Techniques for Facial Expression Recognition using the Kinect

Sherin F. Aly

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

A. Lynn Abbott, Chair
Dhruv Batra
Michael S. Hsiao
Denis Gracanin
Marwan A. Torki

September 23, 2016
Blacksburg, Virginia

Keywords: Facial Expression Recognition, Kinect, Action Units, FACS, EMFACS

Techniques for Facial Expression Recognition using the Kinect

Sherin F. Aly

(ABSTRACT)

Facial expressions convey non-verbal cues. Humans use facial expressions to show emotions, which play an important role in interpersonal relations and can be of use in many applications involving psychology, human-computer interaction, health care, e-commerce, and many others. Although humans recognize facial expressions in a scene with little or no effort, reliable expression recognition by machine is still a challenging problem.

Automatic facial expression recognition (FER) has several related problems: face detection, face representation, extraction of the facial expression information, and classification of expressions, particularly under conditions of input data variability such as illumination and pose variation. A system that performs these operations accurately and in realtime would be a major step forward in achieving a human-like interaction between the man and machine.

This document introduces novel approaches for the automatic recognition of the basic facial expressions, namely, happiness, surprise, sadness, fear, disgust, anger, and neutral using relatively low-resolution noisy sensor such as the Microsoft Kinect. Such sensors are capable of fast data collection, but the low-resolution noisy data present unique challenges when identifying subtle changes in appearance. This dissertation will present the work that has been done to address these challenges and the corresponding results.

The lack of Kinect-based FER datasets motivated this work to build two Kinect-based RGBD+time FER datasets that include facial expressions of adults and children. To the best of our knowledge, they are the first FER-oriented datasets that include children. Availability of children data is important for research focused on children (e.g., psychology studies on facial expressions of children with autism), and also allows researchers to do deeper studies on automatic FER by analyzing possible differences between data coming from adults and children.

The key contributions of this dissertation are both empirical and theoretical. The empirical contributions include the design and successful test of three FER systems that outperform existing FER systems either when tested on public datasets or in realtime. One proposed approach automatically tunes itself to the given 3D data by identifying the best distance metric that maximizes the system accuracy. Compared to traditional approaches where a fixed distance metric is employed for all classes, the presented adaptive approach had better recognition accuracy especially in non-frontal poses. Another proposed system combines high dimensional feature vectors extracted from 2D and 3D modalities via a novel fusion technique. This system achieved 80% accuracy which outperforms the state of the art on the public VT-KFER dataset by more than 13%. The third proposed system has been designed and successfully tested to recognize the six basic expressions plus neutral in realtime using only 3D data captured by the Kinect. When tested on a public FER dataset, it achieved 67% (7% higher than other 3D-based FER systems) in multi-class mode and 89% (i.e., 9% higher than the state of the art) in binary mode. When the system was tested in realtime on

20 children, it achieved over 73% on a reduced set of expressions. To the best of our knowledge, this is the first known system that has been tested on relatively large dataset of children in realtime. The theoretical contributions include 1) the development of a novel feature selection approach that ranks the features based on their class separability, and 2) the development of the Dual Kernel Discriminant Analysis (DKDA) feature fusion algorithm. This later approach addresses the problem of fusing high dimensional noisy data that are highly nonlinear distributed.

This work received support from the Egyptian Ministry of Higher Education, VT-MENA program, and National Institute of Health (NIH) award 1R03HD081070-01A1.

Techniques for Facial Expression Recognition using the Kinect

Sherin F. Aly

(General Audience Abstract)

One of the most expressive way humans display emotions is through facial expressions. The recognition of facial expressions is considered one of the primary tools used to understand the feelings and intentions of others. Humans detect and interpret faces and facial expressions in a scene with little or no effort, in a way that it has been argued that it may be universal. However, developing an automated system that accurately accomplishes facial expression recognition is more challenging and is still an open problem. It is not difficult to understand why facial expression recognition is a challenging problem. Human faces are capable of expressing a wide array of emotions. Recognition of even a small set of expressions, say happiness, surprise, anger, disgust, fear, and sadness, is a difficult problem due to the wide variations of the same expression among different people. In working toward automatic Facial Expression Recognition (FER), psychologists and engineers alike have tried to analyze and characterize facial expressions in an attempt to understand and categorize these expressions. Several researchers have considered the development of systems that can perform FER automatically whether using 2D images or videos. However, these systems inherently impose constraints on illumination, image resolution, and head orientation. Some of these constraints can be relaxed through the use of three-dimensional (3D) sensing systems. Among existing 3D sensing systems, the Microsoft Kinect system is notable because it is low in cost. It is also a relatively fast sensor, and it has been proven to be effective in real-time applications. However, Kinect imposes significant limitations to build effective FER systems. This is mainly because of its relatively low resolution, compared to other 3D sensing techniques and the noisy data it produces. Therefore, very few researchers have considered the Kinect for the purpose of FER. This dissertation considers new, comprehensive systems for automatic facial expression recognition that can accommodate the low-resolution data from the Kinect sensor. Moreover, through collaboration with some Psychology researchers, we built the first facial expression recognition dataset that include spontaneous and acted facial expressions recorded for 32 subjects including children. With the availability of children data, deeper studies focused focused on children can be conducted (e.g., psychology studies on facial expressions of children with autism).

This work received support from the Egyptian Ministry of Higher Education, VT-MENA program, and National Institute of Health (NIH) award 1R03HD081070-01A1.

Dedication

To my greatest supporter, my mom. Your love gave me the strength to do this.

To my first hero, my dad. Everything I have or became is because of you and this is a little thing to make you proud of me.

To my loving husband. I would have not been able to finish this if it were not for your love, help, encouragement, and sacrifices.

It is all for her, my sunshine, Natalia Lili.

To my beloved ones, Rania and Omar.

To my role model aunt Laila and my supportive uncle Fathy Khedr.

To the wonderful Constanza. I will always be in your debt.

To my beautiful family and extended family in Egypt, UAE, Saudi Arabia, Finland, Monaco, Argentina, and Colombia.

Acknowledgments

I am very grateful to my advisor, Prof. A. Lynn Abbott for giving me the opportunity to work in such an interesting area of research and for providing all the resources necessary for me to accomplish this work. I cannot find words to express my appreciation to the great efforts, help, and support he provided me during the entire time of my study at Virginia Tech. His passion for research and intense commitment to his work have been a source of inspiration to me. I will keep all his teachings for the rest of my career.

Many thanks to Dr. Marwan A. Torki for being my VT-MENA advisor. I appreciate his support, encouragement, and valuable advice during the last few years. Thanks for suggesting working with the Kinect.

I would like to express my gratitude to Drs. Michael S. Hsiao, Dhruv Batra, and Denis Gracanin for being in my dissertation committee, and for their valuable input on how to improve this work.

Thanks to the Egyptian government, the VT-MENA program, and the National Institute of Health who supported my work.

Thanks to Dr. Amira Youssef for helping defining this research topic, as well as for her collaboration. Thanks also for providing some of the videos utilized in the realtime system.

Thanks to Dr. Susan White and all the Child Study Center team members at Virginia Tech for welcoming me in their session rooms, helping me conduct my experiments, building VT-KFER and designing the realtime system.

I would like to thank the Center for Embedded Systems for Critical Applications (CESCA) professors for the annual CESCA day event that helped enhance my presentation skills and gave me the opportunity to show and discuss my research with a broader audience. I am very proud of being a CESCA member.

I would like to show my gratitude to Dr. Devi Parikh. Several times I asked to step by her office for brainstorming, and she has been always welcoming and inspiring.

Thanks to all my colleagues who helped in data collection including Imen Tanniche, Kannikha Kolandaivelu, Farzaneh Tb, Delasa Aghamirzaie, Mahi Abdelbar and other VT-MENA students in Blacksburg campus. Thanks to Eman Badr for bringing her kids for data collection. Thanks to Ahmed Ibrahim for his collaboration during my early work with the Kinect and for bringing his

kids for data collection.

Thanks to Evelyn McKeon and Dr. Donald McKeon for their friendship and the love they surrounded me with, which helped me getting over my home sickness feeling. I always felt like home in their house. Thanks for the beautiful flowers you, Evelyn, gave me when I was most in need.

I would like to thank my loving family for the boundless love, patience, and unconditional support they gave to me. Special thanks to my mom and my mother in law for taking care of Natalia day and night to give me a chance to work and sleep. Without your help, I definitely would not have been able to finish by now. Thanks to my aunt Laila and uncle Fathy Khedr for following up my progress and encouraging me as a daughter of their own.

Contents

List of Figures

List of Tables

Abbreviations

ARA Average Recognition Accuracy

ASD Autism Spectrum Disorder

AU(s) Action Unit(s)

DKDA Dual Kernel Discriminant Analysis

EMFACS Emotion Facial Action Coding System

FACS Facial Action Coding System

FER Facial Expression Recognition

FS Feature Selection

HM(s) Heat Map(s)

HOG Histogram of Oriented Gradient

KDA Kernel Discriminant Analysis

LBP Local Binary Pattern

LDA Linear Discriminant Analysis

PCA Principal Component Analysis

ROC Operating Characteristic

ROI region of interest

SP Spatial Pyramid

SP-ULBP Spatial Pyramid-Uniform Local Binary Pattern

ULBP Uniform Local Binary Pattern

Chapter 1

Introduction

1.1 Motivation

Facial expressions are fundamental for effective interpersonal human relations, communications, and survival. Emotions are powerful means for people to interact and clarify agreement or disagreement. Facial expressions are a naturally preeminent way to regulate interactions with the environment and with other people. Typically, the ability to discriminate certain expressions develops very early in childhood, with ability to distinguish basic emotions from static cues appearing as early as 3 months [10, 11]. We, humans, recognize facial expressions almost with no effort.

Automatic Facial Expression Recognition (FER) has interested many researchers due to its various purposes and applications in video games [12, 13], health-care [14, 15], driver safety [16, 17], deceit detection [18], human-computer interaction and robots [19–21], psychology [22], commerce [23–26], affective computing [22, 27], and even advertising [23].

Early studies of facial expressions were pioneered by Ekman in the early 70s [28, 29]. He proposed six prototypical emotions, namely, anger, fear, disgust, happiness, sadness and surprise, each of which corresponds to a unique facial expression. Although the uniqueness and the universality of these emotions are still under dispute within the psychology community [30], most researchers, including us, follow this classification scheme. This dissertation presents various techniques for Kinect-based FER to recognize these universal/basic six facial expressions, plus neutral. Examples of the six basic expressions, taken from [1], are shown in Figure 1.1.

It has been argued that the ability to recognize basic emotions is universal due to the fact that humans detect and interpret faces and facial expressions in a scene with little or no effort. However, developing an automated system that accurately performs facial expression recognition is rather difficult and is still an open problem.

There are two main reasons that explain why, in general, automatic FER is a challenging problem: 1) human faces can express a wide arrange of emotions (even more than the so called six basic

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

Figure 1.1: The 6 basic facial expressions. Pictures are taken from the NimStim Face Stimulus Set [1]. This work is focused on the design of FER systems to detect these expressions, plus the neutral, using the Microsoft Kinect.

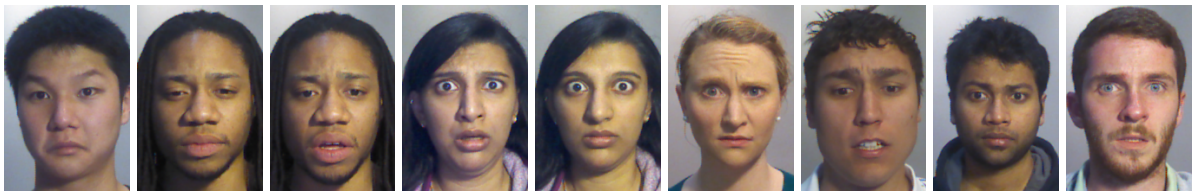


Figure 1.2: Some variations for fear expression over different subjects. Pictures are taken from the VT-KFER dataset [2].

emotions, and including even combinations of emotions such as surprise and happiness); and 2) even for the same emotion, there is high variability between people in how they express it. As an example of the latter case, figure 1.2 shows this variability for the expression corresponding to "fear" among different subjects and even for the same subject.

1.2 Data Sensors for Facial Expression Recognition: State of the Art

Several researchers have considered the development of systems that can perform FER automatically using 2D images or videos [31–36]. Comprehensive surveys in this area include those by Fasel and Luettin [37], Pantic et al. [38] and Zeng et al. [39]. While these 2D facial expression recognition systems have achieved remarkable performance in heavily conditioned environments (e.g., with particular illumination conditions), 2D face expression recognition still faces important challenges such as illumination, image resolution, and pose variations. Some of these issues can be addressed through the use of three-dimensional (3D) sensing systems.

Regarding the use of 3D sensing systems, several researchers have recently started to explore how this dimensional extension can improve FER. They found that the 3D geometry contains ample information about human facial expression [40]. Some researchers have successfully used 3D data based on 2D images, such as multiple views [41] or 3D models for facial expression

Figure 1.3: Example RGB image (left) and corresponding range image (right) captured by the Kinect. This example shows the artifacts in depth maps resulting from glasses in structured light methods.

analysis [42–45]. These models can alleviate the problems caused by different head poses to a certain degree with the assistance of a 3D model or with multiple views of the face. However, since their 3D face models were generated from the 2D images/videos, the ability to handle large head pose variation is inherently limited. A recent comprehensive survey about static and dynamic 3D FER systems, including state of the art employed sensors, can be found in [46]. Other FER surveys can be found in [3, 47, 48].

The acquisition technique used for capturing 3D facial expressions data is significant. Different equipment result in different types of data and thus different development systems and techniques [46]. In addition, the equipment used can affect the level of bother on the subject (e.g., they have to sit for long time without moving), thereby changing their behavior significantly. A variety of devices and techniques have been employed for 3D facial expression data acquisition, including the use of single image reconstruction, photometric stereo, multi-view stereo, and structured light [46].

Structured light method is among the most widely used technologies for acquisition of 3D facial surface [2, 49–54]. In order to extract shape information, one or more encoded light patterns are projected onto the scene and then the deformation of the pattern on the objects' surfaces is measured in order to extract shape information. Unfortunately, the acquired range images can contain holes where points are missing, as well as small artifacts, mainly in areas that cannot be reached by the projected light that are either highly reflective (e.g., eye-glasses) or poorly reflective (e.g., hair, glasses frames, beard). Figure 1.3 shows an example of these artifacts in a range image when the subject is wearing glasses. A range image is a 2D image where the pixel values correspond to distances. If the sensor that is used to produce the range image is properly calibrated the pixel values can be given directly in physical units, such as meters. In the given presented image, and for the rest of this document, the lighter the gray color, the closer the distance to the camera.

Examples of structured light-based sensors are the Minolta Vivid 900/910 series [55], the Inspeck-Mega Capturor II 3D [56] and the Microsoft Kinect™ v1.0 camera [57]. The Kinect sensor is

notable because of its low cost. It is also a relatively fast sensor (at 30 frames per second), and it has been proven to be effective in realtime activities that involve full-body sensing.

A newer version of the Microsoft Kinect sensor has recently been released, the Kinect v2.0. The main difference between v1.0 and v2.0 is in the technology that the image sensor relies on. In v2.0, the image sensor is based on the time-of-flight (TOF) technique. The basic idea behind the TOF cameras is that they measure the time it takes for pulses of laser light to travel from a laser projector, to a target surface, and then back to an image sensor. One of the main advantages of the Kinect v2.0 sensor over v1.0 is that the former can be used in sunlight while the later cannot. This is because Kinect v2.0 has built-in ambient light rejection where each pixel individually detects when that pixel is over saturated with incoming ambient light, and it then resets the pixel in the middle of an exposure. The Kinect 1.0 sensor has no means of rejecting ambient light, and thus, cannot be used in environments prone to near-infrared light sources (i.e. sunlight). This new sensor enables employing the Kinect for outdoor applications that were not possible with the old Kinect [58].

The presented work in this dissertation is utilizing the Kinect v1.0 sensor and next section is providing more details about its specifications.

1.3 The Kinect v1.0 Sensor

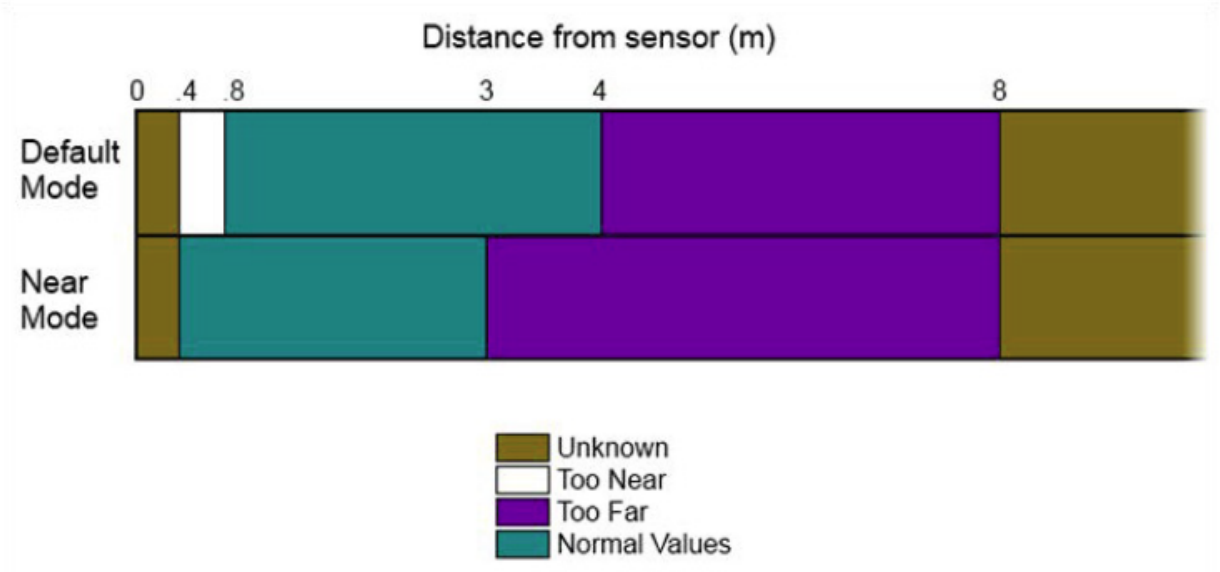
The Microsoft Kinect™ v1.0 sensor is a structured light-based dynamic infrared scanner capable of estimating the 3D geometry of the acquired scene at 30 fps with spatial resolution of 640×480 pixels [59]. The RGB video stream uses 8-bit resolution of 640×480 pixels with a Bayer color filter. The hardware is also capable of resolutions up to 1280×1024, but at 12 frame per second (fps) acquisition rate. The Kinect relies on triangulation between a near-infrared camera and a near-infrared laser source to perform structured-light. Because structured light-based systems use invisible infrared light, there is no perceivable disturbance to the environment during recording. The Kinect device is capable of acquiring a stream of depth images, thus opening the way to dynamic analysis of temporal sequences in 3D. It was commercialized by Microsoft for the Xbox 360 video game console and was developed by PrimeSense with both proprietary and open source drivers.

With respect to other 3D scanning devices, the Kinect is characterized by a low cost and simplicity of use. However, Kinect resolution is still lower than that exhibited by static scanners or by more costly dynamic depth scanners. As a result, very few researchers have considered the Kinect for the purpose of FER.

The acquisition specifications of the Kinect sensor v1.0 (displayed in figure 1.4a) are listed below:

- The nominal accuracy is 1cm depth at 2m of distance;
- Depth images at a resolution of 640×480 pixels are captured at a speed of around 25-30

(a)



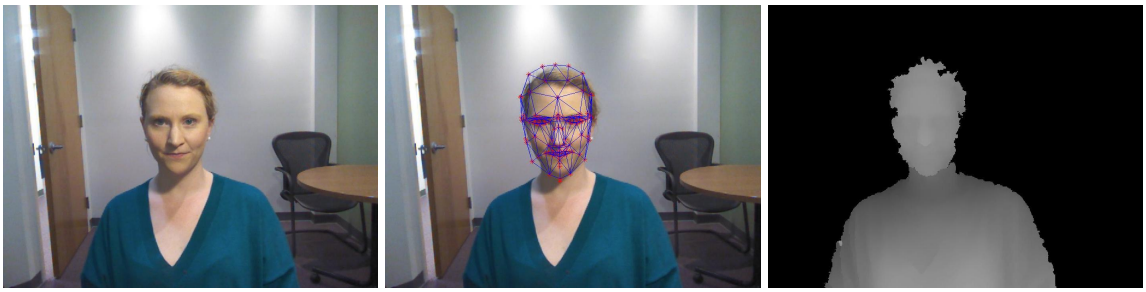
(b)

Figure 1.4: (a) The Microsoft Kinect sensor. (b) The operative range when the sensor is used in the default and near modes.

frames per second (fps);

- RGB color images are synchronized with depth images and captured at a resolution of 640×480 pixels (RGB at 1280×960 is also possible, but at 12fps);
- The operative range is between 0.4 (for the near mode) to 4m (see figure 1.4b).

The Kinect for Windows sensor expands the possibilities with the so called Near Mode, which enables the depth camera to see objects as close as 40cm in front of the sensor. Figure 1.4b summarizes the operative range of the device both for the default and the near mode. Figures 1.5a and 1.5c show sample RGB and depth images output from the Kinect, respectively. The Kinect SDK also enables the extraction and tracking of 121 3D keypoints on the face as shown in Figure 1.5b.



(a) (b) (c)

Figure 1.5: Sample of the Kinect output data. In (a), sample RGB image is shown. In (b), the corresponding projected 3D face mesh is displayed. In (c), depth image resulting from the Kinect sensor is shown.

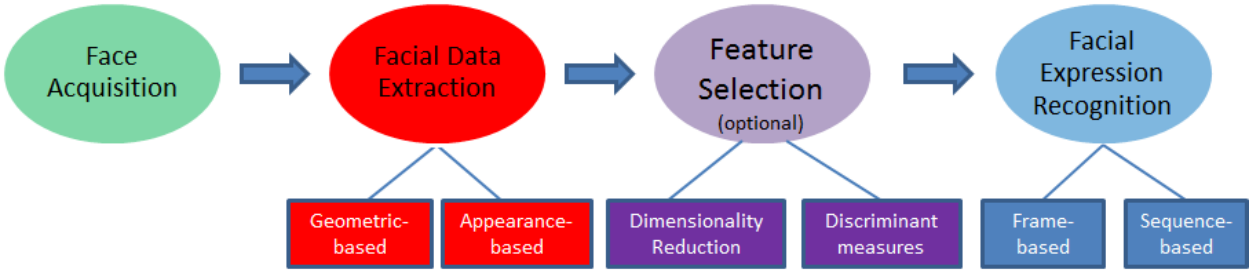


Figure 1.6: Basic structure of a generic FER system [3].

1.4 Basic Structure of Facial Expression Recognition Systems

Facial expression recognition (FER) systems are concerned with measuring face deformation and automatically recognizing the corresponding expression [3]. The general approach to automatic FER consists of four steps: face acquisition, facial data extraction and representation, feature selection, and facial expression recognition. These steps, as well as some of their corresponding approaches, are illustrated in Figure 1.6.

The face acquisition step automatically finds the face location for the input images or sequences. This can be done either by independently detecting the face in each frame, or by detecting the face at the first frame in the sequence and then tracking it through the later frames using some feature tracking approach. The Kinect face tracking SDK engine [60] analyzes input from a Kinect camera and automatically detects the face region in each frame. Therefore, this step is not of concern for this dissertation.

After the face is located, the next step is to extract and represent the facial changes that result from the facial expressions. There are mainly two approaches for feature extraction: geometric feature-based methods and appearance-based methods. The geometric-based facial features represent the shape and locations of facial components such as mouth, eyes, and eye brows. The facial feature

points are extracted and a feature vector is composed, using a previously constructed face model, to represent the face geometry using distance or angle-based techniques. The appearance-based methods apply image filters and/or feature extraction approach (e.g., scale-invariant feature transform (SIFT) [61], GIST [62], local binary pattern (LBP) [63] or Gabor wavelets [64]) to either the whole face or specific face regions to extract the feature vector.

The feature selection step aims to reduce the feature vector by selecting the features that increase the class separability and thus the recognition rate. Dimensionality reduction and discriminant measures are two well known approaches for feature selection.

Facial expression recognition is the final step of automatic FER systems. The facial changes can be identified as facial action units or emotional expressions. Depending on whether temporal information is utilized or not, the recognition approach is classified as frame-based (static) or sequence-based (dynamic). This dissertation presents both static and dynamic FER systems.

1.5 Challenges

Research in 3D facial expression analysis is still in its infant stage. There exist several issues that remain unsolved in this field. This section discusses the current challenges towards building a robust FER system with emphasis on the Kinect sensor. The challenges are either in the high variability of input data, the sensor specifications limitations, or in the limited availability of FER-specific datasets. Each of these challenges is discussed below.

1.5.1 Data Variations Challenge

Although humans recognize facial expressions virtually without effort or delay, reliable expression recognition by a machine is still challenging. To achieve successful recognition performance, expression recognition approaches require some control over the imaging conditions. The controlled imaging conditions typically cover the following challenges:

- Pose of the head: The effect of out-of-plane rotation (non-frontal pose) is more difficult to mitigate than frontal poses, as it can result in wide variability of image views. Further research is needed into pose-invariant expression recognition.
- Environment clutter and illumination: Complex image background pattern, occlusion, and uncontrolled lighting have a potentially negative effect on recognition. These factors would typically make image segmentation more difficult to perform reliably. Hence, they may potentially cause the contamination of feature extraction by information not related to facial expression. Consequently, many researchers use uncluttered backgrounds and controlled illumination, although such conditions do not match the operational environment of some potential applications of expression recognition. The Kinect sensor provides an infrared

camera that can artificially solve the illumination problem by the 3D point clouds that it generates.

- Facial variability: Facial properties contain a high degree of variability due to a number of factors such as differences across people (arising from age, illness, gender, or race, for example), growth or shaving of facial hair, make-up, and blending of several expressions.

1.5.2 Dataset Challenge

In addition to input variations, most existing FER datasets are based on expensive 3D sensors that are not suitable for realtime applications. The lack of a Kinect-based FER dataset makes our mission of building a reliable FER system more difficult.

Moreover, there are many databases that can be used for static 3D facial expression analysis. However, most of them are static. The current trend has shown a shift in researchers interest towards the analysis of facial expression dynamics, as these allow the encoding of temporal cues that are significant for more complex states and expressions. Currently, there exist only two publicly available datasets of dynamic 3D facial samples, namely BU-4DFE [5] and D3DFACS [65]. BU-4DFE was mainly used for facial expression recognition, and the recently published D3DFACS, which contains only 10 subjects, was designed for Action Unit (AU(s)) analysis. In order to test the real generalization capabilities of 4D (3D+time) facial expression recognition system, more databases of posed 4D facial expressions and AU(s)s are needed.

1.5.3 The Kinect Sensor Challenge

A Kinect sensor produces both 2D color video and depth images at 30 fps, combining the best of both worlds. However, in spite of its advantages, the Kinect sensor imposes significant limitations to build effective FER systems. The main problem is that the Kinect sensor is relatively low in resolution and noisy as compared to many other 3D sensing techniques. As a result, very few researchers have considered the Kinect for the purpose of FER. Techniques that can get over such limitations is required for robust FER system using the Kinect.

1.6 Dissertation Contributions

The main contributions of this work are as follows:

1. Efficient techniques for facial expression recognition (FER) that are based on 3D data captured by the Kinect v1.0;

2. A novel feature extraction approach that is based on various distance metrics (DMs), adaptably selected and trained on pool of binary Radial Basis Function (RBF) support vector machine (SVM) classifiers for each expression;

3. An efficient approach to classify the six basic facial expressions based on maximization of SVM posterior probabilities of six binary classifiers;

4. A detailed analysis for the minimum training size required for a robust FER system which contributes in accelerating the training step, especially with large datasets;

5. A framework that addresses the challenges of pose variation in FER systems;

6. A novel FER dataset, VT-KFER. VT-KFER is the first publicly available Kinect-based RGBD+time dataset that not only includes adults (age from 18 to 30) but also children (age from 10 to 17) in various poses and three intensities;

7. In addition to VT-KFER, a second FER dataset was collected. This new dataset is the first RGBD dataset captured by the Kinect that includes 20 children;

8. This dissertation addresses the problem of fusing high dimensional noisy data that contains severe non-linearities (such as what the Kinect sensor produces);

9. A novel feature fusion approach, the dual kernel discriminant analysis (DKDA);

10. A novel ranking-based features selection approach that utilizes five different selection criteria in which features with the greatest class separability are selected;

11. A dynamic realtime FER system that employs discriminative 3D facial features extracted from depth data to automatically recognize the six basic expressions, plus neutral. The proposed approach is fast, relatively robust to pose and illumination variation, of low cost, and outperforms the state of the art;

12. Affective features that are strongly correlated with Emotion Facial Action Coding System (EMFACS) action units (AUs) are utilized. To the best of our knowledge, this is the first work that validates their proposed selected features with EMFACS AUs;

13. This document also introduces a novel semi-automatic labeling approach for AU location in 3D face mesh. The proposed approach generates heat maps that represents the AU location and then utilizes the generated heat maps as ground truth for our AU validation;

14. The proposed realtime FER system has been tested on children and adults. To the best of our knowledge, no other realtime FER system has been tested on children. This makes our system more reliable for many children related applications/studies.

1.7 Dissertation Organization

Chapter 2 presents an extensive literature review for FER datasets (Section 2.1), the previous FER (Section 2.2) and work that analysed action units and their detection approaches (Section 2.3). Chapter 3 introduces our new RGBD+time dataset, its participants (Section 3.2), the procedure taken for recording the data (Section 3.3), its contents (Section 3.4), and a qualitative evaluation for it with a conclusion (Sections 3.5 and 3.6). Chapter 4 presents our novel adaptive feature selection-based FER system where various distance metrics were adaptively selected for better expressions classification. The main procedures followed here is given in section 4.2. The corresponding results with discussion are presented with a conclusion in sections 4.3 and 4.4), respectively. Chapter 5 introduces our multi-modal feature fusion framework which is based on DKDA. Chapter 6 describes a dynamic realtime FER system based on 3D only. The proposed system includes a novel EMFACS validation framework (Section 6.3) and comparative results with state of the art. Chapter 7 presents the dissertation summary.

Chapter 2

Related Work

In the past decade, a great deal of effort has been put into developing automatic approaches for facial expression recognition ([6, 31, 34, 35, 50, 66–73]). Most of these works were based on 2D images or videos (e.g., [31, 34, 35, 50, 67–69, 73]). Only recently, researchers have also considered expression recognition using three-dimensional (3D) data [40, 74–82] and time-varying 3D sequences, sometimes called 4D data [83, 84].

To our knowledge, only a few of these 3D/4D attempts have utilized the Kinect [85], although this sensor has been used for other purposes such as face detection [86], face recognition [72, 87], animation [70, 71] and face tracking [88]. The work in [73] employed the Kinect sensor for a FER system but they only utilized the RGB camera. Other work utilized other sensors with high resolution for 3D FER system [66, 74, 77, 80, 89–94] with high accuracy. The two main comprehensive surveys about 2D and 3D FER can be found in [37] and [46], respectively.

A comprehensive survey of existing 3D FER datasets is presented in section 2.1. The related FER work is presented in Section 2.2 from five main perspectives, the feature extraction and representation (Section 2.2.1), feature selection (Section 2.2.2), feature fusion (Section 2.2.3), classification (Section 2.2.4), and data pruning (section 2.2.5). The research work conducted to study action units and their detection approaches is also introduced in Section 2.3.

2.1 3D Facial Expression Databases

A common publicly available dataset is essential for research on facial expression analysis. This document considers a database dedicated to expression analysis only when it contains at least the 6 basic expressions or/and different AUs of the Facial Action Coding System (FACS) [95]. Although there are a number of popular 2D and 3D facial expression databases accessible for facial expression research, prior to the work described here, no readily accessible Kinect-based FER database with a complete set of basic facial expressions for 3D or 4D expression analysis

was available. The lack of an accredited common database and evaluation methodology makes it difficult to compare, validate, and resolve the issues concerned with 3D facial expression analysis. In this section, the existing 3D databases are reviewed.

A number of standard facial expression recognition databases, containing both 2D and 3D data (e.g., [4, 5, 96–98]), have become available to the facial expression recognition community. Before this work, there were only four publicly available 3D databases that were designed specifically for expression analysis (i.e., the BU-3DFE [4], the BU-4DFE [5], the Bosphorus [99], and the ICT-3DRFE [100]).

The first 3D facial expression dataset [50] consists of six subjects expressing the six basic facial expressions. It was collected using NTSC video equipment and a custom-built system consisting of a camera/projector pair and active stereo using structured light projections. However, the database is not publicly available.

The first systematic effort to collect public 3D facial data for facial expression recognition was in the creation of the BU-3DFE dataset [4]. Examples of the BU-3DFE dataset can be seen in Figure 2.1. Static 3D expressive faces of 100 subjects, displaying the six prototypical expressions at four different intensity levels, were captured using the 3DMD acquisition setup [101]. The models created were of resolution in the range of 20,000 to 35,000 polygons, depending on the size of the subject's face. The database was accompanied by a set of metadata, including the position of 83 facial feature points on each facial model.

The same institution that built the BU-3DFE dataset recorded BU-4DFE [5], the first database consisting of 4D faces (sequences of 3D faces). The database includes 101 subjects and was created using the DI3D (Dimensional Imaging [102]) dynamic face capturing system. It contains sequences of the six prototypical facial expressions with their temporal segments (onset, apex and offset) with each sequence lasting approximately 4 s (examples can be seen in Figure 2.2). The temporal and spatial resolution are 25 frames/s and 35,000 vertices, respectively. Unfortunately, the database provides no AU annotation. Also. as seen in Figure 2.3 the system setup is very huge and may not be suitable for portable realtime applications as the Kinect.

Another publicly available dataset consisting of static 3D facial models is the Bosphorus database [99]. The database was captured using Inspeck Mega Capturor II 3D, which is a commercial structured-light based 3D digitizer device. The database consists of 105 subjects (60 men and 45 women, with the majority of the subjects being Caucasian), 27 of whom were professional actors, in various poses, expressions and occlusion conditions. The subjects expressed the six basic facial expressions, and up to 24 AUs. The database is fully annotated with regards to 25 AUs, split as 18 lower AUs and 7 upper AUs. The texture images are of resolution 1600×1200 pixels while the 3D faces consist of approximately 35,000 vertices. Examples from the dataset can be seen in Figure 2.4.

One of the most recently created facial expression databases is the ICT-3DRFE database [103]. The database consists of 3D data of very high resolution recorded under varying illumination conditions, in order to test the performance of automatic 2D facial expression recognition systems.

