Doctoral Dissertation

Estimation of human emotion by analyzing facial expression transition of image sequences

Prarinya Siritanawan

Supervisor: Kazunori Kotani

School of Information Science
Japan Advanced Institute of Science and Technology

Degree conferment : December, 2015

Abstract

Face is a medium to deliver our emotional message to those around us, and we can read it in the otherwise. It is the first engaged area and the most readable area in non-verbal communication. In addition, face has a greatly influence in our way of interacting with others. Generally, there is a particular set of facial muscles that usually appears under the same emotional context. For example, a person reveals a smile in the happiness moment, or shows a sign of distress by wrinkles around forehead and eyebrows. In computer vision aspect, facial expression analysis is basically a supervised learning by classification or regression trained from labelled data. Many previous methods share a common flaw by assuming that facial expression feature can be modeled from a single image of the most intense part of facial expression. Therefore, the existing methods cannot handle the facial expression with less intensity. In addition, the conventional features such as texture features or geometric features are inconsistently varied for each person face. We created a new computer vision method to estimate the emotional messages from the transition of facial image sequences. Since typical motion based features are sensitive to face alignment errors, we proposed a novel robust temporal feature to measure facial activities. Our proposed feature can represent facial activities across space and time and can detect a subtle action of face. Moreover, in the early studies, researchers defined the categories of emotion by a few English words. These categories have been inherited to present. The limitation of emotion class number often induces impractical descriptions of facial expressions. To describe more complex facial expressions without prior assumption of emotion labeling, we applied our robust temporal feature and discriminative subspace method to automatically learn the underlying muscle activations in form of Action Units (AUs) according to Facial Action Coding System (FACS) standard.

Keywords: Facial expression analysis, Emotion, Robust temporal feature, Facial Action Coding System, Human Machine Interaction

Acknowledgments

I wish to express my sincere gratitude to Assoc. Prof. Kazunori Kotani, the supervisor of this research. I truly appreciate his guidance, and his optimistic thinking to encourage his students to confront and learn their own essence of research profession.

I would like to thank to the second supervisor, Assoc.Prof. Atsuo Yoshitaka, and all committee members, Prof. Jianwu Dang, Prof. Mineo Kaneko, and Assoc.Prof. Hirokazu Tanaka from School of Information Science, and Assoc.Prof. Toru Abe from Tohoku University. Thanks to their critical and useful comments, so I can further improve the quality of this dissertation.

I am also grateful to the minor research supervisor, Prof. Bao Tu Ho from School of Knowledge Science, for the valuable discussion and introduction of the other aspects in machine learning research.

Thank to Dr. Fan Chen, Dr. Hung Viet Nguyen and his wife, Suthum Keerativittayanun, Takuto Watanabe, Tsuyoshi Kobayashi, Masato Fujio, and all colleagues in Kotani laboratory and Yoshitaka laboratory for their kind support over the past years in both academic and private life. Furthermore, I would like to thank to my family and Leela Sattayamas for their patient, encouragement and always being supportive beside me.

Thank to funding for doctoral study supported by JAIST under Graduate Research Program (GRP).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Emotions are written all over our faces. Faces are the first area that engaged in social interaction. Since, they are the most expressive and readable area in nonverbal communications, people can make a judgment about what other feeling based on their facial features. With the advancement of the current computer vision and machine learning technology, this intuition is not limited only to human-human communication. It is possible to develop a system that can observe human facial expressions, interpret their explicit meaning, and response back to a user with a proper decision. From old-fashioned heartless machines to sophisticated robots at the present time, we are closer to the dream of building an intelligent machine that can understand our emotional messages. In the frustrated or sad moments, the machines can be our dear friends to share sympathies, and find the way to solve the problems. This kind of system will greatly encourage the study of human behavior in non-invasive environment, which is usually an exhaustive task in the traditional research.

