Facial image analysis robust to pose variation

(姿勢変化に頑健な顔画像分析技術に関する研究)

March 2021

Koichiro Niinuma

Facial image analysis robust to pose variation

(姿勢変化に頑健な顔画像分析技術に関する研究)

Koichiro Niinuma

Graduate School of Science and Technology

University of Tsukuba

March 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automated facial image analysis (AFA) has attracted substantial attention in recent decades. As AFA can obtain a wide range of information about people, including human identity, expression, age, gender and race, the technology can be applied to a wide range of applications, such as user authentication, border monitoring, social robots, driver monitoring, online learning, gaming, marketing, and medical treatment [83, 95, 52]. In many real-world scenarios, AFA's robustness to unconstrained environments, such as arbitrary pose, illumination change, or occlusion, is crucial.

The objective of this thesis is to improve AFA across pose, which is one of the major challenges encountered by AFA systems. To improve the performance of AFA systems, a great number of approaches have been proposed, including image-based, video-based, and 3D-based approaches (See [96, 83, 91] for face recognition, and [52, 95] for facial expression recognition). Though these versatile approaches have been steadily improving the performance of AFA systems, they still suffer from arbitrary pose situations. To mitigate the problem, another possible solution is to use approaches involving the design of new feature representation or matching methods for arbitrary pose situations. However, these approaches make it hard to take advantage of well-studied versatile approaches because they cannot directly use well-studied existing feature representation or matching methods. We therefore believe that developing methods extending versatile approaches for arbitrary pose situation is important to tackle the problem. Our methods strengthen versatile approaches by utilizing the information specific to arbitrary pose situations. Specifically, we set three AFA applications across pose (face recognition, continuous

authentication, and facial expression recognition) as our targets, and then propose new methods extending existing versatile approaches for arbitrary pose. Experimental results outperform existing works, and show the effectiveness of our methods. Our methods can easily be combined with other versatile approaches to leverage state-of-the-art approaches.

In this chapter, we first overview the representative approaches for AFA in Chapter 1.1. We next discuss the motivation and objective of this thesis in Chapter 1.2. Finally, we provide the organization of this thesis in Chapter 1.3.

1.1 Approaches for automated facial image analysis

We overview three mainstream approaches for AFA: image-based, video-based and 3D-based approaches. We mainly discuss the two representative applications (face recognition and facial expression recognition), but the same discussion is applicable to other AFA applications as well.

Image-based approaches

Image-based approaches that usually analyze one static image are the most well-studied approaches. Compared with video-based or 3D-based approaches, the advantage of image-based approaches is a wide range of applications because neither multiple images or special devices are required. There has been a significant trend of image-based approaches in terms of feature representation. Wang and Deng [91] categorized feature representation for face recognition into four groups according to the time period when they were used: 1) holistic approaches in the 1990s (e.g., Eigenface [86] and Fisherface [13]); 2) handcrafted local descriptors in the early 2000s (e.g., Gabor [57], local binary pattern (LBP) [3]); 3) shallow learning in the early 2010s; and 4) deep learning after 2014. Similarly, for facial expression recognition, Li and Deng [52] classified facial expression recognition into three categories: 1) handcrafted, 2) shallow learning, and 3) deep learning. In 2014, the deep-learning based face recognition approach DeepFace [81] showed compatible results with human performance for the first time. The achievement triggered the transformational shift to deep-learning based approaches, which occurred not only for face recognition but also for other AFA applications. For facial expression recognition, we can see the same shift in the Facial Expression Recognition & Analysis Challenge (FERA

2015 [87], FERA 2017 [88]). In 2015, only a single deep learning method [33] entered the Challenge. It ranked 3rd and 4th in occurrence and intensity detection, respectively. Two years later, deep-learning based approaches dominated the medal podium by a large margin [82, 97]. It is clear that advances in deep learning techniques have been key to the recent huge performance improvement for AFA systems.

Video-based approach

Unlike image-based approaches, video-based approaches use a series of images to recognize the human identification or facial expression. Taskiran et al. [83] review image-based approaches for face recognition, and Li and Deng [52] review them for facial expression recognition. In [83], video-based face recognition approaches are categorized into two groups: set-based methods, and sequence-based methods. In set-based methods, frames of a video are treated as a set of image samples and the temporal order is not considered while sequence-based methods employ the temporal information that exists in a video. The video-based approaches can analyze the information that image-based approaches cannot, such as facial dynamics though applicable scenarios could be limited.

3D-based approach

Many 3D-based approaches have also been proposed. These approaches are becoming a realistic solution for some scenarios as 3D sensors are becoming more accurate, smaller, and inexpensive. Zhou and Xiao [96] surveyed 3D based approaches for face recognition, and categorized them into three groups: pose-invariant, expression-invariant and occlusion invariant recognition. The main objective of each group is to provide more robust recognition algorithms under unconstrained environments.

1.2 Motivation and objective

The objective of this thesis is to improve automated facial image analysis (AFA) across pose, which is one of the major challenges for AFA systems. While AFA has been making significant progress in recent decades, arbitrary pose situations are still challenging because there are specific problematic issues in these situations.

Ding and Tao [25] identify four challenges for face recognition across pose: a) self-occlusion, b) loss of semantic correspondence, c) nonlinear warping of facial textures, and d) accompanied variations in resolution, illumination, and expression.

One consideration to bear in mind in addressing these challenges is that while versatile approaches cannot leverage the information specific to arbitrary pose situations, approaches involving the design of new feature representation or matching methods only for arbitrary pose situations cannot take advantage of well-studied versatile approaches. Therefore, our methods extend versatile approaches by leveraging the information specific to arbitrary pose, rather than proposing new features or matching methods for the specific situations. More specifically, we set three AFA applications across pose (face recognition, continuous authentication, and facial expression recognition) as our targets, and then propose the new methods as described below.

Multi-view face recognition via 3D model based pose regularization

Face recognition across pose is critical for many real-world applications, such as criminal identification, and face tagging. In these applications, there is a strong possibility that the face images are obtained without the cooperation of subjects. To tackle the situation, we propose a fully automatic method for multi-view face recognition. We first build a 3D model from each frontal target face image, and also estimate the pose of a query face image using a multi-view face detector. We next generate synthetic target images to resemble the pose in query face images. We then align the synthetic target images and the query image by applying Procrustes analysis [32], and extract block based MLBP features for face matching. We conducted the experiments on two public-domain database (Color FERET [68] and PubFig [47]) and the Mobile face databases collected using mobile phones. The experimental results show promising results. We also show that the proposed approach can be easily extended to leverage existing face recognition methods for multi-view face recognition.

Continuous authentication using soft biometric traits

Most existing systems authenticate a user only at the initial login session, but this could be a critical security flaw for many systems. To mitigate the problem, a great

number of continuous authentication systems that continuously monitor and authenticate users even after the login session has been proposed. Previous methods for continuous authentication primarily used hard biometric traits, such as fingerprints and faces. However, the use of these biometric traits is not only inconvenient for the user, but is also not always feasible due to the user's posture in front of the system. To solve the problem of the previous methods, we propose a new framework for continuous user authentication that primarily uses soft biometric traits (e.g., facial skin color and clothing). The proposed framework automatically enrolls soft biometric traits every time the user logs into the system, and combines soft biometric matching with the conventional authentication schemes (password and face biometric traits). Experimental results show that the proposed scheme has high tolerance to the user's posture in front of the console for continuous user authentication.

Systematic evaluation for deep-learning based facial expression recognition across pose

In most of the scenarios of facial expression recognition, it is necessary to analyze facial expression from natural communication or reactions that include arbitrary poses. To tackle the situation, many deep-learning based approaches have been proposed. While these approaches showed promising results, the contribution of critical design choices remains largely unknown. To address the problem, we systematically evaluate design choices for deep-learning based facial expression recognition in pre-training, feature alignment, model size selection, and optimizer details. In our experiments, the Facial Expression Recognition and Analysis (FERA 2017) database, which includes synthesized face images with 9 head poses [88], was used. By utilizing all the insights we found, we developed an architecture that exceeds the state-of-the-art on FERA 2017 both in detection of the occurrence and intensity of facial actions. We also report the cross-pose and cross-domain generalizability of our architecture.

1.3 Thesis organization

The rest of this thesis is organized as follows. We propose multi-view face recognition via 3D model based pose regularization in Chapter 2, and continuous authentication using soft biometric traits in Chapter 3. In Chapter 4, we describe systematic eval-

uation for deep-learning based facial expression recognition across pose. Chapter 5 concludes with directions for future research.

Chapter 2

Multi-view face recognition via 3D model based pose regularization

2.1 Introduction

Automated face recognition has attracted tremendous interest in the past decades due to its wide applications [72] including border control, surveillance, criminal identification, login authentication, purchase authentication, and face tagging. Automated face recognition in controlled conditions has shown impressive performance, such as frontal poses, neutral expressions and near uniform illumination. However, automated face recognition in uncontrolled environments remains a challenging problem, such as arbitrary poses, non-uniform illumination, and occlusion [72]. One typical example of automated face recognition across pose is identification or authentication of individuals with face images captured by mobile devices, such as smart phones, handheld terminals, or surveillance cameras. Another well-known application is face tagging provided by many photo storage services, such as Google Photos. Fig. 2.1 shows some examples.

Since the face images are captured without the user's cooperation in many of these scenarios, faces in the query images can be of arbitrary poses. An example where a profile face image resulted in the arrest of a robbery suspect is shown in Fig. 2.1 (a). The arbitrary pose variations have become one of the primary problems

Figure 2.1: Examples of images of face recognition across pose. (a) Profile face image which led to arrest, (b) Non-frontal face image in FERET dataset [68], (c) Non-frontal face image in the Mobile dataset collected using a mobile phone in our laboratory, (d) Non-frontal face image in PubFig dataset [47].

for most existing systems to perform automated face recognition.

To tackle the problem, we propose a new fully automatic multi-view face recognition method:

1. The proposed approach does not require manual landmark annotations or the assumption of known poses within a limited range.

2. The proposed approach achieves higher performance than two baseline matchers (FaceVACS [1] and MKD-SRC [55]) in several scenarios with different pose variations.

3. The proposed approach has the good extensibility. We show that by replacing our MLBP based face matcher with two baseline face matchers.

Table 2.1: A comparison of existing methods for multi-view face recognition.

		Publication	Approach	Pose assumed to be known?	Manual annotation required for non-frontal face image?	Databases used (pose variations)
Pose invariant feature extraction	Holistic	Sharma et al. [74]	Partial Least Squares, Bilinear Model, Canonical Correlation Analysis	Yes	Yes	FERET (±60°) CMU-PIE (±90°) Multi-PIE (±90°)
		Li et al. [50]	Partial Least Squares	Yes	Yes	Multi-PIE(±90°) CMU-PIE (±90°)
		Fischer et al. [29]	Partial Least Squares	Yes	Yes	Multi-PIE (±90°)
		Prince et al. [70]	Tied Factor Analysis	Yes	Yes	FERET (±90°) CMU-PIE (±90°) XM2VTS (±90°)
		Li et al. [49]	Linear Regression	Yes	Yes	FERET (±60°) CMU-PIE(±90°)
		Blanz and Vetter [15]	3D Morphable Model	No	Yes	FERET (±60°) CMU-PIE(±90°)
		Wang et al. [90]	Orthogonal Discriminant Vector	No	No	FERET(±25°) CMU-PIE(±15°) Yale B, AR
	Local	Kanade and Yamada [42]	Subregion Based Probabilistic Model	Yes	Yes	CMU-PIE (±90°)
		Ashraf et al. [7]	Probabilistic Stack-flow	Yes	Yes	FERET (±60°)
		Lucey and Chen [59]	Patch-whole Sparse Registration	No	Yes	FERET (±60°)
		Castillo and Jacobs [17]	Stereo Matching	No	Yes	CMU-PIE (±90°)
		Arashloo and Kittler [6]	Markov Random Field	No	No[1]	CMU-PIE (±90°) XM2VTS
		Liao et al. [55]	Multi-keypoint Descriptor	No	No	PubFig (Arbitrary)
Pose Normalization	To frontal	Chai et al. [18]	Linear Regression	Yes	No	CMU-PIE(±45°)
		Sarfraz and Hellwich [73]	Multivariate Regression	Yes	No	CMU-PIE (±90°) FERET (±60°)
		Li et al. [51]	Morphable Displacement Field	Yes	No	FERET (±60°) CMU-PIE (±90°)
		Teijeiro-Mosquera et al. [84]	Active Appearance Model	No	No	CMU-PIE(±45°)
		Asthana et al. [8]	View Based Active Appearance Model	No	No	FERET(±40°) CMU-PIE(±45°) Multi-PIE(±45°) FacePix(±45°)
		Ding et al. [26]	Random Forest Embedded Active Shape Model	No	No	FERET(±60°) CMU-PIE(±67.5°) CAS-PEAL(±45°)
	To non-frontal	Prabhu et al. [69]	3D Generic Elastic Model	No	No	Multi-PIE (±60°) Video Clips
		Han and Jain [34]	3D Modeling from two images	No	Yes	FERET (±22.5°)
		Our approach	3D Based Pose Regularization	No	No	FERET(±90°) Mobile (±90°) PubFig (Arbitrary)

1 A bounding box is required, but it is not clear if the bounding box is obtained manually or automatically.

2.2 Related work

Table 2.1 shows existing approaches for automated face recognition across pose. These approaches for multi-view face recognition can be grouped into two main categories: Pose invariant feature extraction, and pose normalization.

Pose invariant feature extraction

The approaches usually provide a common representation which maximizes the correlation among subject's face images with different poses. We can further classify them into two groups: (i) holistic representation, and (ii) local representation:

(i) **Holistic representation**: For this group, Linear regression, partial least squares (PLS), Bilinear Model (BLM), Canonical Correlation Analysis (CCA), 3D Morphable Model, are widely used approaches [15, 29, 49, 50, 70, 74, 90]. They obtain a pose-independent representation by projecting face images with different poses into latent spaces. While these approaches can solve the pose variation problem and feature representation at the same time, holistic representation can easily be affected by face deformations due to large pose variations. Additionally, many holistic methods assume that the face poses are known. For example, they directly used the poses provided in the databases to build pose-specific models, and used only the model covering the pose of a testing image for recognition.

(ii) **Local representation**: Local representations extract features from individual patches of a face. In comparison with holistic representation, the approaches are usually more robust to large pose variations. The representative approaches of this category includes Markov Random Field [6], subregion based probabilistic model [42], probabilistic stack-flow [7], patch-whole sparse registration [59], and stereo matching [17]. One of the drawbacks of the category is that most local representation based approaches [6, 7, 17, 42, 59] require manual landmark annotation to establish the local patch correspondence between frontal and non-frontal face images.

Pose normalization

Pose normalization approaches convert face images with different poses into face images with the same pose. Unlike approaches for pose invariant feature extraction that usually involves the design of new feature representation and matching methods, the approaches in this category can directly use existing feature representation and matching methods. Many approaches in this group transform non-frontal face images into frontal images using either 2D or 3D based approaches because face recognition techniques for frontal or near-frontal poses have been widely studied. The representative approaches are Linear and multivariate regressions, Active Shape Model (ASM), and Active Appearance Model (AAM) [8, 18, 26, 51, 73, 84]. However, the recovered frontal face images can be inadequate due to the self-occlusion under large poses as observed in [8]. To avoid the matching of corrupted facial regions in the recovered frontal images, Li et al. [51] utilized occlusion masks.

Instead of recovering frontal images from non-frontal images, a different approach is to generate non-frontal images from frontal images. More specifically, the approaches generate non-frontal views that resemble the poses in testing face images. Park et al. [65] used 3D face data to generate non-frontal views. However, the approach requires 3D sensing that is still expensive and can be slow. Additionally, because 2D images constitute the legacy databases, their 3D images may not be available. Therefore, 3D face models reconstructed from frontal face images can be the substitutions for real 3D faces. To generate non-frontal images from frontal images, typical approaches are 3D Morphable Model and 3D generic elastic model (3D GEM) [34, 69].

Though many approaches in pose invariant feature extraction and pose normalization have been proposed, most face recognition systems cannot perform fully automatic multi-view face recognition; manual landmark annotations and assumption of known poses are required. These requirements limit the application of these methods in real-world scenarios.

2.3 Proposed method

We propose a new fully automatic multi-view face recognition method via 3D model based pose regularization. The proposed approach extends existing face recognition

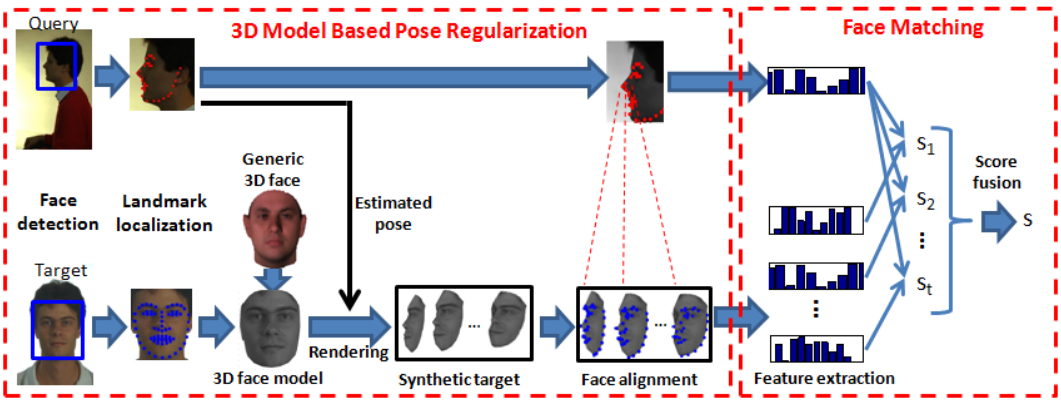

Figure 2.2: An overview of the proposed approach for multi-view face recognition consisting of 3D model based pose regularization and face matching.

systems into multi-view scenarios. As shown in Fig. 2.2, the proposed approach consists of two main modules:

1. 3D model based pose regularization

2. Face matching with block based multi-scale LBP (MLBP) features

Most previous pose normalization approaches transformed non-frontal face images into frontal images. Unlike the existing approaches, the proposed 3D model based pose regularization method generates synthetic target images to resemble the pose variations in query images. Note that it is much easier and more accurate to generate non-frontal images from frontal views than generate frontal images from non-frontal views. This is because detecting accurate landmarks automatically under large pose variations is more difficult. In addition, it is problematic to recover the frontal view for the occluded facial regions, especially when many face areas are significantly occluded under large pose variations.

Our approach is similar to the novel view rendering based on 3D GEM [69], but there are some important differences. First, in our approach, a simplified 3D Morphable Model [15] is used. In addition, our face alignment is performed using Procrustes analysis [32] under large pose variations instead of aligning the synthetic target and query images based on eye positions. For face images with large pose variations, one of the two eyes is occluded. Face alignment based on two eyes does not work under these circumstances. Furthermore, we utilized a face matching

method with blocked MLBP features that provides better robustness against an unconstrained environment, such as face illumination and expression variations. Han et al. [35] suggested that additional face preprocessing methods might be integrated with MLBP to further improve the robustness. Finally, the proposed approach has the good extensibility. We show the extensibility of the proposed framework by replacing our MLBP based face matcher with two face matching systems (FaceVACS [1] and MKD-SRC [55]).

2.3.1 3D Model Based Pose Regularization

Fig. 2.2 illustrates how to perform our 3D model based pose regularization. We first build a 3D model from each frontal target face image. We next estimate the pose of a non-frontal query face image, and generate synthetic target face images that resemble the pose variation of a query face image. By generating the synthetic target face images, we are able to perform holistic face alignment between the target and query images. We align the target and query images by utilizing Procrustes analysis.

3D Modeling from A Frontal Image

Our 3D shape model is obtained from the USF Human ID 3-D database [2]. The database consists of 3D face shape and texture of 100 subjects captured with a 3D scanner. To build a 3D model, a simplified 3D Morphable Model [15] without the texture fitting is utilized. The original 3D face includes 75,972 vertices, but 76 vertices are interactively selected based on the 76 keypoints for efficient computation. We utilize open source Active Shape Model (Stasm [62]) to select the vertices. We can represent the 3D shape of a new face using a PCA model as follows:

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{k=1}^{K} \alpha_k \mathbf{W_k}, \tag{2.1}$$

where $\mathbf{S}$ is the shape of a new 3D face, $\bar{\mathbf{S}}$ is the mean 3D shape of 100 3D faces from the USF Human ID 3-D database, $\mathbf{W_k}$ is the shape eigenvector corresponding to the k-th largest eigenvalue, and α_k is a coefficient for the k-th shape eigenvector.

A 3D face is projected onto a 2D plane under a set of transformations and then a 2D face image is obtained. Based on such a face imaging process, we can recover

the shape of a 3D face from its 2D projection by minimizing the cost function [66]

$$e(\mathbf{P}, \mathbf{R}, \mathbf{T}, s, \{\alpha_k\}_{k=1}^{K}) = ||\mathbf{P_{2D}} - s \cdot \mathbf{PRTS}||_{L^2}, \tag{2.2}$$

where $\mathbf{P_{2D}}$ is a set of facial landmarks that Stasm detects, $\mathbf{P}$ is an orthogonal projection from 3D to 2D, and $\mathbf{R}, \mathbf{T}, s$ indicate the rotation, translation and scaling operations for the 3D face shape $\mathbf{S}$, respectively.

The input frontal face image is directly used as the texture corresponding to the frontal 3D facial shape. When we generate a novel view, the frontal face image is directly mapped to a novel view based on Delaunay triangulation of the 2D facial landmarks. In comparison with a statistical face texture model used in 3D Morphable Model, our mapping method can retain detailed and realistic features that are important for face recognition. In addition, our texture mapping method is more efficient.

