2014

Facial movement based human user authentication

Pengqing Xie
Iowa State University

Recommended Citation

Xie, Pengqing, "Facial movement based human user authentication" (2014). *Graduate Theses and Dissertations*. 14267.
https://lib.dr.iastate.edu/etd/14267

Facial movement based human user authentication

by

Pengqing Xie

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Engineering

Program of Study Committee:
Yong Guan, Major Professor
Daji Qiao
Diane Thiede Rover

Iowa State University

Ames, Iowa

2014

DEDICATION

I would like to dedicate this thesis to my father Xinpo Xie and mother Ling Zhang for their eternal affection and unconditional support. I would also like to dedicate this thesis to my wife Shunhua Pan without whose love and consideration I would not have been able to finish this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to offer my sincere appreciation to my major professor, Yong Guan, who pointed me to the right direction and shed a bright light on my path. I would like to thank my committee members Daji Qiao and Diane Rover for their enthusiasm and support on my research. Additionally, I would also like to thank my office mates for sharing such an inspiring and joyful working environment. Last but not least, I am greatly indebted to those who were willing to participate in my experiments, without whom, this thesis would not have been possible.

ABSTRACT

Face recognition is a form of biometric authentication that has received significant attention during the last decades. Using the human face as a key to security, face recognition technology can be potentially employed in many commercial and law enforcement applications. Despite of the fact that most of the face recognition techniques have greatly developed since the earliest forms, they suffer from spoofing attack which aims at deceiving the sensor by manipulating a face replica. One of the methods to solve this problem is to utilize facial movements.

Facial muscle movements represent facial behavior which makes it unrealistic to be replicated and thus more distinctive. The third dimension of facial data – depth – is also utilized to improve recognition performance and to avoid video-based attack. Apart from security concerns, physiologists and psychologists have discovered the imperative role of facial movements during human face perception. Therefore, a “3D dynamic signature” can be added to augment facial recognition for which relying on static features related to shape and color.

In this thesis, a user authentication method based on spatiotemporal facial movements is proposed. Facial movements are obtained by making a facial expression in front of a 3D camera and are encoded by a standard system. By discretizing motion classifications into values, the problem of face recognition can be reinterpreted as matching two time sequences – probe and gallery – for each facial movement category obtained during the enrollment phase and the verification phase. Experiments have been conducted to show the possibility of discriminating subjects based on their facial movements.

CHAPTER 1. INTRODUCTION

Research Motivation

Security system often has to make authentication decisions to grant access to legitimate users. Currently, the most common method is to use a string of characters and numbers, which are known as the password. However, passwords can be transferred and stolen which often lead to identity theft and fraud. A recent bug called heart bleed[1] has been discovered in the widely used cryptographic software openSSL which opens up a breach allowing hackers to eavesdrop on communication, stealing any information protected from the services including passwords. Security organizations have advised the general public to alter their passwords and maintain dissimilar patterns for different accounts but little attention is received. The underlying reason is that passwords require memorization and are clumsy to manage, not to mention that they are increasingly insecure. The current trend of solving these problems is to use biometric authentication which utilizes body characteristics. Nevertheless, biometric methods suffer from impersonation attacks. For instance, audio clips of the user’s casual conversations can be recorded to deceive a voice recognition system; fingerprints may be collected from leftovers or latents to gain unauthorized access to a fingerprint-based system. Just like the other biometrics, face recognition can be bypassed by manipulating digitized facial images in front of a camera. Figure 1 illustrates such an attempt to spoof face recognition software by using a facial image displayed on a computer screen. Since most face recognition algorithms only detect and track physiological features, they

[1] Details about the heart bleed bug can be found at http://heartbleed.com.

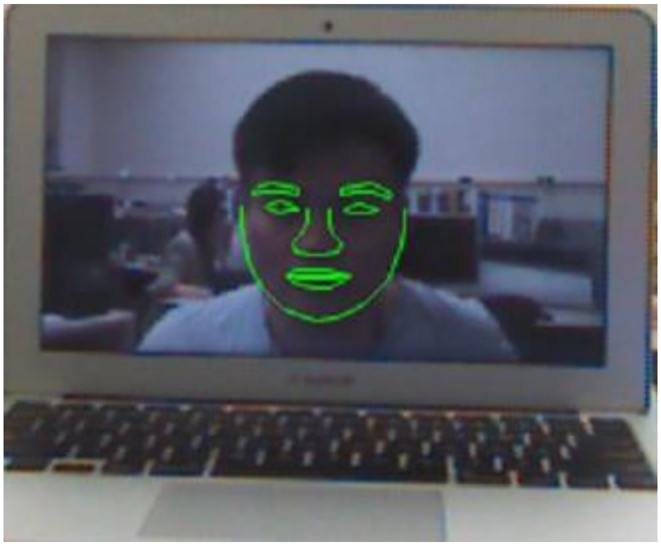


Figure 1. Face recognition software detects a face displayed on a computer screen.

may not be able to distinguish between a real and a counterfeited face [1]. With the advent of social media and search engines, face images are publically exposed all over the internet and can be potentially used to impersonate identities without consent. In an effort to countermeasure such attacks, facial movements are exploited to detect liveness of the displaying face. However, traditional liveness detection techniques potentially suffer from video-based attacks. Thus, we have proposed to add the third dimension – depth – to effectively prevent impersonation attack from all aspects.

Apart from security concerns, facial movements also benefit recognition. Numerous researches have been conducted by physiologists and psychologists on human face perception to understand the role of facial movements in machine recognition. Researchers have already discovered that familiar faces are easier to recognize when conveyed in moving sequence than in still photos [2]. It is argued that human’s ability to perform facial expressions is consisted of both nature and nurture, which encapsulates not only innate biological characteristics, but also learned facial behaviors stimulated by emotions [3]. In agreement to this argument, neural studies have shown that during face perception, not only physiological features are extracted using face specific brain areas, but a consistent neural activity is also stored [4]. This implies that a “dynamic signature” can be employed for face identification.

In this thesis, we propose a user authentication method based on facial movements. Given that each user makes a facial expression in a slightly different way, identities can be differentiated by exploiting facial behaviors or facial movement patterns. Facial behaviors are inherent to each user, thus are more distinctive and cannot be easily replicated.

Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 delivers a formal definition of the problems being solved. Chapter 3 contains a literature review of existing authentication methods as well as related works in attack countermeasures and dynamic face recognition. Chapter 4 comprises the proposed methods for facial movement based authentication. Chapter 5 introduces reference designs for data acquisition and authentication. Chapter 6 details the experimental evaluations conducted with a discussion on the result. Chapter 7 summarizes the thesis and presents potential future improvements. Additionally, Appendix A provides a complete list of the shaping units and action units that have been used in our approach. Appendix B includes the approval document from Institutional Review Board (IRB) for researches involving human subjects.

CHAPTER 2. PROBLEM DEFINITION

The formal problem statement can be formulated as follows: given video images of a user’s face, verify his/her identity by extracting a spatiotemporal facial pattern and comparing it with existing templates in a database. Generally, we expect to be able to discriminate user identities with dynamical facial characteristics which are distinctive and invulnerable to impersonation attacks.

Facial movements are defined by the Facial Action Coding System (FACS) which categorizes facial behaviors by 46 action units (AUs), each of which is anatomically related to a specific set of facial muscles [5]. By characterizing facial movements into discrete values, the problem of face recognition is thus relaxed to matching two sequence – probe and gallery – for each AU acquired in a time interval under controlled environment. To restrict our scope, an existing 3D model is exploited to detect and track faces among video frames. Thus, in this thesis we focus on the modeling and matching of six major AU patterns specifically for an authentication scheme.

Previous face recognition researches related to facial movements either focus on counter-measuring spoofing attacks [6] or coping with facial expression for robustness [7], however, anti-spoofing techniques suffer from video-based attack and, in both cases, face verification process are still carried out with respect to each user’s static physical appearance. Acknowledging both the security concerns and human perception benefits as mentioned in Chapter 1, our approach aims at differentiating user identities based on spatiotemporal facial patterns. Similarly, facial pattern recognition was proposed in [8] [9], nevertheless, a probabilistic path was taken to represent the identity space which may not reveal the true facial pattern. By following the facial encoding standard FACS as defined by physiologists and psychologists, identity space can thus be represented by facial muscle changes revealed by making a facial expression.

Traditional impersonation attacks can be efficiently prevented by exploiting facial movements, however, our approach can still be vulnerable to video-based attacks. Thus, the third dimension is also added acknowledging the fact that depth data can hardly be replayed in front of a 3D camera.

Requirements

In general, our facial movement based user authentication method has to satisfy the following five requirements:

(2.1) Make verification decision based on facial movements.

(2.2) Invulnerable to impersonation attacks.

(2.3) Non-learnable and non-replicable.

(2.4) False positive is less than 10% for the samples gathered under controlled environment.

(2.5) False negative is less than 10% for the samples gathered under controlled environment.

CHAPTER 3. LITERATURE REVIEW

Face recognition is an important technique among biometric methods. Similar to many other biometric techniques, an individual's identity is established based on physiological and/or behavioral characteristics rather than knowledge such as passwords or possession such as ID card. For the past few decades, numerous efforts have been made by the researchers to improve the performance of face recognition, however, less attention have been paid to another important factor, circumvention, which considers the vulnerability under spoofing attacks [10]. In the following sections, we briefly summarize existing authentication methods, evaluate state of the art liveness detection techniques preventing impersonation attacks, and also present related works on the dynamic face recognition approach.

Authentication

Authentication is the binding of an identity to a subject. Based on the subject information used, authentication methods generally have four categories: knowledge-based (what the subject knows), object-based (what the subject has), biometric-based (what the subject is), and location-based (where the subject is) [11]. In the following sections, we examine knowledge-based and biometric-based approaches which are relevant to our research.

Knowledge-based Approach

Knowledge-based methods are the most widely used authentication techniques which generally include the traditional text-based passwords and the more recent graphical-based approach. For decades, password has been used as the standard means for user authentications. Despite of its long existence and popularity, many studies have shown the vulnerabilities of password-based authentication [12] [13] . The major problem is that simple passwords can often be searched, such as using dictionary attack, whereas long, random, changing passwords are difficult to memorize. In an effort to ameliorate this problem, cryptography algorithms are developed to secure the passwords. For instance, the advanced encryption standards (AES), which has been adopted by the U.S. government as the standard, makes it mathematically infeasible to guess the password by encrypting it with a private key as long as 256 bits. However, the heart of any security system is people. Since the private key is too long to be remembered by the users, it has to be stored on a computer protected by another password – which lead us back to the original problem [14]. Another issue associated with passwords is that user often choose passwords consist of personal information, such as birth date, phone number, family name and so on. With the advancement of social media and global search, guessing attack (or linkage attacks) are another foreseeable thread against the usage of password.

Motivated by the fact that humans are able to recall images better than text, graphic-based user authentication has been proposed as an alternative to passwords [15]. Consider, for example, the Windows 8 operating system which has included picture password as a login option. A picture password is comprised of two complimentary parts: a picture chose by the user from user’s own collection and a set of drawing gestures applied on the picture. Coordinates of the gestures are recorded during enrollment and later compared with gestures performed when user attempt to login. Figure 2 illustrate an example of Windows 8 picture password where a straight line gesture is carried out from $(X1, Y1)$ to $(X2, Y2)$. By

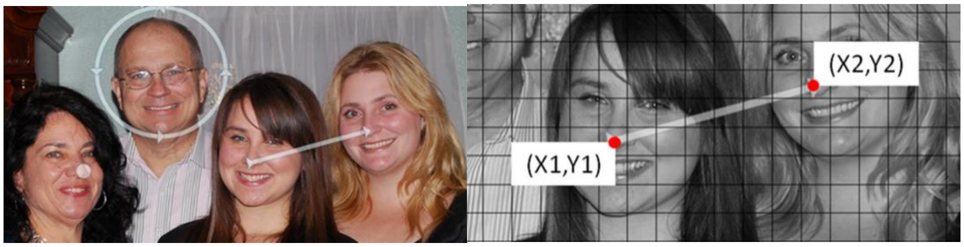


Figure 2. Illustration of picture password on Windows 8. On the left are the common gestures and on the right are the recorded coordinates used for authentication. (Courtesy of MSDN Blog[2])

converting alphanumeric password into coordinates of pixels (including the gesture types), the number of possible combinations can be greatly increased. However, similar to text-based passwords, user often choose weak and predictable graphic passwords which potentially suffer from guessing attacks [16].