The importance of understanding nonverbal communications has been long recognized. Mehrabian reinforces this importance by his studies about the contributed channels of emotions in 1967 [10][11]. The communication channels of emotions are contributed by 55% from facial expressions, 38% from voice elements, and 7% from spoken contents [12]. Although this famous ratio is not a precise estimation for every context, it well underlies the essential role of facial expression in conveying our emotions.

As we have reached the information era, the idea of interpreting an emotional state by sensors and information processing had been widespread in the last two decades. Picard introduces the term *Affective Computing* in 1995 to describe this multi-disciplinary field [13]. The cues of emotions can be collected through various channels such as facial activities, bodily gestures, speech elements, cardiovascular activities, electro-dermal measures, temperature, etc. Some of them require invasive approaches to measure the emotional signals. In this research, we measure the facial activities since it is non-invasive method. Thus, it is more suitable than the other cues to be observed by a robot or machine for interaction.

Facial expressions usually have multiple functions. Give a facial action, it can have several meaning depend on the context at the moment. It can be an indicator of emotion (happy, sad, angry, surprise, fear, and disgust), an indicator of cognitive process (attention, concentration), social communication approach, or speech-related action. Given a

smile face, or more specifically a contraction of zygomatic major muscle, may not always mean a person is in happiness state. Another example of facial expression for communication is a *wink* expression, which has been often found in many western cultures, acting as a message to a target party secretly, or as a flirting sign. We discussed about the evidences and debates of association between facial expression and emotion later in section 1.4.

To clarify the standpoint of this research, the natural relationship between expression and emotion is not 1-1 corresponding map, but in fact it is many-to-many mapping relationship function. This relationship function requires many emotional indicators (many inputs to the relationship function) to improve and calibrate the estimation of emotional states and other semantic meanings of expressions (many outputs from the relationship function). Such a relationship in engineering implementation is indeed a complex estimation system and it is a big challenge in affective computing field.

However the research in affective computing is still in an infant state which may need more solid psychological theorem to infer the emotion pathway model. Moreover, the studies of each emotional cue still have their own problems that require a lot of attentions to be solved. As a simplification of the complex system, we simplify the research problem and scope the limitation to estimate emotions from a single cue - facial expression - with the assumption of 1-1 relationship as the developmental phase of complex emotional estimation realization.

1.2 Overview

The concept of this research is mainly enveloped around the temporal transition of facial expression analysis. The discussion in the introductory chapter of this dissertation will begin with the background in different research domains such as behavioral psychology, cognitive science, and the methodologies in computer vision and machine learning. The author introduces the novel spatiotemporal features for modeling the transition of facial expression by the level of muscle activation.

In previous research, the interpretation of facial expression to features is frequently produced by using a still image. However, a static analysis solely describes a few parts of all possible facial appearances in the sequence. Since a facial expression of emotion is a combination of many appearances, it is possible to see a part of one emotion expression has a similar feature to a part of another emotion expression (Fig. 1.1). In addition, people tend to display a weak intensity of facial expression instead of an exaggerated one. This often leads to the situations that classification methods recognize the input image as another emotion expression.

In order to deal with the problem in the static analysis, the temporal template method is introduced, where the changes of facial expression are observed in sequence and modeled as the feature for classification. The trace or history of changes in facial expression of a specific emotion has a solid pattern. This intuitive assumption can be formulated as a temporal template. The advantage of the temporal template is it can represent the dynamic of facial expression, and avoid the problem from the different durations of expression by packing the temporal features into a fixed-length feature vector. Moreover, the temporal template can be applied with the conventional classification methods. The baseline method of the temporal template is widely known as the Motion History Image (MHI), which describes the order of actions in action recognition [14][15]. Instead, we

