facial Expressions. New York: Cambridge University Press, 1990.

[92] L. ten Brinke, S. MacDonald, S. Porter, and B. O'Connor, "Crocodile Tears: Facial, Verbal and Body Language Behaviours Associated with Genuine and Fabricated Remorse," Law and Human Behavior, 2011.

[93] S. Porter, N. Korva, and A. Baker, "Secrets of the Human Face: New Insights Into the Face and Covert Emotions," Psychology Aotearoa, 2012.

[94] E. A. Haggard and K. S. Isaacs, "Micromomentary Facial Expressions as Indicators of Ego Mechanisms in Psychotherapy," in Methods of Research in

Psychotherapy, L. A. Gottschalk and A. H. Auerback, Eds., ed New York: Appleton Century Crofts, 1996, pp. 154-165.

[95] P. Ekman, E. Rolls, D. Perrett, and H. Ellis, "Facial expressions of emotion: An old controversy and new findings [and discussion]," Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, vol. 335, pp. 63-69, 1992.

[96] P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press, 1978.

[97] P. Ekman, G. Roper, and J. C. Hager, "Deliberate Facial Movement," Child Development, vol. 51, pp. 886-891, 1980.

[98] W. Friesen and P. Ekman, "EMFACS-7: Emotional Facial Action Coding System," 1983.

[99] S. Leal and A. Vrij, "Blinking During and After Lying," Journal of Nonverbal Behavior, vol. 32, pp. 187-194, 2008.

[100] S. Leal and A. Vrij, "The Occurence of Eye Blinks During a Guilty Knowledge Test," Psychology, Crime and Law, vol. 16, pp. 349-357, 2010.

[101] S. Shan, W. Gao, B. Cao, and D. Zhao, "Illumination normalization for robust face recognition against varying lighting conditions," in Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on, 2003, pp. 157-164.

[102] X. Xie and K.-M. Lam, "Face recognition under varying illumination based on a 2D face shape model," Pattern Recognition, vol. 38, pp. 221-230, 2005.

[103] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, et al., "Adaptive histogram equalization and its variations," Computer vision, graphics, and image processing, vol. 39, pp. 355-368, 1987.

[104] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics gems IV, S. H. Paul, Ed., ed: Academic Press Professional, Inc., 1994, pp. 474-485.

[105] F. Essannouni, R. O. Haj Thami, D. Aboutajdine, and A. Salam, "Simple noncircular correlation method for exhaustive sum square difference matching," Optical Engineering, vol. 46, pp. 107004-107004, 2007.

[106] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of basic Engineering, vol. 82, pp. 35-45, 1960.

[107] G. Welch and G. Bishop, "An introduction to the Kalman filter," 1995.

[108] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," Computer vision and image understanding, vol. 61, pp. 38-59, 1995.

[109] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, pp. 681-685, 2001.

[110] P. Viola and M. Jones, "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, pp. I-511-I-518 vol. 1.

[111] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in Systems, Man and Cybernetics, 2005 IEEE International Conference on, 2005, pp. 1692-1698.

[112] A. S. M. Sohail and P. Bhattacharya, "Detection of facial feature points using anthropometric face model," in Signal Processing for Image Enhancement and Multimedia Processing, ed: Springer, 2008, pp. 189-200.

[113] J. F. Cohn, L. I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior," in Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004, pp. 610-616.

[114] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in Proceedings of the 8th international conference on Multimodal interfaces, 2006, pp. 162-170.

[115] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders," Journal of neuroscience methods, vol. 200, pp. 237-256, 2011.

[116] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in Real-time vision for human-computer interaction, ed: Springer, 2005, pp. 181-200.

[117] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, pp. 1940-1954, 2010.

[118] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Segment and recognize expression phase by fusion of motion area and neutral divergence features," in Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, 2011, pp. 330-335.

[119] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," Image and Vision Computing, 2012.

[120] S. Strupp, N. Schmitz, and K. Berns, "Visual-based emotion detection for natural man-machine interaction," in KI 2008: Advances in Artificial Intelligence, ed: Springer, 2008, pp. 356-363.

