PERSONAL IDENTIFICATION FROM FACIAL EXPRESSION

YE NING

NATIONAL UNIVERSITY OF SINGAPORE

2010

PERSONAL IDENTIFICATION FROM FACIAL EXPRESSION

YE NING
(B.Sc., Fudan University, 2005)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE THE DEGREE OF

Doctor of Philosophy

in

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE, 2010

To my parents, who love and support me all the way.

Acknowledgements

I am sincerely thankful to my supervisor, Dr. Terence Sim, for his guidance, support and encouragement from the very beginning of my Ph.D study. Without him, this thesis would not have been possible. I would also like to express my deep gratitude to Dr. Zhang Sheng and Dr. Yu Dan for the many invaluable discussions, which have greatly broadened my understanding about research. Thanks are heartily extended to my seniors, Dr. Miao Xiaoping, Dr. Zhang Xiaopeng and Mr. Rajkumar for all the help they have given to me, especially during the early days of my stay in NUS. I am also happily indebted to my colleagues and friends, Guo Dong, Li Hao, Zhuo Shaojie, Qi Yingyi, Chen Su and Wang Xianjun for all the treasured memories we have shared together. A special thank is given to the aunt who cleans our lab everyday in the early morning, though I do not know her name yet. Finally and most deeply, I owe my thanks to my parents for their eternal love, support and understanding. This work is dedicated to the lovely old couple in my deepest gratitude.

Abstract

Motion-based face recognition is a new member to the family of biometrics. It studies personal characteristics concealed behind facial motions (*e.g.* facial expressions, speech) and uses the information for identity recognition. Research in this field is in its early stage and many questions remain unanswered.

This thesis contributes in two unexplored aspects of motion-based face recognition: the use of facial expression dynamics and cross-expression identification techniques. Two novel approaches are proposed respectively and tested through a series of experiments. The experimental results indicate that facial expression dynamics can be highly discriminative and cross-expression motion-based face recognition is possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The term, motion-based face recognition, is used to refer to a group of biometric techniques which utilize facial motions to recognize personal identities. Motion-based face recognition is a young research area which is motivated by the growing demands from the security industry for more reliable biometrics systems as well as a recent psychological discovery that facial motions can benefit human perception of identity.

1.1 The Goal and the Questions

The ultimate goal of motion-based face recognition is to recognize human identity from any kind of facial motion in any reasonable pose of head and under any reasonable lighting condition. This is an extremely challenging task and is honestly far beyond the reach of existing techniques. In order to eventually achieve this ultimate goal in future, a series of research questions must be answered first, which include and may not be limited to the following ones:

1. Under which condition is motion-based face recognition viable?

2. If it is viable, what features should be used?

3. How discriminating are the features?

The three questions are fundamental to motion-based face recognition. The first question asks about the feasibility. Is motion-based face recognition generally possible, or limited to certain circumstances (*e.g.* a fixed pose or a fixed type of motion), or not possible at all? The second question asks about the methodology. What kind of features can be extracted from facial motion and used for biometrics? And is it possible to rely on just one feature or a set of features designed for different situations is necessary? The last question asks about the uniqueness of the features. Are the features so powerful that they can even tell identical twins apart or are the features so weak that they perform no better than a random guess? This thesis attempts answer these questions (please read Section 1.6 for the contributions of this thesis).

1.2 Relation to Conventional Face Recognition

Different from motion-based face recognition, conventional face recognition relies on static facial appearance, *i.e.* shape and color, to recognize human identity. Even in conventional video-based face recognition, the features are all based on face shape and color rather than facial motions. For the sake of convenience, conventional face recognition will always be referred to using the term, "appearance-based face recognition", hereafter in this thesis.

Motion-based face recognition and appearance-based face recognition share a common foundation of face detection. And the accuracy of face detection, which includes finding an approximate face region as well as locating a set of key points on the face, greatly affects the performance of either group of approaches.

Compared to appearance-based face recognition, motion-based face recognition is expected to be more robust to lighting variation and face makeup - as long as face detection works properly. This expectation has been justified in several experiments, including one which will be reported in this thesis.

Compared to appearance-based face recognition, motion-based face recognition is less well developed and less mature for practical use. This is understandable considering that appearance-based face recognition has been studied for almost 40 years while research on motion-based face recognition primarily started after 2000.

Motion-based face recognition and appearance-based face recognition can be complementary to each other. Motion-based face recognition works on facial motions while appearance-based face recognition works on static mugshots. Motion-based face recognition are less sensitive to lighting variation and face makeup while appearance-based face recognition seems to have higher recognition rate under standard imaging conditions [Chen et al. 2001]. By combining the advantages from both sides, it may be possible to build a more robust and more general face recognition system.

1.3 Background: Biometrics

Motion-based face recognition is a branch of biometrics, the science that studies how to recognize human identity based on biological characteristics. Those biolog-

ical characteristics are called biometric traits. There are two categories of biometric trait, physiological biometric traits and behavioral biometric traits. Typical physiological biometric traits include fingerprint, facial appearance, iris and palm print. Typical behavioral biometric traits include signature, voice and gait. Facial motion is a behavioral biometric trait. For a detailed survey on biometric technology, readers are referred to [Jain et al. 2006].

With existing biometric techniques, fingerprint and iris are considered the most reliable among biometric traits, but both require the cooperation of the subject - either to press his fingers on a fingerprint scanner or position his face right before an iris scanner. In comparison, face recognition can be performed contactlessly and at a distance, which allows for an operation called mass screening. The term, mass screening, means identifying everyone in a crowd simultaneously. Gait recognition also supports mass screening, but face recognition is much more reliable. This advantage makes face recognition the favorite choice in deploying camera-based surveillance systems in public places, *e.g.* at airports and casinos. Motion-based face recognition extends face-oriented biometrics by making use of facial motion, which up until recently has been considered a nuisance.

The discriminating power of a biometric trait can be measured by an FAR-FRR curve in an identity verification test or the Bayes' error rate. The FAR (false accept rate) is the probability of accepting an imposter as a genuine user and the FRR (false reject rate) is the probability of mistaking a genuine user for an imposter. Ideally, both FAR and FRR are zero, *i.e.* no errors are made. For any non-ideal biometric system, lowering one of the error rates means increasing the other. There is a trade-off between the two. Thus, the EER (equal error rate), where the two error rates are equal, is often used to give an overall performance of the system.

The Bayes' error rate is the ideal tool for measuring the discriminating power of a biometric trait, because it is the theoretical minimum error rate that can be achieved with the given biometric trait. Unfortunately, the true Bayes' error rate is usually unknown, because the true probability distribution of the biometric trait value is usually unknown. Thus, various mathematical tools have been proposed to estimate the Bayes' error rate from samples. In this thesis work, the Bayes' error rate is estimated from either 1NN (nearest-neighbor) error [Cover and Hart 1967] or the Bhattacharyya coefficient [Duda et al. 2000]. The choice of the evaluation tools largely depends on the nature of the databases used in the experiments.

1.4 Background: Dynamic Facial Signature

Motion-based face recognition is closely related to and partially motivated by psychological studies on human perception of faces. It is believed in psychology that facial motion helps humans to recognize familiar faces. For unfamiliar faces, contradictory experimental results have been reported [Shepherd et al. 1982; Schiff et al. 1986; Pike et al. 1997; Christie and Bruce 1998; Bruce et al. 1999; Bruce et al. 2001; Hill and Johnston 2001; Thornton and Kourtzi 2002]. How facial motion affects face perception is not known yet. Three major hypotheses exist: supplemental information hypothesis, representation enhancement hypothesis and motion as a social signal hypothesis. Among the three hypotheses, the supplemental information hypothesis is most related to the topic of this thesis. It states that facial motion provides identity-specific *dynamic facial signature* to help face perception [Roark et al. 2003] (please also refer to this article for the definition of the other two hypotheses). To a considerable extent, the purpose of the research on motion-

based face recognition can be considered as finding a computational dynamic facial signature.

1.5 The State of the Art

Research on motion-based face recognition is in its very early stage and only several articles have been published in this area. The reported results are encouraging but still far from applicable in practice. Existing works focus on looking for various motion-based features which can be used for identification. The types of facial motion that have been studied include smile [Pamudurthy et al. 2005; Tulyakov et al. 2007], mouth open [Zhang et al. 2004] and speech [Chen et al. 2001]. A detailed field review will be given in Section 2.2. A main drawback of existing works is that they are all limited to fixed facial motion, which means strictly the same facial motion for training and recognition. This requirement of fixed facial motion leaves a big gap between the state of the art and the ultimate goal of general motion-based face recognition.

1.6 Contributions of the Thesis

This thesis contributes in two unexplored aspects of motion-based face recognition.

1. The use of facial expression dynamics. Existing works merely make use of the point-wise displacement between the neutral face and the final pose of a facial expression and ignore the intermediate dynamics. In Chapter 3, it is argued that the dynamics, specifically smile dynamics, can be highly discriminating.

2. Cross-motion features. Existing works are all limited to fixed facial motion, that is, a human subject must perform a specific facial motion in order to be successfully recognized. This limitation is broken by the technique proposed in Chapter 4, which looks into the micro patterns of facial skin deformation observed during various facial expressions.

Other minor findings include:

- With smile dynamics, lower face is more discriminating than upper face (Section 3.2.4);
- A combination of smile dynamics and facial appearance may help distinguish between identical twins (in Section 3.4.5);
- The proposed cross-motion approach, Local Deformation Profile, can work under extremely heavy face makeup (in Section 4.2.5).

And possible applications include:

- To improve the performance of existing face recognition systems by incorporating the proposed motion-based techniques;
- To build identity-specific facial motion models for computer facial animation or psychological studies by adopting the proposed local deformation profile technique.

Chapter 2

Literature Review

This chapter reviews related literature from both the psychology and the pattern recognition communities. Although the purpose and method of the research in the two communities are very different, regarding the problem of motion-based face recognition, they have two common fundamental questions to answer: *is it possible?* and *how does it work?* Certainly that psychologists study humans and pattern recognition researchers study automated systems to answer those questions, but the findings may benefit and inspire both sides. The importance of this kind of "bridging" has been noticed by some researchers recently [Sinha et al. 2006].

2.1 Psychological Studies

Psychological studies on the role of facial motion in human perception of identity started in 1980's. After more than twenty years of research, it is now widely acknowledged that facial motion can benefit recognition of familiar faces, *i.e.* faces of someone's families, friends, colleagues or faces of celebrities, *etc*. For unfamiliar

faces, the reported results are contradictory and the community has not yet reached a consensus. Thus, this section focuses on the psychological studies regarding the role of facial motion in familiar face recognition by humans, especially those found to be inspiring to research on facial motion as a biometric trait. In all the psychological studies mentioned below, facial motion is a mixture of rigid motion (*i.e.* head motion) and non-rigid motion (*i.e.* facial expression or speech). Non-rigid motion dominates in the mixture in most of the cases.

In order to study the impact of facial motion in recognizing familiar faces, psychologists usually have to first completely or partially hide the facial appearance information from the experiment participants. Otherwise, the participants will easily recognize those faces by just a glance at the static face configuration.

One of the first studies in this field investigated the human ability of recognizing personal identity from pure facial motion. Bruce and Valentine [1988] employed point-light displays of faces so that appearance information was hidden from the audience. In a point-light display, reflective dots were scattered on a moving face and the brightness of the recording was reduced so that only the dots were visible - very much like the technique used in today's vision-based motion capture systems. They found that the participants could recognize the faces of their friends under this display but with a low accuracy (33.5%). Interestingly, similar idea was adopted by Tulyakov *et al*. [2007] in a pattern recognition paper twenty years later. And they reached a similar conclusion, but in pattern recognition/biometrics terminology, that sparse tracker displacements possessed weak discriminating power and could only be used as a *soft biometric trait*, a concept used to refer to a less reliable class of biometric traits which can be used to assist in the decision making process of a primary biometric system [Jain et al. 2004]. Their work will be discussed in more

detail in Section 2.2.

Follow-up research focused on studying the advantage in identification that facial motion may bring, over static faces. Knight and Johnston [1997] asked their participants to recognize famous faces (*e.g.* the faces of celebrities or politicians) from negative videos/images. The faces were better recognized when presented as videos rather than as single static images. Lander *et al.* [2001] ran a similar experiment with pixelized and blurred videos/images. Advantages in recognition were observed when videos were presented to the participants. In other two reported studies, single static images were replaced by multiple static images [Lander et al. 1999] and jumbled videos [Lander and Bruce 2000] (video/image degradation was applied in both experiments). And in both cases, the famous faces presented in normal-ordered videos were better recognized by the participants. In aforementioned experiments, generally, using facial motion videos as stimuli increased recognition accuracy by 5 to 20 percentage points in terms of recognition rate or hit rate.

When normal non-degraded videos/images of famous faces were used in experiment, less reaction time in recognition was observed with videos [Lander and Bruce 2004].