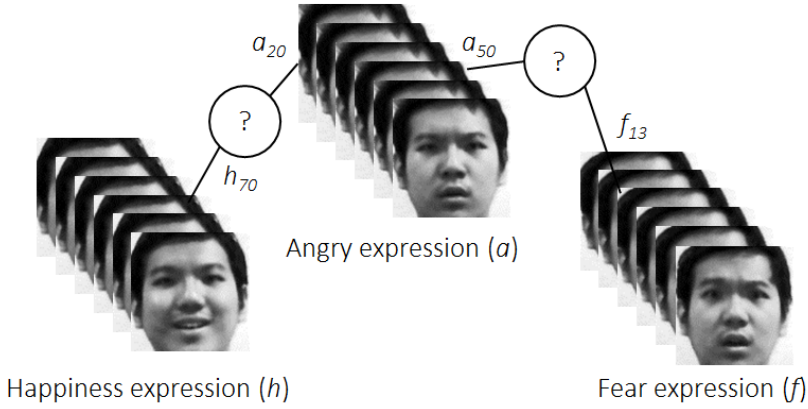


Figure 1.1: Facial expression is the combination of many facial appearances or features in sequence. It is possible to see a part of one emotion expression has a similar feature to a part of another emotion expression.

alternatively represent the temporal information in form of facial muscle activation level by Cumulative Changes of Feature (CCF) (*Publication I.* and *Publication II.*).

However, the presences of small variations such as translation, scaling, or blurriness affects the quality of the CCFs. Therefore, we extend the Gabor features [16] in the CCF framework. The Gabor features have the similar mechanism to the human visual system which they contain both spatial and frequency information. Thus, the proposed feature extraction method gains the advantages of two biological characteristics. Firstly, the ability of recognizing a facial expression of an emotion is associated by the degree of perceptible motions [17]. Secondly, the robustness of the Gabor features. As a result, we can derive the novel extension of the 2D Gabor filter such that it can summarize the pattern of the facial expression over a period of time. We call this temporal feature as a Cumulative Differential Gabor features (CDG) (*Publication III.*).

In classification system, we consider the subspace approaches for estimating the facial expression classes. The objective of the subspace method is to recreate the new basis vectors of data such that we can illustrate the clear visualization from messy data, or extract important components in class separation. Subspace techniques have been widely associated with the research on face features such as face recognition, since they can extract significant components out of an individual face. The well-known subspace methods are principle component analysis (PCA) linear discriminant analysis (LDA))[18], or independent component analysis (ICA)[19]. Facial expression analysis is a class separation problem rather than an individual identification issue. Thus, the role of subspace methods in the facial expression classification is to find the proper subspace that can separate the data between different classes to the highest degree. In this research, we implemented several class separable subspace methods such as linear classifier by Eigenspace Method based on Class features (EMC)[6] and its variants by using kernel (KEMC)[20] or discriminative independent components (ICA with whitened EMC)[7]. In order to evaluate the performance our proposed features and classification methods, we perform the experiments on our original dataset and standard dataset such as Extended Cohn-Kanade (CK+) dataset [4] and Japanese Female Facial Expression (JAFFE) dataset [3]. For each datasets, we choose 70% of samples in the datasets as the training set, and the

rest for testing. The experimental results confirm the feasibility of the features and the separability of classification methods under different simulation conditions including the inter-personal setting.

Another issue of this dissertation concerns the problem of using of basic emotion categorization. Several facial expression recognition studies utilized the emotion categorization according to Ekmans basic emotion categorization [21][22]. Although the advantage of this categorization is the simplicity in the implementation, but the investigation of facial expression is limited to only 6 emotion classes according to the specific combinations of the muscle responses (anger, fear, happiness, disgust, sadness, surprise). However, the facial expressions in our life basis are not confined to such a limited set of basic emotions. There are other variations of facial expressions which cannot be categorized as a particular basic emotion class. In order to explain more complex facial expressions, we proposed a novel action unit (AU) detector following the Ekman's Facial Action Coding System (FACS) [23][2](*Publication VI.*). Our AU detection system use our proposed robust temporal feature CDG combining with a new architecture of classification by ICA with whitened EMC as a binary classifier for each AUs. Therefore we can objectively describe the complex facial expression in the same standard in psychology studies. Furthermore, in order to describe emotions beyond basic categorization, we investigate a feasibility to realize such a system in computer vision. The non-basic emotion can be recognized as mixtures of basic emotions, expansion of emotion categories, or dimensional emotion models. These issues are discussed in the Appendix A of this dissertation.