[121] L. Ding and A. M. Martinez, "Precise detailed detection of faces and facial features," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1-7.

[122] J. L. Moreira, A. Braun, and S. R. Musse, "Eyes and Eyebrows Detection for Performance Driven Animation," in Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on, 2010, pp. 17-24.

[123] M. Chau and M. Betke, "Real time eye tracking and blink detection with USB cameras," Boston University Computer Science Department2005.

[124] E. Missimer and M. Betke, "Blink and wink detection for mouse pointer control," in Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments, 2010, p. 23.

[125] A. Królak and P. Strumillo, "Vision-based eye blink monitoring system for human-computer interfacing," in Human System Interactions, 2008 Conference on, 2008, pp. 994-998.

[126] I. F. Ince and T.-C. Yang, "A new low-cost eye tracking and blink detection approach: extracting eye features with blob extraction," in Emerging Intelligent Computing Technology and Applications, ed: Springer, 2009, pp. 526-533.

[127] T. Danisman, I. M. Bilasco, C. Djeraba, and N. Ihaddadene, "Drowsy driver detection system using eye blink patterns," in Machine and Web Intelligence (ICMWI), 2010 International Conference on, 2010, pp. 230-233.

[128] M. Divjak and H. Bischof, "Eye blink based fatigue detection for prevention of Computer Vision Syndrome," in IAPR Conference on Machine Vision Applications, Tokyo, 2009.

[129] J.-W. Li, "Eye blink detection based on multiple Gabor response waves," in Machine Learning and Cybernetics, 2008 International Conference on, 2008, pp. 2852-2856.

[130] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," Universal Access in the Information Society, vol. 11, pp. 409-419, 2012.

[131] T. Bhaskar, F. T. Keat, S. Ranganath, and Y. Venkatesh, "Blink detection and eye tracking for eye localization," in TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, 2003, pp. 821-824.

[132] M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau, "Real-time eye blink detection with GPU-based SIFT tracking," in Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on, 2007, pp. 481-487.

[133] R. Heishman and Z. Duric, "Using image flow to detect eye blinks in color videos," in Applications of Computer Vision, 2007. WACV'07. IEEE Workshop on, 2007, pp. 52-52.

[134] C. D. N. Ayudhya and T. Srinark, "A Method for Real-Time Eye Blink Detection and Its Application."

[135] A. Panning, A. Al-Hamadi, and B. Michaelis, "A color based approach for eye blink detection in image sequences," in Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on, 2011, pp. 40-45.

[136] K. Arai and R. Mardiyanto, "Comparative Study on Blink Detection and Gaze Estimation Methods for HCI, in Particular, Gabor Filter Utilized Blink Detection Method," in Information Technology: New Generations (ITNG), 2011 Eighth International Conference on, 2011, pp. 441-446.

[137] I. Bacivarov, M. Ionita, and P. Corcoran, "Statistical models of appearance for eye tracking and eye-blink detection and measurement," Consumer Electronics, IEEE Transactions on, vol. 54, pp. 1312-1320, 2008.

[138] J. Wu and M. M. Trivedi, "An eye localization, tracking and blink pattern recognition system: Algorithm and evaluation," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 6, p. 8, 2010.

[139] W. O. Lee, E. C. Lee, and K. R. Park, "Blink detection robust to various facial poses," Journal of neuroscience methods, vol. 193, pp. 356-372, 2010.

[140] I. Choi, S. Han, and D. Kim, "Eye detection and eye blink detection using adaboost learning and grouping," in Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on, 2011, pp. 1-4.

[141] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8.

[142] T. Moriyama, T. Kanade, J. F. Cohn, J. Xiao, Z. Ambadar, J. Gao, et al., "Automatic recognition of eye blinking in spontaneously occurring behavior," in Pattern Recognition, 2002. Proceedings. 16th International Conference on, 2002, pp. 78-81.