Figure 2.1: Sample of the BU-3DFE dataset [4].

Figure 2.2: Sample of the BU-4DFE [5] dataset.

Figure 2.3: 3D dynamic imaging system setup for BU-4DFE [5].

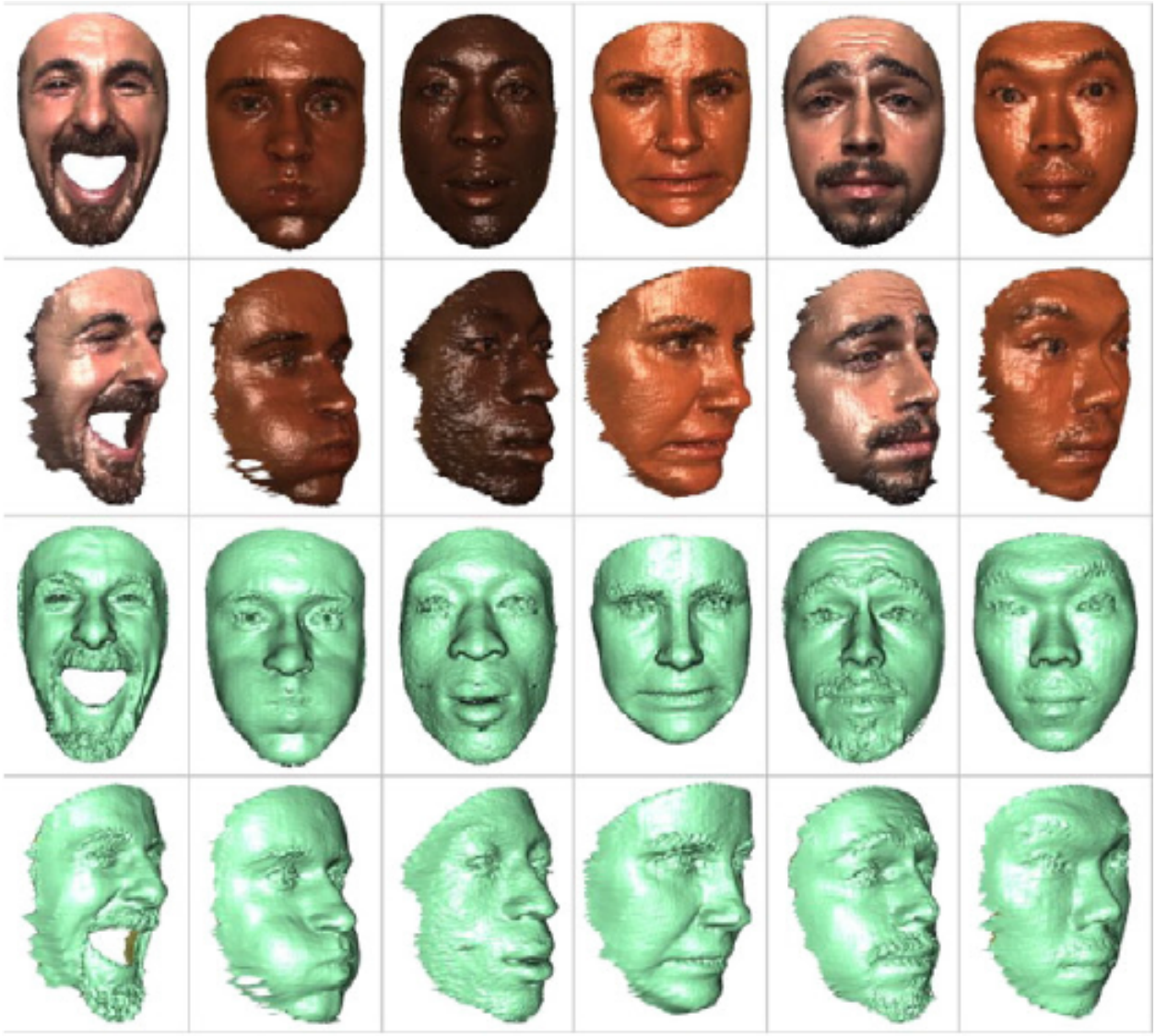


Figure 2.4: Sample of the Bosphorus dataset.

The database contains 23 subjects (17 male and 6 female) and 15 expressions: the six prototypical expressions, two neutral states (eyes closed and open), two eyebrow expressions, scrunched face expression, and four eye gaze expressions (Figure 2.5). Each model in the dataset contains up to 1,200,000 vertices with reflectance maps of 1296×1944 pixels, resolution that corresponds to a detail level of sub-millimeter skin pores. The database also includes photometric information that allows photorealistic rendering. The database is fully annotated with regards to AUs. AUs are also assigned scores between 0 and 1 depending on the degree of muscle activity.

Another available facial expression database that is not widely available is the one presented in [104]. The database consists of 832 sequences of 52 participants (40 males and 12 females). In each sequence, the human subject shows happiness, sadness, disgust, surprise, or neutral expression or display an AU 2 to 4 times for 5 to 10 seconds on average.

However, none of these datasets utilize the Kinect. They were captured using high resolution 3D sensors and/or contain only acted expressions. One of them, [99], even employed actors as subjects.

Some recently developed Kinect-based datasets [105, 106] have been released. For example, the FaceWarehouse database [105] includes raw RGBD data, with sets of 74 landmarks of facial features, such as eye corner, mouth contour and the nose tip. The dataset includes 150 subjects of age range from 7 to 80 years old with 20 expressions and AUs. However, Facewarehouse dataset is mainly designed for animation and does not include the full set of basic expressions. Figure 2.6 illustrates samples of the FaceWarehouse dataset.

The CurtinFaces dataset [106] was captured using the Kinect for the purpose of face recognition under varying poses, expressions, illumination and disguise. It includes 52 male and female subjects, with and without glasses. The dataset has 92 images per subject. This includes 3 frontal poses, 49 images for 7 poses per 7 expressions, 35 images for 5 illumination per 7 expressions, 5 images for sunglasses different poses. It also includes no landmarks.

Other databases, such as VAP [107] and KiFaEx [73], are rarely used for facial expression recognition purposes, although they contain expression variations, mainly due to an incomplete expression set. A summary of well-known FER databases is given in Table 2.1.

To conclude, no existing 3D dataset has been designed for FER, especially using the Kinect, that includes varying poses, children, and/or unscripted data.

2.2 Facial Expression Recognition

In the last decade, much FER research was conducted using either 2D [6, 37, 118], 3D [4, 8, 119, 120], 4D [121], or a combination of 2D and 3D [2, 122, 123]. Recently, many researchers have considered expression recognition using three-dimensional (3D) data [66, 77, 78, 80, 89, 94, 99, 117] and time-varying 3D sequences, sometimes called 4D data [124], using high resolution data

Figure 2.5: Sample of the ICT-3DRFE dataset.

Figure 2.6: Sample of the FaceWarehouse dataset.

sensors.

The use of fast sensors such as the Kinect has not yet been widely used for FER. Some work [73] employed the Kinect for FER but they only utilized the RGB camera. Few researchers [125, 126] have employed the Kinect for realtime FER. Mao et al. [125] combined the features of 6 AUs and 121 feature point positions (FPPs) tracked by the Kinect to automatically recognize the six basic expressions, plus neutral. Malawski et al. [126] proposed a FER approach for 3 expressions (neutral, happiness, and anger), using only depth information provided by the Kinect. Comprehensive surveys that focus on 3D-based FER are available in [27, 46, 127, 128]. This section will describe related 3D FER systems from perspectives of feature extraction, feature selection, feature fusion, classification, training data reduction, and realtime operation.

2.2.1 Feature Extraction and Representation

Many automated 3D FER systems begin by detecting salient points on the face, such as corners of the eyes or mouth. These are often called landmarks or fiducial points or keypoints, and discriminating features can be extracted using these locations. An example database that provides 3D coordinates of facial keypoints is BU-3DFE [4]. Some systems also extract a 3D triangular mesh representation, with vertices of the mesh serving as keypoints. For example, the Kinect SDK produces a 3D mesh of 206 triangles, consisting of 121 vertices linked by 318 edges [129]. FaceWarehouse [105] and VT-KFER [2] are examples of datasets that provide 3D triangular mesh representations of faces.

Distance-based features: Distances between 3D keypoints represent a common feature type for both static and dynamic FER. Euclidean distance is often used, as well as other metrics such as sum of absolute distances. For example, the method developed in [74] uses six characteristic distances that are extracted from the distribution of 11 facial feature points provided in BU-3DFE. They reported an average recognition rate of 91.3%. In [130], a larger number of distances are extracted, achieving a mean rate of 87.8%. Similarly, [75] uses six distances that are related to the movement of particular parts of the face, plus some angles that relate to the shape of the eyes and mouth, achieving an average rate of 90.2%.

In [78], the authors used residues as features, which they define as magnitude and direction of keypoint displacement with respect to the neutral expression. A 2D feature matrix was formed by combining the residue values for all 3 spatial dimensions. The average recognition rate achieved was 91.7%. The authors in [131] computed the displacement of facial keypoints from the neutral expression at five different yaw angles. Then they normalized these distances to zero mean and unit variance, and used the resulting features for classification. The average recognition rate they achieved for different head poses was 66.5%.

The distances between all pairs of available facial keypoints was used as features in [66, 76, 132]. The average expression recognition rates achieved using the BU-3DFE database with these methods were 93.7% [66, 132] and 88.2% [76]. In [40], normalized Euclidean distances between 83

facial feature points in the 3D space were employed. They reported an accuracy of 95% for 6 expressions.

Curvature-based features: Surface curvature is also a common feature type for modeling and recognizing 3D facial expressions. In [77], the surface curvature was computed at each vertex of a triangular face mesh, and classified into 1 of 8 categories, based on the sign of maximum principal curvature and minimum principal curvature. The authors combined the computed curvature features with calculated distances between all pairs of available face points. This approach resulted in an average expression recognition rate of 83.5%.

Similarly, the authors of [85] computed mean curvature and Gaussian curvature over range images captured using the Kinect sensor. They achieved an average recognition accuracy of about 91% for nine facial exercises. Mean curvature was employed again by [133]. Several noise filtering steps were applied to smooth the data and to fill in gaps of missing data. Mean curvatures were then estimated on the 3D face mesh surface and then resampled in the image domain via orthogonal projection. They reported an accuracy of 95.4%.

The authors in [93] computed LBP over range images to approximate 3D curvature. They fused the 3D curvature with LBP features extracted from 2D images, and they reported a recognition rate of 77% using a SVM.

Angle-based features: Geometric angles have also been employed for FER [134]. The angles were computed from triangles of the 3D face mesh extracted by the Kinect SDK. In [2], The authors fused the 3D features with 2D features for 6-class FER. The average accuracy was 80%.

Patch-based features: A patch refers to a small, local region around a point of interest. Patches can be defined for portions of a face mesh [135], or around certain keypoints [82]. For a set of 3D points within a small patch, regression can be performed to fit a surface locally. For instance, the authors in [135] fit a smooth polynomial patch to the local surface at each vertex in a triangular mesh. From the polynomial shape representation, they estimated the principal curvatures and applied classification rules to label each vertex according to curvature type. These curvature labels were use, in turn, to make a decision concerning the facial expression. This approach achieved average expression recognition rate of 83.6%, on an in-house database containing the six basic expressions.

In another effort, patches were fit around manually selected landmarks in a 3D mesh [82]. These patches were used to define surface curves circling these landmarks. Then they computed a square-root velocity function (SRVF) to represent the shape of each curve. Geodesic distance values were computed using these curves, and were used for FER. This method achieved an average recognition rate of 96.1%.

Features based on surface normals: The local orientation of a 3D shape can be represented using surface normals. The authors in [136] used surface normals within a statistical model to capture the variation of 3D shape due to facial expression. Compared to features based on keypoint locations and related distance features, the surface-normal approach showed better accuracy. For a set of four expressions (happiness, anger, fear, and sadness), an average recognition accuracy of 53%

was reported.

Researchers have also combined features based on surface normals with other feature types [90, 137].

Other approaches: Other feature extraction approaches have also been used for 2D and 3D FER. These include principal component analysis (PCA) [83, 92], linear discriminant analysis (LDA) [83, 92, 138], and LBP [93, 94, 137, 139]. Combinations of those features have often been used [83, 90, 117, 137, 139]. In several cases (e.g., [134]), the fusion of several feature types gives better results than using one type of feature alone. Gong et. al. [80] captured the facial surface deformation by encoding the depth differences between a basic facial shape component (i.e., neutral face) and each expressional face with more than 76% recognition accuracy of the six basic expression.

These hand-crafted appearance-based features such as LBP or geometric-based features such as distances and angles, were computed from mostly the locations of facial landmarks. However, these hand-crafted features are not tuned to a specific task, and thus they limit the performance of the classifier learned on these features. Deep learning techniques, on the other hand, allowed a multistage approach in which the features were learned directly from the pixel values in combination with the classifier. Systems that employed features obtained using deep learning techniques such as convolution neural networks (CNN) and deep belief network (DBN) [140–144] showed the state of the art performances over many FER datasets. For instance, Jung et al. [144] employed two models for deep network construction. One deep network to extract temporal appearance features from image sequences, while the other to extract temporal geometry features from temporal facial landmark points. This combined deep networks achieved the state-of-the-art performance in the CK+ (97.25%) and Oulu-CASIA databases (81.46%). Ranzato et. al. [143] employed a DBN with multiple hidden layers for the recognition of 7 expressions. The proposed DBN showed high performance when predicting expression categories from face images with synthetic occlusion. The deep model was generally more robust to these occlusions compared to other methods such as Gabor with PCA and sparse coding.

Transfer learning of features from deep convolutional networks (ConvNets) was successfully applied in facial expression recognition by Xu et al. [145]. Compared with 78.84% accuracy based on distance features and 50.65% accuracy based on Gabor features, they have achieved 80.49% accuracy on an in-house dataset. Transfer learning also has recently been used for cross-dataset facial expression recognition as in [146] and 3D object recognition as in [147, 148]. In the cross-dataset work, the transfer learning approach outperformed the typical approach of training on one dataset and testing on another.

2.2.2 Feature Selection

Feature selection is commonly achieved through dimensionality reduction techniques, such as PCA. PCA has been employed in several 3D facial expression recognition methods [66, 77, 83,

92, 131, 132]. In some of these methods, LDA was applied to create a discriminant subspace [66, 83, 132]. Discriminant measures such as the Fisher criterion [66, 132] and Kullback-Leibler divergence (cross-entropy) [40, 149] have been employed to select the best features for classification. The average recognition rate in [149] with entropy driven features was 88.3%, which was 8% more than achieved using the standard features. The normalized cut-based filter (NCBF) algorithm was used in [77] to select the most relevant geometrically localized features (GLF) and surface curvature features (SCF) with low redundancy. The average recognition rate using nearest neighbor classification was 83.5% on the BU-3DFE dataset. Entropy was utilized for feature selection in [149] to select the most discriminative features caused by facial deformation. In [134], the authors combined five discriminant criteria, including entropy and the t-test, for feature selection with more than 80% average recognition accuracy for 6 expressions.

2.2.3 Feature Fusion

To get better system performance, some authors have combined 2D and 3D data [2, 46, 122]. However, when fusing 2D and 3D features, the system may suffer from the curse of dimensionality [150]. Therefore, reduction approaches such as PCA [151], LDA [152, 153], minimal redundancy-maximal-relevance (mRMR) [121], and other approaches [119, 154] have been proposed. LDA is preferred over PCA in many classification problems such as face recognition because LDA deals directly with discrimination between classes, whereas PCA deals with the data in its entirety without paying attention to the underlying class structure [155]. Although LDA can provides improved facial feature representation and thus better recognition accuracy, it is still a linear approach in nature. When severe non-linearity is involved, this approach is intrinsically poor. To deal with this limitation, nonlinear extensions of LDA such as Kernel Discriminant Analysis (KDA) [156, 157] have been proposed.

2.2.4 Classification

A wide range of classification techniques have been employed in 3D facial expression recognition systems. Support Vector Machines (SVMs) represent one of the most common methods [2, 8, 78, 85, 89, 93, 94, 117, 120, 133, 134, 136, 137, 149], including multi-class SVMs [40, 158]. Other approaches that have been used include linear classifiers [159], LDA-based techniques [135, 160, 161] nearest neighbor (NN) methods [162, 163], clustering algorithms [79], and Maximum Likelihood classifiers [164]. Bayesian classifiers have been also widely used [139]. AdaBoost [133, 139] has been used in conjunction with these classifiers. Neural network classifiers constitute another popular approach. Probabilistic neural networks (PNN), for example, were employed in [74, 130]. Rule-based classifiers, widely used for 2D FER, have also been employed for 3D analysis [92, 165].

2.2.5 Training Data Reduction

To our knowledge, no previous work has investigated how much training data is needed for robust 3D facial expression recognition system. Recently, [166] tackled this problem for the general object recognition paradigm. They investigated whether answering this question will be better by increasing amounts of training data or the development of better object detection models. This dissertation tests how much training data size a giving FER system needs to achieve as high accuracy as when it uses empirical training data size (e.g., use 80% for training and 20% for testing) as other work typically follow.

2.2.6 Kinect-based FER in Realtime

In spite of the recent interest in 3D sensing for FER, the Kinect sensor has not yet been widely used for this purpose. Some work [73] employed the Kinect for FER but only the RGB camera was used. Realtime FER using the Kinect depth camera was considered by [125, 126]. Mao et al. [125] combined the 3D location of facial keypoints with the weights of six animation units, representing the estimated action units by the Kinect SDK, to automatically recognize the six basic expressions, plus neutral. Average recognition rate of 80% for the seven expressions was reported using a Kinect-based dataset of only 10 subjects (adults). Malawski et al. [126] proposed a FER approach for only 3 expressions (neutral, happiness, and anger), using depth information provided by the Kinect. The best recognition accuracy reported was 87%, and it was achieved using AdaBoost-based feature selection and a SVM.

2.2.7 Summary

Table 2.2 summarizes the 3D FER systems that are most closely related to our work. As shown in the table, most of the existing 3D FER systems are based on relatively expensive sensors that are not suitable for realtime operation. In addition, the lower-cost Kinect-based systems summarized here either have not been tested in realtime, utilize 2D modality which is computationally expensive, or have not considered AUs.

2.3 Action Units

Since each facial expression has its own explicit local variations, which are corresponded to AUs, early studies on FER have considered making use of these distinctive spatial regions. The automatic detection of action units from two-dimensional images [168–170] or videos [171–173] has been studied extensively. Some researchers proposed curve fitting of geometric models such as ellipses for AU detection [168, 169]. Statistical approaches such as active shape models (ASM)

and active appearance models (AAM), were extensively employed for facial analysis [170]. Other researchers proposed direct image-based analysis techniques, such as Bartlett et al. [174]. The authors automatically selected Gabor wavelet coefficients from the face image using AdaBoost. A dynamic system based on temporal information extraction from videos was proposed in [172, 173] to analyze the semantic relationships of AUs. Jaiswal and Valstar proposed a combination of Convolutional and Bi-directional Long Short-Term Memory Neural Networks (CNN-BLSTM), which jointly learned shape, appearance and dynamics in a deep learning manner.

Deep learning has been recently employed for AU detection in various work [175–177] with outstanding performance. Zhao et. al. presented Deep Region and Multi-label Learning (DRML) framework for spontaneous facial AU detection. The use of DRML-based framework unified the AU detection by construction, and allowed two seemingly irrelevant tasks, region learning (RL) and multi-label learning (ML), to interact directly. Their approach was able to identify more specific regions for different AUs than conventional patch-based methods. They evaluated their approach on two spontaneous datasets, BP4D [178] that includes 12 AUs and DISFA [179] that includes 8 AUs. The performance was evaluated on two common frame-based metrics: F1-frame and AUC. F1-frame is the harmonic mean of precision and recall. AUC quantifies the relation between true and false positives. Compared to the state of the art deep networks such as AlexNet, DRML outperformed AlexNet in all 12 AUs in AUC (56 on average) and 8 out of 12 using F1-frame (48.3 on average) in BP4D dataset. Better performance was reported on the DISFA dataset as well.

Although deep learning has outperformed many other ML approaches when used for feature classification using 2D modality (and recently using 3D for object recognition [147, 148]), working with deep learning requires specific hardware (e.g., Beowulf clusters, GPUs) to handle such extensive computations and very large datasets for deep training (e.g., BP4D includes around 140,000 face image). Because of the small collected dataset size, using deep learning was not considered.

A limitation of AU detection systems that rely on 2D sensing is that they are subject to fairly tight constraints on illumination and pose of the subject's head. Those systems suffer when the room is not well illuminated, and when the subject is not squarely facing the camera. Researchers [104, 180, 181] have therefore considered AU detection using sensors that capture 3D and 4D (time-varying 3D) data. It has been claimed that 3D face data allows better discrimination of subtle differences of AUs [182].

A few researchers have considered AU detection [40, 84, 135] in conjunction with FER. In one study, 11 posed AUs were recognized in [104] using rule-based classification. In this dissertation, a larger set of AUs are detected and used in a validation framework to support FER.

Table 2.1: Facial expression dataset summary. For each dataset, the table indicates whether it is static (S) or dynamic (D), created specifically for facial expression recognition (FER) or not. The table also provides the type of sensor utilized, the number of subjects (i.e., size), what expressions (Exp.) and Action Units (AU) does it includes (i.e., contents), whether the landmarks (i.e., LM) are included and how many landmarks are there, if any, whether it provides a full annotation (Annot.), non-frontal poses (i.e., NF), and if it public (P) or not.

Name	**S/D**	**FER**	**Sensor**	**Size**	**Contents**	**LM**	**Annot.**	**NF**	**P**
BU-3DFE [4]	S	Y	Multi-view Stereo (3DMD)	100 adults	6 Exp. + 4 intensities	83	N/A	Y	Y
Bosphorus [99]	S	Y	Structured Light (Inspeck Mega Capturor II 3D)	105 adults inc. 27 actors	24 AUs + neutral + 6 Exp. with occlusions	24	AUs	Y	Y
ICT-3DRFE [103]	S	Y	Structured Light (Other)	23 adults	15 Exp. (6 basic + others)	N/A	AUs	N	Y
Benedikt et al. [108]	S	N	Multi-view Stereo (3DMD)	94 adults	Smiles and word utterance	N/A	N/A	N	N
ND-2006 [109]	S	N	Structured Light (Minolta Vivid 910)	888 adults	5 Exp. + Neutral	N/A	N/A	N	Y
CASIA [110, 111]	S	N	Structured Light (Minolta Vivid 910)	123 adults	5 Exp. + Neutral	N/A	N/A	N	Y
Gavdb [112]	S	N	Structured Light (Minolta VI-700).	61 adults	3 Exp.	N/A	N/A	Y	Y
York 3D [113, 114]	S	N	Structured light	350 adults	4 Exp. + Neutral	N/A	N/A	Y	Y
Texas [115, 116]	S	N	Stereo	105 adults	3E + open/closed eyes	25	N/A	N	Y
UPM-3DFE [117]	S	Y	Structured Light (3D Flexscan (V2.6))	50 adults	6 Exp.	32(M)	Y	N	N
BU-4DFE [5]	D	Y	Multi-view Stereo (DI3D)	101 adults	6 Exp.	83	N/A	N	Y
D3DFACS [65]	D	Y	Multi-view Stereo (3DMD)	10 adults inc. 4 FACS experts	Up to 38 AUs	N/A	AU peaks	N	Y
Facewarehouse [105]	S	N	Structured light (Kinect)	150 subjects (7-80)	20 Exp. + AUs	74	Y	N	Y
CurtinFaces [106]	S	N	Structured light (Kinect)	52 adults w/o Glasses	7 Exp.	N/A	Y	Y	Y
VAP [107]	S	N	Structured light (Kinect)	31 adults	4 Exp. + other poses	N	Y	Y	Y
KiFaEx [73]	S	Y	Structured light (Kinect)	20 adults	7 Exp.	N/A	N/A	N	N

Table 2.2: 3D facial expression recognition systems. Only a few systems have utilized data from the Kinect, as indicated in the database column. The next 2 columns summarize the 3D facial features and the classifier that was utilized. The 2D column indicates whether 2D data was used along with 3D data. The S/D column indicates whether the system is static or dynamic. The LM column lists the number of landmark points that were used. Where specified, those points are indicated as being manually (M) or automatically (A) selected. The other cases utilized both manually and automatically selected landmarks. The E/AU column indicates the number of expressions and/or action units that are recognized by the system. (E refers to expression, N refers to neutral, and I refers to intensity.) The NF column indicates whether nonfrontal poses are accommodated, or frontal (F) only. The Accuracy column lists the recognition accuracy reported for each system. The realtime column (RT) indicates whether the system runs in realtime or not.

Reference	Database	3D facial features	Classifier	2D	S/D	LM	E/AU	F/NF	Accuracy	RT
[125]	UJS-KED	3D facial keypoints + weights of six animation units	SVM	Y	D	121 (A)	7 E	NF	80%	Y
[126]	Kinect-based	AdaBoost-based selected distances	SVM	N	S	121 (A)	3 E	F	87%	Y
[134]	VT-KFER	Automatically selected Angles	SVM	Y	S	121 (A)	6 E	NF	80%	N
[8]	in-house Kinect-based /FaceWare-house	9 distance metrics	SVM	N	S	121 (A)	6 E	NF	98%	N
[120]	in-house Kinect-based	121 keypoint coordinates	SVM/K-NN	N	S	121 (A)	6 E	F	81.8%	N
[85]	Kinect	Curvature + line profile DCT + point signature DCT	SVM	N	S	58 (M)	9 E	F	91%	N
[149]	BU-3DFE	Hierarchical facial feature selection using entropy	SVM + voting	N	S	71	6 E	F	88%	N
[131]	BU-3DFE	2D keypoints displacement between emotional and neutral expressions	LDC, QDC, Parzen, SVM, K-NN (with PCA + LDA + LPP)	N	S	N/A	6 E + N + 4 I	NF	N/A	N
[80]	BU-3DFE	depth difference between emotional face and neutral expression	SVM	N	S	N/A	6 E	F	76%	N
[130]	BU-3DFE	6 3D distances	Probabilistic neural network (PNN)	N	S	23	6 E + neutral	F	88%	N
[77]	BU-3DFE	3D keypoints+surface curvature features filtered by normalized cut and fused using PCA	Multiple discriminant analysis	N	S	83	6 E	F	95%	N
[89]	BU-3DFE	Geodesic distances between local curve-based patches	Multi-boosting, SVM	N	S	68	6 E	F	99% / 98%	N
[90]	BU-3DFE	Vertex coordinates, normals, and local curvature	Linear logistic regression	N	S	Auto	6 E	F	90%	N
[94]	BU-3DFE	LBP, multiscale LBP, and local Gabor binary patterns (LGBP)	SVM	Y	S	Patch-based	6 E	NF	71%	N
[78]	BU-3DFE	Spatial displacement(residues) between neutral and expression	SVM	N	S	83	6 E	F	92%	N
[117]	BU-3DFE /UPM-3DFE	Distance + angles + other geometrical features	SVM	N	S	A	7 E	F	92% /89%	N
[137]	BU-3DFE /BU-4DFE	Positions, normals, curvatures, and wavelets + point distribution model	SVM	N	S	12	6 E	F	91% / 74% 6 E (97% 3 E)	N
[136]	Bosphorus	3D surface normal	SVM	N	S	83	6 E	F	53% (happy=100%)	N
[139]	Bosphorus	Multi-scale LBP, shape index, distances between landmarks, and landmark displacement	Dynamic Bayesian net, adaBoost	Y	S	19 (M)	7/16 AUs	F	94% / 86%	N
[93]	Bosphorus	Curvature LBP	Chi square distance-based NN classification, SVM	Y	S	N/A	6 E	NF	77%	N
[133]	Bospshorus /DFAT-504	Curvature and intensity	SVM	Y	S	N/A	25 AUs/19 AUs	NF	97%	N
[74]	BU-3DFE	6 3D distances	PNN	N	S	11	7 E	F	91%	N
[75]	BU-3DFE	Distance and angle	PNN	N	S	83	7 E	F	90%	N
[82]	BU-3DFE	Distances computed over a Riemannian-based shape analysis of local patches represented as closed curves	Linear SVM, AdaBoost, polynomial SVM, and RBF SVM	N	S	24	6 E	F	96% (RBF SVM)	N
[81]	BU-3DFE	Feature point displacement (D), texture (T), range values (R), shape index (i.e., curvature) (SI), and 5 multi-scale LBP operators	Bayesian belief network (BBN)	Y	S	19 (M) vs. (A)	6 E	F	87% / 82%	N
[40]	BU-3DFE	Normalized distances and slopes of the line segments connecting certain 3D facial feature points	Multi-class SVM	N	S	83	6 E	F	87%	N
[132]	BU-3DFE	Normalized distance between all 3D points pairs then PCA for dimensionality reduction, then LDA for optimum discriminative K-1 subspace	PNN	N	S	83	7 E	F	94%	N
[84]	BU-4DFE	Displacement vectors	HMM	Y	D	83 A	6/8 AUs	F	81% / 87%	N
[83]	BU-4DFE	Facial level curve + localized chamfer distances evaluation + PCA + LDA	HMM	N	D	N/A	3 E	F	92%	N
[124]	BU-4DFE	Motion based features extracted using non-rigid registration method with sliding window	GentleBoost + HMM	N	D	N/A	6 E	F	92%	N
[167]	in-house	Combining 23 geometric, appearance and surface deformation measurements	A rule-based approach	Y	D	81	4 E	NF	85%	Y

Chapter 3

VT-KFER: A Kinect-based RGBD+Time Dataset for Spontaneous and Non-Spontaneous Facial Expression Recognition

Human facial expressions have been extensively studied using 2D static images or 2D video sequences. The main limitations of 2D-based analysis are problems associated with large variations in pose and illumination. Therefore, an alternative is to utilize depth information, captured from 3D sensors, which is both pose and illumination invariant. The Kinect sensor is an inexpensive, portable, and fast way to capture the depth information. However, only a few researchers have utilized the Kinect sensor for the automatic recognition of facial expressions. This is partly due to the lack of a Kinect-based publicly available RGBD facial expression recognition (FER) dataset that contains the relevant facial expressions and their associated semantic labels. This chapter addresses this problem by presenting the first publicly available RGBD+time facial expression recognition dataset using the Kinect 1.0 sensor in both scripted (acted) and unscripted (spontaneous) scenarios. Our fully annotated dataset includes seven expressions (happiness, sadness, surprise, disgust, fear, anger, and neutral) for 32 subjects (males and females) aged from 10 to 30 and with different skin tones. Both human and machine evaluation were conducted. Each scripted expression was ranked quantitatively by two research assistants in the Psychology department. Baseline machine evaluation resulted in average recognition accuracy levels of 60% and 58.3% for 6 expressions and 7 expressions recognition, respectively, when features from 2D and 3D data were combined.

3.1 Introduction

This chapter presents the first public Kinect-based RGBD+time facial expression recognition dataset, VT-KFER [2], that includes, for each subject, sequences of both scripted (acted) and unscripted (spontaneous) expressions of the six basic emotions (along with the neutral expression), as shown in figure 3.1. The database consists of different data modalities such as 2D, 2.5D (i.e., depth maps), 3D, and sequence-based (i.e., dynamic) face data. Figure 3.1 displays samples of the 2D and 2.5D modalities of the six basic expressions, plus neutral, from VT-KFER dataset. These expressions are generally accepted to be universal [183]. The dataset also is the first Kinect-based FER dataset that includes children. Currently, the dataset includes seven children of age 10 to 17. This dataset not only can be utilized for facial expression recognition system evaluation but also can be used for face recognition and animation system evaluations.

The rest of this chapter is organized as follows: section 3.2 provides detailed information about the dataset participants; section 3.3 describes the procedure for recording the scripted and unscripted expressions; section 3.4 describes the dataset contents; finally, section 3.5 and section 3.6 introduce a qualitative evaluation for our dataset and the conclusion.

3.2 Participants

A total of 32 participants were recruited from a multicultural community in the USA. The majority of the participants had no previous experience with facial expression related study or research. Demographic information of the proposed VT-KFER dataset can be seen in figure 3.2. There were a total of 14 males and 18 females with a broad diversity in facial appearance, skin tone, ethnic/racial backgrounds including White/Caucasian (11), Asian (14), Middle East/Maghreb (2), Black/African (3), and Hispanic (2). The majority of the subjects were aged between 19 and 25 (14 subjects), with 7 subjects aged from 10 to 18 years old. Seven subjects had a beard, mustache, or short facial hair. One subject wore glasses. 41% (13) of subjects reported playing video games such as Wii and Xbox, 22% (7) had never played video games, and 34% (11) reported playing video games in the past, but not currently.