1.3 Facial Structure

Appearances of face are constructed by several components: bony structure (skull), muscles, fatty tissues, and skin tissues. The frontal view of muscles, which are responsible for facial expression, is shown in Fig. 1.2. The corresponding Action Units (AUs) in Facial Action Coding System are also noted in Table 1.1. The terms to refer the visible changes of face are shown in Fig. 1.3.

Basically, the contraction and relaxation of muscle fibers produce changes of face appearances. When a facial muscle is contracted, the shape of face changes not only the activated area, but also including the other parts of face too since most components are close to each other and several parts are well connected i.e. action of a particular muscle can draw a large skin surface area, fatty tissues, as well as the other muscles under face skin. For example, a contraction of *Orbicularis oculi (pars orbitalis)-AU6*, which is located around the eye orbit, causes both upper and lower facial surface to change i.e. the infraorbital triangle and chin to rise by pulling skin toward the eye. In addition, age, gender, and amount of fat are factors that effect to appearances of faces. A good example of these differences is the visibility of the crow's feet and nasolabial furrow, which the lines are clearly shown in the older persons. These factors should be considered when design an automatic facial expression recognition system.

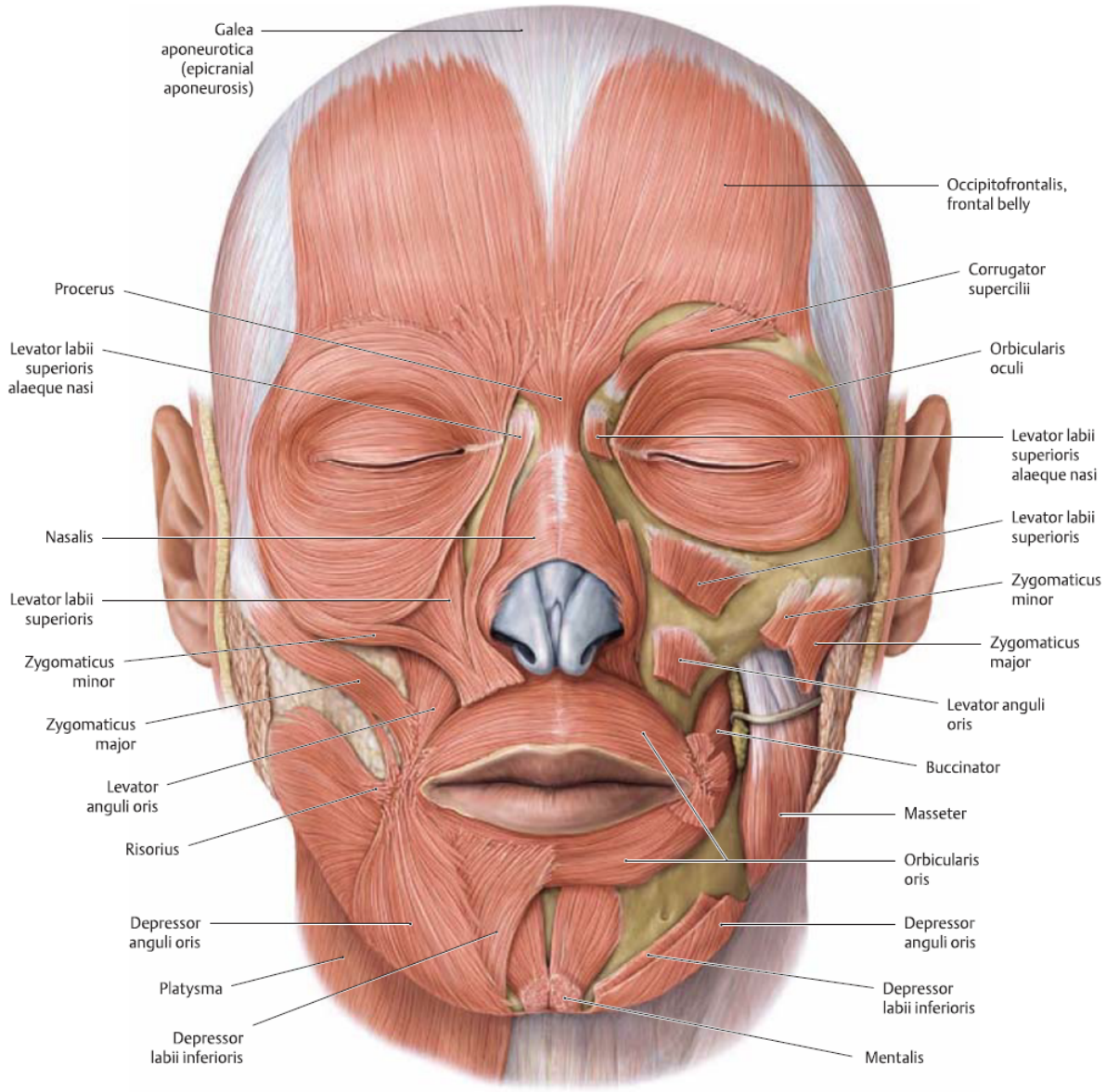


Figure 1.2: The illustration of facial muscles related to facial expression. The outer layer is shown in the right half of the face, while the deeper layer is on the left half (source image from [1]).

Table 1.1: Action Units (AUs) and corresponding facial muscles in Ekman's Facial Action Coding System (FACS) [9].