[143] M. J. Roshtkhari and M. D. Levine, "An On-Line, Real-Time Learning Method For Detecting Anomalies In Videos Using Spatio-Temporal Compositions (under review)," Computer Vision and Image Understanding, 2013.

[144] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 886-893.

[145] J.-L. Fan and B. Lei, "A modified valley-emphasis method for automatic thresholding," Pattern Recognition Letters, vol. 33, pp. 703-708, 2012.

[146] (2013). Blink. Available: http://en.wikipedia.org/wiki/Blink [cited July 10, 2013]

[147] N. Otsu, "A threshold selection method from gray-level histograms," Automatica, vol. 11, pp. 23-27, 1975.

[148] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, pp. 699-714, 2005.

[149] S. Marcos, J. Gómez-García-Bermejo, E. Zalama, and J. López, "Nonverbal communication with a multimodal agent via facial expression recognition," in Robotics and Automation (ICRA), 2011 IEEE International Conference on, 2011, pp. 1199-1204.

[150] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim, "Age estimation using a hierarchical classifier based on global and local facial features," Pattern Recognition, vol. 44, pp. 1262-1281, 2011.

[151] Y.-H. Choi, Y.-S. Tak, S. Rho, and E. Hwang, "Accurate Wrinkle Representation Scheme for Skin Age Estimation," in Multimedia and Ubiquitous Engineering (MUE), 2011 5th FTRA International Conference on, 2011, pp. 226-231.

[152] W.-B. Horng, C.-P. Lee, and C.-W. Chen, "Classification of age groups based on facial features," Tamkang Journal of Science and Engineering, vol. 4, pp. 183-192, 2001.

[153] H. Takimoto, Y. Mitsukura, M. Fukumi, and N. Akamatsu, "Robust gender and age estimation under varying facial pose," Electronics and Communications in Japan, vol. 91, pp. 32-40, 2008.

[154] J.-D. Txia and C.-L. Huang, "Age estimation using AAM and local facial features," in Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP'09. Fifth International Conference on, 2009, pp. 885-888.

[155] J. Hayashi, M. Yasumoto, H. Ito, and H. Koshimizu, "Method for estimating and modeling age and gender using facial image processing," in Virtual Systems and Multimedia, 2001. Proceedings. Seventh International Conference on, 2001, pp. 439-448.

[156] Y. H. Kwon and N. D. V. Lobo, "Age classification from facial images," Computer Vision and Image Understanding, vol. 74, pp. 1-21, 1999.

[157] Y.-H. Choi, Y.-S. Tak, S. Rho, and E. Hwang, "Skin feature extraction and processing model for statistical skin age estimation," Multimedia Tools and Applications, pp. 1-21, 2012.

[158] C.-Y. Chang, S.-C. Li, P.-C. Chung, J.-Y. Kuo, and Y.-C. Tu, "Automatic Facial Skin Defect Detection System," in Broadband, Wireless Computing, Communication and Applications (BWCCA), 2010 International Conference on, 2010, pp. 527-532.

[159] N. Batool and R. Chellappa, "Modeling and detection of wrinkles in aging human faces using marked point processes," in Computer Vision–ECCV 2012. Workshops and Demonstrations, 2012, pp. 178-188.

[160] G. Lemperle, R. E. Holmes, S. R. Cohen, and S. M. Lemperle, "A classification of facial wrinkles," Plastic and reconstructive surgery, vol. 108, pp. 1751-1752, 2001.

[161] J. R. Movellan, "Tutorial on gabor filters," Open Source Document, 2002.

[162] A. Senior, R.-L. Hsu, M. A. Mottaleb, and A. K. Jain, "Face Detection in Color Images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, pp. 696-706, 2002.

[163] Y. Tian, T. Kanade, and J. Cohn, "Robust Lip Tracking by Combining Shape, Color and Motion," presented at the The 4th Asian Conference on Computer Vision, 2000.

[164] M. Nilsson, I. Gertsovich, and J. S. Bartunek, "Mouth open or closed decision for frontal face images with given eye locations," in Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, 2010, pp. 1-6.

[165] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Developing a practical smile detector," in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, 2008.