Generating Synthetic Target Images

Since the recovered 3D facial shape $\mathbf{S}$ from (2.1) and (2.2) is with a frontal pose, novel synthetic target images can be easily generated from a target face image by transforming $\mathbf{S}$ using different translation, rotation, scaling, and projection transformations. Three target face images and their synthetic images under 19 novel views ($\pm 90^\circ$ with an interval of 10°) are shown in Figs. 2.3 (a) and (b). The synthetic target images are generated using our 3D face model. To reduce the pose difference between a target face image and a query face image, we generate the synthetic target face images that resemble the pose of a query face image. Fig. 2.3 (c) shows query images, and the red rectangles in Fig. 2.3 (b) indicate the images selected by our online selection based on the pose estimation for each query image.

To generate the synthetic target images, the poses of query images are required. However, in many practical multi-view face recognition scenarios, we cannot assume that the poses are known. Under these circumstances, automatic pose estimation from arbitrary face images is necessary in order to perform fully automatic face recognition. In our approach, a mixture of tree-structured part models (MT-SPM) [98] is utilized to estimate the pose from each query image. Based on the pose estimation for a query image, only synthetic target face images with similar poses will be generated for face matching.

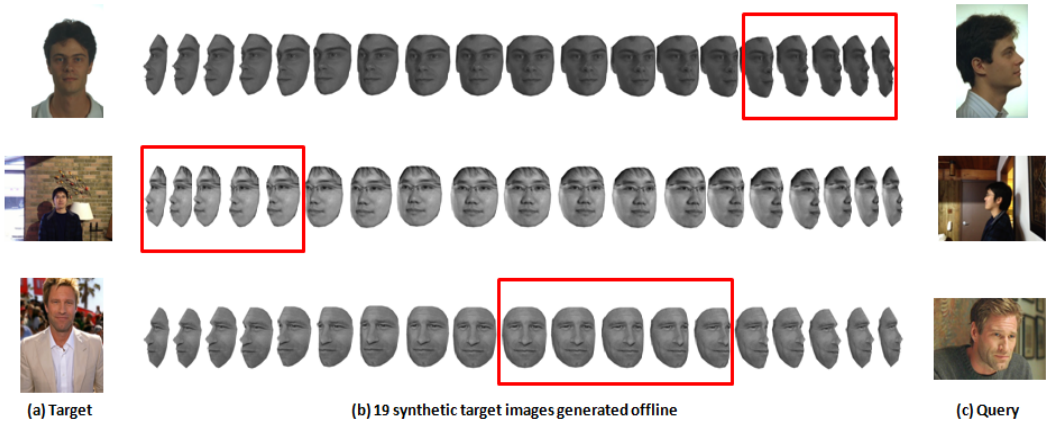

Figure 2.3: Examples of synthetic target images generated by our approach. (a) original target images, (b) synthetic target images, and (c) query images. The red rectangles show the online selection of synthetic target images based on the pose estimation from query images.

Because generating synthetic target images online would increase the computational cost of face matching, we adopt a more efficient strategy. Specifically, after we build a 3D model from each target face image, we generate 19 synthetic target images for each target image offline as shown in Fig. 2.3, and select only synthetic images with similar poses to the query image for matching. Additionally, we select multiple synthetic images with similar poses for matching instead of using only one synthetic image. That is because the pose estimated by MTSPM is prone to error. In our experiments, we select five synthetic target images for face matching. The red rectangles in Fig. 2.3 indicates the selected synthetic target images. According to this strategy, our system can conduct large scale face recognition without increasing the computational cost seriously.

Face Alignment

A widely used approach in face alignment is holistic face alignment based on two eyes. This approach works for frontal or near-frontal face images [69]. However, the face alignment method based on two eyes is problematic for non-frontal poses because one of the two eyes is often not visible under large pose variations. In addition, even when both eyes are visible in non-frontal images, the face alignment

method based on two eyes can generate an artificial increase in the overall size of the face image.

To tackle these problems with face alignment based on two eyes, our approach applies Procrustes analysis [32]. By using Procrustes analysis, we can align the synthetic target and query images based on the facial keypoints from a 3D face model and facial keypoints detected by MTSPM. The numbers of keypoints defined in a 3D face model and MTSPM are different. To establish the keypoint correspondence between a 3D face model and MTSPM, we have manually identified 19 landmarks between two models. We perform Procrustes analysis based on 19 corresponding landmark pairs.

2.3.2 Face Matching

We used a face matching method with block based multi-scale LBP (MLBP) features [64], which provides robustness against face illumination and expression variations. More specifically, MLBP features which are a concatenation of LBP histograms with 8 neighbors sampled at different radii $R = \{1; 3; 5; 7\}$ are used in our experiments. A holistic face image (256×192) is divided into 768 sub-regions (8×8 non-overlapped blocks), and then, MLBP features are extracted from individual blocks and concatenated together.

We extract two MLBP histograms $\mathbf{x}$ and $\mathbf{y}$ with n dimensions from two face images, and then calculate chi-squared distance χ^2 as a measure of similarity between two face images:

$$\chi^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(x_i + y_i)/2} \tag{2.3}$$

where x_i and y_i are the features for i-th bin. A separate distance is calculated for each synthetic image. We have 5 distances for each query image since we generate 5 synthetic target images in our experiments. We calculate the final distance between a target and a query by selecting the minimum of these distances. We also tried some other score fusion methods such as the median or the mean of multiple distances, and decided to use the minimum of multiple distances in our experiments.

2.4 Experiments

2.4.1 Databases

We evaluate our approaches using 3 datasets:

Color FERET

The Color FERET database [68] consists of facial images with multiple poses from 994 subjects. In our experiments, one frontal image (fa) per subject is used as the target, and images with 6 non-frontal poses (ql, qr, hl, hr, pl, pr) are used as the query. While the FERET database has been widely used to evaluate many existing face recognition approaches, its limitation is that the dataset was collected under a well controlled scenario. For example, the participants are required to rotate the head and body to pre-designed directions, and the background and illumination in face images are nearly uniform.

Mobile

We collected a Mobile dataset including 112 subjects using iPhone 4S. Fig. 2.7 shows some examples of the images. By using the dataset, we can evaluate the scenarios of face recognition from images or videos captured using mobile devices. For each subject, we captured one or two frontal face images and around 10 non-frontal face images at several locations inside a building. In our experiments, we use only one frontal face image per subject as a target. In comparison with the FERET database, the Mobile dataset is more challenging in terms of background variation and illumination variation, and motion blurs due to the movement of the hand though it has fewer subjects.

PubFig

The PubFig database [47] includes 200 famous personalities collected from the Internet. In the dataset, 60 subjects are assigned to algorithm development, and the remaining 140 subjects are assigned to algorithm evaluation. We directly evaluate our method using the 140 subjects from the evaluation set because our method is a non-learning based approach. One frontal face image per subject is used as the target set, and 513 non-frontal images with arbitrary poses are used as the query set.

Note that the LFW database [37] also consists of arbitrary pose variations, but the images of many subjects are captured in the same environment.

2.4.2 Baselines

We compare our approach with two baseline systems (for fair comparison, both of the baseline systems are fully automatic in multi-view face recognition, like our approach):

1. Multi-keypoint descriptor based sparse representation (MKD-SRC) [55]
2. FaceVACS [1] (Commercial-Off-The-Shelf (COTS) face matching system)

We evaluate the proposed approach under three scenarios. Note that FaceVACS is not used for large yaw rotations because it cannot enroll any face under the situation as discussed in Chapter 2.4.4:

1. Small yaw rotations (two eyes are visible)
2. Large yaw rotations (only one eye is visible)
3. Arbitrary pose variations.

We perform face verification under the three scenarios because face verification experiments are more convincing than face identification experiments with limited subjects and images. We report the performance in terms of receiver operating characteristic (ROC) curve. The extensibility of the proposed method is also investigated by incorporating pose regularization in our approach with two baseline matchers (MKDSRC and FaceVACS).

2.4.3 Small Yaw Rotations

The results for small yaw rotations are shown in Fig. 2.4. Fig. 2.4 (a) shows the performance of the proposed approach, MKD-SRC, and FaceVACS on the FERET, and Fig. 2.4 (b) shows the performance on Mobile datasets.

Fig. 2.4 (a) shows that, for the FERET database, while there is not a significant difference between the COTS system FaceVACS and multi-view face matcher

MKD-SRC, the proposed approach achieves more than 15% higher face verification rate than MKD-SRC and FaceVACS at 0.1 False Alarm Rate (FAR). Fig. 2.4 (b) shows that, for the Mobile dataset, MKD-SRC is more robust than FaceVACS. However, the proposed approach achieves 20% higher face verification rate than MKD-SRC at 0.1 False Alarm Rate (FAR). Note that the COTS system FaceVACS is mainly designed for near-frontal face recognition. As we discussed above, the FERET database is collected under a more controlled condition than the Mobile database.

Overall, both Fig. 2.4 (a) and (b) indicate that the proposed approach is more effective than MKD-SRC and FaceVACS in handling small pose variations. In addition, the comparison between Fig. 2.4 (a) and (b) demonstrates that both the proposed approach and MKD-SRC are more robust to variations of background and illumination, as well as motion blurs in the Mobile database than FaceVACS.

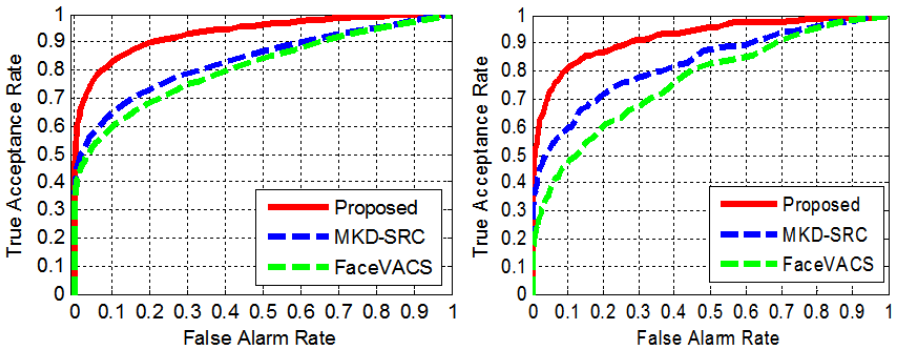

Figure 2.4: Face verification results under small yaw rotations on the (a) FERET and (b) Mobile datasets.

2.4.4 Large Yaw Rotations

The results for large yaw rotations are shown in Fig. 2.5. Under the scenario of large yaw rotations, FaceVACS is no longer available as a baseline because no faces can be enrolled. Fig. 2.5 (a) shows the performance of the proposed approach and MKD-SRC on the FERET, and Fig. 2.5 (b) show the performance on Mobile databases.

While MKD-SRC obtains around 20% verification rates at 0.1 False Alarm Rate (FAR) under large yaw rotations, the propose approach achieves much better per-

formance (50%). However, we can see the performance degradation compared with that under the scenario of small pose variations in Chapter 2.4.3. The performance degradation can be attributed as follows. According to our manual inspection of the images, 90% of the query images are 90 profile images. The eyes and eyebrows have been identified to have the most discriminative ability under small poses [78]. However, these discriminative features are no longer available in 90° profile face images.

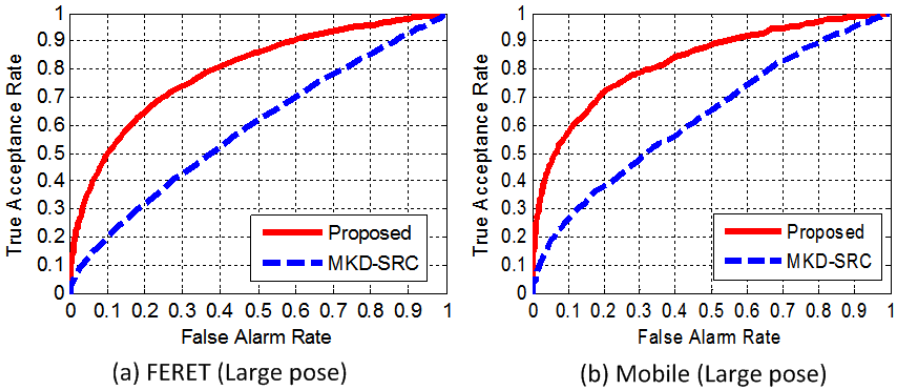

Figure 2.5: Face verification results under large yaw rotations on the (a) FERET and (b) Mobile datasets.

2.4.5 Arbitrary pose variations

The PubFig dataset is more challenging than the other two datasets (FERET and Mobile) for multi-view face recognition systems because face images in the PubFig dataset are taken with non-cooperative subjects in completely uncontrolled situations. Therefore, unlike the FERET and Mobile databases, the dataset includes arbitrary variations in pose, illumination, expression. In this experiment, only face images that can be enrolled by FaceVACS are used as the query set in order to use FaceVACS as a baseline. Though a query set does not contain faces with large pose variations, it covers yaw, pitch, and roll rotations.

Fig. 2.6 shows the performance of the proposed approach, MKD-SRC and FaceVACS. The proposed approach outperforms MKD-SRC; The proposed approach achieved 45% at 0.1 False Alarm Rate (FAR) while MKD-SRC achieves 30%. However, the performance of the proposed approach is nearly the same as the one of the

COTS matcher FaceVACS. This can be attributed to the other types of variations such as expression, illumination and aging. A score level fusion of the proposed approach and FaceVACS achieves a leading performance on PubFig.

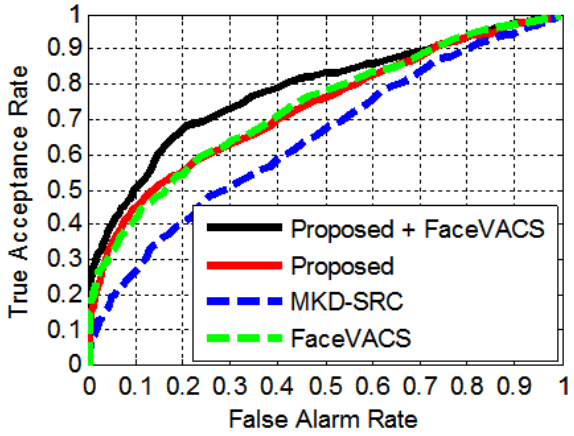

Figure 2.6: Face verification results under arbitrary poses on the PubFig dataset.

2.4.6 Successful examples

Fig. 2.7 shows examples where the proposed approach successfully identifies query and target images at 0.1 False Alarm Rate (FAR). Fig. 2.7 demonstrates that the proposed pose regularization reduces the pose gap between target and query images.

Figure 2.7: Examples of successful matches by the proposed approach at 0.1 False Alarm Rate (FAR) on the FERET, Mobile and PubFig datasets. (a) target images, (b) synthetic target images that lead to correct matching of the query and target images, and (c) query images.

Table 2.2: Face verification performance of MKD-SRC and FaceVACS at 0.1 False Alarm Rate (FAR) without and with the proposed pose regularization (without/with).

Matchers	Small Pose		Arbitrary Pose
	FERET	Mobile	PugFig
MKD-SRC	65.8%/77.0%	58.7%/70.5%	26.7%/30.6%
FaceVACS	60.4%/81.8%	48.3%/75.6%	41.5%/48.0%

2.4.7 Extensibility of the Proposed Approach

By replacing our MLBP based matcher with another matcher, we can easily leverage other sophisticated face matchers. To evaluate the extensibility of our approach, our MLBP based matcher is replaced with MKD-SRC and FaceVACS. The experiments are performed under the scenarios of small and arbitrary pose variations to use FaceVACS. Table 2.2 shows the performance of MKD-SRC and FaceVACS without and with the proposed pose regularization. In all cases, our pose regularization method greatly enhance the performance of the two face matchers. These results demonstrate the good extensibility of the proposed approach.

2.5 Summary

We have proposed a new fully automatic face recognition approach across multiple views. First, we build 3D face models from frontal target face images, and generate synthetic target images to resemble the poses in query face images. We next align synthetic target face images and query face images using Procrustes analysis. Finally, we extract blocked based MLBP features for face matching. The proposed approach does not require manual landmark annotations or known poses so that it can conduct fully automatic multi-view face recognition. Our experimental results on the FERET, Mobile, and PubFig databases demonstrate that the proposed approach outperforms two baseline face matchers. We also show that the good extensibility of the proposed approach.

The limitation of our current approach includes the performance degradation with large yaw rotations compared to that with small yaw rotations. One of the main reasons for this degradation is the difference between our 3D model and the

real 3D shape of each face, which widely affects the performance with large yaw rotations. To mitigate the problem, one possible approach is to have multiple 3D models and select one of them for each face image based on the attributes, such as sex, age, and race.

Chapter 3

Continuous authentication using soft biometric traits

3.1 Introduction

Most computer or network systems require user authentication when the user logs into the systems. Because user authentication is extremely important to protect the system security, a wide range of login authentication methods have been utilized depending on the circumstances. The methods include knowledge-based methods (e.g., passwords), token-based methods (e.g., smart cards), textual and graphical passwords [79], public key infrastructure (PKI), and biometric authentication [30].

However, all of the login methods have a common security flaw. Specifically, they authenticate a user only at the initial log-in session, and do not reauthenticate the user after that. The user is reauthenticated only after the user logs out or there is a substantial time interval between the user's activities on the system. Anyone can access the system resources if the legitimate user leaves the system unattended without logging out. This could be a critical security weakness not only for high-security systems, but also for low-security systems, such as personal computers in a general office environment.

To resolve this common problem in login authentication methods, we need continuous authentication systems that continuously monitor and authenticate the user after the initial login session. While it is strongly desirable that the continuous authentication system has good usability by authenticating a user without his active

cooperation, the available methods for continuous authentication can only provide limited usability. For example, systems that request that a user enter his password frequently for continuous authentication are annoying to the user. Additionally, the method utilizing hard biometric traits is also not adequate. The user will face the inconvenience of limited privilege whenever the system fails to obtain the user's hard biometric trait. We therefore believe that the more appropriate approach to continuous authentication is to utilize biometric traits that can be captured regardless of the user's passive posture in terms of user involvement.

To address this problem in the existing methods for continuous authentication, we propose a new method for continuous user authentication that continuously authenticate a user by combining soft biometric traits and hard biometric traits. In particular, the colors of the user's clothing and face are used as the soft biometric traits for continuous authentication, and PCA-based face features are used as the hard biometric traits for relogin authentication. To the best of our knowledge, the proposed method is the first to use soft biometric traits for continuous authentication.

The block diagram of the proposed continuous user authentication system is shown in Fig. 3.1. The arrow from Mode I to Mode II represents process flow and all other arrows represent possible transitions. Our method also addresses the issues if 1) relogin authentication which handles a short absences of the user, and 2) template update to handle illumination change. These issues are explained in more detail Chapter 3.4.

3.2 Related Work

There have been many proposed methods for continuous user authentication [4, 10, 9, 16, 43, 45, 63, 75]. These studies used one or more hard biometric traits, such as fingerprint and face. Sim et al. [75] and Kwang et al. [48] captured the user's face and fingerprint for continuous authentication. They captured faces using a camera and captured fingerprints using a mouse with a built-in fingerprint sensor. While these two studies showed promising authentication results, their system ran into low availability of the biometric traits. A fingerprint can only be authenticated when the user keeps his finger on the sensor embedded in the mouse. Similarly, when a user is entering a document, the user often needs to turn her head away from the

camera. Another situation where face image is not properly captured is when the user takes a break from typing and does not look directly at the monitor.

Sim et al. [75] defined three criteria for continuous authentication using hard biometric traits:

1. Different reliability of various modalities must be accounted for.
2. Older biometric observations must be discounted to reflect their increasing uncertainty about the continued presence of the legitimate user.
3. User authentication certainty needs to be established at any point of time even when no observation of any of the biometric traits is available.

Sim et al. also proposed a system based on the genuine and imposter matching score densities of face and fingerprint, Ω_{intra} and Ω_{inter}. The decision criterion for a user being genuine (system being safe) is, $P(x_t = safe|Z_t) > P(x_t = compromised|Z_t)$, where $Z_t = \{Z_0, ..., Z_t\}$ denotes the set of biometric observations until time t and x_t denotes the system state (safe or compromised) at time t. A decaying function $p = e^{k\Delta t}$ was introduced to control the influence of a biometric trait, where $k(k < 0)$ is a constant indicating the decaying speed, and t is the elapsed time since the last observation of biometric traits. The drawbacks of the system are summarized as follows:

1. The system needs intraclass and interclass score distributions (Ω_{intra} and Ω_{inter}) of both fingerprint and face biometrics.
2. The system needs a decaying function because continuous stream of hard biometric traits are not always available. The decay function comes at the expense of sacrificing the system security while the decaying function indeed enables continuous authentication. While the authentication decisions can be made to accept the user based on the decaying function, the system may already have been compromised.
3. The assumption $P(x_0 = safe|Z_0) = 1$ is not always valid (i.e., that the user is genuine at the login time). By using hard biometric traits, the user's identity can be verified. However, we cannot assume that the initial state is always safe ($P(x_0 = safe|Z_0) = 0$) in case an attacker logs into the system with a stolen password.

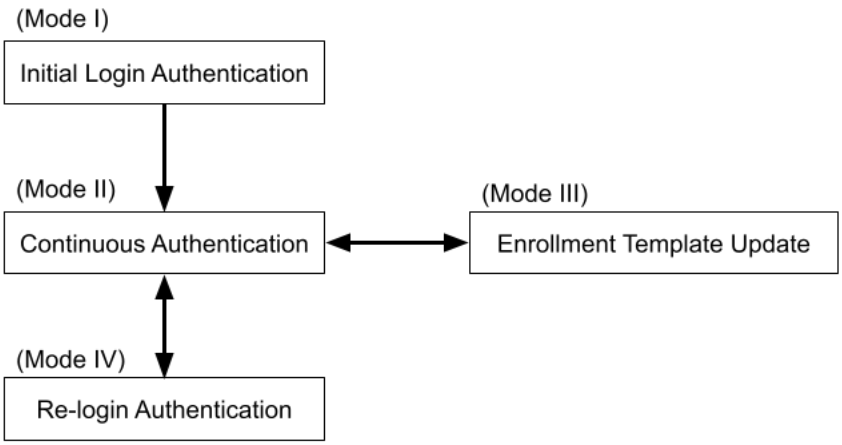

Figure 3.1: Proposed framework using soft biometric traits for continuous user authentication.

