STUDY OF REAL-TIME FACIAL EXPRESSION RECOGNITION

ON NOISY IMAGES AND VIDEOS

by

Myung Hoon Suk

APPROVED BY SUPERVISORY COMMITTEE:

B. Prabhakaran, Chair

Nasser Kehtarnavaz

Kang Zhang

Carlos Busso

This dissertation is dedicated to

my parents and family.

STUDY OF REAL-TIME FACIAL EXPRESSION RECOGNITION ON NOISY IMAGES AND VIDEOS

by

MYUNG HOON SUK, BS, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

COMPUTER ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

December 2018

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. B. Prabhakaran, for his guidance, support and patience. Without his help, I would not have been able to accomplish the completion of my dissertation. I also would like to thank all of my committee members, Dr. Nasser Kehtarnavaz, Dr. Kang Zhang and Dr. Carlos Busso, for their support and valuable advice. I thank all of the wonderful and sincere alumni and present members of the Multimedia Systems and Networks lab for their assistance in my experiments, advice and good discussion during my PhD study.

I will never forget my friends who shared my time at The University of Texas at Dallas, Noun Choi, Dukjin Kim, Wooil Kim, Yohan Jin, Sangjin Hong, Jehun Jeon, Junwon Suh, Donghyun Kim, Hyunbum Kim, and Hyungjae Chang. And I would like to thank Dr. Richard Min for giving me his valuable advice while I was working with him as a TA.

I would also like to specially thank to Steven White, Mitch Butler and Ashok Ramadass in Image Vision Labs and Dr. Feng Yang in Topaz Labs.

Finally, I would like to thank my family standing behind me. They always supported me and encouraged me with their best wishes. I especially cannot express my gratitude enough to my parents, Kwangoh Suk and Philryang Ryu. I appreciate their love. I would like to thank my elder brother and his wife, Changhoon Suk and Shinjae Park. Also I am thankful to Hongtae Jung, Inran Cho, and Chunsil Lee.

Most of all, I would like to thank my wife, Minhee Jung, for her everlasting love and support. She was always there cheering me up and stood by me through the good times and bad. My two sons, Jayden and Caleb, are the source of my greatest joy and happiness.

August 2018

STUDY OF REAL-TIME FACIAL EXPRESSION RECOGNITION

ON NOISY IMAGES AND VIDEOS

Myung Hoon Suk, PhD
The University of Texas at Dallas, 2018

Supervising Professor: B. Prabhakaran, Chair

Automatic facial expression recognition (FER) and emotion recognition have aroused many researchers' interest in a variety of research fields because of an important role in human centered interfaces and the advent of cheap and powerful computer and video camera in the last decade. In addition, the emergence of the smartphones era has aroused considerable interest in the mobile application development in connection with facial expression and emotion recognition.

However, in spite of the enhanced hardware of recent smartphones, mobile applications for processing real-time video should always consider limited resources available in smartphones. The limited processing resources in smartphones still make it difficult to directly adopt the existing facial expression and emotion recognition system from desktops. Most studies for FER have been carried out and evaluated under restricted experimental environment. For instance, some approaches deal with only static images or work with video sequences manually pre-segmented (temporally) for each expression. However, the temporal segmentation of expressions is the most essential element in automatic FER systems as real world applications for real-time video. Also, the real world dataset for FER is different from most conventional datasets which are mainly collected in a limited experimental environment. It is hard to apply models made with datasets collected under lab environment to real world

application. The automatic FER should be capable of satisfying these various types of noisy datasets.

We address several problems for real-time FER on low-power smartphones. First, we presents a real-time FER effectively running on smartphones. The system employs a set of Support Vector Machines (SVM) for neutral expression and 6 basic emotions with 13D geometric facial features including temporal information. We evaluated the performance of the proposed system in terms of speed and accuracy on offline dataset and commercial off-the-shelf smartphones. Second, we present a real-time temporal video segmenting approach for automatic FER applicable in a smartphone. The proposed system uses a Finite State Machine (FSM) for segmenting real time video into temporal phases from neutral expression to the peak of an expression. The system performs FER with SVM on every apex state after automatic temporal segmentation, without any sampling time delay. Third, we present gender-driven ensemble models for FER on smartphones working with a context-sensitive multimedia content recommendation system. Based on the fact that male and female express an emotion with a distinct difference in the horizontal and vertical facial movements, we employ the ensemble model with three weak classifiers trained by gender-specific subsets and a general dataset of facial expression. In the system, users receive feedback by links to multimedia contents such as videos, photos and e-books regarding a current user's emotion. Last, we present an approach using CNN model for FER to accommodate noisy images and videos dataset in real world environment. We adopt FER2013 dataset for training CNN model. We show the CNN model is able to work very well for expression recognition even with real, noisy data that is not used for training.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Facial expressions as a result of motions of facial muscles are a primary means conveying human internal emotional states or intentions to others as a form of non-verbal communication. The study on facial expressions has been conducted in a long time by many scientists such as Darwin, Ekman, Friesen, Hager, Mehrabian, etc. Automatic facial expression recognition and emotion recognition have aroused many researchers' interest in a variety of research fields such as human-computer interaction, robotics, games, education and entertainment because of an important role in human centered interfaces and have been an active research topic particularly with the advent of cheaper and much powerful computer and video camera in the last decade.

At the beginning of the automatic facial expression recognition research, most studies were performed using cleaner dataset under well-controlled laboratory environments, whereas recent research has increasingly focused on more realistic or practical (less clean, noisy) dataset. One of the major factors driving this shift is the advent of smartphones, the emergence of new environments such as Virtual Reality(VR)/Mixed Reality(MR), and using various devices embedding a high resolution small camera.

In the past, such experiments including data acquisition were carried out in a limited space that could not get out of the fixed camera position, but now we can easily acquire image and video data with small, easy-to-carry, high-performance camera-equipped devices anytime, anywhere without any significant spatial constraints. The followings are examples of various environments in which image/video data can be obtained:

- Smartphone
- Indoor / outdoor security camera

- Vehicle dash cam
- Activity camera like GoPro
- Wearable camera
- Drone with camera
- Robot vision
- 2D / 3D camera device for Game and Virtual Reality/Mixed Reality

Moreover, it was a tremendous effort and time to bring the subjects into a limited laboratory environment and collect facial expression data from them, and the data were mostly controlled by the subjects who recognized the experimental intention. However, as in the environments listed above, now it is possible to easily acquire innumerable natural facial expressions in images and videos, which are not intended to be intentionally, by information infrastructure like high speed internet platform and mobile platform, where they are easily acquired and shared inexpensively. Therefore, it is necessary to make more effort and development for research on facial expression recognition in noisy image / video acquired in a diverse environment, and a lot of these research are going on.

In addition, the emergence of the smartphones era has aroused considerable interest in the mobile application development in connection with facial expression and emotion recognition.

Smartphones such as *Apple iPhone series* and *Samsung Galaxy series* have become very popular consumer devices worldwide in recent years. Cameras mounted on smartphones are the most common features. Beyond the basic feature of the camera for photos and videos, advanced computer vision technology applicable to camera-equipped smartphones has the potential to play an important role in the development of emerging applications or innovative changes in the user interface of smartphones.

To use the *Samsung Galaxy* Android phone for an example, the security feature such as '*Face Unlock*' by face recognition is already in use on Android phone, and it has the screen control features such as '*Smart Stay*' and '*Smart Rotation*' by using face detection and eye tracking. These kinds of applications have become feasible because of not only the advances in computer vision technologies but also improved hardware performance on smartphones. People preferred using a remote high performance server for heavy tasks relevant to computer vision instead of processing them on a mobile device itself even a couple of years ago. But now smartphones with the swift advances in processing power and memory has brought computer vision tasks in real time within the bounds of possibility.

Furthermore, smartphones can be utilized as an useful communication channel in affective computing. An understanding of users' emotional states that is an area of affective computing can be exploited as a context-sensitive user interface on smartphones for context-sensitive multimedia applications. Therefore, automatic facial expression and emotion recognition on smartphones can play a key role in building a context-sensitive multimedia application in affective computing.

1.2 Problem Statement

As studies on utilization of emotions in other fields, we have no doubt that natural human-centered interaction through users' emotions would be a good help for smartphones to get advanced features. However, in spite of the enhanced hardware of recent smartphone devices, mobile applications for processing real-time video should always consider limited resources available in smartphone devices. The limited processing resources in smartphones still make it difficult to directly adopt the existing facial expression and emotion recognition system from desktops. For example, the real-time facial expression recognition system running on a smartphone without using the remote servers with high performance for heavy tasks should tackle issues with low computation power causing slow frames per second (FPS) and low

resolution video frames as well as characteristic of handheld mobile devices causing non-stationary camera and varying illumination condition.

Although a variety of research approaches for automatic facial expression recognition have been proposed, most studies have been carried out and evaluated under restricted experimental environment. For instance, some approaches for facial expression recognition deal with only static images or work with video sequences manually pre-segmented (temporally) for each expression. However, the temporal segmentation of expressions is the most basic and essential element in automatic facial expression recognition systems as real world applications for real-time video.

In addition, the issues from recognition tasks of facial expressions have been addressed from different approaches. From the universality hypothesis of facial expressions that humans communicate six basic emotions with the same facial movements regardless of culture and region (Darwin, 1872), most approaches have generally dealt with facial expression irrespective of any prior knowledge on human face such as gender, age, culture, or race. Therefore, not surprisingly among many approaches for facial expression recognition, there are few researches dealing with gender-specific facial expression dataset. However, we are motivated that facial expression recognition can be improved with prior knowledge on subject.

Lastly, since most conventional facial expression recognition dataset are mainly collected in a limited experimental environment, there is a difference from the data under the actual environment. The data actually collected can be cropped faces in low resolution, be recorded under bad lighting conditions, or have scattered backgrounds. In addition, because the facial expressions of subjects in the experimental environment can be unconsciously reflected by the intention of the experiment, it may be different from the natural facial expression. While it is mainly limited to the head-on frontal view in the experimental environment, emotional expression in the real environment can be reflected with the head pose of various angles or partial occlusion of face by hand gesture depending on the intensity of emotion. There is a

need for an automatic facial expression recognition system capable of satisfying these various types of noisy dataset.

1.3 Proposed Approaches

1.3.1 Experiences with Real-time Facial Expression Recognition

In order to tackle the problems in applying the existing approaches for real-time facial expression recognition to low-power smartphones, we propose a simple and effective approach for real-time video based facial expression recognition running on smartphones. The proposed system works on recognizing users' emotion through the front-facing camera mounted on a smartphone. Because the proposed system deals with all the processes in the smartphone device, there is no communication delay to remote servers. Therefore, the real-time mobile emotion recognition system is expected to be a good start for the emergence of a variety of mobile services and applications.

1.3.2 Temporal Segmentation of Video Sequences

We present an efficient approach for real time temporal video segmentation for facial expression recognition on smartphones. The proposed system uses a Finite State Machine (FSM) to temporally segment continuous video in real time into sequences starting with a neutral expression and ending with a peak expression. The FSM based approach employs Lucas-Kanade's optical flow vector (Lucas and Kanade, 1981) based scores for state transitions in a manner that is adaptive to the varying speeds of facial expressions. We have tested our system using the *Samsung Galaxy S3* running Android 4.3 *Jelly Bean* with a frame rate of average 3.7 *fps* (video resolution: 352×288).

In comparison with HMM based approaches, the proposed approach has an advantage. With respect to temporal segmentation, being a dynamic classifier, although HMM can

handle time series data, HMMs cannot directly do temporal segmentation of video sequences. Therefore, HMM can employ a sliding window for N size samples and then the sequence of N size samples for a certain period of time is classified as one of 6 emotions yielding highest probability. For example, the Multi-level HMMs designed for automatic segmentation and recognition of emotions in (Cohen et al., 2003) needs N frames samples in the lower layer HMM. On the other hand, our proposed system using the FSM approach does not require any sliding windows or sampling time. In other words, it runs in real time without any sampling time delay. Experimental results show that it is an appropriate way for real-time mobile applications in terms of speed performance.

For facial expression recognition, the proposed system employs dynamic features such as displacements between neutral and apex states, not dynamic features on a frame-by-frame basis. This makes the system less sensitive frame-to-frame variations.

1.3.3 Gender-driven Facial Expression Recognition on Smartphones for Multimedia Content Recommendation System

In the proposed system employing multimodal approach, the idea using prior knowledge for improving facial expression recognition performance comes from experiential knowledge that people are more likely to recognize expressions of familiar faces much better and more quickly than those of unfamiliar faces. As an example, the evidence about difference of facial expression between gender and age is supported by experimental result of the paper (Houstis and Kiliaridis, 2009). Particularly in the paper, authors found that males had a pronounced vertical movement in the posed smile and lip pucker compared to females. On the other hand, females had a greater horizontal component in the posed smile. From the result about the gender difference in facial expression, we are pretty motivated to improve the facial expression recognition system. To put it concretely, we can recognize facial expressions by classifiers trained with gender-specific dataset depending on previously determined gender

information. Likewise, dataset for a specific age group such as the elderly or children can be used for training specific classifiers. In addition, it is possible to use combinations of genders and ages. (e.g., boys, girls, elderly women, young men, etc.)

For this approach, we present a context-sensitive multimedia content recommendation system that has the capability to understand users' emotional states along with priori knowledge such as gender by gender recognition and facial expression recognition on a smartphone and to recommend multimedia contents associated with users' emotion.

1.3.4 Facial Expression Recognition on Noisy Images and Videos dataset

To create a model that can accommodate noisy images and videos dataset captured under a variety of environments, we select a open resource dataset, FER2013 dataset provided by the Facial Expression Recognition (FER) Challenge 2013 and composed of natural facial expression samples with 48×48 small size image. We create a small and basic CNN model suitable to a small dataset for CK+, while deeper convolutional neural network is built for a large dataset with FER2013. To compensate for the disadvantages of the small amount of data in CK+, we employ data augmentation increasing the amount of training data. On the other hand, In the case of a SVM classifier with 13D geometric features introduced in Chapter 3, it is difficult to apply data augmentation because the 13D feature itself uses normalized data. The proposed CNN models applied to each dataset (CK+ and FER2013) have good accuracy performance, even if these are baseline CNN models. Therefore, it is expected that if the current insufficient amount of dataset used under noisy environment like VR/MR are helped by data augmentation in CNN model, the system can have better performance.

1.4 Contributions of the Dissertation

1.4.1 A Efficient Framework for Real-time Facial Expression Recognition without Temporal Segmentation

The contributions of this work are highlighted as follows:

- Investigation of good feature selection suitable to carry out facial expression recognition on smartphones.
  - Comparison of the performance of 13D geometric features to 6D geometric features as a baseline from (Houstis and Kiliaridis, 2009) in order to verify the enhancement of the proposed features.
  - Comparisons of offline accuracy performance between geometric features and LBP-based appearance features on the extended Cohn-Kanade datasets (CK+).
  - Comparisons of online accuracy performance between geometric features and LBP-based appearance features on commercial off-the-shelf smartphones.
  - Comparisons of online speed performance between geometric features and appearance features on commercial off-the-shelf smartphones.
  - Evaluation of proposed system with CK+ datasets on smartphones.
- An efficient framework for real-time facial expression recognition on smartphones using a set of SVMs for neutral and facial expressions (anger, disgust, fear, happiness, sadness and disgust).
  - The use of neutral expression frame detection for distinguishing expression frames with neutral or slight expression frames.
  - The use of dynamic information of geometric features with displacement between neutral and expression frames. 13D geometric features outperform 6D base features with the increase of 4.8% in terms of recognition rate.

- Implementation of mobile app for Android phones (Suk and Prabhakaran, 2014).

We have implemented a mobile app which is available for download (See the link in the Section 3.3.8) and tested on several Android phones for evaluation. In particular, a *Samsung Galaxy S3* running Android 4.3 *Jelly Bean* (We observe that latest *OpenCV* 2.4.9 is most compatible with *Samsung Galaxy S3* before now.) at a frame rate of average 3.23 *fps* with 640×480 and at average 3.64 *fps* with 352×288. Although the proposed system performs roughly 3 frames per second (FPS), it is still possible to achieve the goal (of facial emotion recognition) with a high degree of accuracy as normal expressions usually last between 0.5 and 4 seconds.

1.4.2 Temporal Segmentation of Video Sequences on Smartphones

The main contribution of our work is real-time, adaptive segmentation of a continuous video of facial expressions into sequences of individual expressions on smartphones. This real-time segmentation is adaptive to the varying speeds with which facial expressions are made. Our approach using FSM is able to handle low frames per second such as 3 *fps* due to low processing power of smartphone devices along with high degree of accuracy of facial emotion recognition.

1.4.3 Gender-driven Facial Expression Recognition on Smartphones for Multimedia Content Recommendation System

The main contributions in this approach include 1) analyzing difference of facial expression between gender on Cohn-Kanade database, 2) applying and evaluating a priori knowledge such as gender for facial expression recognition by building ensemble models from gender-driven weak models separately trained from male, female, and mixed gender dataset, and 3) implementing a context-sensitive multimedia content recommendation system interacting with users of real-time smartphone application.

1.4.4 Facial Expression Recognition on Noisy Images and Videos dataset

The contributions in this approach include 1) building CNN models appropriately for the respective amount of datasets such as small amount of Cohn-Kanade dataset and large amount of FER2013 dataset 2) evaluating the CNN models cross dataset. For small amount of CK+ dataset, the proposed CNN with data augmentation outperforms SVM built with geometric features.

1.5 Dissertation Objective

Automatic facial expression recognition and emotion recognition have aroused in a variety of research field because of an important role in human centered interfaces. The objective of the dissertation is to develop the effective solutions and systems to solve the challenges and issues for recognizing natural facial expressions under a variety of noisy environments.

- How do we make real-time facial expression recognition efficient and reliable to low-power smartphones in terms of speed and accuracy?

- How do we have an efficient automatic facial expression recognition by temporally segmenting continuous video in real time?

- How do we improve facial expression recognition performance by prior knowledge such as gender?

- How do we build a model handling noisy dataset in more realistic environments?

The dissertation objectives are shown in the Figure 1.1

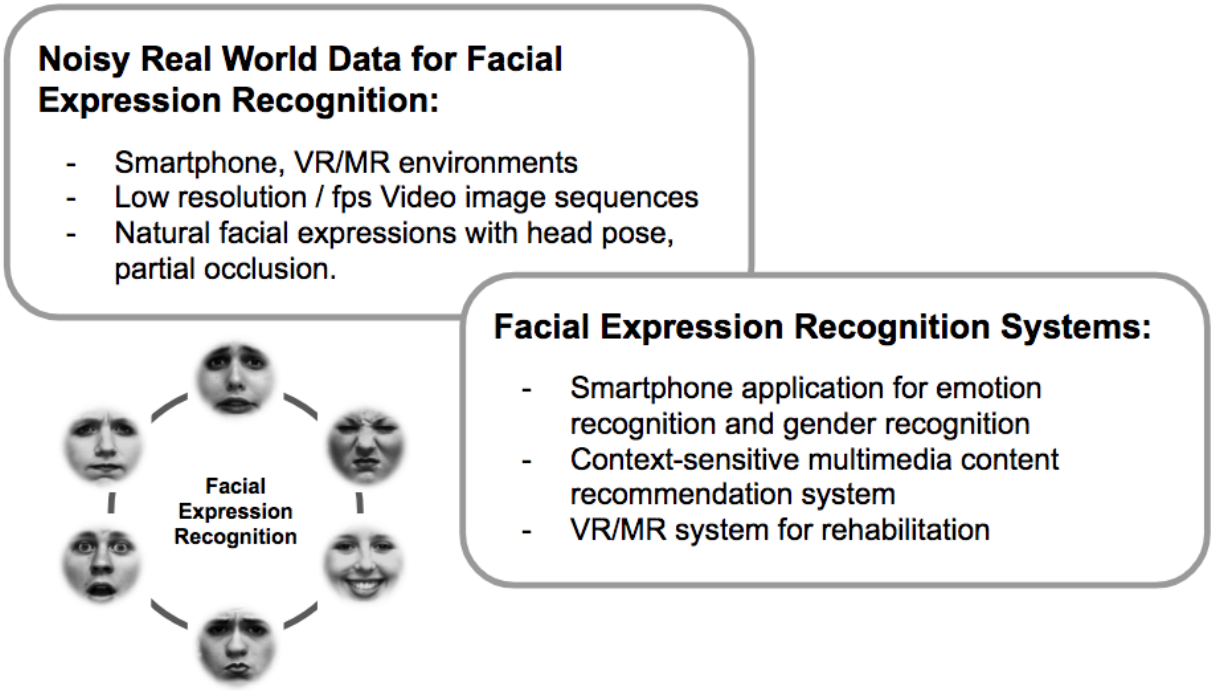


Figure 1.1: The overview of dissertation objectives.

1.6 Organization of the Dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents previous works related to this dissertation. Chapter 3 shows a simple and effective approach for real-time facial expression recognition system running on smartphones. In order to make the efficient framework for smartphones, the investigation of good feature selection are presented and the evaluation of the proposed method on commercial off-the-shelf smartphones. Chapter 4 details the FSM approach for real-time temporal segmentation of video sequences on smartphones. In Chapter 5, we propose an approach for facial expression recognition using multimodal information such as gender. The proposed gender-driven facial expression recognition system is integrated with a context-sensitive multimedia content recommendation system. In Chapter 6, we presents alternative approach using CNN model in order to

handle noisy dataset with the extended experiments. The Chapter 7 concludes the dissertation, and finally future work is discussed in Chapter 8.

CHAPTER 2

RELATED WORK

Since facial expressions convey the human's internal emotional states and play an essential part in human face-to-face communication as a form of non-verbal communication, many studies on facial expression analysis have been carried out with great interest for a long time. Furthermore, automatic facial expression analysis and recognition has received a lot of attention for decades. First of all, Paul Ekman and his colleagues' significant work about the facial expression in the 1970s became the foundation of the existing automatic facial expression recognition system and related research fields (Ekman, 1994). Paul Ekman *et al.* have found six universal emotions (anger, disgust, fear, happiness, sadness, and surprise), and developed Facial Action Coding System (FACS) for categorizing human facial expressions by describing the movements of individual facial muscles with Action Units (AUs) (Ekman, 1994; Ekman and Friesen, 1978).

Facial representation or facial feature is one of important factors in successfully designing facial expression recognition systems because the classifiers for facial expressions in the system are interrelated to facial features. Facial features to be represented are mostly divided into 2 types of groups such as geometric features and appearance features extracted from targeting face.