3.3 Protocol

Our dataset is composed of two types of expressions, unscripted (i.e., spontaneous) and scripted (i.e., acted), for each subject. To record this data, two experimenters were present for each subject (one to operate the machine and one to assist the subject). The experimenters followed a particular protocol for 1) seating each subject and preparing him/her for the session, 2) recording the unscripted expressions, and 3) recording the scripted expressions. Before collecting any data, a consent form (assent for children) along with behavioral scales were always done prior to any data

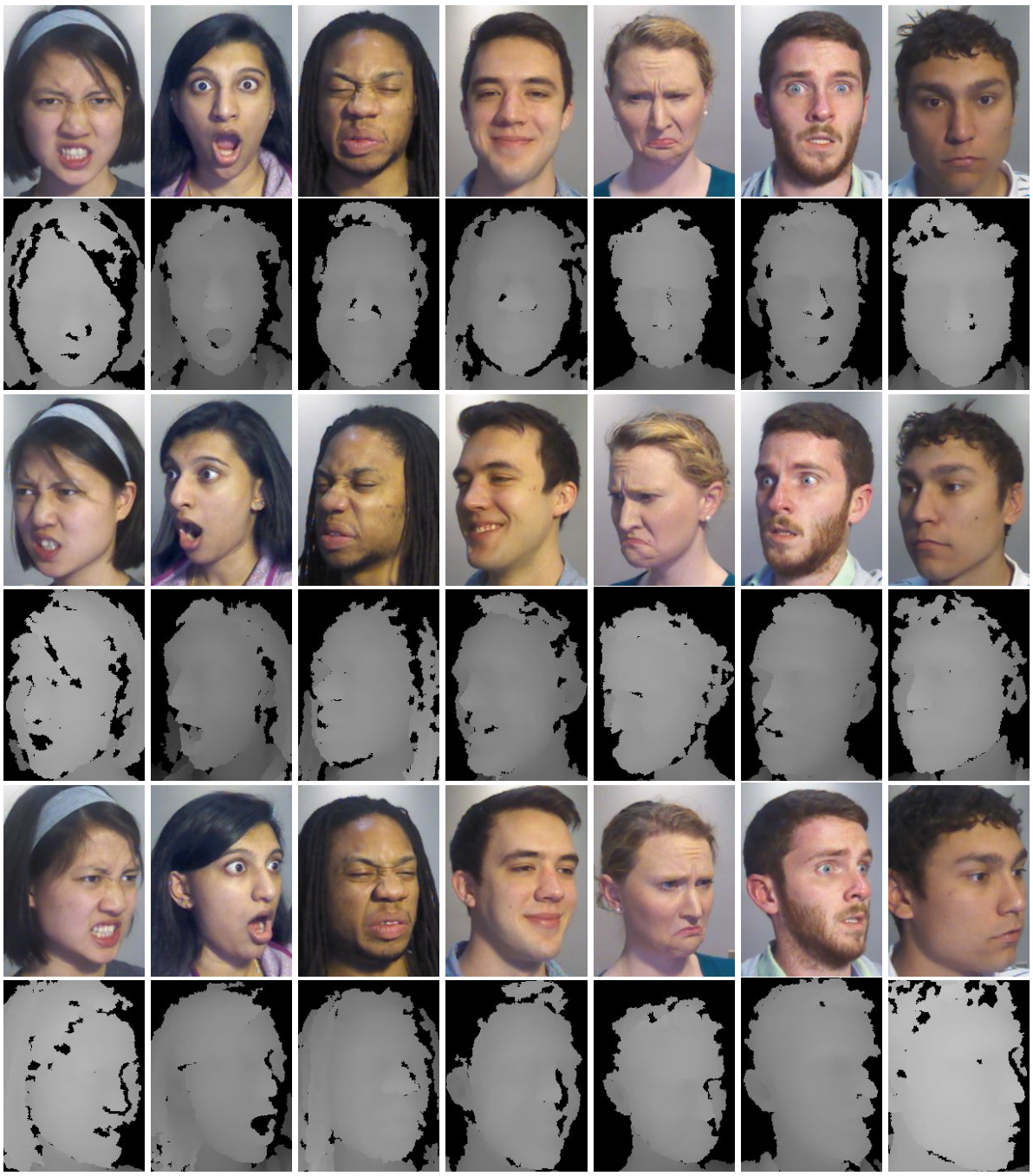


Figure 3.1: Samples of the RGB and corresponding depth images of the six basic facial expressions plus neutral, as contained in the VT-KFER database [2], in frontal ($0°$), left ($45°$), and right ($-45°$) poses. Shown left to right are anger, surprise, disgust, happiness, sadness, fear, and neutral.

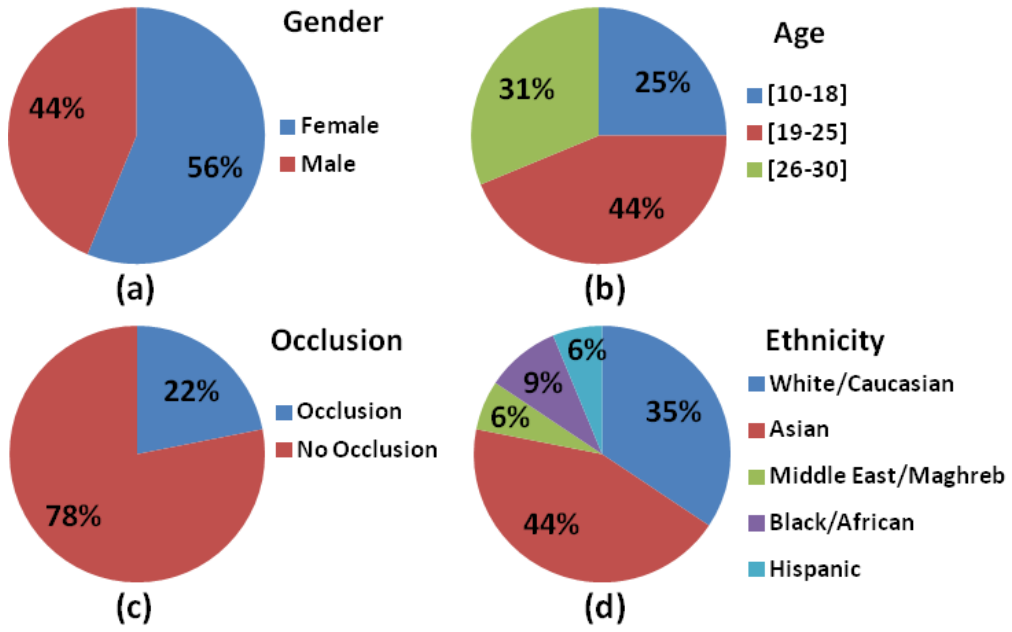


Figure 3.2: Demographics of the VT-KFER dataset by (a) gender, (b) age, (c) occlusion by facial hair, and (d) ethnicity.

collection.

3.3.1 Seating the Participant

After greeting a participant who entered the room, he/she is informed about the experiment: *"Now, I will ask you to watch a computer monitor and make different expressions."* Then, the participant is seated until the automatically detected face mesh captured by the Kinect system was affixed and then we asked them to move closer to the screen slowly until they were about 60cm in front of the screen. We made sure that participants' eyes were in the middle of the screen by moving the chair up or down. We reminded the participant that he/she can stop the session at any time. Finally, we said: *"Please remember that there are no right or wrong answers. Please just try your best. Do you have any questions before we begin?"*

3.3.2 Recording the Unscripted Expressions

The recognition process started by recording the unscripted (spontaneous) expressions after giving these instructions to the participant: *"You will see some pictures, some of which may make you feel certain emotions. Please look at the screen and show (using your face) the emotion. Please do not turn away from the screen or cover your face with your hands while you are seeing the pictures."*

The displayed images were selected, in advance, from the International Affective Picture System (IAPS) [184]. These images were first rated independently by us for emotion intensity, and the ones that had highest agreement between researchers were chosen. A set of 53 images were selected and displayed for 3 seconds each to evoke the six expressions. Three images were picked to evoke the anger expression, and ten images were picked for each of the remaining expressions. The selected images for each expression were sorted in ascending order based on level of evocation of each expression, where level 1 is least evoking and level 10 (3 for anger) is most evoking of emotion. These images were displayed in random order with respect to the emotion aimed to be evoked, but in order of least to most emotion evoking.

3.3.3 Recording the Scripted Expressions

To record the scripted facial expressions, the subjects were asked to show certain expressions by giving them some instructions. These instructions can, generally, be of different types including verbal instructions, text, images, video, live demonstration and combinations of the above. For this dataset, three types of instructions were employed: verbal, images (taken from the NimStim Face Stimulus Set [1]), and live demonstration, given in this order to the subjects. Each type caused a different expression intensity. By intensity I mean the relative degree of displacement, away from a neutral or relaxed facial expression, of the pattern of muscle movements involved in emotional expressions of a given sort [185]. Instructions associated with each level were as follows:

Level 1: verbal (repeated up to 2 tries). The participants were asked to make the six expressions in 3 different poses (front, left, and right), by saying: *"I am going to ask you to make different facial expressions. Please make a (happy) face. Show me (happy)."*

After the frontal poses, the participant was asked to rest his/her face and slowly turn to the left. Then, the participant was asked to slowly turn back to the center and continue to the right.

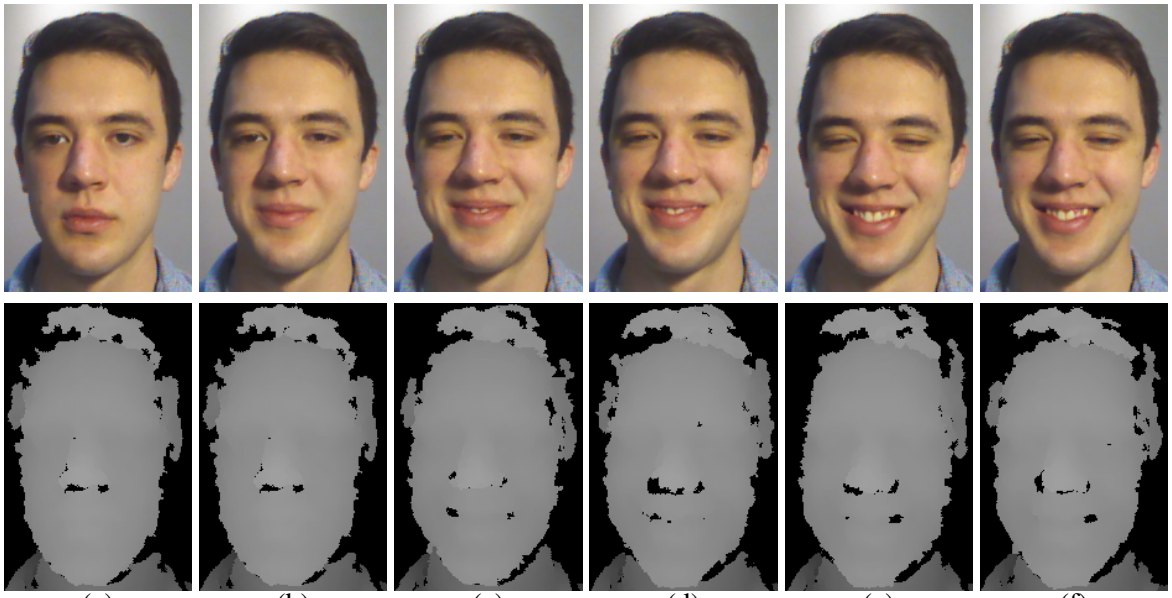
Level 2: images (repeated up to 2 tries). On the screen, a sample expression picture was displayed to the subject, for each expression, and he/she was asked to imitate it (again in 3 poses). The following instructions were used: *"Now please make a (happy) face like it is shown in the picture."*

Level 3: live demonstration (repeated up to 2 tries). The experimenter demonstrated the expression to the subject, and then he/she was asked to imitate the demonstrated expression. The following instructions were used: *"Please make the facial expression I am making. Like this...."*

The order of expressions performed by subjects was selected randomly for each subject. The subjects were not given any prior knowledge of this order.

3.4 VT-KFER Database Contents

VT-KFER includes both scripted and unscripted expressions for 32 subjects. The scripted portion of the dataset is composed of 1,956 sequences of RGB images and depth maps for the six facial expressions in 3 poses, frontal, right, and left. Each sequence starts with a neutral expression, then an onset, followed by a few frames of expression development, and ends with an apex expression. The onset of the action is when the muscular contraction begins and increases in intensity. The apex is usually where the intensity reaches a stable level. The average number of frames per sequence is 6, with maximum of 61 frames and minimum of 2 frames including the neutral. Figure 3.3 provides an example of a frontal happiness facial expression sequence along with its corresponding depth map sequence. The neutral expression is shown in figure 3.3a, then an onset expression is shown in figure 3.3b, then the expression development is shown in figures 3.3c to 3.3e, and finally the apex expression is shown in figure 3.3f. The total number of scripted expressions is 12,317 frames



(a) (b) (c) (d) (e) (f)

Figure 3.3: Sample sequence of frames for a frontal happy face with corresponding depth map. The sequence starts with a neutral face, then an onset frame until it reaches the apex expression.

of the 6 expressions (plus neutral). This data includes expressions performed in all 3 poses. Figure 3.1 shows an example of these three poses from our dataset. Also, VT-KFER includes expressions in 3 different intensities. Figure 3.4 illustrates the three intensities of frontal disgust and anger expressions for the same subject.

Our unscripted dataset includes 32 sequences, one for each subject, that includes the subjects' facial expressions during a session of displaying the 53 IAPS pictures to them in the order described in section 3.3. For each frame, the dataset records the time stamp, the RGB image, the depth map, the depth to color frame mapping, the 3D and 2D facial landmarks, and the face location in the 2D image. There are a total of 61,374 unscripted frames, with an average of 2,116 frames per subject.

(a) (b) (c)

Figure 3.4: An example of the three intensity levels of the disgust (top) and anger (bottom) expressions for the same subject. (a) Subject expression with verbal instructions. (b) Subject expression when an image was displayed to imitate. (c) Subject expression with live demonstration.

Figure 3.5 presents a part of an unscripted sequence recorded while a sequence of disgust stimulus images was displayed to the subject.

Each frame in the dataset includes 121 automatically detected 3D facial landmarks and their corresponding projected 2D keypoints on the RGB images. These keypoints represent the salient keypoints on subjects' faces such as nose tip, inner eye corners, eye browse, and mouth corners, for all poses and intensities. Example landmarks are displayed in figure 3.6 in various poses.

3.5 Evaluation

3.5.1 Human Evaluation

A manual evaluation of the scripted dataset was performed by two research assistants. Each participant's expression was ranked from 0 to 2 based on how well the participant expressed the appropriate emotion. The research assistants coded whether the emotion was appropriate and if yes, how well the participant portrayed the emotion. The rank 0 was assigned when participant did not make any expression, 1 was assigned when participant made a partial emotion, and 2 was assigned when the participant made a full, appropriate expression. If the participant expressed a different emotion, the encoders reported what expression was recognized. The average ranking of participants' expressions according to each intensity is 1.6, 1.8, and 1.9 for verbal stimulus, image stimulus, and live demonstration, respectively. The results for the 32 subjects show enhancement in

Figure 3.5: Part of an unscripted sequence for a subject while disgust stimulus images were displayed.

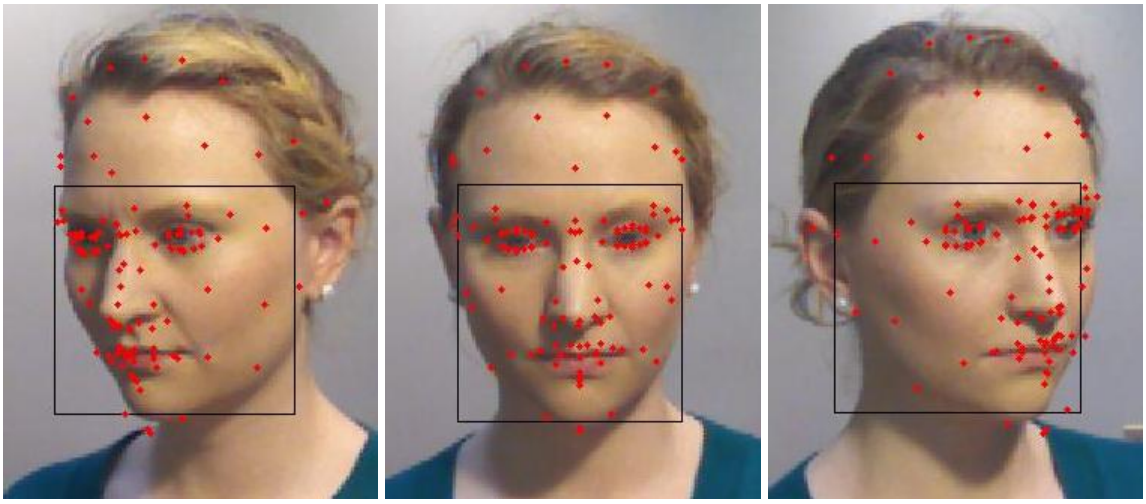


Figure 3.6: The 121 landmarks, automatically detected by the Kinect SDK (highlighted with red dots) in frontal and non-frontal poses, along with automatically detected face locations (highlighted with black boxes).

expression ranking with increased intensity level of instructions. The average recognition accuracy of the two raters was 93.1% for the 6 expressions. Cohen's kappa statistic (κ) was calculated to determine whether agreement between the two raters' judgments on what emotion the participant portrayed was better than chance [186]. There was very good agreement between the two raters' judgments, $\kappa = 0.86$ (95% confidence interval, 0.826 to 0.893), $\rho<0.0005$. Perfect agreement is indicated by $\kappa = 1$, whereas $\kappa = 0$ indicates agreement equivalent to chance. A value of κ higher than 0.81 indicates substantial agreement.

3.5.2 System Evaluation

To automatically validate our FER database, experiments were conducted on face expression recognition using these three baseline feature sets:

1) 2D local binary pattern (LBP): To test the 2D data, LBP features were utilized as described by Shan et al. [6]. To ensure that all facial images are properly localized and aligned, a preprocessing step similar to [187] was applied prior to feature extraction. All RGB scenes were registered to a reference image using control points. After the RGB scene was registered, a 150×110 face region was cropped where the vertical distance between eye location and the upper border of the face bounding box is $0.6d$, where d is the interocular distance. After cropping, the LBP descriptor [6] was extracted by dividing the face image into 6×7 smaller blocks and computing the ULBP descriptor of each block. Then all histograms were concatenated into one, resulting in a feature vector of length 2478. The average accuracy using this approach was 59% and 57.3% for 6 class and 7 class recognition, respectively, using an SVM classifier as described below.

2) Distance-based 3D features: For 3D feature extraction, a distance-based approach computed over the 3D face mesh is adopted as described in [8]. The face mesh is composed of 121 3D keypoints and 206 edges. In contrast to [8], which utilized all of the 206 Euclidean distances, feature reduction is applied to obtain the most discriminative distances only, and to avoid the curse of dimensionality when combined with the LBP descriptor later. Two reduction approaches were utilized. First, a ranking-based feature selection approach automatically selected the best 2 distances that discriminate every pair of classes (e.g., happiness vs. surprise). Then the union of the selected sets is considered as the final reduced feature vector, D_r, of length 18. The features were ranked using 5 criteria [188] (the T-test, Receiver Operating Characteristic (ROC), Bhattacharyya distance, entropy, and Wilcoxon ranking). Only the highly ranked features by the five criteria were considered for recognition. The second reduction approach utilized Principal Component Analysis (PCA) and only the eigenvectors that contained 99% of the energy were selected as the feature vector, D_{pca}, which is of length 11. The concatenation of D_r and D_{pca} is the final 3D feature vector utilized here.

3) Combined 2D+3D features: The 2D and 3D features described above are combined into one feature descriptor.

Classification: For fairness of comparison, this system trained and tested all three feature types using multi-class linear SVMs. Table 3.1 illustrates the leave-one-subject-out cross validation (LOSOCV) average accuracy over all subjects for the 2D, 3D, and 2D+3D features. This system trained the SVMs for both the 6-expression and 7-expression (6+neutral) cases. Results show that combining 2D and 3D features gave the best results in both cases, with average LOSOCV accuracy of 60% and 58.3%, respectively. The confusion matrices (CM) are given in figure 3.7a and 3.7b. The LOSOCV accuracy per testing subject is illustrated in figure 3.7c. The accuracy of each test subject, when trained on the rest of the dataset, varies between 38% and 85% for the 6-class case and 32% to 84% in the 7-class case.

Table 3.1: The LOSOCV average accuracy using 2D, 3D, and 2D+3D features for six and seven expressions using multi-class linear SVMs. The 2D features utilized the LBP features in [6]. The 3D features are based on Euclidean distances from the wire frame mesh in [8].

Feature Type	**6 Classes**	**7 Classes**
2D [6]	59%	57.3%
3D [8, 46]	49.4%	43.7%
2D+3D	**60%**	**58.3%**

3.6 Conclusion

In this chapter, a new RGBD+time facial expression recognition dataset using the Kinect v1.0 sensor has been created. Our dataset includes the six basic expressions and the neutral face for 32 subjects aged from 10 to 30 years in three intensities and various poses. Our dataset includes both scripted (non-spontaneous) and unscripted (spontaneous) expressions for each subject. The availability of both cases for each subject will enable better testing of the capabilities of FER approaches. Moreover, our dataset is the first that includes children, so it may directly aid child-specific studies and applications. Both manual and automated evaluation for the frontal scripted expressions were conducted. The manual evaluation showed enhancement in expression ranking with increased level of instructions. With a kappa coefficient of 0.86, the average recognition accuracy by human evaluators was 93.1% for 6 expressions.The recognition performance achieved using multi-class linear SVMs was 60% and 58.3% for 6 expressions and 7 expressions, respectively, with combined 2D and 3D features.

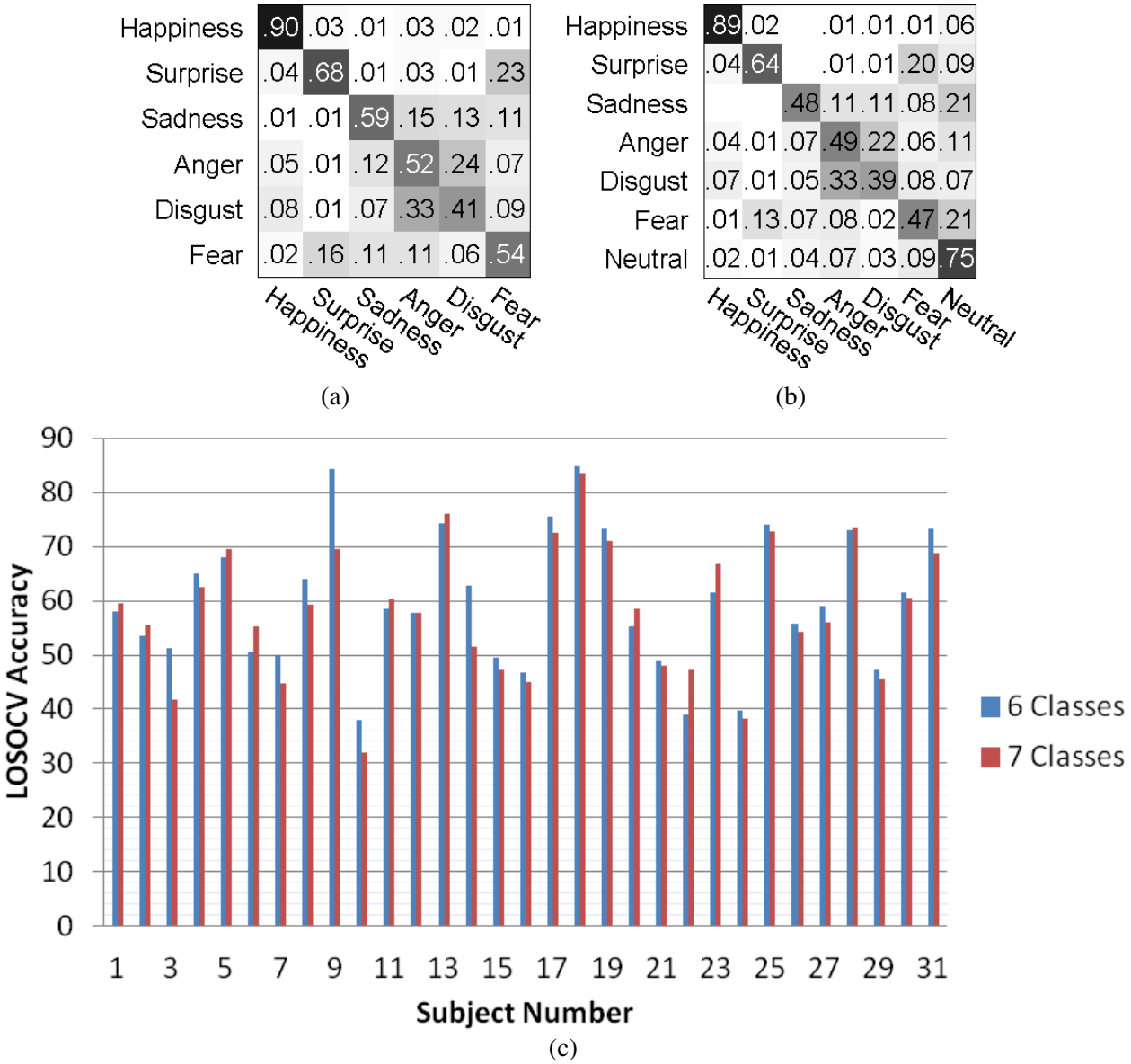


Figure 3.7: (a) Confusion matrix for the case of 6 expressions. (b) Confusion matrix for the case of 7 expressions. Multi-class linear SVMs were used. (c) The LOSOCV accuracy per subject in the 6-expression and 7-expression cases.

Chapter 4

Adaptive Feature Selection and Data Pruning for 3D Facial Expression Recognition using the Kinect

This chapter is concerned with the automatic recognition of pose-varying facial expressions from a low-resolution three-dimensional data sensor, the Kinect. It is an early, preliminary study that was conducted, before VT-KFER dataset was created, on FaceWarehouse dataset and a small in-house dataset of 10 subjects. It introduces a novel, comprehensive framework that employs Delaunay triangulation and a pool of Distance Metrics (DM) for feature extraction. The recognition approach utilizes binary RBF SVMs on DM-based feature matrices. Optimal class models are then chosen automatically, and then utilized in the testing stage. Our experimental results show that automatically tuned DMs for each class outperform a fixed DM approach, especially with non-frontal poses. In addition to the design of the overall framework, this chapter describes the effect of training data pruning, providing insights that could contribute to the reduction of training times for very large datasets.

4.1 Introduction

This chapter describes a new system to recognize the basic facial expressions (happiness, sadness, surprise, fear, anger, and disgust), that can accommodate low-resolution data from the Kinect sensor. Unlike other 3D FER systems [77, 89], this work has particularly addressed the problem of non-frontal poses. Our novel feature extraction/selection approach computes different Distance Metrics (DM) over a 3D triangular mesh, and selects as a feature vector, for every class, the DM that optimizes its recognition results (i.e, adaptive DM-feature selection). As a result, our feature selection is on the category level, not global across all categories as is typical. Using RBF SVM, our system identifies which DM is most suitable for a category by testing its accuracy on a training

set. That feature is then used for that category at test time. The system then makes a final decision through a fusion step that compares the different outputs of the different individual classifiers and gives a final decision based on maximizing the a-posterior probability of the individual SVM classifiers.

As a final step in our assessment, this chapter describes the effects of reduction in size of the training dataset. To our knowledge, no previous work has investigated how much training data is needed for robust 3D facial expression recognition. This chapter presents a comprehensive analysis of the effect of training data reduction on FER system accuracy.

To the best of knowledge, the system described here is the first to utilize 3D sensing from a Kinect sensor for recognition of facial expressions, particularly for varying poses. Previous approaches have considered frontal poses only, or have utilized higher-resolution sensors, or have targeted the problem of face recognition rather than facial expression recognition.

Sections 4.2, 4.3, and 4.4 describe the proposed approach, our experimental results, and concluding remarks, respectively.

4.2 The Proposed Method

Our system is divided into the following stages: 1) feature extraction; and 2) classification.

4.2.1 Feature Extraction

For feature extraction, this system follows a novel distance-based approach that is made of three main steps. First, it applys Delaunay triangulation on the 3D keypoints, provided by the datasets, to get a fixed mesh of triangles, T, that connects the face keypoints. Second, it extracts the edge set, E, from this triangular mesh. Finally, for each edge in E, the proposed system computes its length using 9 DMs, namely: 1) Cosine; 2) Standardized Euclidean; 3) City block; 4) Chebychev; 5) Minkowski; 6) Hamming; 7) Spearman; 8) Correlation; and 9) Jaccard. In other words, a pool of feature matrices, $\{F_{DM_1}, ..., F_{DM_9}\}$, is created in which F_{DM_i} is the feature matrix of the dataset extracted using a corresponding DM. Each matrix F_{DM_i} is of size $n \times m$, where n is the size of the dataset and m is the length of each feature vector (i.e, number of the edges in the edge map, E, extracted from the triangular mesh, T).

For good performance, normalization of these measured values is needed. For each face, all the measured distances are normalized by the Euclidean distance between the inner eye corners. Normalization is a crucial step because it allows us to have relative measurements that are independent of the scale of the face.

4.2.2 Classification

To solve a multi-class classification problem, one way is to use a one-vs-all approach, in which the system trains K binary classifiers, $f_k(x)$, $k = 1, ..., K$, where the data from class k is treated as positive, and the data from all the other classes is treated as negative. However, this can result in regions of ambiguously labeled input space [189]. A common alternative is to pick $y(x) = \arg\max_k f_k(x)$. This approach is simple to implement and works well in practice but it suffers from the class imbalance problem. To have balanced classifiers, for each classifier C_i, the system employed cross validation (CV) to randomly select the negative examples nearly equally from each class C_j, where $j \neq i$ and with total size equal to the positive examples. In our specific case, where the system has large K when using the FaceWarehouse dataset, using one-vs-all is more computationally efficient over other approaches such as one-vs-one. This later approach utilizes $K(K-1)/2$ binary discriminant functions, one for every possible pair of classes. A voting approach is then used to select the final class. One-vs-one has less ambiguously labeled input space but worse computational complexity.

Our approach is divided into three main steps: training, model selection, and testing. *First*, the system utilized the LibSVM [190] library to train $K \times 9$ binary (i.e., one-vs-all) Gaussian kernel RBF SVM classifiers, $M_{i,j}$, $i = 1, ..., K$, $j = 1, ..., 9$. K is the number of classes (expressions) in the dataset, and 9 represents the number of DM employed for feature extraction as described in section 4.2.1. The RBF kernel was selected because it is more flexible than the linear or polynomial, and it is popularly used to solve many classification problems. Each model, $M_{i,j}$, represents a binary classifier for expression i that has been trained using features computed using DM j. The training involves finding the best parameters, C and σ, of each RBF SVM classifier with a Gaussian kernel function. The training will lead to efficient heuristic ways of searching for points in that hyperparameter space. To select the C and σ, this system followed the LibSVM recommendation [191] of using CV over a 2D grid with values $C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$ and $\sigma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$. *Second*, the system employs a model selection procedure that selects the model with the highest training accuracy as the class optimal model, M_i^*, where $i = 1, ..., K$. These models, along with their corresponding feature-type indices, I_i^* (i.e, what DM were used in its training), where $i = 1, ..., K$, are fed into the testing engine. *Finally*, for each test face, the system utilizes the index values provided by the previous step, I_i^*, that indicates which DM is best to be utilized as a metric for feature extraction per class. So the system does not extract all 9 feature vectors for every face in the testing set, but it only extracts the DMs that best fit our models. This list of feature matrices in $[F_1^*, ..., F_K^*]^T$ can have repeated matrices as some DMs can best fit more than one class. The trained SVM models produce an a-posteriori probability $P(c_i|F_i^*)$ that indicates the probability of a given test face in F_i^* being of class c_i. LibSVM employs Platt's method [192] to produce probabilistic outputs. For a final class value, the system further applys a maximization procedure over all these outputs from the SVMs, in which a test face is considered to belong to class c_i if $P(c_i|F_i^*) \geq P(c_j|F_j^*), j = 1, ..., K, P(c_i|F_i^*) > \tau$ (where τ is an empirically determined value), or give -1 as a label otherwise.

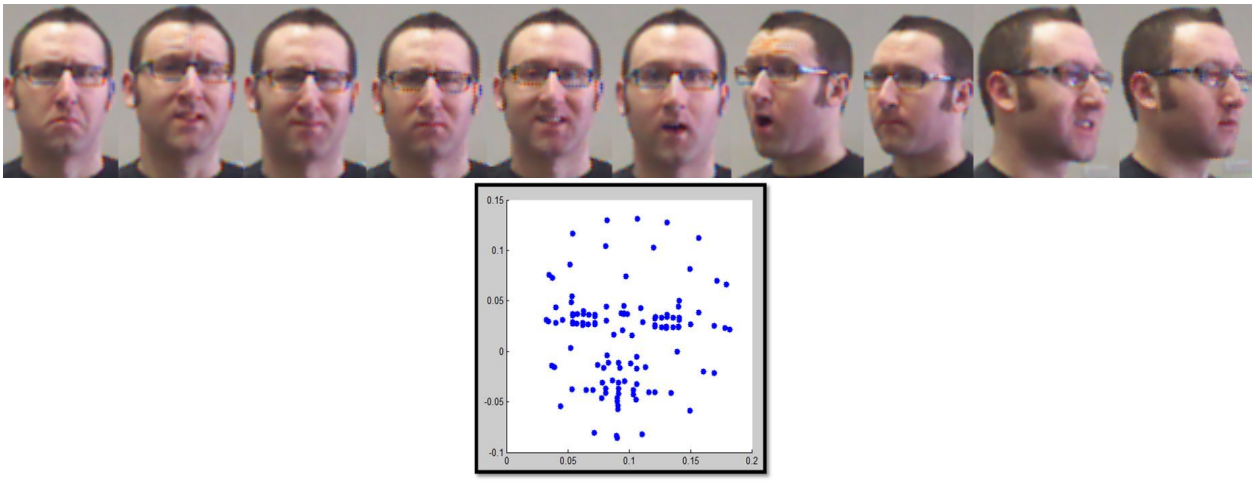


Figure 4.1: Top: Sample images from our in-house dataset for the 6 facial expressions and various poses that it includes. Bottom: For every face in our in-house dataset, 121 3D keypoints, such as eyes and mouth corners, are provided.

4.3 Results and Discussion

4.3.1 Dataset

Since this work was conducted before VT-KFER was built, to recognize the 6 basic facial expressions, it was applied on 3D facial keypoints provided by two RGBD FER datasets, the FaceWarehouse and a small in-house dataset. Many RGBD datasets have been created and used for facial expression recognition, but most of them are not Kinect-based [96–98]. FaceWarehouse (FW) [105] was, at the time this work was conducted, the only publicly available Kinect-based RGBD dataset that includes facial expressions and some action units as well. Since the FW dataset is mainly for computer animation applications, it does not have all the basic facial expressions such as disgust and fear. So, to further examine our system, a small in-house dataset, Our Dataset (OD), was utilized that contains more than 17,000 3D images from 10 subjects with frontal and non-frontal poses of the six universal expressions, namely 1) happiness, 2) sadness, 3) anger, 4) disgust, 5) fear, and 6) surprise. Figure 4.1 show sample images from our in-house dataset. Both datasets provide 3D facial keypoints for each sample. FaceWarehouse provides 73 3D facial keypoints, and our in-house dataset provides 121 keypoints extracted using the Kinect SDK. The proposed approach was tested on these 3D keypoints.

4.3.2 Experimental Setup

For training and testing data selection, the system applied holdout CV for our dataset, which involves random division of the dataset into two non-intersecting parts, training and testing. The experiments were conducted for both datasets with 60%-40% as training-testing partition size.

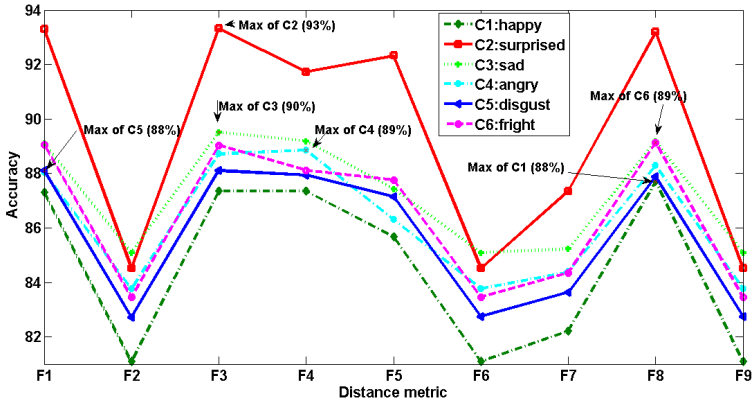


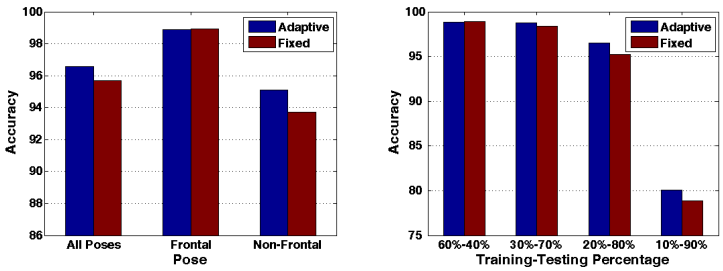
Figure 4.2: Training accuracies using different distance metrics for each of 6 expression classes in our dataset.