3.3 Soft biometric traits for continuous authentication

Soft biometric traits include a wide range of information, such as gender, ethnicity, color of eye/skin/hair, height, weight, and SMT (scars, marks, and tattoos). Jain et al. [39] defined the traits as "those characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals". Jain et al. [39, 40] also shows that the security of login authentication can be improved by combining soft biometric traits with hard biometric traits (e.g., fingerprint, face, iris, palm vein) even though it do not have enough discriminatory information to identify the user. Though the soft biometric traits cannot identify a user uniquely, it can be used to decide whether the user who is currently using the system is the same as the user who initially logged into it. Use of soft biometrics in a continuous authentication system has the following advantages:

1. The system can authenticate a user even when no hard biometric traits are available.

2. The system does not require preregistration of the soft biometric traits. Every time the user logs into the system, our proposed method automatically enrolls the soft biometric traits, and then automatically registers the user by combining the soft biometric traits with the hard biometric traits or the conventional

login authentication (e.g., face recognition authentication, password).

In comparison with the method that only relies on hard biometrics described in Chapter. 3.2, we can summarize the advantages of our system as follows:

1. The system does not need Ω_{intra} or Ω_{inter} to be available.
2. The system does not need the decaying function because the soft biometric traits enable richer observations (Z_t).
3. The assumption $P(x_0 = safe|Z_0) = 1$ is true on our system because soft biometric template is enrolled at each login time.

Note that soft biometric traits do not provide higher security at the login time. For example, a stolen password can still be used. However, our problem formulation starts with correct assumptions. Table 3.1 summarizes the differences between continuous authentication systems using hard and soft biometric traits.

Table 3.1: Summary of differences between the continuous authentication systems using hard biometric traits and soft biometric traits. In the table, Ω_{intra} and Ω_{inter} represent the intraclass and interclass matching score distributions, respectively

	Hard biometrics	Soft biometrics
Confidence of decision with each observation	High to medium	Medium to low
Frequency of observation	Medium to low	High
Pre-registration	Required	Not required
Ω_{intra} and Ω_{inter}	Available	Not available

Our system also satisfies the three criteria for continuous authentication introduced by Sim et al. [75]:

1. Our system uses a reliability factor of the different modalities.
2. Our system uses a time decaying function in relogin authentication mode to make older observations increasingly uncertain.
3. Our system can determine authentication certainty at any point of time.

Note that, in comparison with the hard biometric based continuous authentication systems, the third criterion is not as important in our system because of the high availability of soft biometric traits. The criterion is critical for the hard biometric based continuous authentication system because the hard biometric trait is often unavailable. On the other hand, in our system, the soft biometric is always available in continuous authentication. When the soft biometric is not available, our system moves to relogin authentication mode, and reauthenticates the user using a combination of the time decay function and hard and soft biometrics.

We use the similarity scores to make the authentication decision. Unlike the system proposed by Sim et al. [75], our system does not use Ω_{intra} and Ω_{inter} to make the system more flexible even though we can obtain Ω_{intra} and Ω_{inter} by running continuous authentication sessions with a number of subjects. Our system does not require preregistration of soft biometric traits.

In our system, the color histogram of user's clothing color and face color are used as soft biometrics. As a hard biometric for relogin authentication, the PCA-based face features [86] is used (the PCA-based face matcher can easily be replace with another matcher). Let Z_t^{sf}, Z_t^{hf}, and Z_t^c denote the set of observations of soft-face, hard-face, and clothing color, respectively, at time t and Z_0 be the observation at the login time. The similarity scores for the three biometric traits are represented as follows:

$$S_{softface} = S(Z_t^{sf}, Z_0^{sf}) \tag{3.1}$$

$$S_{hardface} = S(Z_t^{hf}, Z_0^{hf}) \tag{3.2}$$

and

$$S_{clothes} = S(Z_t^{c}, Z_0^{c}) \tag{3.3}$$

where $s(.,.)$ denotes the similarity score based on the Bhattacharyya coefficient [14]. We use the three-dimensional RGB color histograms as the features of soft face and clothes, $f_{softface}$ and $f_{clothes}$. The length of each dimension is l_{soft}. Therefore, the dimensions of feature vectors of both $f_{softface}$ and $f_{clothes}$ are $l_{soft} \times l_{soft} \times l_{soft}$. All feature values are transformed to a one-dimensional vector to calculate the similarity using the Bhattacharyya coefficient. The Bhattacharyya coefficient between two feature vectors a and b of length D_1 is given as

$$\Sigma_{i=1}^{D_1} \sqrt{(a_i b_i)} \tag{3.4}$$

The total soft biometric score S_{cont} is calculated as the weighted sum

$$S_{cont} = wS_{softface} + (1 - w)S_{clothes} \tag{3.5}$$

where w is the weighting factor in combining soft biometric traits of face and clothing (in our experiments, w is set to 0.5). The hard face feature $f_{hardface}$ is a set of Eigen vectors ($length = l_{eig}$) with each Eigen vector of length $l_{hardface}$. The dimensions of feature vectors of $f_{hardface}$ are $l_{eig} \times \_hardface$.

The decision criterion for a user being genuine is simply $S_{cont} \geq t_{cont}$, where t_{cont} is a threshold value. If $S_{cont} < t_{cont}$, the system status moves to relogin authentication mode. The main idea underlying the proposed methods is that the system uses only the soft biometric traits in the continuous authentication mode, and hard biometric traits are used only for the relogin authentication. If we use hard biometric traits in the continuous authentication mode, we will experience the same problem that the existing methods including Sim et al.'s method faced because the hard biometric trait is often unavailable. Instead of identifying each subject at every single instance, our system continuously monitors the user to determine whether the user is the same person who initially logged into the system. The similarity score of the hard face biometric $S_{hardface}$ is used only in the relogin authentication stage as

$$S_{relogin} = F(T_{cur} - T_{reject})S_{cont} \tag{3.6}$$

where $F(\Delta t) = e^{k\Delta t}$ denotes a time decaying function with the decay rate k ($k < 0$), T_{cur} denotes the current time when $S_{hardface}$ is above a threshold, $t_{hardface}$, and T_{reject} denotes the time when the system rejected a user in the continuous authentication mode. In $S_{relogin}$, both hard and soft biometric traits are used to make the relogin process more secure. Note that $S_{relogin}$ becomes small if $T_{cur} - T_{reject}$ is large. If $T_{cur} - T_{reject}$ is larger than a threshold, the system status moves to the initial login authentication mode (mode I). On the other hand, if the user is absent for a short time, ($T_{cur} - T_{reject}$ is small), the user will be accepted again by giving valid soft and hard biometric traits.

In the continuous authentication mode, T_{reject} or T_{cur} is not considered because the system uses S_{cont} as the criterion to accept the user instead of $S_{relogin}$. Overall, relogin authentication must satisfy the following three conditions:

$$S_{hardface} \geq t_{hardface} \tag{3.7}$$

$$T_{cur} - T_{reject} \leq t_{delay} \tag{3.8}$$

and

$$S_{relogin} \geq t_{relogin} \tag{3.9}$$

All the conditions in 3.7, 3.8 and 3.9 are incorporated in equation 3.6.

Fig. 3.2 compares conventional and continuous authentication systems. During the login session, the user logins to the system by inputting his identifying information (e.g., password or hard biometric information), and when the user is authenticated at the login authentication, the system registers soft biometric traits, such as color of user's clothing, as a "one-time" enrollment template. Finally, the system continuously authenticates the user using the enrolled "one-time" soft biometric traits.

Fig. 3.3 shows some example images of user's posture, where any hard biometric traits cannot be captured, such as hard facial biometric information. While continuous authentication systems using only hard biometric traits cannot handle such cases, our system can continuously authenticate the user because some of the soft biometric traits can be continuously observed. In Fig. 3.3, red and green ellipses indicate clothing and facial color histogram. The four different modes of the proposed continuous authentication system shown in Fig. 3.1 is explained in Chapter 3.4.

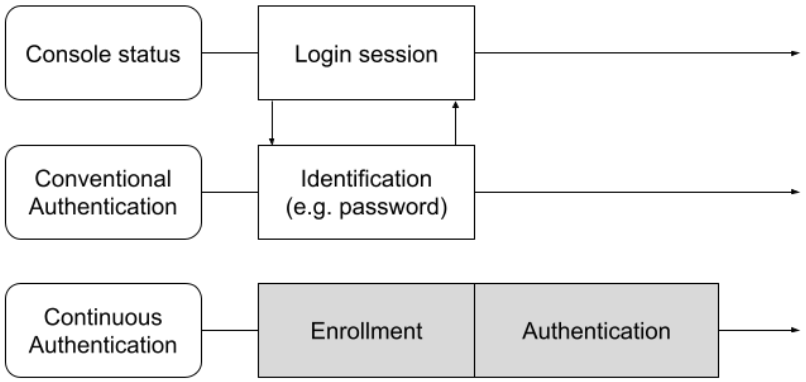

Figure 3.2: Diagram demonstrating the difference between conventional and continuous authentication systems.

(a) (b) (c)

Figure 3.3: Examples of user's posture. Any hard biometric traits cannot be reliably captured. The red ellipses indicate the face regions and the green ellipses indicate clothing region identified by our system.

3.4 Proposed method

We propose a versatile framework combining continuous authentication with conventional authentication. The proposed framework consists of four modes as described in Fig. 3.1. The first mode is initial login authentication (Mode I). During the mode, the user is authenticated using a conventional login authentication method, and a new enrollment template (color histogram of a user's clothing and face) is registered. In the continuous authentication mode (Mode II), the system authenticates the user continuously by using the enrollment templates. For relogin authentication (Mode IV), both hard biometric traits (i.e., face) and soft biometric traits are used to achieve high usability and security.

3.4.1 Initial Login Authentication (Mode I)

The first mode (initial login authentication) consists of the following four steps:

1) Initial authentication: We can use any conventional login authentication method for this step. In our current system, a password-based authentication is used.

2) Face detection: We use Haar classifier [56, 89] to detect a face. We assume that a user is typically looking in the frontal direction during the login session. This assumption is reasonable because the user typically looks at the monitor at the login session to input login information. Additionally, the user wants to be authenticated.

3) Body localization: We estimate the location and size of the user's body with respect to his face based on Jaffre and Joly's method [38]. We assume that

the area under the face is always the user's body and the size of this area is proportional to the one of the face.

4) Template enrollment: We obtain histogram of soft face (face color), and histogram of the clothing color, as well as features of hard face (Eigenface representation [86] of the face) and store them as enrollment templates. We use top 100 Eigenfaces to construct the template of hard face. In order to generate the color histograms of face and clothing, the RGB color space is quantized into 16x16x16 bins.

Fig. 3.4 (a), (b), and (c) depict the intermediate processes of steps 2), 3), and 4), respectively.

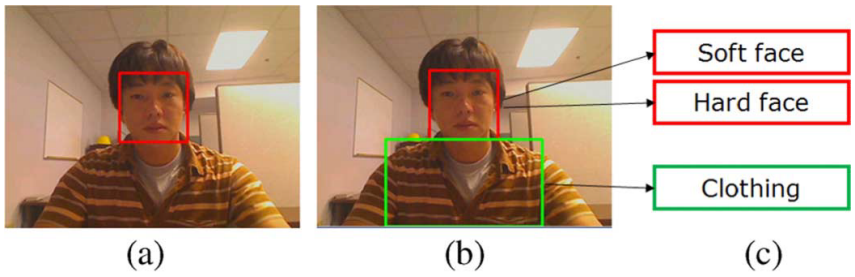

Figure 3.4: Enrollment steps during initial login authentication mode. (a) Face detection, (b) body localization, and (c) registration.

3.4.2 Continuous Authentication (Mode II)

After the user is authenticated in the initial login authentication mode, the system status changes to this mode (continuous authentication mode). In this mode, the system authenticates the user continuously by using the "soft face" and "clothing" enrollment templates registered in the initial login authentication mode (Mode I). The system status moves to Mode III (enrollment template update) when the system recognizes that the user leaves the system. The continuous authentication mode includes the following three steps:

1) Face and body identification using color histograms: by applying the mean shift algorithm [22, 23], the system tracks the face and the body separately. To track them, the histograms registered in the initial login authentication mode

(Mode I) are used, and the similarities $S_{softface}$ and $S_{clothes}$ are calculated. We use the Bhattacharyya coefficient [14] to calculate the similarity between two histograms.

2) Face recognition: in our system, we use a PCA-based face recognition technique (Eigenface) [86] to extract facial features. However, we can easily replace it with other matchers. Note that $S_{hardface}$ is not directly used in continuous authentication but it is stored for use in relogin authentication. Face recognition is executed at regular intervals (1 second).

3) Computing the final similarity: the final similarity S_{cont} is calculated based on equation 3.5. If S_{cont} is below a threshold (t_{cont}), the system enters Mode III to determine whether it is due to user's absence in front of the console or the illumination change.

3.4.3 Enrollment Template Update (Mode III)

When the similarity S_{cont} falls below t_{cont} in continuous authentication mode (Mode II), the system status changes to Mode III. We introduce this mode to reduce the false rejects caused by illumination changes. This process consists of two steps.

1) Illumination change detection: every time S_{cont} is lower than t_{cont} in continuous authentication mode (Mode II), the system analyzes whether: i) there has been a change in the ambient illumination or ii) the user is no longer in front of the system. To detect the illumination change, the simple method of image subtraction is used. We use a pair of images for image subtraction: one just before the time when $S_{cont} \leq t_{cont}$, and the other immediately after the time when $S_{cont} \leq t_{cont}$. The system counts the number of pixels that show a large difference in brightness between the two images, and decides that there has been an illumination change if the difference image shows intensity differences all over the image. Fig. 3.5 shows two image subtractions results; there is an illumination change between Fig. 3.5 (d) and (e), while there is no change between Fig. 3.5 (a) and (b).

2) Enrollment template update: the system updates the user's biometric template when an illumination change is detected in this mode [z_0 in 3.1, 3.2,

and 3.3] to keep successful continuous authentication in the modified operating environment without reauthentication. After the enrollment template update, the system status moves to continuous authentication mode (Mode II) again.

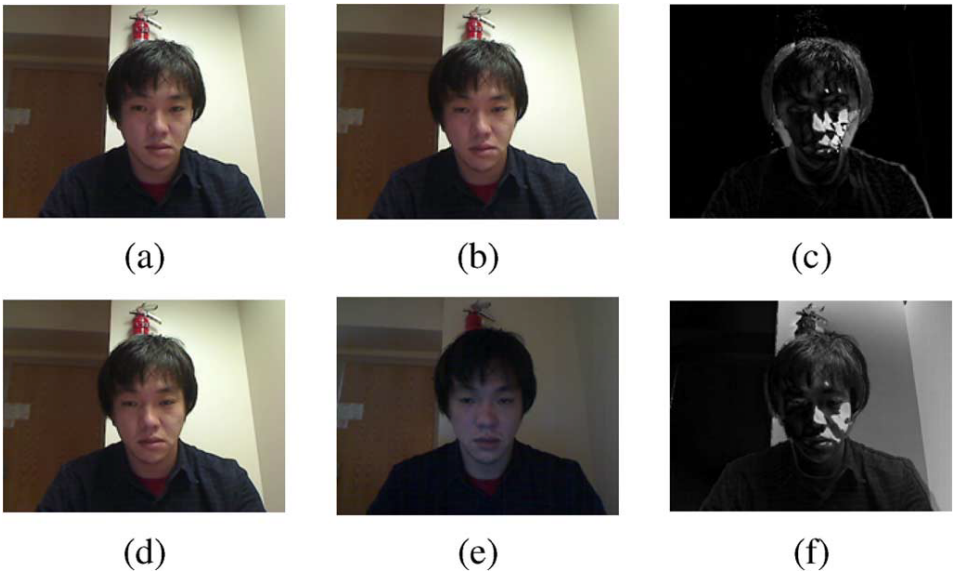

Figure 3.5: Example of image subtraction for illumination change detection. The difference image in (f) shows an illumination change between (d) and (e), but the difference image in (c) does not show a change in illumination between (a) and (b).

3.4.4 Relogin Authentication (Mode IV)

When the system identifies that the user is no longer in front of the system, the status changes to this mode and the system is locked. In this mode, the system tries to reauthenticate the user automatically. The status moves to the continuous authentication mode (Mode II) again if the system detects a user and reauthenticates the user as genuine. The relogin authentication mode includes four steps. Steps 1), 2), and 3) are the same procedures as those used in steps 2), 3), and 4) in Mode I. In step 4) of the relogin authentication mode, the system authenticates the user using both soft (face and clothing color) and hard biometrics (hard face). The similarity score $S_{relogin}$ shown in 3.6 is used to determine whether the user is genuine in the relogin authentication mode.

3.4.5 Overall Flow of Proposed Algorithm

Fig. 3.6 presents a detailed flowchart of the proposed algorithm. We address the problem of existing methods for continuous authentication by using both soft and hard biometric traits. The system enters relogin authentication mode when there is a discontinuity in the similarity scores based on the soft biometric. In the relogin authentication mode, valid soft and hard biometric traits need to be provided.

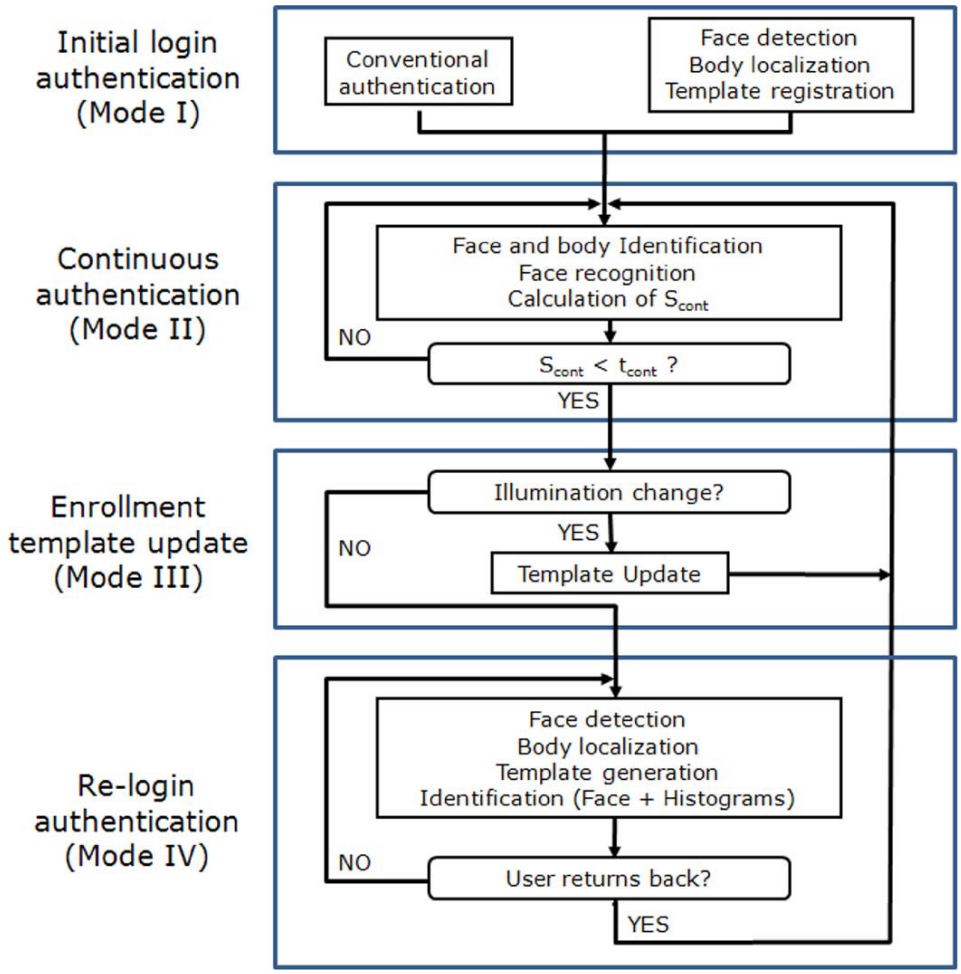

Figure 3.6: Overall flowchart of proposed algorithm.

Figure 3.7: Continuous authentication system setup used in our experiments: laptop with a webcam.

3.5 Experiments

3.5.1 System Configuration

As shown in Fig. 3.7, we used a laptop and a webcam to collect videos for our experiments. The system has the following characteristics that are conducive to users, especially for PC or laptop users

- Robustness to changes in user's posture.
- No requirement for user to preregister.
- Real-time continuous user authentication capability.
- No requirement for a specific background (robust to cluttered background).

3.5.2 Database

Using the system shown in Fig. 3.7, videos of 20 subjects were collected to evaluate the proposed continuous authentication approach. We asked each user to sit in front of the webcam, and perform the following set of actions.

- Scenario A: turning head to the left;

- Scenario B: turning head to the right;
- Scenario C: turning head down;
- Scenario D: leaning back in chair;
- Scenario E: stretching arms;
- Scenario F: walking away.

The video length ranges from 54 to 143 seconds. The frame rate of the videos is 15 frames/s, and the frame size is 640x480 pixels. Fig. 3.8 shows some example images of the videos. The red and green ellipses show the face and body regions automatically tracked by the system.