First, geometric features are relevant to measurement, shape, and locations of facial components such as eyes, eye brows, nose, and mouth. For example, Active Shape Models (ASMs) originally developed by T. Cootes et al. (Cootes et al., 1995) are one of the most popular and robust statistical model-based algorithms to fit landmarks on facial component. ASMs are successfully used to extract geometrical features by measuring lengths and angles of the markers in landmarks fitted on the face image so that the facial features are used as input data for the classifier of facial expression recognition.

Second, appearance features are changes of facial texture or appearance in the local areas (e.g., wrinkles and furrows) or the whole face. Gabor wavelets representation and Local Binary Patterns (LBP) are typical approaches for appearance features. Additionally the Active Appearance Model (AAM) is a well-known computer vision algorithm for hybrid features of shape and appearance (Edwards et al., 1998).

In fact, both geometric and appearance feature based facial representations are most subject to the reliable facial feature detection. Therefore, our proposed system in Chapter 5 employs ASMs for reliable facial feature detection, and we evaluated the performance of geometric feature based and appearance feature based approaches in terms of the implementation of real-time mobile application on commercial off-the-shelf smartphones.

There are many classifiers to be applied for facial expression recognition. For video-based facial expression recognition, classifiers are divided into two large groups: spatial and spatio-temporal approaches (Fasel and Luettin, 2003). Spatial approaches that have been extensively used in many existing facial expression recognition systems are based on frame-by-frame expression detection. The frame-by-frame approach using still images or only single frames in video sequences does not care about motion related information and temporal information. Therefore the classification of facial expressions is processed in accordance with spatial information in a single frame. For example, classifiers such as Neural Network (NN) (Lisetti and Rumelhart, 1998; Padgett and Cottrell, 1996), Bayesian Network (BN) (Cohen et al., 2003), rule-based classifiers (Pantic and Rothkrantz, 2000), Support Vector Machine (SVM) (Bartlett et al., 2001; Littlewort et al., 2002) led to a good success as this approach for facial expression recognition. On the other hand, spatio-temporal approaches are not simple, but usually results in better performance on video sequences than spatial approaches without temporal information. One of the most popular classifiers for the spatio-temporal approach is Hidden Markov Models (HMM) that have been well applied to facial expression recognition in (Cohen et al., 2003; Bartlett et al., 2001; Lien et al., 2000).

Although most studies of facial expression recognition generally have focused on high performance in terms of accuracy rate of recognition, performance issues on the mobile platform have been not too considered. Accordingly our study shows an approach suitable to mobile platform in Chapter 3.

To the best of our knowledge, our real-time video based facial expression recognition system is a rare work on mobile platform. An eBook reader application as a use-case scenario was introduced in (Anand et al., 2012). Although a user controls it with facial expressions as a natural interaction, some Facial Action Units limited to eyes (not various facial expressions) are used as facial gestures. For facial expression system on smartphones, authors employed AAM with a Difference Of Gaussian (DOG) to fix illumination variation problems in (Jo et al., 2011). However, they used only four classes (sadness, surprise, neutral and happiness) for experimental results.

There have been several works for segmenting a facial action into its temporal phases of neutral, onset, apex, and offset. Valstar and Pantic presented a hybrid SVM-HMM classifier for segmenting into temporal phases of Action Units (AUs) (Valstar and Pantic, 2007). The HMM consists of four states (neutral, onset, apex, and offset), and SVM is used as emission probabilities of HMM. In (Pantic and Patras, 2006), the authors performed automatic segmentation of video sequence using a neutral-expressive-neutral sequential facial expression model by finding the presence or absence of facial activity with AU recognizer. In (Valstar and Pantic, 2006), they showed temporal analysis of AUs and automated recognition of facial AUs.

Most approaches focus on making good performance in terms of accuracy rate of recognition without regard to speed issue. However, if the system is needed to run on mobile platform with lower computation power compared to high-performance computers, approaches should be concerned about speed performance. For example, calling multiple classifiers for AUs at the same time or calling a classifier such as NN, SVM, or HMM every frame would impose a burden on real-time video processing on mobile system.

In (Cohen et al., 2003), authors proposed automatic segmentation and recognition of emotions using multi-level HMM. Using un-segmented continuous video, they showed the multi-level HMM performed at the similar recognition rate compared to the emotion-specific HMM. (Cohen et al., 2003) found that although the dynamic classifiers such as HMM are more suited to person dependent expressions, the expression classification accuracy is lower compared to static classifiers. On the other hand, static classifiers are easier to train and implement, but perform poorly when used with frames not at the peak of an expression in video sequences.

Milborrow2008The results of (Cohen et al., 2003) motivated us to design the proposed system in Chapter 4 that (a) automatically segments sequences of expressions from the video stream, and uses dynamic features for facial expression recognition; (b) instead of using HMM (or other spatio-temporal classifiers) static classifiers such as SVM are used; (c) should be practicable solution on smartphones with low computational power. We show that the proposed system can lead to lower sensitivity to temporal patterns of different people's expressions, as well as easier training and implementation.

Facial expressions are mostly treated with six basic emotions regardless of culture and region under the universality hypothesis of facial expressions. However, O. Houstis and S. Kiliaridis in (Houstis and Kiliaridis, 2009) supports the evidence about the difference of facial expressions between different gender or age groups. Therefore, prior knowledge can help to make emotion recognition system enhanced. I. Bisio et al. (Bisio et al., 2013) tried gender-driven emotion recognition through speech signals and showed that a priori knowledge of speaker's gender increases the accuracy of emotion recognition.

The motivation is the use of multimodal information. There are bimodal or multimodal approaches for emotion recognition systems. By bimodal approaches, audio-visual data are used at feature-level for recognizing six emotions in (Huang et al., 1998) and (Chen and Huang, 2000). In (Busso et al., 2004), the authors show the result the complementary rela-

tions of the two modalities such as facial expression and speech and using the fused modalities can improve the performance and the robustness of the emotion recognition system.

We present some recent works related to content recommendations system and other similar systems using facial expressions, emotions, or facial components. I. Arapakis et al in (Arapakis et al., 2009) model users' affective responses with facial expressions and other physiological signals in order to predict topical relevance without explicit user judgments. They mentioned that low-level features such as motion units perform better rather than high-level information such as emotions by classifiers inefficiently modelled. R. Valenti et al in (Valenti et al., 2010) present a visual creativity tool to generate a combination of sounds changing in real time by automatically recognizing facial expressions and tracking facial muscle movements. In (Zhao et al., 2010), Zhao et al discuss a concept of a domain specific music recommendation system based on users' sleep quality, and they present an EEG-based approach for sleep quality measurement. Zhao et al in (Zhao et al., 2011) propose an approach for indexing and recommending videos based on affective analysis of viewers. For building facial expression classifier, they use compositional Haar-like features along with hidden conditional random fields (HCRFs). In our previous work (Mariappan et al., 2012), we presented FaceFetch, a user emotion driven multimedia content recommendation system that recommends multimedia related to users' current emotional state, but the system could not work with fully automatic interaction (the prototype system required a photo taken by clicking a button when the user had a facial expression.) and needs help of the remote server to do the task for facial expression recognition in a photo that the user sent. On the other hand, the system proposed in this paper has the ability to run automatic facial expression recognition in real-time video sequences on smartphones themselves without any remote server and manual interaction.

CHAPTER 3

EXPERIENCES WITH REAL-TIME FACIAL EXPRESSION RECOGNITION IN SMART PHONES[1]

3.1 Proposed System Overview

In order to make real-time facial expression recognition efficient and reliable to low-power smartphones, we consider performances in terms of speed and accuracy. The proposed system employs temporal information for facial features in anticipation of higher accuracy of spatio-temporal features than spatial and static features. For obtaining such dynamic features, the reference point such as a starting video frame including a neutral expression should be needed in video. In other words, the preceding process such as the temporal segmentation for facial expressions in real-time video sequences is additionally required for the starting and peak frames of a facial expression, and facial features with temporal information between these frames are taken. However complex temporal segmentation in video sequences can be costly on mobile devices with low computation power that should be optimized in terms of speed performance in order to run the application smoothly on. For addressing this issue, the proposed system adopts a simple and reliable approach to find both a neutral expression and non-neutral expressions. The system checks the presence of neutral expressions at every frame, but it does not require accurately to identify the peak of expressions in non-neutral expression frames of video sequences. Again, since the proposed approach distinguishes only neutral expressions from non-neutral expressions including both transitional states and apex states, it would lessen the computation load of temporal segmentation task. (Whereas the strict temporal segmentation of a facial expression motion includes at least four temporal

[1]© 2014 IEEE. Reprinted, with permission, from Myunghoon Suk, B. Prabhakaran, "Real-Time Mobile Facial Expression Recognition System - A Case Study," Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on , pp.132,137, June 2014.

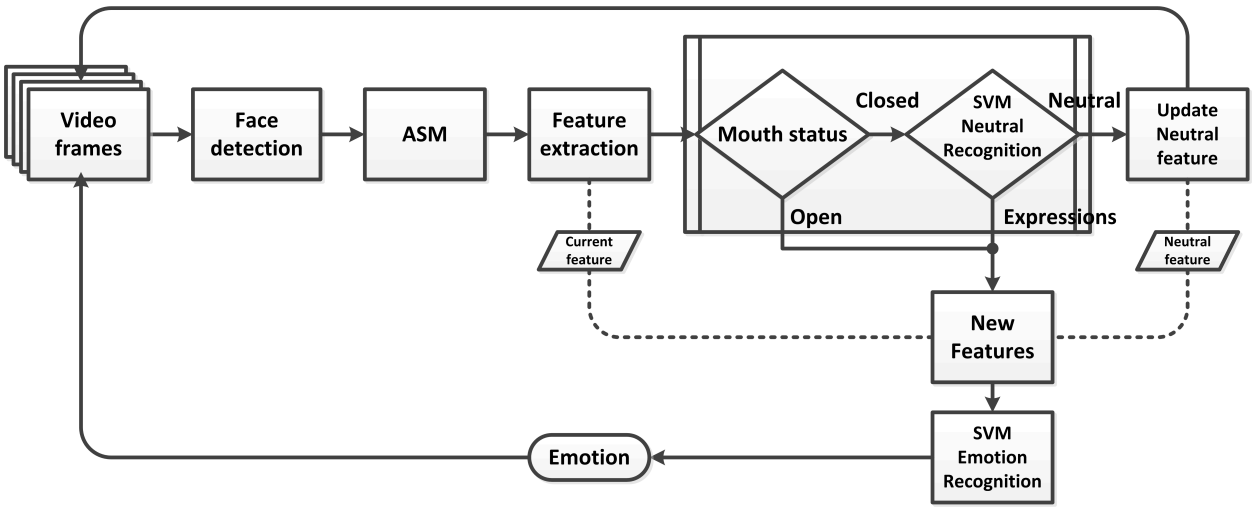


Figure 3.1: The proposed system architecture.

states such as neutral, onset, offset and peak in real-time video sequences. Therefore, it would be a more difficult and time-consuming process.)

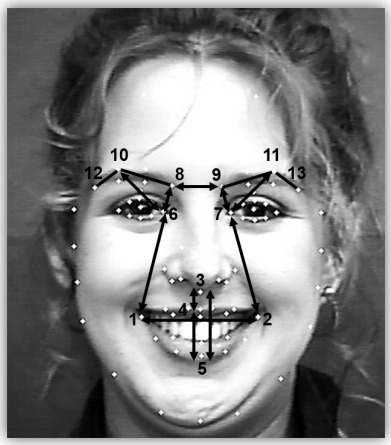
Without saving a sequence of features over a window of video frames or in a temporal video segmentation as time series dataset for classifiers such as HMMs, the dynamic features generated from a pair of neutral and non-neutral expression frames are fed to SVM classifiers. Therefore, the proposed approach makes the system effectively run in real time at a moderate level while avoiding the delay time in the case of sliding window or frequent processing time for frame-by-frame classification without temporal segmentation. For example, even though the system runs with a low frame rate of about 3.0 *fps*, it seems like that it usually does not miss peak frames in practical experiments because most real expressions reach a peak expression within a second and the facial expression on peak state is held for a moment about 1-2 seconds. Therefore the system is able to catch at least a couple of non-neutral frames during a period between starting and ending frames of a facial expression. After finding neutral expression and subsequent non-neutral expressions, the system creates new spatio-temporal facial features, that is, dynamic features with the displacement between the neutral expression feature and current non-neutral expression feature. Comparing to features

with the displacement of frame-to-frame basis, the features with longer gap of time affect the system less sensitive to frame-to-frame variations.

The system architecture for a real-time facial expression recognition on smartphones is shown in Figure 3.1. The proposed system generally follows main steps of ordinary facial expression recognition systems such as face detection, feature extraction, and facial expression classification. As a part of ASM stage followed by the face detection, the face alignment is an essential step to assure more accurate facial features. After ASM stage, coordinates of the markers on the face template are used to generate the static facial features in a current frame of video sequences. In the next step after the feature extraction in Figure 3.1, we employ both the SVM model and mouth status for detecting neutral expressions in a frame. If the current frame includes a neutral expression, the current features are saved as neutral features to create dynamic features with non-neutral expression features later. If it is a non-neutral expression in the current frame, new dynamic features are generated by displacement between the saved neutral features and current features. Then the dynamic features are fed into SVM models for facial expression recognition and finally the emotion recognition result from SVM classifiers are returned as one of 6 emotion classes.

3.2 Detailed Features and Facial Expression Recognition used in the Proposed System

The first step of the system is to grab a frame of video stream on the smartphone. In the current frame, the face is found by face detection using haar-cascade classifier in OpenCV. Although the next step is to extract features from the detected face, the ASM module is required prior to feature extraction because it is a prerequisite to both geometric features and appearance features for face alignment. Also the status of mouth in the neutral expression detection module are based on ASM landmarks. In the following sections, we explain the details of 13D geometric features and LBP appearance features for the system. In Section



	Markers	Descriptions
P1	LCO	Mouth left point
P2	RCO	Mouth right point
P3	Sn	Lower nose
P4	Ulip	Mouth upper point
P5	Llip	Mouth lower point
P6	LC	Left eye inner point
P7	RC	Right eye inner point
P8	LIEB	Left eyebrow inner point
P9	RIEB	Right eyebrow inner point
P10	LUEB	Left eyebrow upper middle point
P11	RUEB	Right eyebrow upper middle point
P12	LOEB	Left eyebrow outer point
P13	ROEB	Right eyebrow outer point

	Features
F1	Distance between P1 and P2
F2	Distance between P1 and P6
F3	Distance between P2 and P7
F4	Distance between P4 and P5
F5	Distance between P3 and P4
F6	Distance between P3 and P5
F7	Distance between P8 and P9
F8	Distance between P6 and P8
F9	Distance between P7 and P9
F10	Distance between P6 and P10
F11	Distance between P7 and P11
F12	Inter-angle of P8, P10 and P12
F13	Inter-angle of P9, P11 and P13

Figure 3.2: Example of landmarks by STASM (Milborrow and Nicolls, 2008), and 13 geometrical facial features by distances and angles among 13 points on one sample of CK+ dataset

3.3, we evaluate and compare the performance of approaches for two different types of feature set so that we choose more suitable features to mobile platform.

3.2.1 Geometric Feature Extraction

STASM (Milborrow and Nicolls, 2008) provides the face detection and ASM implementation where the 77 facial landmarks are located on the detected face. Based on x, y coordinates of landmarks from ASM, 13 high-level facial shape features are generated and normalized in the proposed system. 77 Landmarks (white dots) are shown on the face in the most left figure, 13 points are listed in the middle table, and 13D features are described in the most right table in Figure 3.2. The first 7 points (P1-P7) and 6 features (F1-F6) are the same points and distances as those used in (Houstis and Kiliaridis, 2009). However, the other 6 points (P8-P13) and 7 features (F7-F13) are new additional ones for robustness of the proposed features.

From the proposed system architecture in Figure 3.1, if the detected face has a neutral expression, the 13 facial shape features are saved as neutral features which is used as the reference features for calculating displacement from non-neutral expression features after-

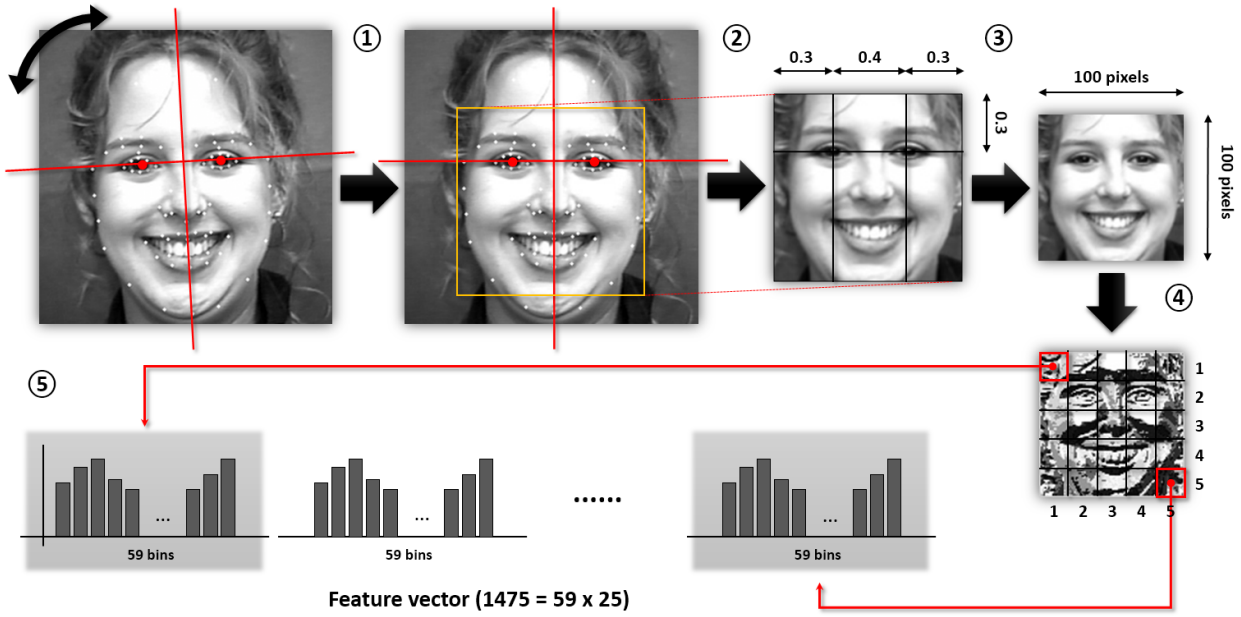


Figure 3.3: Procedure for the LBP features extraction: 1) Rotate, 2) Crop, 3) Scale, 4) LBP and 5) 1475 features.

ward. Otherwise, if it is a non-neutral expression, new dynamic features are created by saved neutral features and current facial features.

3.2.2 Appearance Feature Extraction with Local Binary Patterns (LBP)

For the appearance features representing facial expressions, we employ Local Binary Patterns (LBP). Since LBP was originally introduced in (Ojala et al., 1994), LBP features have been used in texture analysis as well as facial image analysis (Feng et al., 2004; Shan et al., 2009). LBP as good features works well for texture description. For examples, the features are robust to illumination conditions. Most of all, the most important reason why we choose LBP, not other descriptors such as the Histogram of Oriented Gradients (HOG) is for the computational simplicity. The LBP features are extracted by a sequence of steps (See Figure 3.3). Before applying LBP for a face image, the face image should be aligned by rotating, cropping, and scaling. For this pre-process, the positions of left and right eyes tracked by

ASM are used as reference points for rotating and cropping a face image. The 30% of the image size next to the eyes in horizontal and vertical direction are kept when the face image is cropped. Therefore, a final face image is aligned at the eyes with same fixed width and height (100 pixels×100 pixels size). The 100 × 100 face image is converted by the LBP operator thresholding a neighborhood of 8 pixels with a center pixel and labeling the center pixel as number. After labeling a face image into 256 numbers, the labeled face image is divided into 5 × 5 sub-regions. Each sub-region (20 pixels×20 pixels) of the labeled face image is represented by a histogram with 59 bins (59-label LBP (Shan et al., 2009)) by accumulating some patterns into a single bin, instead of 256 bins. Finally the face image is represented by 1475 (59 × 25) length features concatenated by 25 histograms with 59-bins.

In contrast to the 13D geometric features, the LBP features are not converted to dynamic features with the displacement between neutral frame and non-neutral frame in the proposed system. Therefore, LBP features as static features are used in training and testing states.

Additionally, 1475D LBP features can be reduced to 46D features by using feature selection method with F-score in (Chen and Lin, 2006) for better performance in terms of accuracy.

3.2.3 Neutral and Expressions Classification

Basically, a set of SVMs are used for neutral expression and 6 facial expressions classification. SVMs have been already well-known classifiers for pattern recognition tasks such as face recognition and facial expression recognition. The open source library called LIBSVM (A Library for Support Vector Machines) (Chang and Lin, 2011) is used to develop the mobile application for this system.

First, for the neutral expression detection, we check the status of the mouth by one of the 13D features (F4 in Figure 3.2) on detected face every frame. If the mouth is open or a feature (F4) is greater than a threshold number, we intuitively know the face has one

Figure 3.4: Examples of CK+ dataset (Lucey et al., 2010): (left to right) anger, disgust, fear, happiness, sadness, and surprise expression.

of non-neutral expressions such as happiness, surprise, partial anger, partial disgust, and partial fear. However, if the mouth is closed or a feature (F4) is less than the threshold number, the SVM classifier should double-check for distinguishing between 'neutral' class and 'non-neutral' class as we see Figure 3.1. For the neutral expression classifier, a SVM model with Linear kernel function is built by training with neutral faces and others from CK+ dataset (more details for CK+ in Section 3.3.1). We show accuracy of the SVM classifier for neutral expression in Section 3.3.3.

Second, for the facial expression classification, the features are used as an input for SVM classifiers of 6 emotions such as anger, disgust, fear, happiness, sadness, and surprise during expression frames. In the case of 13D geometric features approach, the SVM classifier with Radial Basis Function (RBF) kernel for emotion recognition is built with dynamic features from CK+ dataset with 6 emotions. For LBP features approach, the SVM classifier with Linear kernel is built with 46D features from CK+ dataset with 6 emotions.