AU Number	FACS Name	Muscular Basis
1	Inner Brow Raiser	Frontalis (Pars Medialis)
2	Outer Brow Raiser	Frontalis (Pars Lateralis)
4	Brow Lowerer	Depressor Glabellae; Depressor Supercilli; Corrugator Supercilii
5	Upper Lid Raiser	Levator Palpebrae Superioris

Continued on next page

AU Number	FACS Name	Muscular Basis
6	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
7	Lid Tightener	Orbicularis Oculi, Pars Palebralis
8	Lips Toward Each Other	Orbicularis Oris
9	Nose Wrinkler	Levator Labii Superioris (Alaeque Nasi)
10	Upper Lip Raiser	Levator Labii Superioris (Caput Infraorbitalis)
11	Nasolabial Deepener	Zygomatic Minor
12	Lip Corner Puller	Zygomatic Major
13	Sharp Lip Puller	Levator Anguli Oris (also known as Caninus)
14	Dimpler	Buccinnator
15	Lip Corner Depressor	Depressor Anguli Oris (also known as Triangularis)
16	Lower Lip Depressor	Depressor Labii Inferioris
17	Chin Raiser	Mentalis
18	Lip Pucker	Incisivii Labii Superioris; Incisivii Labii Inferioris
20	Lip Stretcher	Risorius
22	Lip Funneler	Orbicularis Oris
23	Lip Tightener	Orbicularis Oris
24	Lip Pressor	Orbicularis Oris
25	Lips Part	Depressor Labii Inferioris, or Relaxation of Mentalis or Orbicularis Oris
26	Jaw Drop	Masetter; Relaxation of Temporalis and Internal Pterygoid
27	Mouth Stretch	Pterygoids; Digastric
28	Lip Suck	Orbicularis Oris
38	Nostril Dilator	Nasalis (Pars Alaris)
39	Nostril Compressor	Nasalis (Pars Transversa) and Depressor Septi Nasi
41	Lid Droop	Relaxation of Levator Palpebrae Superioris
42	Inner Eyebrow Lowerer (Eye slit)	Orbicularis Oculi
43	Eyes Closed	Relaxation of Levator Palpebrae Superioris
44	Eyebrow Gatherer (Squint)	Orbicularis Oculi (Pars Palpebralis)
45	Blink	Relaxation of Levator Palpebrae and Contraction of Orbicularis Oculi (Pars Palpebralis)
46	Wink	Orbicularis Oculi