[166] D. McDuff, R. El Kaliouby, and R. Picard, "Crowdsourcing Facial Responses to Online Videos," 2012.

[167] S. L. Wang, W. H. Lau, and S. H. Leung, "Automatic lip contour extraction from color images," Pattern Recogn., vol. 37, pp. 2375-2387, 2004.

[168] A. V. Nefian, L. Luhong, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, 2002, pp. II-2013-II-2016.

[169] S. Werda, W. Mahdi, and A. B. Hamadou, "Lip localization and viseme classification for visual speech recognition," arXiv preprint arXiv:1301.4558, 2013.

[170] A. Hulbert and T. Poggio, "Synthesizing a colour algorithm from examples," Sciences, vol. 239, pp. 482-485, 1998.

[171] P. A. Devijver and J. Kittler, Pattern recognition: A statistical approach: Prentice/Hall International Englewood Cliffs, NJ, 1982.

[172] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.

[173] L. Breiman, "Technical note: Some properties of splitting criteria," Machine Learning, vol. 24, pp. 41-47, 1996.

[174] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," Encyclopedia of Database Systems, vol. 3, pp. 532-38, 2009.

[175] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," Image and Vision Computing, vol. 26, pp. 1052-1067, 2008.

[176] F. Bourel, C. C. Chibelushi, and A. A. Low, "Recognition of facial expressions in the presence of occlusion," in Proceedings of the Twelfth British Machine Vision Conference, 2001, pp. 213-222.

[177] S. Polikovsky, Y. Kameda, and Y. Ohta, "Detection and Measurement of Facial Micro-Expression Characteristics for Psychological Analysis," Kameda's Publication, vol. 110, pp. 57-64, 2010.

[178] A. Vrij, Detecting Lies and Deceit: Pitfalls and Opportunities, 2 ed.: John Wiley & Sons, 2008.

# Appendix I. List of Forensic Cases in Our Database

*On Porter's list*

| Name of Suspect | Name of Missing Person | Relationship to Missing Person | Guilt/ Innocence |
|---|---|---|---|
| Audrey Hingston | Eric Hingston | spouse/partner | guilty |
| Darren Vickers | Jamie Lavis | acquaintence | guilty |
| Derek Fleming | Linda Flemming | father | guilty |
| Fadi Nasri | Nisha Patel-Nasri | spouse/partner | guilty |
| Gordon Wardell | Carol Wardell | spouse/partner | guilty |
| Graham Alderton | Three Children | father | guilty |
| Ian Huntley | Holly Wells & Jessica Chapman | acquaintence | guilty |
| Jean Daddow | Terence Daddow | spouse/partner | guilty |
| John Tanner | Rachel McLean | spouse/partner | guilty |
| Karen Matthews | Shannon Matthews | mother | guilty |
| Maxene Carr | Holly Wells & Jessica Chapman | acquaintence | guilty |
| Mike Gifford-Hull | Kirsi Gifford-Hull | spouse/partner | guilty |
| Miles Evans | Zoe Evans | step-father | guilty |
| Mitchell Quy | Lindsay Quy | spouse/partner | guilty |
| Mukhtiar Panghali | Mangit Panghali | spouse/partner | guilty |
| Nick Kay | Rhonda Kay | spouse/partner | guilty |
| Paul Dyson | Joanne Nelson | spouse/partner | guilty |
| Penny Boudreau | Karissa Boudreau | mother | guilty |
| Sion Jenkins | Billy Joe | father | guilty |
| Susan Smith | Michael & Alex Smith | mother | guilty |
| Vincent Shilton | Lisa Blunt | spouse/partner | guilty |
| Agnes Gaylor | Diana Garbott | mother | innocent |
| Alan Symes | Aisling Symes | father | innocent |
| Angela Symes | Aisling Symes | mother | innocent |
| Ed Smart | Elizabeth Smart | father | innocent |

| | | | |
|---|---|---|---|
| Harry Clinch | Sharon Malone | father | innocent |
| Jean Nelson | Joanne Nelson | mother | innocent |
| Joanne Coombs | Natasha Coombs | mother | innocent |
| Katen Patel | Nisha Patel-Nasri | brother | innocent |
| Keith Lunnon | Charlene Lunnon | father | innocent |
| Paula Evans | Zoe Evans | mother | innocent |
| Rodney Stafford | Victoria Stafford | father | innocent |
| Sara Payne | Sarah Payne | mother | innocent |