3.3 Experimental Results

3.3.1 Facial Expression Dataset

For a set of SVM classifiers, we use the extended Cohn-Kanade (CK+) database including 593 video sequences from 123 subjects (Lucey et al., 2010) in training stage. All video samples have already been segmented temporally between neutral frame and peak frame of

Table 3.1: The accuracy results for facial expression recognition on CK+ dataset. The SVM classifiers with RBF kernel are evaluated with 10 folds Cross-validation. (Top) using 6D features and (Bottom) using 13D features.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.689	0.117	0.500	0.689	0.579	0.786	0.390
disgust	0.797	0.060	0.758	0.797	0.777	0.868	0.643
fear	0.480	0.011	0.800	0.480	0.600	0.735	0.426
happiness	0.957	0.046	0.857	0.957	0.904	0.955	0.830
sadness	0.036	0.014	0.200	0.036	0.061	0.511	0.095
surprise	0.940	0.044	0.886	0.940	0.912	0.948	0.849
weighted avg.	**0.761**	**0.053**	**0.730**	**0.761**	**0.734**	**0.854**	**0.636**

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.756	0.068	0.654	0.756	0.701	0.844	0.530
disgust	0.763	0.044	0.804	0.763	0.783	0.859	0.658
fear	0.640	0.018	0.762	0.640	0.696	0.811	0.517
happiness	0.928	0.025	0.914	0.928	0.921	0.951	0.864
sadness	0.500	0.043	0.538	0.500	0.519	0.729	0.315
surprise	0.928	0.031	0.917	0.928	0.922	0.948	0.870
weighted avg.	**0.809**	**0.038**	**0.809**	**0.809**	**0.808**	**0.886**	**0.700**

a facial expression for about 10-60 frames in CK+ dataset. In Figure 3.4, examples of CK+ dataset are shown (anger, disgust, fear, happiness, sadness and surprise expression).

During training SVM for the neutral expression, we used 309 neutral frames for positive set and 327 peak frames from 6 emotions plus contempt emotion for negative set. For 6 facial expressions, we took 309 neutral frames and 309 peak frames except contempt emotion among 7 emotions dataset. All 309 samples are 45 from anger, 59 from disgust, 25 from fear, 69 from happiness, 28 from sadness and 83 from surprise.

Table 3.2: The accuracy results for Neutral expression recognition using 13D geometric features with 10 folds Cross-validation: (top) Linear kernel function, (bottom) RBF kernel function.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Expressions	0.660	0.116	0.843	0.660	0.740	0.772	0.772
Neutral	0.884	0.340	0.734	0.884	0.802	0.772	0.708
weighted avg.	**0.775**	**0.231**	**0.787**	**0.775**	**0.772**	**0.772**	**0.715**

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Expressions	0.809	0.055	0.933	0.809	0.867	0.877	0.847
Neutral	0.945	0.191	0.840	0.945	0.889	0.877	0.822
weighted avg.	**0.879**	**0.125**	**0.885**	**0.879**	**0.878**	**0.877**	**0.834**

3.3.2 Evaluation of 13D Geometric features

As we mentioned in Section 3.2.1, we have added 7 points (P8-P13) and 7 features (F7-F13) to 6D features used in (Houstis and Kiliaridis, 2009). For evaluation of the extended features, we compared the results of facial expression recognition between using 6D and 13D geometric features on CK+ dataset in Table 3.1. The SVMs with RBF Kernel ($Cost$=128.0, γ=0.03125) for 6 classes of emotions are tested by 10-folds cross validation. The accuracy result using 6D features (F1-F6) shown in Figure 3.2 is 76.1%, whereas the extended 13D features result in 80.9%. Therefore this experimental result shows that the proposed system performs better with 13D geometric features.

3.3.3 Neutral Expression Recognition

In the training stage, the SVM classifier for neutral expression recognition is trained with CK+. Using 13D geometrical features, Table 3.2 shows accuracy results from two different kernel functions (top: Linear, and bottom: RBF kernel function) by 10-folds cross validation. While the SVM classifier with RBF has 87.9% of accuracy, the Linear kernel results in 77.5%

Table 3.3: The accuracy results for Neutral expression recognition using 1475 LBP features on CK+ dataset. The SVM classifiers with Linear kernel are evaluated with 10 folds Cross-validation.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Expressions	0.848	0.052	0.942	0.848	0.893	0.800	0.898
Neutral	0.948	0.152	0.862	0.948	0.903	0.800	0.898
weighted avg.	**0.898**	**0.102**	**0.902**	**0.898**	**0.898**	**0.800**	**0.898**

of the accuracy. We employ the Linear kernel function for neutral expression recognition in our proposed system even if the SVM classifier with RBF Kernel results in better accuracy. While the SVM classifier with RBF kernel is apt to be more optimized to CK+ dataset, the SVM with Linear kernel is flexible. Therefore, the Linear kernel is more suitable for the application dealing with data from outside the training dataset (as in our case of using the mobile facial expression application with users not present in CK+ dataset). The SVM classifier built from CK+ is embedded with the mobile application and tested on real-time video on smartphones.

Similarly, the accuracy result for the neutral expression recognition using 1475D LBP features from SVM with the Linear kernel on CK+ dataset is 89.8% from 10-folds cross validation in Table 3.3.

3.3.4 Facial Expression Recognition using 13D Geometric features on CK+

We show the accuracy results for 6 facial expressions recognition in Table 3.4. Table 3.4 shows the accuracy result of 6 facial expression recognition using 13D geometrical features. The dynamic features are created with neutral frame (the first frame of video sample) and apex frame (the last frame of video sample) in CK+ dataset. We obtained average 85.8% of accuracy from 10-folds cross validation.

Table 3.4: The accuracy results for facial expression recognition using 13D geometric features on CK+ dataset. The SVM classifiers with RBF kernel are evaluated with 10 folds Cross-validation.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.778	0.045	0.745	0.778	0.761	0.866	0.612
disgust	0.831	0.044	0.817	0.831	0.824	0.893	0.711
fear	0.720	0.011	0.857	0.720	0.783	0.855	0.640
happiness	0.928	0.021	0.928	0.928	0.928	0.953	0.877
sadness	0.679	0.025	0.731	0.679	0.704	0.827	0.525
surprise	0.964	0.027	0.930	0.964	0.947	0.969	0.906
weighted avg.	**0.858**	**0.030**	**0.857**	**0.858**	**0.857**	**0.914**	**0.763**

Table 3.5: The accuracy results for facial expression recognition using 46 LBP features on CK+ dataset. The SVM classifiers with RBF kernel are evaluated with 10 folds Cross-validation.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.711	0.038	0.762	0.711	0.736	0.693	0.837
disgust	0.898	0.060	0.779	0.898	0.835	0.795	0.919
fear	0.520	0.004	0.929	0.520	0.667	0.677	0.758
happiness	0.986	0.013	0.958	0.986	0.971	0.963	0.987
sadness	0.714	0.028	0.714	0.714	0.714	0.686	0.843
surprise	0.976	0.022	0.942	0.976	0.959	0.943	0.977
weighted avg.	**0.864**	**0.029**	**0.866**	**0.864**	**0.860**	**0.838**	**0.918**

Because 'sadness' is done with relatively small movements and sometimes ambiguous with some other expressions such as anger, disgust and fear, we have the lowest accuracy, 67.9% among 6 emotions as we see Table 3.4.

Table 3.6: The accuracy results for facial expression recognition using 46 LBP features on CK+ dataset. The SVM classifiers with Linear kernel are evaluated with 10 folds Cross-validation.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.711	0.038	0.762	0.711	0.736	0.693	0.837
disgust	0.898	0.044	0.828	0.898	0.862	0.829	0.927
fear	0.840	0.011	0.875	0.840	0.857	0.845	0.915
happiness	0.986	0.008	0.971	0.986	0.978	0.972	0.989
sadness	0.643	0.025	0.720	0.643	0.679	0.651	0.809
surprise	0.976	0.013	0.964	0.976	0.970	0.959	0.981
weighted avg.	**0.883**	**0.022**	**0.881**	**0.883**	**0.882**	**0.861**	**0.931**

Table 3.7: The accuracy results for facial expression recognition using 1475 LBP features on JAFFE dataset. The SVM classifiers with Linear kernel trained by CK+ are used for testing JAFFE dataset.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.700	0.196	0.412	0.700	0.519	0.416	0.752
disgust	0.172	0.123	0.208	0.172	0.189	0.053	0.525
fear	0.125	0.099	0.211	0.125	0.157	0.032	0.513
happiness	0.452	0.046	0.667	0.452	0.538	0.477	0.703
sadness	0.032	0.039	0.143	0.032	0.053	-0.014	0.496
surprise	0.900	0.222	0.443	0.900	0.593	0.532	0.839
weighted avg.	**0.393**	**0.12**	**0.347**	**0.393**	**0.34**	**0.248**	**0.637**

3.3.5 Facial Expression Recognition using LBP on CK+

For the facial expression recognition using LBP features, SVM classifiers are trained with CK+ dataset. While the 13D geometric features are created by the displacement of features between neutral and peak frames of facial expressions, 1475D LBP features are extracted in only 309 peak frames of 6 emotions on CK+. By the feature selection with F-score, 46D

(a) Anger (b) Disgust (c) Fear (d) Happiness (e) Sadness (f) Surprise (g) Neutral

Figure 3.5: Screenshot of facial expressions captured in real time mobile app.

features with high F-score are selected from 1475D features. The accuracy results for 6 facial expressions recognition by SVM with RBF kernel are shown in Table 3.5. The result is average 86.4% of accuracy from 10-folds cross validation. Similarly, Table 3.6 shows the accuracy result by SVM with Linear kernel. The accuracy is 88.3%. Compared to 85.8%, the accuracy result of 13D features in Table 3.4, the facial expression recognition using LBP features has a little bit higher accuracy results. This LBP SVM model built with CK+ is embedded into the mobile app for real-time emotion recognition using LBP.

Additionally we tested with the Japanese Female Facial Expression (JAFFE) Database (Lyons et al., 1998) (Total 183, Anger: 30, Disgust: 29, Fear: 32, Happiness: 31, Sadness: 31, Surprise: 30) for 6 emotion recognition on the SVM classifiers with Linear kernel trained by CK+ in Table 3.7. This shows the results from testing dataset (JAFFE) not present in training dataset (CK+).

3.3.6 Evaluation of Real-time Emotion Recognition Accuracy using 13D Geometric features

For the evaluation of the performance of real-time emotion recognition system running on mobile platform, we developed the mobile application on Android smartphones. The main device for experiments is *Samsung Galaxy S3* (CPU: Quad-core 1.4 GHz Cortex-A9, GPU: Mali-400MP, Android 4.3 *Jelly Bean*). We asked 7 subjects to perform 7 expressions as examples of Figure 3.5, looking at the front-facing camera of the smartphone. It should be noted that they are not professional actors and made almost posed facial expressions,

not spontaneous expressions. And we just considered the expressions of only peak frames except transitional expressions because our system gives us back all results from non-neutral expressions.

Each participant performed the 6 different expressions and the neutral expression 10 times. Therefore, 70 facial expressions are created per participant. The classification confusion matrix for 6 emotions and neutral expression is shown in Table 3.8. The average accuracy of 7 expressions is about 72%.

One good group (disgust, happiness and surprise) has average 92% of accuracy, whereas the other group (anger, fear and sadness) has worse result with average 43% of accuracy. The obvious reason about bad accuracy from expressions such as anger, fear and sadness is that subjects, being nonprofessionals, were uncertain about what to do for such expressions. On the other hand, they easily performed expressions such as disgust, happiness and surprise intuitively.

As the result from CK+ dataset shows, 'sadness' is the most difficult expression for classification because it is likely to be confused with other expressions such as anger, disgust, and fear. Moreover because the 'sadness' usually has closed mouth, it is liable to be classified to neutral class particularly. However, the expressions with both open mouth such as happiness and surprise rarely happen to be misclassified.

3.3.7 Evaluation of Real-time Emotion Recognition Accuracy using LBP

The experiment for real-time emotion recognition accuracy using LBP features on the smartphone is carried out with the same devices and methods as those for the evaluation with the 13D geometric features. The SVM models evaluated and built by 309 CK+ dataset in the previous section are used for the evaluation of real-time emotion recognition using LBP appearance features on smartphones. The confusion matrix for emotion classification is shown in Table 3.9.

Table 3.8: Emotion classification confusion matrix as result of facial expression recognition (%) using 13D geometric features module on *Samsung Galaxy S3*–Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su), and Neutral (Ne).

	An	Di	Fe	Ha	Sa	Su	Ne
An	**41.4**	32.9	10.0	2.9	2.9	5.7	4.3
Di	4.3	**87.1**	2.9	2.9	1.4	1.4	0.0
Fe	7.1	4.3	**57.1**	2.9	20.0	8.6	0.0
Ha	0.0	0.0	0.0	**98.6**	1.4	0.0	0.0
Sa	15.7	17.1	5.7	0.0	**30.0**	2.9	28.6
Su	2.9	1.4	1.4	1.4	1.4	**91.4**	0.0
Ne	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**

Table 3.9: Emotion classification confusion matrix as result of facial expression recognition (%) using LBP features module on *Samsung Galaxy S3*.

	An	Di	Fe	Ha	Sa	Su	Ne
An	**44.3**	20.0	12.9	1.4	0.0	21.4	0.0
Di	0.0	**92.9**	0.0	5.7	0.0	1.4	0.0
Fe	10.0	10.0	**52.9**	2.9	5.7	18.6	0.0
Ha	0.0	1.4	0.0	**98.6**	0.0	0.0	0.0
Sa	14.3	14.3	12.9	7.1	**32.9**	12.9	5.7
Su	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0
Ne	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**

3.3.8 Evaluation of Processing Time using 13D Geometric features

For evaluating the performance of processing time on the proposed system, we used with *Samsung Galaxy S3* as well as other commercial off-the-shelf smartphones. Our proposed system is implemented on Android smartphones because the Android OS is the most popular mobile OS (dominating nearly 85% of the market share in the second quarter of 2014) running on diverse smartphones and tablets.

Table 3.10: Average computation time and standard deviation (milliseconds) taken in the process of modules of the proposed system. Two different smartphones (*Nexus 4* and *Galaxy S3*) are tested.

	Frame Resolution	Frames	Optical Flow	ASM	SVM Neutral	SVM 6 Emotions	Total Frames
Nexus 4	480 x 320	119	69.4±14.3	416.3±74.5	1.1±0.38	0.9±0.67	526.5±79.24
Galaxy S3	640 x 480	146	59.1±7.29	318.3±9.79	1.0±0.15	0.6±0.24	420.2±14.97

The screenshot of the mobile app running on *Galaxy S3* is shown in Figure 3.6. In previous experiments (Suk and Prabhakaran, 2014), two different smartphones, *Samsung Galaxy S3* and *Google Nexus 4* (CPU: Quad-core 1.5 GHz Krait, GPU: Adreno 320, Android 4.4 *KitKat*) are compared to each other in terms of the processing time (See average computation time and standard deviation (milliseconds) taken in the process of the sub-modules in Table 3.10). For one minute, average computation time in each sub-module is taken. The frame resolutions from smartphones may not be the same because most smartphones have different camera capabilities such as picture preview size and ratio of video images.

Figure 3.6: Screenshot of mobile app for real time facial expression recognition.

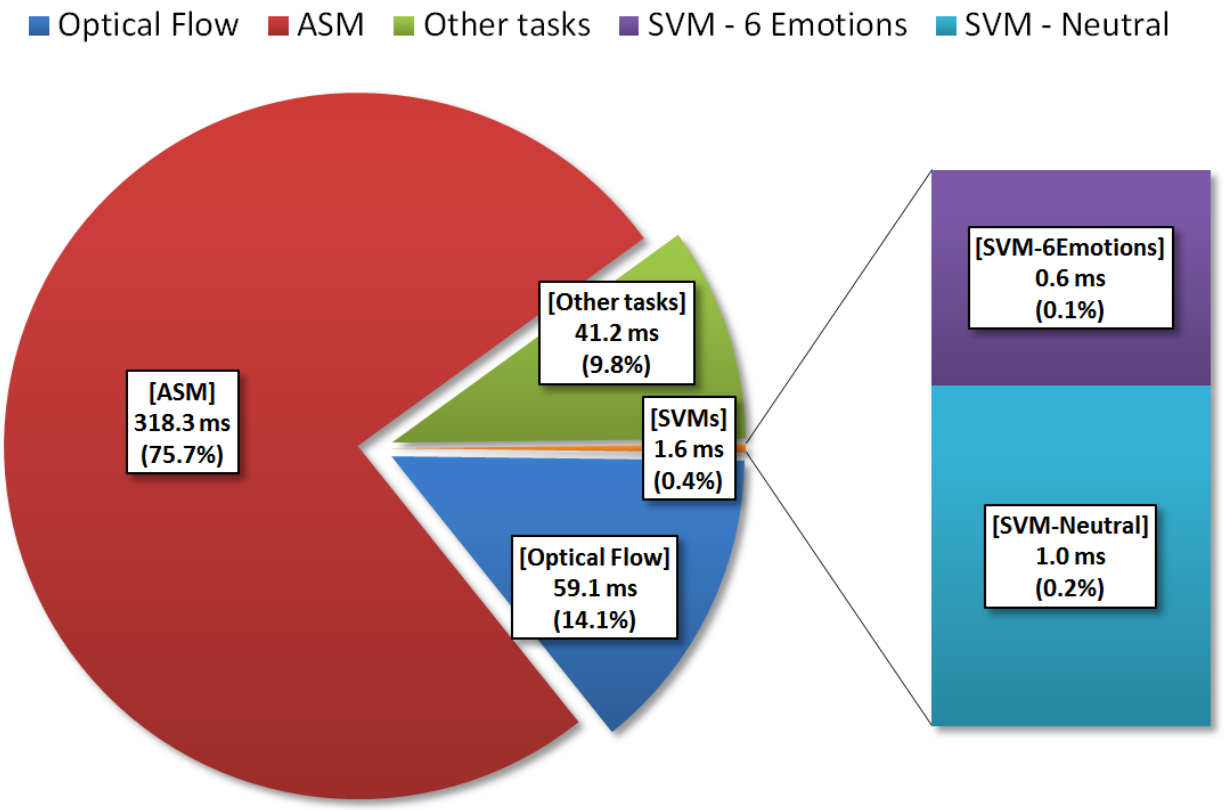


Figure 3.7: Computation time in modules (Optical flow, ASM, SVM-neutral and SVM-emotions) in a grabbed frame when running on *Samsung Galaxy S3*.

In Figure 3.7, the pie chart shows proportions of sub-module's computation time in a single frame grabbed on *Galaxy S3* (it is the same data as *Galaxy S3* in Table 3.10). The ASM module for face detection and fitting landmarks takes the most time-consuming task with 75.7%. However, the SVM classification modules take slight time about 0.1%-0.2%. The rest of tasks includes image and graphics processing with average 9.8%.

There is some room for improvement of computation time by changing parameters such as face size to be searched in the ASM module. The minimum size of the face to be found can be adjusted (e.g. 25% or 50% width of a frame size). Absolutely smaller number of minimum size takes longer time with a small sliding window to find faces. Also from a perspective of the usage with the front-facing camera of smartphones (we can hardly set our faces very small within arm reach), it is reasonable to use large minimum size. We shows

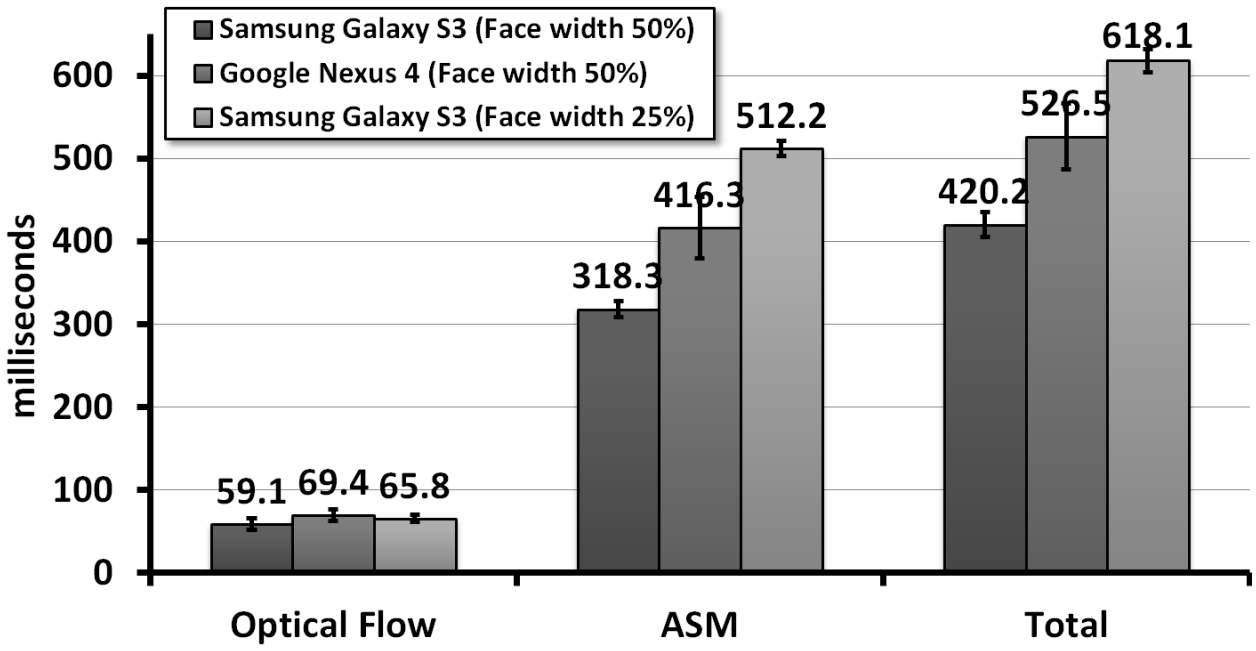


Figure 3.8: Computation time comparison in different parameters such as minimum width of face - 25% and 50% on *Galaxy S3*, and 50% on *Nexus 4*.

the comparison between different minimum sizes of face width (25% and 50%) on *Galaxy S3* in Figure 3.8. Therefore, as we had anticipated, the minimum face width 50% requires less processing time than 20% on ASM module.