Efforts have also been put in studying the type of facial motion which can aid face recognition by humans. Lander and Chuang [2005] tested and compared the face recognition accuracy in using static images, rigid head motion videos, talking videos and facial expression videos as stimuli. The faces to be recognized were personally familiar to the participants (as their teachers, students or colleagues). Videos/images were degraded with lower contrast, higher brightness and image blur to avoid ceiling effect. Compared to using static images, significant advantages

Study	Display of Faces	Major Findings
[Bruce and Valentine 1988]	point-light display	Participants can recognize faces in point-light display, but with low accuracy.
[Knight and Johnston 1997]; [Lander et al. 2001]	negative / pixelized / blurred videos/images	Moving faces were better recognized than static faces.
[Lander et al. 1999]	degraded videos / multi-images	Faces in videos were better recognition than faces in multiple static images.
[Lander and Bruce 2000]	degraded normal-ordered videos / jumbled videos	Faces in normal-ordered videos were better recognized than faces in jumbled videos.
[Lander and Bruce 2004]	normal videos/images	Faces in videos were recognized with less reaction time.
[Lander and Chuang 2005]	degraded videos of facial expressions, talking, rigid head motion and static images	Faces in videos of facial expressions or talking were recognized with the highest accuracy; faces in rigid head motion was better recognized than faces in static images with a small advantage.
[Lander et al. 2006]	degraded videos of natural smile / synthesized smile and static images	Faces in natural smile videos were better recognized than faces in static images, but faces in synthesized smile videos were not.

Table 2.1: Major findings from psychological studies on the role of facial motion in recognizing familiar faces by human

in face recognition were observed when talking videos or facial expression videos were used as stimuli (an increment of 25 to 35 percentage points in recognition rates). Less advantage was observed with rigid head motion videos (an increment of around 10 percentage points in recognition rates). In another work done by Lander *et al*. [2006], they studied the recognition advantages possibly brought by natural smile videos and synthesized smile videos (which were generated using computer graphics techniques). And they found that compared with single static face image, the natural smile videos were better recognized while the synthesized smile videos were not.

Table 2.1 summarizes the major findings from aforementioned psychological studies. Please note that all those studies were about *familiar* face recognition.

For a more detailed field review which covers both familiar and unfamiliar face recognition by human, please refer to [Roark et al. 2003].

From those psychological findings, several conclusions could be drawn and may be useful for related research on motion-based face recognition in pattern recognition and biometrics.

1. Sparse representation of facial motion may not be very discriminative.
2. The benefit brought by facial motion is mostly observable under non-optimal viewing conditions in which appearance information is distorted.
3. Non-rigid facial motion (*i.e.* facial expression, talking) may be more discriminative than rigid motion.

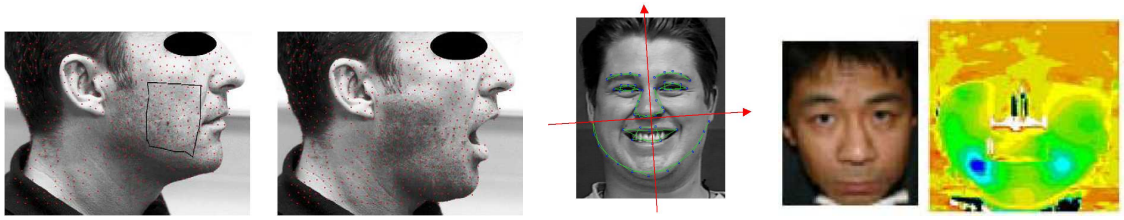
The first conclusion is supported by the work done by Bruce and Valentine [1988]. The second conclusion is based on the fact that in most of the experiments, degraded images/videos have been used. The last conclusion is drawn from the work done by Lander and Chuang [2005].

2.2 Pattern Recognition Studies

In the pattern recognition community, research on motion-based face recognition started primarily after year 2000. Existing works focus on looking for discriminating features from various kinds of facial motions.

2.2.1 Existing Works

Chen *et al*. [2001] concatenated a series of dense optical flow fields computed from a short talking video to make a feature. The vocabulary of the speech was limited to



(a) (b) (c)

Figure 2.1: Motion-based features for face identification used by existing works

two specific words. They claimed that the feature was less sensitive to illuminance variation, compared with traditional facial appearance features. A recognition rate around 87% was reported.

Zhang *et al*. [2004] made use of physical laws (momentum conservation and Hooke's law) to estimate the elasticity of the masseter muscle from a pair of face range images, *i.e*. 3D images (Figure 2.1(a)). The first image was the side view of a neutral face and the second one was the side view of the same face with its mouth open. They claimed that this estimated elasticity could be used as a biometric trait. At a false alarm rate of 5%, a verifiction rate of 67.4% was achieved.

Pamudurthy *et al*. [2005] used a dense displacement field as a feature (Figure 2.1(c)). The field was computed from a pair of face images. The first image was the frontal view of a neutral face and the second image was the frontal view of the same face with a slight smile. They claimed that this feature could be used for identification even under face makeup. No quantitative evaluation of identification performance was reported.

Tulyakov *et al*. [2007] used sparse tracker displacement as a feature. A set of tracker points were defined on a pair of face images. The first image was the frontal view of a neutral face and the second one was the frontal view of the same face with a smile (Figure 2.1(b)). After rigid alignment, the displacements of the tracker

Study	Motion	Input	Feature	Fixed motion?
[Chen et al. 2001]	speech	one video	optical flow	yes
[Zhang et al. 2004]	mouth open	two images	muscular elasticity	yes
[Pamudurthy et al. 2005]	smile	two images	dense displacement field	yes
[Tulyakov et al. 2007]	smile	two images	tracker displacement	yes

Table 2.2: Existing works in motion-based face recognition

points were calculated and stacked to form a long feature vector. They said that this feature could be used as a soft biometric trait [Jain et al. 2004]. An equal error rate around 0.4 was reported.

2.2.2 Research Gaps

Table 2.2 summaries existing works in motion-based face recognition, from which two big gaps are noticeable.

First, those studies which deal with smile (*i.e.* [Pamudurthy et al. 2005] and [Tulyakov et al. 2007]) exploit only the displacements between a neutral face and a final smiling face and ignore the intermediate dynamics. In Chapter 3, it is argued that the smile dynamics can be highly discriminating.

Second, existing works are all limited to fixed facial motion, that is, a human subject must perform a specific facial motion in order to be successfully recognized. This limitation is overcome by the technique proposed in Chapter 4, which looks into the micro patterns of facial skin deformation observed during various facial expressions.

Chapter 3

A Fixed-Motion Method: Smile Dynamics

This chapter describes a study on using smile dynamics for identification. A novel motion-based feature, smile dynamics, is proposed. The experimental results indicate that this feature is highly discriminating. Efforts are also made in combining smile dynamics with facial appearance to yield a hybrid feature with even greater discriminating power.

Compared with existing works, this study is novel in two aspects:

1. Proposes the first technique which makes use of the dynamics of a facial expression for personal identification;

2. Makes the first attempt in combining facial motion with facial appearance for personal identification.

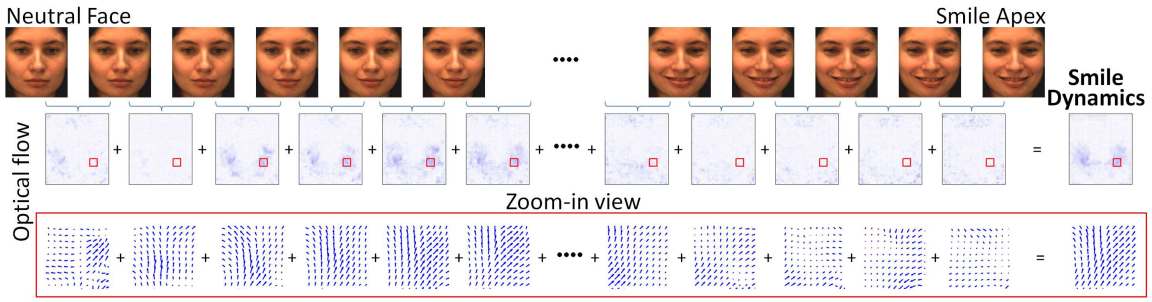


Figure 3.1: Smile dynamics is defined as the sum of a series of optical flow fields which are computed from the pairs of neighboring frames of a smile video.

3.1 Smile Dynamics

Smile dynamics[1] is defined as the sum of the motion fields which are extracted from a smile (Figure 3.1). Given a frontal-view smile video which starts from a neutral face, smile dynamics is computed in following steps:

1. A set of key points are located on the neutral face in the first frame (Figure 3.2(a));
2. The set of key points are tracked throughout the rest of the video;
3. Faces are aligned and cropped from the video;
4. Optical flow fields are computed from each pair of sequential cropped face images;
5. Smile intensity is computed for each face image based on its offset from neutral face;
6. The face image with the greatest smile intensity, *i.e.* smile apex, is detected (Figure 3.2(b));

[1]This work was published in [Ye and Sim 2008].

7. The optical flow fields between the neutral face and the smile apex are summed pixel-wisely. The sum is called smile dynamics (Figure 3.1).

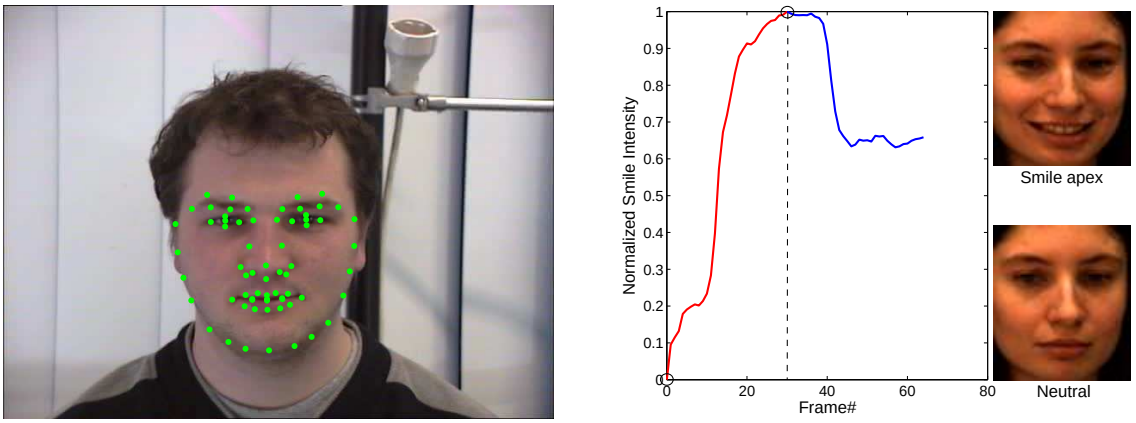
In current implementation, STASM [Milborrow and Nicolls 2008] is used for face detection and localization (Step 1). Lucas-Kanade optical flow estimation [Lucas and Kanade 1981] with pyramidal refinement is used in Step 2 and 4. In Step 3, faces are aligned based on the positions of eyes by a 2D similarity transformation. The size of the cropped face images is 81 by 91 pixels. Smile intensity is defined as follows,

$$\tau(k) = \sum_{z \in pixels} \| \sum_{i=1}^{k} f_i(z) \|_2, \tag{3.1}$$

where $f_i(z)$ denotes the 2D motion vector estimated on pixel z and between the $(i-1)$-th and the i-th sequential face images; $\| \cdot \|_2$ denotes l^2-norm; $\tau(k)$ is the smile intensity of the k-th face image; the 0-th face image is of a neutral facial expression. The smile intensity grows during the phase of smiling but drops during the phase of relaxing. Because the motion observed during relaxing will cancel the motion accumulated during smiling. Figure 3.2(a) shows an example of face localization and Figure 3.2(b) shows an example of smile intensity (normalized to 0 to 1 for convenience of representation). Smile dynamics is defined as the sum of motion fields between the neutral face and the smile apex,

$$u = \sum_{i=1}^{K} f_i, \quad K = \arg\max_k \tau(k), \tag{3.2}$$

where K-th frame contains the smile apex. The fixed-motion assumption implies that the intensity of smile apex is approximately consistant for the same subject across different video recordings. Identification from smiles of largely varying in-



(a) (b)

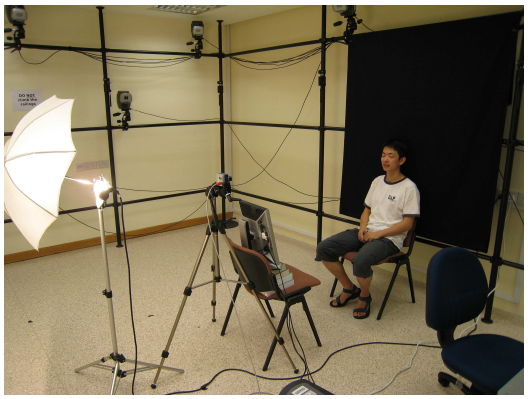
Figure 3.2: (a) Face localization result; (b) Normalized smile intensity: the red and the blue curves illustrate the neutral-to-smile period and the smile-to-neutral period, respectively; the neutral face and smile apex images are shown on the right.

tensity (e.g. laughing *vs*. smirking) is essentially a cross-motion problem. There are three reasons for choosing this temporally compressed representation of smile dynamics. First, compared with a whole set of optical flow fields, the dimension of the data is significantly reduced. Second, this representation requires no temporal alignment. And last also the most important, this representation preserves sufficient discriminating power for the task of personal identification - as shown in the experiments (Section 3.2).