For data pruning experiments, different percentages for the training-testing division, such as 30%-70%, 20%-80%, and 10%-90%, which involve smaller training set sizes were employed. The data-pruning experiments were conducted over the frontal-pose data only.

4.3.3 Experimental Results

Class-adaptive model selection: Figure 4.2 shows significant variation in training accuracy for each of the 9 DMs on the 6 classes of OD, which demonstrates our claim of achieving better classification with different DM per class.

System accuracy: The authors in [85] employed Kinect for FER. Using manually located face regions for feature extraction and SVMs for classification, they reported an average recognition accuracy of about 91% for nine facial exercises, recorded for 11 subjects and a total of 696 images. Our experimental results show an Average Recognition Accuracy (ARA) on the FW dataset of 99.97% for 20 classes. The overall accuracy was calculated by taking a weighted average of each class accuracy. The weight of each class is computed as the number of test samples belonging to class C_i divided by the total number of test samples. To show the effectiveness of our adaptive tuning approach, its results were compared to the case where auto selection for best features was not employed and only a fixed DM feature vector was utilized for recognition in all classifiers. The features computed using correlation DM were selected, which gave the highest training accuracy for most of the classes in both datasets, for all binary SVM classifiers. Results for the FW dataset show similar accuracy for the fixed and adaptive approaches. However, for our in-house dataset, our optimal DM selection approach achieved ARA of 96.56% vs. 95.67% using correlation DM (Figure 4.3a). When employed for only frontal-pose data, the systm achieved ARA of 98.87% vs. 98.94% when using correlation DM. However, a significant increase was noted in accuracy for the proposed approach (i.e., 95.12%) over the fixed DM case (i.e., 93.70%) for the case of non-frontal pose data, which reflects the contribution of utilizing adaptive DM feature vector over a fixed one



(a) (b)

Figure 4.3: Adaptive vs. fixed system results for (a) various poses and with (b) data pruning.

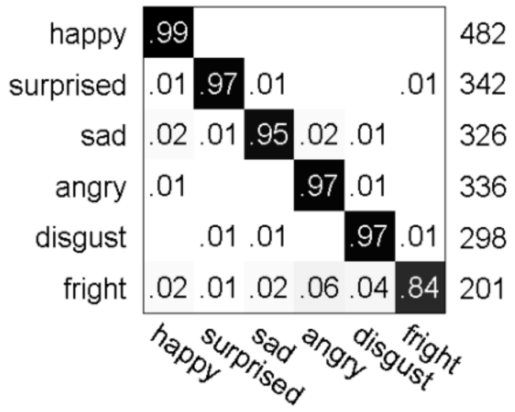
for all classes. Figure 4.4 shows the confusion matrices for our in-house dataset when using only frontal (4.4a) and non-frontal (4.4b) data. It is obvious how challenging the non-frontal poses are, with lower recognition accuracies compared to the frontal cases.

Data pruning: Our data pruning experiment shows that using as little as 20% of the data for training could lead to more than 96% accuracy (Figure 4.3b). Smaller training data sizes will dramatically reduce the system accuracy (as in the case with using 10%-90% ratios). Figure 4.4 displays the confusion matrices for frontal poses for the case of using 30%-70% (4.4c) and 10%-90% (4.4d) of the data for training-testing, respectively. It is obvious how the accuracy was reduced dramatically when only 10% of the data was used for training. Figure 4.4 shows the confusion matrices of the system for cases of frontal (4.3a) and non-frontal (4.3b) poses.

4.4 Conclusion

This chapter has introduced a preliminary study for automatic facial expression recognition using low-resolution, 3D data. Rather than relying on a set of features produced by particular distance metrics (as done by most existing facial expression classifiers), the proposed framework automatically tunes itself to the given 3D data by identifying the best distance metric that maximizes the accuracy for each class. This tuning is particularly important when designing robust recognition systems that have to deal with low resolution data of varying poses.

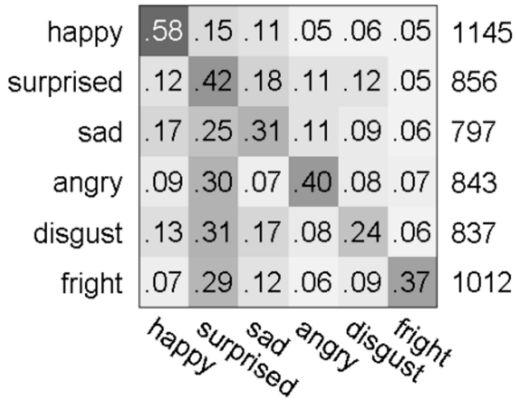
Using the adaptive tuning approach that is presented here, our system has demonstrated improvements in accuracy as compared with a non adaptive case. The performance of the proposed framework was analyzed on two different Kinect-based datasets, FaceWarehouse and a small in-house dataset of 10 subjects (this was before VT-KFER was built), and observed high recognition rate for our approach over the fixed DM ones especially in non-frontal poses. We further demonstrated successful training of the system under reduction of training set size down to 20% of the overall dataset.



(a) Frontal pose with 60% training-40% testing

(b) Non-frontal pose with 60% training-40% testing

(c) Frontal pose with 30% training-70% testing



(d) Frontal pose with 10% training-90% testing

Figure 4.4: System confusion matrices under varying poses and training-testing partitioning. At the right of each row, the number of test faces that actually belong to this class is displayed.

Chapter 5

A Multi-modal Feature Fusion Framework for Kinect-based Facial Expression Recognition using Dual Kernel Discriminant Analysis (DKDA)

This chapter presents a multi-modal feature fusion framework for Kinect-based FER. The framework extracts and pre-processes 2D and 3D features separately. The types of 2D and 3D features are selected to maximize the accuracy of the system, with the Histogram of Oriented Gradient (HOG) features for 2D data and statistically selected angles for 3D data giving the best performance. The sets of 2D features and 3D features are reduced and later combined using a novel Dual Kernel Discriminant Analysis (DKDA) approach. Final classification is done using SVMs. The framework is benchmarked on the VT-KFER dataset, which includes data for 32 subjects (in both frontal and non-frontal poses and two expression intensities) and 6 basic expressions (plus neutral), namely: happiness, sadness, anger, disgust, fear, and surprise. The experimental results show that the proposed combination of 2D and 3D features outperforms simpler existing combinations of 2D and 3D features, as well as systems that use either 2D or 3D features only. The system also outperforms LDA-transformed and traditional KDA-transformed systems, with an average accuracy improvement of 10%. It also outperforms the state of the art by more than 13% in frontal poses.

5.1 Introduction

Although more advantageous for many applications, depth information from low resolution sensors such as the Kinect 1.0 suffer from noise and inaccurate depth measures [193], especially in non-frontal poses. Therefore, the fusion of 2D with 3D data can address these issues and increase the performance of the recognition system, but at the same time significantly increase the dimen-

sionality and complexity of the problem. Therefore, finding the most representative features that can discriminate the various facial expressions becomes a vital step for robust FER.

This chapter presents a multimodal FER system that extracts the most discriminating 2D and 3D features from data captured by the Kinect 1.0 and then transforms them, independently, in a more efficient feature space where data non-linearity is considered and addressed. The main interest of this chapter is to find the best representation for multimodal data that can be later used in realtime systems. This system utilized the VT-KFER dataset [2] for testing the proposed approach as, to the best of knowledge, it is the most comprehensive Kinect-based FER dataset available today.

The rest of this chapter is organized as follows. Section 5.2 presents our contribution in more detail. Section 5.3 introduces our proposed methodology. Section 5.4 discusses the experimental setup. Section 5.5 illustrates the results and section 5.6 gives a conclusion.

5.2 Contribution

The system proposed in this chapter provides two main contributions: first, this system addresses the problem of fusing high dimensional noisy data that contains severe non-linearities (such as what the Kinect sensor produces). In contrast to previous work where features extracted from different modalities are concatenated and then transformed using a single kernel, the system applies the feature transformation on 2D and 3D features independently using dual Gaussian kernels, to achieve the optimal feature projection for each modality. Then, the system combines the transformed features into a single representation. As the second primary contribution, this chapter presents a novel feature selection approach that is applied to the 3D feature set. The proposed approach combines five ranking criteria into one to select the most discriminative features for the recognition problem.

5.3 Methodology

The proposed system is illustrated in Figure 5.1. It is composed of the following steps: 1) feature extraction, 2) feature selection, 3) feature fusion using DKDA, and 4) classification. In brief, it first extracts 2D and 3D features. Then, it employs a feature selection approach on the 3D features. The resulting 2D and 3D features are then fed to our proposed feature fusion approach. Finally a linear SVM is trained and tested for recognition. The figure indicates the Histogram of Oriented Gradient (HOG) and angles as examples of 2D and 3D features, respectively.

The rest of this section is organized as follows. Section 5.3.1 presents the feature extraction approach. The feature selection approach is introduced in section 5.3.2. The proposed DKDA approach for feature fusion is described in section 5.3.3 where σ_1 and σ_2, the DKDA parameters, are explained. Finally, the employed classification approach is presented in section 5.3.4.

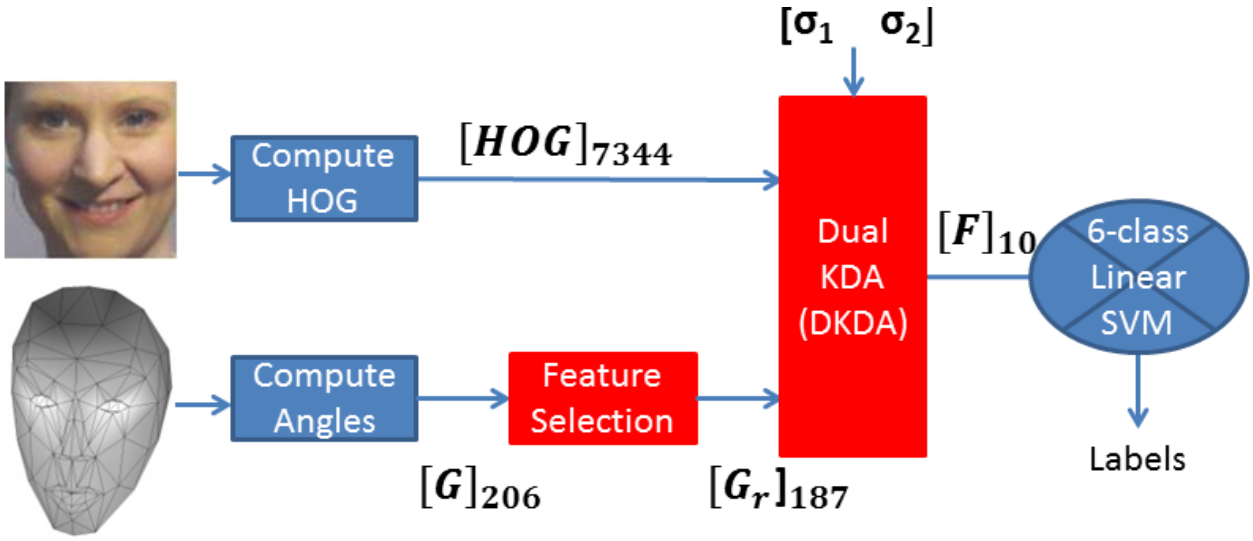


Figure 5.1: System overview. HOG is indicated as an example of 2D features, and "angles" represent the 3D features. G is the set of angles computed over the 3D face mesh, and G_r is the set of statistically selected angles. σ_1 is the Gaussian kernel's parameter of the first KDA that maximizes the recognition accuracy of the HOG features, and σ_2 is Gaussian kernel's parameter of the second KDA that maximizes the recognition accuracy of the selected 3D features, G_r. Vector lengths are indicated as subscripts following brackets.

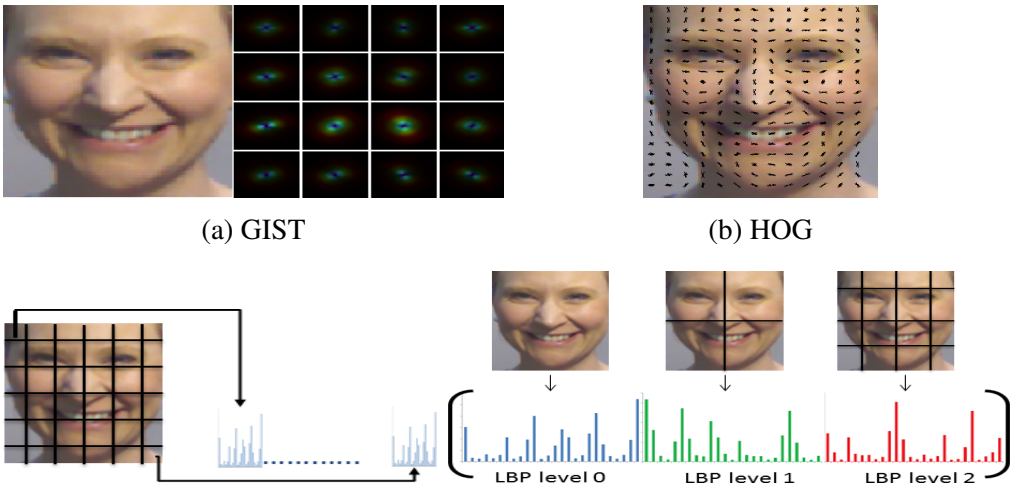
5.3.1 Feature Extraction

For 2D feature extraction, this system utilizes HOG [194], GIST [195], LBP [6], spatial pyramid (SP)-ULBP (SP-ULBP) [196], and PCA-transformed SP-ULBP features. For 3D feature extraction, this system utilizes angles, Euclidean distances, and mean curvature [133]. Each of these features types is described briefly below.

GIST: This is a global image descriptor that provides low dimensional representation of an image, and does not require any form of segmentation. A GIST descriptor is computed over the entire facial image resulting in a feature vector $GIST \in \mathbb{R}^{512}$. Figure 5.2a illustrates a sample GIST descriptor for happiness facial expression.

HOG: The Histogram of Oriented Gradient (HOG) descriptor [194] partitions a given image into 8×8 pixel blocks and computes a histogram of gradient orientations in each block. This results in a descriptor $HOG \in \mathbb{R}^{7344}$. Figure 5.2b illustrates sample HOG features plotted over a happy face image.

LBP: The Local Binary Pattern (LBP) descriptor [6] encodes the local texture and global shape of face images. The LBP descriptor, by Shan et al. [6], is computed by first equally dividing the face images into small 7×6 regions as shown in Figure 5.2c. Then the uniform LBP features are extracted from each sub-region and concatenated into a single, spatially enhanced feature histogram, resulting in the descriptor $LBP \in \mathbb{R}^{2478}$. An LBP descriptor is called uniform if the binary



(a) GIST (b) HOG

(c) LBP (d) SP-ULBP

Figure 5.2: 2D feature extraction on a sample input image of the happiness expression. (a) GIST descriptor visualization; (b) HOG descriptor visualization; (c) LBP descriptor proposed by Shan et al. [6]; and (d) spatial pyramid uniform LBP (SP-ULBP) of 2 levels.

pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. This condition reduces the descriptor dimensionality from 256 to 59 using a neighborhood of $P = 8$ sampling points on a circle of radius $R = 1$ pixel.

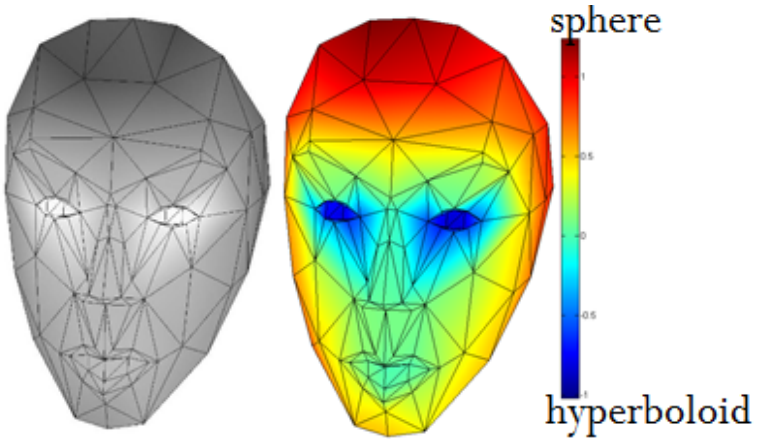
SP-ULBP: This system utilizes a 2-level spatial pyramid of uniform LBP descriptor [197], termed here as *SP-ULBP*, over the face image. The proposed descriptor is composed of the concatenation of the uniform LBP descriptors computed 1) over the whole face (i.e., L_0), 2) over four quarters of the face image with the 4 resulting histograms concatenated (i.e., L_1), and 3) over 16 equally sized blocks of the face image with the 16 resulting histograms concatenated. This step results in a descriptor [*SP-ULBP*] $\in \mathbb{R}^{1239}$. Figure 5.2d illustrates the [*SP-ULBP*] feature extraction process.

PCA-transformed SP-ULBP: After extracting the [*SP-ULBP*] descriptor, principal component analysis (PCA) is applied and the largest eigenvalues containing 99% of the energy are selected, resulting in a descriptor [*SP-ULBP*]+PCA $\in \mathbb{R}^{379}$.

Euclidean distance: (D) is computed over a predefined 3D triangular mesh that includes 206 triangles and 318 edges provided by the Kinect predefined face mesh, and shown in figure 5.3a. This results in a feature vector $D \in \mathbb{R}^{318}$.

Angles: Geometric features such as angles enclose important cues for facial expression discrimination. The system computes the angles on the same triangular mesh, where one angle, selected randomly, is computed per triangle. A feature vector $G \in \mathbb{R}^{206}$, is then computed from each 3D face mesh.

Curvature: The surface curvature is an important representation of 3D facial expression models [96] and an indicator for surface shape classification in range image analysis. The system computes



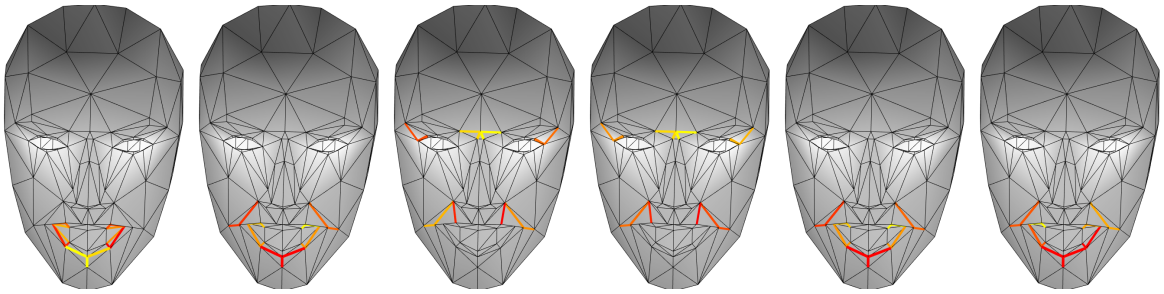
(a) 3D face model (b) Mean curvature

Figure 5.3: (a) The 3D face mesh model utilized in 3D feature extraction. (b) The mean curvature of the sample 3D face in (a).

the mean curvature over the 3D face model. Figure 5.3b illustrates the mean curvature values plotted over a 3D model. The figure is color coded according to the curvature value which encodes the surface shape (e.g., sphere or hyperboloid).

5.3.2 Feature Selection

The goal of the feature selection (FS) step is to find the features that minimize within-class variations of expressions while maximizing between-class variations. This section presents an automatic feature selection approach based on ranking features according to their class separability criteria. The proposed system applies this feature selection procedure on the computed 3D angle features, G. Our approach utilizes five criteria to rank the features, namely: t-test, ROC, Bhattacharyya distance, relative entropy, and Wilcoxon signed-rank test. Each criterion is computed for each pair of classes, independently, to rank the features from most significant to least significant. For each pair of classes, the approach selects the top ranked s features by each criterion. This results in a total of $s \times 5$ non-unique features (i.e., some features are selected as the top s discriminant features by more than one criterion). They are then sorted by their frequency, from the most frequently selected by the five criteria to the least frequently, and then the first s features are selected as the best features that discriminate the corresponding pair of classes. With a total of $\frac{k(k-1)}{2}$ pairs of classes, where $k = 6$ is number of classes, a unique set of features is then composed from the union of the $s \times \frac{k(k-1)}{2}$ features to compose the final discriminating features for the 6-class problem. In this work, $s = 128$ (almost 50% less than G), which gave best performance among the sets of $[2, 4, 8, 2^4, 2^5, 2^6, 2^7, 200]$ features in the benchmarking scenario. This process results in the feature vector G_r of size greater than or equal to s. For example, in the case of frontal data, $G_r \in \mathbb{R}^{187}$. Figure 5.4 illustrates the selection of the $s = 8$ best angles that discriminate happiness vs. surprise. The output of the five criteria are illustrated in figures 5.4a to 5.4e, and the best set



(a) t-test (b) ROC (c) Entropy (d) Bhattach. (e) Wilcoxon (f) Best

Figure 5.4: Examples of detected features. (a)-(e) The best 8 angle features selected using the five criteria to discriminate happiness vs. surprise. (f) The best selected angles using all five criteria. The selected features are color coded using autumn color model based on significance, the more significant the feature is, the darker the color it takes.

of angles selected by our approach is illustrated in figure 5.4f. Each of those criteria is described next.

Statistical hypothesis testing: The t-test [198] offers statistical evidence about the difference of the mean values of a single feature in the various classes. The larger the difference, the more discriminating the feature is. Let x_i, $i = 1, 2, \ldots, N$, be the sample values of the feature in class c_1 with mean μ_1, and for the other class c_2 we have $y_i, i = 1, 2, \ldots, N$, with mean μ_2. N is the length of the feature vector. This approach assumes equal variance for both classes (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$). To decide whether the feature is discriminating or not, the hypotheses H_0 and H_1 is tested:

$$\begin{aligned} H_1 &: \Delta\mu = \mu_1 - \mu_2 \neq 0; \\ H_0 &: \Delta\mu = \mu_1 - \mu_2 = 0. \end{aligned} \tag{5.1}$$

A t-test is performed, to accept either hypothesis, using:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{2}{N}}\sqrt{\frac{1}{2N-2}(\sum_{i=1}^{N}(x_i - \bar{x}) + \sum_{i=1}^{N}(y_i - \bar{y}))}}, \tag{5.2}$$

where $\bar{x}$ and $\bar{y}$ are the unbiased estimates of the mean of the features in the two classes. If the value of t is within a predefined confidence interval, the hypothesis H_0 is accepted and thus the feature is not significant and can be discarded during the selection phase. Otherwise, the hypothesis H_1 is accepted and the feature is significant. In this work, the t-test is used to rank our features based on the t function value instead of just accepting or rejecting the previous hypothesis. Figure 5.4a illustrates the selected best $s = 8$ angles that discriminate happiness vs. surprise using the t-test criterion.

ROC: The ROC is another approach that provides information about the overlap between the classes. We usually are concerned with the area between the ROC curve and the random selection curve. This area varies between zero, for complete overlap, and 1/2, for complete separation,

and it thus it can be utilized as a measure of the class discrimination capability of the particular feature [198]. Figure 5.4b illustrates the selected best $s = 8$ angles that discriminate happiness vs. surprise using the ROC criterion.

Relative entropy (divergence): Relative entropy can measure the class separability of features and thus can be used to rank them. Relative entropy is a form of the Kullback-Leibler distance measure between density functions [198]. The divergence d_{ij} between class c_i and class c_j has the following properties: $d_{ij} \geq 0$, $d_{ij} = 0$ if $i = j$, and $d_{ij} = d_{ji}$. Assuming that classes c_i and c_j are normally distributed with $N(\mu_i, \Sigma_i)$ and $N(\mu_j, \Sigma_j)$, respectively, and the components of the feature vectors are statistically independent, the computation of d_{ij} for the general case can be simplified to:

$$\begin{aligned} d_{ij} &= \tfrac{1}{2} trace\{\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I\} \\ &+ \tfrac{1}{2}(\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j). \end{aligned} \tag{5.3}$$

Figure 5.4c illustrates the selected best $s = 8$ angles that discriminate happiness vs. surprise using this entropy criterion.

Bhattacharyya distance: This assumes that the two class distributions are normal with $N(\mu_i, \Sigma_i)$ and $N(\mu_j, \Sigma_j)$, respectively. It can be computed as follows:

$$\begin{aligned} B = \tfrac{1}{8}(\mu_i - \mu_j)^T(\tfrac{1}{2}(\Sigma_i + \Sigma_j))^{-1}(\mu_i - \mu_j) \\ + \tfrac{1}{2}\ln \frac{\det(\frac{1}{2}(\Sigma_i + \Sigma_j))}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}. \end{aligned} \tag{5.4}$$

Figure 5.4d illustrates the selected best $s = 8$ angles that discriminate happiness vs. surprise using Bhattacharyya distance criterion.

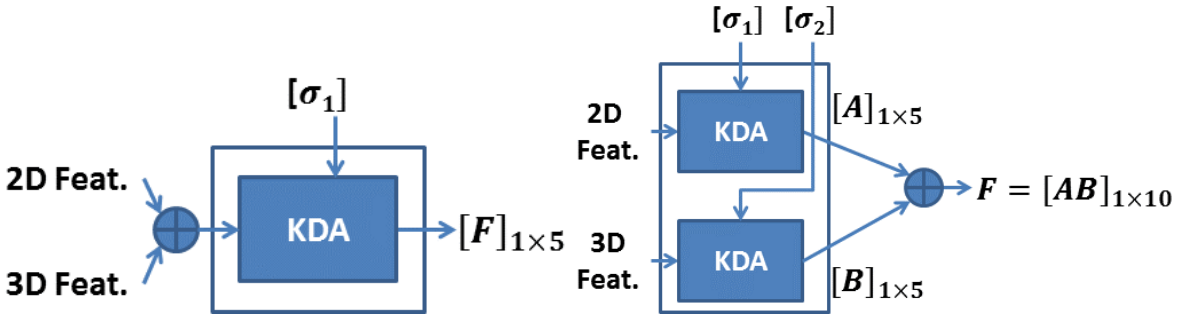
Mann–Whitney U-test (Wilcoxon): This is a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other. It is of greater efficiency than the t-test on non-normal distributions, such as a mixture of normal distributions. It can be computed as follows:

$$U_i = n_i n_j + \frac{n_i(n_i + 1)}{2} - R_i, \tag{5.5}$$

where n_i and n_j are the number of samples of class c_i and class c_j, respectively. R_i is the sum of ranking in sample i. The sum of all the ranks equals $\frac{N(N+1)}{2}$ where N is the total number of observations. Figure 5.4e illustrates the selected best $s = 8$ angles that discriminate happiness vs. surprise using this criterion.

5.3.3 Dual Kernel Discriminant Analysis (DKDA)

This work extends the traditional Kernel Discriminant Analysis (KDA) for feature reduction in which the extracted 2D and 3D features are transformed jointly, and apply KDA on those 2D



(a) KDA (b) DKDA

Figure 5.5: Comparison of KDA and the proposed Dual KDA (DKDA). σ_1 and σ_2 are the Gaussian kernels parameters optimized during training.

and 3D features independently. This results in two disjoint KDA kernels (hence the Dual KDA (DKDA)), with each of these kernels being characterized by parameters σ_1 and σ_2. This parameter is tuned for maximizing the accuracy of each kernel independently. The output of these kernels generates the set of transformed features to be used for FER. The difference between the traditional KDA approach and the proposed DKDA is illustrated in Figure 5.5. The typical KDA approach proposed by Baudat et al. [156, 157] is summarized, which is further extended to build the DKDA system. Assume we have a set of m samples $X = [x_1, x_2, ...x_m]$ with $x_i \in \mathbb{R}^n$ with n features belonging to c classes. KDA extends LDA to the nonlinear case by considering the problem in a feature space F induced by some nonlinear mapping $\phi : \mathbb{R}^n \rightarrow F$. ϕ can be any positive semi-definite kernel function. This work considers the Gaussian kernel function on two samples x and y, represented as feature vectors, is defined as follows:

$$\langle \phi(x), \phi(y) \rangle = K(x, y) = exp(\frac{-\|x - y\|^2}{2\sigma^2}). \tag{5.6}$$

The optimal feature transformation by KDA is given by

$$v^* = \arg\max \frac{v^T S_b^\phi v}{v^T S_t^\phi v}, \tag{5.7}$$

where v is the projective function in the feature space F. S_b^ϕ and S_t^ϕ are the between-class and total scatter matrices in the feature space, respectively. Equation (5.7) can be solved by solving the eigen-problem $S_b^\phi v = \lambda S_t^\phi v$. Because the eigenvectors are linear combinations of $\phi(x_i)$ [156, 157], there exist coefficients α_i such that

$$v = \sum_{i=1}^{m} \alpha_i \phi(x_i). \tag{5.8}$$

In [156], it was proven that (5.7) is equivalent to

$$\alpha^* = \arg\max \frac{\alpha^T K W K \alpha}{\alpha^T K K \alpha} \tag{5.9}$$

which, similarly, can be solved as an eigen-problem using $KWK\alpha = \lambda KK\alpha$, where K is the kernel matrix ($K_{ij} = K(x_i, x_j)$) of size $m \times m$, and W is defined as

$$W_{ij} = \begin{cases} \frac{1}{m_k}, & \text{if } \{x_k, x_j\} \in c_k \\ 0, & \text{otherwise} \end{cases} \tag{5.10}$$

To get a stable solution of this eigen-problem, the matrix KK must be nonsingular. If K is singular, one possible solution proposed by [157] is to use the eigen-decomposition of K as follows

$$K = UDU^T = U_r D_r U_r^T, \tag{5.11}$$

where D is the diagonal matrix of sorted eigenvalues and U is the matrix of normalized eigenvectors associated with D. D_r is the diagonal matrix of the nonzero eigenvalues and U_r is the first r columns of U. By substituting K in (5.9), we get

$$\alpha^* = \arg\max \frac{(D_r U_r^T \alpha)^T U_r^T W U_r (D_r U_r^T \alpha)}{(D_r U_r^T \alpha) U_r^T U_r (D_r U_r^T \alpha)}. \tag{5.12}$$

If we let $\beta = D_r U_r^T \alpha$, by optimizing

$$\beta^* = \arg\max \frac{(\beta)^T U_r^T W U_r (\beta)}{(\beta) U_r^T U_r (\beta)} \tag{5.13}$$

we get the leading eigenvectors of matrix $U_r^T W U_r$. Then, α can be computed as follows:

$$\alpha = U_r D_r^{-1} \beta. \tag{5.14}$$

After computing α, v is computed and normalized using $v^t v = \alpha^t K \alpha = 1$. Then we can compute the projections of test points onto the eigenvectors v using (5.8).

5.3.4 Classification

To solve the multi-class classification problem, this system uses a multi-class linear SVM classifier where one k-class linear SVM is trained and tested using Leave-p-Sequence-Out (LpSO) cross-validation. A set of $100 - p\%$ of the sequence of frames is randomly selected for training, and $p\%$ is used for testing, with no intersection between the two sets. The system uses the implementation of linear SVM provided by the LIBSVM toolbox [190]. The training involves finding the best parameter, C, that represents the penalty parameter of the error term of the linear SVM classifier. In other words, the C parameter tells the SVM how much you want to avoid misclassifying each training example. For large values of C, the SVM optimizer will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. In contrast, a very small value of C will cause the optimizer to look for a larger-margin separating

hyperplane, even if that hyperplane misclassifies more points. The training will lead to efficient heuristic ways of searching for points in that hyperparameter space. To select C, the system follows the LIBSVM recommendation of using cross-validation over a 1D grid with values $C \in \{2^{-5}, ..., 2^{15}\}$.

5.4 Experimental Setup

The experiments are conducted on the VT-KFER dataset. As described in chapter 3, to evoke emotions, VT-KFER employed three types of stimuli: verbal (i.e., I_1), images (i.e., I_2), and live demonstration (i.e., I_3), given in this order to the subjects. Each stimulus type caused a different expression intensity. The dataset includes 3,908 frames of I_1 (2,129 frontal frames), 3,656 frames of I_2 (2,092 frontal frames), and 2,234 frames of I_3 (772 frontal frames). In this chapter's experiments, two expression intensities, I_1 and I_2, in frontal pose only were considered. The data of I_3, where live demonstration was used, was neglected here because the amount of data was very small compared to frontal data of I_1 and I_2. We employed leave-p-sequence-out (LpSO) cross-validation where p=20% (i.e., $100 - p\%$ of the sequences were randomly chosen for training and $p\%$ are left for testing). For each expression, VT-KFER dataset includes various sequences per subject. Using LpSO cross-validation, it possible that a sequence for a particular subject appeared in the training set, and a different sequence for that same subject appeared in the testing set.

5.5 Experimental Results

There are three factors that need to be considered: 1) the 2D feature type, 2) the 3D feature type, and 3) the DKDA parameters (σ_1, σ_2). Section 5.5.1 presents the experiments conducted for selecting the best 2D/3D features and DKDA kernel parameters. Section 5.5.2 presents comparative results to other FER systems.

5.5.1 Control Experiments

2D/3D feature selection: To test which 2D feature type to use in the proposed framework, five 2D feature types are compared, namely HOG [194], GIST [195], LBP [6], SP-ULBP [196], and SP-ULBP+PCA. Figure 5.6 illustrates the average Leave-p-Sequence-Out (LpSO) cross validation accuracy for each feature type in various poses (i.e., frontal and non-frontal) and two expression intensities. Results show that HOG features demonstrated the highest average LpSO cross validation accuracy over the other 2D features.