Evaluation with Additional Smartphones: In previous experiments (Suk and Prabhakaran, 2014), we tried to test with other more recent smartphones such as *HTC One* (CPU: Quad-core 1.7 GHz Krait 300, GPU: Adreno 320, Android 4.4 *KitKat*) and *Samsung Galaxy S4* (CPU: Quad-core 1.9 GHz Krait 300, GPU: Adreno 320, Android 4.4 *KitKat*). However, we failed to run the app with an issue that *OpenCV* Library 2.4.8 implemented using non-public Android API might not support the camera of particular smartphones on which the vendor had modified some parts of Android OS. On the contrary, *Google Nexus 4* had no problem although it used the same versions as Android 4.4 and *OpenCV* library 2.4.8. Therefore, since we revised the mobile application with *OpenCV* library 2.4.9 release which fixed the issues, we have improved the performance of processing time with the

Samsung Galaxy S3 as well as can run other smartphones such as *Samsung Galaxy S4*. As additional experiments, Table 3.11 shows average computation time and standard deviation (milliseconds) by 13D geometric features. We took the evaluation of processing time with smartphones such as *Google Nexus 4*, *Samsung Galaxy S3*, *Samsung Galaxy S4*, *Samsung Galaxy Tab 4* (CPU: Quad-core 1.2 GHz Cortex-A7, GPU: Vivante, Android 4.4.2 *KitKat*) and *HTC Evo* (CPU: Dual Core 1 GHz Scorpion, GPU: Adreno 200, Android 4.1.1 *Jelly Bean*). Compared to the result in Table 3.10, the mobile app updated by the latest *OpenCV* manager (2.4.9) which fixed issues in Native Camera is used on all the smartphones. The updated mobile app has improved the speed performance from 420.2ms to 309.6ms on *Galaxy S3*. Computation times are shown in Table 3.11. While the frame resolution of *Galaxy S4* is 12.5% bigger than that of *Galaxy S3* (from the ratio of 307,200 pixels to 345,600 pixels), the computation times of ASM in *Galaxy S4* are 34.6% and 66.5% longer than those of *Galaxy S3* in Table 3.10 and Table 3.11 respectively. Because of the fact that higher spec smartphones such as *Galaxy S4* give bad performance in terms of computational speed, we still doubt *OpenCV* 2.4.9 compatibility with some of specific Android smartphones.

Evaluating the Possible Use of Optical Flow in Mobile Phones: In the evaluation for the computation time, we have added a Lucas-Kanade's optical flow module (Lucas and Kanade, 1981) to see how the optical flow can affect the performance of the system. In contrast with a stationary camera installed general systems, hand-held mobile devices have issues such as camera shake causing relative head movements. Moreover, head movements arising out of spontaneous facial expressions can lead to degradation of facial expression recognition accuracy. Therefore, the Lucas-Kanade's optical flow can help to overcome the negative effect of head movement on smartphones. In our experiments, the optical flow module takes average 59ms (14%) in a single frame on *Galaxy S3* (See Figure 3.7). The integration with the optical flow vector will be taken up as a future work because the additional task with optical flow module makes the system slow down.

Table 3.11: Average computation time and standard deviation (milliseconds) taken in the process of modules of the proposed system (with revised version with *OpenCV* 2.4.9). Additional smartphones (*Samsung Galaxy S4*, *Samsung Galaxy Tab 4* and *HTC Evo*) are tested. *Galaxy S3* runs at average 3.23 *fps*.

	Nexus 4	Galaxy S3	Galaxy S4	Galaxy Tab 4	HTC EVO
Frame Resolution	576x432	640x480	720x480	640x480	576x432
# of Frames	116	183	136	132	166
Optical Flow	62.0±17.3	46.8±11.5	36.9±21.9	42.6±18.0	32.3±14.0
ASM	500.8±50.3	257.3±21.6	428.3±49.7	372.8±16.3	325.8±30.1
SVM (Neutral)	1.1±0.5	1.1±0.5	0.9±0.5	1.6±0.6	0.9±0.4
SVM (6 Emotions)	1.0±0.8	0.8±0.5	0.7±0.6	1.3±0.6	0.7±0.5
Total time	**569.4±54.3**	**309.6±28.4**	**473.4±59.9**	**422.6±26.0**	**368.1±38.1**

Demo of the Proposed Mobile App: The proposed mobile application for facial expression recognition can be downloaded from: `https://www.dropbox.com/s/716dlnwka4irc8y/EmotionRecognitionRT.apk`

3.3.9 Evaluation of Processing Time using LBP

Table 3.12 shows average computation time and standard deviation (milliseconds) taken in the process of LBP features module on the smartphone. The LBP module consists of pre-process for face alignment, LBP operation, and calculating histogram from LBP. For example, it takes $13.9 \pm 2.5(ms)$ for face alignment, $282.6 \pm 55.9(ms)$ for LBP operation, and $83.8 \pm 40.9(ms)$ for calculating histogram among the LBP+SVM module on *Samsung Galaxy S3*.

For the evaluation of speed performance, we show average computation time and standard deviation (milliseconds) taken in the process of the LBP features module in Table 3.12.

Table 3.12: Average computation time and standard deviation (milliseconds) taken in the process of modules of LBP features based system (with revised version with *OpenCV* 2.4.9). Additional smartphones (*Samsung Galaxy S4*, *Samsung Galaxy Tab 4* and *HTC Evo*) are tested. It is also to be noted that 'Total time' is average computation time for all frames, not just the sum of average computation time of 3 sub-modules (ASM, Neutral SVM, and LBP+SVM) in the table. *Galaxy S3* runs at average 2.33 *fps*.

	Nexus 4	Galaxy S3	Galaxy S4	Galaxy Tab 4	HTC EVO
Frame Resolution	576x432	640x480	720x480	640x480	576x432
# of Frames	79	144	91	89	99
Optical Flow	65.4±19.5	38.5±13.7	55.9±16.4	38.0±29.6	69.5±21.5
ASM	483.8±46.7	255.7±25.3	414.1±55.5	377.8±19.0	319.6±34.2
SVM (Neutral)	3.4±15.0	1.1±0.3	0.9±0.4	1.8±0.7	1.0± 0.3
LBP+SVM (6 Emotions)	625.2±64.2	465.4±43.9	656.0±86.2	773.9±102.9	948.7±137.2
Total time	**795.0±320.4**	**429.4±216.5**	**731.0±323.9**	**719.3±391.8**	**780.9±478.1**

3.3.10 Evaluation of Facial Expression Recognition using 13D Geometric features with CK+ dataset on Smartphone

As an additional evaluation, the CK+ dataset is evaluated in the proposed system running on a smartphone. The CK+ video set has 225 samples (anger: 26, disgust: 33, fear: 18, happiness: 60, sadness: 16, surprise: 72). As mentioned in Section 3.3.1, all video samples have been temporally segmented from neutral expression to peak expression. The average accuracy of 6 emotions recognition with CK+ dataset on a smartphone is about 46.2% by classifying correctly 104 of 225 if we decide by majority of expression in a video sample. The result of majority rule from detected expressions must be worse than that in Table 3.8 where we ignored transitional expressions and accepted only peak expression during experiment. While smartphone devices with limited resource cause real-time video streaming at low frame rate and hardly ever take frames including transitional expressions such as onset states, CK+

videos recorded at high frame rate used for testing on smartphones take many ambiguous transitional frames between neutral and peak expression.

CHAPTER 4

REAL-TIME FACIAL EXPRESSION RECOGNITION ON SMARTPHONES[1]

4.1 Proposed System Overview

In Figure 4.1, the proposed system for automatic facial expression recognition consists of two main modules:

- **Temporal Segmentation of Video Sequences:** For segmentation of continuous video sequences into a series of expressions, the proposed system adopts a Finite State Machine (FSM) model. This model uses dynamic features of input video extracted by using the Lucas-Kanade optical flow method as triggering condition for each transition. Because the FSM consists of a finite number of states representing neutral (start/accept state), onset, apex, and offset state, the systems can recognize the current state at any given time. The FSM is a simplified version of HMM where probabilities of state transition are equal to zero or one.

- **Facial Expression Recognition:** The first step of facial expression recognition is automatic feature extraction by Active Shape Models (ASM) and feature normalization. Subsequently, we use a static classifier, Support Vector Machines (SVM) for recognizing the facial expression at every apex state.

4.1.1 Temporal Segmentation for Facial Expression in Video Sequences

Finite State Machine as a model of time: The main module for temporal segmentation is built by finite state machine, $M = (Q, \Sigma, \delta, q_0, F)$ where a finite set of states

[1]©2015 IEEE. Reprinted, with permission, from Myunghoon Suk, B. Prabhakaran, "Real-time Facial Expression Recognition on Smartphones," Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, pp.1054,1059, January 2015.

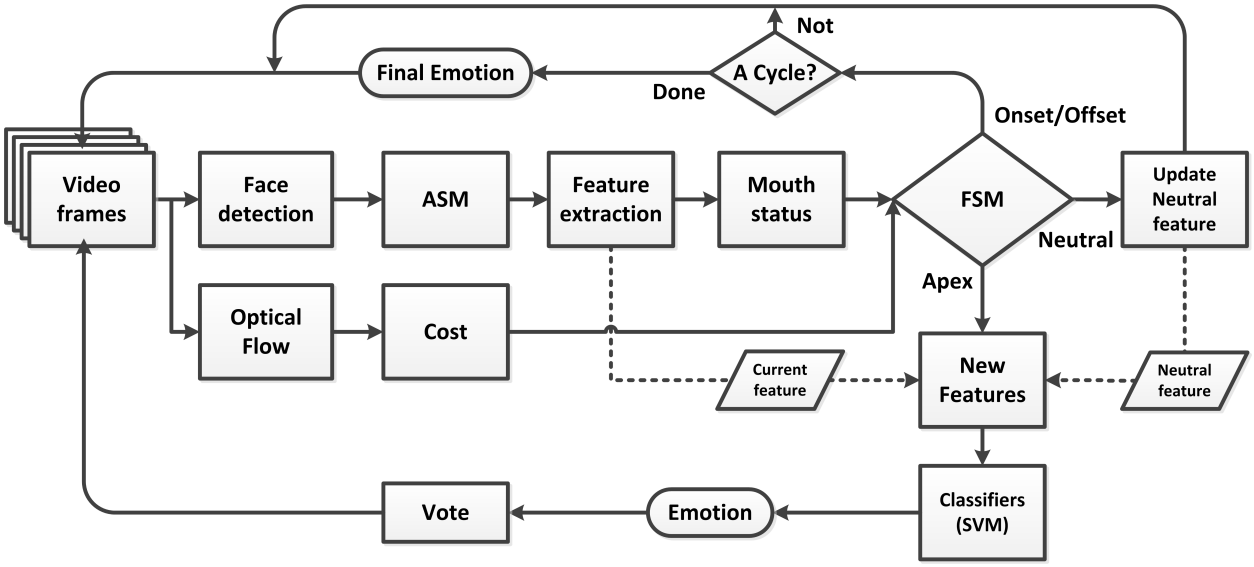
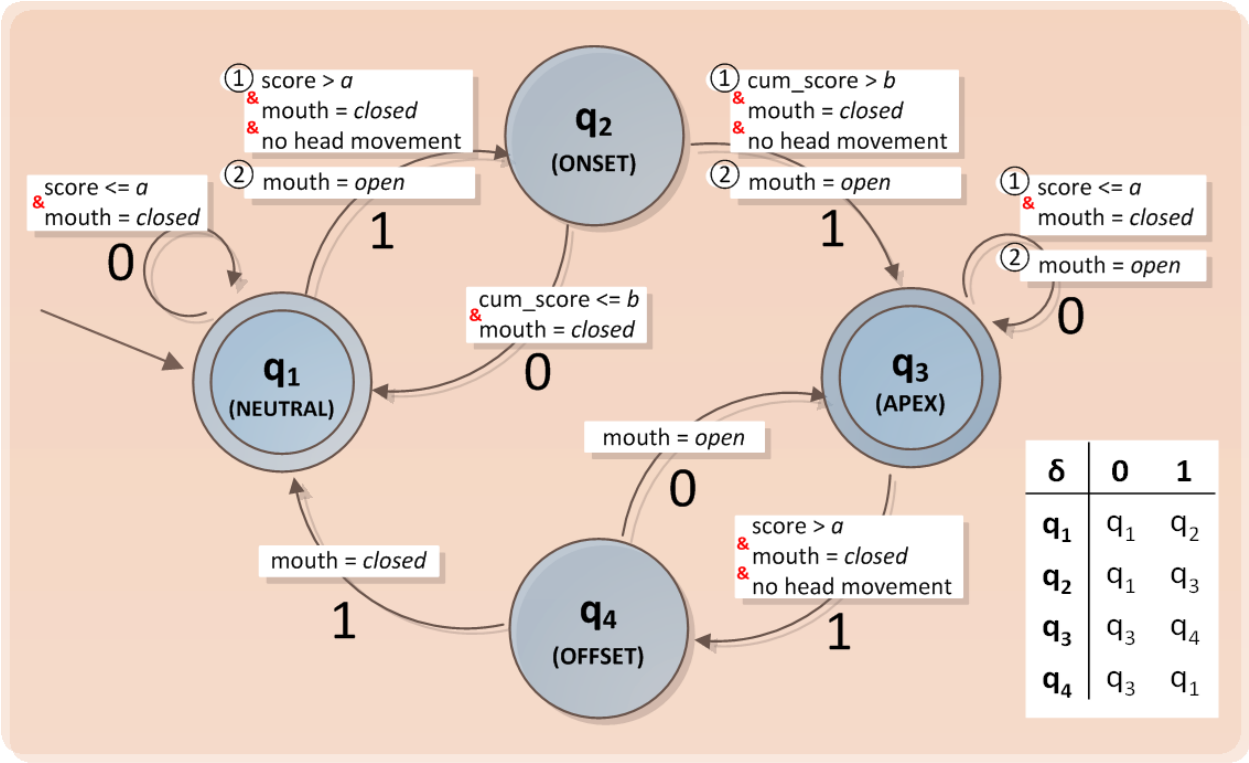


Figure 4.1: Flow diagram for proposed system.

($Q = \{q_1, q_2, q_3, q_4\}$), a finite set of input symbols ($\Sigma = \{0, 1\}$), a start state ($q_0 = q_1$), a set of accept states ($F = \{q_1, q_3\}$), and a transition function where δ is defined by the state transition table (Figure 4.2 shows the proposed FSM and (a)-(g) are examples of video frames corresponding to each state). The language recognized by M is the regular language given by the regular expression $0^*1(01)^*10^*1(01)^*1$, where '*' is the Kleene star. A complete facial expression typically involves starting from a neutral state, going to the peak expression, and coming back to neutral state. Hence, the proposed FSM is designed with two final or accept states: neutral and apex. Because an accept state is q_1 (neutral state), the FSM accepts an input video sequence including a cycle of temporal phases (from neutral, onset, apex, offset to neutral states) in a facial expression. Examples of strings accepted by this state machine are 1111 (minimum length), 01111, 001111, ..., 101111, 10101111, ..., 11011, 110011, ..., 111011, ..., etc. Moreover, in the case of q_3 as an accept state, the FSM accepts an input string for facial expression recognition, and the recognition task is performed on q_3 (apex state). The minimum set of states for the FSM representing temporal phases of a facial expression are a neutral state, two transition states (onset and offset) and an apex state. In order to move from neutral to apex state, at least two video frames meeting the



(a) q_1 (b) q_2 (c) q_3 (d) q_3 (e) q_3 (f) q_4 (g) q_1

Figure 4.2: State diagram and the state transition table of a finite-state machine (M) for facial expression segmentation in proposed system. (a)-(g) are examples of video frames corresponding to states in FSM. (a) and (g) Neutral, (b) Onset, (c)-(e) Apex, (f) Offset.

condition of each transition such as (neutral and onset) and (onset and apex) are needed. Two transition states are to act as a buffer against erroneous '*score*' so that we make the FSM system for temporal segmenting more robust to fickle environment.

Conditions of transition: The proposed FSM uses the combination of the following conditions for transition: (a) '*score*'; (b) '*cum_score*'; (c) *mouth* status and (d) *head movement* status.

***score*:** is obtained by counting number of dynamic features detected in the bounding box

on the face by Lucas-Kanade's optical flow vectors. The '*score*' is defined as the following:

$$\frac{dynamic\,features\,\#}{total\,features\,\#}(0 \leq score \leq 1) \tag{4.1}$$

The '*total features*' is the number of total optical flow features in the face bounding box (N), and the '*dynamic features*' is the number of optical flow features with movement in the face bounding box ($0 \leq dynamic\,features \leq N$), meeting the following condition:

$$\sum_{i=1}^{N} \left(count_i = \begin{cases} 1 & if\,(distance_i > threshold_b) \\ 0 & otherwise \end{cases} \right) \tag{4.2}$$

The '$distance_i$' is the distance of i^{th} optical flow feature between two image frames. The '$distance_i$' is obtained by the following way:

$$\frac{\sqrt{\left(P_i'^x(t) - P_i'^x(t-\Delta t)\right)^2 + \left(P_i'^y(t) - P_i'^y(t-\Delta t)\right)^2}}{\Delta t} \tag{4.3}$$

Where $P_i'(t) = \dfrac{P_i(t)}{w}$ $(0 \leq P_i'(t) \leq 1)$ is the normalized x or y coordinate of $i^{th}(0 \leq i \leq N)$ optical flow feature at time t with w of width of face bounding box in original image frame. Also $P_i(t)$ is x or y coordinates of i^{th} optical flow feature at time t. t is a current time and Δt is time difference between two image frames. For example, two successive frames have $t - (t-1) = 1$.

The proposed system employs the Lucas-Kanade optical flow for detecting the movement of facial muscle or anatomical component on face in video sequence. Although many optical flow features can be tracked in video sequences, only some of the features tracked in face area are selected for temporal segmentation of facial expressions, as we focus on the face area selected by Haar cascade facial detection module with OpenCV in each frame. All the features tracked by Lucas-Kanade method in face area are identified as either a static or dynamic feature depending on how far it has moved. The displacement of a feature tracked between two consecutive video frames is calculated. In Eq. (4.2), if the ratio of

the displacement to face size is more than a particular threshold ($threshold_b$), the feature is tagged as 'dynamic'. Otherwise, it is considered as a static feature. After tagging all the features in face area in each video frame, score in regard to motion detection of facial expression is calculated as the ratio of the number of dynamic features to the total number of features in face area in Eq. (4.1). The score is used as one of transition conditions. Zero or small score can be considered as no facial action or false motion detection by feature tracking error.

***cum_score*:** is the accumulated score used in a sequence from neutral state to apex state, while the '*score*' is treated between successive frames. Therefore, FSM in Figure 4.2 shows transitions between states triggered by comparing '*score*' and '*cum_score*' to thresholds such as 'a' or 'b' determined with the experimental result.

***mouth*:** The status of mouth in current state is an additional condition. Because most features in both neutral and apex states are static in a moment about a few frames, it is somewhat difficult to distinguish between them by only optical flow. If the system mistakenly identifies apex state as neutral state, following transitions are incorrectly conducted in FSM model.

In order to clearly distinguish between neutral and apex expression, the status of the mouth (open or closed) is the clearest possible key feature than the others such as eye, eyebrows, and nose. Open mouth should not exist in neutral expression typically. Therefore, double-checking open mouth in the neutral state helps with distinction between neutral and apex state. The status of mouth is determined by features based on ASM (explained in Section 4.1.2).

Likewise, at the onset state, if the mouth is open, the proposed FSM is firmly convinced of going from onset to apex regardless of '*score*' or '*cum_score*'. However, in cases of expressions with no open mouth, the transition depends on the combination of '*score*' and '*cum_score*' condition.

***head movement*:** When features are tagged as 'dynamic', we should double-check if the head motion is causative of the dynamic features. If the head movement occurs, we need to exclude the features from 'dynamic' tagged features. For example, if the ratio of dynamic features to all features tracked in face area is greater than 50%, the motion of facial features is regarded as head movement.

The proposed system keeps estimating facial expressions every frame in the apex state as explained in Section 4.1.3. After leaving apex state, the state machine returns to neutral state (final state) through offset state. With a single cycle of sequence of states, a facial expression in video sequences can be clearly segmented.

4.1.2 Feature Extraction

Feature creation from ASM landmarks: For feature extraction, the proposed system applies Active Shape Models (ASM) to each frame in video sequences. Facial features such as fiducial points are extracted to represent facial geometry from ASM landmarks. Using STASM (Milborrow and Nicolls, 2008), 77 facial landmarks are located on a face. Not only the x, y coordinates of landmarks themselves can be used for facial shape features, but also new features such as distances between two particularly selected markers are generated as high-level facial shape features based on the x, y coordinates of landmarks.

Feature normalization (scale, translate, and rotation): After acquiring facial features from ASM, the facial features should be normalized geometrically because all face images are not captured at the same distance from the camera, and some faces may be out of upright. The geometrical normalization is performed to get the normalized x, y coordinates of landmarks.

4.1.3 Facial Expression Recognition

The basic classifier we employ in the proposed system is the Support Vector Machine (SVM). SVMs are already popular in many of pattern recognition tasks including face recognition

and facial expression recognition. We used the popular open source machine learning library called LIBSVM (Chang and Lin, 2011). First, for SVM with Radial Basis Function (RBF) kernel, we try to get the best value γ and C by parameter selection with the greedy search method provided by the LIBSVM library, and then trained the SVM classifiers with the parameters from training data set comprising the displacement of the normalized geometrical features between neutral and apex expression. For both taking advantage of extremely efficient SVMs and temporal dynamics in video sequences, the recognition task is performed only during the stay on apex frames detected through the FSM, and uses temporal dynamics between neutral and apex frames.

Recognizing emotions per apex frame: In video sequences, the system checks if the current frame is on neutral or apex. If it is a neutral frame, the system saves current features as neutral features and keeps up-to-date neutral features during neutral states. In FSM, the system checks current state at every frames. If the current state is on apex, the system extracts apex features and creates a new feature vector as relative displacement between neutral and apex features. Whenever the system meets apex states during a single facial expression, the feature vector is fed into SVM models to classify facial expression.

Final emotion estimation in a temporal segmentation of facial expression: Eventually the final decision for facial expression in a single temporal segmentation of video sequences is determined by the majority of facial expressions counted in apex states.

4.2 Experimental Results

We carried out the following experiments: (a) performance of the proposed system in terms of computational time; (b) facial expression recognition using the extended Cohn-Kanade data set; (c) real-time segmentation and recognition of posed expressions by 5 different subjects looking at the front facing camera on a smartphone.

Table 4.1: Given a video frame of video sequences, average processing time and standard deviation (milliseconds) for main modules of the proposed system.