Suppose the video resolution is $w \times h$ pixels, then u is a $2wh \times 1$ column vector. In order to reduce the data dimension, PCA (Principal Components Analysis [Duda et al. 2000]) is applied,

$$v = P_d(u - \bar{u}), \tag{3.3}$$

where the matrix P_d consists of the first d principal components (arranged as rows); $\bar{u}$ is the sample mean. v is used in the experiments.



(a) Video recording setup

(b) A participant in action

Figure 3.3: Smile video collection

3.2 Discriminating Power Analysis

This section examines the discriminating power of smile dynamics by multi-class distribution separability. A high class separability will suggest a high discriminating power of smile dynamics.

3.2.1 The Dataset

The dataset used in this experiment consists of 341 smile video clips which are collected from 10 human subjects, 30 to 40 clips each. Each clip records the facial motion of one subject performing a smile. The expression begins with a neutral face, moves to a smile, and then back again to the neutral expression. Videos were recorded at 15fps under the resolution of 768 by 1024 pixels using a Unibrain Fire-i 701c firewire camera. The subjects were asked to perform their own smiles as naturally as they could. Before recording, a sample smile video was shown to the subject to remind him/her of the proper intensity of the smile (in order to avoid too small or too big smiles). Also, an LCD display was placed before the subject to

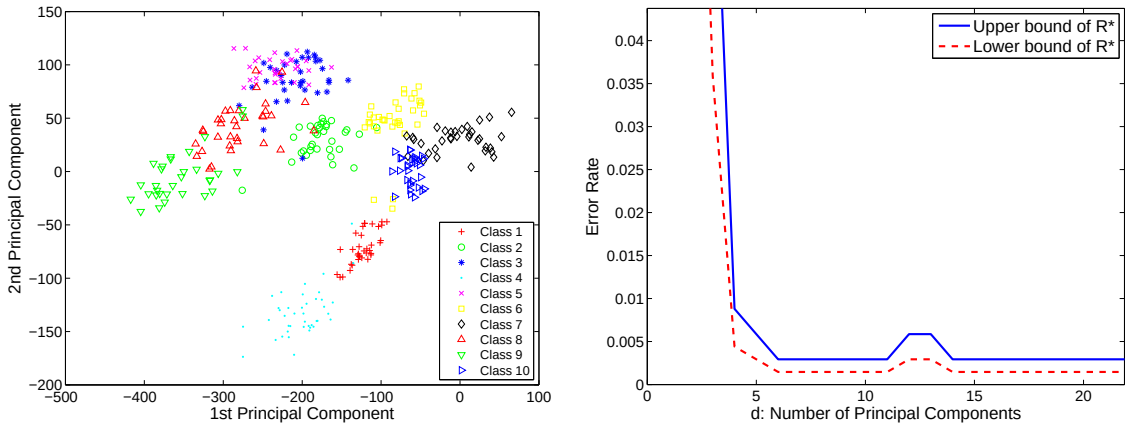
let the subject see himself/herself during recordings, because it was found that the subjects smiled more naturally when they were able to see themselves. The subjects would take a rest after every 4 or 5 times of recordings. The whole recording was conducted in two sessions over two days to avoid fatigue. Figure 3.3(a) and 3.3(b) show the video recording setup and a participant in recording, respectively.

3.2.2 Data Visualization

Figure 3.4(a) visualizes the smile dynamics extracted from the dataset, after projected to the first two principle components. Although the first two principal components preserve only 35.36% of the total energy, the projected features from the 10 classes (i.e. 10 subjects) form visually well-separated clusters (except for Class 3 and Class 5). Quantitative analysis of the class separability is carried out by estimating the Bayes' error rate using the 1NN error rate (single nearest neighbor error rate).

3.2.3 The Bayes' Error Rate

The Bayes' error rate is the theoretical minimum rate any classifier can achieve. Therefore, the ideal way of measuring class separability is to calculate the Bayes' error rate based on the underlying probability distributions of those classes. However, directly calculating the Bayes' error rate is difficult in practice, because the calculation requires the probability density functions which are generally unknown in most applications. Various methods have been proposed to estimate the Bayes' error rate from a set of observations. The approach proposed by Cover and Hart [Cover and Hart 1967] is taken in this study. They have proved that, when the



(a) (b)

Figure 3.4: Class separability studies: (a) Data visualization after projected to 2D space; (b) The band of R^*: the Bayes' error rate R^* is bounded by the blue curve and the red dashed curve (Eq.(3.5));the horizontal axis denote the number of principal components d used in dimension reduction (Eq.(3.3)).

number of samples N approaches infinity, the following inequality holds,

$$R^* \leq R \leq R^* \left(2 - \frac{1}{\alpha} R^*\right), \tag{3.4}$$

or

$$\alpha - \sqrt{\alpha^2 - \alpha R} \leq R^* \leq R, \tag{3.5}$$

where

$$\alpha = \frac{M-1}{M} \tag{3.6}$$

and M is the number of classes (with current dataset, $M = 10$); R^* denotes the Bayes' error rate; R denotes the 1NN error rate, which is defined as,

$$R = \frac{|\{v|\theta(v) \neq \theta(v_{nn})\}|}{N}, \tag{3.7}$$

where v is the feature computed from Eq.(3.3); $\theta(\cdot)$ denotes the labeling function; v_{nn} denotes the nearest neighbor of v; $|\cdot|$ denotes the set size; and N denotes the number of data points. In other words, R is the fraction of the sample whose class labels are different from those of their nearest neighbors.

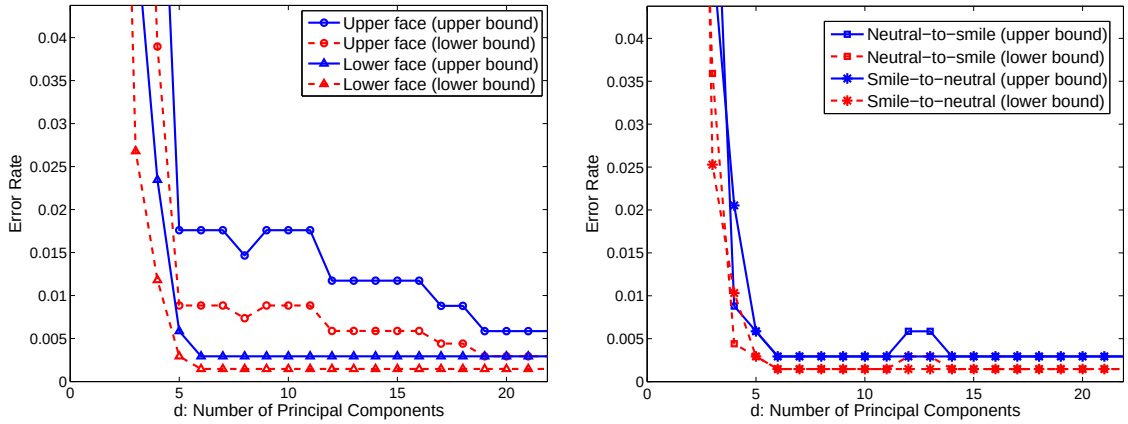
As proved in [Cover and Hart 1967], the bounds given by Eq.(3.5) are tight. Although in real-world applications, it is impossible to get infinite number of samples (with current dataset, $N = 341$), it is a reasonable practice to indirectly measure the Bayes' error rate using the 1NN error rate.

Figure 3.4(b) shows the band of the Bayes' error rate R^* estimated by Eq.(3.5) at $M = 10$. The horizontal axis denotes the number of principal components d used in dimension reduction (Eq.(3.3)). The upper and lower bounds of R^* drop to 0.0029 and 0.0015 respectively at $d = 6$. After $d > 6$, both curves are largely flat, with minor ripples[2]. Such a low error rate suggests clear separation between the underlying probability distributions of the 10 classes, which suggests a high class separability of the extracted features. In other words, the feature is highly discriminating.

3.2.4 Upper Face *vs.* Lower Face

This subsection examines the features generated from upper-face regions and lower-face regions separately to investigate which part of the face is more discriminating. Figure 3.5(a) shows the experiment results. It can be seen that the upper bound of the lower-face error rate (the blue curve with triangles) is always equal to or lower than the lower bound of the upper-face error rate (the dashed red

[2]The Bayes' error rate never increases as more principal components are involved, because the extra components can always be ignored if including them in classification would decrease the discriminating power. The curves shown in Figure 3.4(b) are estimated bounds of the Bayes' error rate, so they may go up and down as the number of principal components goes up.



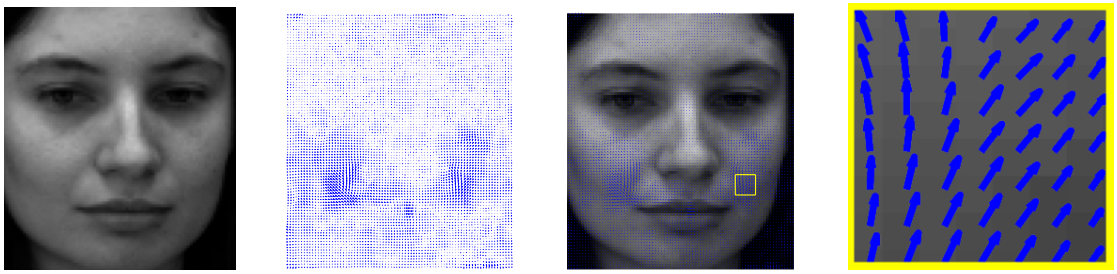
(a) Upper face *vs*. lower face (b) Neutral-to-smile *vs*. smile-to-neutral

Figure 3.5: More class separability studies: upper face *vs*. lower face and smiling *vs*. relaxing

curve with dots). The lower-face error rate can be less than the upper-face error rate by as much as 3 times at $d = 10$. This observation implies that the lower face is more discriminating than the upper face. This stands in contrast to static face recognition, where it has been shown that the upper face is more discriminating [Gross et al. 2001; Ekenel and Stiefelhagen 2006].

3.2.5 Neutral-to-Smile *vs*. Smile-to-Neutral

This subsection examines the features generated from the neutral-to-smile period (the red curve in Figure 3.2(b)) and from the smile-to-neutral period (the blue curve in Figure 3.2(b)) separately to investigate which period of motion is more discriminating. Figure 3.5(b) shows the experiment results. It can be seen that the two upper bounds (the two blue curves) overlap each other almost everywhere, so do the two lower bounds (the two red dashed curves). This observation implies that neutral-to-smile motion and smile-to-neutral motion provide almost the same



(a) Appearance (b) Dynamics (c) Hybrid (d) Zoom-in

Figure 3.6: The three types of features examined in Section 3.3: readers may want to zoom in on (b) (c) to see the motion flows clearly.

amount of information about identity.

3.3 Combining Smile Dynamics with Facial Appearance: A Hybrid Feature

This section, together with next section, report a study on combining smile dynamics with conventional facial appearance. The result of this combination is a novel hybrid feature, whose discriminating power is greater than that of either facial appearance or smile dynamics[3].

Three different features are examined and compared in this study: smile dynamics, facial appearance and a hybrid feature (Figure 3.6). Smile dynamics has been introduced previously in Section 3.1 (more specifically, defined in Eq.(3.2)) and will be denoted as u^m in this study. PCA is applied to reduce the dimension of the data,

$$v^m = P^m_{k^m}(u^m - \overline{u^m}), \tag{3.8}$$

[3]This work was published in [Ye and Sim 2009].

where $P^m_{k^m}$ is the projection matrix which consists of the first k^m principal components; $\overline{u^m}$ denotes the sample mean. Similarly, the facial appearance feature v^a is computed as,

$$v^a = P^a_{k^a}(u^a - \overline{u^a}), \tag{3.9}$$

where u^a denote a column vector made by stacking all the pixel values of the first frame of a video clip, which is a static neutral face image (Figure 3.6(a)). Finally, the hybrid feature is computed as a weighted mixture of facial appearance and smile dynamics,

$$u^h = \begin{pmatrix} (1-w)u^a/\alpha \\ wu^m/\beta \end{pmatrix}, \tag{3.10}$$

$$v^h = P^h_{k^h}(u^h - \overline{u^h}), \tag{3.11}$$

where $0 \leq w \leq 1$ denotes the weight of smile dynamics in the hybrid feature; α and β are two scalars used for normalizing the scales of u^a and u^m, respectively (in implementation, α and β are set to the medians of the l^2-norm of all u^a and all u^m, respectively); $P^h_{k^h}$ is the projection matrix which consists of the first k^h principal components; $\overline{u^h}$ denotes the sample mean.