This chapter tests applying feature transformations such as LDA [153] or KDA [157] on 2D features before fusing them with 3D features. Figure 5.7 shows that the use of HOG features under

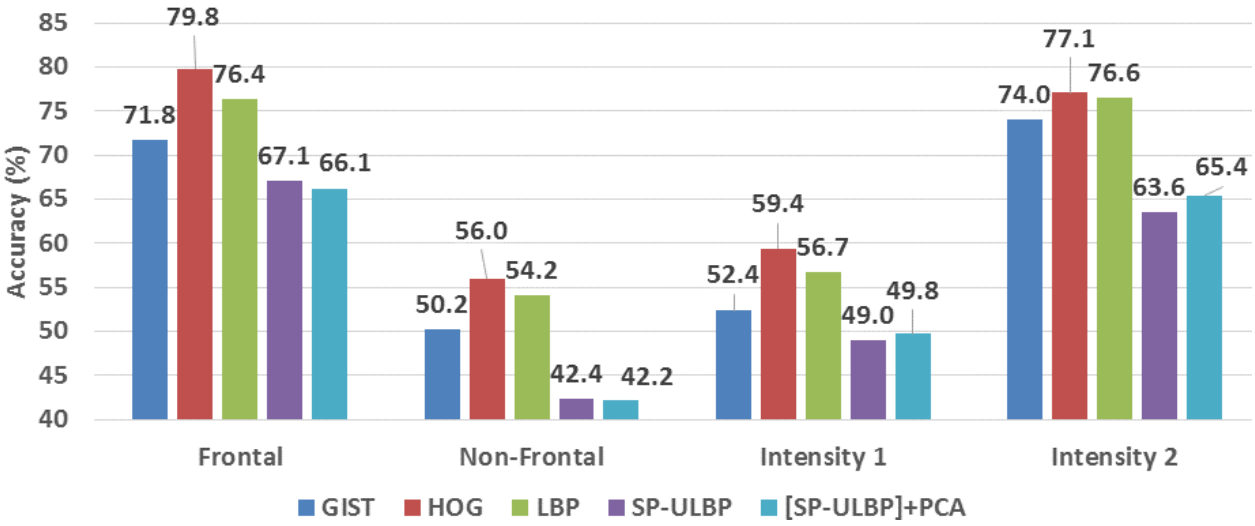


Figure 5.6: Average L-p-SO accuracy of various 2D features under different poses and expression intensities. HOG features resulted in the best performance in all cases. A dramatic decrease in accuracy is noted from frontal to non-frontal poses when only the 2D features are used.

no transformation still achieves the maximum accuracy over the cases when either LDA or KDA was used. In some cases, such as SP-ULBP and SP-ULBP+PCA, feature transformation using KDA enhanced the accuracy over both when no transformation was conducted and when LDA transformation was conducted. However, since HOG features under no transformation achieved the maximum accuracy, the proposed framework uses the HOG features as the 2D feature type.

To test which 3D feature type to utilize in our framework, three 3D features are compared, namely angles, Euclidean distance, and mean curvature [133]. Figure 5.8 illustrates the average LpSO cross validation accuracy for each feature type in various poses (i.e., frontal and non-frontal) and two expression intensities. The results show that, on average, the selected angle features (i.e., Angles+FS or G_r) achieves the maximum accuracy over other 3D features. Note the dramatic decrease in accuracy from frontal to non-frontal when only 2D data is used compared to when 3D data is used. This emphasizes the fact that 3D data is relatively pose-invariant which of great importance to any FER application.

KDA/DKDA σ effect: Figure 5.9 shows the effect of varying the KDA parameter, σ_1, when KDA is applied on the best 2D, 3D, and concatenated 2D and 3D features selected from previous control experiments (HOG, G_r). In addition, the variation of one space DKDA parameter, σ_1, is compared while fixing the other, σ_2, when DKDA is applied on these 2D and 3D features. The results show that 3D-only (i.e., orange line) and 2D+3D (i.e., green line) features under KDA are dramatically decreased when small values of σ_1 are employed (i.e., $1 \leq \sigma_1 \leq 20$) than with higher values of σ_1. Also, 2D-only features under KDA report better results over the 3D-only cases and even the 2D+3D cases. However, DKDA on 2D+3D features reports the highest performance over all σ_1 values. For illustration purposes in the figures, the second kernel parameter σ_2 is fixed. In fact, both σ_1 and σ_2 were varied to exhaustively search in the range $[.6, .7, ..., 1, 10, 15, ..., 130]$ to find

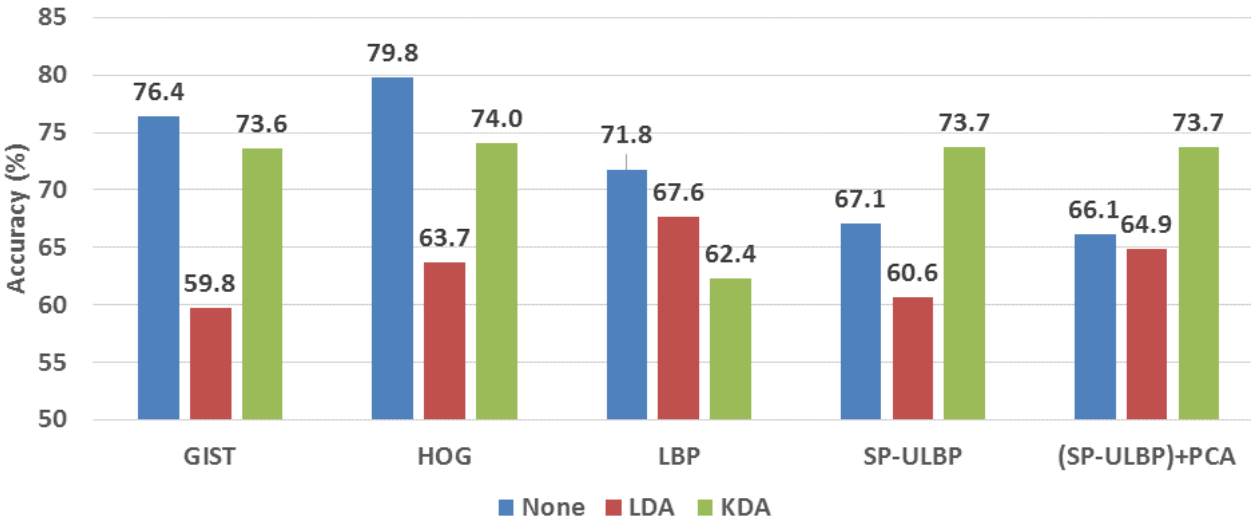


Figure 5.7: Average L-p-SO accuracy of various 2D features under no reduction vs. LDA and KDA in frontal pose. HOG features under no transformation have the best performance.

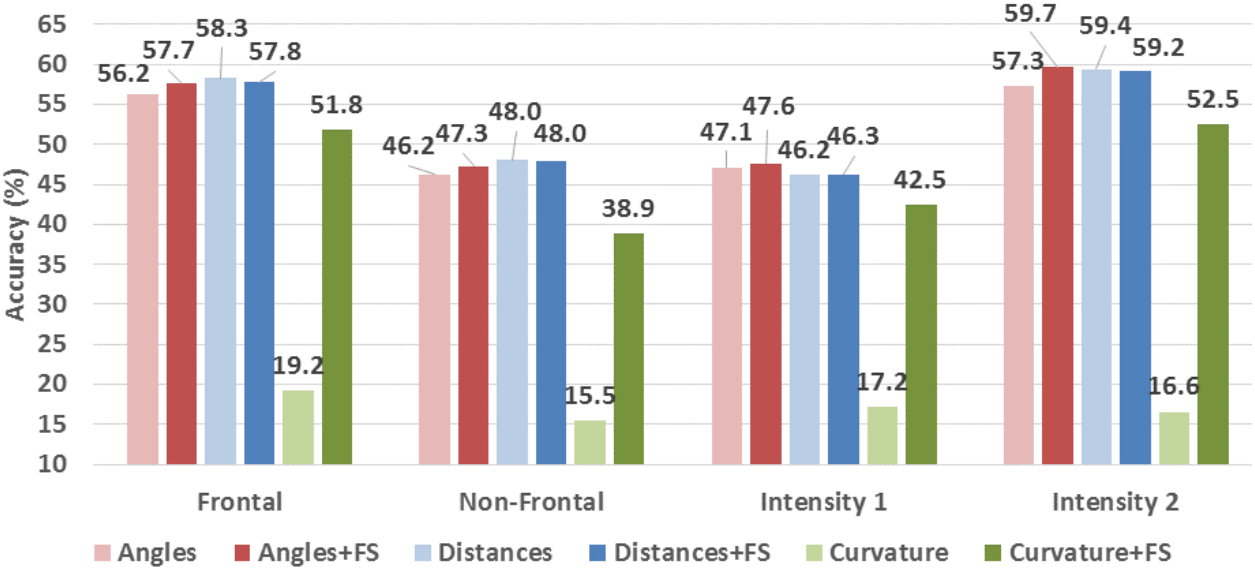


Figure 5.8: Average accuracy using 3D features under different poses and expression intensities. On average, selected angles (i.e., Angles+FS or G_r) result in the best accuracy. "+FS" means that feature selection approach was applied on the corresponding feature.

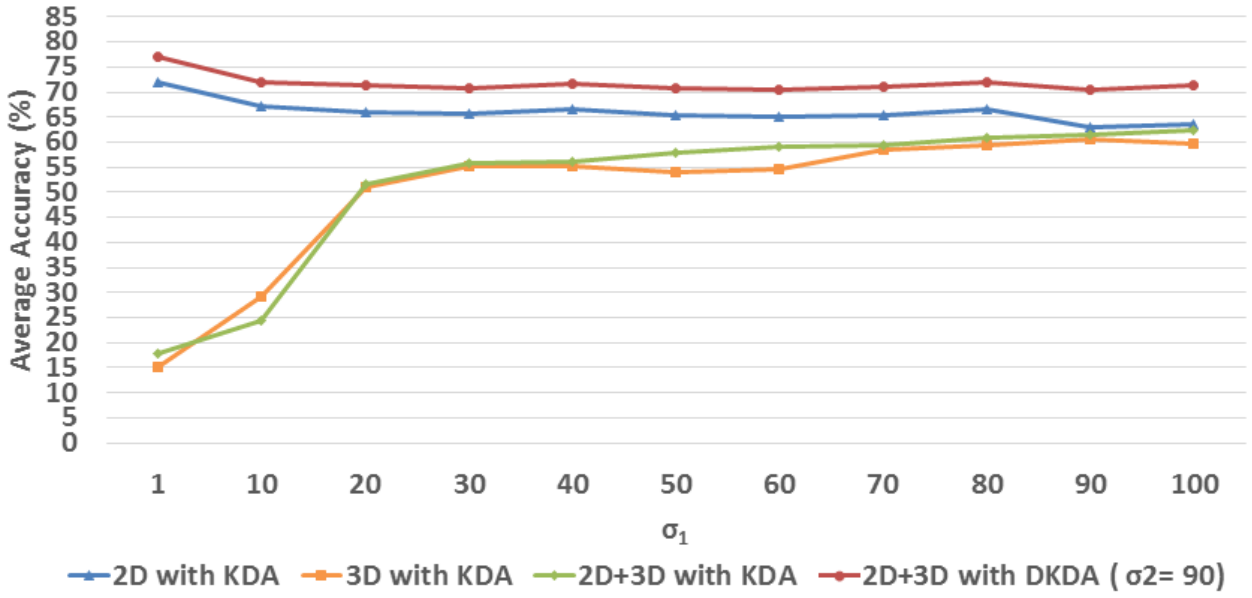


Figure 5.9: The effect of varying KDA Gaussian kernel parameter σ_1 when applied on 2D, 3D, and 2D+3D features vs. when varying DKDA first kernel parameter σ_1 while fixing the other (σ_2 = 90). Note how the proposed DKDA approach achieves better performance over existing alternatives using KDA.

the best DKDA kernel parameters. This approach selects the $[\sigma_1, \sigma_2]$ parameters that achieve the best performance. The results show that best $[\sigma_1, \sigma_2]$ values are [.95, 96], [.81, 93], [.7, 98], [.99, 97] for frontal, non-frontal, I_1, and I_2, respectively.

5.5.2 Comparative Experiments

Experimental results show that our proposed DKDA approach outperforms the typical KDA feature transformation regardless of the 2D and 3D features used. Figure 5.10 illustrates the accuracy results of various 2D+3D features combinations when KDA and DKDA were utilized.

This work also compared the performance of our proposed approach against three feature fusion approaches: 1) concatenating 2D and 3D data, 2) concatenating 2D and 3D features and then applying LDA, and 3) concatenating 2D and 3D features and then applying KDA. Figure 5.11 illustrates the comparative results in various poses and two expression intensities. Our proposed approach outperforms all other approaches in frontal, non-frontal, and first expression intensity. Comparative results were achieved in the case of second expression intensity.

In addition, this system results are further compared to state-of-the-art FER systems on the VT-KFER dataset. Table 5.1 presents different 2D, 3D, and 2D+3D based systems. The effect of DKDA feature transformation and fusion approach is shown in comparison to the state-of-the-art

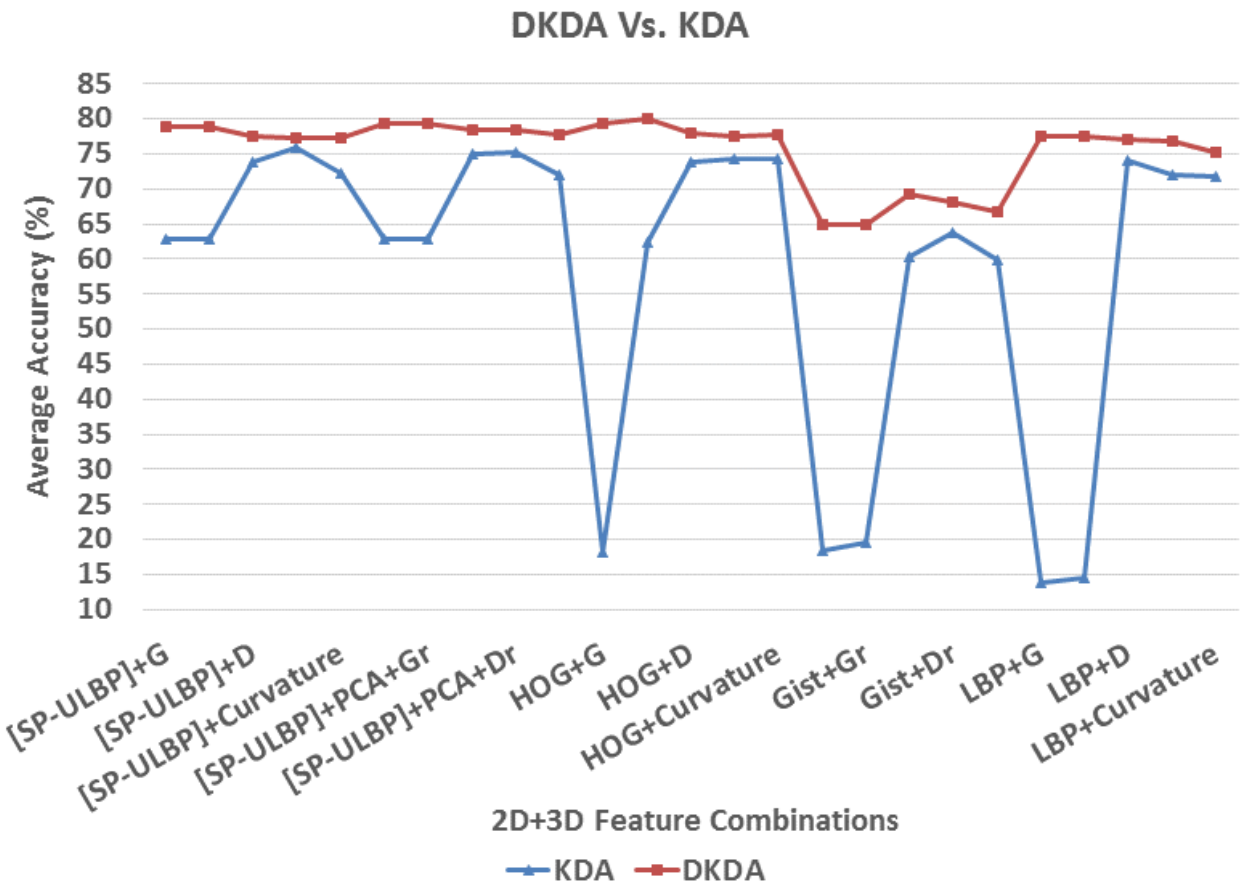


Figure 5.10: Performance of the proposed DKDA approach vs. KDA. With any combination of 2D+3D feature type, our proposed DKDA approach is better than KDA.

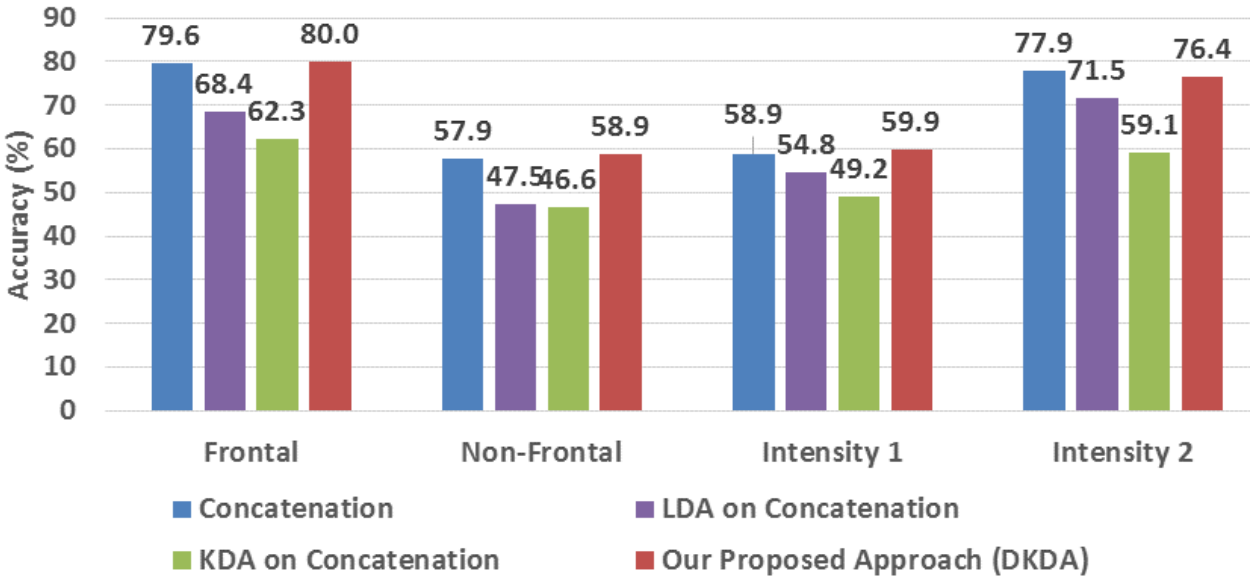


Figure 5.11: Fusion of HOG and selected angles (G_r) using DKDA vs. feature concatenation, LDA-based concatenation, and KDA-based concatenation in frontal, non-frontal, and two expression intensities. DKDA outperforms other feature fusion approaches in frontal, non-frontal and first expression intensity with competitive results in the second expression intensity. Note that the pose-base data includes both intensities while the intensity-based data includes frontal pose data only.

Table 5.1: Quantitative comparison with state-of-the-art FER systems on VT-KFER dataset using leave-p-sequence-out. These results were achieved using frontal poses data only.

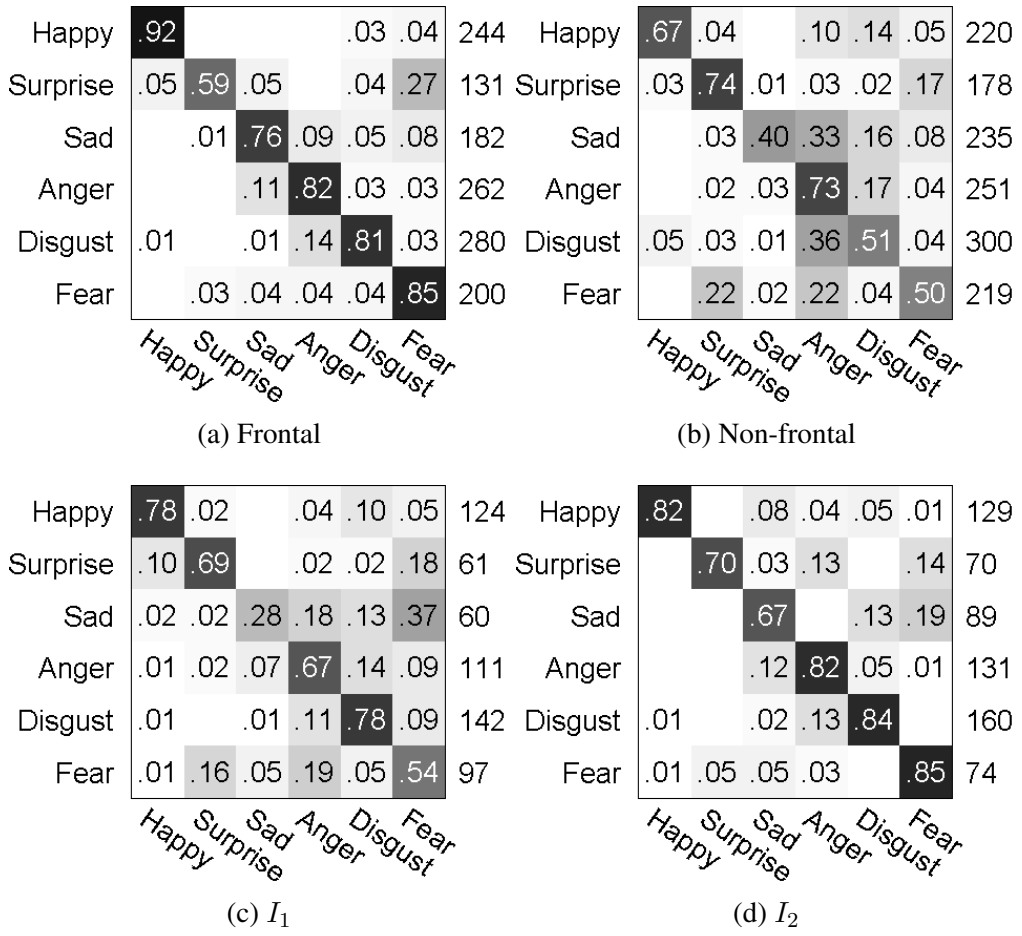
System	Modality	Average Accuracy
[8]	3D	49.4%
[6]	2D	59%
[2]	2D+3D	60%
[196]	2D	67%
Proposed System	**2D+3D**	**80%**

methods. The proposed framework using DKDA outperforms all other systems with significant difference.

The confusion matrices for frontal, non-frontal, I_1 and I_2 for our proposed system are shown in Figure 5.12. The non-frontal pose expressions are more confused with each other than the frontal cases. For example, 14% of the disgust expressions were confused with anger in the frontal pose, while this percentage increased to 36% with non-frontal poses. On the other hand, the increase of expression's intensity of I_2 data over I_1 data resulted in higher accuracy for I_2 data over the I_1 data, especially the anger and sadness expressions.

5.6 Conclusion

This chapter presented a novel Kinect-based multimodal FER system using dual kernel discriminant analysis (DKDA). In contrast to existing feature fusion approaches where 2D and 3D features are concatenated and then transformed, DKDA utilizes dual Gaussian kernels to transform and then combine 2D and 3D features independently. A novel feature selection approach was also presented where only the most discriminant 3D features are selected. When tested on various combinations of 2D and 3D features, the proposed system significantly outperforms other feature reduction approaches in all poses and expression intensities. The system utilized a multi-class linear SVM for classification, and Leave-p-Sequences-Out (LpSO) cross validation for training. Using HOG and selected 3D angles, the proposed DKDA-transformed features resulted in average recognition rates of 80% and 59% on frontal and non-frontal poses, respectively. With varied expression intensities, a 60% average accuracy was achieved for expression intensities evoked by verbal instruction, and 76% in expression intensities evoked by seeing images. The proposed approach outperformed other FER systems by more than 13% for frontal pose only.



(a) Frontal (b) Non-frontal

(c) I_1 (d) I_2

Figure 5.12: Resulting confusion matrices from our proposed approach in case of (a) frontal, (b) non-frontal, (c) I_1, and (d) I_2.

Chapter 6

Dynamic 3D Facial Expression Recognition using the Kinect

The system presented in previous chapter, although has showed great results on VT-KFER, the computation of 2D feature such as the HOG is computationally exhaustive and requires long time (around 4sec per image) on regular PCs, which may not be applicable for realtime operation that runs 30 fps. This chapter describes an innovative facial expression recognition (FER) system that operates at realtime rates using three-dimensional (3D) data only. The emotions of interest are happiness, sadness, anger, fear, surprise, disgust, and neutral. Taking sensor data from the Microsoft Kinect, the system extracts geometric features from a triangular-mesh representation of a person's face. Relatively simple feature types are utilized in order to achieve the desired processing speed. Using a subset of these features, the system identifies the emotion being expressed. Prior to training, feature selection was conducted based on a statistical evaluation of ability to discriminate pairs of expressions. The chosen features were used to train a pool of multiclass classifiers (nearest neighbor classifiers and support vector machines) on individual frames from the Kinect. After training the individual classifiers, seven different subsets of the classifier pool were identified as being best for each of the seven expressions. During operation, each expression is considered in turn, using its particular model (subset of the classifier pool). In the interest of robust operation, typically several input frames in sequence are processed independently, and a voting procedure is employed to make the final decision concerning the emotion being expressed. Training and testing were performed with the VT-KFER dataset, using a "leave-p-sequence-out" and "leave-p-subject-out" cross-validation approaches, where $p = 20\%$. In "leave-p-subject-out" cross-validation, $100-p\%$ of the subjects are selected for training and the rest for testing. All the sequences of the subjects that were selected for training are in the training set and the sequences in the testing are of subjects that the system has not seen before. One the other hand, "leave-p-sequence-out" cross-validation divides the dataset sequences, with no consideration to subjects, into $100-p\%$ for training and $p\%$ for testing. This results in testing sequences that are of subjects the system may has seen before. Additional testing was performed using an in-house dataset of 20 children with ages ranging from 9 to 12 years. For testing using VT-KFER, the average system accuracy was 66.7% using leave-

sequence-out and 57.3% using leave-subject-out. For the dataset of children, the average accuracy was 68.7% for a reduced set of emotions. In addition to recognition of different expressions, the system can operate in a mode that verifies whether a particular emotion is being expressed. In this verification mode the decision by the system is binary, and system accuracy rises to 89.2% using leave-sequence-out on VT-KFER with the full set of emotions, and to 72.5% for the in-house dataset with the reduced set of emotions. Tests conducted using VT-KFER indicate that the new approach outperforms previous FER systems when using 3D data alone by more than 6% compared to other 3D-based approaches and 9% compared to 2D+3D approaches. For further validation of the proposed FER approach, this chapter compares the automatically extracted 3D features with Action Units (AU) that have been proposed as part of the EMFACS. With the exception of one emotion (disgust), the geometric features that were identified as most significant show good correlation with at least one AU that is associated with each emotion.

6.1 Introduction

In recent years, a few researchers have developed 3D FER systems using the Kinect, which is fast and relatively inexpensive [2,8,120,125,126,134,199]. To our knowledge, none of the previous 3D FER systems have considered AU analysis or have been tested in realtime situations on children.

This chapter introduces a dynamic (i.e., sequence-based) realtime FER system that automatically recognizes the basic expressions that were illustrated in Figure 3.1. This system is based on 3D sensing alone, and employs statistically selected geometric features to improve system performance. In order to decide what emotion is being expressed, the system employs a 2-step voting approach over a pool of multiclass classifiers. This approach has been tested two Kinect-based datasets that include both adults and children. The inclusion of facial expressions of children is unusual, having received very little attention in the computer vision community [200].

Another unusual aspect of this work is a comparison with distinct muscle activations known as Action Units (AU), which were tabulated as part of the Facial Action Coding System (FACS) [9,201]. Of the 44 AUs introduced by the FACS, a set of 18 AUs has been associated with common facial expressions by the EMFACS [202]. Examples of these AUs are given in Figure 6.1, and short descriptions are provided in Table 6.1. As shown in Table 6.2, particular AUs have been associated with each of the basic emotions by Matsumoto and Ekman [7]. For example, anger is often expressed by by furrowing the brow (AU4) and tightening the lips with teeth displayed (AU22, AU23, and AU24). This chapter describes a validation procedure by which AUs in Table 6.2 have been detected in 3D data from the Kinect, and compared with discriminative geometric features that have been chosen to perform FER. The chapter therefore explores the relation between geometric 3D features and corresponding facial movements that are associated with particular expressions of emotion.

Our contributions in this chapter can be summarized as follows: 1) We propose a dynamic realtime FER system that employs discriminative geometric 3D features to recognize the six basic

expressions, plus neutral, automatically. A pool of classifiers has been trained using machine-learning (ML) techniques, and a two-level voting approach is employed to make the final decision from a sequence of several frames. Our approach is fast, of low cost, and is relatively robust to pose and illumination variations. It outperforms the state of the art by more than 4% using the VT-KFER dataset. 2) In addition to testing with VT-KFER, which contains expression data for adults and children, the system was also tested with another dataset of 20 children in the age range of 9 to 12 years. To our knowledge, no other realtime FER systems have been tested so extensively using data from children. 3) Finally, a validation step was performed that compared the results from our feature selection step with EMFACS AUs. Good agreement was observed for all but one of the standard emotions.

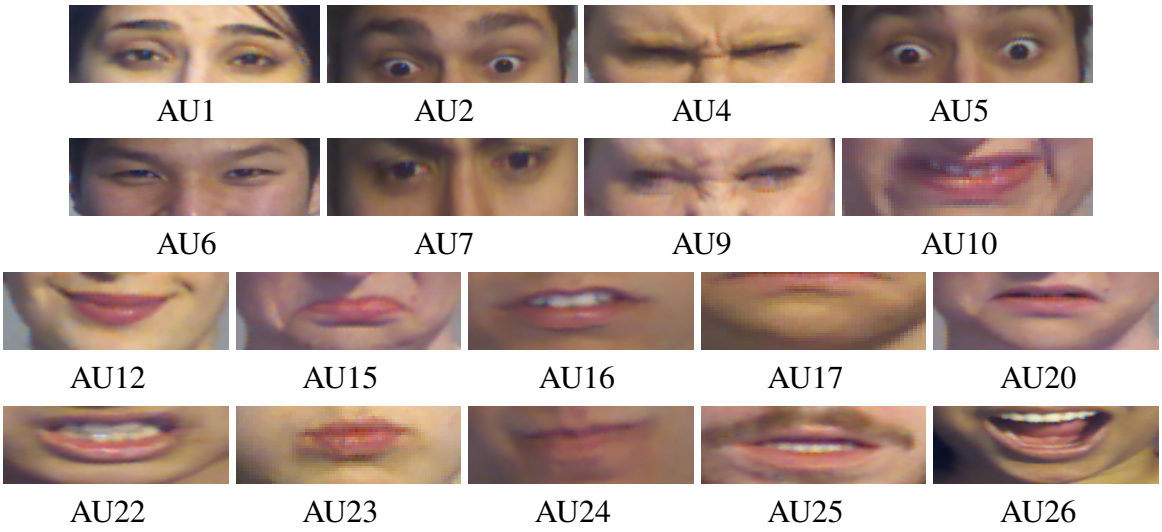


Figure 6.1: Examples of facial Action Units that are related to common emotions, according to EMFACS [7]. These images were taken from the VT-KFER dataset [2].

The rest of this chapter is organized as follows. Section 6.2 presents our FER methodology. The EMFACS validation framework is presented in Section 6.3. Section 6.4 and section 6.5 discuss the experimental setup and results. Finally, section 6.6 presents concluding remarks.

6.2 Methodology

This section describes our approach to realtime FER using 3D sensing only. A block diagram of the system is given in Figure 6.2. The rest of this section describes the major components of feature extraction, feature selection, and classification.

Table 6.1: Descriptions of Action Units shown in Figure 6.1 [9].

Action Unit	Description	Facial Muscles
AU1	Inner Brow Raiser	Frontalis, Pars Medialis
AU2	Outer Brow Raiser	Frontalis, Pars Lateralis
AU4	Brow Lowerer	Corrugator Supercilii, Depressor Supercilii
AU5	Upper Lid Raiser	Levator Palpebrae Superioris
AU6	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
AU7	Lid Tightener	Orbicularis Oculi, Pars Palpebralis
AU9	Nose Wrinkler	Levator Labii Superioris Alaquae Nasi
AU10	Upper Lip Raiser	Levator Labii Superioris
AU12	Lip Corner Puller	Zygomaticus Major
AU15	Lip Corner Depressor	Depressor Anguli Oris
AU16	Lower Lip Depressor	Depressor Labii Inferioris
AU17	Chin Raiser	Mentalis
AU20	Lip Stretcher	Risorius Platysma
AU22	Lip Funneler	Orbicularis Oris
AU23	Lip Tightener	Orbicularis Oris
AU24	Lip Pressor	Orbicularis Oris
AU25	Lips Part	Depressor Labii Inferioris or Relaxation of Mentalis, or Orbicularis oris
AU26	Jaw Drop	Masseter, Relaxed Temporalis and Internal Pterygoid

Table 6.2: Associations of emotions with Action Units, from Matsumoto and Ekman [7]. The parentheses indicate AUs that are optional.

Emotion	Action Units
Happiness	6, 12
Surprise	1, 2, 5, 25 or 26
Sadness	1, (4), 15, (17)
Anger	4, 5 and/or 7, 22, 23, 24
Disgust	9 and/or 10, (25 or 26)
Fear	1, 2, 4, 5, 7, 20, (25 or 26)

6.2.1 Feature Extraction

Because of the realtime requirement of this system, fast but effective feature extraction is mandatory. The goal is to detect features that relate directly to changes in shape of the face. We have therefore chosen four geometric 3D feature types that can be computed quickly from the 3D face mesh that is fit to range data from the Kinect. Each of the feature types utilized in our system is described briefly. As introduced earlier in this chapter, the mesh consists of 206 triangles, with 121 landmark points and 318 edges.