Frame Resolution	# of Frames	Optical Flow	ASM	SVM	HMM	Total
352x288	856	16.5±5.5	248.2±28.4	0.7±0.5	-	268.5±29.6
352x288	851	17.4±5.1	253.1±27.9	-	20.0±13.5	280.9±33.0
640x480	682	38.8±15.4	258.5±22.1	0.7±0.5	-	302.7±26.6
640x480	530	52.5±28.5	332.3±87.1	-	36.0±29.2	407.7±108.4

4.2.1 Analyzing the Proposed system's Performance in Computation time

Because the proposed system is aiming at working for real-time mobile application, it is important to retain at least reasonable speed performance. Therefore, to show the performance, we evaluated the proposed system on a smartphone, *Samsung Galaxy S3* (CPU: Quad-core 1.4 GHz Cortex-A9, GPU: Mali-400MP, Android 4.3 *Jelly Bean*). The processing times are obtained during performing a sequence of facial expressions about 1 minute. Table 4.1 shows the elapsed time and standard deviations (milliseconds) for the process of each module in a video frame. Compared to a SVM classifier, we also tested with a HMM classifier already trained by pre-segmented video sequences of the same CK+ dataset. Although HMM is a representative classifier handling time series, it still requires a finite sequence of frames as input. That means we need to have temporally segmented video sequences before using HMM in testing stage. While SVM needs only two frames such neutral and apex states, HMM uses a sequence of all frames between neutral and apex states including transitional states as input. The first and third rows of Table 4.1 are the results using SVM as a classifier on apex state, but with different frame resolutions. The second and fourth rows result from HMM classifier. There are core modules such as Optical flow, ASM and SVM or HMM as classifiers in FSM. The ASM module is to find a face and extract 77 landmarks from ASM. The ASM module is the most time-consuming module among three core modules in our

proposed system, as it takes about 80-90% of the total processing time. The optical flow module is to calculate a cost value by checking dynamic features. SVM takes less than a millisecond whereas HMM takes more time. The decision time for state transition in FSM module is negligible. The entire processing time per a frame varies depending on fitting time for good positions of landmarks in ASM, number of features detected in optical flow and prediction time of a classifier in apex state. The total processing time per a video frame (352×288) is 268.5 ms or 3.72 *fps* on average.

4.2.2 Experiments on Cohn-Kanade Facial expression Dataset

Standard public dataset for facial expression recognition for training: The extended Cohn-Kanade (CK+) database (Lucey et al., 2010) is a public and popular dataset for facial expression recognition–Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa) and Surprise (Su), but each sample has been already segmented in video images. Each sample has a subject with a facial expression from neutral to apex state. Therefore basically in order to use Cohn-Kanade dataset for training SVM model as a baseline classifier, neutral frame (first frame) and apex frame (last frame) in a video sample are chosen, and then feature vectors are created by displacement of features extracted by the frames.

Continuous video sequence re-created from CK+ for testing: In order to evaluate our system on the CK+ dataset, it is absolutely necessary to use continuous video sequences including an expression with a cycle from neutral, through apex, back to neutral state. However, each video sample in CK+ dataset starts with a neutral frame and ends with peak frame. Therefore, we re-created continuous CK+ video sequences by attaching reversed video sequence to the end of the original video sample so that a video sample includes all four temporal phases (neutral, onset, apex and offset). Finally we combined all the re-created video samples sequentially into a continuous video sequence. A combined video has 225 samples (anger: 26, disgust: 33, fear: 18, happiness: 60, sadness: 16, surprise: 72).

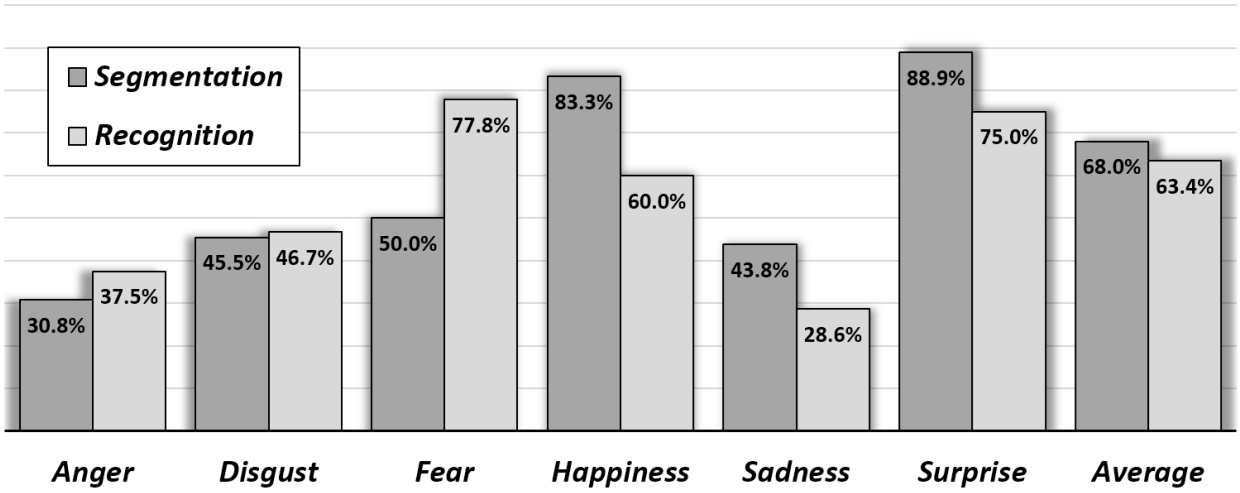


Figure 4.3: The accuracy of segmentation and recognition performance with CK+ dataset on the proposed system on mobile.

Temporal segmentation and Facial expression recognition Results with CK+ Dataset on a Smartphone

For the performance evaluation of temporal segmentation with CK+ dataset, the above mentioned continuous CK+ video is used in the proposed system running on a smartphone. First, we evaluate the accuracy of temporal segmentation on the continuous video sequence, and then the accuracy of facial expression recognition is evaluated in successfully segmented video sequences.

The results of temporal segmentation and recognition with CK+ dataset on a smartphone are presented in Figure 4.3. Average accuracy of all six emotions for temporal segmentation is 68.0% and the accuracy of each emotion is also shown in Figure 4.3. For the accuracy of facial expression recognition, we take in consideration only samples successfully segmented from a continuous video samples. For example, 153 samples are successfully segmented (68.0%) out of 225 in CK+ video sequence in temporal segmentation evaluation, and then 97 of 153 samples (63.4%) are correctly classified in the evaluation of facial expression recognition.

When the proposed system runs continuous CK+ video sequences on *Samsung Galaxy S3*, the processing time per a frame (352×288) is 449.7 ms or 2.22 *fps* on average.

Table 4.2: 6 Emotions classification confusion matrix as result of facial expression recognition (%) by using (a) SVM and (b) HMM in FSM on *Samsung Galaxy S3*–Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa) and Surprise (Su).

(a)

	An	Di	Fe	Ha	Sa	Su
An	**50.9**	24.6	17.5	5.3	1.8	0.0
Di	6.0	**76.1**	1.5	4.5	6.0	6.0
Fe	11.9	8.5	**55.9**	1.7	10.2	11.9
Ha	0.0	0.0	0.0	**100.0**	0.0	0.0
Sa	12.3	5.3	14.0	0.0	**63.2**	5.3
Su	2.9	5.8	10.1	7.2	0.0	**73.9**

(b)

	An	Di	Fe	Ha	Sa	Su
An	**47.1**	2.0	21.6	11.8	2.0	15.7
Di	5.8	**69.2**	7.7	5.8	9.6	1.9
Fe	6.3	4.2	**62.5**	2.1	8.3	16.7
Ha	3.4	0.0	15.3	**67.8**	1.7	11.9
Sa	17.4	2.2	13.0	0.0	**50.0**	17.4
Su	3.2	0.0	6.3	1.6	1.6	**87.3**

4.2.3 Real-time Temporal Segmentation and Facial Expression Recognition on a Smartphone

For the evaluation of the proposed system in terms of temporal segmentation and facial expression recognition accuracy, we had 5 subjects to perform 6 different facial expressions while looking at the front-facing camera on the smartphone. It should be noted that they are not professional actors and despite their best effort, it was more like posed expressions, not spontaneous expressions. We considered final expressions after completing a temporal segmentation from neutral, apex to neutral state. Each participant performed the 6 different posed emotions as well as the neutral expression 10 times (thereby creating 70 facial expres-

sions per participant). Table 4.2 shows the confusion matrix for emotion classification. The average accuracies of 6 expressions in the proposed system are around 70.6% from SVM (a) and 65.2% from HMM (b) which uses a sequence of feature vector between neutral and apex states on apex state instead of SVM.

CHAPTER 5

GENDER-DRIVEN FACIAL EXPRESSION RECOGNITION ON SMARTPHONES FOR MULTIMEDIA CONTENT RECOMMENDATION SYSTEM

5.1 Analysis of Gender Difference

First, we analyzed gender difference in facial expressions. After applying ASM to 309 faces in extended Cohn-Kanade dataset (details in section 5.3.1) for obtaining landmarks on face, we calculated percentage changes of the distance of points between neutral and peak expression. We define the same points and distances as those used in (Houstis and Kiliaridis, 2009). For example, Figure 5.1 shows the points: (1) right commissure point (RCO), (2) left commissure point (LCO), (3) subnasale (Sn), (4) midpoint of the upper lip on the edge of the vermilion zone (Ulip), (5) midpoint of lower lip on the edge of the vermilion zone (Llip), (6) right inner canthus (RC), (7) left inner canthus (LC), (1)-(2) LCO_RCO, (1)-(6) RC_RCO, (4)-(5) Ulip_Llip, (3)-(4) Sn_Ulip, (3)-(5) Sn_Llip. Like the result in (Houstis and Kiliaridis, 2009), we observe that females have a pronounced horizontal movement comparing with males in LCO_RCO of happiness expression shown in Figure 5.2. On the other hand, males have a greater movement in Ulip_Llip than female. In other facial expressions we have similar results to Figure 5.2. Based on this study, we propose a gender-driven emotion recognition approach as described below.

5.2 Proposed System Architecture for Gender-driven Facial Expression Recognition

The proposed system for facial expression recognition is composed of common sequence of processes in general facial expression recognition systems, but our system has added a gender recognition process for gender-driven models. In Figure 5.3, video sequence frames coming

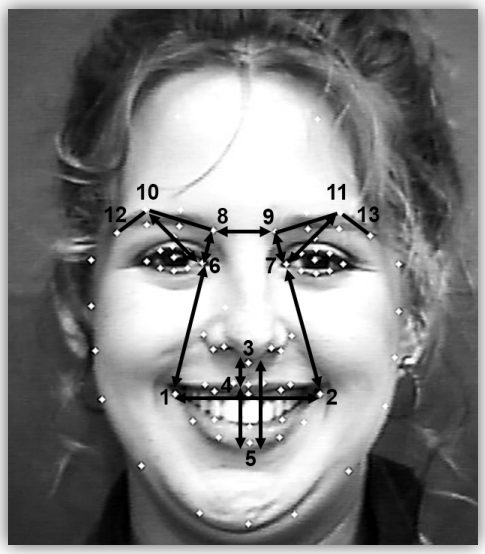


Figure 5.1: Example of landmarks by STASM (Milborrow and Nicolls, 2008) and geometric features on CK+ dataset

from a smartphone camera are processed in sequential modules such as face detection, feature extraction, neutral expression recognizer, gender recognizer, and facial expression recognizer using the ensemble learning of gender-driven SVM classifiers. For the ensemble model for facial expression recognition, three different weak classifiers (male, female and general) are separately trained and then the ensemble mode is built with a majority voting method from these weak classifiers. In the following sections, we describe a sequence of processes for the gender-driven facial expression recognition.

5.2.1 Facial Feature Extraction for Facial Expression Recognition

The face acquisition is a prerequisite for the facial expression system. Incorrect face detection results in erroneous facial expression recognition. The face detection in our system is based on Haar feature-based cascade classifier (Lienhart and Maydt, 2002). The face detector covers a wide variety of faces with angles, but works well in frontal view of faces. After a face is detected in input image, the location information of the face is passed to the next step. For facial expression recognition, facial geometric features are extracted with ASM.

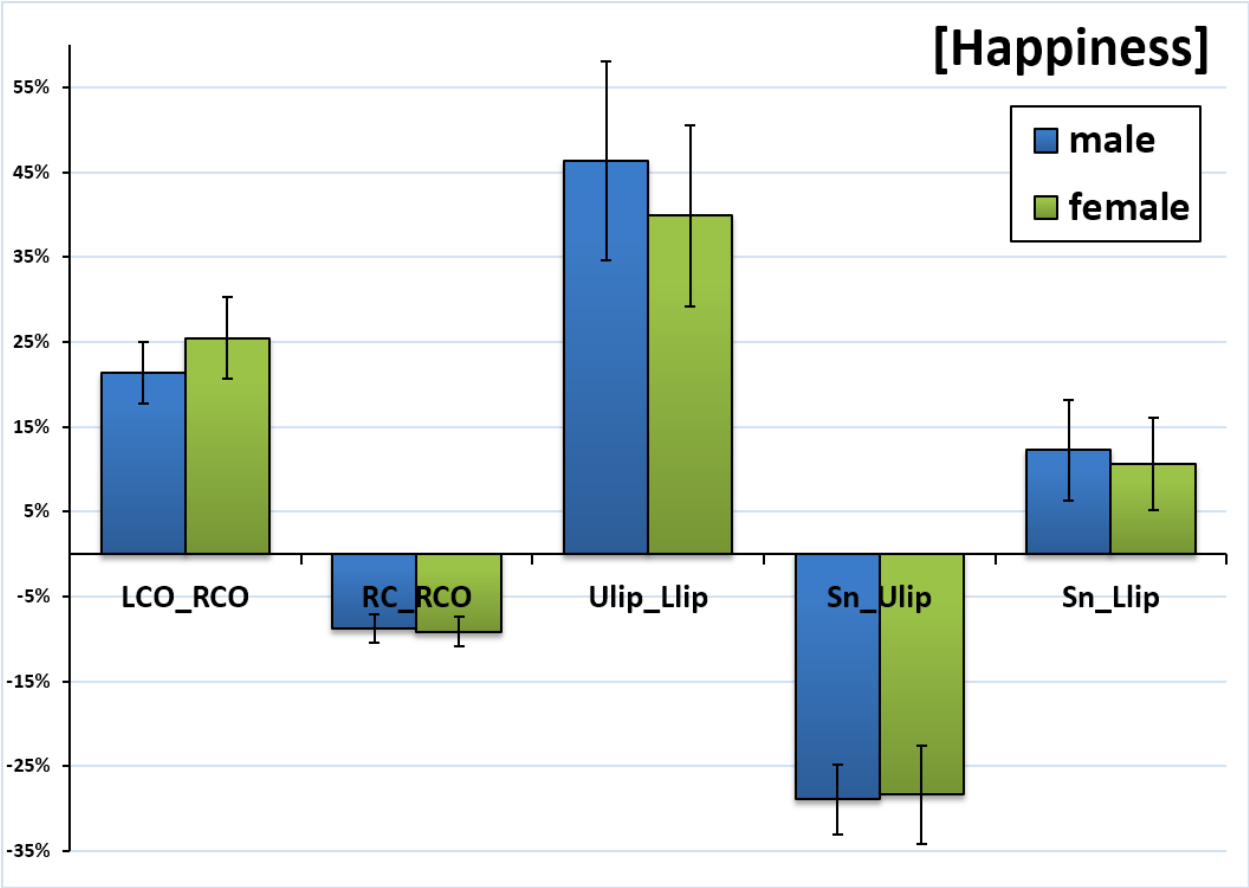


Figure 5.2: Percentage changes of the distances between neutral and happiness expression: each column shows the mean value with standard deviation, and positive percentage is an increase of the distance, while negative is a decrease by distance.

The ASM locates a template mask on the face and iteratively searches the best locations for the landmarks. Using STASM (Milborrow and Nicolls, 2008), 77 facial landmarks are located in faces. Based on the x, y coordinates of landmarks, new features such as distances between two particularly chosen markers are generated as high-level facial shape features such as LCO_RCO and Ulip_Llip. Then we finally obtain features by the displacement in percentage between neutral face and expressive face. The face images and features should be normalized geometrically for both gender recognition and facial expression recognition because all face images are not captured at the same distance from the camera, and some faces may be out of upright. For the normalization, the scale ratio and head's roll angle are

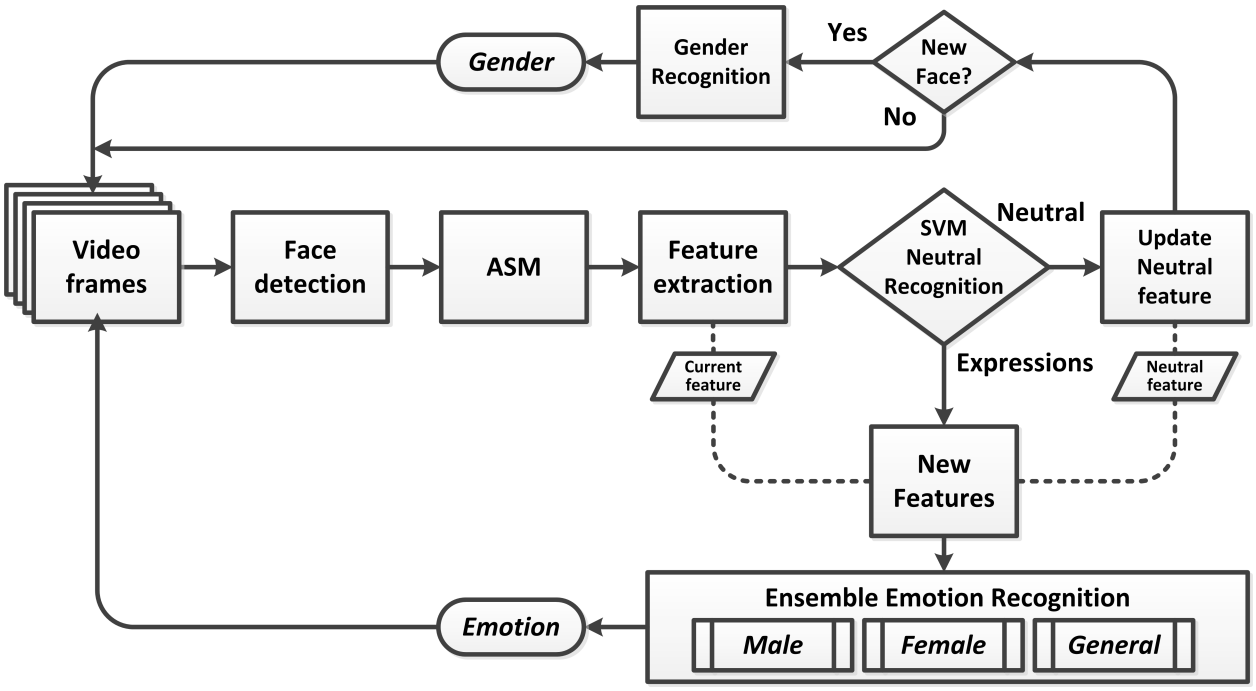


Figure 5.3: The system architecture for facial expression recognition and gender recognition modules in real-time video sequence on smartphones.

calculated from the distance and slope of two eyes. The features normalized with scale and rotation can be fed to facial expression recognition model.

5.2.2 Gender Recognition

Gender recognition is an essential part of our proposed facial expression recognition system. To build weak classifiers, we evaluated different gender recognizers in section 5.3.2. The proposed system employs a Fisherfaces recognizer for the gender recognition. Fisherfaces is the subspace representation of faces as basis vectors when Linear Discriminant Analysis (LDA) is used for face recognition (Belhumeur et al., 1997). In gender recognition module, faces normalized and cropped in the 100×100 size are used for classification of male and female by the gender recognizer based on Fisherfaces. In Figure 5.3, the gender recognition

is performed only once when a new face is detected in video sequences as well as the face has neutral expression.

5.2.3 Basic Classifier for Facial Expression Recognition

The basic classifier we employ in the proposed system is SVM that has been already popular in many pattern recognition tasks including face recognition and facial expression recognition. We used popular open source library called LIBSVM (Chang and Lin, 2011). For facial expression recognition, three SVM classifiers with radial basis function (RBF) kernel are built from different train dataset of male faces, female faces, and all male and female faces. Likewise, the SVM classifier for neutral expression recognition is built with two classes such as neutral and others from CK+ dataset. This neutral expression classifier is used to get neutral features in test stage on smartphones so that the displacement between the neutral features and current facial expression features is fed into SVM classifiers. On the other hand, in the train stage, we choose two frames such as neutral face (the first frame of a video sample) and peak expression face (the last frame of a video sample) to get the displacement as facial expression features in a video sequence sample of CK+ dataset.

5.2.4 Ensemble of Gender-driven SVM Classifiers

An ensemble technique is to produce a strong classifier by combining multiple weak classifiers. Because an ensemble is a supervised learning algorithm, it can be trained and used to make predictions. In order to get better results from an ensemble classifier, it is important to seek a significant diversity among the weak models to be combined. For examples of an ensemble method, there are bagging, boosting and stacking. We propose a facial expression recognition system to be improved by combining gender-driven weak classifiers. But each weak classifier is trained separately by subset of facial expression database including male, female, and general group and all weak classifiers participate in producing an ensemble model.

Figure 5.4: Examples of CK+ dataset: (left to right) anger, disgust, fear, happiness, sadness, and surprise expression

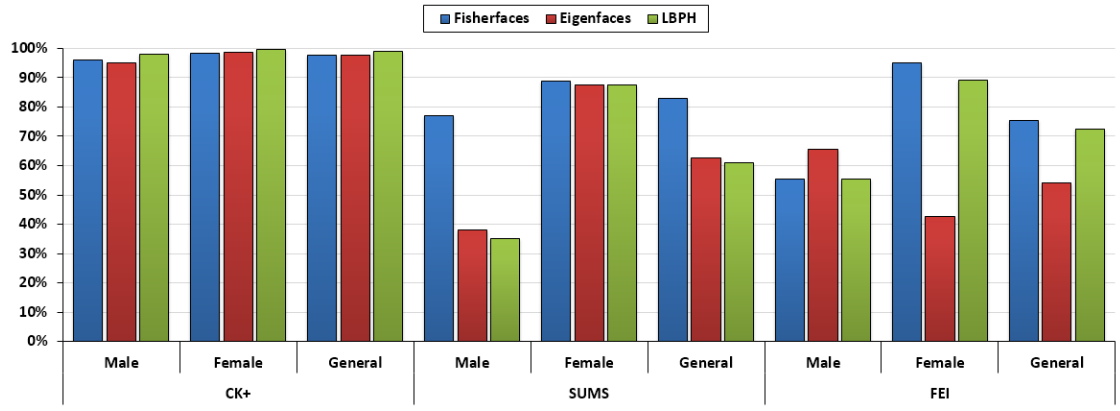
As combination rules in ensemble model, there are many methods such as majority voting, average of probabilities, product of probabilities, maximum probability, minimum probability and median. Among them, we adopt majority voting and average of probabilities due to better experimental results in section 5.3.3. In the system, the ensemble model eventually outputs one of six facial expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise). In section 5.3.3 the ensemble model with majority voting outperforms other voting methods as well as bagging and boosting on CK+ dataset for facial expression recognition.