*Collected by the authors*

| Name of Pleader | Name of Missing Person | Relationship to Missing Person | Guilt/ Innocence |
|---|---|---|---|
| Biurny Peguero | William McCaffrey | accuser | guilty |
| Carlos Perez-Olivo | Peggy Perez-Olivo | husband | guilty |
| Dave Hawk | Debbie Hawk | ex-husband | guilty |
| Diane Downs | Stephen Daniel | mother | guilty |
| Jerry Sandusky | 52 children | acquaintence | guilty |
| Mark Hacking | Lori Hacking | husband | guilty |
| Matt Baker | Kari Baker | husband | guilty |
| Melanie McGuire | Bill McGuire | wife | guilty |
| Phil Spector | Lana Clarkson | husband | guilty |
| Robert Smith | Keisha Abrahams | stepfather | guilty |
| Sam Parker | Theresa Parker | husband | guilty |
| Scott Peterson | Laci Peterson | husband | guilty |
| Tracie Andrews | Lee Harvey | wife | guilty |
| Travis Forbes | Kenia Monge | stranger | guilty |
| William Walsh | Leah Walsh | husband | guilty |
| Aaron Young | Zahra Baker | stepmother's ex-husband | innocent |
| Adam Baker | Zahra Baker | father | innocent |
| Amanat Khan | Aisha Khan | father | innocent |
| Avtar Kolar | Avtar Kolar and Carole Kolar | father | innocent |

| | | | |
|---|---|---|---|
| Bob Dowler | Milly Dowler | father | innocent |
| Damon Van Dam | Danielle Van Dam | father | innocent |
| David Yeates | Joanna Yeates | father | innocent |
| Deborah Irwin | Lisa Irwin | mother | innocent |
| Diena Thompson | Somer Thompson | mother | innocent |
| Erin Runnion | Samantha Runnion | mother | innocent |
| Gary Coombs | Natasha Coombs | father | innocent |
| George Anthony | Caylee Anthony | grandfather | innocent |
| Gerry McCann | Madeleine McCann | father | innocent |
| Jeremy Irwin | Lisa Irwin | father | innocent |
| Kate McCann | Madeleine McCann | mother | innocent |
| Kirk Turner | Jennifer Turner | husband | innocent |
| Michelle Kirwan | Avtar Kolar and Carole Kolar | mother | innocent |
| Sally Dowler | Milly Dowler | mother | innocent |
| Sarah Smith | Elizabeth Brown | daughter | innocent |
| Teresa Scott | Carmen Thomas | mother | innocent |
| Tonya Craft | children | teacher | Innocent |

# Appendix II. Pseudo-code of Constructing a Random Forest

Below is the pseudo-code of constructing an RF containing $k$ decision trees using dataset $D$, and each instance in $D$ has $F$ features. Denote the number of classes by $H$.

**CONSTRUCT_RANDOM_FOREST($D$, $x$)**

FOR $i = 1$ TO $k$ DO

Randomly select $\omega$ samples from $D$ with replacement, forming the training data $D_i$ for constructing $tree_i$

Create root node $q_i$ containing data $D_i$

SPLIT_NODE($q_i$)

END

**SPLIT_NODE**($\boldsymbol{q}$)

IF all data at node $q$ belong to the same class THEN

Return

ELSE

Randomly select $f = log_2 F + 1$ features without replacement, forming candidate features for splitting the node $\mathbb{F} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_f\}$

Compute the best splitting rule $r^*$ = BEST_SPLIT($q, \mathbb{F}$)

Split dataset $D$ at node $q$ into two subsets $D_L$ and $D_R$ according to splitting rule $r^*$