Biometric-based Approach

Identity theft is becoming a major concern in the modern society, especially in United States where credits are being heavily utilized. During the year of 2013, approximately 280,000 incidents of identity thefts had been reported to Consumer Sentinel Network (CSN)[3]. Numerous methods have been proposed to prevent identity theft, one of them being biometrics, which has already been adopted by many law enforcement applications. For instance, criminal fingerprints are being collected and compared against suspects; facial photos are being taken at the border customs to verify travelers' identities. Because biometric characteristics are inborn, they requires no memorization (as needed by passwords), no

[2] More details about Windows 8 picture password can be found at http://blogs.msdn.com/b/b8/archive/2011/12/16/signing-in-with-a-picture-password.aspx

[3] More details about CSN identity theft report can be found at https://www.ncjrs.gov/spotlight/identity_theft/facts.html

physical possessions (as required by token-based methods), and most importantly, they are usually unique. Thus biometrics are invulnerable to traditional knowledge-based attacks such as guessing attacks and dictionary attacks.

Researchers of biometric-based authentication seem to exhibit a very diverse interest. Examples of biometric characteristics being studied include: DNA, ear, face, facial thermogram, hand thermogram, hand vein, fingerprint, hand geometry, iris, palmprint, retina, signature, and voice [10]. Taking a different route from traditional biometric methods based on physical features, behavioral biometrics are gaining increasing interests in this field. Behavioral biometrics such as key stroke dynamics [17] are advantageous because of its non-obtrusiveness [18]. Additionally, assuming that feature traits are quantified accurately, behavioral biometrics are invulnerable to impersonation attack as experienced by traditional methods. Consider gait recognition, for example, which is the closest biometric method to our approach in terms of feature extraction and template matching, aims at extracting spatiotemporal gait pattern based on accelerometry [19]. In an attempt to attack a gait authentication system by imitating legitimate user's walking pattern (commonly known as minimum-effort attack), Davrondzhon et al. have demonstrated that the attack does not significantly increase the chance of being accepted [20]. Since biometric methods usually have to contend with a variety of problems such as noisy data and false positives, a multi-model biometric system has been introduced. By combining multiple sources of information, a multimodal system is able to improve matching performance and possibly prevent spoofing attacks [21].

Face Liveness Detection

The problem of spoofing attack can be prevented by detecting face liveness. It serves as an added layer to existing face recognition to perceive imposters with a fake face after the verification stage. Typically, there are two types of liveness detection indicators: texture and motion. For each category, related methods are presented below followed by a discussion on the advantages and disadvantages.

Texture-based Approach

Texture-based approach assumes that real faces have a higher frequency component. Based on the assumption that photos used for spoofing are typically printed on papers, fake faces can therefore be differentiated by detecting lower frequency component and texture loss due to print failures and overall image blur [6]. In the work of Gahyun Kim et al. texture features are extracted and analyzed using a popular texture descriptor called Local Binary Pattern (LBP) whereas frequency features are obtained by 2-D discrete Fourier transform. These two types of features are fused using Support Vector Machine (SVM) classifier as the decision maker [22]. A similar approach was taken by Nalinakshi et al. where regions of eye, lip, chin and forehead are extracted using LBP and analyzed using mean and standard deviation; if there is any variation in either region, then face is considered alive else not alive [23]. Figure 3 illustrates variations that have been detected within eye, lip, forehead, and chin regions.

Texture-based approaches are easy to implement and require no user collaboration. However, it is still possible to attack using a video displayed on a high-resolution screen as opposed to a printed paper. Therefore, Allan et al. proposed to measure the “visual rhythm”

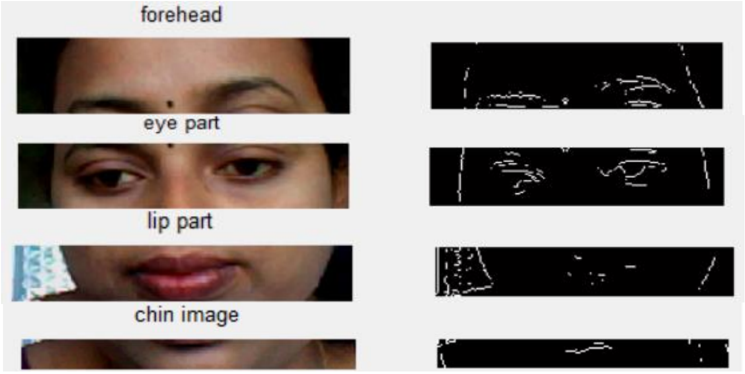


Figure 3. Texture-based approach detecting variations in eyes, lip, forehead, and chin. (Courtesy of Nalinakshi et al. [23]**)**

describing the Fourier spectrum noise that had been introduced due to image/video encoding [24]. Nevertheless, this approach relies heavily on the noise pattern and thus requires a diverse training database.

Motion-based Approach

Most other researches that have been conducted on this field interprets face liveness as motions, thus they focus on detecting differing facial components in videos or even real-time. Bao et al. introduced a method based on the optical flow field which is a common method used for estimating motion direction in videos [25]. The basic idea is that optical flow field for 2D objects can be represented by a projection transformation from the reference field and thus whether the face is planar or not can be determined by measuring the difference between fields. Inspired by the optical flow approach, Andre et al. used image background region as the reference field, and directly compared motion direction with face region using optical flow correlation (OFC) [26]. OFC works by exploiting the fact that genuine faces move with respect to the background whereas faces in printed-photos move along with the background. Similarly, Kollreider et al. used the optical flow of lines (OFL) of certain face parts to detect face liveness [27]. By assuming that a 3D face generates greater motion at central face parts such as nose compared to outer parts such as ears, face liveness can be detected by calculating the motion difference between the two parts. Figure 4 illustrates the motion generated by different focused face parts. Their experimental results have shown that, due to natural and unintentional human behavior, small head rotations will always occur whenever a live face is presented in front of a camera, thus horizontal OFL between two parts can be computed to decide liveness. Taking a similar human natural behavior path, Gang et al. employed Conditional Random Fields (CRFs) to describe the blinking of eyes to determine a live face [28].

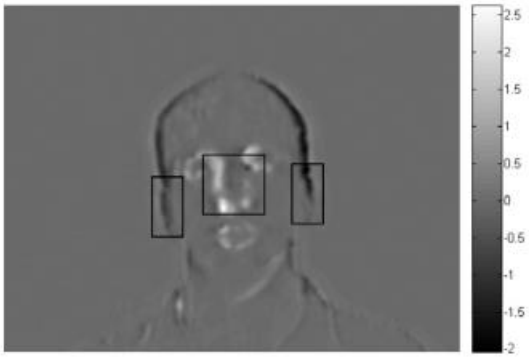


Figure 4. Horizontal OFL focused on two face parts - nose and ears. (Courtesy of Kollreider et al. [27])

The fundamental basis of all motion-based approaches is that static facial images move significantly different from real faces which are 3D objects, and the motion pattern generated by planar objects can be differentiated from motions resulted by real human faces. When using motion analysis, it is difficult to spoof by 2D face images. However, it may face issues under video-based attacks (such as OFC). It may also suffer when there is low motion information due to different user behavior. For instance, eye closity may not be detected if user stares at the camera for a certain amount of time without blinking eyes, if such time takes longer than the timeout threshold, false negative occurs. As mentioned by Saptarshi and Dhrubajyoti, motion-based approaches may also fail when sophisticated attacks are performed such as using 3D sculpture face model [6]. For example, the aforementioned OFL approach cannot distinguish between a horizontally rotating 3D sculpture face and natural rotations of a real face.

Our approach lies within the category of motion, however, excluding the disadvantages. Firstly, since we require facial movements to be generated by either user cooperation or emotion stimulus[4], lack of motion data is not an issue. Secondly, replicating facial movements on a 3D sculpture model is an impossible task. It is worth noting that our strategy is similar to the use of encryption algorithm which makes passwords intrinsically harder to be cracked by imposters. Ultimately, since we are combining holistic facial movements in aid of face recognition, we eliminate the need to implement accessary face liveness detection unit.

Dynamic Face Recognition

Numerous researches have been conducted on dynamic face recognitions to exploit spatiotemporal facial movements. In the work of Tistarelli et al. physiological and behavioral cues for face recognition were derived from neural activation and infant behavioral studies [8]. Face recognition is being discussed from a human's point of view - how does a human recognize a face in social situations. Functional magnetic resonance imaging (fMRI) studies have revealed that visual tasks play a more influential role during face analysis and

[4] Emotion stimulus will be explained in chapter 5.

recognition, where not only face-specific brain areas were involved, but a coherent neural activity was stored devoting to motion perception and gaze control [4]. This implies that in addition to static features related to shape and color, a “dynamic signature” can also be utilized to augment face recognition. For implementation, a dynamic face model consisted of multiple Hidden Markov Models (HMMs) is constructed to categorize facial expressions. Figure 5 illustrates the overall process of the authors’ approach. As a first step, video frames are clustered into x number of spatial HMMs each of which represents a facial expression, where x is determined by Bayesian Classifier. Next, emission probabilities of each cluster are trained using all the video sequences. Finally, the transition matrix and initial state probabilities are trained with respect to the temporal evolution to form a “Psuedo Hierarchical HMM (PH-HMM)”. Thus, the recognition process can be carried out by

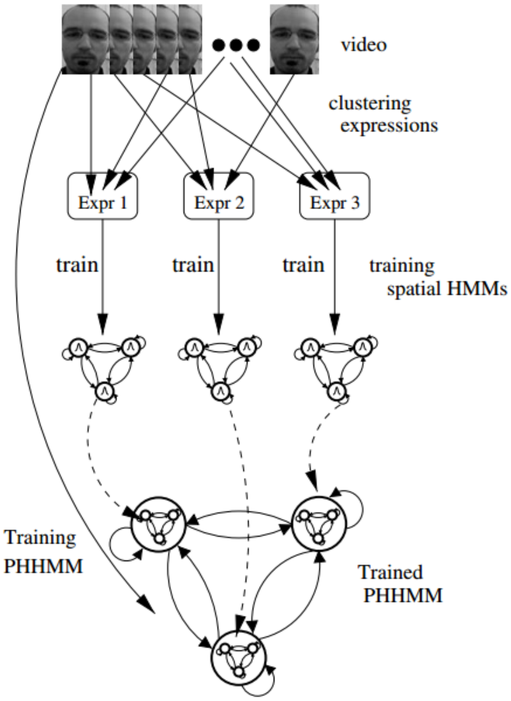


Figure 5. Multi-HMM approach for dynamic face recognition. (Courtesy of Tistarelli et al. [8])

measuring similarity between two stochastic sequences. Similar face recognition techniques based on HMM have been studied in [29], [30], [31], and [32].

It is worth noting that during clustering (as in HMM-based approach) video frames are not in time sequence, meaning that the temporal information is being ignored where each video frame is labeled independently based on a discrete number of facial expressions. Conversely to their approach, we want to model motion behavior in a continuous manner such that the "true" facial moving pattern is revealed. By utilizing the Action Units defined by FACS which was developed by psychologist and physiologist, we can characterize facial movements into discrete continuous values or sequence rather than categories.