5.3 Experimental Results

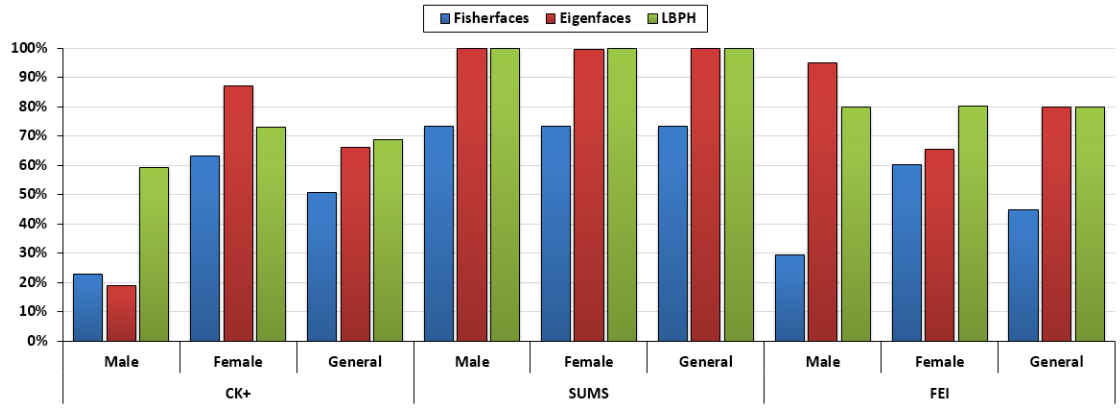
5.3.1 Facial Expression Dataset

We use the extended Cohn-Kanade (CK+) dataset (Lucey et al., 2010). CK+ includes 593 sequences from 123 subjects. Each sequence begins with the neutral frame and ends in the peak frame of facial expressions under duration about from 10 to 60 frames. The peak expression of each sequence is fully FACS coded by certified FACS coders. Figure 5.4 shows examples of CK+ dataset. For facial expression, we used 309 neutral frames and 309 peak frames in 6 emotions except contempt emotion among 7 emotions. 309 samples include 45 in anger, 59 in disgust, 25 in fear, 69 in happiness, 28 in sadness, and 83 in surprise. Among total 309 samples of 6 emotions, female samples are 211 and male samples are 98. We used total 327 peak expression samples including contempt for gender model.

Table 5.1: Gender recognition results on three different databases (CK+, SUMS, FEI) by different recognizers (Fisherfaces, Eigenfaces, Local Binary Patterns Histograms) trained by (top) CK+ dataset and (bottom) SUMS dataset



Methods (trained by CK+)	Datasets for test								
	CK+			*SUMS*			*FEI*		
	Male	Female	All	Male	Female	All	Male	Female	All
Fisherfaces	96.0%	98.2%	97.6%	77.0%	88.9%	83.0%	55.6%	95.0%	75.5%
Eigenfaces	95.0%	98.7%	97.6%	38.0%	87.4%	62.7%	65.7%	42.6%	54.0%
LBP Hist	98.0%	99.6%	99.1%	35.0%	87.4%	61.2%	55.6%	89.1%	72.5%



Methods (trained by SUMS)	Datasets for test								
	CK+			*SUMS*			*FEI*		
	Male	Female	All	Male	Female	All	Male	Female	All
Fisherfaces	22.8%	63.3%	50.8%	73.5%	73.5%	73.5%	29.3%	60.4%	45.0%
Eigenfaces	18.8%	87.2%	66.1%	100.0%	99.5%	99.8%	94.9%	65.3%	80.0%
LBP Hist	59.4%	73.0%	68.8%	100.0%	100.0%	100.0%	79.8%	80.2%	80.0%

Table 5.2: Fisherfaces based gender (male and female) classification rates on CK+ dataset (Male: 101, Female: 226).

Gender Class	Classification		
	True	False	Accuracy
Male	97	4	96.04%
Female	222	4	98.23%
Total	**319**	**8**	**97.55%**

5.3.2 Gender Recognition Result

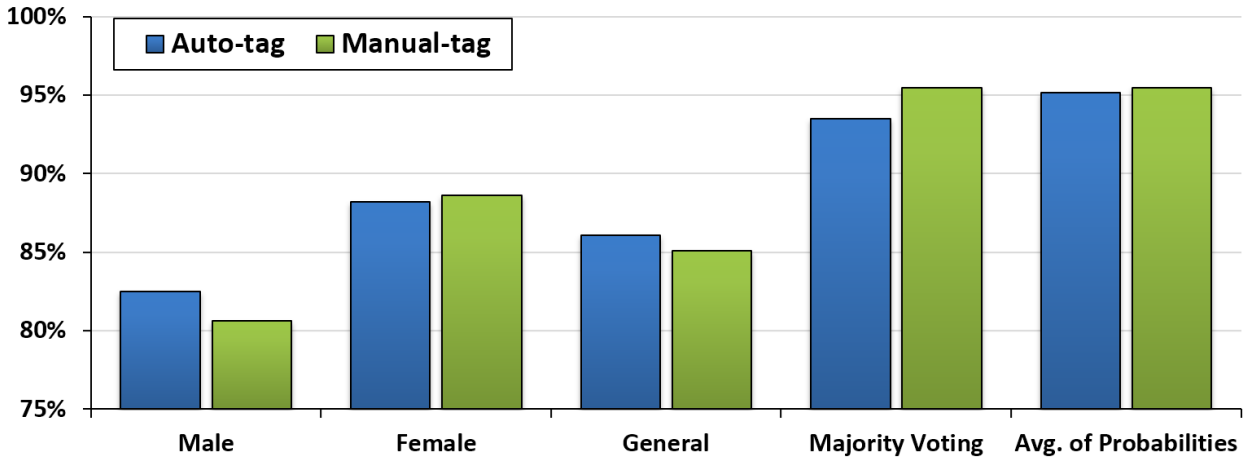
In order to use a gender classifier for the system, we compared three different face recognizers such as Fisherfaces, Eigenfaces, and Local Binary Patterns Histograms available in current OpenCV API (2.4.9). In addition, we had the evaluation of gender classifiers cross over three face databases such as CK+ (101 male and 226 female), Stanford university medical student (SUMS) (200 male, 200 female) and FEI face database (99 male, 101 female) (Thomaz and Giraldi, 2010). Table 5.1 shows the accuracy results according to the combination of three testing databases and three recognizers on CK+ and SUMS databases used for training. The classifier for gender recognition is implemented by the FisherFaceRecognizer in OpenCV API and is trained by faces in 327 peak expression frames. Table 5.2 shows the result of testing with 327 faces using gender recognizer trained from the same dataset. The dataset classified by gender recognizer are used for training the gender-driven models subsequently.

5.3.3 Facial Expression Recognition Result

First, individually weak SVM classifiers for facial expression recognition are trained by gender-specific dataset (male dataset, female dataset, and all combined dataset). In order to select the best value of parameters (γ and C) for SVM using RBF kernel, we used a grid parameter selection tool in LIBSVM. We built the most well-trained SVM models

according to gender-specific dataset using the best parameters selected by the tool. After training SVMs with the parameters, we test all weak SVM classifiers with 10 folds cross validation evaluation, and each model is tested with particular dataset according to the model. For example, the male (female) model is tested with male (female) dataset comprised of six expressions, and the general model is tested with the mixed dataset that includes both males and females showing the six different emotions. In Table 5.3, the accuracy numbers for facial expression recognition are listed. We also have different facial expression recognition

Table 5.3: Comparison of accuracy from different models: (Top) male, female, general model (mixed from both male and female), (Bottom) ensemble models (majority voting and average of probabilities) trained with dataset gender-tagged by hand and dataset automatically tagged by gender detector.



Gender Tagging Method	SVM models for Emotion Classification		
	Male base model	*Female base model*	*General base model*
Manual-Tagging	80.61%	88.63%	85.11%
Auto-Tagging	82.47%	88.21%	86.08%

Gender Tagging Method	Ensemble voting methods for Emotion Classification	
	Majority Voting	*Avg. of Probabilities*
Manual-Tagging	95.47%	95.47%
Auto-Tagging	93.53%	95.15%

results from gender datasets classified by hand and by gender detector. However, there are no significant differences between the two groups in gender tagging methods. The emotion classifier trained by female dataset has better performance than that by male dataset. We infer that facial expressions of female are more distinguishable than male facial expressions.

For ensemble models, we have experiments with average of probabilities and majority voting as combination rules. In comparison with weak models, the ensemble model results in better accuracy about 6.84-14.79%. The detailed rates such as true positive rate, false positive rate, precision, recall, etc. are shown in Table 5.4. Mostly, disgust, happiness, and surprise emotions are well classified, but fear and sadness are difficult emotions to be classified.

Table 5.5 shows the accuracy results in combination of weak classifiers and test subsets as different features set such as 154d and 13d). 154d is a feature vector as displacement of x and y coordinates of 77 markers. 13d is a feature vector of high-level features shown in Figure 5.1 and additional features. In comparison of two different features, 154d features set produces better performance on some weak classifiers and the ensemble classifier with majority voting. In Table 5.5 and Figure 5.5 we show the results with cross over evaluation between different gender datsets.

The proposed ensemble model shows better accuracy rates on average than those resulting from combination of features in (Lucey et al., 2010) as a baseline method: (the accuracy rates from the proposed ensemble model are shown inside parenthesis.) Angry–75.0% (95.6%), Disgust–94.7% (91.5%), Fear–65.2% (84.0%), Happy–100% (100%), Sadness–68.0% (89.3%), Surprised–96.0% (98.8%). Additionally Figure 5.6 shows comparison of some reported methods using static analysis approach on the same database (Shan et al., 2005; Bartlett et al., 2003; Littlewort et al., 2004; Tian, 2004).

Table 5.4: The accuracy results for facial expression recognition: (top to bottom) (1) male model, (2) female model, (3) general model, and (4) ensemble model.

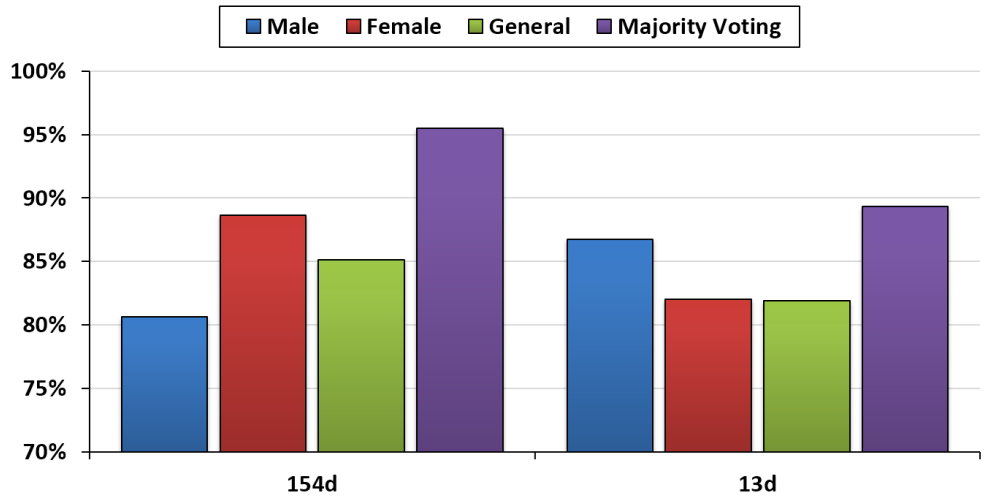
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.786	0.048	0.733	0.786	0.759	0.869	0.607
disgust	0.684	0.051	0.765	0.684	0.722	0.817	0.584
fear	0.429	0.033	0.500	0.429	0.462	0.698	0.255
happiness	1.000	0.077	0.769	1.000	0.870	0.962	0.769
sadness	0.667	0.011	0.857	0.667	0.750	0.828	0.602
surprise	0.897	0.014	0.963	0.897	0.929	0.941	0.894
weighted avg.	**0.806**	**0.040**	**0.809**	**0.806**	**0.802**	**0.883**	**0.695**

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.806	0.050	0.735	0.806	0.769	0.878	0.621
disgust	0.825	0.035	0.846	0.825	0.835	0.895	0.731
fear	0.944	0.021	0.810	0.944	0.872	0.962	0.769
happiness	0.939	0.006	0.979	0.939	0.958	0.966	0.933
sadness	0.684	0.021	0.765	0.684	0.722	0.832	0.552
surprise	0.981	0.000	1.000	0.981	0.991	0.991	0.986
weighted avg.	**0.886**	**0.019**	**0.890**	**0.886**	**0.887**	**0.934**	**0.814**

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.711	0.064	0.653	0.711	0.681	0.823	0.506
disgust	0.780	0.036	0.836	0.780	0.807	0.872	0.694
fear	0.720	0.014	0.818	0.720	0.766	0.853	0.612
happiness	0.986	0.017	0.944	0.986	0.965	0.984	0.934
sadness	0.679	0.025	0.731	0.679	0.704	0.827	0.525
surprise	0.964	0.022	0.941	0.964	0.952	0.971	0.917
weighted avg.	**0.851**	**0.029**	**0.851**	**0.851**	**0.850**	**0.911**	**0.758**

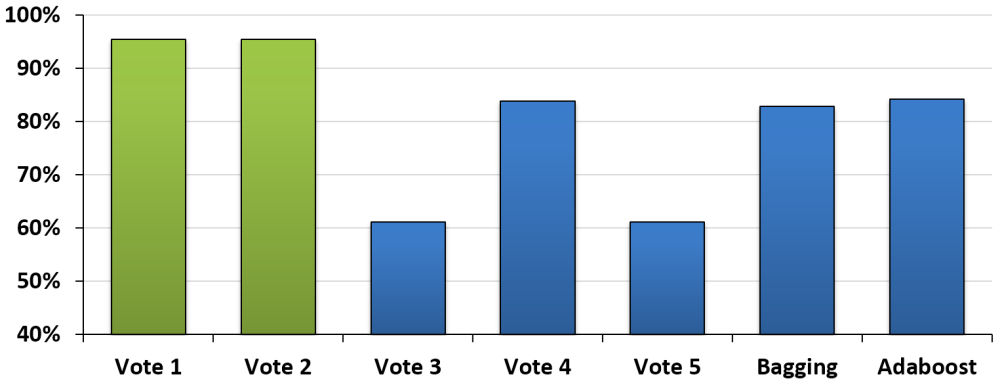
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.844	0.011	0.927	0.844	0.884	0.917	0.805
disgust	0.932	0.020	0.917	0.932	0.924	0.956	0.867
fear	0.960	0.000	1.000	0.960	0.980	0.980	0.963
happiness	1.000	0.004	0.986	1.000	0.993	0.998	0.986
sadness	0.964	0.018	0.844	0.964	0.900	0.973	0.817
surprise	0.988	0.000	1.000	0.988	0.994	0.994	0.991
weighted avg.	**0.955**	**0.008**	**0.956**	**0.955**	**0.955**	**0.973**	**0.921**

Table 5.5: Comparison of accuracy from different models: (Top) male, female, general model (mixed from both male and female), (Bottom) ensemble models with different voting methods on CK+ dataset with 154D and 13D features.



Models trained by Gender	Datasets for test (154D)			Datasets for test (13D)		
	Male	*Female*	*General**	*Male*	*Female*	*General**
Male	80.6%	61.6%	65.0%	86.7%	73.9%	82.2%
Female	76.5%	88.6%	90.6%	71.4%	82.0%	91.3%
General*	84.7%	97.6%	85.1%	83.7%	87.7%	81.9%

**** General : All (Male + Female)***



Datasets for test	Ensemble Voting Methods for Emotion Classification				
	Majority Voting	*Average of Probabilities*	*Product of Probabilities*	*Maximum of Probabilities*	*Minimum of Probabilities*
154D	95.5%	95.5%	61.2%	83.8%	61.2%
13D	89.3%	88.7%	73.5%	84.8%	73.5%

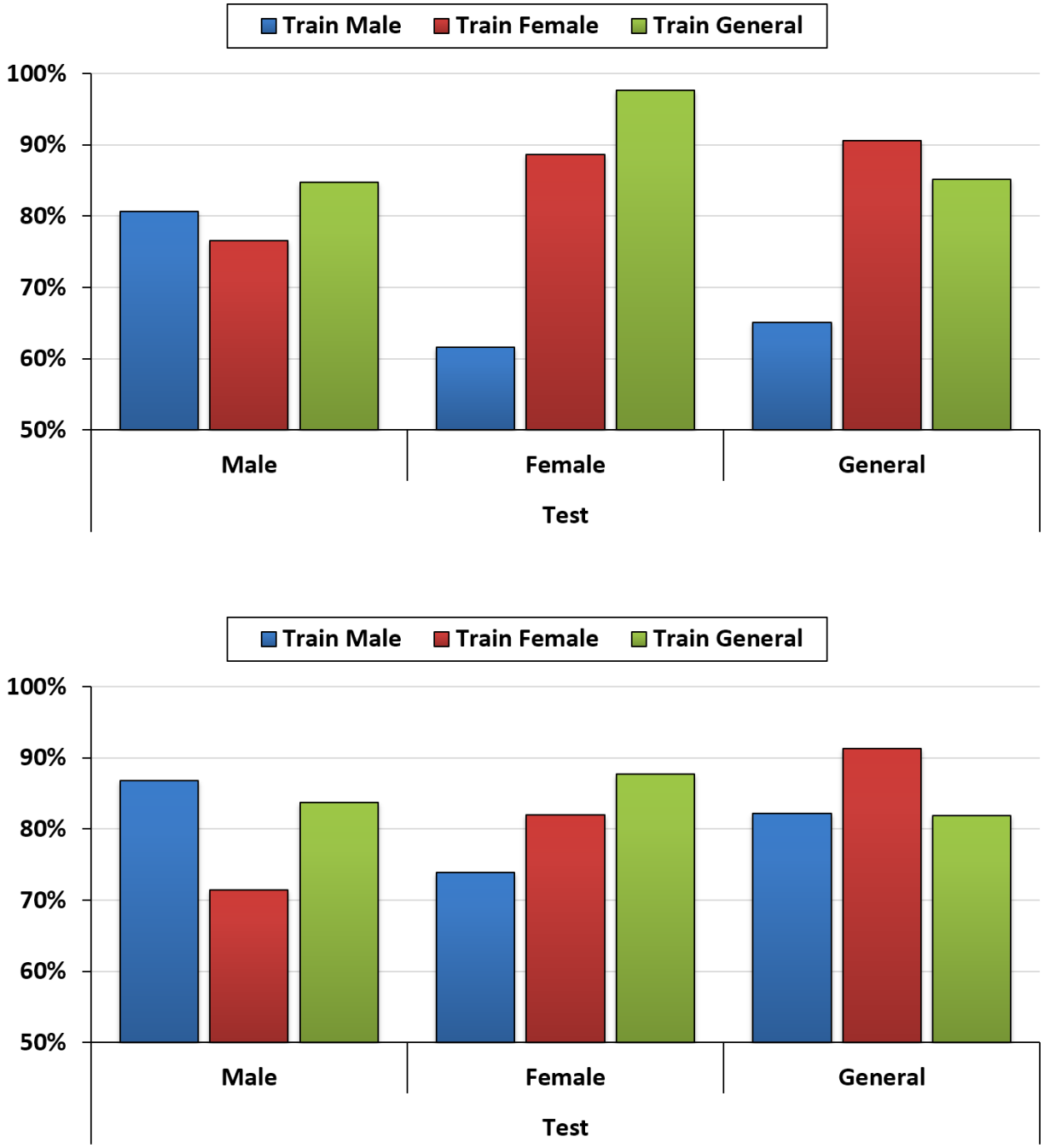


Figure 5.5: Comparison of accuracy from different models (10 folds cross validation evaluation): male, female, general model (mixed from both male and female) and ensemble model with majority voting using (Top) 154D features and (Bottom) 13D features.

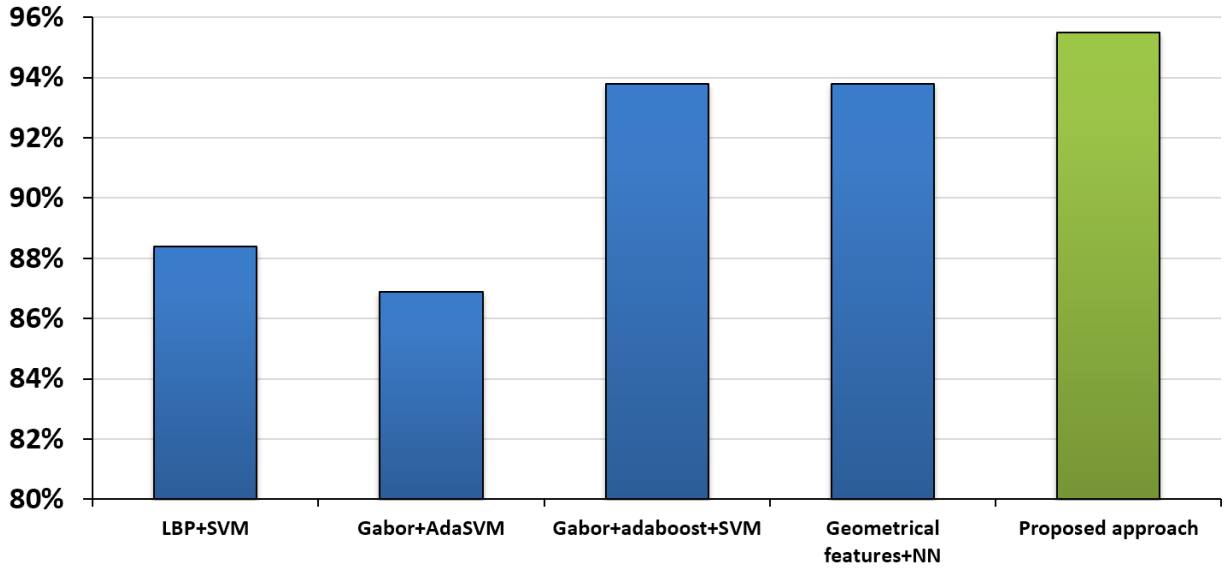


Figure 5.6: Comparison of different approaches using static analysis for facial expression recognition. LBP+SVM (Shan et al., 2005), Gabor+AdaSVM (Bartlett et al., 2003), Gabor+adaboost+SVM (Littlewort et al., 2004), Geometrical features+NN (Tian, 2004). Proposed system is Geometrical features + SVM + Ensemble with Majority voting.