3.4 Face Verification Test and Comparison

In section, the three features (static facial appearance, smile dynamics, hybrid) are tested in turn for face verification. The performance are evaluated and compared.

3.4.1 The Dataset

With the previous dataset (Section 3.2.1), ceiling effect is observed in the experiment with facial appearance feature. Thus, in this evaluation, more data are included. Specifically, the smile videos from three different databases are merged into one dataset. The three databases are the FEEDTUM video database [Wallhoff 2006], the MMI face database [Pantic et al. 2005] and the previous smile video dataset. The FEEDTUM database contains 18 subjects, with three smile videos per subject. The MMI database contains 17 subjects, with one to 16 smile videos per subject. After eliminating unusable videos (mainly due to excessive out-of-plane head motion), the whole dataset consists of 45 subjects and 435 videos in total. Each video clip is a frontal-view recording of a subject performing a facial expression from neutral to smile and back to neutral.

3.4.2 Genuine Distance and Impostor Distance

Face verification performance can be measured by the statistical separability between the distribution of genuine distance and the distribution of impostor distance. Given a set of feature vectors with identity labels, the genuine distance set D_G and the impostor distance set D_I are defined as follow,

$$D_G = \{\|v_i - v_j\|_2\}, \quad L(v_i) = L(v_j), \quad i \neq j, \tag{3.12}$$

$$D_I = \{\|v_i - v_j\|_2\}, \quad L(v_i) \neq L(v_j), \quad i \neq j, \tag{3.13}$$

where v_i and v_j are two feature vectors; $L(v_i)$ and $L(v_j)$ are the identity labels of v_i and v_j, respectively; $\|\cdot\|_2$ denotes the l^2-norm. From the dataset (Section 3.4.1),

5886 genuine distances and 88509 impostor distances are extracted, *i.e.* $|D_G| = 5886$, $|D_I| = 88509$.

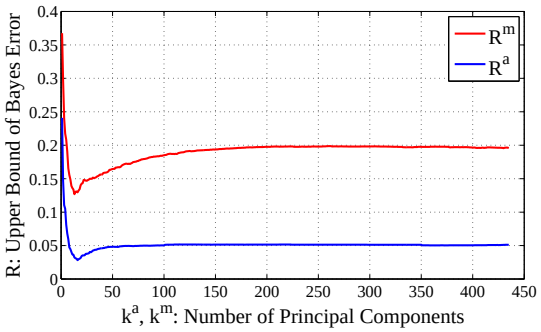
The separability of the two distributions underlying those two distance sets indicates the discriminating power of the feature. The Bayes' error rate is the ideal tool for measuring the separability, because it is the theoretical minimum error rate that any classifier can achieve in classifying the two distances. However, computing the Bayes' error rate directly is difficult in practice, because the exact probability density functions are usually unknown. In this experiment, Bhattacharyya coefficient [Duda et al. 2000] is used to estimate the Bayes' error rate,

$$\rho = \sum_x \sqrt{p_{D_G}(x) p_{D_I}(x)}, \tag{3.14}$$

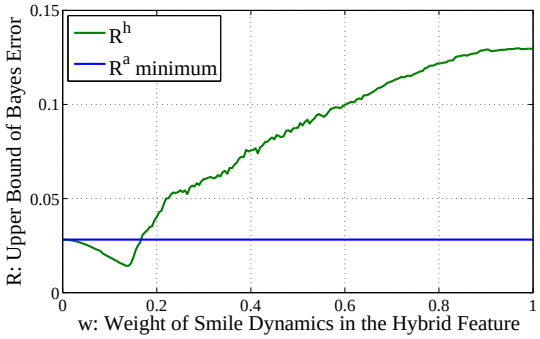
where ρ is the Bhattacharyya coefficient; $p_{D_G}(x)$ and $p_{D_I}(x)$ denote the two discrete probability density functions underlying D_G and D_I, respectively. In implementation, $p_{D_G}(x)$ and $p_{D_I}(x)$ are approximated using the histograms constructed from D_G and D_I, respectively. $0 \leq \rho \leq 1$, where $\rho = 0$ implies a complete separation between the two distributions and $\rho = 1$ implies a complete overlap between the two distributions. The smaller the ρ is, the more separable the two distributions are and therefore the more discriminative the feature is. Bhattacharyya coefficient is an upper bound of the Bayes' error rate in two-category classification problems,

$$R = \rho/2 \geq E_{Bayes}. \tag{3.15}$$

Thus, in this study, R, *i.e.* the upper bound of Bayes' error, is used as the measurement of the face verification performance. Note that $0 \leq R \leq 0.5$ where a smaller R



(a) R^a versus R^m (Eq.(3.15)): the horizontal axis denotes the number of principal components used in dimension reduction (k^a is in Eq.(3.9) and k^m is in Eq.(3.8)); R^a hits its minimum of 0.028 at $k^a = 16$; R^m hits its minimum of 0.127 at $k^m = 13$.



(b) R^h (Eq.(3.15)) with varying w (Eq.(3.10)): k^h (Eq.(3.11)) is fixed to be 16; the dashed blue line denotes the minimum of R^a (see Figure 3.7(a)); R^h hits its minimum of 0.014 at $w = 0.135$.

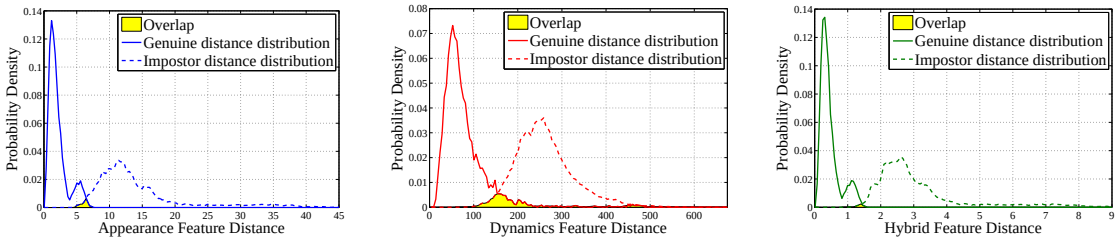
Figure 3.7: Face verification performance evaluation and comparison

indicates a better performance. R^a, R^m, R^h are used to denote the measurement computed from the holistic facial appearance feature (v^a), the smile dynamics feature (v^m) and the hybrid feature (v^h), respectively.

3.4.3 Appearance feature *vs.* smile dynamics feature

Figure 3.7(a) shows R^a and R^m with varying dimensions of the feature vectors (k^a in Eq.(3.9) and k^m in Eq.(3.8)). R^a hits its minimum of 0.028 at $k^a = 16$. R^m hits its minimum of 0.127 at $k^m = 13$. And almost at any dimension, R^a is at least three times smaller than R^m. This observation implies that, with respect to the current dataset, the face verification performance with appearance feature can be at least three times better than the performance with smile dynamics feature.

Figure 3.8(a) and Figure 3.8(b) show the distributions of genuine distance and impostor distance computed from the appearance feature vectors and the smile dynamics feature vectors at $k^a = 16$ and $k^m = 13$, respectively. It can be seen



(a) The distributions of genuine distance and impostor distance estimated from the appearance feature vectors at k^a = 16 (Eq.(3.9)).

(b) The distributions of genuine distance and impostor distance estimated from the smile dynamics feature vectors at k^m = 13 (Eq.(3.8)).

(c) The distributions of genuine distance and impostor distance estimated from the hybrid feature vectors at k^h = 16, w = 0.135.

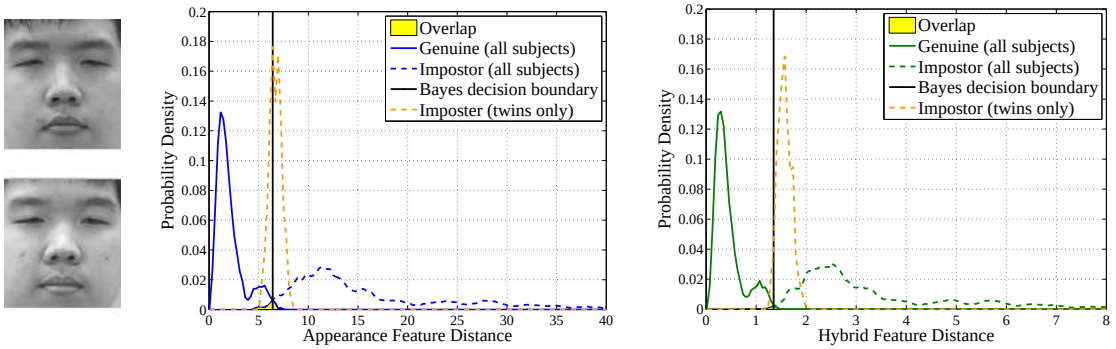
Figure 3.8: Distributions of genuine distance and impostor distance

that the overlap in Figure 3.8(a) is much smaller than the overlap in Figure 3.8(b). Since the overlap is directly related to the Bayes' error rate [Duda et al. 2000], this observation also implies that the appearance feature is more discriminative than the smile dynamics feature.

3.4.4 Appearance feature *vs.* hybrid feature

Since the appearance feature outperforms the smile dynamics feature considerably, the hybrid feature is compared with the appearance feature only.

Figure 3.7(b) shows R^h with varying w (the weight of the smile dynamics feature in the combination, see Eq.(3.10)). w is varied from 0 to 1 with an increment of 0.005 in each step. And since the appearance feature performs best at k^a = 16, k^h is fixed to 16 so that the comparison between the appearance feature and the hybrid feature is fair. In Figure 3.7(b), it can be seen that R^h keeps going down as w grows, until w = 0.135. After that, adding more smile dynamics causes R^h to increase. At w = 0.135, R^h hits its minimum of 0.014, which is a half of 0.028, the minimum of R^a. This observation implies that with current dataset, when w = 0.135, the hybrid



(a) Twins (b) Appearance feature (c) Hybrid feature

Figure 3.9: An attempt on the identical twins problem

feature can be twice more discriminative than the appearance feature.

Figure 3.8(c) shows the distributions of genuine distance and impostor distance computed from the hybrid feature at $k^h = 16, w = 0.135$. Compared with Figure 3.8(a), the overlap (the bright yellow region) between the two distributions becomes smaller, which implies stronger discriminating power of the hybrid feature compared with the appearance feature.

3.4.5 An Attempt on the Identical Twins Problem

The identical twins problem is *the Holy Grail* in face recognition. In the following experiment, an attempt is made on it. Although the identical twin dataset is too small (only one pair of identical twin brothers) to draw any statistically convincing conclusion, the experiment results do suggest that smile dynamics may help in distinguishing identical twins.

Around 20 smile video clips are collected from each of a pair of identical twin brothers (Figure 3.9(a)). The data is added to the previous dataset (Section 3.4.1). Two Bayes' classifiers are trained on the whole dataset. One of them uses the

	FRR (twins)	FAR (twins)	FRR (ordinary subjects)	FAR (ordinary subjects)
Appearance feature based	0%	25.53%	1.34%	1.51%
Hybrid feature based	0%	2.11%	0.48%	0.48%

Table 3.1: FRRs and FARs of two Bayes classifiers applied on the identical twins data

appearance feature and the other uses the hybrid feature. The two classifiers are tested on the same dataset as used for training. For each classifier, two sets of FAR (False Accept Rate) and FRR (False Reject Rate) are computed, one from classifying the data of all the ordinary subjects only and the other from classifying the data of the identical twins only.

The FARs and FRRs have been shown in Table 3.1. The most interesting results are found in the third column of the table, the two FARs computed from classifying the data of the identical twins. The FAR (twins) of the hybrid-feature-based classifier is smaller than the FAR (twins) of the appearance-feature-based classifier by an order of magnitude. Please note that the FAR (twins) represents the chances of mistaking one twin for the other. Visually, by comparing Figure 3.9(b) and Figure 3.9(c), it can be seen that with the hybrid feature, the distribution of impostor distance between the twin brothers shifts towards the right side of the Bayes decision boundary (ideally, the distribution of impostor distance should be all to the right side of the decision boundary so that the FAR is zero).

Readers may ask why not train the classifiers on the data of twins only and then test their performance and compute the FAR (twins). The reason is that in the real world, a face recognition system can never know beforehand if the two

faces in question are from a pair of identical twins or not. If the system knows that they are identical twins, then it already knows that they are from two different persons. Thus, the system will never choose to use a classifier trained specifically for identical twins. The best researchers can do is to build one system and try to make it applicable to both ordinary people and identical twins. This is the way that has been followed in this study.