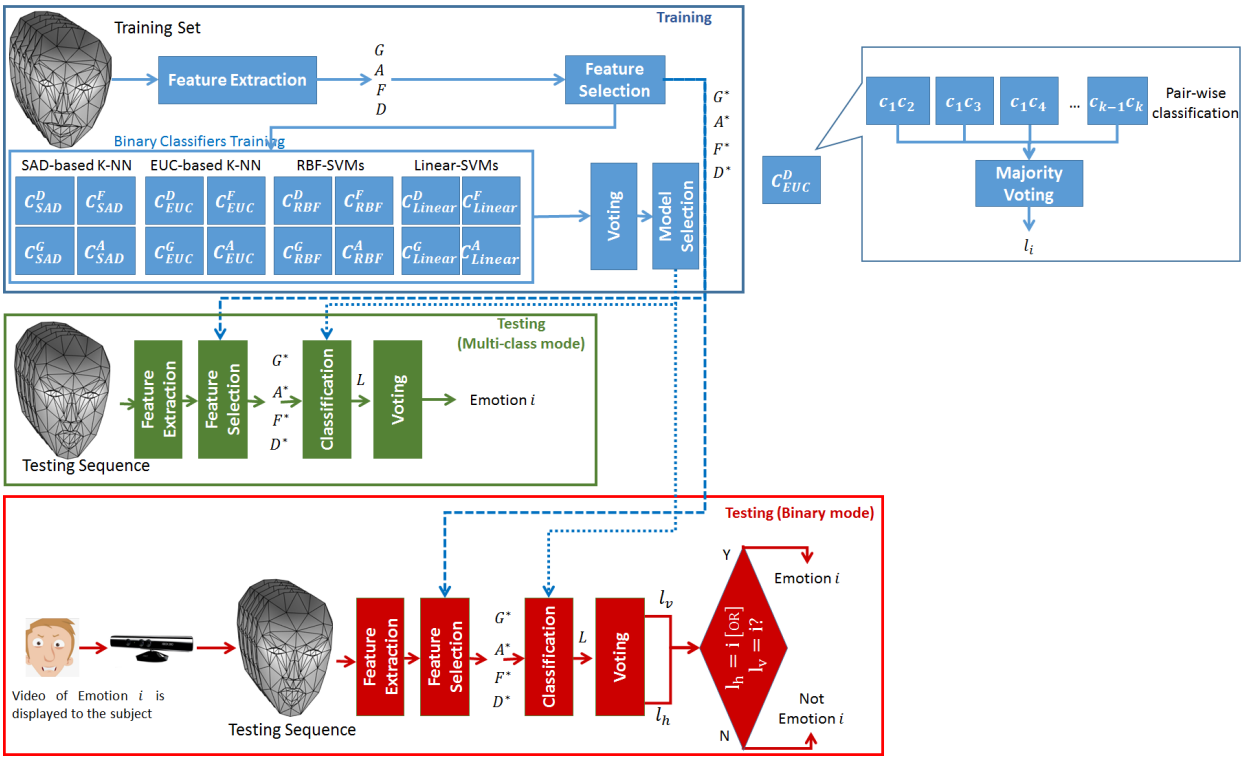


Figure 6.2: Diagram of the proposed FER system, showing both training and testing. During training, a pool of multiclass classifiers are generated. After each is trained individually, an exhaustive comparison determines the combination of those classifiers (known as a model) that best recognizes each expression through voting. During testing and normal operation, those particular models are used separately to test for each emotion of interest. To improve performance at this stage, additional voting is conducted over a sequence of input data frames. Our system is tested in two modes, 1) multi-class where the system responds with a label for 1 of k classes, and 2) binary mode, where the system responds with either the input sequence is of class i or not. The internal processing at each multi-class classifier used in training our system (e.g., C^D_{EUC}) is shown in the top right side.

Euclidean distances: The lengths of all edges in the mesh are computed. This results in a feature matrix $[D]$ of size $r \times 318$, where r is the size of training set.

Angles: Interior angles are computed for all triangles in the mesh. We utilize two feature vectors based on these angles. The first is composed of the values of all three interior angles from each triangle. The second feature vector is a reduced set, consisting of a single angle from each of the 206 triangles (the same that were used in previous chapter). (The particular set of angles was determined at random, in advance.) Two feature matrices, $[F]$ of size $r \times 618$ and $[G]$ of size $r \times 206$, result from all 3D face meshes in the training set.

Triangle surface areas: We also utilize the surface areas of the face triangle mesh. A feature

matrix, $[A]$ of size $r \times 206$, is computed for the 3D face meshes in the training set.

6.2.2 Feature Selection

The goal of the feature selection step is to identify a reduced set of features that minimize within-class variations while maximizing between-class variations to distinguish k classes. We employ our feature selection approach presented in section 5.3.2. It is an automatic approach based on ranking features according to statistical class separability criteria [134]. We employ this procedure on the extracted features as follows.

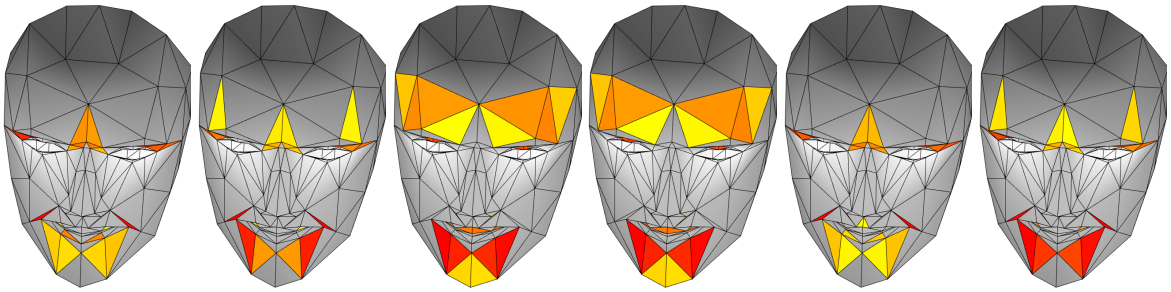
The approach utilizes five criteria to rank the features, namely: t-test, ROC, Bhattacharyya distance, relative entropy, and Wilcoxon signed-rank test. Each criterion is computed separately for each pair of classes, in order to rank the features from most significant to least significant. For each pair of classes, our approach selects the top-ranked s features according to each criterion, where s was chosen empirically. (Typically $s = 32$ in this work.) This results in a set of $s \times 5$ features. These features are combined and sorted according to their frequency, from the most frequently selected by the five criteria to the least frequent. Then, the top s most frequently selected features are chosen as the best features that discriminate the corresponding pair of classes.

The approach is applied to the feature matrices D, F, G, and A of the training set. For k classes, there are $k(k-1)/2$ class pairs. For each class pair, the optimal feature matrices are determined. Let $D^*_{i,j}, F^*_{i,j}, G^*_{i,j}$, and $A^*_{i,j}$ represent the best sets of features, respectively, for a given pair of expressions i and j.

Figures 6.3a to 6.3e illustrate the best $s = 32$ surface area features that best discriminate happiness vs. surprise selected by each criterion. Figure 6.3f shows the most frequently selected features by the five criteria. These are used here as the final "effective features". The features are colored based on their rank; the darker the color, the more significant the feature is. Since each criterion has a different metric, the five criteria may generate different results. However, some features may be selected by more than one criterion. In this case, those features are considered more significant than the other features that were selected by only one criterion. In the illustrated example of happiness vs. surprise, the five criteria agreed that the significant face patches to distinguish happiness vs. surprise are around the eyebrows and mouth. Therefore, our approach results in the features that are most frequently selected by the five criteria. This result agrees with how surprise and happiness are performed.

6.2.3 Classification

Our system works in two classification modes: 1) multi-class mode, and 2) binary mode. In the multi-class mode, the system responds with a label i, where $i \in 1, 2, ..., k$. In the binary mode, the input sequence is tested to check whether it belongs to emotion i (output label is i) or not (output label is $-i$). The main difference between the two modes is in the voting step. Generally,



(a) t-test (b) ROC (c) Entropy (d) Bhattach. (e) Wilcoxon (f) Best

Figure 6.3: The best 32 triangle surface area features selected from the whole face mesh to discriminate happiness vs. surprise. (a)-(e) The selected surface areas by the five criteria. (f) The most frequently selected surface area features by the previous five criteria that are considered here for classification. The features are color coded based on significance. The more significant the triangle surface area is to discriminate the two classes, the darker the color it takes.

the classification paradigm (illustrated in Figure 6.2) is divided into 1) model training, 2) model selection, and 3) testing. Each of these steps is described below.

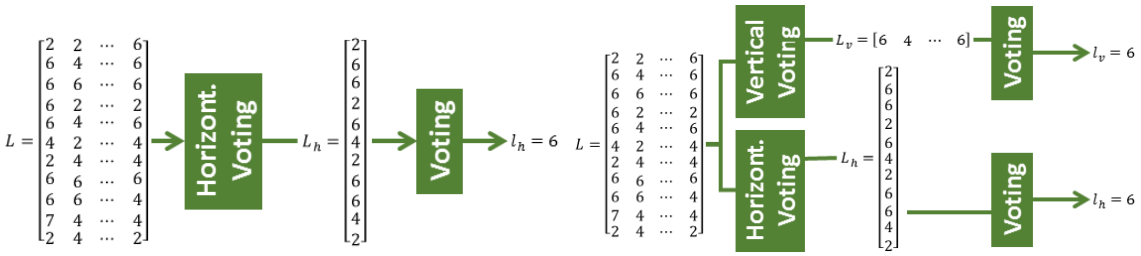
Model training: Each of the four feature types, selected in the previous step, is used to train a pool of 16 classifiers (shown as solid blue boxes in Figure 6.2), in which each classifier selects 1 of k classes. The system presented here employed the following well known classifiers: 1) 1-nearest neighbor (1-NN) classifier based on Euclidean distances (C^f_{EUC}), where f represents one of the feature types, 2) 1-NN classifier based on sum of absolute difference (C^f_{SAD}), 3) linear SVM classifier (C^f_{Linear}), and 4) RBF SVM classifier (C^f_{RBF}). The internal processing of each classifier is illustrated in Figure 6.2. The multi-classification of k classes is achieved here by combining $\frac{k(k-1)}{2}$ binary classifiers, (c_i, c_j), over a voting approach where k is number of classes, $i = 1, 2, ..., k-1$, $j = i+1, ..., k$, and $i \neq j$. K-NN and SVM have been widely used in many FER systems with high accuracy.

Model selection: In both multi-class and binary modes, all possible combinations of the 16 classifiers output labels vectors, are tested to select the one combination that best classifies each expression independently. The combination with the highest training accuracy is the one that is selected for testing, referred to as the "model".

For each combination, the label vectors of the classifiers are concatenated to create a label matrix L of $d \times m$ dimension. Each row in L represents a sequence frame and each column represents a classifier response. A two step voting approach is performed on L to find the final label.

The voting approach employed in the multi-class mode is illustrated in Figure 6.4a, where a horizontal voting is applied on L followed by a second majority voting step to give the final label $i \in \{1, 2, ..., k\}$.

In the binary mode, the voting approach works as follows. A vertical and horizontal voting are applied on L resulting in L_v and L_h, respectively, where L_h is the resulting vector of the horizontal



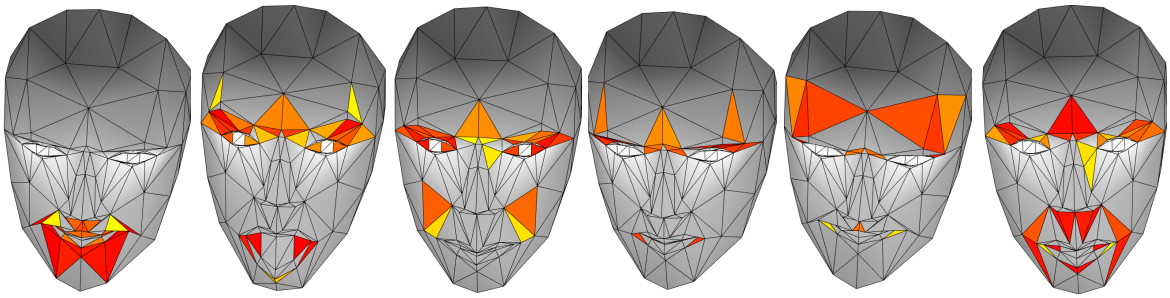
(a) (b)

Figure 6.4: (a) The two-step voting approach for multi-class mode classification with example input and output labels for demonstration. (b) The two-step voting approach for binary mode classification with example input and output labels for demonstration.

voting (voting over m classifiers) and is of size $m \times 1$. L_v is the resulting labels vector of the vertical voting (voting over the sequence of frames) and is of size $d \times 1$. Then a second majority voting is applied on L_v and L_h to produce one label per each, l_v for vertical voting and l_h for horizontal voting. To select the best combination for class i, a final decision for the combined classifiers is equal to i if either l_v or l_h is equal to i. Otherwise, the final label is $-i$ (i.e., not emotion i). An example that illustrates how this voting approach is working in the binary mode is given in Figure 6.4b.

With 16 classifiers, the total number of possible combinations is computed as $\sum_{j=1}^{16} \binom{16}{j} = 65535$. For each class i, the 65535 classifier combinations were tested, using the training set, to find the best combination that maximize the classification accuracy of class i. The best model for class i is composed of the concatenation of m_i classifiers, where $m_i \in 1, 2, ..., 16$ and $i \in 1, 2, ..., k$. Note that each class has different combination depending on which maximizes the corresponding class training accuracy. For example, the combination that best classifies c_1 may be $C^D_{EUC}\ C^G_{RBF}\ C^A_{EUC}$ while for c_2 it is $C^A_{Linear}\ C^D_{SAD}\ C^F_{RBF}$,..etc.

Testing: In realtime, certain videos were displayed to the subjects and we asked them to show what they thought the boy in the video was feeling. This is to test the ability for subjects to interpret the displayed expressions. For efficiency in realtime, only the features needed for the selected models were computed. For instance, in anger case, only angles (G) and Euclidean distance (D) need to be computed. Next, using the index of the selected features indicated in the training, the best features, D^*, A^*, F^*, and G^*, are selected and then are fed into the corresponding classifier in the selected model as shown in Figure 6.2. The classification using the model selected in training generates a label matrix L of size $d \times m_i$. To get a single label from L, we apply the corresponding voting approach on L based on the mode. In the multi-class mode testing, the final label is given by the voting approach in Figure 6.4a. In the binary mode testing, the input sequence of d frames is considered of class i if either l_h or l_v is equal to i.



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

Figure 6.5: Best 32 surface areas that discriminate (a) happiness, (b) surprise, (c) sadness, (d) anger, (e) disgust, and (f) fear vs. neutral.

6.3 EMFACS Validation

6.3.1 Overview

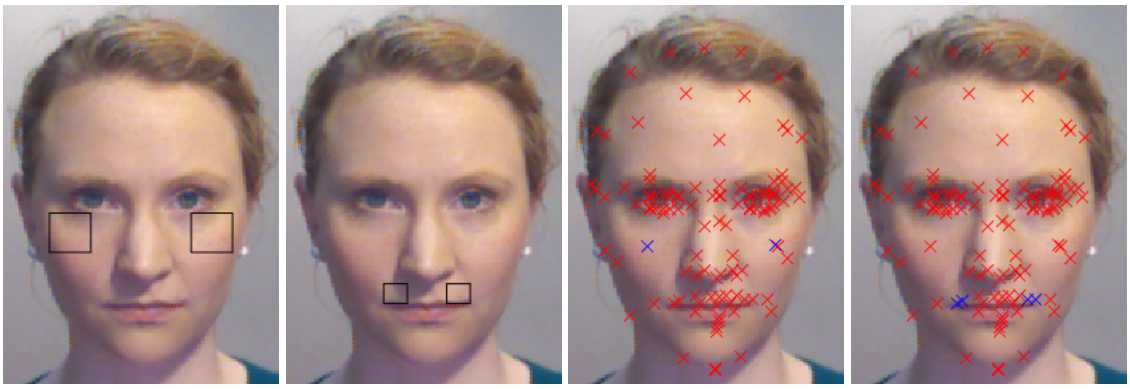
As shown in Section 6.2, our system is based on employing selected features that best discriminate pairs of emotions to train a pool of multi-class classifiers. This section presents a novel framework to validate these selected features with the AUs related to the EMFACS. We show, using the surface area features that discriminate happiness, surprise, sadness, fear, disgust, and anger versus neutral (Figure 6.5), that the selected features strongly correlate with at least one corresponding AU of the AUs related to the EMFACS. The correlation between the selected features and the EMFACS AUs validates the significance and discriminative ability of the features. We illustrate how at least one AU that is related to each expression, according to EMFACS, is strongly correlated to the selected features of corresponding expression.

We test the correlation with some of the AUs presented in Table 6.2. A total of 15 AUs are considered. AU22, AU23, and AU24 are in same location and thus same heat map (HM(s)). Therefore, one of them will only be considered here, AU23. The proposed ground truth is semi-automatically generated HMs that represent the 15 AUs. To the best of our knowledge, this the first work that propose a semi-automatic approach for AU ground truth generation in 3D.

6.3.2 Heat Map Generation

To validate the selected features with EMFACS AUs, a ground truth for the AU locations in the 3D face mesh is a necessity. The proposed ground truth here is a set of semi-automatically generated heat maps for each AU. The steps for ground truth generation are described below.

Step 1: AU ROI localization in 2D images: For each AU listed in Table 6.1, the region of interest (ROI) was manually selected (i.e., AU location) in the face RGB image. Figure 6.6 illustrates the selection of AU6 and AU12. The combination of AU6 and AU12 represent the happiness facial



(a) AU6 ROI (b) AU12 ROI (c) AU6 (d) AU12

Figure 6.6: (a) AU6 (Cheek Raiser) and (b) AU12 (Lip Corner Puller) ROI shown as black rectangles on a sample neutral face RGB image. The 2D keypoints, out of 121 detected by the Kinect SDK, intersecting with ROI of (c) AU6 and (d) AU12 are highlighted in blue crosses and the rest are in red.

expression according to the EMFACS.

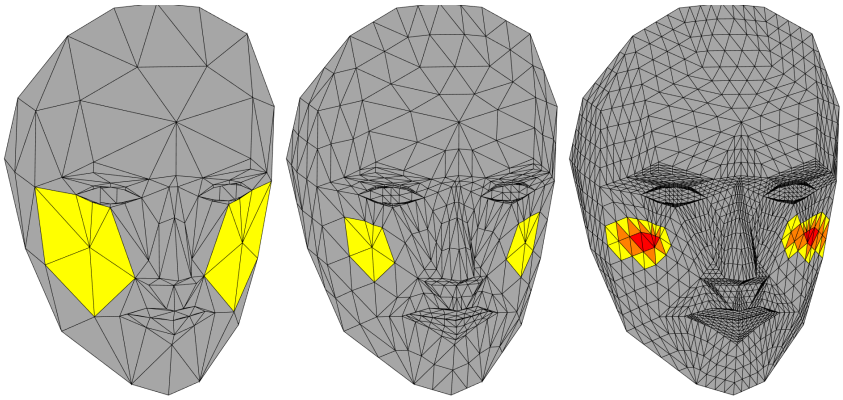
Step 2: Intersection of ROI-2D facial keypoints: The Microsoft Kinect SDK [129] provides the locations of 121 2D facial keypoints on the corresponding RGB frames. Giving these 2D keypoints coordinates, the proposed approach finds the intersection between the manually selected ROI and the 2D keypoints. Figures 6.6c and 6.6d illustrate the 2D keypoints lying in the AU6 and AU12 ROI, respectively. The 2D facial keypoints are plotted over a neutral face where the keypoints inside the ROI are plotted in blue and the rest in red.

Step 3: Generation of AU heat maps: From the 2D keypoints, we know the corresponding 3D keypoints. A corresponding HM can be generated by labeling the triangles that connect those vertices according to their significance. A triangle that has all its vertices inside the ROI is of more significance than if only 2, 1, or none of its vertices are in the ROI. Therefore, the significance is quantized into 4 levels with values that vary from 0 (least significant) to 3 (most significant). In other words, for AU_i, where $i = 1, ..., 15$, the corresponding HM ranking values are as follows

$$0 \leqslant R_j^{AU_i} \leqslant 3, \forall j = 1, ..., N, \tag{6.1}$$

where N is number of triangles. In the coarse face mesh, $N = 206$ while in the fine face mesh $N = 3296$. Figure 6.7a illustrates the AU6 HM generated from the coarse face mesh of the Kinect SDK. The triangles are color coded based on significance where the more significant the triangle is, the darker the color it takes. Non significant triangles are colored in gray.

Step 4: Coarse-to-Fine subdivision: For better AU representation, the 3D face mesh is subdivided into finer level of resolution. The original face mesh provided by Kinect SDK is composed of 121 keypoints and 206 triangles. Each triangle in this face mesh is subdivided into 4 smaller triangles by 1) locating the midpoint between each 2 vertices of the triangle, and 2) connecting



(a) AU6 (L1) (b) AU6 (L2) (c) AU6 (L3)

Figure 6.7: Heat maps for AU6 in 3 levels of resolutions. The triangles are labeled with respect to their closeness to the region of interest (ROI) for this AU. The closer the triangle to the ROI, the higher significance and the darker the color. Triangles take label 3 (most significant) if the three vertices lie inside the ROI, 2 if two vertices lie inside the ROI, 1 if only 1 vertex lies inside the ROI, and 0 if no vertices lie inside the ROI (not significant).

each midpoint with the other midpoints. This process will generates 4 smaller triangles from the original one. We call the resulting mesh a mid-fine face mesh of level 2 that is composed of 438 vertices and 824 triangles. We reapply this subdivision process in order to reach a fine representation of 1690 vertices and 3296 triangles. We call the resulting mesh a fine face mesh or level 3. After the subdivision, we repeat steps 2 and 3 to generate the new HMs of the fine mesh representation. Figures 6.7b and 6.7c illustrate AU6 heat maps of level 2 and 3, respectively. The HMs of the 15 AUs are illustrated in Figure 6.8.

Step 5: combine related AU HMs: To generate each expression HM, R^{H_k}, where $k = 1, ..., 6$, we combine the HMs of each corresponding AU together. The triangles in R^{H_k} are ranked similarly as we ranked R^{S_i}. In other words,

$$0 \leqslant R_j^{H_k} \leqslant 3, \forall j = 1, ..., N, \tag{6.2}$$

Figure 6.9 illustrates the combined HMs of the AUs related to the 6 expressions according to [7]. The HMs are illustrated in fine face mesh representation

6.3.3 Feature Ranking Quantization

The proposed feature selection approach ranks the features according to their discriminative ability. A ranking vector, for each expression, is generated to represent the rank of each triangle. Since the FS approach is applied on the coarse face mesh, the resulting ranking vectors, R^{S_k}, where $k = 1, .., 6$, will be of values:

$$0 \leqslant R_j^{S_k} \leqslant N, \forall j = 1, .., N, k = 1, .., 6, \tag{6.3}$$

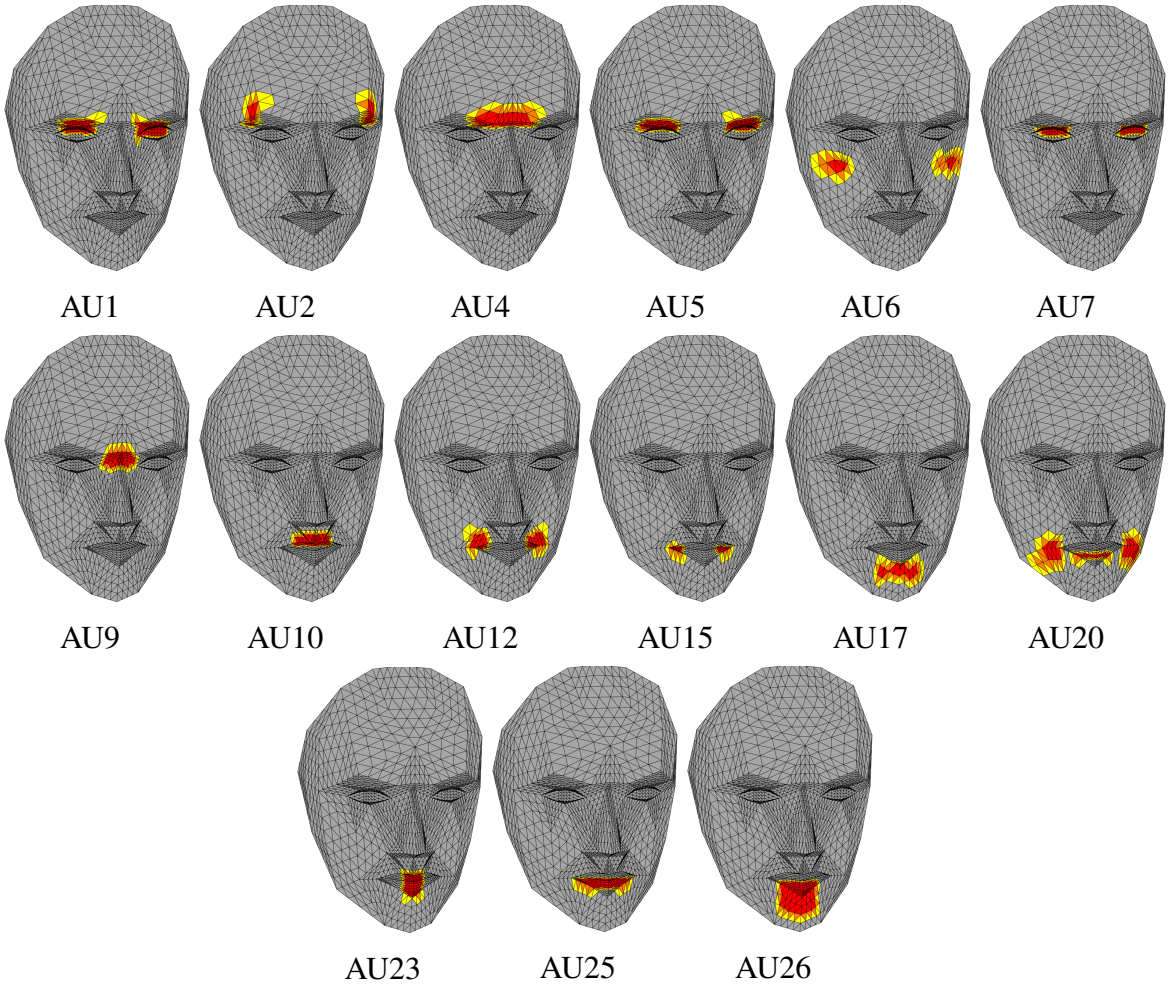


Figure 6.8: The heat maps for the 15 AUs related to the six basic expressions according to EMFACS given in [7].

where $N = 206$ and $R_j^{S_k} = 0$ if triangle j is not selected as significant one. Figure 6.5 illustrates the resulting selected features of the best $s = 32$ surface areas that discriminate happiness, surprise, sadness, anger, disgust, and fear vs. neutral. The mesh is color coded based on the ranking vectors R^{S_k} values. The more significant the feature is, the darker the color a triangle has. For R^{S_k} to be comparable to the heat maps ranking R^{AU_i} and R^{H_k}, we quantize the ranking values in R^{S_k} into 4 levels, as follows:

$$Q_j^{S_k} = \begin{cases} 3, & 1 \leq R_j^{S_k} \leqslant N/3 \\ 2, & 1 + N/3 \leq R_j^{S_k} \leqslant 2 * N/3 \\ 1, & 1 + 2 * N/3 \leq R_j^{S_k} \leqslant N \\ 0, & R_j^{S_k} = 0 \end{cases}, \forall j = 1, .., N. \tag{6.4}$$

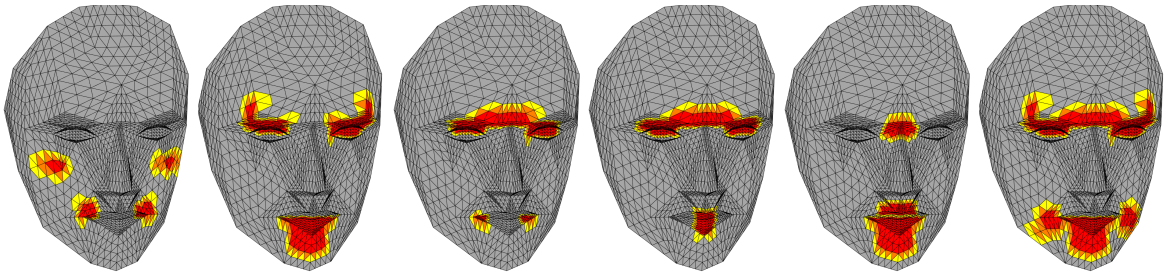


Figure 6.9: The combined heat maps of EMFACS AUs related to the six basic expressions.

6.3.4 FS-HM Agreement

To validate the discriminative ability of our selected features, we compute the agreement between the resulting ranking and the heat maps. Figure 6.10 illustrates the steps of FS-HM agreement computation for a sample happiness expression.

We first compute the agreement between the quantized ranking of selected features, Q^{S_k}, and each AU heat map, R^{AU_i}, independently, using matched filter technique. In signal processing, to detect the presence of a known signal or a template in an unknown signal, a matched filter is obtained by correlating the known signal, or the template, with the unknown signal [203]. In this chapter, we employ the matched filter technique to detect the presence of the related EMFACS AUs (i.e., template) in our selected features (i.e., unknown signal). This will provide an insight of which AU is mostly activated with each expression. The correlation coefficient ρ_{xy} between a template signal x and an unknown signal y is computed as follows:

$$\rho_{xy} = \frac{|r_{xy}|}{\sqrt{r_{xx} r_{yy}}} \tag{6.5}$$

where

$$r_{xy} = \sum_{n=1}^{N} x(n)y(n), \tag{6.6}$$

$$r_{xx} = \sum_{n=1}^{N} x(n)x(n), \tag{6.7}$$

$$r_{yy} = \sum_{n=1}^{N} y(n)y(n), \tag{6.8}$$

and $0 \leqslant \rho_{xy} \leqslant 1$. If ρ_{xy} is close to 0, this means that signal x is uncorrelated to signal y. Otherwise, if ρ_{xy} is close to 1, this means that both signals are correlated. We consider $x = R^{AU_i}$, and $y = Q^{S_k}$.

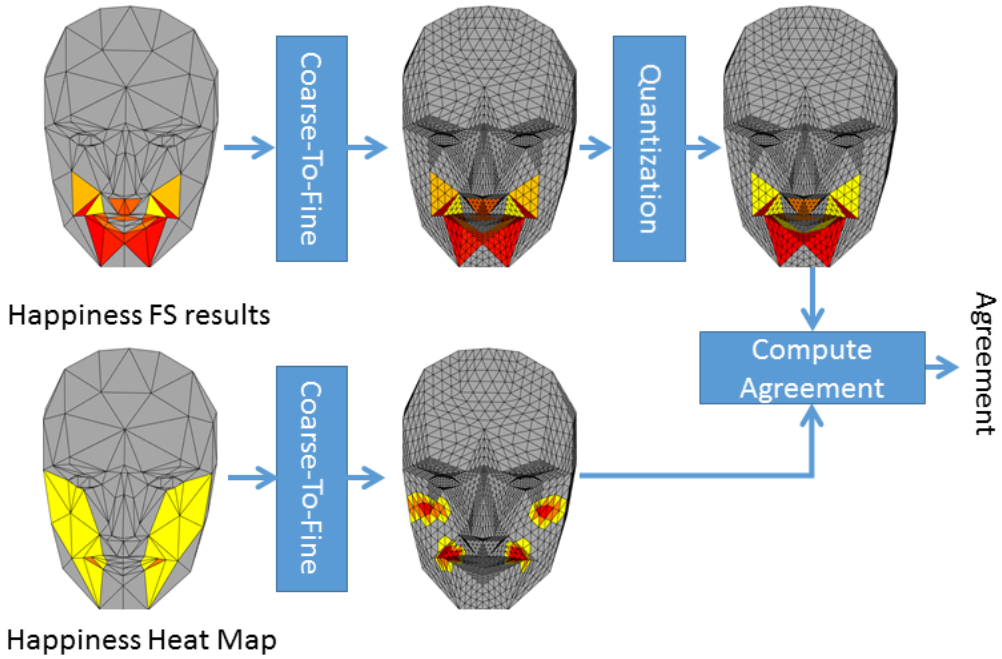


Figure 6.10: Computing the agreement between the feature selection (FS) results (upper) and the generated heat maps (lower) of happiness expression.

Second, we compute the agreement between the quantized selected features, Q^{S_k}, and the combined related AU HMs of each expression, R^{H_k}. We employ an agreement metric based on a one-to-one triangle ranking equality. A triangle-to-triangle agreement vector, L, between Q^{S_k} and R^{AU_i} is computed as follows:

$$L_j = \begin{cases} 1, & Q_j^{S_k} = R_j^{AU_i} \\ 0, & otherwise \end{cases}, \forall j = 1, ..., N. \tag{6.9}$$

The overall agreement/similarity, a, is computed as:

$$a = \frac{\sum_{j=1}^{N} L_j}{N}. \tag{6.10}$$

6.4 Experimental Setup

VT-KFER dataset: The proposed system was trained and tested on VT-KFER using both "leave-p-sequence-out" and "leave-p-subjects-out" cross-validation, where p=20%. As indicated earlier in this chapter, the "leave-p-sequence-out" cross-validation randomly selects $100 - p\%$ of all sequences in the dataset for training and the rest for testing where sequences for the same subject may be in the training and testing. On the other hand, "leave-p-subjects-out" randomly selects

$100 - p\%$ of the subjects for training and the rest for testing where each subject's sequences either are in the training or the testing. Our approach is sequence-based, where the decision is given based on d input frames. For VT-KFER, d varies depending on the length of the recorded sequence. On average, $d = 6$.

FEET dataset: For further testing to our approach, a second Kinect-based FER dataset was collected from children only, namely FEET dataset. FEET dataset includes 20 children, males and females, aged from 9 to 12. In the experiments conducted on the FEET dataset, the system was trained on data from VT-KFER and tested using the FEET dataset. The dataset contains RGBD sequence data of frontal poses for the four facial expressions, happiness, anger, fear, and neutral. In this work, we only use the depth data for our system. Since FEET dataset includes only 4 expressions, the training on VT-KFER dataset was conducted in two modes, $k = 7$ and $k = 4$. When $k = 7$, the multi-class classifiers was trained to recognize the six basic expressions plus neutral. While when $k = 4$, the multi-class classifiers are trained to recognize only the 4 expressions in FEET: happiness, anger, fear, and neutral. The frames were recorded while a video is shown to the subjects. We selected the last d frames from the recorded frames as we expect it is when the subject made the proper expression. For FEET dataset, $d = 11$ for all sequences. The length of the sequence was picked empirically as a tradeoff between speed and accuracy. It was chosen not to be too long so the system respond very late and not too short so it does not include much variety/change.

Using two researchers from psychology, both VT-KFER and FEET datasets were human coded, blindly (i.e., the coders did not know either the subjects or which emotion the subject was asked to make), to identify the correctly expressed emotion. Each frame has been labeled either as correctly expressed (correct=1) or not (correct=0). In addition, correct expressions were rated according to how confident the rater is from the correctness of this expression, where 0 means not confident and 1 means confident. In our experiments, we only employed the correct expressions with confidence.

6.5 Results

The experimental results of the proposed FER approach on VT-KFER and FEET datasets are presented. Also, the results of EMFACS AUs validations are illustrated.