5.4 Context-Sensitive Multimedia Content Recommendation System

We present a context-sensitive multimedia content recommendation system in Figure 5.7. The module for user's gender and emotion recognition runs on a smartphone. (See Figure 5.8 In order to avoid high processing time in video sequences, the gender recognition needs to be performed whenever a new face is detected in the smartphone camera. Figure 5.9 shows an example of contents such as videos, photos, and ebooks with some description. The contents of this example result from a user who is man and acts happy expression. For now, the recommendation system server performs the keyword based searching and already has configuration for similar keywords. (e.g., man, men, boys, male, etc.) For the demo application, the system server uses the YouTube Data API 2.0 to search for YouTube videos, the Flickr API for Flickr photos, and Google Books API to perform a volumes search. Because of API functional limitation for searching, current system performs a search with some combination of keywords. However, the server would be enhanced much more if the

server could search contents with statistic data. (e.g., searching sad movies women love, funny videos men in the 15 - 20 age range like, etc.) Furthermore, the recommendation system provides support for extensibility with additional information by other recognition modules as well as user's saved preference.

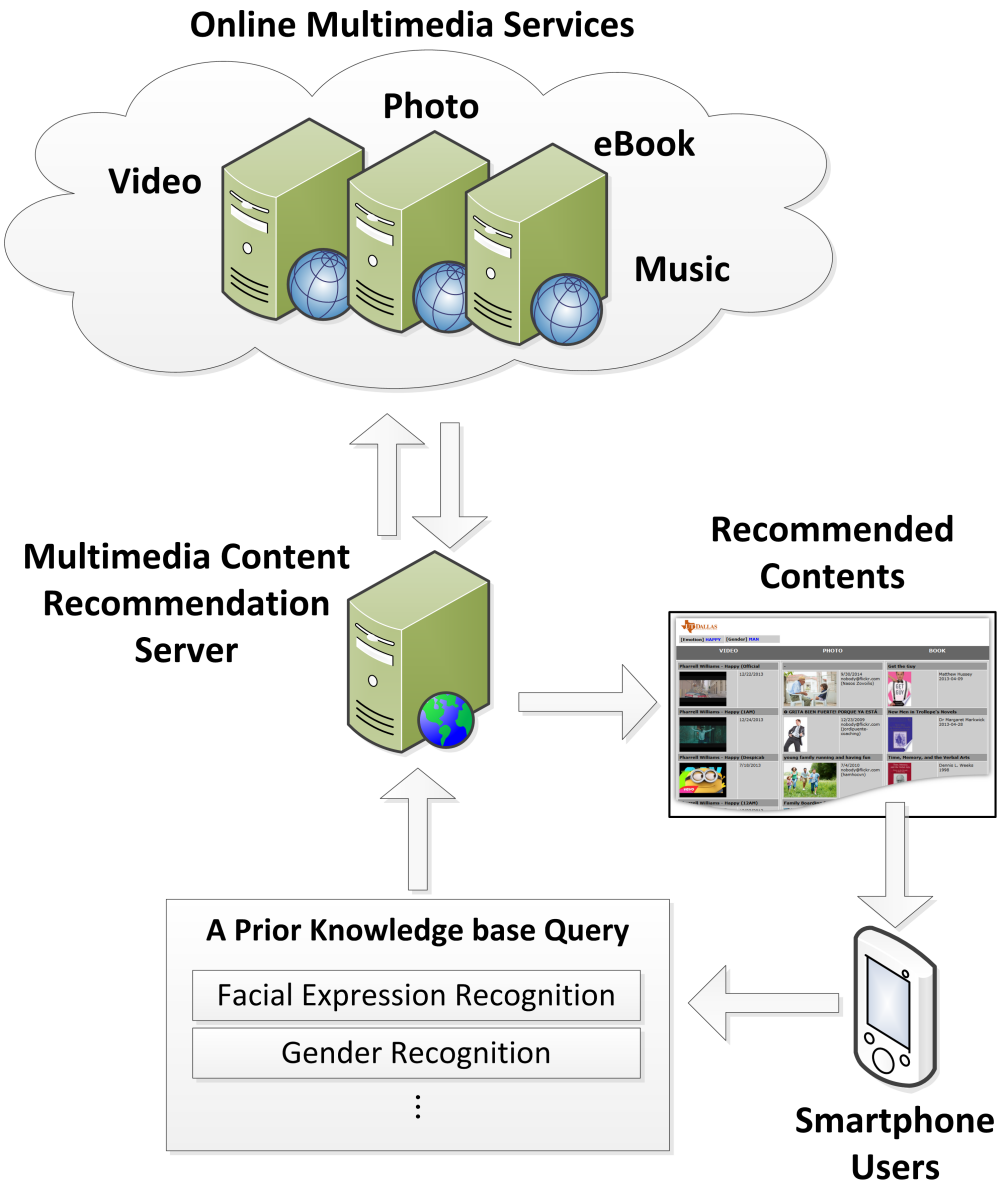


Figure 5.7: System Architecture of Context-Sensitive Multimedia Content Recommendation System.

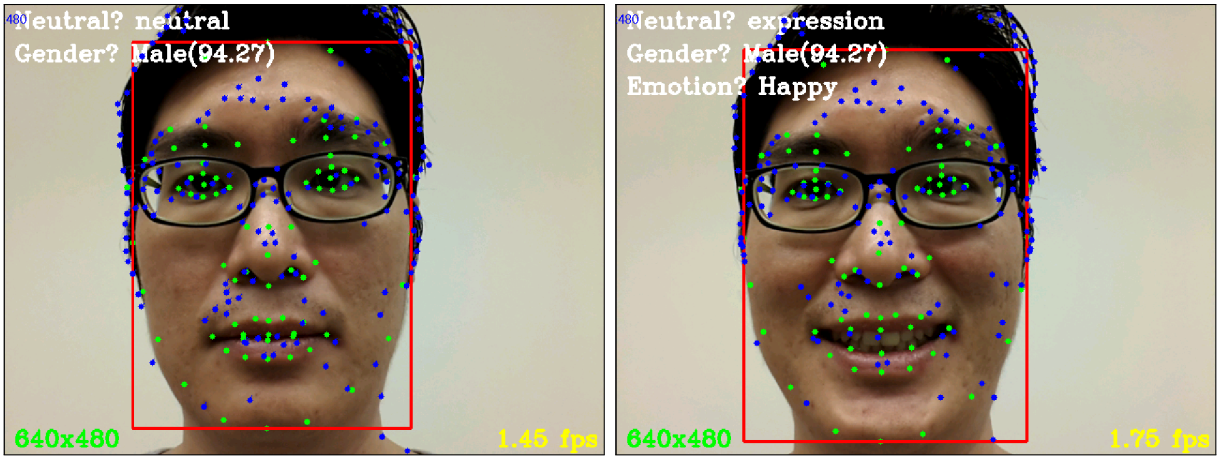


Figure 5.8: Screenshots of gender and emotion recognition on smartphones. (Left) Neutral and (Right) Happy.

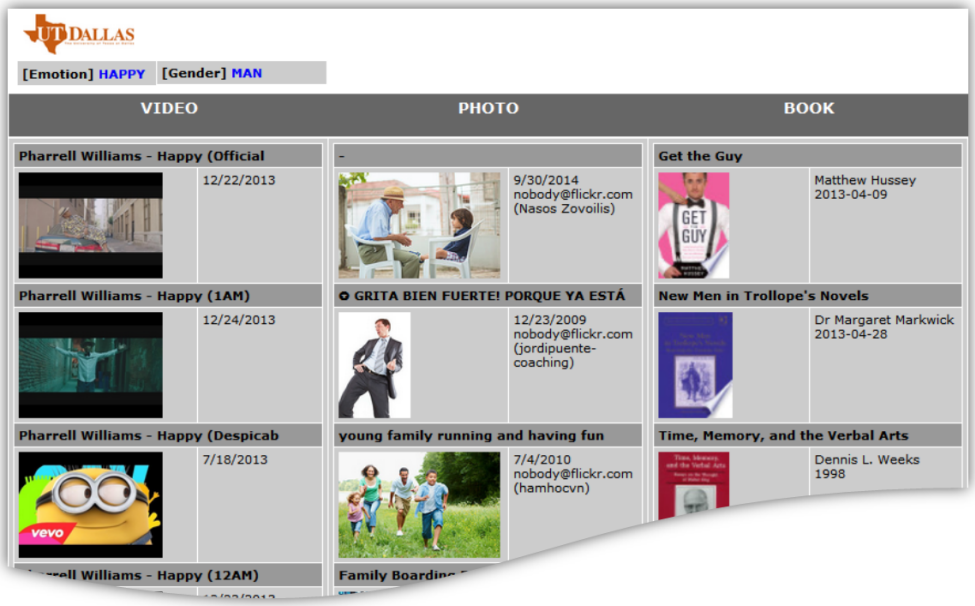


Figure 5.9: Screenshot of contents a user receives feedback from multimedia content recogmendation system. list of videos from YouTube, photos from flickr, and eBooks from Google Books.

CHAPTER 6

EXPERIMENTAL STUDY FOR EXPANDED NOISY DATASET

6.1 Proposed System Overview

In this Chapter, we present CNN models to handle extended noisy dataset for facial expression recognition. Real world data is very challenging because they are captured under a variety of uncontrolled environments and facial expressions are not unintended. We selected large scale FER2013 dataset as real world noisy data. We build a state-of-the-art Deep CNN model and train the dataset.

First of all, we evaluate small and basic CNN model on CK+ dataset, and compare the results between CNN model and SVM model trained on CK+. Moreover, a larger CNN model with deeper layers is built with FER2013 data, and we evaluate the performance for 7 classes of facial expression recognition. Lastly, we evaluate CNN model with cross dataset (CK+ and FER2013).

6.2 Example of Noisy Images and Videos

In reality, real world data for facial expression recognition differ greatly from research data that is controlled by lab environment.

Natural and spontaneous facial expression: In many cases, facial expressions of subjects may be reflected by the experimenter's intention, but the natural facial expressions in everyday life are not posed and not all the same according to emotions and they may have a wide variety of differences in intensity. The captured face is not always in an upright, frontal position, but has head pose in various angles. Depending on the emotional expression, or depending on the intensity of the emotions, the head may be rolled up or bowed. In addition to facial expressions, hand gestures also mask the face. It is very difficult to obtain such

as natural facial expressions data in laboratory environment, and it is clear that the system that does not sufficiently reflect such data will easily fail under practical conditions.

Images under various circumstances: The facial features extracted from image are also heavily influenced under different circumstances. For example, lighting and background around the face have a lot of influence on the facial feature itself. However, existing systems are not robust by lack of dataset in outdoor/indoor or day/night time environments. Therefore, research is needed to find a better way to accept and process noisy image and video data in more realistic environments.

6.3 State-of-the-art Deep Convolutional Neural Networks (CNN)

Deep learning has recently become a big trend in many fields. One of the biggest achievements in Deep Convolutional Neural Networks (CNN) is in a field of computer vision and machine learning. There are two major contributing factors to the development of Deep Learning. First, recently, since computer performance, especially Graphics Processing Unit (GPU) performance, has been further enhanced, the myriad of parameters that Deep learning needs to calculate can be handled quickly by these GPU capabilities. Secondly, digital devices such as digial camera, action camera (e.g., *GoPro*) and smartphone are able to generate numerous image / video data anytime and anywhere easily shared by Social networking service (SNS) or several internal channels by using advanced Internet speed and cheap cloud services. Therefore, the most important key element for Deep learning training, innumerable dataset can be provided. In addition, good open source library and resources such as Tensorflow and Microsoft COCO dataset (Lin et al., 2014) are provided for development of Deep learning. This helps many researchers to be more involved and to make better outcomes. In this chapter, we compare the results of previous experiments and recent experiments using state-of-the-art deep learning, and present a new direction.

6.4 Experimental Results

The FER2013 dataset, which is a lot of noisy image data with various subject and uncontrolled facial expressions, compared to CK + with focus on existing research, is introduced in section 6.4.1. In section 6.4.2, we compare CNN model with CK + dataset with SVM results for FER. Section 6.4.3 shows the CNN model with FER2013 dataset. In section 6.4.4, the CNN models trained on CK+ and FER2013 are evaluated cross dataset.

6.4.1 FER2013 Dataset

We use the extended FER2013 dataset (Goodfellow et al., 2015) as noisy dataset for extended experiment. FER2013 dataset includes 48×48 grayscale face images – 28,709 facial expression images for the training set, 3,589 images for public test set. The FER2013 dataset was provided for Facial Expression Recognition Challenge of ICML 2013 Workshop on Representation Learning. FER2013 dataset has 7 categories (Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral). Compared to CK+, FER2013 dataset consists of the wider range of unposed images, and differences in head pose or angle, expression intensity, occlusion, and illumination under realistic environment.

6.4.2 Evaluation of Facial Expression Recognition Accuracy of CNN model and SVM on CK+ Dataset

We build baseline CNN model for CK+ dataset. (See Table 6.1). The most important consideration for this baseline CNN model layers is to devise a simple network so that even small data sets like CK+ can be fully trained. Therefore, the 64x64 size input image is fed into the baseline network which consists of 4 Conv2D layers, 3 MaxPooling layers, 2 Dropout layers with 0.2 rate, and Rectified Linear Unit (ReLU) and softmax as activation function. During all the CNN training, we increase the input dataset by data augmentation

Table 6.1: CNN model on CK+ dataset. The accuracy for 6 facial expression classes (anger, disgust, fear, happiness, sadness, surprise) is 96.67% on CK+ training dataset

Layer	Output Shape	Param #
Conv2D	(None, 64, 64, 32)	320
Conv2D	(None, 62, 62, 32)	9,248
MaxPooling	(None, 31, 31, 32)	0
Dropout	(None, 31, 31, 32)	0
Conv2D	(None, 31, 31, 64)	18,496
MaxPooling	(None, 15, 15, 64)	0
Conv2D	(None, 15, 15, 128)	73,856
MaxPooling	(None, 7, 7, 128)	0
Flatten	(None, 6272)	0
Dense	(None, 128)	802,944
Dropout	(None, 128)	0
Dense	(None, 6)	774

such as rotation, width/height shift, shear, zoom, and horizontal flip. After training this CNN model on CK+ dataset, we obtain 96.67% accuracy for 6 emotion classes on CK+ training set. Figure 6.1 shows CNN model accuracy over training time. (Left graph is model accuracy on train set, while right graph results from validation set). Because both training set and validation set accuracy curves are converging similarly, this model does not seem overfitted. Table 6.2 (a) is a confusion matrix evaluated by CK+ test set (20% of CK+ all dataset) with 6 categories. The average accuracy is 93.7%. Table 6.2 (b) is another confusion matrix evaluated by CK+ test set with 7 categories with neutral. The average accuracy is 90.4%. Compared to the accuracy with 6 classes, the reason for the drop in accuracy of 7 classes is that the neutral faces are misclassified into anger and sad classes. Also this CNN result is very interesting that CNN baseline model seems pretty good

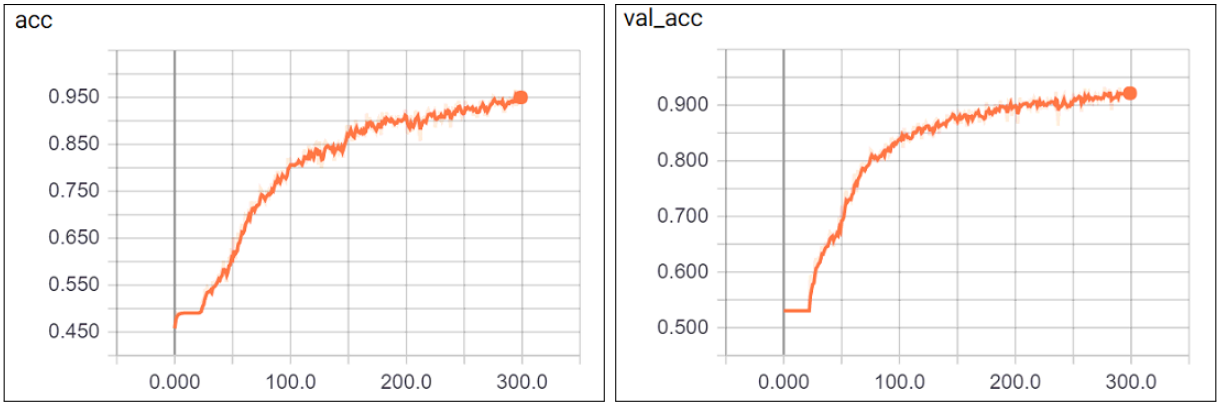


Figure 6.1: CNN model accuracy over training period (300 epochs) on CK+ train (Left) and validation (Right) dataset.

in comparison with SVM model's result, 85.8% with 13D features from section 3.3.4. We presume that data augmentation for training set is a great help to improve the result.

6.4.3 Evaluation of Facial Expression Recognition Accuracy of CNN model on FER2013 Dataset

We build CNN model for FER2013 dataset. (Table 6.3 shows the example for the model layers). These model layers are more complicated than baseline CNN model for CK+ dataset in Table 6.1. These model layers are initially borrowed by the basic network in (Yu and Zhang, 2015), but more simplified by reducing 5 convolutional layers into 3 layers, and use max and average pooling instead of stochastic pooling. 48x48 size input image is fed into the network which consists of 3 Conv2D layers, 1 MaxPooling layer, 2 AveragePooling layers, 2 Dropout layers with 0.2 rate, and Parametric Rectified Linear Units (PReLU) and softmax as activation function. Like CNN training in section 6.4.2, we also increase the input dataset by data augmentation such as rotation, width/height shift, shear, zoom, and horizontal flip. We train CNN model on FER2013 dataset. Figure 6.2 shows model accuracy over training period. Left graph is model accuracy on train set, while right graph results from validation set. We evaluate CNN model on FER2013 dataset. Table 6.4 shows confusion matrix with overall 63.8% accuracy for 7 emotion classes including neutral on FER2013 test dataset.

Table 6.2: Emotions classification confusion matrix as result of facial expression recognition (%) by using CNN Model on CK+ test dataset (20% of all dataset): (a) overall average accuracy 93.7% on 6 categories (b) 90.4% on 7 categories–Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su) and Neutral (Ne).

(a)

	An	Di	Fe	Ha	Sa	Su
An	**88.9**	11.1	0.0	0.0	0.0	0.0
Di	0.0	**83.3**	8.3	8.3	0.0	0.0
Fe	0.0	0.0	**80.0**	20.0	0.0	0.0
Ha	0.0	0.0	0.0	**100.0**	0.0	0.0
Sa	0.0	0.0	0.0	0.0	**100.0**	0.0
Su	0.0	0.0	0.0	0.0	0.0	**100.0**

(b)

	An	Di	Fe	Ha	Sa	Su	Ne
An	**33.3**	22.2	11.1	0.0	0.0	0.0	33.3
Di	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0
Fe	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0
Ha	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0
Sa	0.0	0.0	0.0	0.0	**50.0**	0.0	50
Su	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0
Ne	0.0	0.0	4.8	0.0	0.0	0.0	**95.2**

6.4.4 Evaluation of Facial Expression Recognition Accuracy of CNN model Cross-Dataset

In this section, we evaluate and compare CNN models trained from both CK+ and FER2013 cross dataset. Table 6.5 (a) shows confusion matrix with overall 26.6% accuracy of FER2013 public test dataset for 6 emotion classes not including neutral category on CNN model trained

Table 6.3: CNN model on FER2013 dataset. The accuracy for 7 facial expression classes (anger, disgust, fear, happiness, sadness, surprise, neutral) is 63.8% on FER2013 test dataset

Layer	Output Shape	Param #
Conv2D	(None, 44, 44, 64)	1,664
PReLU	(None, 44, 44, 64)	123,904
ZeroPadding2D	(None, 48, 48, 64)	0
MaxPooling2D	(None, 22, 22, 64)	0
ZeroPadding2D	(None, 24, 24, 64)	0
Conv2D	(None, 22, 22, 64)	36,928
PReLU	(None, 22, 22, 64)	30,976
ZeroPadding2D	(None, 24, 24, 64)	0
Conv2D	(None, 22, 22, 64)	36,928
PReLU	(None, 22, 22, 64)	30,976
AveragePooling2D	(None, 10, 10, 64)	0
ZeroPadding2D	(None, 12, 12, 64)	0
AveragePooling2D	(None, 5, 5, 64)	0
Flatten	(None, 1600)	0
Dense	(None, 1024)	1,639,424
PReLU	(None, 1024)	1,024
Dropout	(None, 1024)	0
Dense	(None, 1024)	1,049,600
PReLU	(None, 1024)	1,024
Dropout	(None, 1024)	0
Dense	(None, 7)	7,175
Activation	(None, 7)	0

by CK+. Table 6.5 (b) is a confusion matrix with average 25.7% accuracy on CK+ CNN model for 7 categories with FER2013 public test dataset. On the other hand, Table 6.6 shows overall 75.9% accuracy with CK+ test set on CNN model trained by FER2013. Comparing

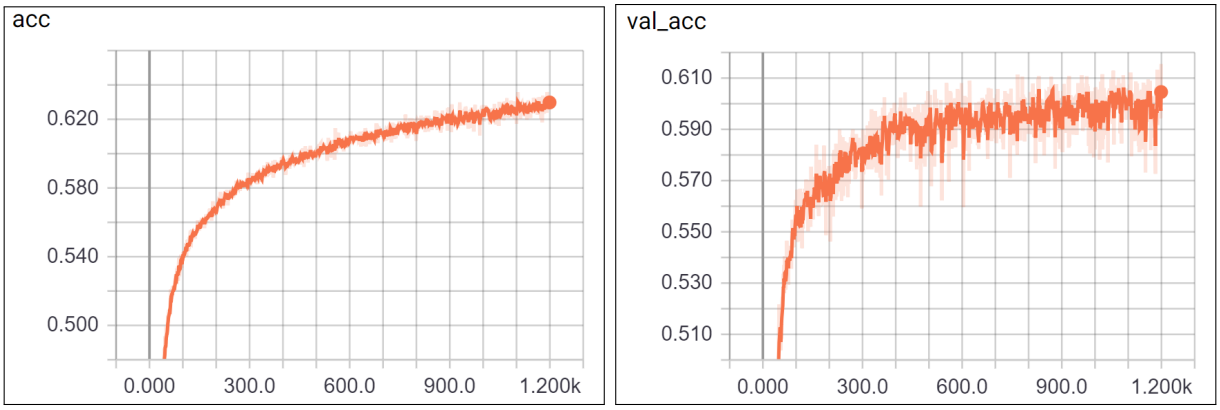


Figure 6.2: CNN model accuracy over training period (1,200 epochs) on FER2013 train (Left) and validation (Right) dataset.