Creating two child nodes $q_L$ and $q_R$ for $D_L$ and $D_R$, respectively

SPLIT_NODE($q_L$)

SPLIT_NODE($q_R$)

END

**BEST_SPLIT**($\boldsymbol{q}, \mathbb{F}$)

Initialize the "goodness-of-split" function $\mathbb{G} = 0$

Compute data portions at node $q$: $\boldsymbol{p} = (p_1, p_2, \dots, p_H)$, where $p_h = \frac{number\ of\ samples\ belong\ to\ class\ h\ at\ node\ q}{number\ of\ samples\ at\ node\ q}$

**FOR** $j = 1$ **TO** $f$ **DO**

Denote feature $\mathcal{F}_j \in \{\mathcal{F}_j{}^1, \mathcal{F}_j{}^2, \dots, \mathcal{F}_j{}^{\mathrm{B}}\}$

**FOR** $\beta = 1$ **TO** B **DO**

Create a split rule $s_j{}^{\beta}: \mathcal{F}_j \le \mathcal{F}_j{}^{\beta}$

Partition data $D$ at node $q$ into two subsets $D_L$ and $D_R$ according to $s_j{}^{\beta}$

Compute data portions in each subset:

$$P_L = \frac{number\ of\ samples\ in\ D_L}{number\ of\ samples\ in\ D}, P_R = 1 - P_L$$

$$\boldsymbol{p}_L = (p_{1,L}, p_{2,L}, \dots, p_{H,L}),$$

$$p_{h,L} = \frac{number\ of\ samples\ belong\ to\ class\ h\ in\ D_L}{number\ of\ samples\ in\ D_L},$$

$$\boldsymbol{p}_R = (p_{1,R}, p_{2,R}, \dots, p_{H,R}),$$

$$p_{h,R} = \frac{number\ of\ samples\ belong\ to\ class\ h\ in\ D_R}{number\ of\ samples\ in\ D_R}$$

Compute impurity function:

$$\phi(\boldsymbol{p}_L) = \sum_{h=1}^{H} p_{L,h}(1 - p_{L,h})$$

$$\phi(\boldsymbol{p}_R) = \sum_{h=1}^{H} p_{R,h}(1 - p_{R,h})$$

Compute "goodness-of-split" of splitting rule $r_j{}^{\beta}$:

$$\mathbb{G}(r_j{}^{\beta}, D) = \phi(\boldsymbol{P}) - P_L\phi(\boldsymbol{p}_L) - P_R\phi(\boldsymbol{p}_R)$$

**IF** $\mathbb{G}(r_j{}^{\beta}, D) > \mathbb{G}$ **THEN**

$$\mathbb{G} = \mathbb{G}(r_j{}^{\beta}, D), \quad r^* = r_j{}^{\beta}$$

**END**

**END**

**END**

Return $r^*$

# Appendix III. Test Accuracy, TPR and TNR as FTV Size Changes

The highest accuracy is in bold.

| FTV size | test accuracy | test TPR | test TNR |
|---|---|---|---|
| 5 | 75.38% | 65.52% | 83.33% |
| 6 | 75.38% | 65.52% | 83.33% |
| 7 | 73.85% | 65.52% | 80.56% |
| 8 | 75.38% | 65.52% | 83.33% |
| 9 | 73.85% | 65.52% | 80.56% |
| 10 | **76.92%** | 68.97% | 83.33% |
| 11 | 70.77% | 58.62% | 80.56% |
| 12 | 73.85% | 65.52% | 80.56% |
| 13 | 70.77% | 65.52% | 75.00% |
| 14 | **76.92%** | 72.41% | 80.56% |
| 15 | **76.92%** | 68.97% | 83.33% |
| 16 | **76.92%** | 72.41% | 80.56% |

| | | | |
|---|---|---|---|
| 17 | **76.92%** | 65.52% | 86.11% |
| 18 | 73.85% | 72.41% | 75.00% |
| 19 | 72.31% | 72.41% | 72.22% |
| 20 | 72.31% | 62.07% | 80.56% |