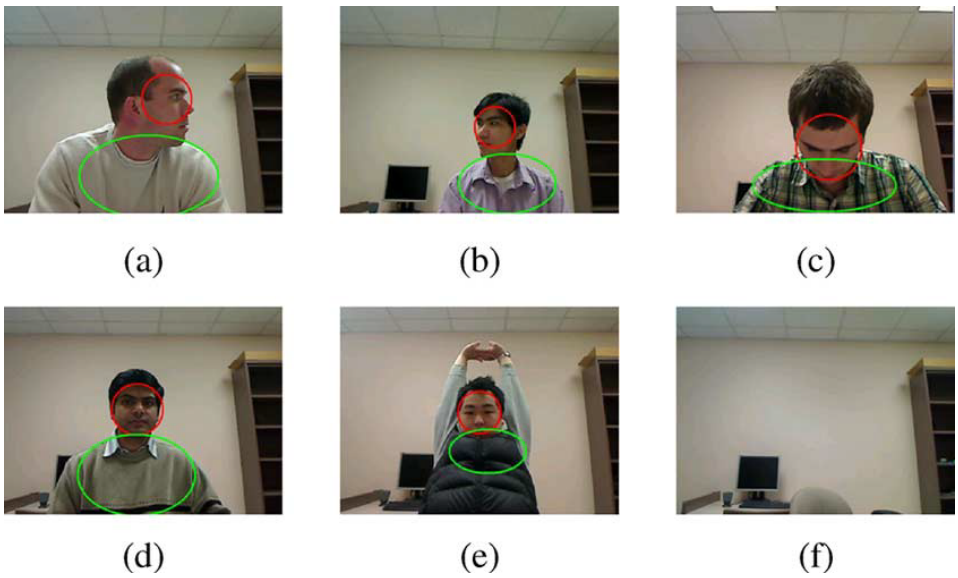

Figure 3.8: Example of video frames. The red ellipses indicate the face regions and the green ellipses indicate clothing region identified by our system. (a) Turning head to left; (b) turning head to right; (c) turning head down; (d) leaning back in chair; (e) stretching arms; and (f) walking away.

3.5.3 Performance Evaluation

We use false accept (FA) and false reject (FR) to measure our system's performance. The false accept and false reject are defined below:

- False Reject (FR): The system wrongly identifies that a user is not in front of the console even though the user is still in front of the camera. False rejects lower the usability of the system.
- False Accept (FA): The system incorrectly identifies that a legitimate user is in front of the console even though the legitimate user is not in front of the camera. False accepts lower the security of the system.

Various performance metrics such as Time to Correct Reject (TCR), Probability of Time to Correct Reject (PTCR), Usability, and Usability-Security Characteristic Curve (USC) were proposed by Sim et al. [75] for continuous authentication using hard biometric traits. However, they are not suitable to evaluate the continuous authentication system using soft biometric traits. Because more frequent observations on a user's biometric traits are available in the proposed continuous authentication system, we use false accept (FA) and false reject (FR) for each event (e.g., turning head away) to measure our system's performance rather than the delayed time until a correct decision is made. Our system can make an immediate decision since the continuous and frequent soft biometric observation is available in our system.

Table 3.2: Performance evaluation (False Reject Rate and False Accept Rate) of the continuous authentication system

Scenario	False Reject Rate	False Accept Rate
A) Turning head to the left	0%(= 0/20)	0%(= 0/20)
B) Turning head to the right	0%(= 0/20)	0%(= 0/20)
C) Turning head down	10%(= 2/20)	0%(= 0/20)
D) Leaning back in a chair	5%(= 1/20)	0%(= 0/20)
E) Stretching arms overhead	10%(= 2/20)	0%(= 0/20)
F) Walking away	0%(= 0/20)	0%(= 0/20)

Table 3.2 shows our experimental results based on data collected on 20 users. There is a small number of false rejects in Scenario C (10%), D (5%) and E(10%). We have selected the parameters ($w = 0.5$, $t_{cont} = 0.6$, $t_{hardface} = 0.8$, and $t_{relogin} = 0.6$) after trying several different threshold values. The main factors of false rejects are:

- Because of the illumination variations, the color histogram of the user's clothing significantly changed. We typically observed this problem when the color

of the user's clothing is white as shown in Fig. 3.9. Clothing that is white in color is more susceptible to change in illumination compared to other colors, especially in scenarios D and E.

- The user's face is completely occluded. In this case, no color histogram of the face could be computed [Fig. 3.16 (c)].

(a) (b)

Figure 3.9: Example of False Reject (FR). (a) Enrollment. (b) Authentication.

Fig. 3.10 shows the changes in similarity values ($S_{clothes}$, $S_{softface}$, $S_{hardface}$, and S_{cont}) while the user performs various actions in front of the webcam according to our scenarios. The corresponding video frames are shown in Fig. 3.11. In Fig. 3.10, green lines indicate the transition of $S_{clothes}$, red lines indicate the transition of $S_{softface}$, blue lines indicate the transition of $S_{hardface}$, and black dots represent $S_{hardface}$. Hard face authentication is only performed every 1 second. The range of similarity scores ($S_{clothes}$, $S_{softface}$, $S_{hardface}$, and S_{cont}) is [0, 1], and a higher score represents a better matching. Fig. 3.12 shows a plot similar to Fig. 3.10, but a Commercial-Off-The-Shelf (COTS) face matching system FaceVACS [1] was used instead of our PCA-based methods. Since both Eigenface and FaceVACS show similar performance, Eigenface was used in the remaining experiments (Figs. 3.13, 3.15, 3.17, 3.19, 3.22, and 3.25).

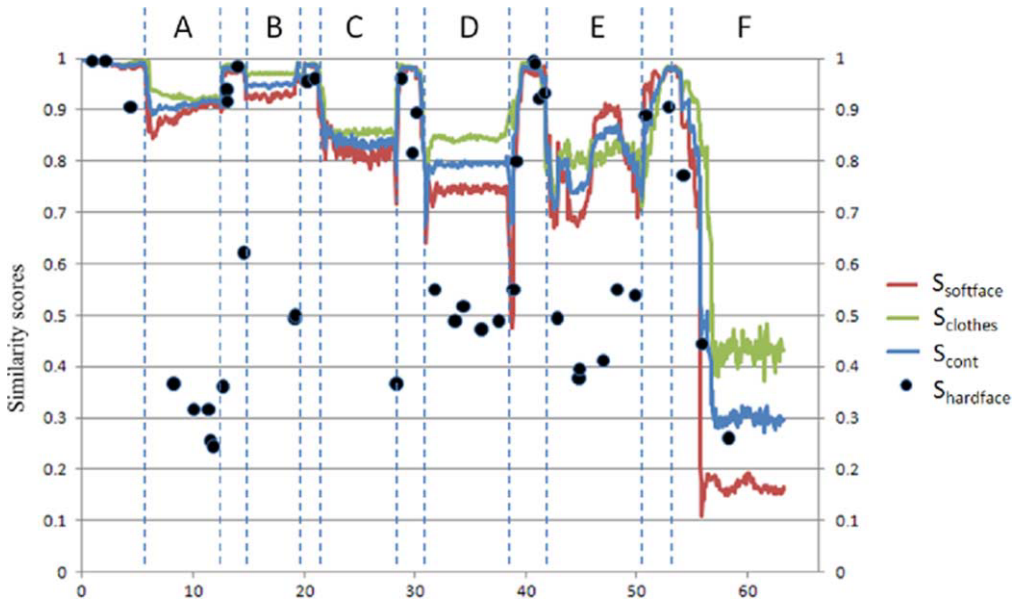

Figure 3.10: Example 1 of similarity scores versus time graph. Eigenface is used to calculate hardface.

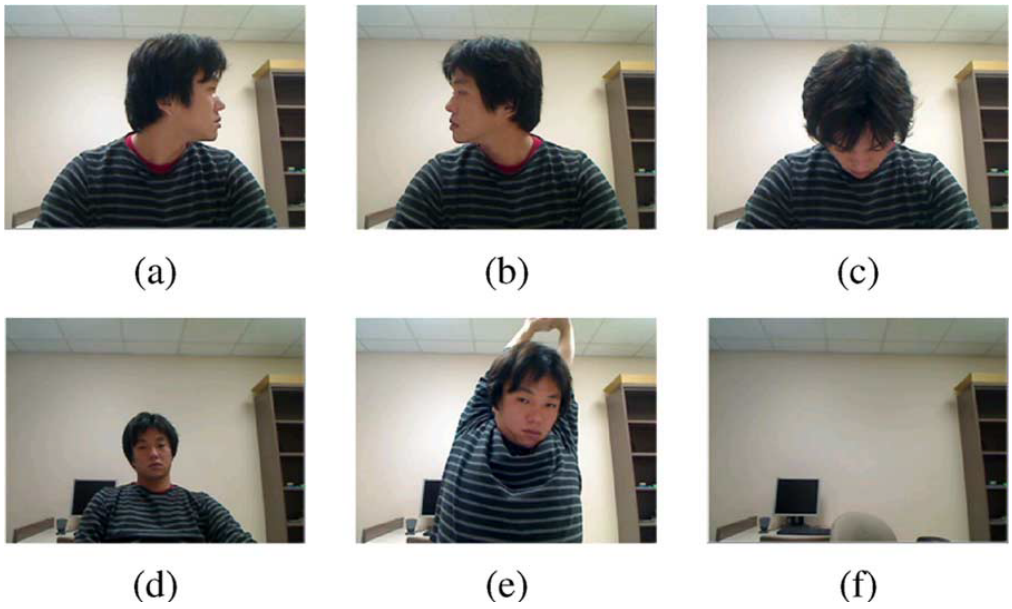

Figure 3.11: Examples of images used to generate the graphs in Figs. 3.10 and 3.12 . (a) Turning head to left; (b) turning head to right; (c) turning head down; (d) leaning back in chair; (e) stretching arms; and (f) walking away.

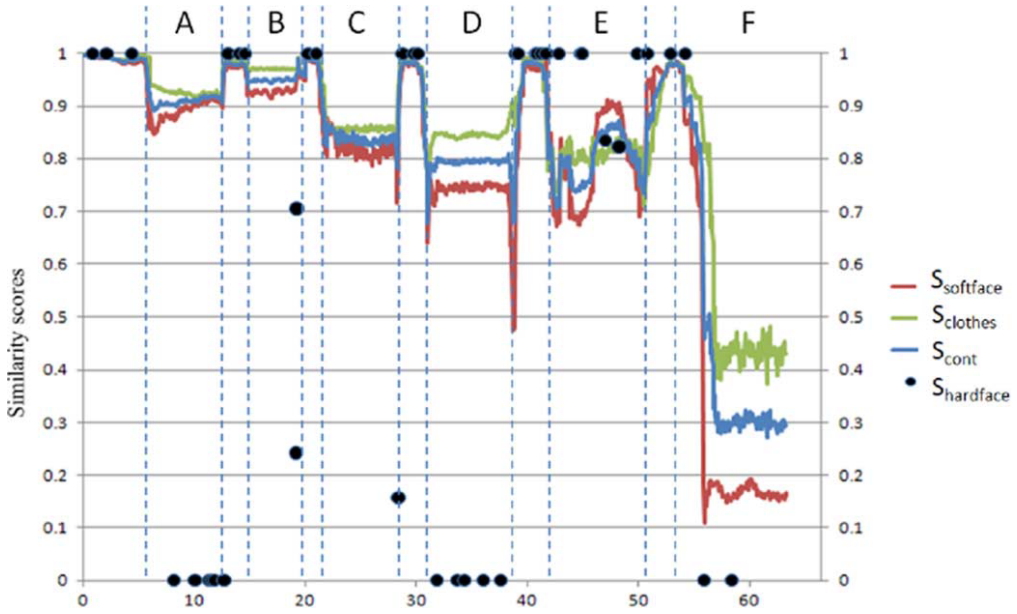

Figure 3.12: Example 1 of similarity score versus time graph. FaceVACS is used to calculate hardface.

Figs. 3.13 and 3.15 show the results for the different users. The corresponding video frames are shown in Figs. 3.14 and 3.16, respectively. In Figs. 3.10 and 3.13, the similarities $S_{clothes}$, $S_{softface}$, and $S_{hardface}$ remain high regardless of the user's posture (scenarios A–E), and they go down rapidly as soon as the user walks away from the console (scenario F) while the hard face similarity is not very stable depending on the user's posture. The results demonstrate the advantage of our approach using soft biometric traits for continuous authentication. In Fig. 3.15, false reject (FR) occurs during scenario C. This is because the system failed to track both the face and body correctly as the user was looking down. Fig. 3.16 (c) shows the corresponding input video frames leading to FR. The user's face is completely occluded.

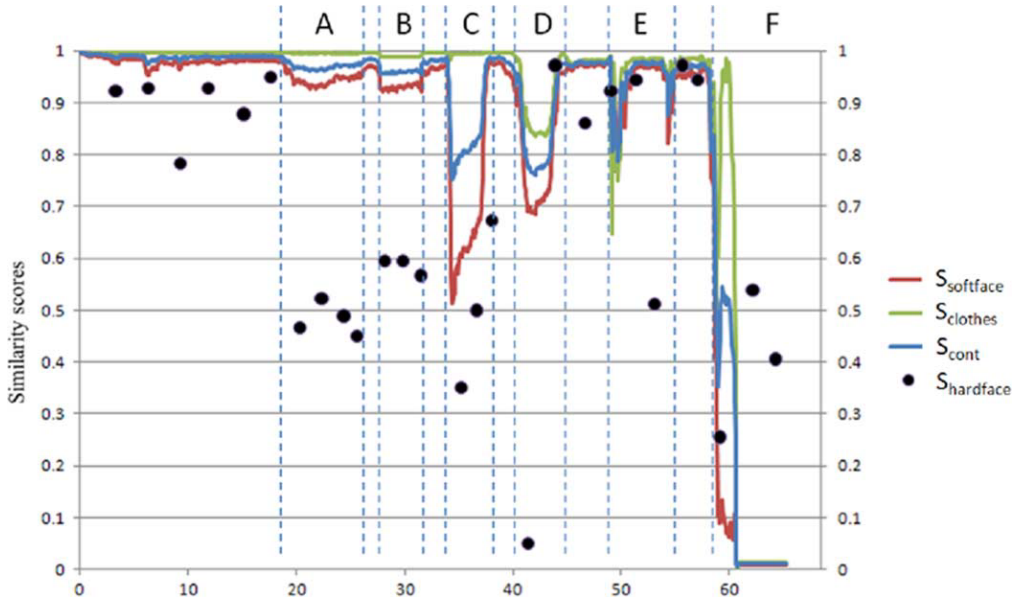

Figure 3.13: Example 2 of similarity scores versus time graph. Eigenface is used to calculate hardface.

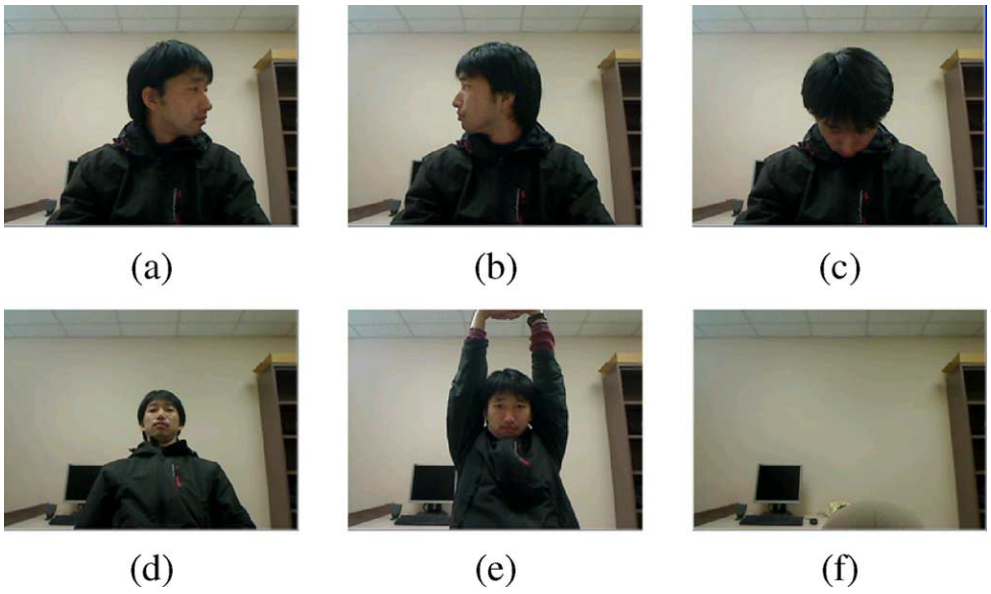

Figure 3.14: Examples of images used to generate the graphs in Fig. 3.13. (a) Turning head to left; (b) turning head to right; (c) turning head down; (d) leaning back in chair; (e) stretching arms; and (f) walking away.

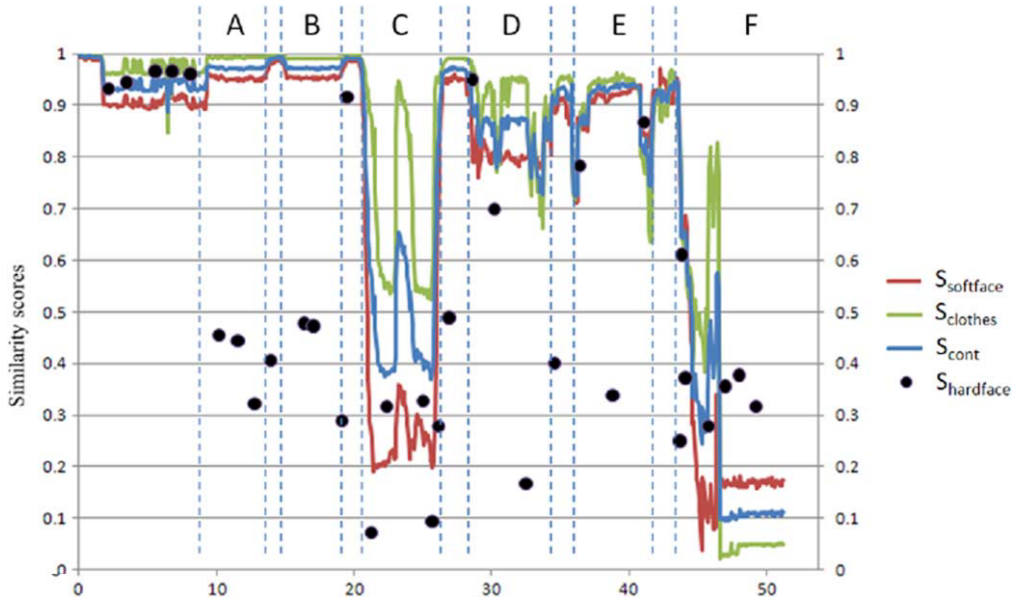

Figure 3.15: Example 3 of similarity scores versus time graph. Eigenface is used to calculate hardface.

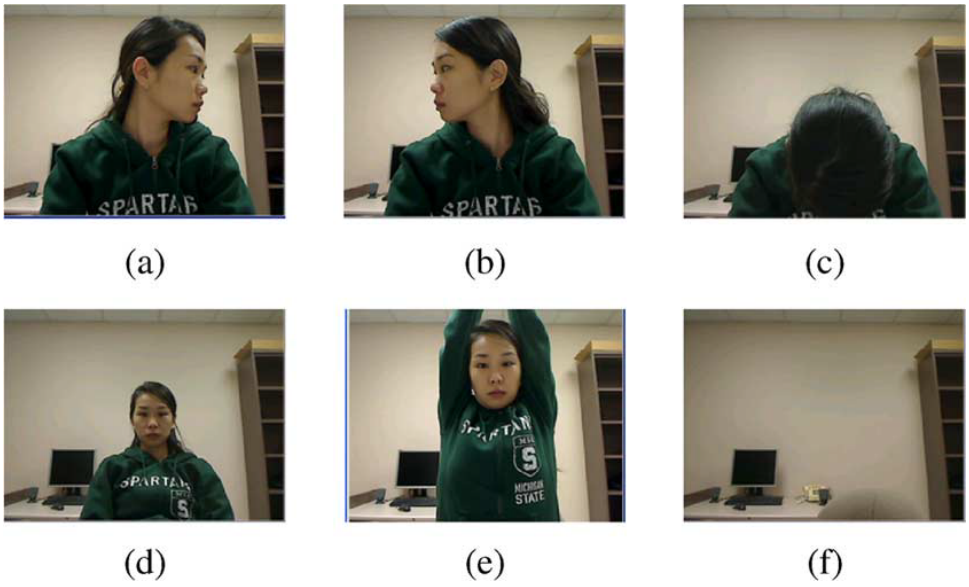

Figure 3.16: Examples of images used to generate the graphs in Fig. 3.15. (a) Turning head to left; (b) turning head to right; (c) turning head down; (d) leaning back in chair; (e) stretching arms; and (f) walking away.

To further demonstrate the robustness of the proposed system, we also performed the following additional experiments: 1) illumination change detection, 2) relogin authentication, 3) occlusion, and 4) laptop with a built-in camera.

Illumination Change Detection

Fig. 3.18 show two frames of the user where where illumination change is observed. For the scenario in Fig. 3.18, the results of various similarity computation over time without and with enrollment update are shown in Fig. 3.17 (a) and (b), respectively.