3.5 Summary

This chapter has reported the studies on smile dynamics as a biometric trait as well as its combination with facial appearance. The combination yields a novel motion-appearance hybrid feature. The contributions of the studies include:

1. The first technique which makes use of the dynamics of a facial expression (smile) for personal identification;

2. The first motion-appearance hybrid feature for personal identification.

The major findings with the smile dynamics include:

1. Smile dynamics is discriminative enough to be used as a biometric trait;

2. With smile dynamics, the lower face is more discriminative than the upper face;

3. The relaxing phase (from smile apex to neutral face) is as discriminative as the smiling phase (from neutral face to smile apex).

The major findings with the motion-appearance hybrid feature include:

1. The hybrid feature is more discriminative than either smile dynamics or facial appearance;

2. The hybrid feature may help distinguish identical twins.

Chapter 4

A Cross-Motion Method: Local Deformation Profile

In Chapter 3, a certain type of facial motion, smile, is assumed. That method, smile dynamics, can be easily extended to any other type of facial motions - within the fixed-motion constraint that the types of facial motion used in training and identification are the same. If the constraint is broken, then the method will fail. Because the decision boundary learned from one type of facial motion will not be applicable to another (such a situation is more or less similar to using a decision boundary learned from fingerprints to differentiate two palm prints). In this chapter, this constraint is relaxed. The question to be answered in this chapter is: is it possible to train an algorithm on one type of facial motion and then ask it to identify people from another type of facial motion (cross-motion)? - provided that the two types of facial motion are at least locally similar in some part of the face. The answer to this question, as to be shown, is *yes*. Specifically, in this chapter, a novel approach is proposed to overcome the fixed-motion limitation, by investigating local skin

deformation patterns exhibited during facial motions[1].

This cross-motion approach is able to extract identity evidence from various types of facial motion, as long as those facial motions are at least, in some part of the face, locally similar to the facial motions used in training. This technique is named Local Deformation Profile (referred to as LDP hereafter). It is the first cross-motion method in the field. This approach is tested through several experiments conducted over a video database of facial expression. The experiment results demonstrate the potential of LDP to be used for biometrics. Moreover, in one of the experiments, the performance of LDP is evaluated under extremely heavy face makeup, showing its usefulness to recognize faces in disguise.

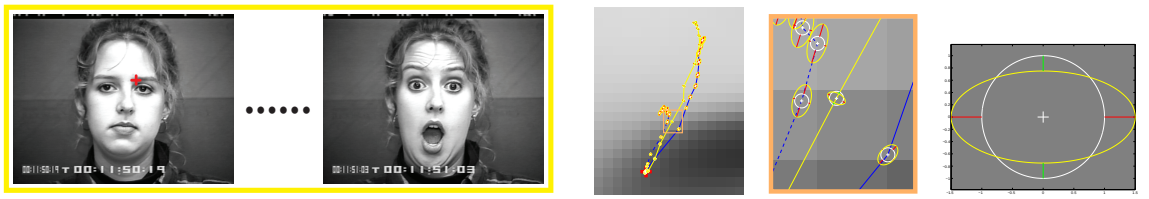
4.1 Methodology

The hypothesis behind LDP is that different faces can exhibit different deformation patterns (*i.e.* stretch and compression) when undergoing the same motion, due to the slight but not negligible individual difference in the physical property of facial materials (skin, muscle, *etc.*). This hypothesis can be considered as an engineering counterpart of the *supplemental information hypothesis* in psychology research (Section 1.4). Based on this hypothesis, the strategy is to look for identity evidence from parts of the faces where similar motion (displacement) are observed. The restriction to "similar motion" is important, because a human face is a non-linear and anisotropic elastic surface. A difference in either the direction or the magnitude of the displacement can cause a significant difference in the deformation patterns observed. This "similar motion" restriction means: in comparing the facial motions

[1]This work is to be published in [Ye and Sim 2010].

(a) Video clip (the solid blue line in (d)): subject 74, surprise

(b) Video clip (the dashed blue line in (d)): subject 74, fear



(c) Video clip (the solid yellow line in (d)): subject 52, surprise

(d) Local deformation pattern

(e) Zoom-in

(f) Deformation indicator

Figure 4.1: An example of local deformation pattern: (a)(b)(c) are the three video clips from which the deformation patterns of a specific point (marked using red cross) are computed; the motion trajectories and the deformation patterns of this point are illustrated in (d), after being aligned to the mean face shape; in (d), the lines represent the motion trajectories and the ellipses are deformation indicators which are computed at each video frame; (f) shows an enlarged deformation indicator; the white cross denotes the deformation center; the white circle represents the undeformed state; the yellow ellipse describes the deformed state; the major/minor axes of the ellipse represent the two principal deformation directions detected, with a red line segment representing a stretch and a green line segment representing a compression.

from two video clips, it is not required that the entire facial motions are the same, but it is required that the motions are locally similar at least in some parts of the faces.

Figure 4.1 illustrates an example of individual differences in local deformation patterns observed in facial expressions. In the figure, video (a) and (b) are from the same human subject but with different facial expressions. Video (a) and (c) are from different human subjects but with the same facial expression. Note in (d): although

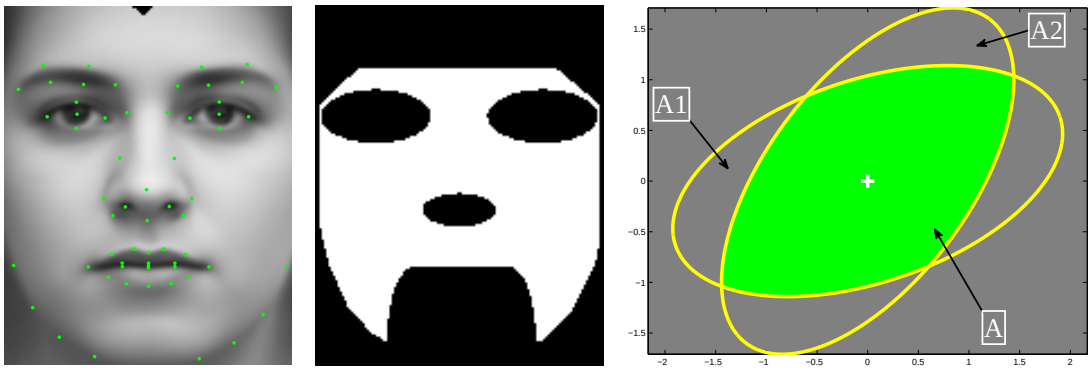
the solid blue line (video (a)) and the solid yellow line (video (c)) are more similar in terms of the trajectory (probably due to the same undergoing expression) the solid blue line (video (a)) shares more similarity with the dashed blue line (video (b)) in terms of the deformation patterns, as suggested by the shape of the ellipses (probably due to the same identity). The facial expressions of surprise and fear are globally different, but they are locally similar around the eyebrows. Finally, please note that this is an example for a better understanding of the hypothesis. The arguments made in this study are NOT based on this single instance.

4.1.1 Representation of Deformation Patterns

The Right Cauchy-Green deformation tensor [Bowen 1989] is employed to describe deformation patterns. The neutral face is used as the initial state. Let u denote the 2D displacement field which changes the neutral face to a specific deformed face, then the deformation tensor C is computed as,

$$C = \nabla u^T \nabla u + \nabla u^T + \nabla u + I, \tag{4.1}$$

where I is an identity matrix; ∇ is the gradient operator. Although in physics both u and C are supposed to be continuous in space, in the numerical implementation, u is defined on each pixel in the neutral face and thus, so is C. The two orthogonal eigenvectors of C give the two principal deformation directions. And the square-root of the corresponding eigenvalue measures the deformation magnitude. If the eigenvalue is smaller than one, a compression is observed. If the eigenvalue is larger than one, a stretch is observed. Such a deformation pattern can be well represented by an ellipse (Figure 4.1(f)). The direction and length of the major/minor axes of the



(a) Mean face (b) Ω: Region of interest (c) Overlap of two ellipses

Figure 4.2:

ellipse are determined by the eigenvectors and the square-roots of the eigenvalues of C, respectively.

4.1.2 From Facial Motion Videos to LDP

The LDP of a human subject is defined to be a set of deformation-displacement pairs. Mathematically,

$$\mathscr{P} = \{(C_{x,t}, u_{x,t})\}, \tag{4.2}$$

where x denote a pixel in the shape-normalized neutral face image of the subject and t is an index. Please note that there is no order for the elements in set $\mathscr{P}$. The index t is used to refer to different deformed state of the face. $u_{x,t}$ excludes rigid head motion. Figure 4.3(a) illustrates two LDPs on a specific pixel x. A facial motion video of N frames which starts with a neutral face can provide $(N-1)\times|\Omega|$ pairs of deformation and displacement for the LDP of the subject, where $|\Omega|$ denotes the number of pixels within the region of interest Ω. Given a frontal-view facial motion video which is assumed to start with a neutral face, LDP is extracted in the following steps:

1. Use a face detection and localization algorithm to find a set of key points on the neutral face (the first frame of the video);
2. Remove any rigid head motion from the video;
3. Crop the face region from the video to get a cropped face image sequence;
4. Track each pixel in the region of interest (Figure 4.2(b)) on the neutral face (the first cropped face image) throughout the image sequence to obtain its displacement in each frame;
5. Warp the displacement fields defined on the neutral face using a transformation which normalizes the face shape to a given mean face shape (Figure 4.2(a));
6. From the shape-free displacement fields, construct LDP (Eq.(4.1) and Eq.(4.2)).

In current implementation, the STASM [Milborrow and Nicolls 2008] library is used for Step 1. Step 2 is skipped, because the videos in current dataset (from the Cohn-Kanade Facial Expression Database [Kanade et al. 2000]) contain very slight and negligible head motion. However, Step 2 can be difficult in general settings. In Step 3, face images are resized to 128 by 160 pixels. Lucas-Kanade optical flow estimation [Lucas and Kanade 1981] with pyramidal refinement is used for tracking (Step 4). Figure 4.2(a) shows the mean face and its key points, which is computed from all the neutral faces found in current dataset. And in the experiment, LDP is computed from the pixels inside the region of interest only (Figure 4.2(b)). There are two reasons for using this region of interest. First, some regions of the face, like the forehead, are not enclosed by the key points. During warping (Step 5), the displacement vectors in those regions are extrapolated, which introduces extra

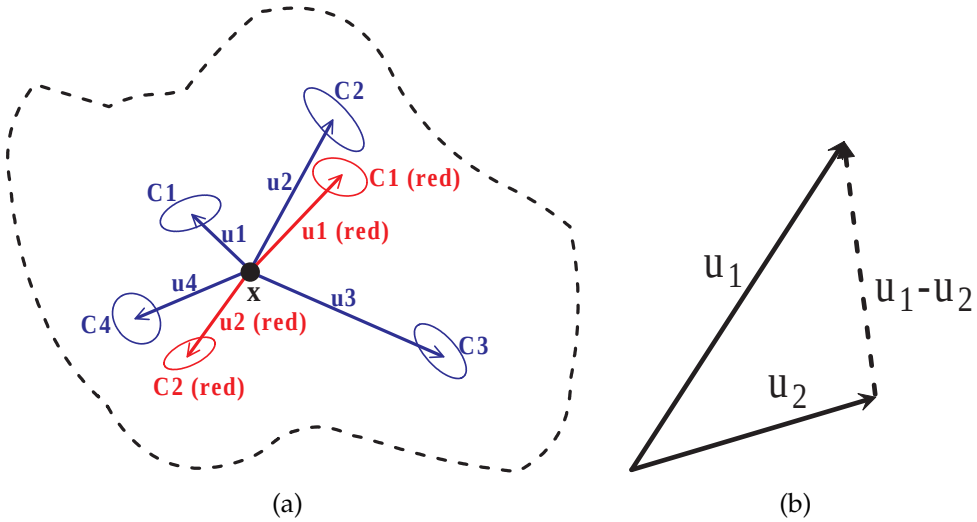


Figure 4.3: (a) Matching the red LDP against the blue LDP on pixel x: an LDP is a set of deformation-displacement pairs (Eq.(4.2)). Suppose the red LDP is being matched against the blue LDP, firstly, for each u in red, a closest u in blue must be found and then the similarity between their corresponding C can be measured. Thus, in this particular example, $C1(red)$ will be compared with $C2$ and $C2(red)$ will be compared with $C4$. (b) A relative vector difference measurement: $r = |u_1 - u_2|/(|u_1| + |u_2|)$.

	low s_m	high s_m
low s_d	not sure	imposter detected
high s_d	not sure	genuine detected

Table 4.1: An intuitive understanding of s_m and s_d

distortion. Second, some regions of the face, like the eyes and chin, can be occluded or move out of the image in some facial motion. For example, eyes will be occluded when blinking occurs and chin will move out of the image with a wide open mouth (*e.g.* in a facial expression of surprise). Please note that LDP is purely motion-based and does not contain any appearance information (color, shape, *etc.*).