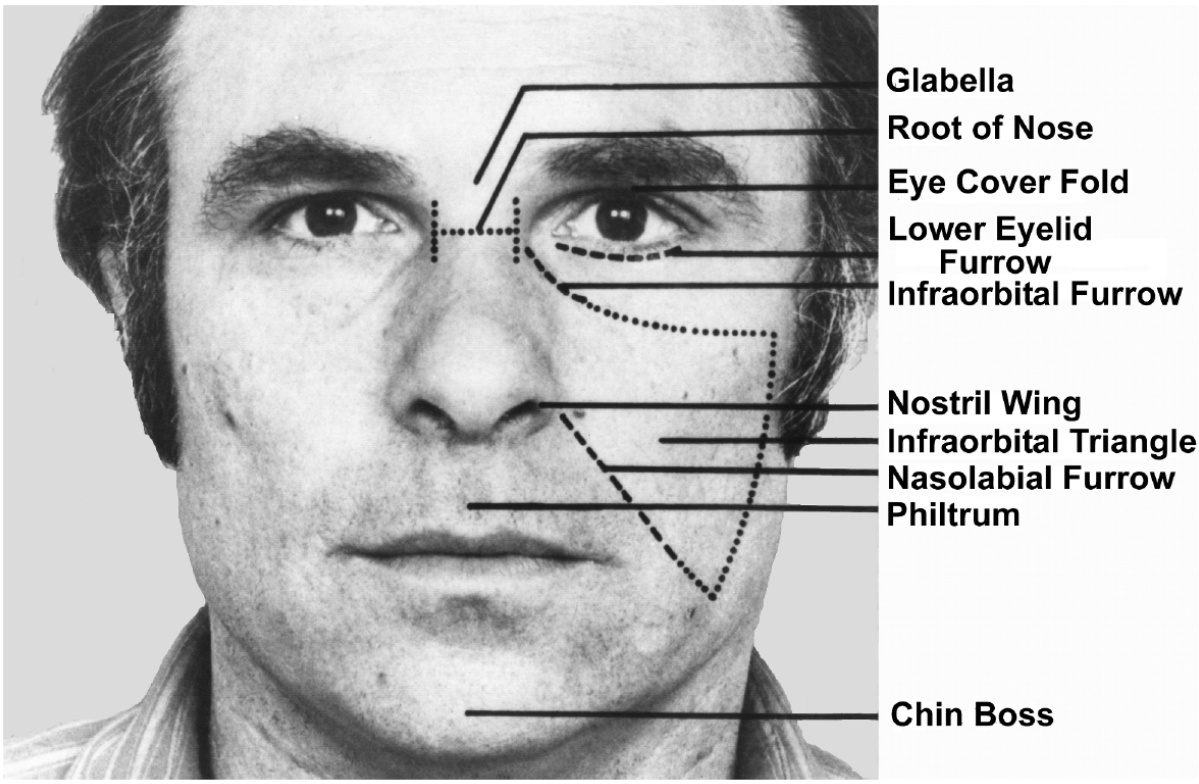


Figure 1.3: Location of facial components (source image from FACS manual [2])

1.4 Facial expression and emotion

Emotion study is one of the multi-century prolonged debates. From classical philosophy of mind in ancient Greek to modern psychology theories, the definitions of emotions and their characteristics have been developed with heterogeneous opinions across professions.

The relation of facial expression and emotion has been discussed in many contemporary studies. Emotions are often assumed to influence the autonomic nervous system [24][25], and therefore produces involuntary actions of faces [26][27]. This statement has been supported by the founding in blinded persons who can express the expression of pleasure and unpleasant, even though they had never seen any facial expression in their life before [28].