The closest research to our approach in nature is the one conducted by Biuk and Loncaric where a pattern trajectory in the Eigen space (principle component analysis) was built from a sequence of face images rotating from -90 °to +90 °(Figure 6) [33]. The prototype trajectory was later compared against incoming samples using a distance measurement of two sequence. Nevertheless, trajectories were coordinated in the Eigen space which might not represent the actual facial movements. Additionally, distance measurement was accumulated in Euclidean, and thus would degrade rapidly with noise and sensitive to time variations [34]. Instead, our approach utilizes the Longest Common Subsequence (LCSS) method to accommodate inputs that are noisy and time-shifted.

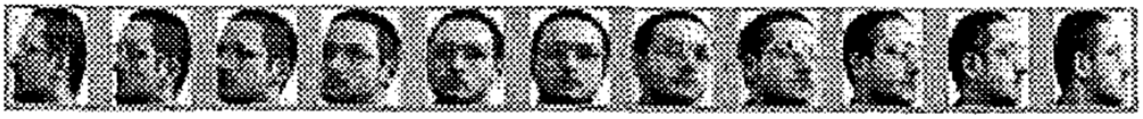


Figure 6. Face images rotating from -90 °to +90 °used for Eigen space analysis. (Courtesy of Biuk and Loncaric [33])

CHAPTER 4. METHODOLOGY

Facial movement based authentication can be decomposed into two phases: enrollment and verification. During the enrollment phase, facial movements are being collected and characterized using the famous Facial Action Coding System (FACS). 3D dynamic face recognition is then carried out using the obtained facial action units to construct a representative model consisted of sequence and matching thresholds. In the verification stage, facial action units are again collected and compared with existing legitimate models to generate a similarity score. In the following sections, we detail the proposed methods for each step.

Facial Movement Characterization

In order to utilize facial movements for authentication, it is necessary to find a way to systematically categorize facial displacements with respect to their muscular formation. After reviewing a few existing encoding system, the Facial Action Coding System (FACS) was chosen as our characterization method because of its physiological and psychological foundation. In this section, we provide a brief overview of FACS and its implementation.

Facial Action Coding System

The Facial Action Coding System (FACS) provides a way to encode facial muscles from slight different instant changes in appearance [5]. The anatomical basis suggests that the nerves controlling the muscles are instinctively linked and related to each other whereas the psychologists state that natural facial expressions are constructed under emotional feelings [35]. With the research foundation from both physiologist and psychologists, FACS has been established as a computed automated system to produce temporal profiles of each facial movement in videos. It has already been proven useful to psychologist for facial expression recognition analysis [3] [36] [37]. Since in our case we tend to focus on retrieving and matching individual action unit trajectories in sequence, FACS is converted to a verification tool rather than a recognition system.

Action Units Implementation

Action Units (AUs), as defined by FACS, are the fundamental actions of individual muscles or groups of muscles. FACS has defined 46 actions units in total, however, we have restricted our scope to six major AUs that are prominent in describing facial expressions [36] [37]. A complete list of the six AUs used in our experiment and their muscular basis can be found in Table 1 where individual muscle can be visualized[5] in Figure 7.

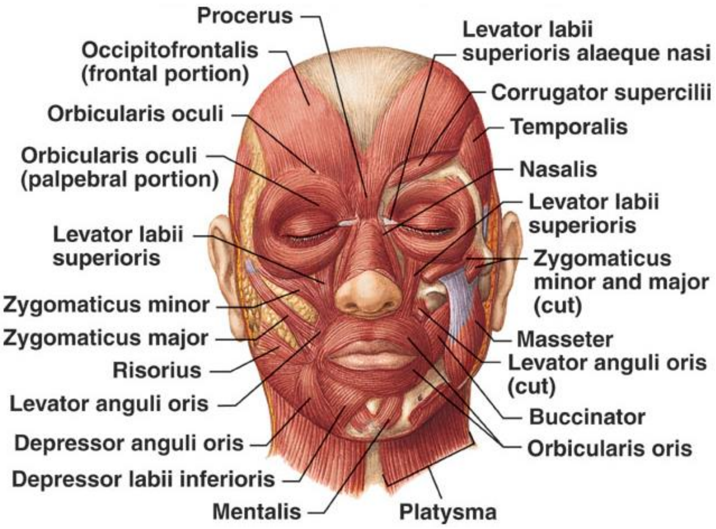


Figure 7. Muscles of Human Face.
(Courtesy of McGraw-Hill [35]**)**

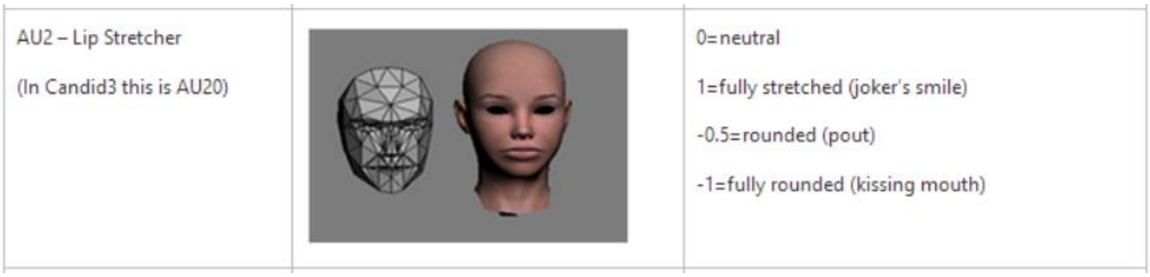
[5] Some of the muscles listed in Table 1. are not included in Figure 7.

Table 1. The six Action Units (AUs) used in our experiment and their muscular basis.

AU Name and Number	Number in FACS	Muscular Basis
Neutral Face	0	N/A
Upper Lip Raiser - 0	10	levator labii superioris, caput infraorbitalis
Jaw Lowerer - 1	26/27	masseter; relaxed temporalis, internal pterygoid - 26 pterygoids, digastric - 27
Lip Stretcher - 2	20	risorius, platysma
Brow Lowerer - 3	4	depressor glabellae, depressor supercilii, corrugator supercilii
Lip Corner Depressor - 4	13/15	levator anguli oris - 13 depressor anguli oris - 15
Outer Brow Raiser - 5	2	frontalis

We have employed the Face Tracking SDK[6] from Microsoft for extracting the six AUs from Kinect for Windows. A similar approach to [38] has been taken to recognize the six action units. In total, 87 feature points are being tracked by the Active Appearance Model (AAM) [39] utilizing a 3D morphable model called Candide-3 to characterize shape and texture with shape units (SUs) [40]. As a result, each AU is discretized into values ranging from -1 to +1, representing the displacement from a neutral expression. For instance, for the lip stretcher, -1 is interpreted as fully rounded, 0 represents neutral, whereas +1 means fully stretched. Figure 8 illustrate the lip stretcher animated on the Candide-3 model. A complete list of SUs and AUs being used can be found at Appendix A. Therefore, for face recognition,

[6] More details about the Face Tracking SDK can be found at http://msdn.microsoft.com/enus/library/jj130970.aspx



AU2 – Lip Stretcher (In Candid3 this is AU20)		0=neutral 1=fully stretched (joker's smile) -0.5=rounded (pout) -1=fully rounded (kissing mouth)

Figure 8. The lip stretcher animated on the Candide-3 model. (Courtesy of Microsoft Face Tracking SDK)

a template and a sample can be obtained by acquiring a sequence for each AU described above. In the next section, we focus on modeling such sequence and measuring the similarities between them to suit our need for authentication.

3D Dynamic Face Recognition

By representing facial movements using six AUs defined by FACS, we can obtain six AU sequence to represent the identity model. The problem of face recognition is thus relaxed into a pattern recognition problem which usually involves matching two sequence [41]. In the following section, we briefly discuss the benefits and practical use of 3D, and then a detailed explanation of our methods will be followed.

3D

Intuitively, depth information adds robustness to face recognition by representing the real facial appearance. Since head-orientation may occur, the third dimension helps approximate frontal view of the face to compensate 2D-specific problems such as pose and illumination. Before the emergence of affordable and accurate 3D scanners, numerous efforts have been made to recover the depth by either triangulation [42] or perception [43] [44].

However, it is clear that neither of these methods can be compared to a 3D camera which is capable of obtaining depth and color streams at the same time.

In our experiment, the Kinect camera obtains the depth data by performing the following steps: 1) an IR emitter projects a random speckle pattern to be perceived by the IR receiver; 2) a reference plane reflecting the true distance to the camera plane is generated in the center surrounded by object planes; 3) a triangulation between the reference and the object plane is carried out to calculate the distance from point to point [45]. Figure 9 illustrate Kinect's depth sensing capability. It is worth to mention that several environmental factors have to be considered when using depth data obtained in this way. Firstly, depth accuracy will degrade with distance, thus a working depth range should be controlled between 1m to 1.5m to allow effective recognition. Secondly, since depth measurements are most accurate on the reference plane, user's face should be guided to fit entirely inside a centering window. Thirdly, due to the reflective nature of IR, objects absorbing IR lights should be removed on user's face to ensure an accurate reading and also eliminate occlusions.

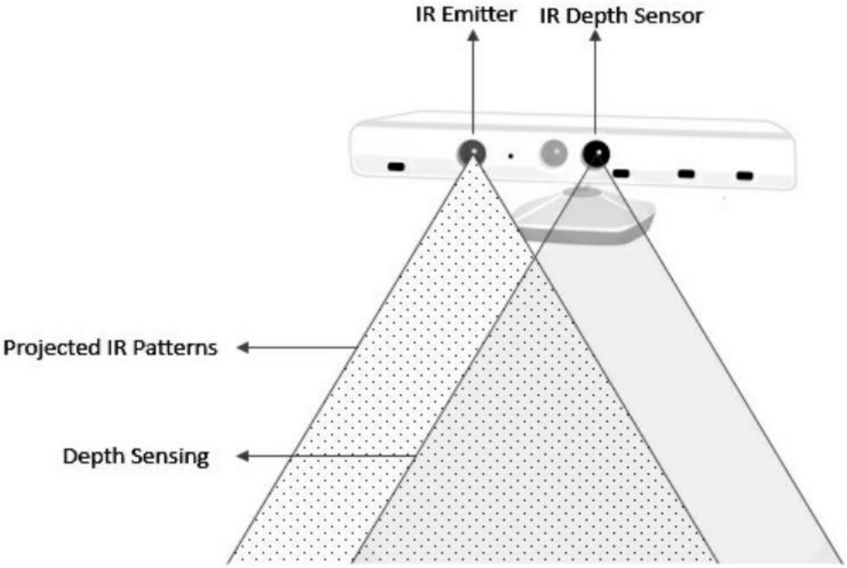


Figure 9. Kinect depth sensing.

Similarity Measurement

Similarity measurement is the heart of many sequence data mining applications and it is also applied in gait recognition techniques – another biometric authentication method based on behavioral traits – where the 3-axis acceleration sequence are matched [46]. As we have seen in Chapter 3, distance of two still face images can be estimated by performing Principle Component Analysis (PCA), however, measuring the distance between two trajectories is a different problem. In general, there are three most commonly used methods for matching two sequence [47]:

1. Euclidean Distance (ED)

   ED is a popular approach for defining similarity of sequence mainly because of its simplicity – it directly calculates the point to point distance. However, ED is prone to noise and does not accommodate variations in time phase, thus is rarely used in practical applications.

2. Dynamic Time Warping (DTW)

   DTW is a method that allows an elastic shifting of the time axis to accommodate sequences which are similar, but out of time phase. It is based on dynamic programming which was proved to be a very reliable method. However, it does not obey the triangular inequality, and again prone to noise.