6.5.1 FER Results

VT-KFER: Since VT-KFER includes sequences of the 6 classes only (no neutral sequences), the results on VT-KFER do not include the neutral case. We used two cross-validation techniques, the leave-p-subject-out and leave-p-sequence-out, where p=20%. Experimental results using VT-KFER dataset in multi-class and binary modes are provided in Tables 6.3 and 6.4, respectively. On average, the system accuracy increased around 10% and 3% when trained and tested using the leave-p-sequence-out over the leave-p-subject-out cross validation, in the multi-class and binary

Table 6.3: The average recognition accuracy of our FER approach on VT-KFER, in the multi-class mode using both leave-p-subject-out (column 1) and leave-p-sequence-out (column 2) cross validation.

Expression	**leave-p-subject-out**	**leave-p-sequence-out**
Happiness	96%	86.2%
Surprise	73.1%	73.5%
Sadness	61.5%	66.7%
Anger	63.3%	66.7%
Disgust	16.7%	51.7%
Fear	33.3%	52%
Average	**57.3**%	**66.7**%

Table 6.4: The average recognition accuracy of our FER approach on VT-KFER in the binary mode using both leave-p-subject-out (column 1) and leave-p-sequence-out (column 2) cross validation.

Expression	**leave-p-subject-out**	**leave-p-sequence-out**
Happiness	86.8%	87.3%
Surprise	93.4%	90.5%
Sadness	88%	86.8%
Anger	82.6%	89.4%
Disgust	85.6%	88.9%
Fear	85%	92.1%
Average	**86.9**%	**89.2**%

modes, respectively. With the exception of the happiness expression, the system in the binary mode achieves higher accuracies than in the multi-class mode by approximately 26% on average for the leave-p-sequence-out. The confusion matrices for our system in multi-class mode when tested on VT-KFER using leave-p-sequence-out, and leave-p-subject-out cross validation are illustrated in Figures 6.11a and 6.11b, respectively. The confusion matrices for happiness, surprise, sadness, anger, disgust, and fear binary classifiers using leave-p-subject-out cross-validation are given in Figures 6.11c to 6.11h, respectively.

FEET: Since the FEET dataset includes only 4 expressions, happiness, fear, anger, and neutral, this approach was tested on FEET dataset using: 1) multi-class classifiers (e.g., C^{D}_{RBF}) trained on 7 classes from data in VT-KFER ($k = 7$), and 2) multi-class classifiers trained on the corresponding 4 classes of FEET from data in VT-KFER ($k = 4$). Table 6.5 introduces the results on FEET dataset in both multi-class and binary modes. Results of the binary mode system show that anger, fear, and neutral binary classifiers trained and tested using $k = 7$ multi-class classifiers outperform the case when $k = 4$. However, happiness binary classifier achieves average recognition accuracy of 79.6% on FEET dataset using multi-class classifiers of $k = 4$ compared to 71% in case of $k = 7$. The confusion matrices for the four best happiness, anger, fear, and neutral binary classifiers tested

Table 6.5: The average recognition accuracy of our FER approach on FEET dataset, in both the multi-class and the binary modes. Column 1 shows our system in the multi-class mode results on FEET dataset. Column 2 and 3 show the system recognition accuracy in the binary mode when it was trained on VT-KFER using multi-class classifiers that recognizes $k = 4$ and $k = 7$ classes, respectively. The corresponding average and median values at each case are given in the last row.

Expression	**Multi-class mode**	**Binary mode** ($k = 4$)	**Binary mode**($k = 7$)
Happiness	84.6%	79.6%	71%
Anger	58.8%	70.1%	72%
Fear	63.6%	64%	66.1%
Neutral	67.9%	76.2%	76.2%
Average	68.7%	72.5%	71.3%
Median	65.8%	73.2%	71.5%

Table 6.6: Quantitative comparison with state-of-the-art FER systems tested on VT-KFER using both leave-p-sequence-out and leave-p-subject-out cross validation, where p=20%.

System	**Modality**	**Leave-p-sequence-out**	**leave-p-subject-out**
[196]	2D	67%	-
[8]	3D	49%	-
[2]	2D+3D	60%	-
[6]	2D	59%	46%
[134]	2D+3D	80%	56%
Our System (Multi-class mode)	**3D**	**66.7%**	**57.3** %
Our System (Binary mode)	**3D**	**89.2%**	**86.9**%

on FEET dataset are provided in Figure 6.11.

Quantitative comparison with state-of-the-art FER systems on VT-KFER is given in Table 6.6. Using multi-class mode and leave-p-sequence-out, our approach achieves average accuracy of 66.7%, which outperforms all other 3D-based approaches tested on VT-KFER. The increase in accuracy in the realtime system over the Dual-KDA work [134] is due to the use of sequence-based classification instead of frame-based classification and the use of expressions that were human-coded as correct with confidence. Using the binary mode testing, our proposed approach achieves 89.2%, using leave-p-sequence-out, which outperforms all listed 2D, 3D, and 2D+3D approaches including the state of the art on VT-KFER with more than 9%. Using leave-p-subject-out, our system in the binary mode achieves average recognition accuracy of 86.9%.

Computation time in realtime: Our system was built using C# developed using Visual studio 2012 and Matlab R2016a where the data acquisition is performed in C# environment and then the data processing is performed in Matlab. To read the input $d = 11$ frames, in C#, for realtime recognition, it takes 0.3 sec, on average. It takes 1 sec, on average, for FEET system to compute the response using Matlab on Intel Core i7 laptop that runs Windows 7. The time was computed

starting from the time Matlab framework receives the buffered frames till the final label is given.

6.5.2 EMFACS Validation Results

Figure 6.13 presents the result of computing the correlation coefficient between selected features to discriminate happiness, surprise, sadness, anger, disgust, and fear vs. neutral and 15 AUs in EMFACS that are related to the six expressions. The AUs related to each expression are indicated with red arrows.

As shown in Figure 6.13, for each expression except disgust, at least one of the AU combinations indicated by EMFACS in Table 6.2 showed highest correlation with the selected features for this expression. For example, in case of happiness, AU12 HM had the highest correlation with the selected features that discriminate happiness vs. neutral. Also, in case of surprise and fear, AU25 had the highest correlation. However, only in case of disgust, the correlation of the related AUs came in the fourth place after three unrelated AUs.

Figure 6.14 presents the agreement values a (computed in Equation 6.10) between the proposed approach vs. EMFACS AUs ground truth for both coarse and fine mesh representation. Results show that fine mesh representation is of higher agreement with the ground truth than the coarse representation, for all expressions. The agreement rate achieved for the six basic expressions vs. neutral ranges from 55.9% to 82%.

6.6 Conclusion

This chapter has presented a novel Kinect-based system that automatically recognizes facial expressions at realtime rates using input sequence of frames (dynamic). The system buffers a sequence of frames and gives the final decision based on a voting approach over the sequence labels. Only 3D data is used in this system. The expressions of interest are happiness, surprise, sadness, disgust, fear, anger, and neutral. Two Kinect-based datasets were utilized for training and testing the system. The first, VT-KFER [2], consists of 32 subjects aged 10 to 30. The second is an in-house dataset containing facial-expression data for 20 children, with ages ranging from 9 to 12 years. To the best of our knowledge, this system is the first realtime FER system that has been tested with children.

Using leave-p-subject-out cross-validation with VT-KFER, where 80% of the subjects were used for training and 20% for testing, our system in the multi-class mode achieved 57.3% average recognition accuracy and 86.9% using the binary mode. Using leave-sequence-out cross-validation with VT-KFER, where 80% of the sequences were used for training and 20% for testing, a higher, as expected, average recognition accuracy of 66.7% was achieved using the multi-class mode and 89.2% was achieved using the binary mode. The increase of the system accuracy using the leave-sequence-out cross validation may be due to the presence of different sequences for the same

subjects in both the training and testing. When tested on the separate dataset of 20 children, 68.7 % was achieved using the multi-class mode and 72.5% accuracy was achieved using the binary mode.

An interesting aspect of this work is a validation step using Action Units, as proposed by EMFACS [7]. An approach was developed to generate AU ground truth (heat map) semiautomatically, relative to triangular mesh representations produced by the Microsoft Kinect SDK. A heat map is generated for each AU related to emotions as proposed by Matsumoto and Ekman [7]. The similarity between the selected features and the ground truth was computed using matched filter-based approach. Results show that for all expressions (except disgust), at least one of EMFACS related AUs is of strong correlation to the selected features.

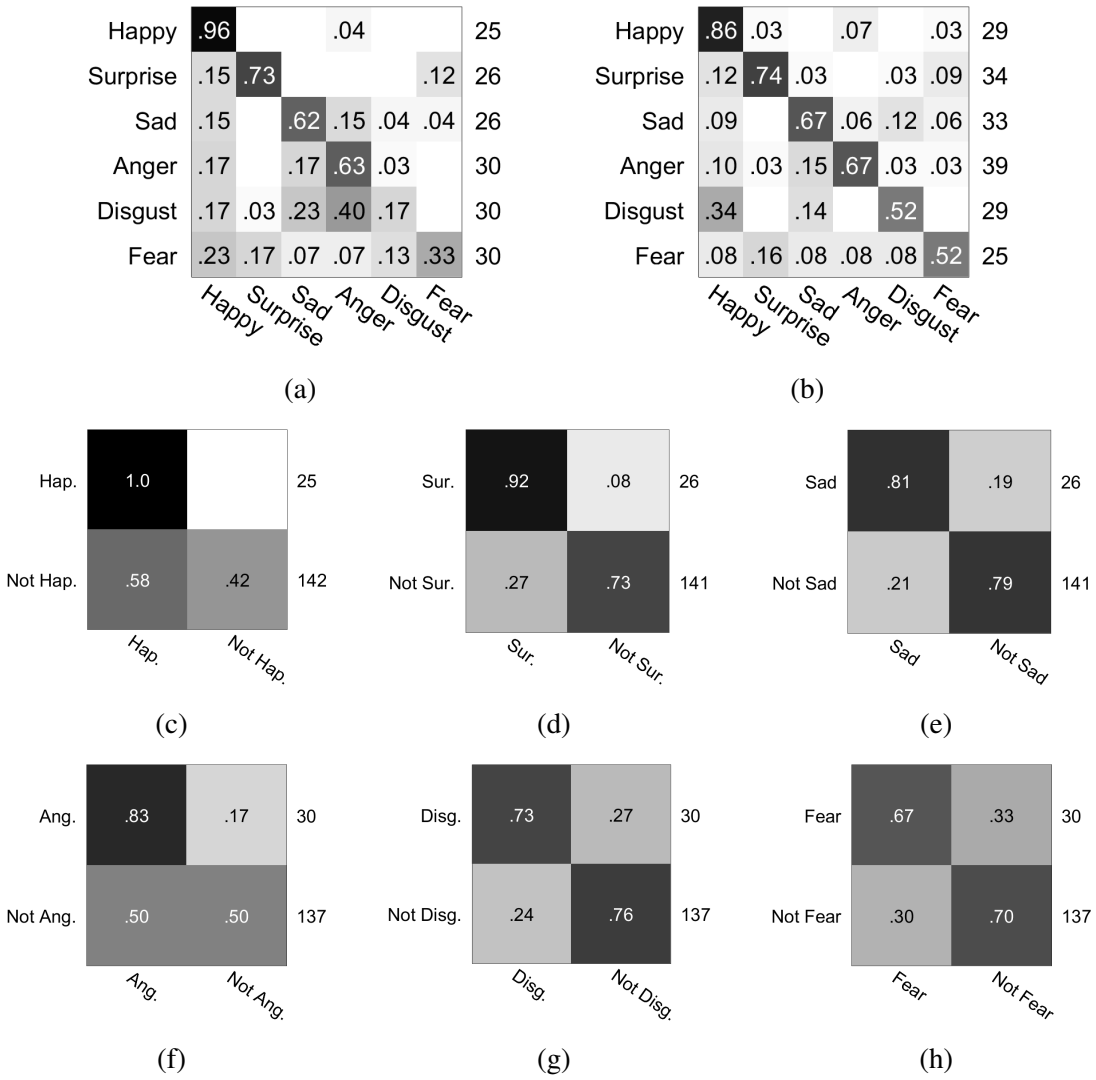


Figure 6.11: The confusion matrices for our system in multi-class mode when tested on VT-KFER using (a) leave-p-sequence-out, and (b) leave-p-subject-out cross validation as given in Table 6.3. Figures (c-h) illustrate the confusion matrices for happiness, surprise, sadness, anger, disgust and fear binary classifiers tested on VT-KFER using leave-p-subject-out cross validation, as given in Table 6.4.

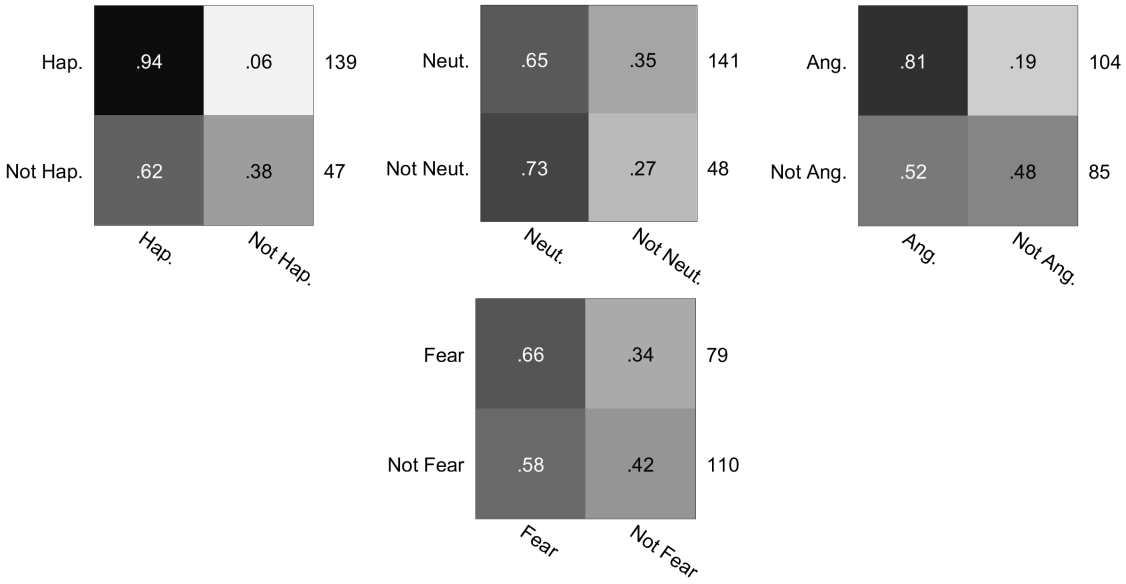


Figure 6.12: The confusion matrices for the happiness ($k = 4$ case), neutral ($k = 7$ case), anger ($k = 7$ case), and fear ($k = 7$) binary classifiers tested on FEET as given in Table 6.5.

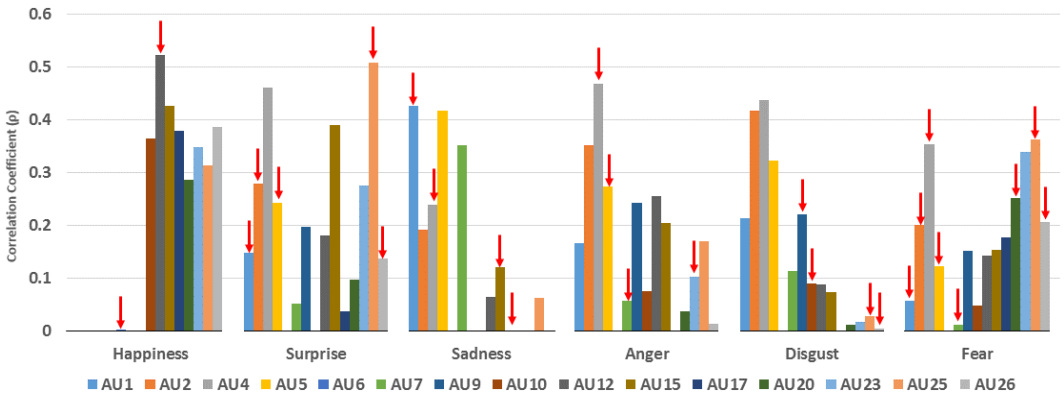


Figure 6.13: Correlation coefficient between selected features that discriminate happiness, surprise, sadness, anger, disgust, and fear vs. neutral and the HMs of EMFACS AUs. For each expression, we marked the related EMFACS AUS with red arrows above. Ideally, the AUs with red arrows above should have the maximum correlation compared to other AUs.

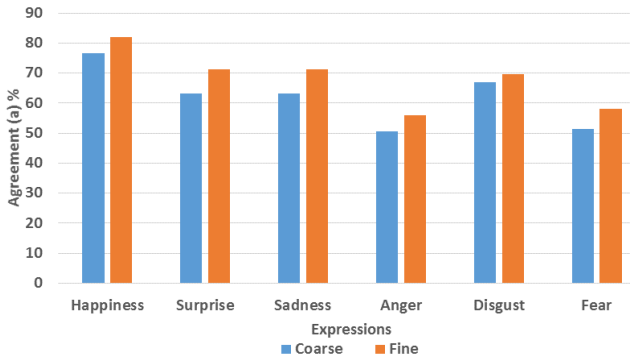


Figure 6.14: Coarse vs. fine face mesh representation agreement/similarity percentage between proposed selected features and corresponding heat maps.

Chapter 7

Summary

Facial expressions are of paramount importance in the daily interactions between humans. A person's face conveys essential information regarding emotion, mood, drowsiness, as well as many other cues related to events such as stress or well-being. Facial expressions make it possible to express and perceive unspoken emotional and mental states. The ability for recognizing facial expressions is so vital for human interactions, that it is one of the first abilities developed soon after birth. Indeed, it has been found that the ability to discriminate certain expressions develops very early in childhood, with ability to distinguish basic emotions from static cues appearing as early as 3 months.

Given the importance of facial expressions in our daily lives, it is not difficult to understand why it has been a field of study in varied areas including psychology, linguistics, neuroscience, and computer science, among others. Of special interest to us is the ability for recognizing facial expressions using computational systems. This is due to the opportunities and improvements that automatic Facial Expression Recognition (FER) could provide in applications such as human-computer interaction, affective computing, video games, advertising, education, driver safety, health care, deceit/intent detection, to mention just a few.

Although humans recognize facial expressions with almost no efforts, for computers, automatic FER is still a very challenging problem. At the core of the most important challenges for creating automatic FER systems are:

- High variability of the input data (e.g., changes in pose and illumination conditions);
- limitations in current sensor technologies;
- limited availability of FER-specific datasets captured by the Kinect for research purposes; and
- high complexity in FER systems that penalizes realtime processing.

The work presented in this dissertation directly addresses these challenges and provides solutions that make automatic FER a reality using a low-cost sensor, namely, the Kinect. The work focuses on the automatic recognition of what have been defined as the six basic expressions, namely, happiness, surprise, sadness, anger, disgust, and fear, plus neutral. FER is accomplished on both frontal and non-frontal poses.

A summary of the solutions to each of the challenges presented in this dissertation follows.

7.1 Addressing the Challenges of Automatic FER

7.1.1 Alleviating Limitations in Data Sensor Technologies

Due to the long-term availability and ubiquitousness of 2D systems, several researchers have addressed the development of systems that can perform FER automatically using 2D images or videos. However, these systems inherently impose constraints on illumination, image resolution, and head pose. Some of these constraints can be relaxed through the use of three-dimensional (3D) sensing systems. Various 3D sensors have been utilized in FER systems, but most of these sensors are expensive, and of a high resolution that may not be practical for realtime applications.

Driven by a need for low-cost and fast 3D sensors that aimed in better gaming experiences, Microsoft introduced the Kinect 1.0 sensor in 2010. This type of sensor provides an attractive alternative for FER due to its low cost, portability, and high acquisition rates (30 frames per second). However, in spite of these advantages, the Kinect is relatively low in resolution and noisy, as compared to many other 3D sensing techniques. This has narrowed the number of researchers that have considered the Kinect for the purpose of FER.

7.1.2 Building the VT-KFER Dataset to Provide a Solution to the Lack of Kinect-based FER Dataset

A common publicly available dataset is essential for research on facial expression analysis. The first obstacle that was found towards building automatic FER system using Kinect, was the lack of a reliable Kinect-based FER dataset. The only publicly available Kinect-based datasets available before this dissertation provided limited facial expressions, (E.g., FaceWarehouse provides only 4 expressions including neutral in frontal pose only).

To address this challenge, we built VT-KFER, the first RGBD+time FER dataset using the Kinect with both frontal and non-frontal poses. This dataset contains 32 subjects from 9 to 30 years old. It has been thoroughly used for the evaluation of the FER algorithms presented in this dissertation, and is already calling the attention of researchers world wide (more than 10 requests for using VT-KFER have been received since its release in May, 2015).

Chapter 3 includes details about VT-KFER, the human and computer evaluation conducted.

7.1.3 Addressing the Data Variation Challenges

Data variation is one of the main challenges towards robust automatic FER system. Pose variation greatly affects the performance of most FER systems. Typical distance-based FER approaches rely on a set of features produced by particular distance metrics.

This dissertation proposes a framework that automatically tunes itself to the given 3D data by identifying the best distance metric that maximizes the accuracy for each class. This tuning is particularly important when designing robust recognition systems that have to deal with low resolution data of varying poses. When tested on an in-house dataset of 10 subjects, the adaptive distance metric (DM) selection approach achieved average recognition accuracy (ARA) of 96.6% vs. 95.7% using single/fixed DM approach. When employed for only frontal-pose data, ARA of 98.9% was observed for the proposed approach vs. 98.9% when using single DM. However, a significant increase was noted in accuracy for the proposed approach (i.e., 95.1%) over the fixed DM case (i.e., 93.7%) for the case of non-frontal pose data, which reflects the contribution of utilizing adaptive DM feature vector over a fixed one for non-frontal pose FER.

More details regarding this system are in chapter 4.

7.1.4 Tackling Low-resolution and Noisy Data for FER Using the Kinect

The Kinect v1.0 sensor produces relatively low-resolution and noisy data as compared to other 3D sensors. This makes FER using this type of sensor a very challenging problem. As an example, when applying existing FER approaches developed by other researchers on VT-KFER, a maximum accuracy of 49.4% was achieved using 3D data only and 60% when 2D and 3D data were combined.

Inspired by the fact that a combined 2D + 3D approach seemed to increase the accuracy, this dissertation considered this combination/fusion as a potential solution for compensating the lack of resolution and the pollution of the data. However, this significantly increases the dimensionality and complexity of the problem when large feature vectors such as LBP and HOG are used. Therefore, finding the most representative features that can discriminate the various facial expressions becomes a vital step for a robust FER.

This dissertation presents two contributions to alleviate these limitations presented by the Kinect sensor and data fusion of large feature vectors. First, it proposes a novel ranking-based feature selection approach that combines five different selection criteria in which features with the greatest class separability are selected. This document applies this selection approach on computed geometric angles over a predefined 3D face mesh. Experimental results show that selected angles outperform full set of angles not only in frontal and non-frontal poses but also in two expression

intensities.

Second, the dissertation presents the Dual Kernel Discriminant Analysis (DKDA) feature fusion approach. DKDA is an extension to the traditional Kernel Discriminant Analysis (KDA) feature reduction approach, in which the 2D and 3D features are transformed jointly. Instead, KDA is applied on those 2D and 3D features independently. This results in two disjoint KDA kernels (hence the Dual KDA (DKDA)), with each of these kernels being characterized by a parameter σ. This parameter is tuned for maximizing the accuracy of each kernel independently. The output of these kernels generates the set of transformed features to be used for FER. Using HOG and selected 3D angles, the proposed DKDA-transformed features achieved average recognition rates of 80% (the current state of the art on VT-KFER using 2D+3D data) and 59% on frontal and non-frontal poses, respectively. With varied expression intensities, a 60% average accuracy was achieved for expression intensities evoked by verbal instruction, and 76% in expression intensities evoked by seeing images.

Chapter 5 elaborates more regarding these two approaches and their corresponding results.

7.1.5 Fulfilling Realtime Operation Requirements

Because of the realtime requirement of the proposed system in this dissertation, fast and effective feature extraction is necessary. The goal is to utilize features that relate directly to changes in shape of the face and that are pose and illumination invariant. Therefore, this system has chosen four geometric 3D feature types that can be computed quickly from the 3D face mesh that is fit to range data from the Kinect. To make our system faster, we applied our feature selection approach during training phase to find the most significant features for classification and employ them during testing in realtime. A pool of k-nearest neighbor-based and SVM-based multi-class classifiers were trained and using a model selection approach, this system adaptively selected the best combination for each of the seven classes of interest. The adaptive approach was inspired by our system that automatically tune the distance metric according to the class, which outperform fixed DM approach specially for non-frontal poses. To analyze the facial expression dynamics, our system buffers a testing sequence of length d frames for classification.

The classification step of the proposed system was designed to run in two modes, 1) a multi-class mode where the given input of frames are classified into k classes, and 2) a binary mode where the system indicates either the input sequence of frames are of certain emotion or not. When tested on VT-KFER using "leave-p-subject-out" ($p = 20\%$) cross-validation, an average recognition accuracy (ARA) of 57.3% was achieved using the multi-class mode and 86.9% was achieved using the binary mode. Using "leave-p-sequence-out" cross-validation with VT-KFER, the ARA was 66.7% in the multi-class mode and 89.2% in the binary mode. When tested on a second dataset of 20 children, 68.7% was achieved using the multi-class mode and 72.5% accuracy was achieved using the binary mode.

A detailed description of our realtime system is given in chapter 6.

7.2 Summary of Contributions

The significant empirical and theoretical contributions presented in this dissertation can be summarized as follows:

1. This dissertation developed a novel FER framework that automatically tunes itself to the given 3D data by identifying the best distance metric that maximizes the accuracy for each class. The proposed tuning significantly enhances the system's accuracy over typical approaches where fixed distance metric is employed especially with data in non-frontal poses.

2. This dissertation built VT-KFER, the first dynamic Kinect-based facial expression recognition dataset, where RGBD+time data (i.e., RGB + depth + time) in 3 corresponding poses (i.e., frontal, left, and right) and 3 levels of expression intensities are captured for 32 subjects (age from 9 to 30 years old). To the best of our knowledge, we are providing the first dataset that includes children. Our new dataset not only provides acted (i.e., non-spontaneous) facial expression but also spontaneous expressions for every subject. Although the study of spontaneous expressions is out of the focus of this dissertation, we aimed to provide a FER dataset that can help wider range of FER researchers.

3. The Dual Kernel Discriminant Analysis (DKDA) feature fusion technique was proposed and successfully tested to combine a wide range of well known 2D and 3D approaches. This novel approach addresses the problem of fusing high dimensional noisy data that contains severe non-linearities. Experimental results show that DKDA outperform LDA and KDA approach in all poses and expressions intensities.

4. A novel statistically-based feature selection approach was introduced. The approach automatically selects the most discriminative features that best discriminate two given classes.

5. A novel Kinect-based 3D FER system was developed and successfully tested off-line using data from VT-KFER and in realtime on 20 TD children aged 9 to 12 years old. The system aims to be used as a clinical aid in children ASD therapies in the future.

7.3 Future Work

The work presented here provides novel solutions to the FER problem. Because this is an important problem, there is always room for improvement targeting automated FER featuring equal or better quality of that provided by human subjects. Based on our FER solutions, the following are some avenues for potentially improving the accuracy and understanding of automated FER systems that make use of 2D and/or 3D data:

- *Adding a second layer to the classifiers targeting refinements for outliers*. The features for training the classifiers in the real-time system were selected based on analyses considering the entire VT-KFER dataset of 32 subjects. It is expected, however, for a few individuals to behave different than the general population. These individuals, which are outliers, may not show the same (or similar) intensity on one or more facial expressions than the general population. To alleviate the negative effect that classifying expressions on these individuals may have (i.e., potential penalty on accuracy due to the difficulty on deciding for outliers), we could add a second layer to the classifiers. This second layer would be trained using the outliers. The overall FER system could be then built weighting the classification results rendered by the "general population" classifier and the "outlier population" classifier.

- *Analyzing similarities and differences between FER on adults and FER on children*. The goal of this dissertation was to develop a general FER system. For the creation of such a system, we used the VT-KFER dataset. As mentioned above, the VT-KFER dataset contains data from both adults and children. The number of adults, however, is much larger than the number of children (14 adults vs. 8 children). If we assume that the characteristics of expressions made by adults and children are different, then the FER systems presented here were created/tuned in an environment biased towards adult made facial expressions. Additional work should target: 1) Analyzing the characteristics of expressions made by adults and children, and 2) if the above mentioned assumption holds, design and/or tune FER that consider the differences between expressions made by adults and children.

- *Studying the accuracy-complexity tradeoff for the dynamic FER system as a function of the number of frames used for classification*. Recall that, in the dynamic system, classification decisions are made based on individual results on a set of frames that capture a subject's expression. In the work presented here, the number of frames in the set was selected empirically. The criterion used for this selection was finding the number of frames that delivered a desired accuracy and yet work in real time. Additional work could be done to better understand the relation between number of frames and accuracy-complexity. The ultimate goal would be to find curves that render the optimal number of frames given a complexity/accuracy target.

- *Employing deep learning-based FER framework*. This dissertation presented novel approaches for each step of FER system (i.e., feature extraction, feature selection, feature fusion, and classification). Recently, convolutional neural networks (CNNs) and other deep leaning-based frameworks has proofed great success in the development of automated systems that accomplish these tasks at once with outstanding results. These deep learning approaches have an ability to automatically extract useful representations from raw data adjusted to the classification problem (e.g., image data). We could take advantage of deep learning technologies for the FER problem. For example, instead of employing each of these steps individually with engineered features such as the ones used in this dissertation, we could transfer features from deep convolutional networks and apply all those steps at once. In order to accomplish this, we would need an increased number of training data. We

expect this to happen as the VT-KFER dataset incorporates new subjects and/or include other FER datasets.

- *Building cross-dataset FER systems for increased accuracy*. As mentioned in the previous item, there is already a decent number of 2D datasets for taking advantage of deep learning. We could exploit this and go beyond basic deep learning approaches. For example, combining multiple FER datasets for training and testing could enhance the performance and prove the generalization of the approach.

7.4 Publications

The work presented in this dissertation has been published in the following venues/journals:

1. **Sherin Aly**, Lynn Abbott, Andrea T. Wieckowski, Susan White, Dynamic 3D Facial Expression Recognition using the Kinect, to be submitted to IEEE Transactions on Affective Computing Journal.
2. **Sherin Aly**, Lynn Abbott, Marwan Torki, A Multi-modal Feature Fusion Framework for Kinect-based Facial Expression Recognition using Dual Kernel Discriminate Analysis (DKDA), IEEE Winter Conference on Applications of Computer Vision (WACV), 2016.
3. **Sherin Aly**, Andrea Trubanova, Lynn Abbott, Susan White, Amira Youssef, VT-KFER: A Kinect-based RGBD+Time Dataset for Spontaneous and Non-Spontaneous Facial Expression Recognition, International Conference of Biometrics (ICB), 2015.
4. **Sherin Aly**, Amira Youssef, Lynn Abbott, Adaptive Feature Selection And Data Pruning For 3D Facial Expression Recognition Using The Kinect, International Conference of Image Processing (ICIP), 2014.
5. Amira E. Youssef, **Sherin F. Aly**, Ahmed S. Ibrahim, and A. Lynn Abbott, "Auto-Optimized Multimodal Expression Recognition Framework Using 3D Kinect Data for ASD Therapeutic Aid," International Journal of Modeling and Optimization vol. 3, no. 2, pp. 112-115, 2013.

7.5 Awards

Parts of this dissertation have received the following awards/recognitions:

1. International Conference of Biometrics (ICB2015) Doctoral Consortium;
2. Poster Presentation Bronze Award, 31^{st} Annual Graduate Student Assembly (GSA) Research Symposium, Virginia Tech, March 2015;
3. Poster Competition Scholarship Award, Second Place, Fifth Interdisciplinary Research (IDR) Symposium, Virginia Tech, February 2015.

Bibliography

[1] "NimStim face stimulus set," http://www.macbrain.org/resources.htm, accessed: 10-01-2014.

[2] S. Aly, A. Trubanova, A. L. Abbott, S. White, and A. Youssef, "VT-KFER: A Kinect-based RGBD+Time dataset for spontaneous and non-spontaneous facial expression recognition," in *International Conference on Biometrics (ICB)*, May 2015, pp. 90–97.

[3] Y. Tian, T. Kanade, and J. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*. Springer, 2005, pp. 247–275.

[4] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *International Conference on Automatic Face & gesture recognition (FGR)*. IEEE, 2006, pp. 211–216.

[5] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2008, pp. 1–6.

[6] C. Shan, S. Gong, and P. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[7] D. Matsumoto and P. Ekman, "Facial expression analysis," *Scholarpedia*, vol. 3, no. 5, p. 4237, 2008.

[8] S. Aly, A. Youssef, and L. Abbott, "Adaptive feature selection and data pruning for 3D facial expression recognition using the Kinect," in *International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1361–1365.

[9] P. Ekman and W. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[10] G. Young-Browne, H. M. Rosenfeld, and F. D. Horowitz, "Infant discrimination of facial expressions," *Child Development*, pp. 555–562, 1977.

[11] N. Soken and A. Pick, "Infants' perception of dynamic affective expressions: do infants distinguish specific expressions?" *Child development*, pp. 1275–1282, 1999.

[12] S. J. Kirsh and J. R. Mounts, "Violent video game play impacts facial emotion recognition," *Aggressive Behavior*, vol. 33, no. 4, pp. 353–358, 2007.

[13] S. Bakkes, C. T. Tan, and Y. Pisan, "Personalised gaming," *Journal of Creative Technologies*, vol. 3, 2012.

[14] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *International Conference on Autonomous Agents and Multi-agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[15] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 664–674, 2011.

[16] M. Jabon, J. Bailenson, E. Pontikakis, L. Takayama, and C. Nass, "Facial expression analysis for predicting unsafe driving behavior," *Pervasive Computing*, no. 4, pp. 84–95, 2010.

[17] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," in *International Workshop on Human-Computer Interaction*. Springer, 2007, pp. 6–18.

[18] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[19] Z. Duric, W. D. Gray, R. Heishman, F. Li, A. Rosenfeld, M. J. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1272–1289, 2002.

[20] L. Maat and M. Pantic, "Gaze-X: adaptive, affective, multimodal interface for single-user office scenarios," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 251–271.

[21] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu, "Robovie: an interactive humanoid robot," *Industrial robot: An International Journal*, vol. 28, no. 6, pp. 498–504, 2001.

[22] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[23] "Emotient," http://www.emotient.com, accessed: 2016-07-12.

[24] "Affectiva," http://www.affectiva.com, accessed: 2016-07-12.

[25] "Realeyesit," http://www.realeyesit.com, accessed: 2016-07-12.

[26] "Kairos," http://www.kairos.com, accessed: 2016-07-12.