In Fig. 3.17 (a), the similarity values of soft biometric traits decrease rapidly as soon as the illumination change occurs. On the other hand, in Fig. 3.17 (b), due to template update, the similarity values of soft biometric traits remain high even after the illumination change. Note that the scores of face recognition also decline after the illumination change. To tackle the problem, we can replace our PCA-based face matcher with a more advanced face recognition engine, but it will still fail with large pose variations, as shown in Fig. 3.12. Another example is shown in Fig. 3.19 (corresponding video frames are shown in Fig. 3.20). In this case, the user also shifted his position along with the illumination change. The transition of the results of various similarity computations without and with enrollment update, are shown in Fig. 3.19 (a) and (b), respectively. Fig. 3.19 (b) demonstrates that the system is able to successfully identify the user even after the illumination change with only slight fluctuations in the similarity scores of soft biometric traits. The false illumination detection rates using the same data used in Section 3.5.3 are shown in Table 3.3. In the experiment, we do not observe any false detection due to illumination change.

Table 3.3: False reject (FR) rates in the presence of illumination change

Scenario	False Reject Rate
A) Turning head to the left	0%
B) Turning head to the right	0%
C) Turning head down	0%
D) Leaning back in a chair	0%
E) Stretching arms overhead	0%
F) Walking away	0%

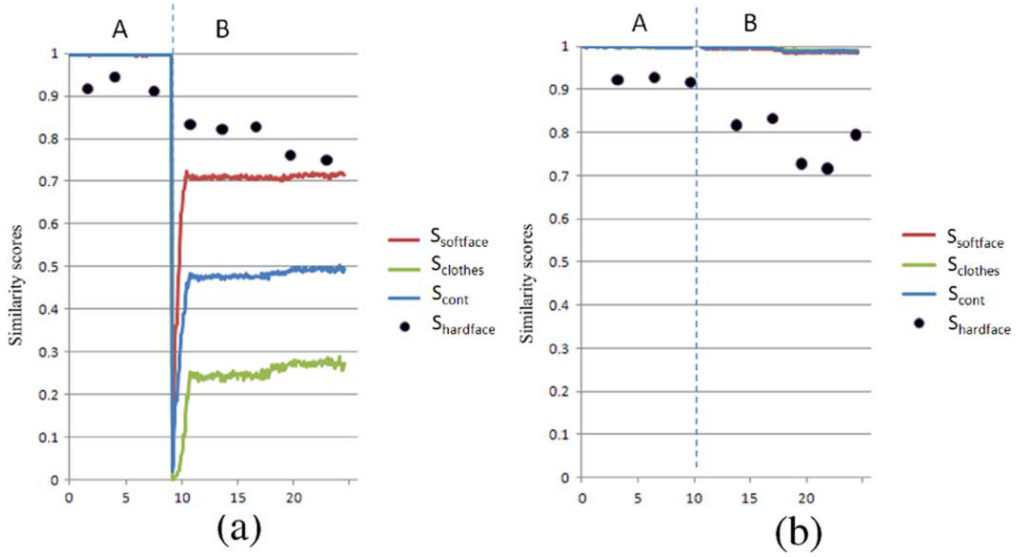

Figure 3.17: Example 1 of similarity scores versus time graphs with and without enrollment update. (a) Without enrollment update. (b) With enrollment update.

Figure 3.18: Examples of images before and after the illumination change used to generate graphs in Fig. 3.17. (a) Dark room. (b) Bright room.

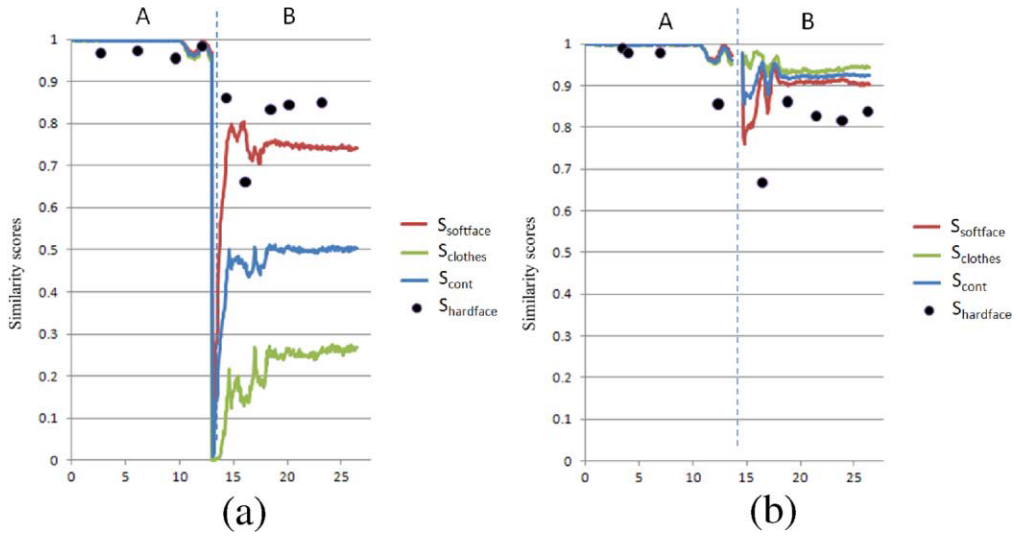

Figure 3.19: Example 2 of similarity scores versus time graphs with and without enrollment update. (a) Without enrollment update. (b) With enrollment update.

Figure 3.20: Examples of images before and after the illumination change used to generate graphs in Fig. 3.19. (a) Dark room. (b) Bright room.

Relogin Authentication

We evaluated the proposed relogin authentication method using video clips. The video clips illustrate the following scenario: a) an authorized user logs in, b) the user leaves the console (without logging out), and then c) another user (an impostor) appears in the field of view of the webcam. Fig. 3.21 shows this scenario. In Fig. 3.21, the colored ellipses show that the system recognized the valid user, and the black-and-white images show that the system recognized the absence of the valid user. As shown in Fig. 3.21, the system successfully recognized both the valid user and the impostor.

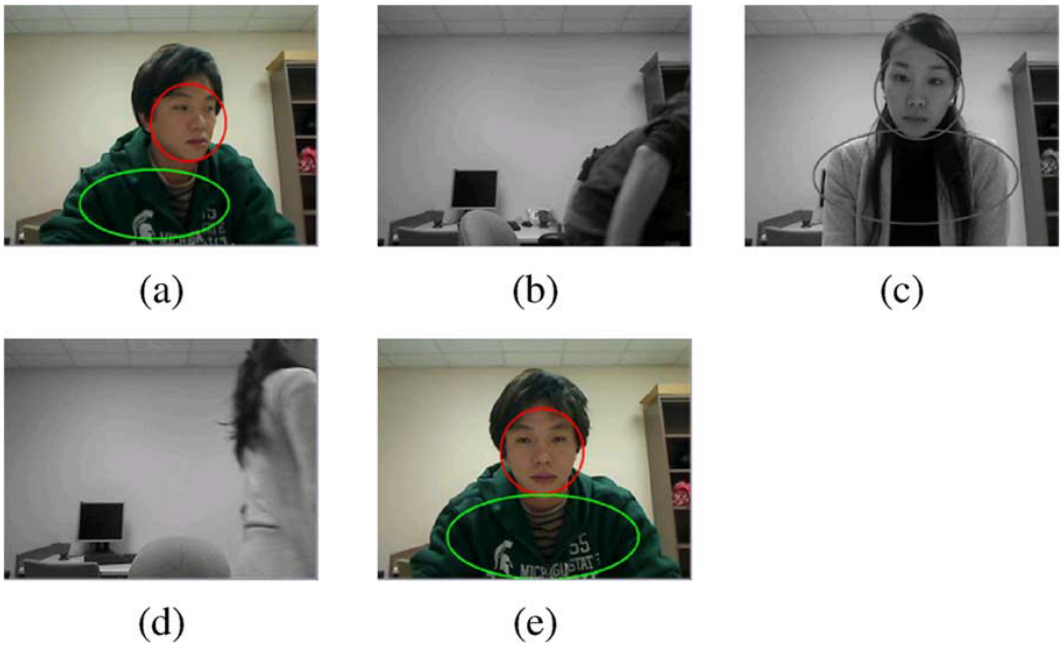

Figure 3.21: Example results of relogin authentication experiments. (a) Authentic user; (b) authentic user walks away; (c) imposter user; (d) imposter user walks away; and (e) authentic user returns.

Occlusion

We partly evaluated occlusion in earlier experiments (turning head down to occlude both face and clothing in Figs. 3.10, 3.13, and 3.15), but we conducted experiments with a more explicit occlusion scenario with a paper file.

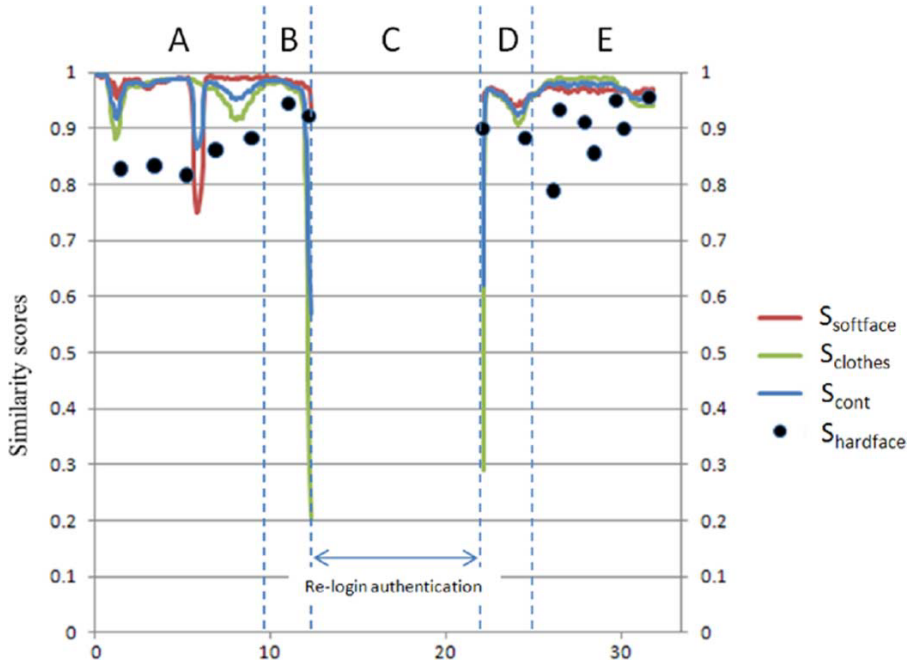

Figure 3.22: Example of similarity scores versus time graph with occlusion.

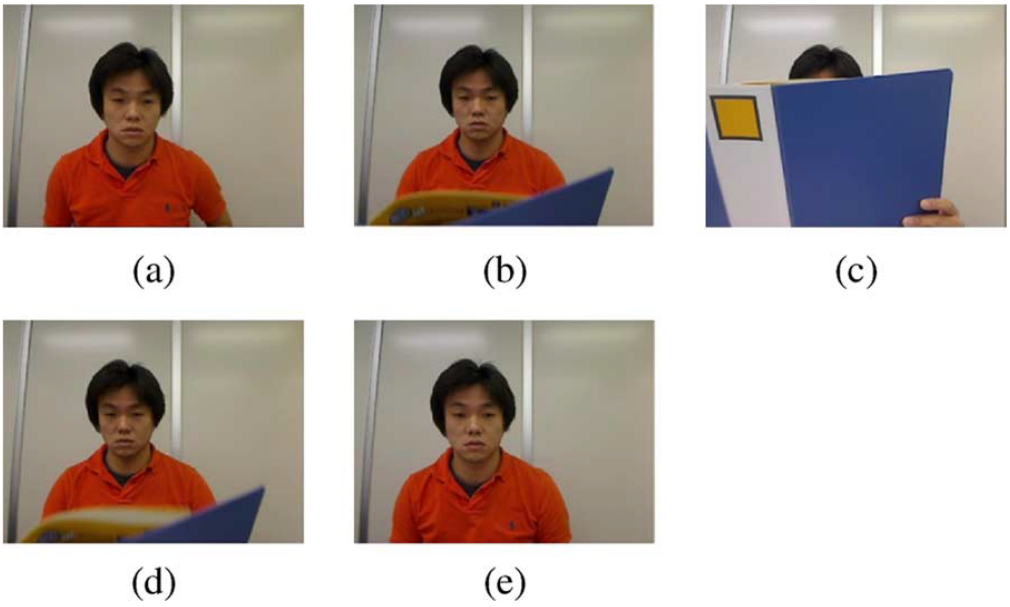

Figure 3.23: Examples of images used to generate the graph in Fig. 3.22. (a), (b), (c), (d), and (e) correspond to time instants A, B, C, D, and E in Fig. 3.22.

Fig. 3.22 show the results. The corresponding video frames are shown in Fig 3.23. The results indicate that the proposed system successfully authenticated the user with the occlusion event. While the user is occluded, the system status is in relogin authentication mode. After the user's face and clothes become available, he is reauthenticated successfully, and the system status moves back to continuous authentication mode. In the current system, if the occlusion occurs for a long time, the user will not be accepted by the system. In this case, the system status moves to initial login mode, and the user needs to start over from the initial login mode. Note that, in Fig. 3.22, the hard face biometric has similar performance as the soft biometrics because the face is frontal in most of the video frames.

Laptop With Built-In Webcam

We have also conducted experiments using a built-in webcam while in the previous experiments we used a webcam externally mounted on the laptop screen.

Fig. 3.24 shows the laptop with a built-in camera (red ellipse). The frame rate and the image size are the same as the ones with a external webcam (frame rate: 15 frames/s, image size: 640x480 pixels). The images captured from the built-in webcam are slightly blurry and show low saturation. Regardless of that, Figs. 3.25 and 3.26 show that our system successfully authenticates the user continuously.

(a) (b)

Figure 3.24: (a) Laptop with a built-in webcam and (b) close up view of the built-in camera.

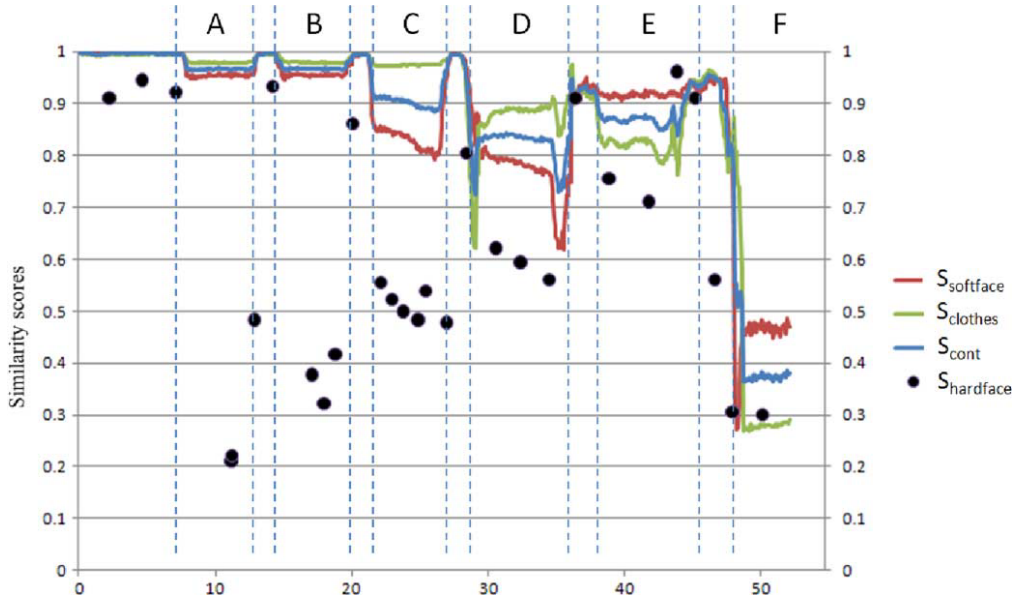

Figure 3.25: Example of similarity scores versus time graph for the built-in webcam.

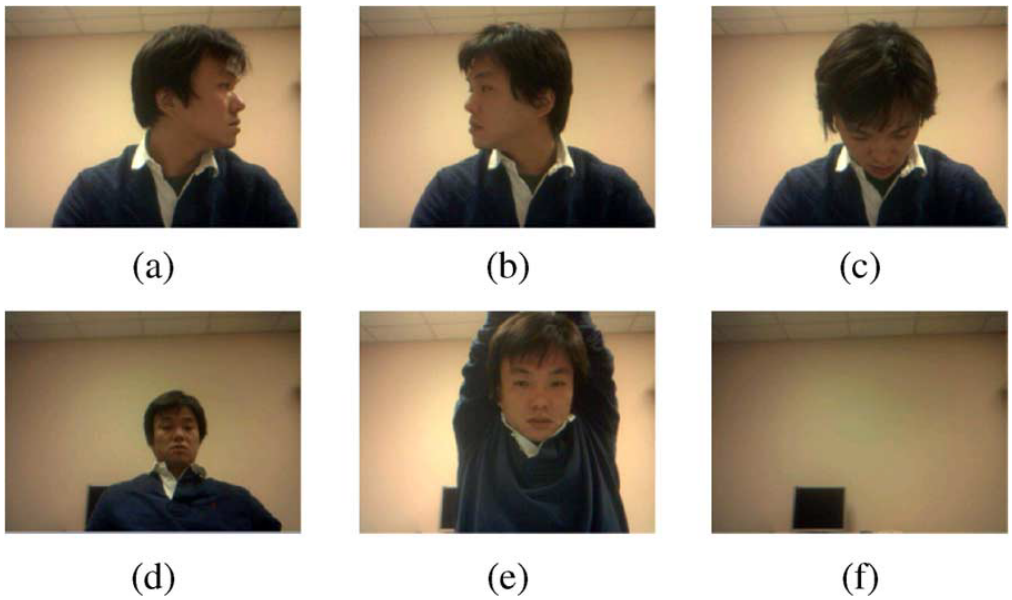

Figure 3.26: Example of images from built-in webcam used to construct the similarity score versus time graph of Fig. 3.25. (a) Turning head to left; (b) turning head to right; (c) turning head down; (d) leaning back in chair; (e) stretching arms; and (f) walking away.

3.5.4 System Attacks

This section discusses vulnerability of the proposed system, and possible approaches to improve the systems' vulnerabilities.

1. **Initial login time with stolen password**: In case a stolen password is used in the login process, utilizing multiple login authentication methods, such as hard biometric authentication, is a practical method to prevent the system attack. It should be noted, however, this approach does not prevent all the attacks because the hard biometric trait itself can be compromised.

2. **Continuous authentication mode**: If an attacker has soft biometric traits that are very similar to those of the authorized user, the attacker can breach the system. To minimize the problem, we introduce the relogin authentication mode. Whenever the system recognizes that the legitimate user is not in front of the console, the system status moves to the relogin authentication mode, and the reauthentication using both hard and soft biometrics are required. In addition, when there is a sudden lighting change, the system first checks whether it is due to the change in the soft biometric trait (absence of the user) or the lighting change as shown in Chapter 3.4.3. If the change is due to the absence of the user, the system enters relogin authentication mode. If the change is due to the lighting condition, the system automatically updates the templates (soft and hard biometric traits), and continuously authenticates the user in the continuous authentication mode.

3. **Relogin authentication mode**: When the user has very similar soft biometric traits (i.e., clothing and face color) and face appearance, an attacker can breach the system at the relogin authentication mode. To mitigate the problem, the time decaying function is used in the relogin authentication mode. This can block an attacker after a certain time lapse.

3.6 Summary

To tackle the problem of existing methods for continuous authentication, we have proposed a new framework that uses soft biometric traits as well as hard biometric traits. The proposed framework enables the system to effectively use soft biometric

traits by enrolling a new template every time the user logs into the system. Face color and clothing color histograms are used as soft biometrics. Because the soft biometric traits are available regardless of the user's posture, the proposed framework is robust to the variation of user's posture. In addition, it has the capability for enrollment template update mode and relogin authentication. Experimental results demonstrate the effectiveness of the proposed framework for continuous user authentication.

Chapter 4

Systematic evaluation of design choices for deep facial action coding across pose

4.1 Introduction

Emotion recognition technologies have made a significant contribution in a wide range of applications including remote communication, online education, products evaluation, and social robots. Facial action units (AU) that correspond to discrete muscle contractions have been widely used to recognize human emotion. Individually or in combinations, they can account for nearly all possible facial expressions of emotion.

The performance of automated facial affect recognition systems have improved steadily in detection of the occurrence and intensity of facial actions. While early work focused on relatively controlled laboratory settings, more recent work emphasizes less-constrained in-the-wild scenarios [20, 52, 95]. For facial affect recognition systems, robustness to pose variation is essential since frontal face views cannot be assumed in less constrained settings. The Facial Expression Recognition and Analysis 2017 (FERA 2017) provided the first common protocol to evaluate robustness to pose variation [88]. In FERA 2017, deep-learning based approaches have shown the best performance in sub-challenges ([82] for occurrence detection, [97] and intensity estimation).

The advantage of modern deep learning techniques is clear, but little is known about critical design choices among them. Most studies use default or ad-hoc parameters provided by the deep learning frameworks, and neglect to examine the effect of different parameter settings on facial action unit (AU) detection. Therefore, little is known about the relative contribution of different design choices in pre-training, feature alignment, model size, and optimizer details.

In this study, we specifically focus on design choices in two scenarios that are important for real-world applications.

1. Robustness to pose variation: Until recently, most systems were concerned with near-frontal face views. However, because pose variation is common in real-world settings, robustness to pose variation is critical.

2. Transfer to new domains: Many real-world applications are required to be applicable in new contexts. Therefore, it is crucial for systems to perform well in the domains to which they may be applied, as well as the domains from which they come. However, evaluation of domain transfer in AU systems is relatively new [21, 28].