Intuitively, people can infer what other feeling by looking on their faces, as well as hundreds of scientific studies support the idea that faces are emotion indicators. However, it is an open question whether the ability to recognize emotion from faces is biologically embedded, or environmental evolved.

For the idea of being biological embedded ability, supporters claimed that the functionalities of facial expression have been inherited thought evolutions of species. Earlier researchers found that dogs and chimpanzees also used facial expression in communication (and recent research also found a similar trend in horses [29]). Back to a hundred years ago, Charles Darwin observed a specific set of facial expressions in human and animals. He noticed the universality in expressing and recognizing those expressions [30]. Motivated by this pioneer study, Paul Ekman and his colleagues did the cross cultural experiments on the universality of facial expressions. They presumed that there are sets

of facial muscles actions which are highly correlated to the emotional states. These facial expressions of emotions can be universally recognized by every human regardless to their cultures or conceptual contexts. The conversion of facial muscles responses and emotions are noted in Ekman and Friesen's FACS investigator guide [9] (See Table 1.2). Their series of cross cultural study became the primary evidences in current affective computing research.

On the contrary, the feasibility of direct interpretation from facial expressions to emotions is doubted. The universality assumption has been opposed by several renowned researchers, for example, James Russell [31][32] noticed the problem of using English words to represent the emotions, which can be lost in translations. Emotion words in one language may be absent in another language [33]. In addition, the conventional method of proving the universality in previous studies had been done by asking subjects to match a photograph of facial expression to a limited set of emotion words. Therefore, the subjects are unintentionally guided, and this skewed the results. To prove this argument, Gendron et al. revised a methodology used in Ekman's by asking subjects to group the face images by themselves instead of limiting the choice of labels, and found the negative results on universal issue [34].

In the midst of the inconclusive debates, the facial expression recognition trait is indeed both innate and cultural dependency. The formulation of facial expression and emotion should not assumed with single-mind that everyone produces and reads facial expression in the exact same way, or in the opposite direction. In correspond to this issue, we anticipate that there should be a middle path to compromise both assumptions, and the way to archive the personalized imprint of facial expression of emotion can be formulated by computer vision and machine learning approaches.

Beside the above issues, facial expressions of emotions are often distorted by social rules or traditions. Faces are masked by another expression to cover their true feeling (e.g. smile in a frustrated situation), de-intensifying (action is too weak to be detected) or over-intensifying (very strong expression found in television acting), or even expressionless (poker face). In some circumstance, the suppressed emotional expression may subtlety leak to outside. Ekman called this involuntary behavior as a *micro-expression*. Nevertheless there is no certain evidence to support this kind of leaked expression whether it is biological or environmental characteristics, as well as if it is universal across culture or not.

To avoid the confusion, there are two terms that we need to clarify. The first term is the *emotion*, and the second one is *expression*. The *emotion* term refers to the feeling of a person toward a particular event, object, or experience. These feelings can be observed by the *expression* of the physiological responses. They can be non-voluntary responses (flight-or-fight), or voluntary actions (social communication purposes). In this research, we interpreted a facial expression as a message of emotion that human can read.

1.5 Research trends and motivations

Most of facial expression recognition systems nowadays are benefited from the advancement of face recognition research including face detection and face tracking techniques. Therefore, these both face recognition and facial expression recognition system also shares their mutual problems such as mis-alignment errors (translation rotation scaling), head

Table 1.2: Conversion between Emotions and Action Units (AUs) in Facial Action Coding System (FACS) [9]. Notice that (*) means in this combination the AU may be at any level of intensity. The capital letter after AU number refers to the least required intensity of action (varying from A-very weak to E-very strong)