3. Longest common Subsequence (LCSS)

   LCSS is more robust than DTW under noisy conditions. It focuses on the most common part of two sequences eliminating impacts of noises and outliers (outlier means data at the beginning and ending). It allows sequences to be stretched by setting the time and space thresholds without rearranging the order but allowing some features to be unmatched.

To conclude, it has been proven that LCSS is more reliable than ED and DTW when it comes to measuring similarity between two trajectories [34] [47] [48]. Figure 10 illustrates the quality matching of LCSS compared to other methods in the presence of noise. We can observe that, ED is extremely inflexible in matching and DTW performs excessive and spurious matching. Suiting our need for face authentication, LCSS has been formulated as follows:

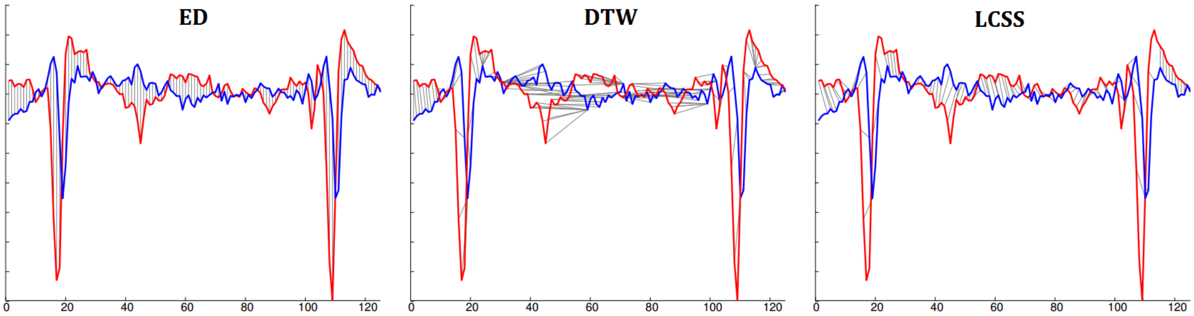


Figure 10. LCSS pattern matching compared to ED and DTW.
(Courtesy of Vlachos et al. [34])

Given an integer δ controlling flexibility in time and a real number $\varepsilon \in [0,1]$ constraining matching in space, two sequence $A = ((a_{x,1}), \dots, (a_{x,m}))$ and $B = ((b_{x,1}), \dots, (b_{x,n}))$ with size m and n, respectively, let $Next(A) = (a_{x,1}), \dots, (a_{x,m-1})$ be the function to eliminate the last element from sequence A, then $LCSS_{\delta,\varepsilon}(A, B)$ can be defined as:

$$LCSS_{\delta,\varepsilon}(A,B) \begin{cases} 0 \qquad if\ A\ or\ B\ is\ empty \\ 1 + LCSS_{\delta,\varepsilon}(Next(A), Next(B)) \\ \qquad if\ \ |a_{x,m} - b_{x,n}| < \varepsilon \\ \qquad and\ |m-n| \le \delta \\ \max(LCSS_{\delta,\varepsilon}(Next(A), B), \\ \quad LCSS_{\delta,\varepsilon}(A, Next(B)) \qquad otherwise \end{cases} \qquad \textbf{(1)}$$

Additionally, by employing the concept of dynamic programming, a hash table $H(i,j)$ where $i \in [1,m], j \in [1,n]$ can be used for storing previous LCSS results such that:

$$LCSS_{\delta,\varepsilon}(A,B)\begin{cases} H(i,j) & if\ H(i,j) \neq -1 \\ LCSS_{\delta,\varepsilon}(A,B) & otherwise \end{cases}$$

Assuming that values being stored in $H(i,j)$ are all initialized to -1, therefore, whenever $LCSS_{\delta,\varepsilon}(A,B)$ is evaluated in equation (1) previously hashed result is always checked first. Thus, the computation time of LCSS $O(\delta(m+n))$ can be greatly reduced by trading off a space complexity of $O(m * n)$. Finally, the similarity $\mathrm{S}_{A,B}$ between the two sequence A and B can be defined as follows:

$$\mathrm{S}_{A,B} = \frac{\mathrm{LCSS}_{\delta,\varepsilon}(\mathrm{A},\mathrm{B})}{\min\big(length(A), length(B)\big)} \qquad (\mathbf{2})$$

Thus the similarity score is generated by normalizing LCSS result with respect to the smaller size of two sequence.

Modeling

After the enrollment samples become available, an identity model has to be constructed to create a user profile. There are numerous ways to build the model from simply picking the first sample to averaging over the entire set, it is clear that a more robust approach has to be taken to accommodate erroneous samples. Acknowledging the fact that the identity model will be later compared with probe samples using LCSS, we have proposed a voting mechanism based on the similarity results generated by LCSS. Specifically, LCSS is performed on each sample against every other sample, the one with the highest average similarity get picked. Thus, each AU is modeled by its most representative sequence based on LCSS results.

Adaptive Time and Space Controls

In addition to storing the most representative sequence, our model also includes two LCSS control parameters: time and space, which controls the flexibility in matching two sequence. These two parameters are adaptively computed from each enrollment sample by the following two steps:

1. From each enrollment sample, find the maximum value and its time index.
2. The space control parameter ε and time control parameter δ are approximated by:

$$\varepsilon = \left(\sum_{1}^{N} \max(M_{1\ldots N}) - M_i \right) \frac{1}{N}$$

$$\delta = \left(\sum_{1}^{N} \max(T_{1\ldots N}) - T_i \right) \frac{1}{N}$$

where M_i the maximum AU value of sample i, T_i the time index of M_i within sample i, N is the number of samples used for building the model, max is a function selecting the maximum value from an array of AU values.

In other words, control thresholds are obtained by averaging time and space distances to the absolute upper bound values, and they are indeed an approximation of the real position where the most representative sample will lie on. Therefore, we can guarantee that the model is the nearest neighbor to all given samples. Figure 11 illustrates the estimated position of the chosen sequence Q bounded by time and space thresholds ε and δ where A represents an unmatched probe.

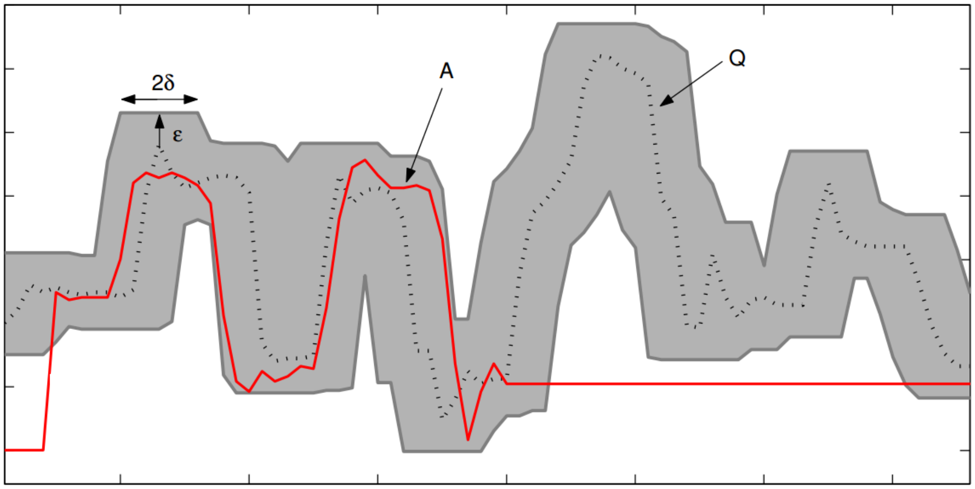


Figure 11. Estimated gallery Q lies in the center of an envelope bounded by matching thresholds ε and δ, where A is an unmatched probe. (Courtesy of Vlachos et al. [34]**)**

Combined AU Model and Result

One representative sequence is selected from each sample based on the LCSS characteristics. In total six sequence along with their own time and space thresholds are combined together to form the final model. In such a way, each AU is validated based its own model, which result in a much accurate similarity measurement. During the authentication phase, each AU model is generating a similarity for the incoming sample. It is worth noting that it is nearly impossible to distinguish the model from illegitimate samples by focusing on only one similarity. The reason is that the subjects are all performing similar tasks, thus sharing certain degree of similarity in term of behavior. In order to reveal the distinction, we have decided to fuse the similarities resulted from each AU model comparison all together to intensify dissimilarity which in turn improve false positive rate.

The fused similarity is used for making the decision. The similarities are being fused using the weighted sum model to exploit priori knowledge such as which facial expression was performed to generate the AUs. Since different AUs may respond differently to various expressions, a greater weight can be assigned to certain AUs. For instance, if the AUs were acquired by smiling, then the lip stretcher and eyebrow raiser are given more weights since their values change significantly during smiling. Therefore, the authentication decision model can be seen in the following equation:

$$Score = \sum_{i=1}^{6} w_i a_i$$

where w_i denotes the relative importance weight of each AU and a_i indicates the similarity result obtained by matching each pair of AU probe and gallery.

This approach is analogous to multimodal biometric authentication where a combination of biometrics is used to for authentication and can usually lead to an improved recognition rate [21]. Ultimately, all samples used for enrollment are evaluated using the above decision model and the lowest score obtained is added to the model as the decision threshold Θ such that:

$$\begin{aligned} if \quad & Score\ (probe) \geq \theta \\ then \quad & probe \cong gallery \end{aligned}$$

That is, if an incoming probe sample's similarity score is greater than or equal to gallery's decision threshold Θ then they are considered similar, thus authorizing the user for security access.

To summarize, after modeling user profile will contain:

1. Six most representative sequences for each AU
2. Time and space threshold
3. Similarity decision threshold

CHAPTER 5. USE SCENARIOS

In this chapter, data acquisition methods and an authentication scheme are proposed to apply dynamic face recognition for authentication purpose. At the end, we provide a theoretical evaluation against potential adversaries to offer some insight on the security of our approach.

Data Acquisition Method

Since the performance of our authentication system relies heavily on the quality of facial movement inputs, effective data collection methods have to be considered. In the following section, we have proposed three methods to generate the desired data. During the enrollment phase, the user will be instructed to perform one of the following tasks: 1) imitating exemplar photos; 2) watching hilarious videos; 3) pronouncing words, all of which are meant to reveal facial movements either artificially (1 and 3) or naturally (3). A brief discussion on the benefits and drawbacks of each task will be provided following method description.

Imitation-based Design

Imitation-based method was employed used by Shichuan at el. in their research of compound facial expressions [3]. Figure 12 shows one of the compound facial expressions – happily disgusted which is a result of two emotion stimulus: happy and disgust. Exemplars were provided to help the participant produce a facial expression. The author did emphasize that the participant was not instructed to look exactly the same as exemplar photos, but rather encouraged to express each emotion as clear as possible. In this case, basic emotions (expressions) are chosen to simplify the imitations.

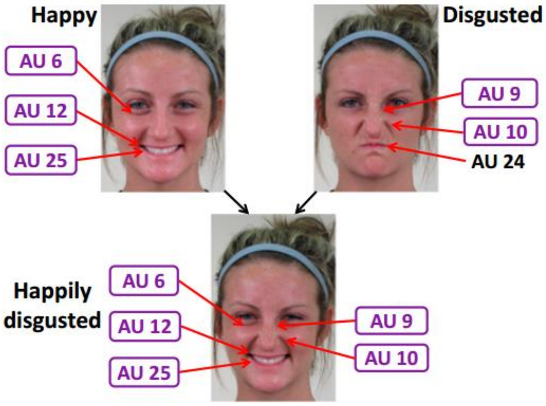


Figure 12. An example of compound facial expression – happily disgusted. (Courtesy of Shichuan at el. [3]**)**

However, each facial expression that had been derived from the participants was solely relying on their ability to perform. An imaginary situation was suggested and a sample picture was given for each expression but, predictably, many participants in this experiment failed to clearly express the emotion as the muscle changes were not a result of instinctive reaction but rather forcefully conveyed. Acknowledging the fact that most of us cannot make a particular facial expression without emotional stimulus, except for actors, this method is retained from practical use of authentication system.