To tackle these two questions, we systematically evaluated the combinations of different components and their parameters in a modern deep-learning based pipeline. Our design choices include pre-training practices, image alignment for pre-processing, training set sizes, optimizers, and learning rates. Informed what we found, we developed an architecture that outperforms state-of-the-art methods on both the occurrence and the intensity sub-challenges of FERA 2017 [88]. Our architecture also achieved state-of-the art in cross-domain generalizability to the Denver Intensity of Spontaneous Facial Action (DISFA) dataset [61]. We also report evaluation of cross-pose generalizability and performance of cross-domain generalizability. Our architecture performs well on the unseen views and domains. We visualize occlusion sensitivity maps in order to understand and interpret at which facial regions our architecture looks to detect specific AUs at specific poses. The occlusion sensitivity maps demonstrate that our architecture attends to meaningful facial regions for different poses and AUs.

Table 4.1: Comparison of the design choices from existing methods using the FERA 2017 dataset. F_1 scores are reported for occurrence detection, and Intraclass Correlation coefficients (ICC) are reported for intensity detection. Best scores are denoted in bold. N/A denotes not applicable; N/R denotes not reported.

	Design choice						Performance	
	Normalization	Architecture	Pre-training	Training set size per model	Optimizer	Learning rate	Occurrence performance (F_1 score)	Intensity performance (ICC)
Valstar et al. [88]	Facial landmarks	Shallow	n/a	n/r	n/a	n/a	0.452	0.217
Li et al. [54]	Facial landmarks	Hybrid	VGG-Face[1]	26,582	n/a	n/a	0.498	n/a
Batista et al. [12]	Face position	Deep	none[2]	1,321,472	Adam	10^{-3}	0.506	0.399
He et al. [36]	Resizing[3]	Hybrid	none	146,847	n/r	n/r	0.507	n/a
Tang et al. [82]	Face position[4]	Deep	VGG-Face	440,541 $+\alpha$[5]	SGD	10^{-3}	**0.574**	n/a
Ertugrul et al. [27]	Face position	Deep	none	1,321,623	Adam	10^{-3}	0.525	n/a
Li et al. [53]	Facial landmarks	Deep	ImageNet-VGG-VD19	260,000 $+\alpha$[6]	SGD	10^{-4}	n/a[7]	n/a
Amirian et al. [5]	Facial landmarks	Shallow	n/a	n/r	n/a	n/a	n/a	0.295
Zhou et al. [97]	Resizing	Deep	ImageNet-VGG-VD16	54,000	SGD	10^{-4}	n/a	**0.446**

[1] A VGG pre-trained model was used to extract features, but not used for classification.

[2] A VGG pre-trained model was used to detect faces, but not used for classification.

[3] Face detection was used for train and validation partition, but not for test partition.

[4] Face position was not directly used, but facial images were cropped by using morphology operations including binary segmentation, connected components labeling and region boundaries extraction.

[5] After down sampling to 440,541 images, Tang et al. increased the number of samples to balance positive and negative samples.

[6] Li et al. increased the number of samples to balance positive and negative samples.

[7] In their paper, Li et al. reported F1 scores only on validation partition.

4.2 Related work

A great number of approaches have been proposed for action unit (AU) analysis [20, 24, 60]. While most of them assume that face orientation has been relatively frontal, some methods tackle non-frontal pose [46, 80, 41, 71, 85]. However, the lack of a common protocol has undermined comparisons. The FERA 2017 Challenge [88] provided the first common protocol to compare approaches in terms of robustness to pose variation. The FERA 2017 dataset contains synthesized face images with 9 head poses as shown in Fig. 4.1. The training set of the FERA 2017 dataset consists of the BP4D database [93], which includes digital videos of 41 participants Both the development and test sets consists of BP4D+ [94]. The development set includes digital videos of 20 participants, and the test set includes digital videos of 30 participants. By using the FERA 2017 dataset, we can evaluate two tasks: occurrence detection and intensity estimation. For the occurrence detection, 10 AUs were labelled; for the intensity estimation, 7 AUs were labelled.

A wide range of approaches have been evaluated using the FERA 2017 dataset. Table 4.1 compares the performance and design choices of the methods proposed in the FERA 2017 challenges and two more recent approaches from Ertugrul et al.[27] and Li et al.[53]. For FERA 2017, F_1 score was used to evaluate performance for occurrence detection, and Intraclass Correlation (ICC) was used to evaluate intensity estimation.

Several comparisons are noteworthy in Table 4.1. As for normalization, we can observe the difference between shallow approaches and deep learning approaches. Precise face alignment using facial landmarks was used for shallow approaches while simple face alignment using face position or resized images are often used for deep learning approaches. In addition, for architecture, deep learning approaches performed better than shallow approaches, and deep learning approaches with pre-trained models achieved better performance than ones without pre-trained model. For both of the sub-challenges, deep learning approaches with a pre-trained model were used by the methods achieving the best performance (Tang et al. [82] for occurrence detection, and Zhou et al. [97] for intensity estimation). As for training set size, a different number of training images were used by each method. Either Adam or SGD was used as optimizer, and learning rate varied between 10^{-3} and 10^{-4}. The comparison of the existing methods indicates the effectiveness of deep learning

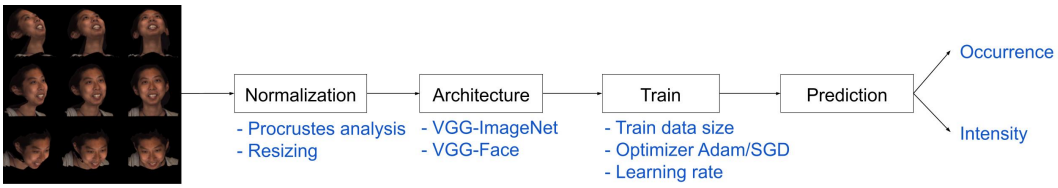

Figure 4.1: An overview of our experimental design. Blue color denotes design choices and parameters for systematic evaluation.

approaches, especially the ones using pre-trained models, for this task. However, a different fixed configuration was used by every approach, and the key parameters are unknown.

To address the problem, we systematically investigate the key parameters for both AU occurrence and intensity estimation, and show the optimal configuration.

4.3 Methods

In this study, we investigate the effect of the different components and parameters by systematically evaluating design choices of deep-learning based facial expression analysis, and provide best practices that researchers can use for training deep learning methods for this task.

An outline of our experimental design is shown in Fig. 4.1. On the basis of the outline, we systematically changed parameters and design choices. Key elements are represented in blue color in Fig. 4.1. In every experiment, we explored the effect of optimizer choice and parametric variation of an additional key parameter. Table 4.2 shows our baseline configuration. In this work, the PyTorch framework was used perform the experiments.

4.4 Experiments

4.4.1 Normalization

Two image normalization methods are compared: Procrustes analysis and Resizing.

1. **Procrustes analysis**: To apply detailed face alignment, we used Procrustes analysis [32] though many existing methods used a face alignment approach

Table 4.2: Baseline configuration

Design choice	Baseline configuration
Normalization	Procrustes analysis
Pre-trained architecture	VGG-ImageNet
Training set size	5,000 images per each pose each AU
Optimizer and learning rate	Adam with lr=5×10^{-5} SGD with lr=5×10^{-3}
Fine-tune layer	From the third convolutional layer
Dropout	0.5

based on eye locations. A limitation of a face alignment approach based on eye locations is that alignment error increases for landmarks farther away from the eye region. This limitation is especially problematic for our evaluation because the FERA 2017 dataset includes a wide range of pose variations. To mitigate the problem, we use Procrustes analysis. More specifically, we first extracted 68 facial landmarks from each image using the dlib face tracker [44], and then applied a Procrustes transform between the extracted landmarks and a frontal looking template. The size of the template covers a bounding box of 224x224 pixels that match the receptive field of the VGG network.

2. **Resizing**: In the second case, each image was resized to 224x224 pixels that corresponds to the receptive field of the VGG network.

The F_1 scores and ICC averages for all nine poses for each AU are shown in Fig. 4.2. The results for two optimizers are shown separately (the left figures for Adam optimizer, and the right figures for SGD optimizer). The results indicate that the difference between Procrustes analysis and Resizing is small (1% or less) though the performance with Procrustes analysis is slightly better than the one with Resizing. One possible explanation for the small difference is that the VGG network has enough capacity to learn all the nine different poses.

4.4.2 Pre-trained architecture

In this experiment, we evaluate pre-trained architecture. Pre-trained architecture is a well-known technique to train deep learning models because training deep models

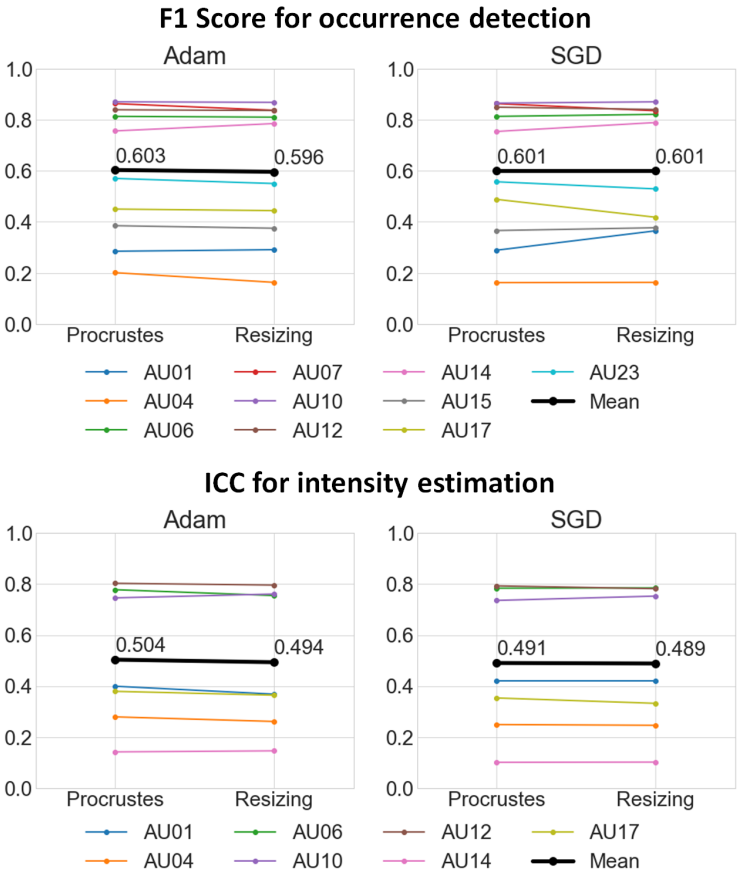

Figure 4.2: Results on FERA 2017 Test partition with two normalization methods.

from scratch is time-consuming and the amount of training data at hand may impede good performance. In addition, as we discussed in Chapter 4.2, approaches with pre-trained models show better results than ones without pre-trained models for the task. The technique has two steps: 1) select a model that was trained on large scale benchmark datasets (source domain), and 2) fine-tune it on the data of our interest (target domain).

Although it is known that this practice is effective, little is known about how the type of data in the source domain influences the performance of fine-tuning in the target domain. To investigate the question, two models that were trained on very different domains were selected in our experiments: VGG-16 trained on ImageNet [77] and VGG-Face [67]. The final layers of each network are replaced with a 2-length one-hot representation for AU occurrence detection. Similarly, the final layers are

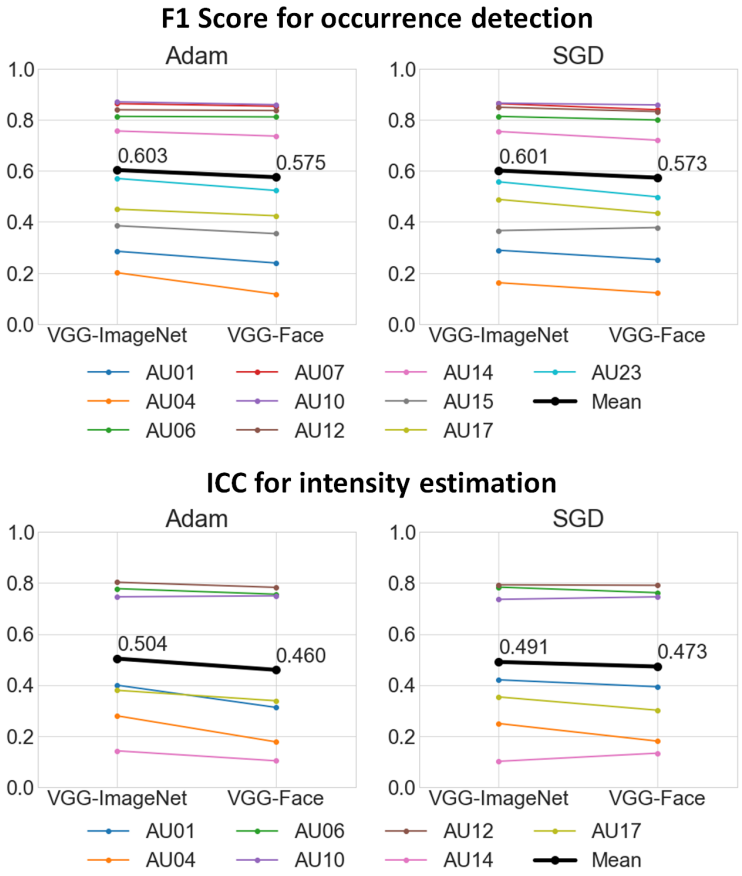

Figure 4.3: Results on FERA 2017 Test partition with two pre-trained architecture.

replaced with a 6-length one-hot representation for the intensity estimation task. For both tasks, separate models were trained for each AU, resulting 10 and 7 models for AU occurrence detection and AU intensity estimation, respectively. Our models were fine-tuned for 10 epochs. We validated performance on the validation partition, then reported results on the subject-independent test partition.

Fig. 4.3 shows the results. These results indicate that models pre-trained on ImageNet achieved better performance than the VGG-Face ones. While VGG-Face was trained on face images for identification, ImageNet includes many non-face images for image classification. We attribute this result as follows. VGG-Face learned to actively ignore facial expression in order to recognize the face. As a result, VGG-ImageNet (a generic image representation) is more suitable for the task.

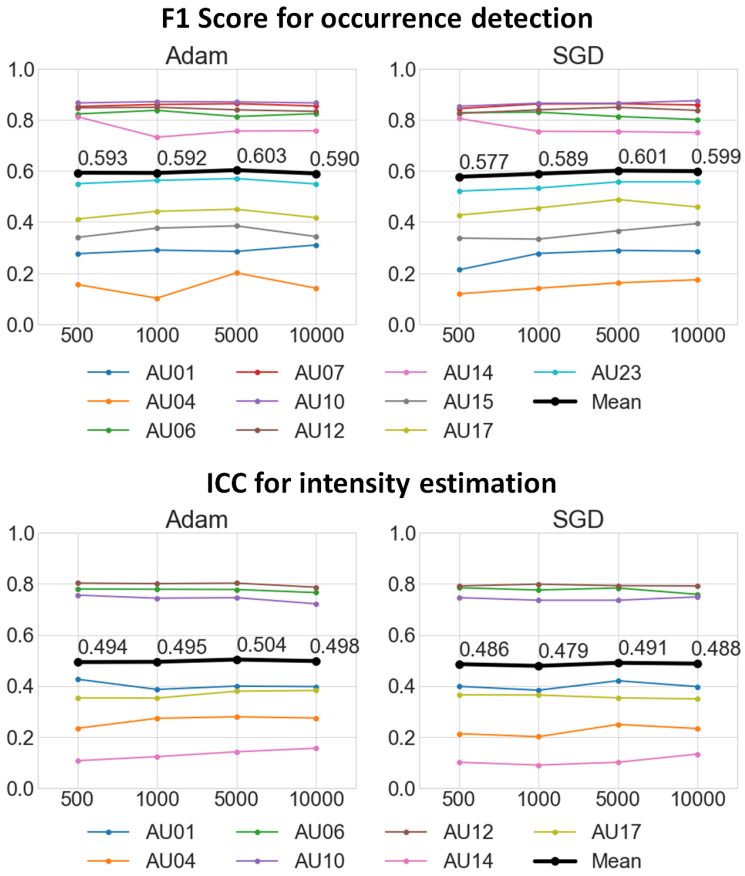

Figure 4.4: Results on FERA 2017 Test partition with different number of train set size.

4.4.3 Training set size

Chu et al. [19] found that multi-label stratified sampling was advantageous over naive sampling strategies for AU detection. We employ this strategy, and explore the effect of different training set sizes on the performance. More specifically, we down-sampled the majority class and up-sampled the minority class to build a stratified training set. We applied this procedure for each pose and each AU. For example, in the case of AU occurrence detection, a $5,000$ training set size means that $5,000$ images with AU present and $5,000$ images without AU present were randomly selected for each pose and for each AU, resulting in $90,000$ images in total (=$5,000$ images x 2 classes x 9 poses). We repeated the same stratifying procedure with the

six classes of the intensity sub-challenge. In this case, a $5,000$ training set size indicates that $5,000$ images were randomly selected from the six classes (not present, and A to E levels) for each pose and for each AU, resulting in $270,000$ images in total (=$5,000$ images x 6 classes x 9 poses).

Fig. 4.4 indicates that the training set size has minor influence on the performance. We use $5,000$ images in the rest of our experiments because scores peaked at $5,000$ images.

4.4.4 Optimizer and learning rate

We investigated the impact of two different optimizers (SGD and Adam) and learning rates (LR) on the performance. All of the deep learning based methods shown in Table 4.1 used SGD or Adam. We varied the learning rates, but we used the default values used in PyTorch for the other optimizer parameters: betas=(0.9, 0.999) without weight decay for Adam, and no momentum, no dampening, no weight decay and no Nesterov acceleration for SGD.

Fig. 4.5 indicates two of note. First, the optimal learning rate is largely different depending on the choice of optimizer. For Adam, LR=5×10^{-5} showed the best results, while LR=0.01 showed the best performance for SGD. Secondly, if the optimal learning rates for each optimizer are used, the performance differences between Adam and SGD are negligible.

When the learning rate was set to a large value, some models did not converge and predicted the majority class for all samples. In this case, ICC converges to zeros. However, this should not be interpreted as chance performance. As variation in predicted intensity values reduces, the ICC metric loses predictive power. It is also worth noting that Zhou et al. [97] used SGD with LR=10^{-4} for the AU intensity estimation task. Our results indicate that their performance could be improved by using Adam optimizer or SGD optimizer with larger learning rate. Tang et al. [82] used SGD with LR=10^{-3}, but they also applied momentum. Our additional experiments revealed that when momentum is used for SGD, smaller learning rate is preferable for optimal performance. More specifically, when we used the same parameters as Tang et al. [82] reported for SGD (momentum=0.9, weight decay=0.02) F_1 score peaked at 0.596 using LR=10^{-4}. The results indicate that their learning rate is close to optimal. However, SGD without momentum further

improves F_1 score to 0.609 with LR=0.01.

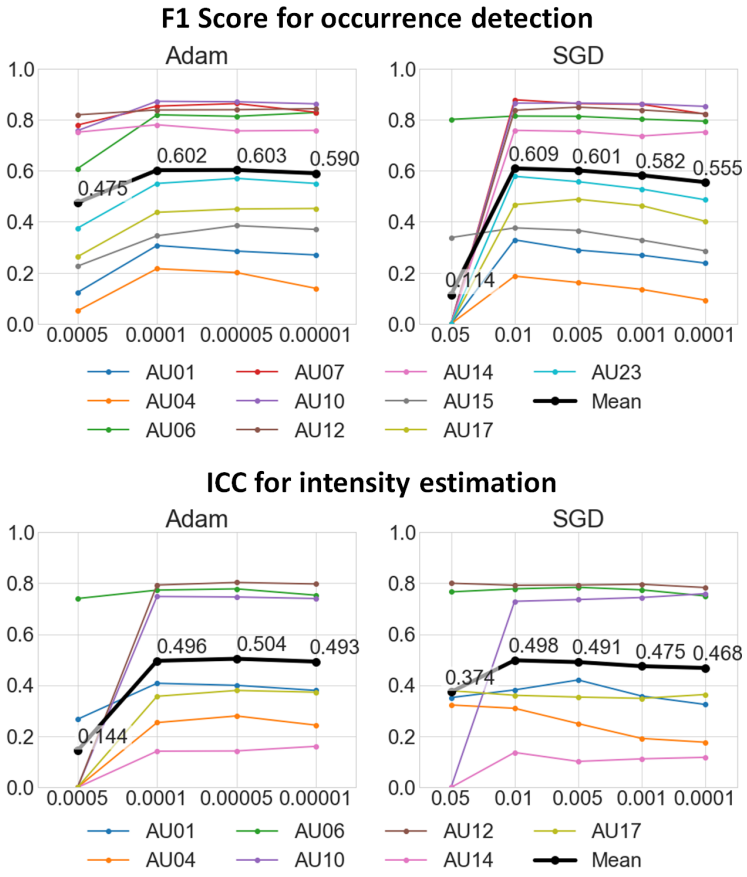

Figure 4.5: Effect of learning rates and choice of optimizers on the FERA 2017 Test partition.

4.4.5 Comparison with existing methods

The optimal parameters of our models are almost the same for the two tasks as shown in Table. 4.3. For AU occurrence detection, SGD with LR=0.01 gave the best result ($F_1 = 0.609$), while for AU intensity estimation, Adam with LR=5×10^{-5} reached the best performance (ICC = 0.504). Table 4.4 shows F_1 score and Accuracy for occurrence detection, and Table 4.5 shows ICC for intensity estimation.