4.1.3 Similarity between Two LDPs

Figure 4.3(a) shows an example of comparing two LDPs at a specific pixel. In order to measure the similarity between two LDPs, both deformation similarity and motion similarity have to be considered. Intuitively, while a small deformation similarity (*i.e.* big difference in deformation patterns) suggests a difference in identity, a small motion similarity (*i.e.* big difference in displacement vectors) will suggest that the deformation patterns are not comparable at all, because the two deformation patterns are caused by very different local motions. And different motions will result in different deformation patterns even when they are performed by the same face, since human face is a non-linear and anisotropic elastic surface. In this sense, motion similarity can be considered as a confidence score about the deformation similarity measurement. Only when motion similarity is high will it be possible to obtain reliable results from the deformation similarity measurement. Table 4.1 summarizes this intuitive understanding of deformation similarity s_d and motion similarity s_m. The overall s_d and s_m are computed as weighted averages of local deformation similarity and local motion similarity which are measured on each pixel, respectively,

$$s_d = \sum_{x \in \Omega} w(x) s_d(x), \quad s_m = \sum_{x \in \Omega} w(x) s_m(x), \tag{4.3}$$

$$w(x) = \frac{s_m(x)}{\sum_{x \in \Omega} s_m(x)} \ , \tag{4.4}$$

where x denote a pixel in the region of interest Ω (Figure 4.2(b)). Normalized motion similarity serves as the weight. The computation of local deformation/motion similarity is explained in the following subsections.

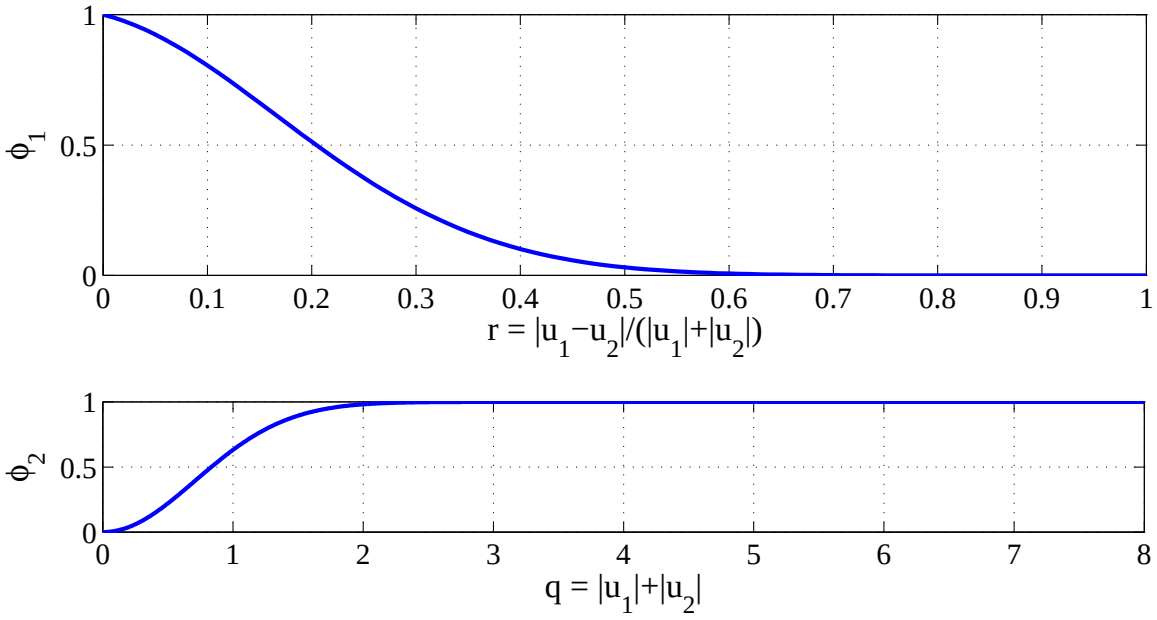


Figure 4.4: ϕ_1: penalty on motion similarity due to large vector difference (Figure 4.3(b)); ϕ_2: penalty on motion similarity due to small displacement. Please read the part of *Local Deformation Similarity* in Section 4.1.3 for details.

Local Deformation Similarity

Based on the elliptical representation (Section 4.1.1), function ψ for comparing two deformation patterns is defined as,

$$\psi(C_1, C_2) = \frac{A}{A_1 + A_2}, \tag{4.5}$$

where A_1 and A_2 are the areas of the ellipses which represent the two deformation patterns, C_1 and C_2, respectively; A is the area of the overlap of the two ellipses after being translated to be concentric. The computation of A has been elaborated in Appendix A. Figure 4.2(c) shows an example of A_1, A_2 and A. Now, to match $\mathscr{P}_A = \{(C^A_{x,t}, u^A_{x,t})\}$ against $\mathscr{P}_B = \{(C^B_{x,t}, u^B_{x,t})\}$ on pixel x, it has to be found first that for each local motion u^A_{x,t_1} in $\mathscr{P}_A$, the most similar local motion in $\mathscr{P}_B$ (on the same

pixel). Mathematically,

$$\tilde{t_2}(u^A_{x,t_1}) = \arg\max_{t_2} \phi(u^A_{x,t_1}, u^B_{x,t_2}) \quad (4.6)$$

$$\phi(u_1, u_2) = \phi_1(u_1, u_2) \cdot \phi_2(u_1, u_2) \quad (4.7)$$

$$\phi_1(u_1, u_2) = (1 - r)exp(-r^2/\sigma_1^2), \quad (4.8)$$

$$r = |u_1 - u_2|/(|u_1| + |u_2|), \quad (4.9)$$

$$\phi_2(u_1, u_2) = 1 - exp(-q^2/\sigma_2^2), \quad (4.10)$$

$$q = |u_1| + |u_2|, \quad (4.11)$$

where $|\cdot|$ denotes l^2-norm; r is a commonly used relative measurement of vector difference (Figure 4.3(b)); ϕ_1 alters the value of r so that larger difference will be penalized more severely (Figure 4.4); ϕ_2 is a penalty for small displacement vectors (Figure 4.4); ϕ_2 is necessary because when the displacement is small, the deformation pattern is also slight and thus does not provide much personal characteristics; σ_1 and σ_2 are two parameters which are set to 0.3 and 1.0 in all of the experiments; ϕ is the motion similarity between two displacement vectors; $0 \leq \phi \leq 1$ and a bigger value indicates a higher motion similarity. Since motion similarity can be considered as the confidence of the deformation similarity measurement (Section 4.1.3, first paragraph), the motion similarity scores are converted to normalized weights,

$$w(x, t_1) = \frac{\phi(u^A_{x,t_1}, u^B_{x,\tilde{t_2}(u^A_{x,t_1})})}{\sum_i \phi(u^A_{x,i}, u^B_{x,\tilde{t_2}(u^A_{x,i})})}. \quad (4.12)$$

And finally, the local deformation similarity at pixel x is computed as the weighted average,

$$s_d(x) = \sum_{t_1} w(x, t_1)\psi(C^A_{x,t_1}, C^B_{x,\tilde{t_2}(u^A_{x,t_1})}). \tag{4.13}$$

Please note that this local deformation similarity measurement is not symmetric. It is assumed here that $\mathscr{P}_A$ is being matched against $\mathscr{P}_B$.

Local Motion Similarity

Given two LDPs, $\mathscr{P}_A = \{(C^A_{x,t}, u^A_{x,t})\}$ and $\mathscr{P}_B = \{(C^B_{x,t}, u^B_{x,t})\}$, the local motion similarity measured on a pixel x is computed as follows,

$$s_m(x) = \sum_{t_1} w(x, t_1)\phi(u^A_{x,t_1}, u^B_{x,\tilde{t_2}(u^A_{x,t_1})}), \tag{4.14}$$

where w_{t_1}, ϕ, $\tilde{t_2}$ are defined in previous section (the part of *Local Deformation Similarity* in Section 4.1.3).

Verification Score

The verification score is defined to be the multiplication of overall motion similarity and deformation similarity,

$$s(\mathscr{P}_A, \mathscr{P}_B) = s_m(\mathscr{P}_A, \mathscr{P}_B) \cdot s_d(\mathscr{P}_A, \mathscr{P}_B). \tag{4.15}$$

No threshold is set on s_m, because such a threshold may vary from case to case, depending on the types of facial motion covered in training/testing. $s_m(\mathscr{P}_A, \mathscr{P}_B)$ denotes the motion similarity measured with $\mathscr{P}_A$ against $\mathscr{P}_B$, similarly $s_d(\mathscr{P}_A, \mathscr{P}_B)$. And $s(\mathscr{P}_A, \mathscr{P}_B)$ is the verification score of matching $\mathscr{P}_A$ against $\mathscr{P}_B$, with $0 \leq$

$s(\mathscr{P}_A, \mathscr{P}_B) \leq 1.0$. A higher score means a higher similarity in identity. We adopt a simple verification rule in our experiment: if $s(\mathscr{P}_A, \mathscr{P}_B) \geq \theta$, subject A is considered as the same person as subject B; if $s(\mathscr{P}_A, \mathscr{P}_B) < \theta$, subject A is considered as a different person from subject B. In the experiments, θ is varied from 0.0 to 1.0 to draw a FAR-FRR (false accept rate vs. false reject rate) curve.

4.2 Experiments

For the experiments reported in this section, a set of videos from the Cohn-Kanade Facial Expression Database [Kanade et al. 2000] is used. The Cohn-Kanade database has been chosen for three reasons. First, it is AU-coded (Action Unit) so that it is possible to run analysis on the relation between motion similarity and AU score distance (Section 4.2.2). Second, among all available databases, it provides the best balance between facial motion variation and identity variation while other databases contain either much motion variation from only a few subjects or only one or two facial motions from many subjects. An ideal database for motion-based face recognition experiment should contain both rich facial motion variation and rich identity variation, but unfortunately, such a database does not exist. Third, rigid head motion is very slight and negligible in most of the videos in the Cohn-Kanade database.

4.2.1 The Dataset

The Cohn-Kanade database contains facial expression videos collected from 97 subjects. All video clips have been AU-coded using FACS (Facial Action Coding System [Ekman and Friesen 1978]). The AU scores are translated to expression

(a) surprise (b) anger (c) happy (d) sad (e) fear (f) disgust

Figure 4.5: Examples of the six basic facial expressions

tags in the experiment. Each subject has three to eight recordings, which include all or part of following facial expression/motion: happy, sad, surprise, fear, anger, disgust and mouth-open. Most subjects have only one recording for each facial expression/motion. Rigid head motion is not noticeable in most of the recordings.

4.2.2 Experiment 1: Pair-wise Cross-Expression Face Verification

In this experiment, the performance of LDP is evaluated in a "train on one expression, test on another" setting, *i.e.* training on one type of facial expression and testing on another type of facial expression. The six basic facial expressions of emotion [Ekman 1972] are covered in this experiment, namely, happy, sad, surprise, fear, anger and disgust. This is a very challenging experimental setting, because one type of facial motion can not provide much discriminant information and also the six expressions are actually very different from each other. In order to have a fair comparison, all the pair-wise tests must be run over the same set of subjects. Thus,

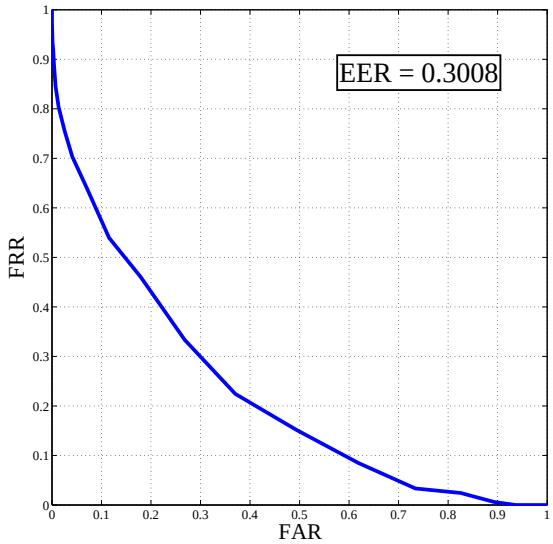
only the subjects whose video data cover all the six expressions are picked. This results in a dataset which includes 11 subjects with each subject having one video recording for each of the six facial expressions. Figure 4.5 shows two sets of examples of the six basic facial expressions. Table 4.2 lists the pair-wise cross-expression face verification performance in terms of the EER (equal error rate). Column averages are also included in the table. Among all the six facial expressions, happy has the smallest column average, which suggests happy the best choice when only one facial expression is allowed to be used for training. Another interesting fact about the table is that it is somehow symmetric. For example, the EERs of sad-anger and anger-sad are both small while the EERs of fear-anger and anger-fear are both large. This kind of symmetry is expected, because LDP is more sensitive to the similarity between the training/testing motions rather than the motions themselves. In this sense, a large EER in the table may also suggest a less similarity between the pair of involved facial expressions. Figure 4.6(a) provides an overall FAR-FRR curve of the pair-wise cross-expression face verification. The overall EER is 0.3008, which is significantly above chance (*i.e.* EER=0.5). Moreover, it is even better than the previous result reported in fixed-motion experiment (EER=0.4 by Tulyakov *et al*. [2007]). And please note that Tulyakov *et al*. [2007] required fixed facial motion, while our result is from cross-expression face verification, which is much more difficult and has never been reported before.