Video-based Design

In order to capture facial expressions, a corresponding stimulus has to present. For instance, in order for an identity to feel happily disgusted, probably the identity needs to watch a comedy that is disgusting to some extent. Thus, in this design we propose to play videos from different emotional category for generating corresponding facial expressions used for authentication. A typical authentication scenario would be: whenever an identity is claimed, a series of hilarious videos are played to stimulate a nature facial reaction on the identity’s face. Only in this way can we guarantee that the true facial behavior is being captured by the camera.

Nonetheless, it may be viable if the expression has only to be obtained once, however, this method may not be applicable to an authentication system - where the same expressions have to be repeated whenever the identity is claimed. Since the identity will eventually become accustomed to certain stimulus, it would be unrealistic to explore new ones to arouse the same emotion again. Not to mention that, finding a stimulus for someone to show a specific facial expression itself is a challenging task since every person may react differently to each stimulus.

Pronunciation-based Design

As an alternative to the previous two methods, we have proposed a method based on pronunciation which similar to the data collection mechanism in voice authentication techniques [49]. We have examined each of the 26 English letters to see which ones will result in most significant chances in facial muscles. Each letter is being pronounced repeatedly, and is compared against each other. As expected, vowel sounds are the most significant contributor to muscle changes during pronunciations whereas consonants have little impact on facial displacement. The 26 letters can therefore be classified into groups with respect to their vowel sounds. For instance, the letters “A”, “H”, “J”, “K” can be grouped together since their pronunciation involve the vowel sound **/ei/**. The complete groups for seven vowels can be seen in Table 2.

Table 2. Grouping 26 English letters by their vowel sounds.

Vowel	Letters having the same vowel
/ei/	A, H, J, K
/i:/	B, C, D, E, G, P, T, V, Z
/e/	F, L, M, N, S, X, Z
/ai/	I, Y
/əƱ/	O
/u:/	Q, U, W
/a:/	R

Ideally, facial muscle movements can be revealed by letting the user pronounce a letter from the 7 vowel categories; however, several drawbacks can be easily observed when it comes to practical data collection. Firstly, single vowel sound will only generate facial displacement momentarily making it hard to be captured in real-time. Secondly, it is difficult to go back to neutral state while repeating a single vowel sound, thus it may result in residuals in the "move back" stage pushing the next pronunciation to an undefined initial state. Therefore, a combination of consonant and vowel sound is proposed to relax this problem. We have decided to use "WE" in our experiment since the sound of /w/ will likely reset the face to a known state before each pronunciation whereas "E" being the major contributor to muscle changes. Our experiment result, which will be shown in Chapter 6, supports our statement made above.

Authentication Scheme

A typical authentication scheme is composed of two phases: enrollment and verification. In the following sections, we uncover the methods we have proposed for each phase followed by a discussion on the potential attacks against our design.

Enrollment

During the enrollment phase, the user is asked to follow instructions in a restrictive manner. Acknowledging the environmental factors we had mentioned in Chapter 4, the registration process is carried out in a relatively dark environment to ensure that the intensity of depth sensor is not being affected. The user needs to remove any objects on their face, such as glasses and hats, to eliminate occlusions and reflection. The user is guided to move accordingly to fit the entire face in a rectangle, which is chosen to guarantee an optimal accuracy for the depth measurement. Finally, the user is instructed to perform one of the tasks as described in the data collection section. In the subsequent steps, facial action units are computed from each video frame resulting in a number of temporal trajectories. Based on best average similarities against the other trajectories, one of these trajectories is determined to represent the identity model to guarantee that it is the nearest neighbor to all given trajectories. Additionally, the time and space matching thresholds used for LCSS similarity measurement are adaptively computed from enrollment trajectories and are stored along with the model. The decision threshold is determined from the lowest similarity score.

Verification

When it comes to making an authentication decision, the user is again asked to perform one of the data gathering tasks as described in the previous section. A number of AU sequence are extracted and compared with existing legitimate models using LCSS with the time and space threshold stored in the models. Similarity results are generated and a combined matching score is calculated and compared with the decision threshold to determine the final decision.

Adversary Model

Regardless of the performance, authentication systems are of no practical use if they are vulnerable to spoofing attacks. In order to make it more reliable and secure, the circumvention factor has to be evaluated. In this section, we attempt to evaluate our proposed methods against possible attacks, potential hazards are provided at the end.

Potential attacks against face authentication system can be broadly divided into two types: indirect attack and direct attack. Indirect attack or software attack requires priori knowledge of the system component but can be powerful if deliberately implemented. Efficient software attack has been proposed by Javier et al. to evaluate the vulnerability of face authentication systems [50]. Input facial images were synthesized statistically and the hill-climbing algorithm was adopted to mathematically calculate a converging similarity score. However, the authors assumed to have the access to the evaluation score of the matching function, which is normally not exposed while system is running. Thus, we focus on discussing another type of attack – direct attack.

Direct attack or spoofing attack is the most commonly used method, which aims at fooling the camera with a face replica of the legitimate user. We have already seen in Chapter 3 that previous liveness detection approaches against spoofing attacks may fail under certain conditions. However, our method is invulnerable to spoofing attack because of the following two reasons:

1. Facial movements are exploited for authentication, thus preventing photo playbacks.
2. Depth data are utilized during recognition, therefore forbidding video playbacks.

Traditional playback attacks can be efficiently prevented by employing dynamic features but facial movements still can be played back to the camera if a video is recorded. That is one of the reasons why the third dimension is added since it is nearly impossible to replay 3D facial movements. It is worth noting that if we have only utilized 3D static features, the adversary could still possibly produce a 3D facial mask to spoof our system [1]. By combining 3D and face dynamics together, the secureness of our face authenticator can be greatly increased.

Theoretically, our authenticator can be bypassed using both a replica of facial movements and a replica of depth. One way is to animate facial expression on a face sculpture made with artificial muscle tissues such that an actual duplicate of facial appearance is created. In this case, emotion stimulus (video-based design) can be employed to countermeasure such attack acknowledging the fact that facial expressions stimulated by emotion are intrinsically linked with underlying neural, which cannot be easily replicated. However, the video-based design is left as a future work given the difficulty to stimulate emotions repeatedly. Another potential attack should aim at fooling the depth sensor such that a video-based attack can be carried out with depth replica. Consider, for example, the most commonly used infrared (IR) sensor which measures depth either by a process of triangulation or observing structured lights. In either way, it is possible to deflect or even absorb the infrared lights and replay the depth data by projecting artificial lights from the same direction of the camera. We believe that this type of attack can be potentially prevented if the verification environment is strictly controlled.

CHAPTER 6. EVALUATION AND RESULT

Experiments had been conducted to evaluate our methods. The experiments were carried out with the Kinect camera and the Face Tracking SDK. To begin with, we investigated the action units generated by making a facial expression. Next, ten participants were invited to undergo the proposed authentication scheme. Similarities between the sequences obtained from the same subject were measured and examined using LCSS. Finally, similarity scores were generated with one gallery subject and nine probe subjects. Result are summarized following each section.

Action Unit Analysis

Six action units (AUs) were considered in our experiment with each AU expressed as a numeric value varying between -1 and +1. These coefficients represent deformations of the 3D mask caused by the moving parts of the face (mouth, eyebrows, and so on). Experiments were first conducted to evaluate action units' responses to a facial expression. Using the pronunciation-based design we have proposed for data collection, two types of samples were generated by making the sound "E" and "WE". Each pronunciation process was carried out from the neutral state to fully stretched state (Figure 13). Figure 14 presents the sample data collected by pronouncing each word twice with a one second delay in between.

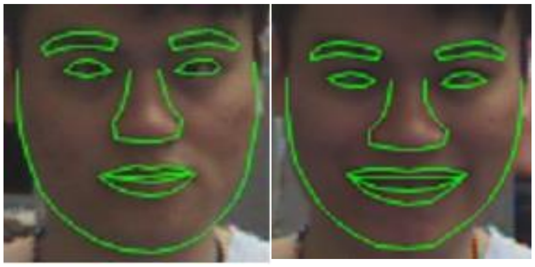


Figure 13. Pronunciation of "E" with superimposed feature points. On the left is the neutral face whereas on the right is the fully stretched face.

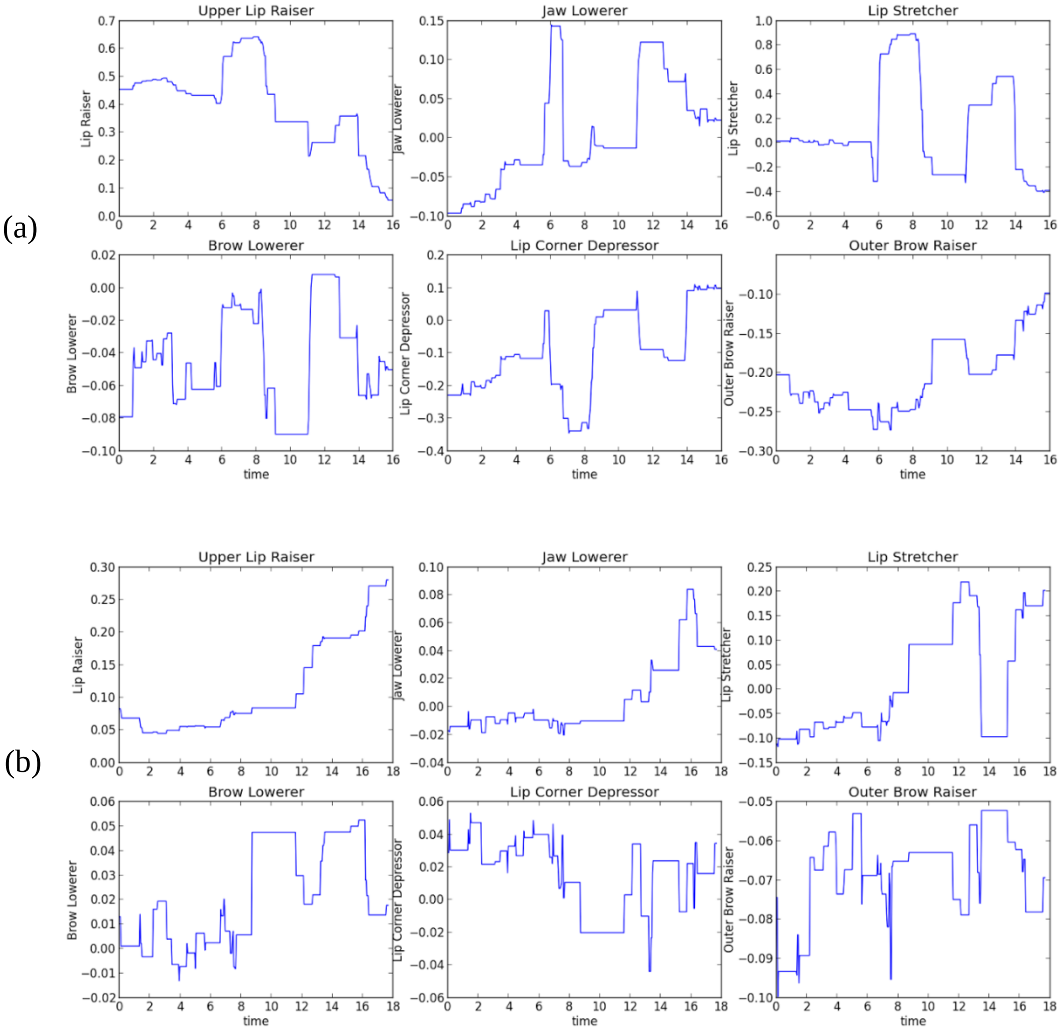


Figure 14. Plots of six AU numeric weights with respect to the time frame in response to the pronunciation of (a) "WE" and (b) "E". Note the difference between two lip stretchers, which is the major contributor to facial muscle changes.