Table 4.6 and 4.7 show the comparison of our method and the state-of-the-art on the AU occurrence detection and the AU intensity estimation, respectively. Our

Table 4.3: Optimal parameters of design choices for occurrence detection and intensity estimation

Design choice	Occurrence detection	Intensity estimation
Normalization	Procrustes analysis	Procrustes analysis
Pre-trained architecture	VGG-ImageNet	VGG-ImageNet
Training set size	5,000 images (each pose each AU)	5,000 images (each pose each AU)
Optimizer and learning rate	SGD with lr=0.01	Adam with lr=5×10^{-5}

method outperforms the state-of-the-art. We note a few key differences that contributed to this achievement. The main difference with Tang et al. [82] is pre-trained architecture. Tang et al. used VGG-Face pre-trained model while we used ImageNet pre-trained model. The key difference with Zhou et al. [97] is learning rate. Zhou et al. used SGD with small learning rate while the combination of our optimizer and learning rate is optimal. Li et al. [53] also evaluated their method for AU occurrence detection using the FERA 2017 dataset, but they reported performance only on the Validation partition. Their best F_1 score (0.522) is 9% lower than ours (0.611) on the Validation partition.

Table 4.4: F1 scores and Accuracy of our model for occurrence detection under 9 facial poses on FERA 2017 Test partition.

F1 score										
Pose	1	2	3	4	5	6	7	8	9	Mean
AU01	0.358	0.292	0.272	0.353	0.346	0.366	0.312	0.314	0.345	0.329
AU04	0.247	0.208	0.129	0.254	0.226	0.217	0.131	0.135	0.133	0.187
AU06	0.808	0.803	0.788	0.828	0.830	0.811	0.829	0.821	0.811	0.814
AU07	0.887	0.886	0.864	0.877	0.883	0.885	0.871	0.875	0.878	0.878
AU10	0.859	0.867	0.864	0.868	0.872	0.868	0.872	0.870	0.841	0.865
AU12	0.821	0.830	0.850	0.830	0.843	0.850	0.833	0.847	0.828	0.837
AU14	0.756	0.737	0.742	0.758	0.776	0.771	0.787	0.759	0.735	0.758
AU15	0.422	0.419	0.369	0.408	0.379	0.357	0.357	0.340	0.336	0.376
AU17	0.453	0.485	0.493	0.461	0.492	0.486	0.482	0.430	0.416	0.466
AU23	0.568	0.588	0.577	0.611	0.597	0.588	0.557	0.568	0.545	0.578
Mean	0.618	0.612	0.595	0.625	0.624	0.620	0.603	0.596	0.587	0.609
Accuracy										
Pose	1	2	3	4	5	6	7	8	9	Mean
AU01	0.855	0.835	0.862	0.834	0.850	0.871	0.811	0.831	0.844	0.844
AU04	0.961	0.949	0.900	0.944	0.926	0.919	0.944	0.937	0.907	0.932
AU06	0.828	0.828	0.807	0.847	0.847	0.838	0.838	0.819	0.805	0.829
AU07	0.850	0.848	0.813	0.837	0.841	0.844	0.828	0.832	0.836	0.837
AU10	0.830	0.836	0.834	0.834	0.838	0.835	0.833	0.829	0.798	0.830
AU12	0.793	0.807	0.834	0.808	0.821	0.832	0.817	0.831	0.802	0.816
AU14	0.708	0.690	0.686	0.709	0.722	0.718	0.733	0.694	0.676	0.704
AU15	0.814	0.802	0.785	0.795	0.780	0.792	0.782	0.753	0.711	0.779
AU17	0.744	0.790	0.787	0.764	0.789	0.794	0.778	0.754	0.739	0.771
AU23	0.769	0.782	0.764	0.803	0.782	0.775	0.786	0.790	0.757	0.779
Mean	0.815	0.817	0.807	0.818	0.820	0.822	0.815	0.807	0.788	0.812

Table 4.5: ICC of our model for intensity estimation under 9 facial poses on FERA 2017 Test partition.

Pose	1	2	3	4	5	6	7	8	9	Mean
AU01	0.441	0.449	0.400	0.403	0.436	0.433	0.353	0.354	0.333	0.400
AU04	0.305	0.278	0.250	0.294	0.333	0.281	0.317	0.244	0.216	0.280
AU06	0.779	0.786	0.787	0.787	0.788	0.786	0.762	0.776	0.754	0.778
AU10	0.763	0.750	0.738	0.759	0.763	0.768	0.734	0.722	0.720	0.746
AU12	0.799	0.812	0.815	0.809	0.813	0.812	0.795	0.797	0.777	0.803
AU14	0.144	0.161	0.162	0.137	0.143	0.153	0.124	0.141	0.126	0.143
AU17	0.393	0.396	0.403	0.394	0.388	0.382	0.383	0.359	0.319	0.380
Mean	0.518	0.519	0.508	0.512	0.523	0.516	0.495	0.485	0.464	0.504

Table 4.6: F1 scores for occurrence detection results on FERA 2017 Test partition.

	Valstar et al. [88]	Li et al. [54]	Batista et al. [12]	He et al. [36]	Ertugrul et al. [27]	Tang et al. [82]	Our
AU01	0.147	0.215	0.219	0.198	0.196	0.263	**0.329**
AU04	0.044	0.044	0.056	0.043	0.067	0.118	**0.187**
AU06	0.630	0.755	0.785	0.747	0.766	0.776	**0.814**
AU07	0.755	0.805	0.816	0.784	0.791	0.808	**0.878**
AU10	0.758	0.810	0.838	0.816	0.840	**0.865**	**0.865**
AU12	0.687	0.753	0.780	0.809	0.819	**0.843**	0.837
AU14	0.668	0.750	0.747	0.691	**0.764**	0.757	0.758
AU15	0.220	0.208	0.145	0.208	0.247	0.362	**0.376**
AU17	0.274	0.286	0.388	0.398	0.349	0.424	**0.467**
AU23	0.342	0.356	0.286	0.374	0.413	0.519	**0.578**
Mean	0.452	0.498	0.506	0.507	0.525	0.574	**0.609**

Table 4.7: ICC for intensity estimation on FERA 2017 Test partition.

	Valstar et al. [88]	Amirian et al. [5]	Batista et al. [12]	Zhou et al. [97]	Our
AU01	0.035	0.169	0.228	0.307	**0.400**
AU04	-0.004	0.021	0.057	0.147	**0.280**
AU06	0.461	0.509	0.702	0.671	**0.778**
AU10	0.451	0.590	0.710	0.735	**0.746**
AU12	0.518	0.615	0.732	0.793	**0.803**
AU14	0.037	-0.027	0.104	**0.147**	0.143
AU17	0.020	0.190	0.260	0.319	**0.380**
Mean	0.217	0.295	0.399	0.446	**0.504**

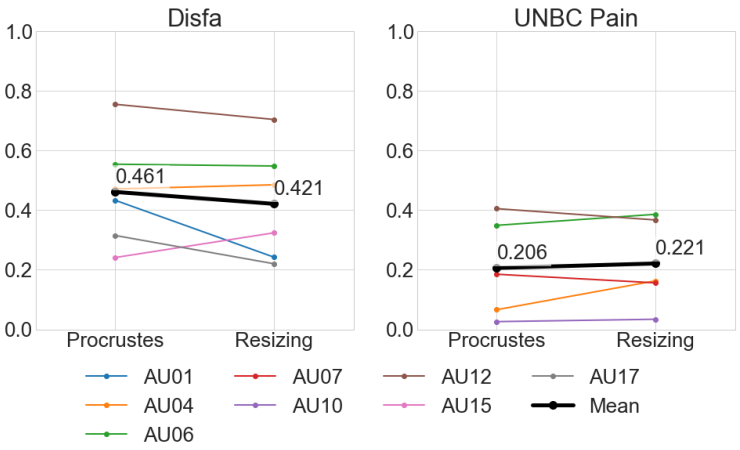

Figure 4.6: F_1 scores for occurrence detection on DISFA and UNBC Pain with two normalization methods.

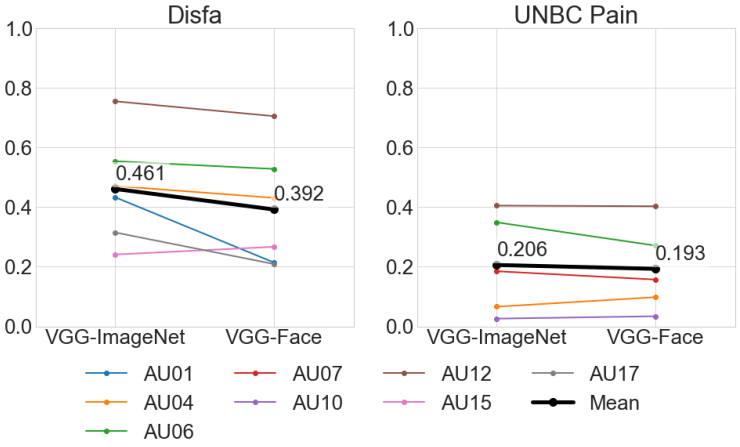

Figure 4.7: F_1 scores for occurrence detection on DISFA and UNBC Pain with two pre-trained architecture.

4.4.6 Cross-domain evaluation

To evaluate the generalizability of our method to unseen conditions, we report performance on the Denver Intensity of Spontaneous Facial Action (DISFA) [61] and UNBC McMaster Pain [58] datasets. In the experiments, we investigate the impact of differences in environments including illumination, cameras, orientation of

Table 4.8: Comparison of cross-domain performance to DISFA dataset for occurrence detection.

	Accuracy		2AFC		F1		AUC
	Our	Ghosh et al. [31]	Our	Ghosh et al. [31]	Our	Baltrušaitis et al., [11]	Our
AU01	**0.932**	0.838	**0.714**	0.660	0.475	-	0.787
AU04	0.806	**0.833**	0.723	**0.740**	0.531	-	0.809
AU06	**0.860**	0.703	0.758	**0.870**	0.567	-	0.867
AU12	**0.859**	0.624	0.859	**0.873**	**0.742**	0.700	0.934
AU15	**0.823**	0.752	**0.671**	0.617	0.253	-	0.761
AU17	**0.738**	0.689	**0.742**	0.585	**0.361**	0.260	0.823
Mean	**0.836**	0.740	**0.745**	0.724	**0.488**	-	0.830

Table 4.9: Cross-domain performance to DISFA dataset for intensity estimation.

	ICC
AU01	0.533
AU04	0.560
AU06	0.451
AU12	0.747
AU17	0.319
Mean	0.522

the face, quality and diversity of the training data. Note that we did not perform fine-tuning on the target domain in these experiments.

We evaluated occurrence detection and intensity estimation performance of our system, but both DISFA and UNBC Pain were annotated with AU intensity labels. To create binary AU occurrence labels, we thresholded the 6-points intensity values at A-level (the AU is present if A-level or higher). In these experiments, the baseline configuration with Adam optimiser is used (See Table 4.2). To detect a face from each image, we used the built-in face detector in dlib [44]. As for Resizing, to include whole faces, we extended the boxes of detected face positions by 30%, and

Table 4.10: Cross-domain performance to UNBC Pain dataset for occurrence detection and intensity estimation.

Occurrence Detection			Intensity Estimation	
	F1	AUC		ICC
AU04	0.195	0.863	AU04	0.152
AU06	0.249	0.720	AU06	0.262
AU07	0.188	0.784	AU07	-
AU10	0.028	0.743	AU10	0.018
AU12	0.405	0.785	AU12	0.388
Mean	0.213	0.779	Mean	0.205

then cropped and resized the boxes to 224x224 pixels.

Figs. 4.6 and 4.7 show the results. The F_1 scores with two normalization methods, Procrustes analysis and resizing, are shown in Fig. 4.6 and, the F_1 scores with two pre-trained architectures, VGG-ImageNet and VGG-Face, are shown in Fig. 4.7. The results for DISFA show that Procrustes analysis and VGG-ImageNet show better performance. However, the results show that the differences are small for UNBC Pain.

Table 4.8 and Table 4.9 show the best results on DISFA for occurrence detection and intensity estimation, respectively. Similarly, Table 4.10 shows the best results on UNBC Pain for both tasks. In these experiments, the previously trained CNN models reported in Chapter 4.4.5 were used. Table 4.8 compares our approach with other cross-domain methods for occurrence detection on DISFA. Both Ghosh et al. [31] and Baltrušaitis et al. [11] used BP4D to train their model, and thresholded AU intensity values at A-level to create binary events. For a fair comparison, we also report Accuracy and 2AFC scores, that Ghosh et al. [31] used. Our approach outperforms their method in both metrics. Baltrušaitis et al. [11] report cross-domain scores only for two AUs (AU 12 and AU17). Our models show better performance for both AUs. These results show the robustness of our model for cross-domain situation.

The performance on UNBC Pain is much lower than the one on DISFA. We attribute this as follows. The image size of UNBC Pain (320x240 or 352x240) is

smaller than the other two datasets (FERA2017: 1024x1024, and DISFA: 1024x768), and the base rates on UNBC Pain is also small (DISFA: 13.3%, and UNBC Pain: 7.2%). Additionally, facial expressions on UNBC Pain are mainly associated with pain, and the correlation among AUs differs from that of FERA2017 and DISFA. Table 4.8 and 4.10 also show AUC for occurrence detection.

4.4.7 Cross-pose evaluation

To evaluate the generalization of our method to unseen poses, we also performed cross-pose experiments. Our baseline configuration with Adam optimizer is used. In the experiments, two types of experiments are reported:

1. The architecture was trained using eight of the nine poses of training set and tested with the remaining pose of test set (Fig. 4.8)

2. The architecture was trained using one pose of training set and tested with nine poses of test set (Fig. 4.9).

,

Fig. 4.8 shows the differences between models trained with eight poses and with all nine poses as the results of the first experiment. In the figure, zero values means that the performance between two models are the same and plus values indicate that the performance with eight poses is better than the one obtained with nine poses. The black lines indicate mean differences. Fig. 4.8 shows that the difference of the performance between nine poses and eight poses is small though the performance with nine poses is slightly better than that with eight poses. These results demonstrate that our architectures can generalize well to unseen poses.

Fig. 4.9 shows the results of the second experiment. Each 3x3 matrix shows the performance of each model. Each cell of a matrix shows the performance of each pose, and the blue rectangle indicates a pose that was used to train a model. A cell of each matrix corresponds to the pose in the same cell given in Fig. 4.1. For example, with respect to a model trained with Pose1, the F_1 score is 0.604 when we test it with Pose1 of test set, and the F_1 score is 0.446 when we test it with Pose9 of test set. The figure shows that we obtained maximum results within-pose. When the models are tested with the poses in the neighboring cells, small performance decreases are observed. However, the performance is largely decreased

when a model is tested with largely different poses.

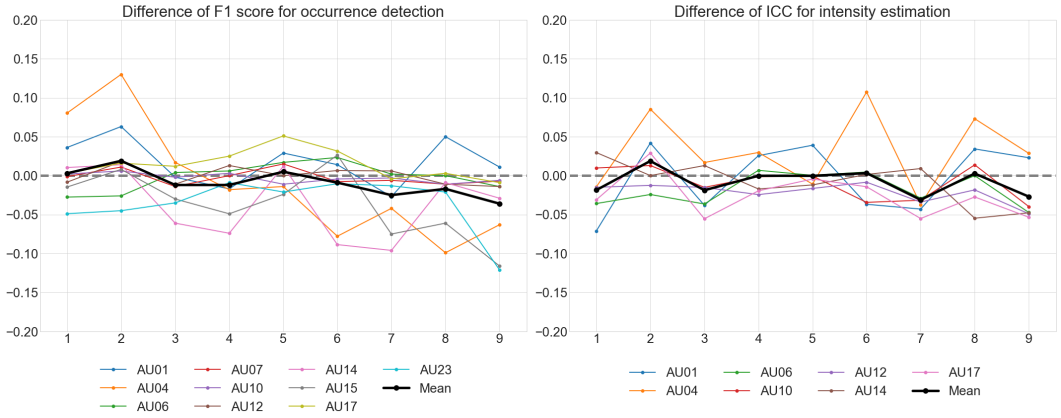

Figure 4.8: Performance difference between models trained with eight poses and with nine poses. Horizontal axis shows each pose, and vertical axis indicates performance difference between two models.

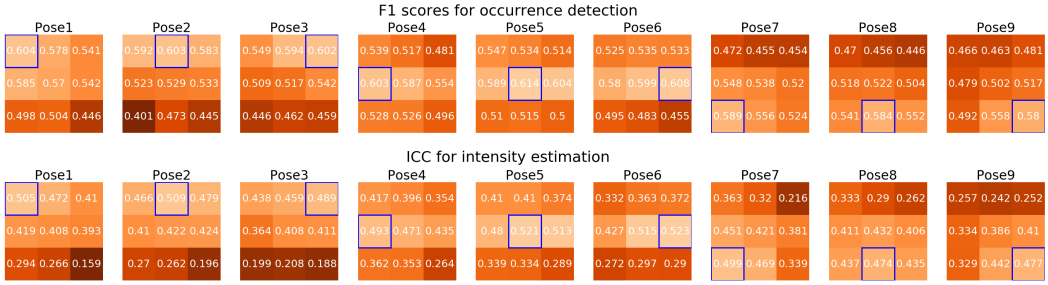

Figure 4.9: F_1 scores and ICC for models using one pose of training set and test them with nine poses of test set. We report only mean values.

4.4.8 Occlusion sensitivity maps

In order to understand and interpret our architecture, we generated Occlusion Sensitivity Maps [92] for each pose and each AU. An occlusion patch having size 45x45 with Gaussian random noise were used. The patch was slid over the original image of size 224x224 with a stride 15. For each AU each pose, 200 images were selected: 100 images that contained the specific AU, and 100 images that did not contain it. The 200 images per each AU each pose were tested to obtain accuracy values.

Fig. 4.10 shows the generated maps. In these maps, darker red colors represent lower accuracy values. Significant regions are the ones colored with red because their occlusion causes the largest decrease in the accuracy.

The results show that significant regions are correctly localized. We can also see that significant regions in Fig. 4.10 are off to the left side even for frontal faces. This seems to be reasonable because the pitch and yaw rotations of images in FERA2017 datasets is in one direction as shown in Fig. 4.1. We also created a mirrored image for each image in training sets, and trained models using the combination of images with nine original poses and nine mirrored poses. Fig. 4.11 shows the results. As we can see, significant regions are in a center location.

4.4.9 Saliency maps

To compare the learned features, saliency maps using Vanilla Backpropagation [76] were generated. Fig. 4.12 shows the results. For each AU each pose, we selected 200 images: 100 images that contained the specific AU, and 100 images that did not contain it. We then obtained a mean image of saliency maps from the 200 images. Brighter areas indicate the areas that are more important for the classifier to detect the related AU. Fig. 4.12 indicates that important regions are better localized for the VGG-ImageNet compared to VGG-Face.

4.4.10 ResNet

We performed the experiments using ResNet50 pre-trained on ImageNet to examine the impact of different deep learning architectures. Our baseline configuration was used except for the fine-tune layer. In this experiment, the network was fine-tuned from the first layer. The results, as shown in Fig. 4.13, there is a small difference between VGG16 and ResNet50. While ResNet50 (0.516) shows better performance than VGG16 (0.504) for intensity estimation, VGG16 (0.609) shows better performance than ResNet50 (0.591) for occurrence detection.

4.5 Summary

The aim of this study is to investigate the key parameters for both AU occurrence and intensity estimation, and show the optimal configuration especially for across

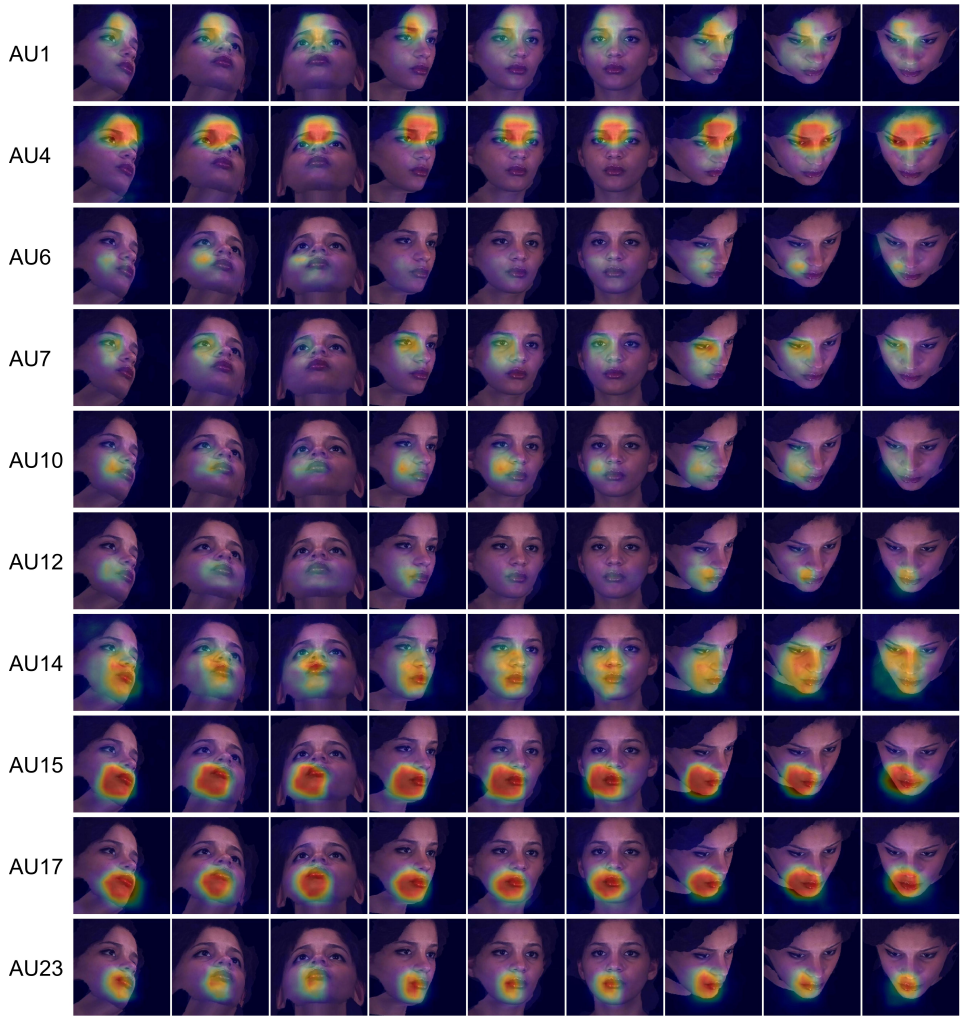

Figure 4.10: Occlusion sensitivity maps for each pose each AU. We trained models with our baseline configuration.

pose situation. To achieve the goal, we evaluated the combinations of different components and parameters.