[27] A. A. Salah, N. Sebe, T. Gevers *et al.*, "Communication and automatic interpretation of affect from facial expressions," *Affective Computing and Interaction*, pp. 157–183, 2010.

[28] P. Ekamn and W. Friesen, "Facial action coding system (facs): manual," 1978.

[29] P. Ekman, W. V. Friesen, and C. P. Press, *Pictures of facial affect*. Consulting Psychologists Press, 1975.

[30] R. Birdwhistell, "Background considerations to the study of the body as a medium of expression," *The body as a medium of expression. New York: EP Dutton and Co., Inc. p*, pp. 36–58, 1975.

[31] S. Jain, S. Bagga, R. Hablani, N. Chaudhari, and S. Tanwani, "Facial expression recognition using local binary patterns with different distance measures," in *Intelligent Computing, Networking, and Informatics*. Springer, 2014, pp. 853–862.

[32] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, pp. 356–361.

[33] P. Shen, S. Wang, and Z. Liu, "Facial expression recognition from infrared thermal videos," in *Intelligent Autonomous Systems 12*, 2013, pp. 323–333.

[34] D. M. Deriso, J. Susskind, J. Tanaka, P. Winkielman, J. Herrington, R. Schultz, and M. Bartlett, "Exploring the facial expression perception-production link using real-time automated facial expression recognition," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 270–279.

[35] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, "Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in *Facial and Bodily Expressions for Control and Adaptation of Games Workshop (ECAG)*, 2008, p. 3.

[36] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," vol. 16, no. 1, 2007, pp. 172–187.

[37] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[38] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[40] H. Tang and T. S. Huang, "3D facial expression recognition based on automatically selected features," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2008, pp. 1–8.

[41] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1449–1461, 2004.

[42] B. Braathen, M. S. Bartlett, G. Littlewort, E. Smith, and J. R. Movellan, "An approach to automatic recognition of spontaneous facial actions," in *International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2002, pp. 360–365.

[43] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod, "Model-based face tracking for view-independent facial expression recognition," in *International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2002, pp. 287–293.

[44] Z. Wen and T. S. Huang, "Capturing subtle facial motions in 3D face tracking," in *International Conference on Computer Vision*. IEEE, 2003, pp. 1343–1350.

[45] L. Zalewski and S. Gong, "Synthesis and recognition of facial expressions in virtual 3d views," in *International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2004, pp. 493–498.

[46] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[47] A. Danelakis, T. Theoharis, and I. Pratikakis, "A survey on facial expression recognition in 3D video sequences," *Multimedia Tools and Applications*, pp. 1–39, 2014.

[48] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3D facial expression recognition: A perspective on promises and challenges," in *International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*. IEEE, 2011, pp. 603–610.

[49] C. Beumier and M. Acheroy, "3D facial surface acquisition by structured light," in *International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging*, 1999.

[50] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3D facial expression analysis in videos," in *Analysis and Modelling of Faces and Gestures*. Springer, 2005, pp. 293–307.

[51] M. B. Vieira, L. Velho, A. Sa, and P. C. Carvalho, "A camera-projector system for real-time 3D video," in *Conference on Computer Vision and Pattern Recognition-Workshops (CVPR Workshops)*. IEEE, 2005, pp. 96–96.

[52] F. Tsalakanidou, F. Forster, S. Malassiotis, and M. G. Strintzis, "Real-time acquisition of depth and color images using structured light and its application to 3D face recognition," *Real-Time Imaging*, vol. 11, no. 5, pp. 358–369, 2005.

[53] S. Zhang and P. Huang, "High-resolution, real-time 3D shape acquisition," in *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*. IEEE, 2004, pp. 28–28.

[54] S. Zhang and S.-T. Yau, "High-resolution, real-time 3D absolute coordinate measurement based on a phase-shifting method," *Optics Express*, vol. 14, no. 7, pp. 2644–2649, 2006.

[55] "Minolta Vivid 910," http://www.konicaminolta.com/instruments/products/3{D}/non-contact/vivid910/features.html, accessed: 2013-10-22.

[56] "Inspeck Mega Capturor II Digitizer," http://www.inspeck.com, accessed: 2013-10-22.

[57] "MS Kinect," http://www.xbox.com/en-GB/Kinect, accessed: 2013-10-22.

[58] "The science behind Kinects or Kinect 1.0 versus 2.0," http://www.gamasutra.com/blogs/DanielLau/20131127/205820/The_Science_Behind_Kinects_or_Kinect_10_versus_20.php, accessed: 10/22/2014.

[59] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, *Time-of-flight cameras and microsoft Kinect*. Springer, 2012.

[60] "MS Kinect sdk," http://www.microsoft.com/en-us/{K}inectforwindows/, accessed: 2013-10-22.

[61] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[62] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1011139631724

[63] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European Conference on Computer Vision*. Springer, 2004, pp. 469–481.

[64] T. S. Lee, "Image representation using 2d gabor wavelets," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[65] D. Cosker, E. Krumhuber, and A. Hilton, "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2296–2303.

[66] H. Soyel and H. Demirel, "Optimal feature selection for 3D facial expression recognition using coarse-to-fine classification," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 18, no. 6, pp. 1031–1040, 2010.

[67] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction." in *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, vol. 5. IEEE, 2003, pp. 53–53.

[68] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 160–187, 2003.

[69] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 1998, pp. 454–459.

[70] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," in *ACM SIGGRAPH 2011 papers*, New York, NY, USA, 2011, pp. 77:1–77:10.

[71] "Replicating the emotions of a facial expression on a furhat robot face using a Kinect input," http://www.csc.kth.se/utbildning/kth/kurser/DD143X/dkand13/Group6Gabriel/report/report_david_thomas.pdf, accessed: 2014-1-27.

[72] B. Li, A. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Workshop on Applications of Computer Vision (WACV)*. IEEE, 2013, pp. 186–192.

[73] B. Seddik, H. Maamatou, S. Gazzah, T. Chateau, and N. Ben Amara, "Unsupervised facial expressions recognition and avatar reconstruction from kinect," in *International Multi-Conference on Systems, Signals Devices (SSD)*, March 2013, pp. 1–6.

[74] H. Soyel and H. Demirel, "Facial expression recognition using 3D facial feature distances," in *International Conference Image Analysis and Recognition*. Springer, 2007, pp. 831–838.

[75] X. Li, Q. Ruan, and Y. Ming, "3D facial expression recognition based on basic geometric features," in *International Conference on Signal Processing*. IEEE, 2010, pp. 1366–1369.

[76] U. Tekguc, H. Soyel, and H. Demirel, "Feature selection for person-independent 3D facial expression recognition using nsga-ii," in *International Symposium on Computer and Information Sciences*. IEEE, 2009, pp. 35–38.

[77] T. Sha, M. Song, J. Bu, C. Chen, and D. Tao, "Feature level analysis for 3D facial expression recognition," *Neurocomputing*, vol. 74, no. 12, pp. 2135–2141, 2011.

[78] R. Srivastava and S. Roy, "3D facial expression recognition using residues," in *IEEE Region 10 Conf. TENCON*, 2009, pp. 1–5.

[79] S. Ramanathan, A. Kassim, Y. Venkatesh, and W. S. Wah, "Human facial expression recognition using a 3D morphable model," in *International Conference on Image Processing*. IEEE, 2006, pp. 661–664.

[80] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3D face by exploring shape deformation," in *International Conference on Multimedia*. ACM, 2009, pp. 569–572.

[81] X. Zhao, D. Huang, E. Dellandréa, and L. Chen, "Automatic 3D facial expression recognition based on a Bayesian Belief Net and a statistical facial feature model," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 3724–3727.

[82] A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Local 3D shape analysis for facial expression recognition," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 4129–4132.

[83] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*. IEEE, 2011, pp. 414–421.

[84] Y. Sun, M. Reale, and L. Yin, "Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition," in *International Conference on Automatic Face & Gesture Recognition(FG)*. IEEE, 2008, pp. 1–8.

[85] C. Lanz, B. Olgay, J. Denzler, and H.-M. Gross, "Automated classification of therapeutic face exercises using the Kinect," in *International Conference on Computer Vision Theory and Application*, 2013.

[86] R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. Moeslund, and G. Tranchet, "An RGB-D database using microsoft's kinect for windows for face detection," in *International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, 2012, pp. 42–46.

[87] R. Yampolskiy, B. Klare, and A. Jain, "Face recognition in the virtual world: Recognizing avatar faces," in *International Conference on Machine Learning and Applications (ICMLA)*, vol. 1, 2012, pp. 40–45.

[88] T. Baltrusaitis, P. Robinson, and L. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2610–2617.

[89] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *Pattern Recognition*, vol. 44, no. 8, pp. 1581–1589, 2011.

[90] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for 3D facial expression recognition," in *International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 1270–1275.

[91] X. Lu, A. K. Jain, and S. C. Dass, "3D facial expression modeling for recognition," in *Defense and Security*. International Society for Optics and Photonics, 2005, pp. 113–121.

[92] I. Mpiperis, S. Malassiotis, V. Petridis, and M. G. Strintzis, "3D facial expression recognition using swarm intelligence," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 2133–2136.

[93] S.-Y. Chun, C.-S. Lee, and S.-H. Lee, "Facial expression recognition using extended local binary patterns of 3D curvature," in *Multimedia and Ubiquitous Engineering*. Springer, 2013, pp. 1005–1012.

[94] S. Moore and R. Bowden, "The effects of pose on facial expression recognition," in *British Machine Vision Conference (BMVC)*, 2009, pp. 1–11.

[95] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, 1978.

[96] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*. Springer, 2008, pp. 47–56.

[97] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *International Conference on Automatic Face Gesture Recognition*, 2008, pp. 1–6.

[98] R. Habibu, M. Syamsiah, M. M. Hamiruce, and S. M. Iqbal, "UPM-3D facial expression recognition database (UPM-3DFE)," in *PRICAI 2012: Trends in Artificial Intelligence*. Springer, 2012, pp. 470–479.

[99]

[100] G. Stratou, A. Ghosh, P. Debevec, and L. Morency, "Effect of illumination on automatic expression recognition: a novel 3d relightable facial database," in *International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*. IEEE, 2011, pp. 611–618.

[101] "3DMD 3D Face Capturing ," http://www.3dmd.com/3dmdface.html, accessed: April, 2014.

[102] "DI4D 4D Capture Systems," http://www.di3d.com/products/4dsystems/, accessed: April 2014.

[103] G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency, "Exploring the effect of illumination on automatic expression recognition using the ICT-3DRFE database," *Image and Vision Computing*, vol. 30, no. 10, pp. 728–737, 2012.

[104] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+ 3D facial action and expression recognition," *Pattern Recognition*, vol. 43, no. 5, pp. 1763–1775, 2010.

[105] C. Chen, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: a 3D facial expression database for visual computing," *Transactions on Visualization and Computer Graphics*, 2013.

[106] B. Y. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Workshop on Applications of Computer Vision (WACV)*. IEEE, 2013, pp. 186–192.

[107] R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet, "An RGB-D database using Microsoft's Kinect for windows for face detection," in *International Conference on Signal Image Technology and Internet Based Systems (SITIS)*. IEEE, 2012, pp. 42–46.

[108] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 40, no. 3, pp. 449–460, 2010.

[109] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "Using multi-instance enrollment to improve performance of 3D face recognition," *Computer Vision and Image Understanding*, vol. 112, no. 2, pp. 114–125, 2008.

[110] C. Zhong, Z. Sun, and T. Tan, "Robust 3D face recognition using learned visual codebook," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–6.

[111] "Casia 3D face database," http://www.cbsr.ia.ac.cn/english/3DFaceDatabases, accessed: April 2014.

[112] A. Moreno and A. Sanchez, "Gavabdb: a 3D face database," in *Proc. 2nd COST275 Workshop on Biometrics on the Internet*, 2004, pp. 75–80.

[113] T. Heseltine, N. Pears, and J. Austin, "Three-dimensional face recognition using combinations of surface feature map subspace components," *Image and Vision Computing*, vol. 26, no. 3, pp. 382–396, 2008.

[114] "York 3D database," http://www-users.cs.york.ac.uk/nep/research/3Dface/tomh/3DFaceRecognition.html, accessed: April 2014.

[115] S. Gupta, K. Castleman, M. Markey, and A. Bovik, "Texas 3D face recognition database," in *Southwest Symposium on Image Analysis & Interpretation (SSIAI)*. IEEE, 2010, pp. 97–100.

[116] "Texas 3D Face Recognition Database," http://live.ece.utexas.edu/research/texas3dfr, accessed: April 2014.

[117] H. Rabiu, M. I. Saripan, S. Mashohor, and M. H. Marhaban, "3D facial expression recognition using maximum relevance minimum redundancy geometrical features," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–8, 2012.

[118] A. Garg and R. Bajaj, "Facial expression recognition & classification using hybridization of ICA, GA, and neural network for human-computer interaction," *Journal of Network Communications and Emerging Technologies (JNCET)*, vol. 2, no. 1, 2015.

[119] A. Azazi, S. L. Lutfi, I. Venkat, and F. Fernández-Martínez, "Towards a robust affect recognition: Automatic facial expression recognition in 3D faces," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3056–3066, 2015.

[120] A. Youssef, S. Aly, A. Ibrahim, and A. L. Abbott, "Auto-optimized multimodal expression recognition framework using 3D Kinect data for ASD therapeutic aid," *International Journal of Modeling and Optimization*, vol. 3, no. 2, p. 112, 2013.

[121] X. Mingliang, A. Mian, L. Wanquan, and L. Ling, "Automatic 4D facial expression recognition using DCT features," in *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2015, pp. 199–206.

[122] A. Moeini and H. Moeini, "Multimodal facial expression recognition based on 3D face reconstruction from 2D images," in *Face and Facial Expression Recognition from Real World Videos*, ser. Lecture Notes in Computer Science, vol. 8912. Springer, 2015, pp. 46–57.

[123] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[124] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.

[125] Q. Mao, X. Pan, Y. Zhan, and X. Shen, "Using Kinect for real-time emotion recognition via facial expressions," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 4, pp. 272–282, 2015.

[126] F. Malawski, B. Kwolek, and S. Sako, "Using Kinect for facial expression recognition under varying poses and illumination," in *International Conference on Active Media Technology*. Springer, 2014, pp. 395–406.

[127] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications," *Transactions on pattern Analysis and Machine Intelligence*, 2016.

[128] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[129] J. Webb and J. Ashley, *Beginning Kinect Programming with the Microsoft Kinect SDK*. Apress, 2012.

[130] H. Soyel and H. Demirel, "3D facial expression recognition with geometrically localized facial features," in *International Symposium on Computer and Information Sciences (ISCIS)*. IEEE, 2008, pp. 1–4.

[131] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. Huang, "A study of non-frontal-view facial expressions recognition," in *International Conference on Pattern Recognition (ICPR)*, Dec 2008, pp. 1–4.

[132] H. Soyel and H. Demirel, "Optimal feature selection for 3D facial expression recognition with geometrically localized facial features," in *International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (ICSCCW)*. IEEE, 2009, pp. 1–4.

[133] A. Savran, B. Sankur, and M. Bilge, "Facial action unit detection: 3D versus 2D modality," in *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*. IEEE, 2010, pp. 71–78.

[134] S. Aly, A. L. Abbott, and M. Torki, "A multi-modal feature fusion framework for Kinect-based facial expression recognition using dual kernel discriminant analysis (DKDA)," in *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 71–78.

[135] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 1399–1406.

[136] H. Ujir, M. Spann, and I. H. M. Hipiny, "3D facial expression classification using 3D facial surface normals," in *International Conference on Robotic, Vision, Signal Processing & Power Applications*. Springer, 2014, pp. 245–253.

[137] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3D/4D facial expression analysis: An advanced annotated face model approach," *Image and Vision Computing*, vol. 30, no. 10, pp. 738–749, 2012.

[138] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *European Conference on Computer Vision (ECCV)*. Springer, 2008, pp. 58–71.

[139] X. Zhao, E. Dellandréa, L. Chen, and D. Samaras, "AU recognition on 3D faces based on an extended statistical facial feature model," in *International Conference on Biometrics: Theory Applications and Systems (BTAS)*. IEEE, 2010, pp. 1–6.

[140] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Deep learning for emotion recognition in faces," in *International Conference on Artificial Neural Networks*. Springer, 2016, pp. 38–46.

[141] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep Convolutional Neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

[142] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1749–1756.

[143] J. Susskind, V. Mnih, G. Hinton *et al.*, "On deep generative models with applications to recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 2857–2864.

[144] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in Deep Neural Networks for facial expression recognition," in *International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 2983–2991.

[145] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *International Conference on Natural Computation (ICNC),*. IEEE, 2015, pp. 702–708.

[146] R. Zhu, T. Zhang, Q. Zhao, and Z. Wu, "A transfer learning approach to cross-database facial expression recognition," in *2015 International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 293–298.

[147] L. A. Alexandre, "3D object recognition using convolutional neural networks with transfer learning between input channels," in *Intelligent Autonomous Systems*. Springer, 2016, pp. 889–898.

[148] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 681–687.

[149] K. Yurtkan and H. Demirel, "Feature selection for improved 3D facial expression recognition," *Pattern Recognition Letters*, vol. 38, no. 0, pp. 26 – 33, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513004182

[150] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[151] I. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2002.

[152] J. Yang, A. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[153] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990.

[154] B. Scholkopft and K. Mullert, "Fisher discriminant analysis with kernels," *Neural Networks for Signal Processing IX*, vol. 1, p. 1, 1999.

[155] A. Martínez and A. Kak, "PCA versus LDA," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[156] G. Baudat and F. Anouar, "Generalized Discriminant Analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[157] D. Cai, X. He, and J. Han, "Speed up Kernel Discriminant Analysis," *The VLDB Journal*, vol. 20, no. 1, pp. 21–33, 2011.

[158] N. Vretos, N. Nikolaidis, and I. Pitas, "3D facial expression recognition using Zernike moments on depth images," in *International Conference on Image Processing (ICIP)*. IEEE, 2011, pp. 773–776.

[159] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for 3D facial expression recognition," in *International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 1270–1275.

[160] M. Rosato, X. Chen, and L. Yin, "Automatic registration of vertex correspondences for 3D facial expression analysis," in *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2008, pp. 1–7.

[161] T. Sha, M. Song, J. Bu, C. Chen, and D. Tao, "Feature level analysis for 3D facial expression recognition," *Neurocomputing*, vol. 74, no. 12, pp. 2135–2141, 2011.

[162] P. Wang, C. Kohler, F. Barrett, R. Gur, and R. Verma, "Quantifying facial expression abnormality in schizophrenia by combining 2d and 3d features," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.

[163] Y. Venkatesh, A. K. Kassim, and O. R. Murthy, "Resampling approach to facial expression recognition using 3D meshes," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 3772–3775.

[164] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-D face and facial expression recognition," *Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 498–511, 2008.

[165] F. Tsalakanidou and S. Malassiotis, "Robust facial action recognition from real-time 3D streams," in *Computer Vision and Pattern Recognition Workshops*, 2009, pp. 4–11.

[166] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do we need more training data or better models for object detection?" in *British Machine Vision Conference (BMVC)*, 2012, pp. 1–11.

[167] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Pattern Recognition*, vol. 43, no. 5, pp. 1763 – 1775, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320309004786

[168] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.

[169] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[170] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. De la Torre, and J. Cohn, "Aam derived face representations for robust facial action recognition," 2006.

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

[172] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007.

[173] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting coded dynamic features for facial action units and facial expression recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–6.

[174] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[175] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.

[176] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.

[177] M. Liu, S. Li, S. Shan, and X. Chen, "AU-aware deep networks for facial expression recognition," in *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–6.

[178] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3D dynamic facial expression database," in *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[179] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[180] Y. Li, J. Chen, Y. Zhao, and Q. Ji, "Data-free prior model for facial action unit recognition," *Transactions on Affective Computing*, vol. 4, no. 2, pp. 127–141, 2013.

[181] G. Sandbach, S. Zafeiriou, and M. Pantic, "Local normal binary patterns for 3D facial action unit detection," in *International Conference on Image Processing (ICIP)*. IEEE, 2012, pp. 1813–1816.

[182] A. Savran, B. Sankur, and M. T. Bilge, "Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units," *Pattern Recognition*, vol. 45, no. 2, pp. 767–782, 2012.

[183] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[184] P. Lang and M. Bradley, "The international affective picture system (iaps) in the study of emotion and attention," *Handbook of Emotion Elicitation and Assessment*, vol. 29, 2007.

[185] U. Hess, S. Blairy, and R. E. Kleck, "The intensity of emotional facial expressions and decoding accuracy," *Journal of Nonverbal Behavior*, vol. 21, no. 4, pp. 241–257, 1997.

[186] A. Viera and J. Garrett, "Understanding interobserver agreement: the kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, 2005.

[187] V. Štruc and N. Pavešić, "The complete Gabor-Fisher classifier for robust face recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 31, 2010.

[188] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Elsevier/Academic Press, 2006. [Online]. Available: www.summon.com

[189] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, 2006, vol. 1.

[190] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[191] C. Hsu, C. Chang, and C. Lin, "A practical guide to support vector classification," *Available at http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf*, 2009.

[192] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[193] K. Khoshelham and S. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[194] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, pp. 886–893.

[195] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[196] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *International Conference on Automatic Face & Gesture Recognition & Workshops (FG)*. IEEE, 2011, pp. 921–926.

[197] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, pp. 2169–2178.

[198] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, 4th ed. Academic Press, 2008.

[199] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3D facial expression database for visual computing," *Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

[200] S. Jain, B. Tamersoy, Y. Zhang, J. K. Aggarwal, and V. Orvalho, "An interactive game for teaching facial expressions to children with autism spectrum disorders," in *International Symposium on Communications Control and Signal Processing (ISCCSP)*. IEEE, 2012, pp. 1–4.

[201] J. Cohn, Z. Ambadar, and P. Ekman, "Observer-based measurement of facial expression with the facial action coding system," *The Handbook of Emotion Elicitation and Assessment*, pp. 203–221, 2007.

[202] W. Friesen and P. Ekman, "Emfacs-7: Emotional Facial Action Coding System," *Unpublished manuscript, University of California at San Francisco*, 1983.

[203] P. M. Woodward, *Probability and Information Theory, with Applications to Radar: International Series of Monographs on Electronics and Instrumentation*. Elsevier, 2014, vol. 3.

Appendix A

VT-KFER Dataset Landmarks Description

In this appendix we will provide a detail description to the 121 keypoints provided in VT-KFER dataset by the Kinect. The 121 keypoints are automatically detected by the Kinect SDK. Figure A.1 illustrates a sample 2D face image with the 121 keypoints plotted over it. Out of these 121 keypoints, table A.1 describes 71 main landmarks in salient parts of the face.

Table A.1: Key Facial Features and the corresponding index in the dataset

Feature	**Idx**	**Feature**	**Idx**
Top Skull	0	Outer Top Right Pupil	67
Top Right Forehead	1	Outer Bottom Right Pupil	68
Middle Top Dip Upper Lip	7	Outer Top Left Pupil	69
Above Chin	9	Outer Bottom Left Pupil	70
Bottom Of Chin	10	Inner Top Right Pupil	71
Right Of Right Eyebrow	15	Inner Bottom Right Pupil	72
Middle Top Of Right Eyebrow	16	Inner Top Left Pupil	73
Left Of Right Eyebrow	17	Inner Bottom Left Pupil	74
Middle Bottom Of Right Eyebrow	18	Right Top Upper Lip	79
Above Mid Upper Right Eyelid	19	Left Top Upper Lip	80
Outer Corner Of Right Eye	20	Right Bottom Upper Lip	81
Middle Top Right Eyelid	21	Left Bottom Upper Lip	82
Middle Bottom Right Eyelid	22	Right Top Lower Lip	83
Inner Corner Right Eye	23	Left Top Lower Lip	84
Under Mid Bottom Right Eyelid	24	Right Bottom Lower Lip	85
Right Side Of Chin	30	Left Bottom Lower Lip	86
Outside Right Corner Mouth	31	Middle Bottom Upper Lip	87
Right Of Chin	32	Left Corner Mouth	88
Right Top Dip Upper Lip	33	Right Corner Mouth	89
Top Left Forehead	34	Bottom Of Right Cheek	90
Middle Top Lower Lip	40	Bottom Of Left Cheek	91
Middle Bottom Lower Lip	41	Above Three Fourth Right Eyelid	95
Left Of Left Eyebrow	48	Above Three Fourth Left Eyelid	96
Middle Top Of Left Eyebrow	49	Three Fourth Top Right Eyelid	97
Right Of Left Eyebrow	50	Three Fourth Top Left Eyelid	98
Middle Bottom Of Left Eyebrow	51	Three Fourth Bottom Right Eyelid	99
Above Mid Upper Left Eyelid	52	Three Fourth Bottom Left Eyelid	100
Outer Corner Of Left Eye	53	Below Three Fourth Right Eyelid	101
Middle Top Left Eyelid	54	Below Three Fourth Left Eyelid	102
Middle Bottom Left Eyelid	55	Above One Fourth Right Eyelid	103
Inner Corner Left Eye	56	Above One Fourth Left Eyelid	104
Under Mid Bottom Left Eyelid	57	One Fourth Top Right Eyelid	105
Left Side Of Cheek	63	One Fourth Top Left Eyelid	106
Outside Left Corner Mouth	64	One Fourth Bottom Right Eyelid	107
Left Of Chin	65	One Fourth Bottom Left Eyelid	108
Left Top Dip Upper Lip	66		

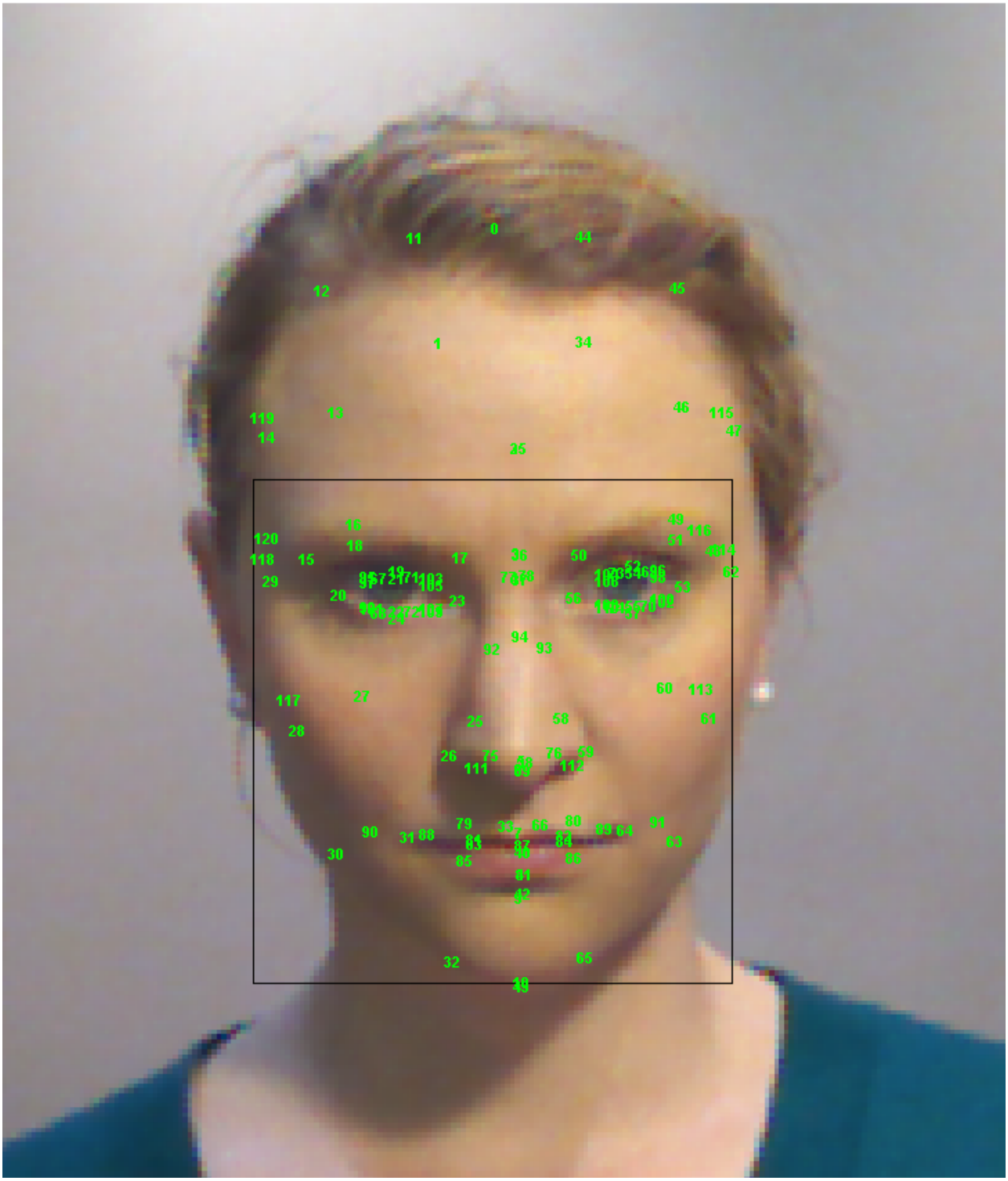


Figure A.1: The 121 landmarks automatically detected by Kinect SDK. The landmarks are numbered from 0 to 120.

Appendix B

VT-KFER Dataset File Structure and Data Format

A brief description of the VT-KFER dataset hierarchy along with the contents of both scripted and unscripted data is provided below.

B.1 Dataset Hierarchy

Figure B.1 illustrates VT-KFER main hierarchy.

Subjects are numbered from 001 to 032. Since the order of expressions displayed to each subject is selected randomly, this corresponding order is save for the scripted and unscripted sessions in a separate file named *ExpressionList*.txt. The main contents of both scripted and unscripted sessions are provided below.

B.2 Folders contents description

For a sample subject 001, the following data is saved:

1. ExpressionList.txt: contains the order of expressions displayed to the corresponding subject in the unscripted and scripted sessions, respectively.
2. Unscripted: Contains data from the unscripted session. A sequence of frames is saved while stimuli images are shown to each subject.
3. Scripted: Contain data from the scripted session, which recorded after the unscripted session.

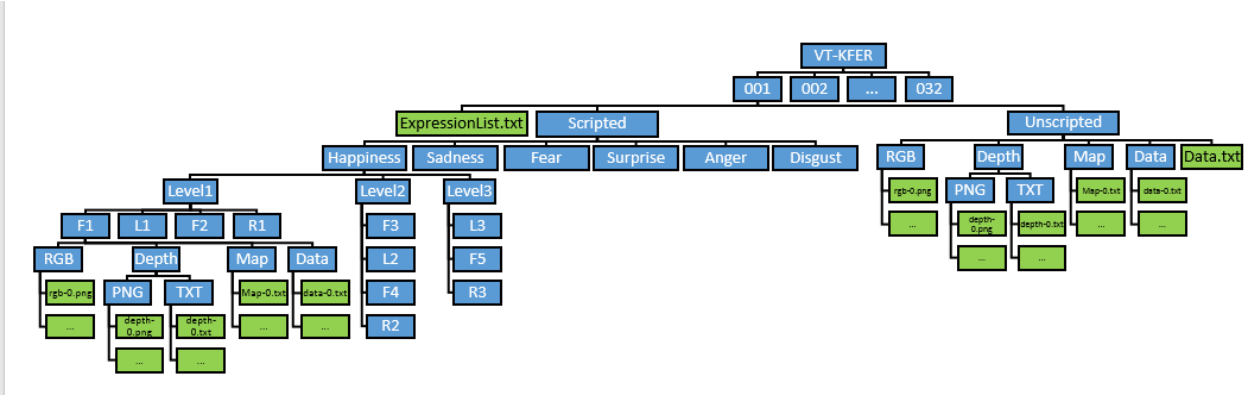


Figure B.1: VT-KFER dataset hierarchy. Blue boxes refer to folders. Green boxes refer to files. Note that the Map folder includes a sub Txt folder where all the .txt files are saved. It is removed from this hierarchy for simplicity.

B.2.1 Unscripted Session

For the unscripted expressions, a RGBD sequence (RGBD+time) of the subject emotions is recorded, where RGB refer to the red, green, and blue channels of the 2D image of the scene, and D refers to the depth. The RGB, Depth, and landmark information are separately saved as follows.

1. **RGB**: This folder contains all the RGB image sequence for the unscripted session with each image is named with a unique timestamp.

2. **Depth**: This folder contains the depth data sequence saved in both gray-scale intensities (i.e., depth map) and depth values in millimeters (mm) format.

3. **Data**: This folder contains sequence of data.timeStamp.txt files where timestamp is the time stamp of the corresponding frame. Each file includes the 121 3D landmarks points coordinates in the corresponding 3D scene and the projected 2D landmarks coordinate in the 2D image. It also includes the ground truth coordinates of the face location in each scene. Each data-timeStamp.txt file is of the next form:

    - Color frame timestamp;
    - Color frame number;
    - Corresponding emotion invoking Image displayed to the subject at this particular time stamp indicating which emotion set (i.e., happiness, sadness, surprise, etc) and which order (a number from 01 to 10 where 01 is the least invoking and 10 is the most invoking);
    - the 121 3D keypoint on the face saved as a list of x, y, and z coordinates,

- the 121 2D projected keypoints on the face saved as a list of x and y coordinates;
- the 2D coordinate of the location of the rectangle surrounding the face in the RGB scene, saved as the upper left corner (i.e, x and y coordinate of the upper left corner), the width, and height.

4. **Map**: To register the 3D and 2d scenes, the mapping information for each depth frame to the corresponding RGB frame is recorded for each frame pairs. For each depth pixel in the depth frame, the corresponding RGB image coordinate is recorded.

B.2.2 Scripted Session

For the scripted expressions, the following data is record at each pose:

1. **RGB**: It contains the sequence of images for this pose, starting by a neutral face and ending by an apex. The images are saved in the form: Rgb-timeStamp.png: where timeStamp corresponds to the image frame's timeStamp value which is unique for every frame per session. The sequence is ordered by this timestamp value.
2. **Depth**: It contains the corresponding depth data saved in two forms, as follows:
   - PNG: gray-scale Png images for the depth channel saved in the form: Depth-timeStamp.png (i.e., Depth image of timeStamp saved as a gray scale image).
   - Bin: binary files contains the raw depth data in mm, saved in the form: Depth-timeStamp.bin (i.e., Raw Depth values in millimeters (mm) saved as an array of short integers (where the depth is stored in two bytes).
3. **Map**:contains the mapping from depth to RGB. For each depth pixel, the corresponding RGB image coordinate is saved.
4. **Data**: This folder contains the data-x.txt files for each captured frame. Each data-x.txt file is composed of 1 line of data of the same form of the unscripted part, except that the displayed image name is replaced with a dash;