By observation, we can see that for both of the pronunciation the lip stretcher served as the major contributor to facial muscle displacements whereas the rest of the AUs resulted in a less significant response. Thus the obtained result corresponds to our initial statement made in Chapter 4 where a greater weight could be assigned to certain AUs based on the prior knowledge of the performed facial expressions.

Furthermore, it is clear that the sound “WE” generated much cleaner and useable AUs as opposed to “E” where a residual was remained prior to neutral state. Thus, the experiment result complies with our assumption made earlier in Chapter 5 where a combination of consonant and vowel sound delivers a more quality result compared to pronouncing the vowel sound itself.

Finally, we observe that there exist some noises in the plots:

1. We have outliers which are either smaller or large than the average values.
2. Time index varies from one sample to another.

These types of noises were either introduced by face tracking or AU extraction or simply due to the fact that most of the subjects could not perform the same facial expression consistently. However, we do observe that a certain pattern can be pulled out from these samples to represent an identity as long as the method we use for measuring similarity – LCSS – is capable of accommodating the two forms of noises mentioned above.

Authentication

Ten subjects were invited to participate in our experiment. Each subject was asked to pronounce the word “WE” 30 times where each pronunciation was carried out within 4 seconds. Including the time spent on computations, 26 frames per second was achieved. Figure 15 provides an overview of the program flow we had proposed for authentication experiments. During the enrollment phase, the Kinect camera was first initialized for color and depth stream acquisition. Next, the Face Tracking thread was initiated to actively learn the shape units (SUs) from a user’s face. Once the SUs had converged, which might take up to two minutes, the enrollment process was started to repeatedly collect AU sequence with a

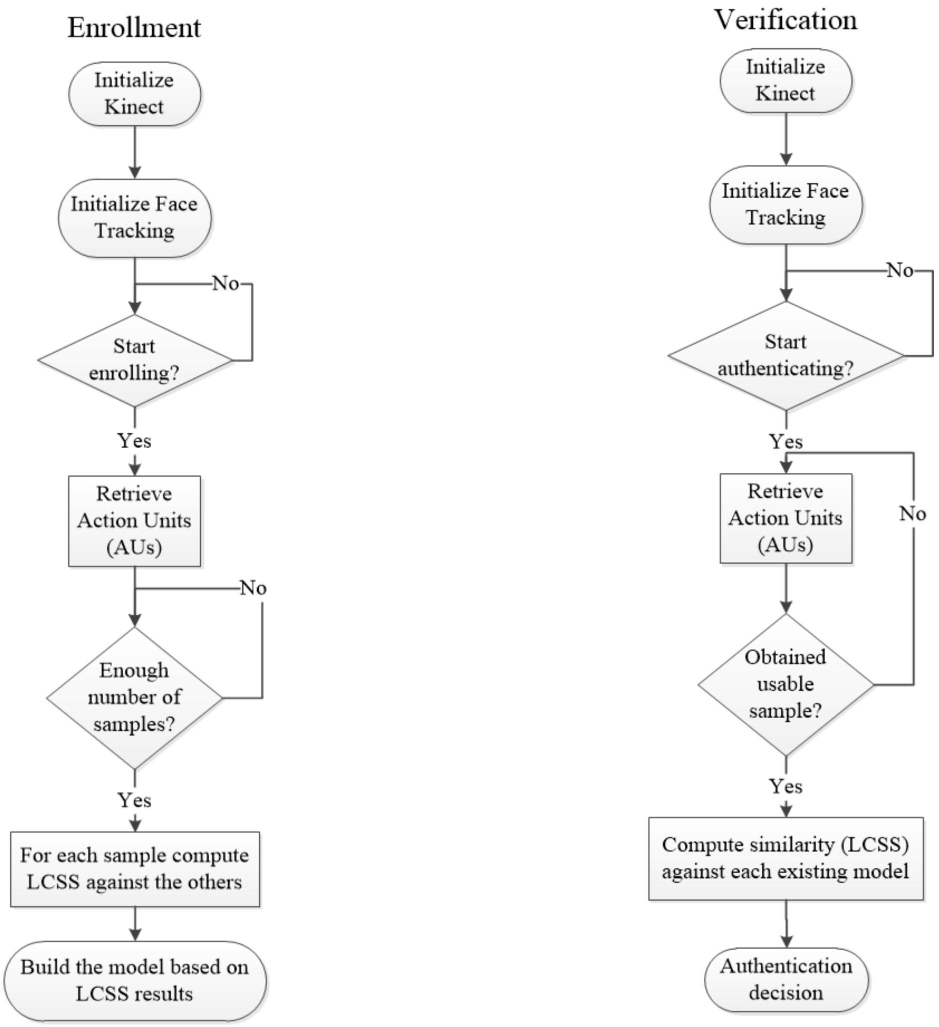

Enrollment
Initialize Kinect
Initialize Face Tracking
No
Start enrolling?
Yes
Retrieve Action Units (AUs)
No
Enough number of samples?
Yes
For each sample compute LCSS against the others
Build the model based on LCSS results
Verification
Initialize Kinect
Initialize Face Tracking
No
Start authenticating?
Yes
Retrieve Action Units (AUs)
No
Obtained usable sample?
Yes
Compute similarity (LCSS) against each existing model
Authentication decision

Figure 15. Authentication experiment program flow.

4 seconds interval. After retrieving enough number of sequence samples, a subset of the samples was used for building the model whereas the remaining ones were retained for validation. For model construction, the LCSS time and space thresholds were first calculated based on average distance to the maximum values, then for each candidate sample, LCSS was computed against every other candidate to generate a similarity matrix. Finally, considering the best average similarity, one most representative sequence was stored with time and space thresholds to characterize the identity model. Figure 16 shows the selected gallery sequence for the Lip Stretcher along with other candidate sequences.

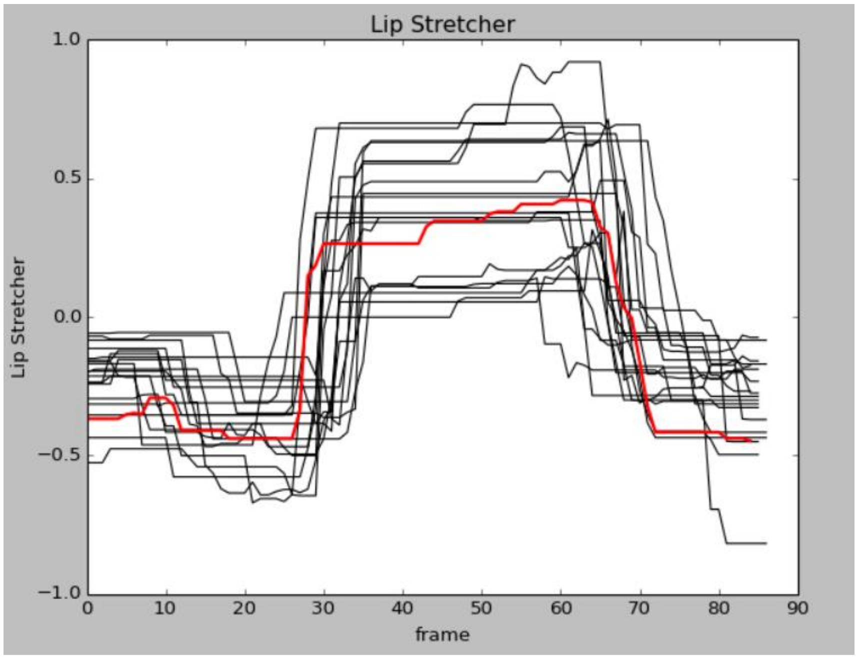


Figure 16. The selected gallery sequence for the Lip Stretcher.

Similar to the enrollment phase, the verification phase was proposed for collecting probe samples. However, since we had decided to perform cross validation, probe samples were already obtained during the enrollment phase. Therefore, the verification stage was carried out completely offline to effectively evaluate our algorithm. For each probe sample, six different AU sequence were extracted and compared against existing gallery models to generate six similarity results. The similarity results were being fused using the weighted summation model to exploit priori knowledge of performed facial expression. In this case, since the AUs were acquired by pronouncing the word “WE” (similar to smiling), lip stretcher was assigned more weight.

Figure 17 displays the final result of our experiment. In this experiment, subject 1 was chosen to be the gallery which was also compared against itself with its own 10 validation samples. We can observe that similarities above 0.85 were achieved for subject 1 whereas most of the other subjects exhibited similarities below 0.85 (except for sample 3 of subject 6

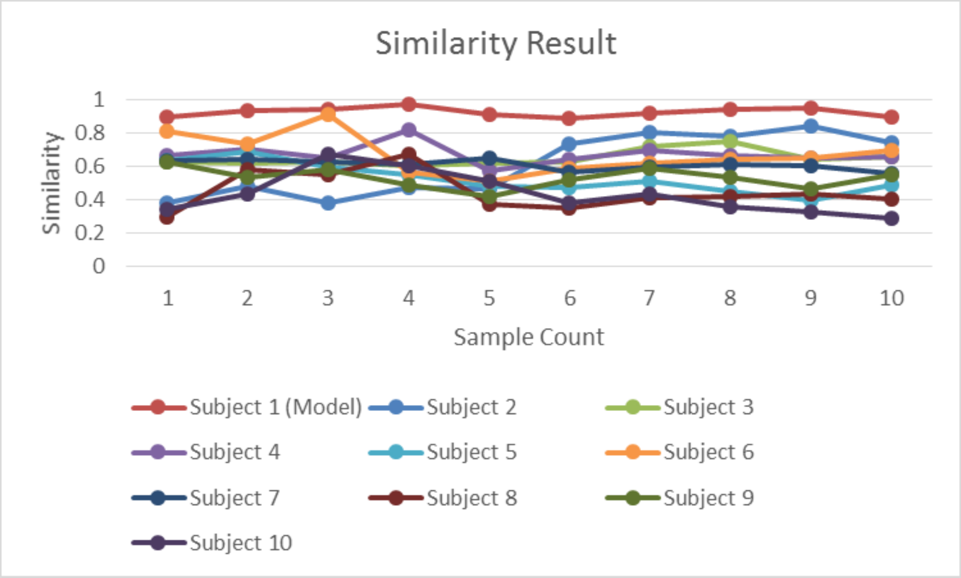


Figure 17. Similarity result for ten subjects. Subject 1 was chosen as the gallery.

which was a false positive). Since decision threshold was set by the model, which was 0.85, thus we could authenticate subject 1 and reject the others effectively.

Cross validations were performed on all the samples with each subject selected to be the gallery model respectively. The evaluation result has been presented in Table 3. Overall, a 95% recognition rate was achieved with a 7% false positive rate.

Table 3. Experiment evaluation result.

Authentication Decision		***Test Condition***	
		NO	**YES**
Test Outcome	**YES**	7% (False positive)	95%
	NO	93%	5% (False negative)

Discussion

Down Slope Sampling

Since the subjects could intentionally control their facial muscle during up slope process, we believe that up slope (neutral to maximal) facial movements might be similar for all subjects whereas down slope (maximal to neutral) ought to be distinctive. Experiment had been conducted using only the down slopes for sampling and modeling. However, no better result than using the entire sequence as a whole was observed. It was mainly because the down slope happened too fast that the camera was not able to catch the full details. In our experiment, the down slope lasted for only 5 to 6 frames for most of the samples collected which is approximately 0.2 second (since the program could achieve up to 26 frames per second).

Real-time Authentication

We had proposed a real-time verification scheme but it was not implemented. We believe that once our approach become more robust and mature after offline validations, a fully automatic authentication scheme can be easily developed later on.