Our findings range from the optimal pre-trained models (e.g., generic pre-training outperformed face-specific models optimized for recognition) to best practices in tuning optimizers. By utilizing all these insights, our architecture outperforms state-of-the-art performance on both tasks. We also show that our architecture performs well on unseen poses, and new domains.

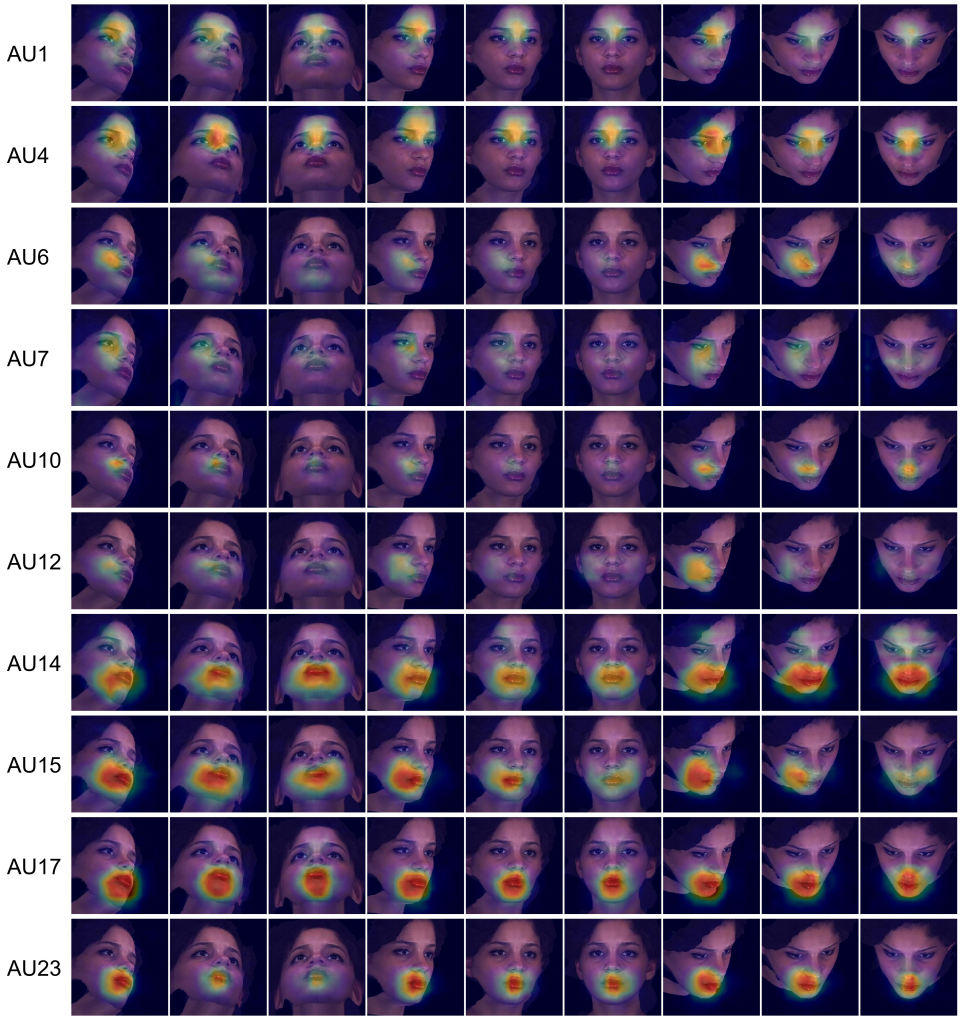

Figure 4.11: Occlusion sensitivity maps for each pose each AU. We trained models using images with eighteen poses (nine original poses and nine mirrored poses).

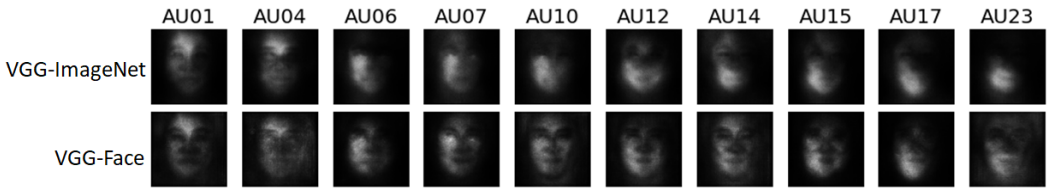

Figure 4.12: Saliency maps extracted using Vanilla Backpropagation.

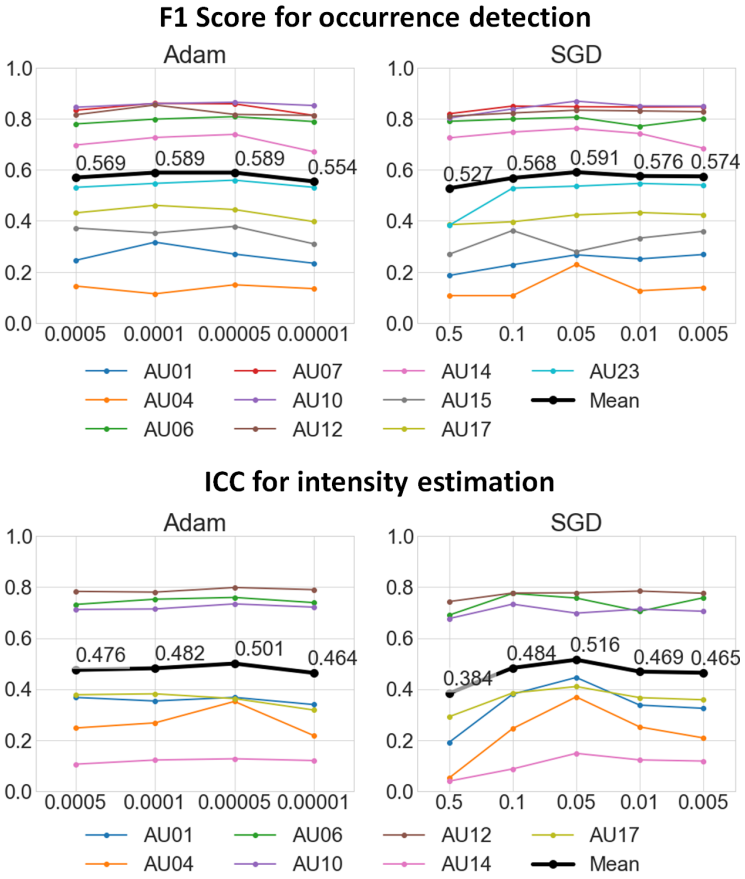

Figure 4.13: Effect of learning rates and choice of optimizers for ResNet50 on the FERA2017 Test partition.

Chapter 5

Conclusion

In this research, we investigate automated facial image analysis (AFA) across pose for three representative applications for AFA: face recognition, continuous authentication, and facial expression recognition. We proposed a different approach for each application according to the features of the applications.

First, we have addressed the problem of fully automatic face recognition across multiple views. The proposed approach builds 3D face models from frontal target face images, and uses them to generate synthetic target images to resemble the poses in query face images. By aligning synthetic target and query images by applying Procrustes Analysis, the proposed approach can leverage a wide range of well-studied matchers. A MLBP based face matcher is used by our approach, but we also show the good extensibility of our approach by replacing our MLBP based matcher with MKD-SRC and FaceVACS.

We have next proposed a new framework for continuous authentication. Since existing approaches for continuous authentication used hard biometric traits, they suffered from the low availability of the biometric traits. To tackle the problem, our framework primarily uses soft biometric traits (facial skin color and clothing color) to continuously authenticate the user. The system is robust with respect to user's posture in front of the workstation and it also has the capability for enrollment template update and relogin authentication. Our approach uses PCA-based face recognition, but we can easily replace it with more sophisticated matchers to make the system more robust. Experimental results demonstrate that the system is able to successfully authenticate the user continuously with high tolerance to the user's

posture.

Finally, we have addressed the design choices of deep-learning based facial expression recognition. More specifically, we systematically evaluated design choices in pre-training, feature alignment, model size selection, and optimizer details. By utilizing all the insights, we developed an architecture that exceeds state-of-the-art on FERA 2017. The architecture achieved a 3.5% increase in F_1 score for occurrence detection and a 5.8% increase in ICC for intensity estimation. To evaluate the generalizability of the architecture to unseen poses, we performed experiments across pose in FERA 2017. To evaluate the generalizability of the architecture across domains, we performed experiments in DISFA and the UNBC Pain Archive.

Our future work includes expanding the capabilities of the proposed frameworks to handle other unconstrained environments, such as illumination change, occlusion and aging. Another important issue is the limited dataset size for AFA. To tackle the problem, it is not realistic to collect enough real images for all real-world applications. Therefore, we believe that it is critical to leverage image generation technologies to tackle the problem.

Acknowledgements

I first wish to express my deepest gratitude to my supervisor, Prof. Kazuhiro Fukui. He is always supportive, and his advice is always concise, appropriate and easy to understand. His kind support was tremendously helpful to complete the thesis.

I also wish to express sincere appreciation to my thesis committee, Prof. Keisuke Kameyama, Prof. Yutaka Satoh, Prof. Hotaka Takizawa, and Prof. Itaru Kitahara for taking time to review my thesis and giving helpful comments.

I would also like to thank my co-authors: Prof. Anil K. Jain, Prof. Unsang Park, Prof. Hu Han, Prof. Jeffrey Cohn, Prof. László Jeni, and Prof. Itir Onal Ertugrul. While doing research with them in Michigan and Pittsburgh, I completed a major part of the research work.

Finally, I wish to express my thanks to my family to support and make the opportunity to pursue Ph.d degree. Especially under the covid-19 situation, their support was essential.

Bibliography

[1] FaceVACS Software Developer Kit Cognitec Systems GmbH.

[2] USF DARPA HumanID 3D Face Database. Courtesy of Professor Sudeep Sarkar, University of South Florida.

[3] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[4] A. Altinok and M. Turk. Temporal integration for continuous multimodal biometrics. *Workshop on Multimodal User Authentication*, page 131–137, 2003.

[5] M. Amirian, M. Kächele, G. Palm, and F. Schwenker. Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[6] S. R. Arashloo and J. Kittler. Energy normalization for pose-invariant face recognition based on MRF model image matching. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 33(6):1274–1280, 2011.

[7] A. B. Ashraf, S. Lucey, and T. Chen. Learning patch correspondences for improved viewpoint invariant face recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[8] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. V. Rohith. Fully automatic pose invariant face recognition via 3D pose normalization. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 937–944, 2011.

[9] A. Azzini and S. Marrara. Impostor users discovery using a multimodal biometric continuous authentication fuzzy system. *Lecture Notes in Artificial Intelligence*, 5178:371–378, 2008.

[10] A. Azzini, S. Marrara, R. Sassi, and F. Scotti. A fuzzy approach to multimodal biometric continuous authentication. *Fuzzy Optimal Decision Making*, 7:243–256, 2008.

[11] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2015.

[12] J. C. Batista, V. Albiero, O. R. P. Bellon, and L. Silva. AUMPNet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, 1997.

[14] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[15] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 25(9):1063–1074, 2003.

[16] C. Carrillo. Continuous biometric authentication for authorized aircraft personnel: A proposed design. *Master's thesis, Naval Postgraduate School*, 2003.

[17] C. D. Castillo and D. W. Jacobs. Wide-baseline stereo for face recognition with large pose variation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 537–544, 2011.

[18] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose invariant face recognition. *IEEE Transaction on Image Processing*, 16(7):1716–1725, 2007.

[19] W.-S. Chu, F. De la Torre, and J. F. Cohn. Learning facial action units with spatiotemporal cues and multi-label sampling. *Image and vision computing*, 81:1–14, 2019.

[20] J. F. Cohn and F. De la Torre. Automated face analysis for affective computing. In *The Oxford handbook of affective computing*, page 131. 2014.

[21] J. F. Cohn, I. O. Ertugrul, W.-S. Chu, J. M. Girard, L. A. Jeni, and Z. Hammal. Chapter 19 - affective facial computing: Generalizability across domains. In *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 407 – 441. Academic Press, 2019.

[22] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 24(5):603–619, 2002.

[23] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 25(5):564–577, 2003.

[24] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1548–1568, 2016.

[25] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology*, 7(3):1–42, 2016.

[26] L. Ding, X. Ding, and C. Fang. Continuous pose normalization for pose-robust face recognition. *IEEE Signal Processing Letters*, 19(11):721–724, 2012.

[27] I. O. Ertugrul, L. A. Jeni, and J. F. Cohn. FACSCaps: Pose-independent facial action coding with capsules. *IEEE Computer Vision and Pattern Recognition Workshops*, pages 2130–2139, 2018.

[28] Itir Onal Ertugrul, Jeffrey F. Cohn, L. A. Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):158–171, 2020.

[29] M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Analysis of partial least squares for pose-invariant face recognition. *IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2012.

[30] A. K. Jain P. Flynn and A. A. Ross. Handbook of biometrics. *Eds. New York: Springer*, 2007.

[31] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.

[32] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[33] A. Gudi, H. E. Tasli, T. M. Den Uyl, and A. Maroulis. Deep learning based FACS action unit occurrence and intensity estimation. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–5, 2015.

[34] H. Han and A. K. Jain. 3D face texture modeling from uncalibrated frontal and profile images. *IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2012.

[35] H. Han, S. Shan, X. Chen, and W. Gao. A comparative study on illumination preprocessing in face recognition. *Pattern Recognition*, 46(6):1691–1699, 2013.

[36] J. He, D. Li, B. Yang, S. Cao, B. Sun, and L. Yu. Multi view facial action unit detection based on CNN and BLSTM-RNN. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[37] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained

environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[38] G. Jaffre and P. Joly. Costume: A new feature for automatic video content indexing. *Adaptivity, Personalization and Fusion of Heteogeneous Information (RIAO)*, pages 314–325, 2004.

[39] A. K. Jain, S. C. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? *SPIE*, 5404:561–572, 2004.

[40] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. *Lecture Notes in Computer Science (LNCS)*, 3072:731—-738, 2004.

[41] L. A. Jeni, A. Lorincz, T. Nagy, Zs. Palotai, J. Sebok, Z. Szabo, and D. Takacs. 3D shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, 30(10):785 – 795, 2012.

[42] T. Kanade and A. Yamada. Multi-subregion based probabilistic approach toward pose-invariant face recognition. *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, pages 954–959, 2003.

[43] H.-B. Kang and M.-H. Ju. Multi-modal feature integration for secure authentication. *International Conference on Intelligent Computing*, page 1191–1200, 2006.

[44] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[45] A. Klosterman and G. Ganger. Secure continuous biometric-enhanced authentication. *Carnegie Mellon University, Tech. Rep. CMU-CS-00-134*, 2000.

[46] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision (IJCV)*, 83(2):178–194, 2009.

[47] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2009.

[48] G. Kwang, R. H. Yap, T. Sim, and R. Ramnath. A usability study of continuous biometrics authentication. *Lecture Notes in Computer Science (LNCS)*, 5558:828–837, 2009.

[49] A. Li, S. Chai, and W. Gao. Coupled bias-variance tradeoff for cross-pose face recognition. *IEEE Transaction on Image Processing*, 21(1):305–315, 2012.

[50] A. Li, S. Shan, X. Chen, and W. Gao. Cross-pose face recognition based on partial least squares. *Pattern Recognition Letters*, 32(15):1948–1955, 2011.

[51] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. *European Conference on Computer Vision (ECCV)*, 1:102–115, 2012.

[52] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv:1804.08348*, 2018.

[53] W. Li, F. Abtahi, Z. Zhu, and L. Yin. EAC-Net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(11):2583–2596, 2018.

[54] X. Li, S. Chen, and Q. Jin. Facial action units detection with multi-features and -AUs fusion. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[55] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 35(5):1193–1205, 2013.

[56] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. *IEEE International Conference on Image Processing (ICIP)*, 1:900–903, 2002.

[57] Chengjun Liu and Harry Wechsle. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

[58] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. *IEEE*

International Conference on Automatic Face & Gesture Recognition and Workshops (FG), pages 57–64, 2011.

[59] S. Lucey and T. Chen. A viewpoint invariant, sparsely registered, patch based, face verifier. *International Journal of Computer Vision (IJCV)*, 80(1):58–71, 2008.

[60] B. Martinez, M. F. Valster, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017.

[61] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA : A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[62] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *European Conference on Computer Vision (ECCV)*, 5305:504–513, 2008.

[63] F. Monrose and A. D. Rubin. Keystroke dynamics as biometrics for authentication. *Future Generation Computer Systems*, 16:351–359, 2000.

[64] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 24(7):971–987, 2002.

[65] U. Park, H. Chen, and A. K. Jain. 3D model-assisted face recognition in video. *Conference on Computer and Robot Vision*, pages 322–329, 2005.

[66] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 32(5):947–954, 2010.

[67] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision Conference*, 2015.

[68] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(10):1090–1104, 2000.

[69] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose invariant face recognition using 3D generic elastic models. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 33(10):1952–1961, 2011.

[70] S. J. D. Prince, J. H. Elder, J. Warrell, and F. M. Felisberti. Tied factor analysis for face recognition across large pose differences. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 30(6):970–984, 2008.

[71] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 35(6):1357–1369, 2013.

[72] S. Z. Li and A. K. Jain (eds.). Handbook of face recognition, second edition. *Springer*, 2011.

[73] M. S. Sarfraz and O. Hellwich. Probabilistic learning for fully automatic face recognition across pose. *Image and Vision Computing*, 28(5):744–753, 2010.

[74] A. Sharma, M. A. Haj, J. Choi, L. S. Davis, and D. W. Jacobs. Robust pose invariant face recognition using coupled latent space discriminant analysis. *Computer Vision and Image Understanding (CVIU)*, 116(11):1095–1110, 2012.

[75] T. Sim, S. Zhang, R. Janakiraman, and S. Kumar. Continuous verification using multimodal biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(4):687–700, 2007.

[76] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations(ICLR) Workshop*, 2014.

[77] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

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

[79] X. Suo, Y. Zhu, and G. Owen. Graphical passwords: A survey. *Annual Computer Security Applications Conference (ACSAC)*, 2005.

[80] S. Taheri, P. Turaga, and R. Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 306–313, 2011.

[81] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[82] C. Tang, W. Zheng, J. Yan, Q. Li, Y. Li, T. Zhang, and Z. Cui. View-independent facial action unit detection. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[83] Murat Taskiran, Nihan Kahraman, and Cigdem Eroglu Erdem. Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106, 2020.

[84] L. Teijeiro-Mosquera, J. L. Alba-Castro, and D. Gonzalez-Jimenez. Face recognition across pose with automatic estimation of pose parameters through AAM-based landmarking. *International Conference on Pattern Recognition (ICPR)*, pages 1339–1342, 2010.

[85] Z. Tősér, L. A. Jeni, A Lőrincz, and J. F. Cohn. Deep learning for facial action unit detection under large head poses. *European Conference on Computer Vision Workshop*, pages 359–371, 2016.

[86] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[87] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. FERA 2015-second facial expression recognition and analysis challenge. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8, 2015.

[88] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. FERA 2017 - addressing head pose in the

third facial expression recognition and analysis challenge. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[89] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–511–I–581, 2001.

[90] J. Wang, J. You, Q Li, and Y. Xu. Orthogonal discriminant vector for face recognition across pose. *Pattern Recognition*, 45(12):4069–4079, 2012.

[91] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv:1804.06655*, 2020.

[92] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *European Conference on Computer Vision (ECCV)*, pages 818–833, 2014.

[93] X. Zhang, L. Yin, J. F. Cohn, S Canavan, M. Reale, A. Horowitz, and J. M. Girard. BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[94] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016.

[95] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, pages 1–27, 2019.

[96] Song Zhou and Sheng Xiao. 3d face recognition: a survey. *Human-centric Computing and Information Sciences*, 8(35), 2018.

[97] Y. Zhou, J. Pi, and B. E. Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2017.

[98] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012.

List of publications

International journal

1. **K. Niinuma**, U. Park, and A. K. Jain, "Soft Biometric Traits for Continuous User Authentication", IEEE Transactions on Information Forensics and Security (TIFS), Vol. 5, No. 4, pp. 771-780, 2010.

2. T. Takai, A. Maki, **K. Niinuma**, and T. Matsuyama, "Difference sphere: An approach to near light source estimation", Computer Vision and Image Understanding, 113(9), 966-978, 2009.

International conference

1. **K. Niinuma**, L. A. Jeni, I. O. Ertugrul, and J. F. Cohn, "Synthetic Expressions are Better Than Real for Learning to Detect Facial Actions", IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-10, 2021.

2. **K. Niinuma**, L. A. Jeni, I. O. Ertugrul, and J. F. Cohn, "Unmasking the Devil in the Details: What Works for Deep Facial Action Coding?", British Machine Vision Conference (BMVC), pp. 1-12, 2019.

3. **K. Niinuma**, H. Han, and A. K. Jain, "Automatic Multi-view Face Recognition via 3D Model Based Pose Regularization", IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1-8, 2013.

4. **K. Niinuma** and A. K. Jain, "Continuous User Authentication Using Temporal Information", Proc. of SPIE Defense, Security, and Sensing: Biometric Technology for Human Identification VII, 76670L, pp. 1-10, 2010.