Table 6.4: Emotions classification confusion matrix as result of facial expression recognition (%) by using CNN Model. The accuracy for 7 facial expression classes (anger, disgust, fear, happiness, sadness, surprise, neutral) is 63.8% on FER2013 test dataset

	An	Di	Fe	Ha	Sa	Su	Ne
An	**49.5**	0.6	11.1	5.6	12.0	4.1	17.1
Di	30.4	**41.1**	1.8	1.8	12.5	0.0	12.5
Fe	7.7	0.4	**39.7**	2.6	17.1	13.3	19.2
Ha	1.3	0.1	2.6	**84.7**	2.4	2.7	6.3
Sa	7.8	0.2	11.3	3.7	**47.8**	2.9	26.3
Su	1.9	0.2	7.5	2.9	2.2	**80.5**	4.8
Ne	3.8	0.0	5.3	7.7	9.7	1.8	**71.7**

between Table 6.5 (b) and Table 6.6, CNN model trained by FER2013 outperforms model by CK+.

6.5 Conclusion

With the development of technology at great speed, we can easily gather tremendous amount of image/video data recorded by a variety of environments which are very different from clean but relatively very small amount of dataset controlled under lab environment. In this

Table 6.5: Emotions classification confusion matrix as result of facial expression recognition (%) for FER2013 test dataset by using CNN model trained on CK+ dataset: (a) overall average accuracy 26.6% on 6 categories (b) 25.7% on 7 categories–Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su) and Neutral (Ne).

(a)

	An	Di	Fe	Ha	Sa	Su
An	**3.0**	25.9	17.1	9.0	12.2	32.8
Di	1.8	**41.1**	12.5	1.8	12.5	30.4
Fe	1.0	12.3	**25.6**	4.4	14.5	42.1
Ha	0.4	10.7	26.0	**27.0**	3.6	32.2
Sa	3.7	18.1	21.9	7.7	**13.5**	35.2
Su	0.7	5.8	12.8	3.1	5.8	**71.8**

(b)

	An	Di	Fe	Ha	Sa	Su	Ne
An	**0.6**	36.8	8.1	13.3	0.0	31.0	10.1
Di	0.0	**53.6**	8.9	16.1	0.0	12.5	8.9
Fe	0.0	24.4	**11.9**	13.9	0.2	37.7	11.9
Ha	0.0	20.6	8.6	**46.0**	0.0	20.9	3.9
Sa	0.0	29.7	12.9	15.9	**0.3**	28.5	12.7
Su	0.0	12.0	4.1	5.8	0.0	**68.7**	9.4
Ne	0.2	19.9	15.0	15.7	0.2	27.7	**21.4**

Chapter, we explored FER2013 dataset as such noisy dataset with unposed natural facial expression. The large-scale FER2013 dataset is adequate and sufficient for use as as training data in CNN. We compared accuracy performance of CNN model with SVM model with 13D geometric features (introduced in Chapter 3) based on CK+ dataset. In contrast with small and limited amount of feature set which has been already normalized per size, angle and

Table 6.6: Emotions classification confusion matrix as result of facial expression recognition (%) for CK+ dataset by using CNN model trained on FER2013 dataset: The overall average accuracy is 75.9% on 7 categories–Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su) and Neutral (Ne).

	An	Di	Fe	Ha	Sa	Su	Ne
An	**22.2**	0.0	11.1	0.0	15.6	2.2	48.9
Di	67.8	**3.4**	0.0	5.1	3.4	0.0	20.3
Fe	8.0	0.0	**40.0**	8.0	16.0	16.0	12.0
Ha	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0
Sa	0.0	0.0	10.7	0.0	**42.9**	0.0	46.4
Su	2.4	0.0	0.0	1.2	0.0	**89.2**	7.2
Ne	0.3	0.0	0.3	1.6	2.3	1.0	**94.5**

position, we were able to use data augmentation such as rotation, width/height shift, shear, zoom, and horizontal flip in order to increase amount of training dataset for CNN during training. The result of small and simple CNN model (93.7%) outperforms SVM model with 13D geometric features (85.8%) Moreover, deeper CNN model works well in Public Test dataset of FER2013. (63.8% for 7classes) Moverover, we evaluated the performance of CNN models cross dataset. The CNN model built by FER2013 dataset works well on CK+ dataset, but CNN model by CK+ dataset does not perform well the FER 2013 dataset. Therefore, we can say the CNN model is more appropriate to handle large amount of noisy dataset in real world like FER2013 dataset as well as even small a mount of dataset like CK+ dataset.

We can expect more and more noisy but natural dataset will be collected because of advance of technology sharing numerous image/video via platforms such as high speed network, cloud services, and SNS. Then we can improve CNN model with innumerable amount of natural facial expression data.

CHAPTER 7

CONCLUSION

In Chapter 3, we present the real-time video based facial expression recognition system good enough to run on commercial off-the-shelf smartphones. In order to find appropriate features in terms of both accuracy and speed performance for smartphones with relatively low hardware specification, we evaluated the performance of SVM classifiers built by 13D geometrical features or LBP appearance features. Moreover the proposed system employs a reliable SVM, well-known for facial expression recognition with high performance of accuracy and speed. The dynamic features generated by the displacement between neutral frame and non-neutral frames are simple and effective enough to be used in real-time sequence video on smartphones. Although ASM module searching faces and fitting landmarks on detected face is a costly process in terms of time, the ASM provides accurate facial features and the system can process the facial expression recognition task in video sequences. We show the proposed system runs with good performances in terms of speed and recognition accuracy by evaluating the computation time and accuracy offline dataset (CK+) and online on various commercial off-the-shelf smartphones in Section 3.3.

In Chapter 4, we have proposed an efficient approach for real-time temporal video segmentation for facial expression recognition running on smartphones. We show experimental results that the proposed system employing FSM instead of a sliding window or sampling time for temporal segmentation runs on a low-powered smartphone with well-balanced performance between speed and recognition accuracy, compared to existing system on high-powered computer. Although the FSM approach can work with any other classifiers for emotion recognition on apex state, we employ SVM models with dynamic features on non frame-by-frame basis because of merits of a static classifier being less sensitive to temporal patterns on different people as well as good practical results well known to the computer learning community. In our approach, SVM prediction does not always happen every frame

and it is carried out on apex states so that the mobile system lightens the computation burden. Moreover, SVM as a static classifier is reinforced with dynamic features made of displacement between facial features on neutral and apex states. As seen in comparison of SVM and HMM in experimental result, SVM is more efficient than HMM and more suitable to our system for smartphones. Our efficient approach runs smoothly on the smartphone with good performances in terms of speed and recognition accuracy shown in section 4.2.

In Chapter 5, we propose a novel approaches for facial expression recognition based on gender-driven ensemble model and introduce a prototype of context-sensitive multimedia content recommendation system with the ability to detect and respond to emotions of users. All the core modules such as facial expression recognition and gender recognition perform on smartphones themselves, and then query based on user's information including current emotion state and prior knowledge are sent to a remote web server of the recommendation system. Eventually users can receive a list of multimedia contents such as videos, photos, ebooks from the server. The priori knowledge helps to recognize accurate emotion as well as find appropriate content for the user. We have experimental results that a gender-driven facial expression recognition can outperform other state-of-the-art approaches. For future work, the ensemble model can be added to models with occluded faces or other prior knowledge like age and race.

In Chapter 6, we present CNN models with large scale FER2013 dataset captured under noisy realistic environment. Compared to overall accuracy of 13D SVM model proposed in Chapter 3, CNN model outperforms overall accuracy. We employ data augmentation for training dataset in order to avoid overfitting. Also though the cross dataset evaluation, we prove that it is hard to apply the model trained by controlled facial expressions in CK+ dataset to FER2013's samples with noisy natural facial expression. On the other hand, CNN models based on FER2013 dataset produces good result with CK+. If we obtain more datset and make deeper model, the CNN model can be improved in the future.

CHAPTER 8

FUTURE WORK

- Adaptive thresholding for the state transition in the proposed FSM for temporal segmentation. In the proposed FSM presented in Chapter 4, we used fixed values such as 'a', 'b', and '$threshold_b$' (See Figure 4.2). Particularly '$threshold_b$' for tagging as 'dynamic' is compared to '$distance_i$', the distance of optical flow features between two video frames. Therefore, while a dynamic feature in slow motion has short distance between two frames of video sequences, fast motion makes the same dynamic feature large displacement at the same FPS. Conversely lower FPS can produce dynamic features with big displacement, but faster FPS has features with small movement between frames. In the previous work, we have determined the thresholds with the experimental result on testing smartphones. However it may not good in all the conditions where the proposed system runs at different FPSs or people make facial expressions at different speeds. Therefore the system should calculate the threshold according to different FPSs so that the system for temporal video segmentation for facial expression recognition can automatically use different thresholds on varying environments such as video resolutions, FPSs, and hardware performance of smartphones.

- Open mouth causing unexpected behaviors on FSM. Another problem in the proposed system in Chapter 4 is the mouth status in FSM model. The mouth status is a straightforward condition for state transition on facial expressions with open mouth. The mouth status is determined by the distance between middle points of upper and lower lips. Sometimes we found some unexpected erroneous state transitions on FSM by the following causes: mouth markers incorrectly tracked by ASM and wrong threshold for recognizing open mouth and open mouth when talking. Therefore, the system needs to be improved by modified design.

- To improve recognition performance on facial expressions in the presence of occlusions, the presence of occlusions is one of the challenging problems in computer vision. Likewise the system for smartphones can be easily exposed to varying environment including occlusions. By making the facial expression recognition system robust to the presence of occlusions, the system can improve the recognition performance on facial expressions even under unexpected situations on smartphones.

- To improve performance on CNN models for facial expression recognition on a lack of data amount from VR/MR environment, more dataset is required to train without overfitting. Therefore, it is possible to employ other approach like Generative adversarial networks (GANs) as data augmentation. Also it would be better to make CNN model with deeper layers with more annotated dataset.

REFERENCES

Anand, B., B. Navathe, S. Velusamy, H. Kannan, A. Sharma, and V. Gopalakrishnan (2012, Jan). Beyond touch: Natural interactions using facial expressions. In *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, pp. 255–259.

Arapakis, I., I. Konstas, and J. M. Jose (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, New York, NY, USA, pp. 461–470. ACM.

Bartlett, M. S., B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T. J. Sejnowski, and J. R. Movellan (2001). Automatic analysis of spontaneous facial behavior. In *IN THE EYE OF THE*. Oxford University Press.

Bartlett, M. S., G. Littlewort, I. Fasel, and J. R. Movellan (2003, June). Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, Volume 5, pp. 53–53.

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

Bisio, I., A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone (2013, Dec). Gender-driven emotion recognition through speech signals for ambient intelligence applications. *Emerging Topics in Computing, IEEE Transactions on 1*(2), 244–257.

Busso, C., Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *in Sixth International Conference on Multimodal Interfaces ICMI 2004*, pp. 205–211. ACM Press.

Chang, C.-C. and C.-J. Lin (2011, May). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. 2*(3), 27:1–27:27.

Chen, L. and T. Huang (2000). Emotional expressions in audiovisual human computer interaction. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, Volume 1, pp. 423–426 vol.1.

Chen, Y.-W. and C.-J. Lin (2006). Combining svms with various feature selection strategies. In I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh (Eds.), *Feature Extraction*, Volume 207 of *Studies in Fuzziness and Soft Computing*, pp. 315–324. Springer Berlin Heidelberg.

Cohen, I., N. Sebe, L. Chen, A. Garg, and T. S. Huang (2003). Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pp. 160–187.

Cootes, T. F., C. J. Taylor, D. H. Cooper, and J. Graham (1995, January). Active shape models–their training and application. *Comput. Vis. Image Underst. 61*(1), 38–59.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray. The original was published 1898 by Appleton, New York. Reprinted 1965 by the University of Chicago Press, Chicago and London,.

Edwards, G., C. Taylor, and T. Cootes (1998). Interpreting face images using active appearance models. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 300–305.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychology Bulletin 115*(2), 268–287.

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

Fasel, B. and J. Luettin (2003). Automatic facial expression analysis: a survey. *Pattern Recognition 36*(1), 259 – 275.

Feng, X., A. Hadid, and M. Pietikainen (2004). A coarse-to-fine classification scheme for facial expression recognition. In A. Campilho and M. Kamel (Eds.), *Image Analysis and Recognition*, Volume 3212 of *Lecture Notes in Computer Science*, pp. 668–675. Springer Berlin Heidelberg.

Goodfellow, I. J., D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio (2015). Challenges in representation learning: A report on three machine learning contests. *Neural Networks 64*, 59 – 63. Special Issue on "Deep Learning of Representations".

Houstis, O. and S. Kiliaridis (2009). Gender and age differences in facial expressions. *Eur J Orthod 31*(5), 459–66.

Huang, T. S., L. S. Chen, H. Tao, T. Miyasato, and R. Nakatsu (1998). Bimodal emotion recognition by man and machine. In *ATR Workshop on Virtual Communication Environments*.

Jo, G.-S., I.-H. Choi, and Y.-G. Kim (2011). Robust facial expression recognition against illumination variation appeared in mobile environment. In *Proceedings of the 2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering*, CNSI '11, Washington, DC, USA, pp. 10–13. IEEE Computer Society.

Lien, J. J.-J., T. Kanade, J. F. Cohn, and C.-C. Li (2000). Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems 31*(3), 131–146.

Lienhart, R. and J. Maydt (2002). An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, Volume 1, pp. I–900–I–903 vol.1.

Lin, T., M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft COCO: common objects in context. *CoRR abs/1405.0312*.

Lisetti, C. L. and D. E. Rumelhart (1998). Facial expression recognition using a neural network. In *Proceedings of the Eleventh International Florida Artificial Intelligence Research Society Conference*, pp. 328–332. AAAI Press.

Littlewort, G., M. Bartlett, I. Fasel, J. Susskind, and J. Movellan (2004, June). Dynamics of facial expression extracted automatically from video. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pp. 80–80.

Littlewort, G., I. Fasel, M. S. Bartlett, and J. R. Movellan (2002). Fully automatic coding of basic expressions from video. Technical report, Tech. rep.(2002) U of Calif., S.Diego, INC MPLab.

Lucas, B. D. and T. Kanade (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, San Francisco, CA, USA, pp. 674–679. Morgan Kaufmann Publishers Inc.

Lucey, P., J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101.

Lyons, M., S. Akamatsu, M. Kamachi, and J. Gyoba (1998, Apr). Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 200–205.

Mariappan, M., M. Suk, and B. Prabhakaran (2012, Dec). Facefetch: A user emotion driven multimedia content recommendation system based on facial expression recognition. In *Multimedia (ISM), 2012 IEEE International Symposium on*, pp. 84–87.

Milborrow, S. and F. Nicolls (2008). Locating facial features with an extended active shape model. In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, ECCV '08, Berlin, Heidelberg, pp. 504–513. Springer-Verlag.

Ojala, T., M. Pietikainen, and D. Harwood (1994, Oct). Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, Volume 1, pp. 582–585 vol.1.

Padgett, C. and G. W. Cottrell (1996). Representing face images for emotion classification. In M. Mozer, M. I. Jordan, and T. Petsche (Eds.), *NIPS*, pp. 894–900. MIT Press.

Pantic, M. and I. Patras (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 36*(2), 433–449.

Pantic, M. and L. Rothkrantz (2000). Expert system for automatic analysis of facial expressions. *Image and Vision Computing 18*(11), 881 – 905.

Shan, C., S. Gong, and P. W. McOwan (2005, Sept). Robust facial expression recognition using local binary patterns. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, Volume 2, pp. II–370–3.

Shan, C., S. Gong, and P. W. McOwan (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing 27*(6), 803 – 816.

Suk, M. and B. Prabhakaran (2014, June). Real-time mobile facial expression recognition system - a case study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Thomaz, C. E. and G. A. Giraldi (2010, June). A new ranking method for principal components analysis and its application to face image analysis. *Image Vision Comput. 28*(6), 902–913.

Tian, Y.-L. (2004, June). Evaluation of face resolution for expression analysis. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pp. 82–82.

Valenti, R., A. Jaimes, and N. Sebe (2010). Sonify your face: Facial expressions for sound generation. In *Proceedings of the International Conference on Multimedia*, MM '10, New York, NY, USA, pp. 1363–1372. ACM.

Valstar, M. and M. Pantic (2006). Fully automatic facial action unit detection and temporal analysis. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pp. 149–149.

Valstar, M. F. and M. Pantic (2007). Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In M. Lew, N. Sebe, T. Huang, and E. Bakker (Eds.), *Human-Computer Interaction*, Volume 4796 of *Lecture Notes in Computer Science*, pp. 118–127. Springer Berlin Heidelberg.

Yu, Z. and C. Zhang (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, New York, NY, USA, pp. 435–442. ACM.

Zhao, S., H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji (2011). Video indexing and recommendation based on affective analysis of viewers. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, New York, NY, USA, pp. 1473–1476. ACM.

Zhao, W., X. Wang, and Y. Wang (2010). Automated sleep quality measurement using eeg signal: First step towards a domain specific music recommendation system. In *Proceedings of the International Conference on Multimedia*, MM '10, New York, NY, USA, pp. 1079–1082. ACM.

BIOGRAPHICAL SKETCH

Myunghoon Suk received a BS degree in mechanical engineering from Sungkyunkwan University, South Korea, in 1999. He entered the Digital Media Lab at Information and Communications University (ICU), and graduated with a MS degree from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2005. While he worked for Digital Media Lab and Ubiquitous Computing Lab, IBM Korea, he developed a strong interest in the HCI and Computer Vision, and he decided to study abroad to learn more in this field. In the Multimedia Systems and Networks lab at The University of Texas at Dallas, he worked with human motion recognition using motion capture data and video data and facial expression recognition on smartphones. He has a well-balanced experience in academic research and industry. During the PhD studies, as an intern, he worked for Image Vision Labs and Samsung Research America in Texas. He worked for Image Vision Labs as a scientific staff/software engineer and has been working for Topaz Labs in Addison, Texas as a software engineer.

CURRICULUM VITAE

Myung Hoon Suk

August 31, 2018

Contact Information:

Department of Computer Engineering
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080-3021, U.S.A.

Email: mhsuk@utdallas.edu

Educational History:

B.S., Mechanical Engineering, SungKyunKwan University, 1999
M.S., Digital Media, Korea Advanced Institute of Science and Technology, 2005
Ph.D., Computer Engineering, University of Texas at Dallas, 2018

Study of Real-time Facial Expression Recognition on Noisy Images and Videos
Ph.D. Dissertation
Computer Engineering Department, University of Texas at Dallas
Advisor: Dr. B. Prabhakaran

Employment History:

Senior Software Engineer, Topaz Labs, LLC, Addison, TX, November 2016 – present
Software Engineer, Image Vision Labs, INC, Anna, TX, January 2015 – October 2016
Research Assistant and Teaching Assistant, University of Texas at Dallas, January 2008 – December 2014
Software Engineer Intern, Samsung Research America, Richardson, TX, June 2014 – July 2014
Scientific Staff Intern, Image Vision Labs, INC, Anna, TX, June 2012 – August 2013
Software Developer, EzTogether, INC, Korea, February 2007 – June 2007
Software Developer, Ubiquitous Computing Lab, IBM Korea, September 2005 – February 2006
Software Developer, AILeaders, INC, Korea, December 1999 – January 2003

Publications:

1. **Myunghoon Suk**, B. Prabhakaran, "Real-time Facial Expression Recognition on Smartphones," Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, pp.1054,1059, January 2015.

2. **Myunghoon Suk**, B. Prabhakaran, "Real-Time Mobile Facial Expression Recognition System - A Case Study," Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on , pp.132,137, June 2014.

3. Mahesh Mariappan, **Myunghoon Suk** and Balakrishnan Prabhakaran, "Facial Expression Recognition Using Dual Layer Hierarchical SVM Ensemble Classification," Proceedings of IEEE International Symposium on Multimedia (ISM 2012), Irvine, CA, December 2012.

4. Mahesh Mariappan, **Myunghoon Suk** and Balakrishnan Prabhakaran, "FaceFetch: A User Emotion Driven Multimedia Content Recommendation System Based on Facial Expression Recognition," Proceedings of IEEE International Symposium on Multimedia (ISM 2012), Irvine, CA, December 2012.

5. **Myunghoon Suk**, Ashok Ramadass, Yohan Jin, B. Prabhakaran, "Video Human Motion Recognition Using Knowledge-based Hybrid Method Based On Hidden Markov Model," ACM Transactions on Intelligent Systems and Technology, Volume 3 Issue 3, Article No. 42, May 2012.

6. **Myunghoon Suk**, Ashok Ramadass, Yohan Jin, B. Prabhakaran, Multimedia, "Video Human Motion Recognition Using Knowledge-Based Hybrid Method," International Symposium on, pp. 65-72, 2010 IEEE International Symposium on Multimedia, 2010.

7. Ashok Ramadass, **Myunghoon Suk**, Balakrishnan Prabhakaran, "Feature extraction method for video based human action recognitions: extended optical flow algorithm," accepted by the 35th ICASSP, Dallas, Texas, USA, March 14-19, 2010.

8. Yohan Jin, **Myunghoon Suk**, and B.Prabhakaran, "3D Human Motion Control Through Refined Video Gesture Annotation," Chapter in Edited Book, Handbook of Digital Media in Entertainment and Arts, Springer, 2009.

9. Yohan Jin, **Myunghoon Suk**, B. Prabhakaran, Bhavani Thuraisingham, "Synthesize Virtual World Motions From 2D Video Recognition," UTD CS Technical Report (UTDCS-20-08), 2008.