CHAPTER 7. SUMMARY AND FUTURE WORK

Summary

Considering emerging security concerns of currently existing facial recognition techniques and also inspired by the work of physiologists and psychologists on the use of facial movement for face perception, we had proposed an authentication method based on spatiotemporal facial patterns. We have demonstrated that our method is able to satisfy the five requirements (2.1 – 2.5) as mentioned in Chapter 2.

In chapter 3, we included a literature review for existing researches. An overview of authentication had been given at the beginning to compare and evaluate different methods. Subsequently, we had shown that existing liveness detection techniques might fail to counter measure spoofing attacks under various conditions. Finally, we indicated that previous dynamic approaches only aimed at improving recognition rate which might not respect the true facial behavior.

In chapter 4, we detailed the proposed methods for our approach. Facial movements were encoded by the Facial Action Coding System (FACS) where each action unit (AU) characterized a set of anatomically related facial muscles. We briefly covered the benefits and practical use of the third dimension. For authentication, an identity model is constructed by extracting multiply AU sequence from enrollment samples, and one most representative sequence was selected for each AU based on best average LCSS similarity results. Along with the chosen sequence, time and space thresholds for LCSS matching were also adaptively computed and stored in the model. Similarity between the gallery and probe was computed by LCSS, and the final decision is determined by fusing similarities results produced by all AUs.

In chapter 5, we delivered the use scenario for authentication. Three data acquisition methods had been proposed and an authentication scheme was given. We had explained how to prevent impersonation attacks by employing 3D dynamic features. An evaluation on potential hazards were also given to provide some insight on the security of our approach.

In chapter 6, we explained the experiments that had been conducted. The proposed pronunciation-based design was implemented as the data acquisition method to evaluate the AU data collected. Sequence similarities were then measured on the same subject to evaluate LCSS results. Finally, ten subjects were invited to participate in our experiment, and the results had shown that our method was able to differentiate user identities based on facial behavior.

Future Work

For experimental purpose, only six action units were considered in this thesis. A more robust approach can be taken by utilizing all 46 action units as described in the FACS. The combinational model formed by all the action units can then represent the universal facial behavior for each individual, thus leading to a more secure and non-replicable identification.

We have demonstrated data acquisition using the pronunciation-based design, however we believe that video-based design is more desirable given the fact that facial expressions stimulated by emotion stimulus represent the true facial behaviors. Further investigation has to be made considering the complexity to discover differing emotion stimulus for repetitive authentication.

One pitfall can be foreseen in long run is that user may not be able to gain access to the system due to changes in facial appearance, i.e. aging. This problem can be potentially prevented by performing a periodic profile refreshment every time the user is authenticated, thus adapting the identity model gradually.

Voice recognition provides an alternative to face recognition. Since we have employed an authentication scheme which involves pronunciations, voice recognition can be easily adopted to our system to serve as an alternative - in case the face recognition does not perform well under certain circumstances - or even combined with facial features to compose a multi-modal biometric authentication system. Although it does make implementation of the system more complicated, it is believed that accuracy and false positive rate can be effectively improved since the errors will be uniformly distributed among each feature metrics and thus resulting in a smaller error rate after a weighted summation. It is worth to mention that the Federal Bureau of Investigation (FBI) is working towards the Next Generation Identification System[7] (NGI) which combines multiple biometric features together for criminal identification.

[7] More details about the Next Generation Identification System (NGI) can be found at http://www.fbi.gov/aboutus/cjis/fingerprints_biometrics/ngi

REFERENCES

[1] B. Biggio, Z. Akhtar, G. Fumera, G. Marcialis and F. Roli, "Security evaluation of biometric authentication systems under real spoofing attacks," *Biometrics, IET,* vol. 1, no. 1, pp. 11-24, 2012.

[2] B. KNIGHT and A. JOHNSTON, "The role of movement in face recognition," *Vision cognition,* vol. 4, pp. 265-274, 1997.

[3] S. Du, Y. Tao and A. M. Martinez, "Compound facial expressions of emotion," *PNAS,* 2014.

[4] J. Haxby, E. Hoffman and M. Gobbini, "The distributed human neural system for face perception," *Trends in Cognitive Sciences,* vol. 4, no. 6, pp. 223-233, 2000.

[5] P. Ekman and W. Friesen, Facial Action Coding System, Palo Alto: Consulting Psychologists Press, 1978.

[6] S. Chakraborty and D. Das, "An overview of face liveness detection," *International Journal on Information Theory,* vol. 3, no. 2, pp. 11-25, 2014.

[7] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, "Face recognition: A literature survey," *ACM Compuer Survey,* vol. 35, no. 4, pp. 399-458, 2003.

[8] M. Tistarelli, M. Bicego and E. Grosso, "Dynamic face recognition: From human to machine vision," *Image and Vision Computing,* vol. 27, no. 3, pp. 222-232, 2009.

[9] Y. Li, *Dynamic face models: construction and applications,* Ph.D. thesis, Queen Mary, University of London, 2001.

[10] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 14, no. 1, pp. 4-20, 2004.

[11] M. Bishop, "Authentication," in *Introduction to Computer Security*, Addison Wesley Professional, 2004, pp. 187-212.

[12] R. Morris and K. Thompson, "Password security: A case history," *Communications of the ACM,* vol. 22, no. 11, pp. 594-597, 1979.

[13] D. L. Jobusch and A. E. Oldehoeft, "survey of password mechanisms," *computer security,* vol. 8, no. 8, pp. 675-689, 1989.

[14] L. O’Gorman, "Comparing Passwords, Tokens, and Biometrics," *In proceedings of the IEEE*, vol. 91, no. 12, pp. 2021-2040, 2003.

[15] X. Suo, Y. Zhu and G. S. Owen, "Graphical passwords: a survey," *Proceedings of the 21st Annual Computer Security Applications Conference,* pp. 463-472, 2005.

[16] "A Survey on Recognition-Based Graphical User Authentication Algorithms," *International Journal of Computer Science and Information Security*, vol. 6, no. 2, pp. 119-127, 2009.

[17] F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication," *Future Generation Computer Systems,* vol. 16, pp. 351-359, 2000.

[18] R. V. Yampolskiy and V. Govindaraju, "Behavioural biometrics: a survey and classification," *International Journal of Biometrics,* vol. 1, no. 1, pp. 81-113, 2008.

[19] J. J. Kavanagh and H. B. Menz, "Accelerometry: A technique for quantifying movement," *Gait & Posture,* vol. 28, no. 1, pp. 1-15, 2008.

[20] D. Gafurov, E. Snekkenes and P. Bours, "Spoof Attacks on Gait Authentication System," *IEEE Transactions on Information Forensics and Security,* vol. 2, no. 3, pp. 491-502, 2007.

[21] A. Ross and A. K. Jain, "Multimodal Biometrics: An Overview," *12th European Signal Processing Conference,* pp. 1221-1224, 2004.

[22] G. Kim, S. Eum, J. K. Suhr, D. I. Kim, K. R. Park and J. Kim, "Face liveness detection based on texture and frequency analyses," in *5th IAPR International Conference on Biometrics*, New Delhi, 2012.

[23] N. B. G, S. M. Hatture, M. S.Gabasavalgi and R. P. Karchi, "Liveness Detection Technique for Prevention of Spoof Attack in Face Recognition," *International Journal of Emerging Technology and Advanced Engineering,* vol. 3, no. 12, pp. 627-633, 2013.

[24] A. da Silva Pinto, H. Pedrini, W. Schwartz and A. Rocha, "Video-Based Face Spoofing Detection through Visual Rhythm Analysis," in *25th SIBGRAPI Conference on Graphics, Patterns and Images*, Washington, DC, USA, 2012.

[25] W. Bao, H. Li, N. Li and W. Jiang, "A liveness detection method for face recognition," in *International Conference on Image Analysis and Signal Processing*, Taizhou, 2009.

[26] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *International Joint Conference on Biometrics*, Washington, DC, 2011.

[27] K. Kollreider, H. Fronthaler and J. Bigun, "Non-intrusive liveness detection by face images," *Image and Vision Computing,* vol. 27, no. 3, pp. 233-244, 2009.

[28] G. Pan, Z. Wu and L. Sun, "Liveness detection for face recognition," in *Recent Advances in Face Recognition*, Vienna, Austria, I-Tech, 2008, p. 236.

[29] F. Tsalakanidou, S. Malassiotis and M. Strintzis, "Face localization and authentication," *IEEE Transactions on Image Processing,* vol. 14, no. 2, pp. 152-168, 2005.

[30] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov model," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[31] A. V. Nefian, "An embedded HMM-based approach for face detection and recognition," in *IEEE International Conference on the Acoustics, Speech, and Signal Processing*, Washington, DC, 1999.

[32] A. Nefian and M. H. III, "Hidden Markov Models for Face Recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Seattle, Washington, 1998.

[33] Z. Biuk and S. Loncaric, "Face recognition from multi-pose image sequence," in *2nd International Symposium on Image and Signal Processing and Analysis*, Pula, 2001.

[34] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance," in *the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, 2003.

[35] D. N. Shier, J. L. Butler and R. Lewis, "Muscular System," in *Hole's essentials of human anatomy and physiology*, McGraw-Hill, 2011, pp. 176-213.

[36] A. Kapoor, Y. Qi and R. Picard, "In Fully Automatic Upper Facial Action Recognition," in *IEEE international workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[37] S. Lucey, A. Ashraf and J. Cohn., "Investigating spontaneous facial action recognition," in *Face Recognition*, InTECH Education and Publishing, 2007, pp. 275-286.

[38] Y. Sun, M. Reale and L. Yin, "Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition*, Amsterdam, 2008.

[39] M. Zhou, L. Liang, J. Sun and Y. Wang, "AAM based face tracking with temporal matching and face segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, 2010.

[40] J. Ahlberg, "CANDIDE-3 – an updated parameterized face," Dept. of Electrical Engineering, Linköping University, Sweden, 2001.

[41] R. Chellappa, A. Veeraraghavan and G. Aggarwal, "Pattern Recognition in Video," in *International Conference on Pattern Recognition and Machine Intelligence*, Kolkata, INDIA, 2005.

[42] A. M. Bronstein, M. M. Bronstein and R. Kimmel, "Expression-invariant 3D face recognition," in *international conference on Audio- and video-based biometric person authentication*, 2003.

[43] A. Saxena, J. Schulte and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *international joint conference on artifical intelligence*, San Francisco, CA, 2007.

[44] A. K. R. Chowdhury and R. Chellappa, "Face reconstruction from monocular video using uncertainty analysis and a generic model," *Computer Vision and Image Understanding - Special issue on Face recognition,* vol. 91, no. 1-2, pp. 188 - 213, 2003.

[45] K. Khoshelham and S. O. Elberink, "Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications," *Sensors,* vol. 12, pp. 1437-1454, 2012.

[46] D. Ioannidis, D. Tzovaras, I. G. Damousis, S. Argyropoulos and K. Moustakas, "Gait Recognition Using Compact Feature Extraction Transforms and Depth Information," *IEEE transactions on information forensics and security,* vol. 2, no. 3, pp. 623 - 630, 2007.

[47] Z. Zhang, K. Huang and T. Tan, "Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes," in *International Conference on Pattern Recognition*, Washington, DC, 2006.

[48] H. Liu and M. Schneider, "Similarity measurement of moving object trajectories," in *ACM SIGSPATIAL International Workshop on GeoStreaming*, New York, NY, 2012.

[49] S. Institute, "Shedding some light on voice authentication.," 2003.

[50] J. Galbally, C. McCool, J. Fierrez, S. Marcel and J. Ortega-Garcia, "On the vulnerability of face verification systems to hill-climbing attacks," *Pattern Recognition,* vol. 43, no. 3, pp. 1027-1038, 2010.