In this experiment, the relation between the motion similarity (*i.e.* s_m) and the AU score distance has also been investigated. The term, AU score distance, is used to refer to the distance between the AU scores of two video recordings of facial expression (all video recordings in the Cohn-Kanade database are AU-coded). To compute the AU score distance, a vector is used to encode the AU score of a video

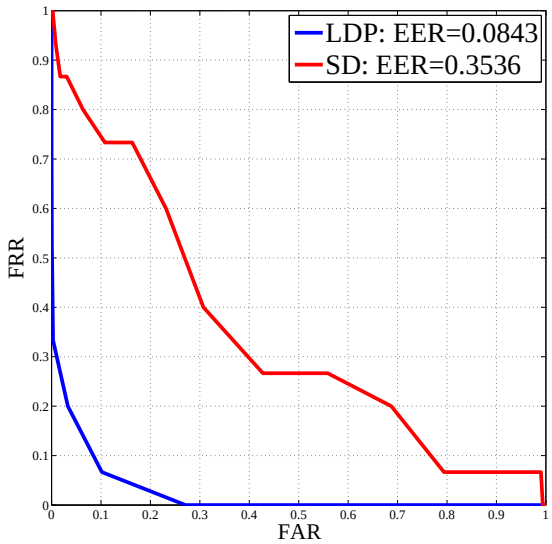
test \ train	surprise	anger	happy	sad	fear	disgust
surprise	N/A	0.2318	0.2000	0.3273	0.2818	0.2864
anger	0.2727	N/A	0.2955	0.1773	0.4182	0.3455
happy	0.3091	0.3636	N/A	0.4045	0.2182	0.1545
sad	0.3227	0.1182	0.3500	N/A	0.3318	0.3955
fear	0.2500	0.4318	0.2000	0.3364	N/A	0.2545
disgust	0.3136	0.3318	0.2455	0.3182	0.2273	N/A
column avg.	0.2936	0.2954	0.2582	0.3127	0.2955	0.2873

Table 4.2: Experiment 1 pair-wise cross-expression face verification result: the equal error rates

recording. Each dimension of the vector corresponds to one AU. If an AU is not activated in this video, 0 is given to the corresponding dimension. Otherwise, the magnitude of a dimension is extracted from the corresponding AU magnitude. There are five degrees of AU magnitude, marked from 'a' to 'e' in FACS. In the experiment, 'a' is translated to 0.1, 'b' to 0.2, 'c' to 0.4, 'd' to 0.7, 'e' to 0.9 and if there is no magnitude marked (meaning at least 'b'), 0.5 is used. This rough quantization is based on the explanation in the 2002 version of FACS manual [Ekman et al. 2002]. The AU score distance is computed as the Euclidean distance between the vectors. Then the Pearson correlation coefficient between the motion similarity and the AU score distance is computed. The result, -0.35, implies that the motion similarity and the AU score distance are correlated, but the correlation is not strong. The correlation is negative because larger AU score distance means less similarity. Considering that AU score is also a measurement of local facial motion (though not a quantitative measurement in nature), it is not surprising to see the presence of this correlation. However, the correlation is weaker than expected. There may be two reasons behind this. First, AUs are not independent from each other in terms of observable local motion. For example, AU1 (inner brow raiser) and AU2



(a) Experiment 1: overall FAR-FRR plot of pairwise cross-expression face verification



(b) Experiment 2: FAR-FRR plot from face verification using fixed facial expression

Figure 4.6: FAR-FRR plots for Experiment 1 and 2

(outer brow raiser) overlap in the middle part of the eyebrow region and both can cause wrinkles in the forehead region. This kind of dependence is hard to measure and is not covered in the quantization of the AU scores. Second, s_m is not a linear measurement of local motion similarity with respect to the difference in displacement. Instead, s_m stems from altering the relative vector difference measurement r using function ϕ_1 (Eq.4.6) and ϕ_1 is non-linear (Figure 4.4). Since the correlation between s_m and AU score distance is weak, AU scores are not used to guide the experiments.

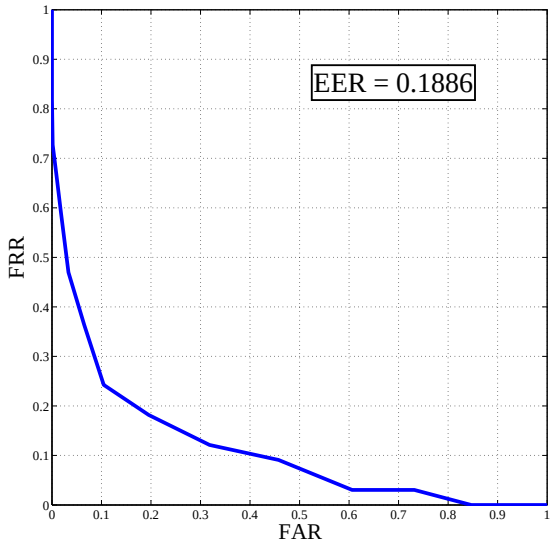
4.2.3 Experiment 2: Fixed Facial Expression

This experiment tests the performance of LDP in fixed facial motion. This allows us to compare with the smile dynamics of Chapter 3. In the Cohn-Kanade dataset, there are 15 subjects who have a second recording of happy facial expression (with

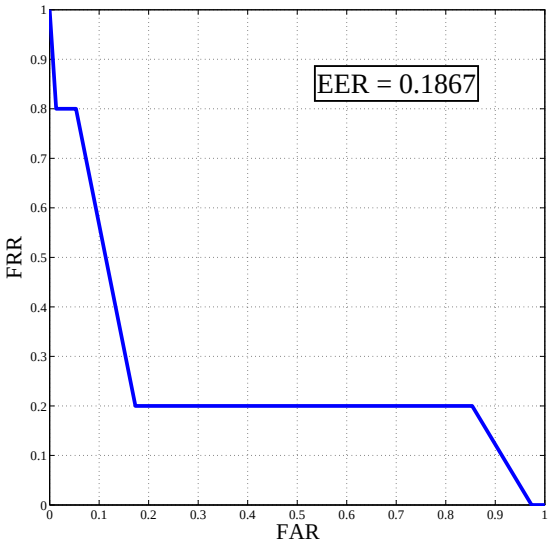
mouth open). In this experiment, these 15 video clips are verified against the 88 first recordings of happy expression (with mouth closed) from 88 subjects (including the 15 subjects). Figure 4.6(b) shows the FAR-FRR curve (the blue line) from this test. The EER is around 0.0843, which is significantly better than the previous result reported by Tulyakov *et al*. [Tulyakov et al. 2007] (EER=0.4). For comparison, smile dynamics (Chapter 3) is tested on the same set of data (the red line in Figure 4.6(b)). The performance of smile dynamics is much worse than LDP in this test. The main reason seems to be that in this dataset, the first recording (used for training) and the second recording (used for testing) of happy facial expression are different around the mouth (which is closed in the first recording and open in the second one). And smile dynamics seems to be highly sensitive to even small difference between the facial motion used in training and testing.

4.2.4 Experiment 3: Using More Facial Expressions for Training

This experiment is conducted over the same set of videos as that used in Experiment 1 (11 subjects, 6 expressions from each, see Section 4.2.2). But this time, LDP is trained on five out of six facial expressions and test on the one remaining facial expression. The experiment is repeated six times and leave each facial expression out in turn. The purpose of this experiment is to verify if there will be a boost in performance when more motion data is used for training. The test result is to be compared with the result obtained in Experiment 1. Figure 4.7(a) plots the overall FAR-FRR curve obtained from this experiment. The EER is around 0.1886, which shows a considerable improvement over the results from Experiment 1 (EER=0.3008).



(a) Experiment 3: overall FAR-FRR plot from "train on more" cross-expression face verification



(b) Experiment 4: FAR-FRR plot from face verification under heavy face makeup

Figure 4.7: FAR-FRR plots for Experiment 3 and 4

4.2.5 Experiment 4: Face Verification under Heavy Face Makeup

One of the major benefits that face recognition researchers are expecting from motion-based approaches is the ability of identification even when appearance information is severely distorted, for instance, by extreme lighting conditions or by heavy face makeup. This experiment tests the performance of LDP under extremely heavy face makeup (Figure 4.8). The dataset from Experiment 1 (11 subject, six facial expressions from each, see Section 4.2.2) is re-used. But this time, all expressions are used for training. In addition, a group of painted face videos are collected. Specifically, three sets of facial expression recordings are collected from five subjects, the first and the second normal face video sets and the painted face video set. The six basic facial expressions are covered in each set. The first normal face video set is used for training. Thus, in total, there are 16 reference subjects.

(a) surprise (b) anger (c) happy (d) sad (e) fear (f) disgust

Figure 4.8: An example of facial expressions with heavy face makeup: several sets of these data from five subjects are collected for the experiment. The faces of all subjects are painted with the same pattern which is commonly seen in Beijing Opera.

The painted face video set is used in testing. Figure 4.7(b) shows the FAR-FRR curve of this test. The EER is around 0.1867, which is reasonably good considering that the painting can actually reduce the accuracy of STASM (for face localization) as well as the accuracy of optical flow estimation, even though LDP itself does not contain any appearance information. In the worst case that either STASM or optical flow estimation completely fails, LDP, which is built upon the two routines, shall completely fail as well. For comparison, the experiment is repeated with the second normal face video set used for testing. EER=0 is obtained this time, which means all subjects are successfully recognized. Although this can be due to the small size of the testing dataset (five subjects), please note that the motion information contained in this test (six expressions for both training and testing) is much richer than those in any of the previous experiments.

Unlike in Section 3.2, experiments are not run here to compare the discriminat-

ing power of the upper face and the lower face for LDP. The reason is that for LDP which one is more discriminating really depends on the two sets of facial motion used for training and testing respectively. If the two sets of facial motion share more similar local motion in the upper face than in the lower face, then the upper face is more discriminating than the lower face, and vice versa.

4.3 Discussion

Several conclusions can be drawn from the experimental results.

First, and most important: cross-expression motion-based face recognition is possible. In the first and the third experiments (Section 4.2.2 and Section 4.2.4), although the facial expressions used for training and the facial expressions used for testing are completely different, LDP shows above-chance verification performance (EER=0.3008 and EER=0.1886). Moreover, the performance is even better than the previously reported performance from a fixed expression face verification test [Tulyakov et al. 2007] (EER=0.4). This finding implies the possibility of developing a general motion-based face recognition algorithm which can identify a human subject from any kind of facial motion. Such an algorithm will be very useful to enhance and promote face recognition technology in practice, because facial motion is a nuisance to existing still-image face recognition systems while in contrast, motion-based approaches are exploiting facial motion as a biometric trait. A combination of the two may give a more robust face recognition system.

Second, LDP can help face recognition under heavy face makeup. In the fourth experiment (Section 4.2.5), although all the subjects have their faces completely painted in testing (not in training), LDP shows an above-chance face verification

performance (EER=0.1867). This suggests that LDP is effective even when the face is disguised. However, extremely severe appearance distortion can still impact the performance of LDP, because appearance distortion can jeopardize the extraction of motion features, even though LDP does not contain any appearance information. When the second set of normal unpainted faces is used in testing, EER=0.0 is obtained in the experiment. This shows that face makeup does increase EER (from 0.0 to 0.1867).

Third, there seems to be an approximate relation between the performance of LDP and the similarity between the facial motion covered by training and testing videos. In Experiment 1 (Section 4.2.2), EER is 0.3008 when the training videos contain only one type of facial expression and the testing videos contain another type of facial expression (most un-similar motions). In Experiment 3, EER is 0.1886 when the training videos contain five facial expressions and the testing video contains the remaining one facial expression (higher chances of local motion similarity due to the more facial expressions covered in training). In Experiment 2, EER is 0.0843, when both training and testing videos contain only the facial expression of happy (almost same facial motion). And finally, in Experiment 4, EER is 0 when both training and testing videos contain the six facial expressions (almost same facial motions). Although currently, the dataset is too small to support any quantitative analysis over the relation between the performance and the training-testing motion similarity, there is probably a very close relation. Because LDP looks for identity evidence from face regions where similar local motions are observed and intuitively, *globally* similar motions should also be locally similar. A further implication of this approximate relation is that if someone wants to train an LDP which can be used for recognizing a human subject from any kind of facial motion,

the LDP must be built from a set of videos that contains all possible *local* facial motions of this subject. This idea introduces one interesting question: does there exist a smallest set of facial expression that captures all possible local motions? Such a set can be named as *Minimum Spanning Set for Facial Motion*, whose investigation is left for future work. On the other hand, this approximate relation also suggests the extent, to which LDP would be viable, that is: if the facial motion covered by training and testing videos share no similarity at all, then LDP will fail. In other words, LDP requires the training and the testing facial motion to be locally similar somewhere on the face. And the more local similarity they share, the more discriminative LDP will be.