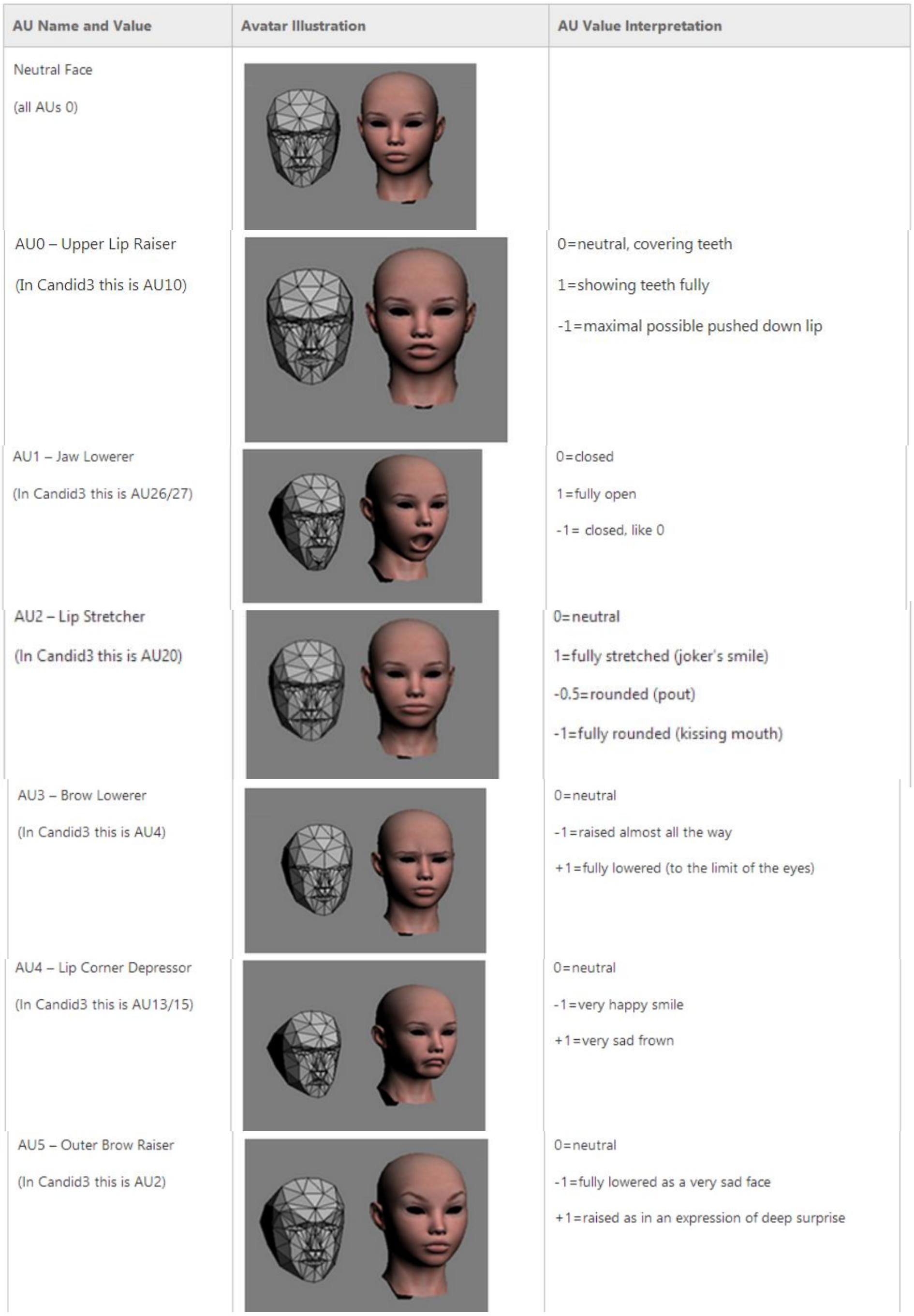
APPENDIX A. SHAPE UNITS AND ACTION UNITS

Shape units (SUs) define shape parameters of a neutral face - eyes, mouth, nose, thus remain unchanged over time. The Face Tracking SDK tracks the following 11 SUs (Table 4.). Each SU specifies the vertices it affects and the displacement (x,y,z) per affected vertex. In order to guarantee an accurate measurement of the SUs, the Face Tracking SDK needs to learn the SUs in real-time which takes approximately 2 minutes.

Table 4. The complete list of SUs.

SU Name	SU number in Candide-3
Head height	0
Eyebrows vertical position	1
Eyes vertical position	2
Eyes, width	3
Eyes, height	4
Eye separation distance	5
Nose vertical position	8
Mouth vertical position	10
Mouth width	11
Eyes vertical difference	n/a
Chin width	n/a

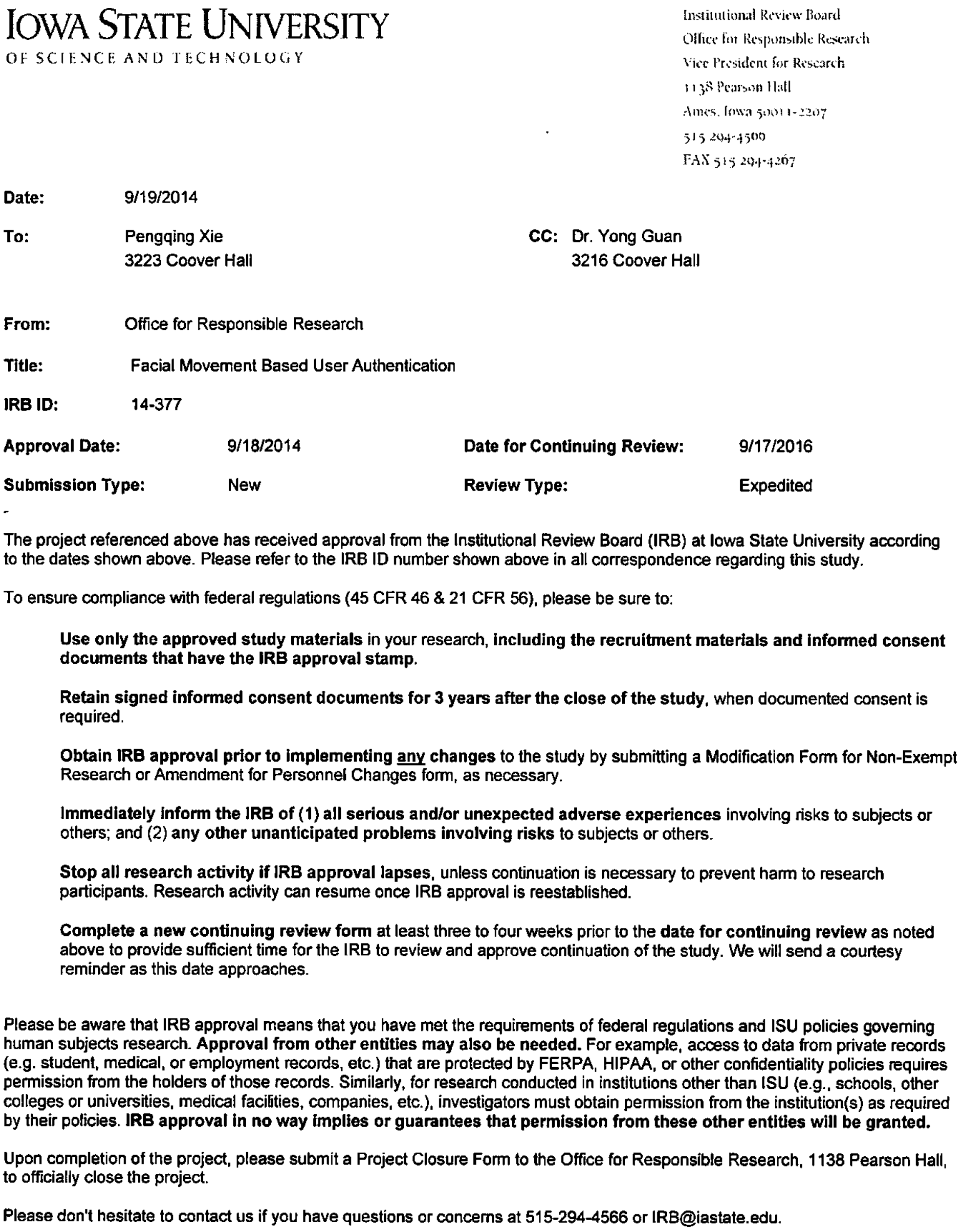
Action units (AUs) are defined by the Facial Action Coding System (FACS), each of which is anatomically related to the contraction of a specific set of facial muscles movement. Figure 18 contains a complete list of the six AUs being tracked by the Face Tracking SDK animated on the Candide-3 model and their corresponding range of values.



AU Name and Value	Avatar Illustration	AU Value Interpretation
Neutral Face (all AUs 0)		
AU0 – Upper Lip Raiser (In Candid3 this is AU10)		0=neutral, covering teeth 1=showing teeth fully -1=maximal possible pushed down lip
AU1 – Jaw Lowerer (In Candid3 this is AU26/27)		0=closed 1=fully open -1= closed, like 0
AU2 – Lip Stretcher (In Candid3 this is AU20)		0=neutral 1=fully stretched (joker's smile) -0.5=rounded (pout) -1=fully rounded (kissing mouth)
AU3 – Brow Lowerer (In Candid3 this is AU4)		0=neutral -1=raised almost all the way +1=fully lowered (to the limit of the eyes)
AU4 – Lip Corner Depressor (In Candid3 this is AU13/15)		0=neutral -1=very happy smile +1=very sad frown
AU5 – Outer Brow Raiser (In Candid3 this is AU2)		0=neutral -1=fully lowered as a very sad face +1=raised as in an expression of deep surprise

Figure 18. The complete list of AUs.

APPENDIX B. INSTITUTIONAL REVIEW BOARD APPROVAL



IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
1138 Pearson Hall
Ames, Iowa 50011-2207
515 294-4566
FAX 515 294-4267

Date: 9/19/2014

To: Pengqing Xie
3223 Coover Hall

CC: Dr. Yong Guan
3216 Coover Hall

From: Office for Responsible Research

Title: Facial Movement Based User Authentication

IRB ID: 14-377

Approval Date: 9/18/2014

Date for Continuing Review: 9/17/2016

Submission Type: New

Review Type: Expedited

The project referenced above has received approval from the Institutional Review Board (IRB) at Iowa State University according to the dates shown above. Please refer to the IRB ID number shown above in all correspondence regarding this study.

To ensure compliance with federal regulations (45 CFR 46 & 21 CFR 56), please be sure to:

- **Use only the approved study materials** in your research, **including the recruitment materials and informed consent documents that have the IRB approval stamp.**
- **Retain signed informed consent documents for 3 years after the close of the study**, when documented consent is required.
- **Obtain IRB approval prior to implementing any changes** to the study by submitting a Modification Form for Non-Exempt Research or Amendment for Personnel Changes form, as necessary.
- **Immediately inform the IRB of (1) all serious and/or unexpected adverse experiences** involving risks to subjects or others; and (2) **any other unanticipated problems involving risks** to subjects or others.
- **Stop all research activity if IRB approval lapses**, unless continuation is necessary to prevent harm to research participants. Research activity can resume once IRB approval is reestablished.
- **Complete a new continuing review form** at least three to four weeks prior to the **date for continuing review** as noted above to provide sufficient time for the IRB to review and approve continuation of the study. We will send a courtesy reminder as this date approaches.

Please be aware that IRB approval means that you have met the requirements of federal regulations and ISU policies governing human subjects research. **Approval from other entities may also be needed.** For example, access to data from private records (e.g. student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **IRB approval in no way implies or guarantees that permission from these other entities will be granted.**

Upon completion of the project, please submit a Project Closure Form to the Office for Responsible Research, 1138 Pearson Hall, to officially close the project.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.