4.4 Summary

This chapter has reported a study on a novel motion-based face recognition approach, the Local Deformation Profile (LDP). LDP is the first approach in the field which can be used for cross-motion face recognition tasks. That is, with LDP, it is possible to learn human identity from one type of facial motion and later verify human identity from another type of facial motion - as long as the two types of facial motion are locally similar in some parts of the face. The performance of LDP has been evaluated through several experiments conducted over a facial expression video database. The experimental results have shown its potential of being a biometric trait. Moreover, LDP can also help face recognition when extremely heavy face makeup is present.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

As stated in Introduction, Chapter 1, regarding motion-based face recognition, there are three basic questions to be answered:

1. Under which condition is motion-based face recognition viable?
2. If it is viable, what features should be used?
3. How discriminating are the features?

The studies reported in Chapter 3 and 4 provide partial answers to the questions, which have been summarized in Table 5.1. Smile Dynamics and Local Deformation Profile are the two novel facial motion-based features which have been proposed in this thesis. Compared with other techniques in the field (Chapter 2), they are novel in two main aspects:

- Smile Dynamics: the first approach which makes use of facial expression dynamics for identification;

- Local Deformation Profile: the first approach which can carry out cross-motion identification.

Other minor contributions of the thesis include:

- Shows that with smile dynamics, lower face is more discriminating than upper face, which is different from the conclusions by conventional face recognition studies (in Section 3.2.4);
- Proposes the first motion-appearance hybrid feature and shows that it is more discriminating than either of its components, *i.e.* motion-based feature and appearance-based feature (in Section 3.3);
- Shows that the hybrid feature may help distinguish identical twins (in Section 3.4.5);
- Shows that local deformation profile can work under extremely heavy face makeup (in Section 4.2.5).

5.2 Future Work

Motion-based face recognition is a young research area. There is still a large amount of work to be done, which include and may not be limited to the following ones,

1. A good facial motion database is still lacking. Existing facial motion databases are all collected for the purpose of facial expression recognition. Compared with an experiment on facial expression recognition, an experiment on motion-based face recognition requires more variation in facial motion and also larger number of recordings for each type of facial motion.

Feature	Working Condition	Performance
Smile Dynamics	• Frontal view; • Facial motion of smile; • Proper lighting for face detection and tracking to work.	Estimated Bayes' error rate: between 0.0015 and 0.0029
Local Deformation Profile	• Frontal view; • Any facial motion; • Proper lighting for face detection and tracking to work.	Performance grows as the local motion similarity between training and testing facial motions increases; estimated equal error rate varies between 0.08 and 0.3 depending on the mutual local motion similarity.

Table 5.1: Answers to the questions: summary of the features

2. The research area of 3D motion-based face recognition is worthy of exploration. Extending existing technique to 3D may not be hard. But data acquisition and post-processing can be more tricky and complicated in 3D.

3. The existence of a *Minimum Spanning Set for Facial Motion* (please read the last paragraph in Section 4.3 for details) is worthy of further study. If such a set of facial motions can be found, it will become much easier for researchers to develop and deploy motion-based face recognition systems in practice.

4. Motion-based face recognition under different head poses has not been studied. Considering that head pose is still a huge problem for conventional face recognition after so many years of research, it will not be an easy task for motion-based approaches, too.

5. Telling apart identical twins by using facial motion cues is an interesting topic. A really small-scale experiment has been reported in this thesis, but it is merely a preliminary study. To further explore this area, a large dataset of facial motion videos from identical twins is essential. However, such a dataset will be hard to get, considering that identical twins are not easy to find in large numbers.

Bibliography

Bowen, R. M. 1989. *Introduction to Continuum Mechanics for Engineers*. Plenum Press.

Bruce, V., and Valentine, T. 1988. When a nods as good as a wink: The role of dynamic information in facial recognition. *Practical Aspects of Memory: Current Research and Issues 1*, 169–174.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P., Burton, A., and Miller, P. 1999. Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied 5*, 339–360.

Bruce, V., Henderson, Z., Newman, C., and Burton, M. 2001. Matching identities of familiar and unfamiliar faces caught on cctv images. *Journal of Experimental Psychology: Applied 7*, 207–218.

Chen, L.-F., Liao, H.-Y., and Lin, J.-C. 2001. Person identification using facial motion. In *Proceedings of International Conference on Image Processing*.

Christie, F., and Bruce, V. 1998. The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition 26*, 780–790.

Cover, T., and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 21–27.

Duda, R. O., Hart, P. E., and Stork, D. G. 2000. *Pattern Classification (2nd ed.)*. Wiley Interscience.

Ekenel, H. K., and Stiefelhagen, R. 2006. Block selection in the local appearance-based face recognition scheme. In *Proceedings of IEEE Computer Vision and Pattern Recognition Workshop*.

Ekman, P., and Friesen, W. V. 1978. *Facial Action Coding System*. Consulting Psychologists Press, Inc.

Ekman, P., Friesen, W. V., and Hager, J. C. 2002. *FACS Manual (2002)*.

Ekman, P. 1972. Universals and cultural differences in facial expression of emotion. In *Nebraska Symposium on Motivation*, University of Nebraska Press, Lincoln, Nebraska, J. Cole, Ed., 207–283.

Gross, R., Shi, J., and Cohn, J. F. 2001. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*.

Hill, H., and Johnston, A. 2001. Categorizing sex and identity from the biological motion of faces. *Current Biology 11*, 880–885.

Jain, A., Dass, S., and Nandakumar, K. 2004. Can soft biometric traits assist user recognition? In *Proc. SPIE, Biometric Technology for Human Identification*.

Jain, A., Ross, A., and Pankanti, S. 2006. Biometrics: a tool for information security. *Information Forensics and Security, IEEE Transactions on 1*, 2 (June), 125–143.

Kanade, T., Cohn, J. F., and Tian, Y. 2000. Comprehensive database for facial expression analysis. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*.

Knight, B., and Johnston, A. 1997. The role of movement in face recognition. *VISUAL COGNITION 4*, 265–273.

Lander, K., and Bruce, V. 2000. Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology 12*, 259–272.

Lander, K., and Bruce, V. 2004. Repetition priming from moving faces. *MEMORY & COGNITION 32*, 640–647.

Lander, K., and Chuang, L. 2005. Why are moving faces easier to recognize? *Visual Cognition 12*, 429–442.

Lander, K., Christie, F., and Bruce, V. 1999. The role of movement in the recognition of famous faces. *Memory & Cognition 27*, 974–985.

Lander, K., Bruce, V., and Hill, H. 2001. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied cognitive psychology 15*, 101–116.

Lander, K., Chuang, L., and Wickham, L. 2006. Recognizing face identity from natural and morphed smiles. *The Quarterly Journal of Experimental Psychology 59*, 801 – 808.

Lucas, B. D., and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In *Proc. DARPA Image Understanding Workshop*.

Milborrow, S., and Nicolls, F. 2008. Locating facial features with an extended active shape model. In *Proceedings of European Conference on Computer Vision*.

Pamudurthy, S., Guan, E., Mueller, K., and Rafailovich, M. 2005. Dynamic approach for face recognition using digital image skin correlation. In *Proc. Audio- and Video-Based Biometric Person Authentication*.

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. 2005. Web-based database for facial expression analysis. In *Proc. ICME*.

Pike, G., Kemp, R., Towell, N., and Phillips, K. 1997. Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition 4*, 409437.

Roark, D. A., Barrett, S. E., Spence, M., Abdi, H., and O'Toole, A. J. 2003. Memory for moving faces: Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and Cognitive Neuroscience Reviews 2*, 15–46.

Schiff, W., Banka, L., and de Bordes-Galdi, G. 1986. Recognizing people seen in events via dynamic "mug shots". *American Journal of Psychology 99*, 219–231.

Shepherd, J. W., Davies, G., and Ellis, H. 1982. *Identification evidence : a psychological evaluation*. Aberdeen University Press.

Sinha, P., Balas, B., Ostrovsky, Y., and Russell, R. 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE 94*, 11 (Nov.), 1948–1962.

Thornton, I., and Kourtzi, Z. 2002. A matching advantage for dynamic human faces. *Perception 31*, 113–132.

Tulyakov, S., Slowe, T., Zhang, Z., and Govindaraju, V. 2007. Facial expression biometrics using tracker displacement features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Wallhoff, F., 2006. Facial expressions and emotion database. www.mmk.ei.tum.de/~waf/fgnet/feedtum.html, Technische Universität München.

Ye, N., and Sim, T. 2008. Smile, you're on identity camera. In *Proceedings of International Conference on Pattern Recognition*.

Ye, N., and Sim, T. 2009. Combining facial appearance and dynamics for face recognition. In *Proceedings of International Conference on Computer Analysis of Images and Patterns*.

Ye, N., and Sim, T. 2010. Towards general motion-based face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Y., Kundu, S., Goldgof, D., Sarkar, S., and Tsap, L. 2004. Elastic face - an anatomy-based biometrics beyond visible cue. In *Proceedings of International Conference on Pattern Recognition*.

Appendix A

Overlap of Two Concentric Ellipses

Write the first ellipse in standard form,

$$\frac{x^2}{a_1^2} + \frac{y^2}{b_1^2} = 1. \tag{A.1}$$

And write the second ellipse in parametric form, which has been rotated anti-clockwisely around the origin by an angle of θ,

$$\begin{cases} x &= a_2 \cos t \cos\theta - b_2 \sin t \sin\theta \\ y &= a_2 \cos t \sin\theta + b_2 \sin t \cos\theta \end{cases}. \tag{A.2}$$

Let

$$c = a_2 \cos\theta, \quad d = b_2 \sin\theta, \quad e = a_2 \sin\theta, \quad f = b_2 \cos\theta, \quad r = \tan(t/2), \tag{A.3}$$

because

$$\cos t = \frac{1 - \tan^2(t/2)}{1 + \tan^2(t/2)}, \quad \sin t = \frac{2\tan(t/2)}{1 + \tan^2(t/2)}, \tag{A.4}$$

Eq.(A.2) can be re-written as,

$$\begin{cases} x &= \dfrac{c - 2dr - cr^2}{1 + r^2} \\ \\ y &= \dfrac{e + 2fr - er^2}{1 + r^2} \end{cases} . \tag{A.5}$$

Substituting Eq.(A.5) to Eq.(A.1) yields

$$[b_1(c - 2dr - cr^2)]^2 + [a_1(e + 2fr - er^2)]^2 = [a_1 b_1(1 + r^2)]^2, \tag{A.6}$$

which is a quartic equation of r and can be solved by Ferrari's method. In implementation, a symbolic solver provided by Matlab is employed to help solve this equation. If Eq.(A.6) has zero or two roots, the overlap area

$$A = \min(A_1, A_2), \tag{A.7}$$

where

$$A_1 = \pi a_1 b_1, \quad A_2 = \pi a_2 b_2 \tag{A.8}$$

are the areas of the first and the second ellipses, respectively. If Eq.(A.6) has four roots, the overlap area

$$A = 2 \cdot \min(B_1, B_2), \tag{A.9}$$

where

$$B_1 = \frac{1}{2}(t_{1,2} - t_{1,1})a_1 b_1 + \frac{1}{2}(t_{2,3} - t_{2,2})a_2 b_2, \tag{A.10}$$

$$B_2 = \frac{1}{2}(t_{2,2} - t_{2,1})a_2 b_2 + \frac{1}{2}(t_{1,3} - t_{1,2})a_1 b_1, \tag{A.11}$$

and

$$0 \leq t_{2,i} = 2 \cdot \arctan r_i < 2\pi, \tag{A.12}$$

$$0 \leq t_{1,i} = \arctan(\frac{a_1}{b_1}\frac{y_i}{x_i}) < 2\pi, \tag{A.13}$$

$$x_i = a_2 \cos t_{2,i} \cos\theta - b_2 \sin t_{2,i} \sin\theta, \tag{A.14}$$

$$y_i = a_2 \cos t_{2,i} \sin\theta + b_2 \sin t_{2,i} \cos\theta, \tag{A.15}$$

$$i = 1, 2, 3, 4, \tag{A.16}$$

where (x_i, y_i) denote the four intersection points, which are assumed to be ordered anti-clockwisely; $t_{1,i}$ and $t_{2,i}$ denote the corresponding parameters of the four intersection points in the parametric forms of the first and the second ellipses, respectively; r_i denote the four roots of Eq.(A.6).