Emotion	Prototypes	Major Variants
Surprise	1+2+5B+26 1+2+5B+27	1+2+5B 1+2+26 1+2+27 5B+26 5B+27
Fear	1 +2+4+5*+20*+25, 26, or 27 1+2+4+5*+25, 26, or 27	1+2+4+5*+L or R20*+25, 26, or 27 1+2+4+5* 1+2+5Z, with or without 25, 26, 27 5*+20* with or without 25, 26, 27
Happy	6+12* 12C/D	
Sadness	1+4+11+15B with or without 54+64 1+4+15* with or without 54+64 6+15* with or without 54+64	1+4+11 with or without 54+64 1+4+15B with or without 54+64 1+4+15B+17 with or without 54+64 11+15B with or without 54+64 11+17
	25 or 26 may occur with all prototypes or major variants	
Disgust	9 9+16+15, 26 9+17 10* 10*+16+25, 26 10+17	
Anger	4+5*+7+10*+22+23+25,26 4+5*+7+l0*+23+25,26 4+5*+7+23+25, 26 4+5*+7+17+23 4+5*+7+17+24 4+5*+7+23 4+5*+7+24	Any of the prototypes without any one of the following AUs: 4, 5, 7, or 10.

pose variation, changing light conditions, occlusion, noise, blurriness. These problems dramatically change the appearances of face textures to the point that the system cannot fit them to the original training data. This degrades the system performance, or even fails the whole process. Some of them can be solved by performing normalization, or relying on the robust feature extraction methods. These issues will be discussed later in the content of this dissertation. Instead, in this section, we describe the issues we recognized as the significant trends of facial expression recognition in the computer vision field.

1.5.1 Natural facial expression

In the most recent survey of affective computing [35][36][37], they have a mutual sentiment that the analysis in computer vision research is mostly lack of natural response data. Most of them are posed by actors in strictly controlled environments. Therefore, the trend of the facial expression research has been recently focused on the spontaneous facial expression. The spontaneous facial expressions are what people usually express in our daily life. They are often composed of the various facial appearances and have weaker intensities than typical appearances in movies and other kinds of media. The spontaneous facial expressions have several characteristic that difference from the posed facial expression especially timing of the onset and offset [38]. The smiling case is often studied due to its basic usage in social communication, and often treated as the signature of happiness emotion. They found the different in elicited smile and social smile [39]. The spontaneous smile is found its differences in timing, velocity and amplitude comparing to the posed smile expression [40]. In addition, the spontaneous smile often has a sustained region with multiple peak [41].

1.5.2 Temporal feature analysis

Typically, the previous works assumed that the characteristic of facial expression can be captured from a still image in the same way that human can understand face from a photo. However, it has been suggested that human may understand the facial expression of emotion from the continuous sequences better than the still images [17]. Since facial expression is a transition of its appearance from one state to another, observing them in time series is inevitable as we mentioned about the significant of timing in the spontaneous smile cases.

Moreover, facial expression recognition system is usually trained from the peak of exaggerated facial expressions. In the case that an input face image has a weaker intensity of expression, the system cannot give a precise estimation of emotion class. In accordance to our preliminary study (Appendix E), the clustering of texture features of weak intensity expressions usually yields poor performances due to less structural differences between those subtle facial expression. Therefore, the sequential modeling techniques that involving with clustering methods such as Hidden Markov Model (HMM) and its family might not be suitable for measuring subtle face transition.

Another usual method for modeling the temporal character of facial expression is motion based method. Generally, motion vectors can be estimated from a neutral face to a designated face of a same person. However, it is difficult to initiate a proper neutral face template. In addition, the method frequently captures something else besides actual expression especially head motions. Even if a subject is stay still, it is apparent that

the strong motion vectors are appeared at the boundary of face due to face alignment errors. So, we suggest to model a spatiotemporal feature that can enhance facial motion actions. Furthermore, the feature should be compatible with the subtle changes of facial expression. This temporal issue will be discussed in the later chapter.

1.5.3 Emotion interpretation via facial action units

An appearance change at a particular area of face can be invoked by the remote muscle afar from the observed area as we discussed earlier. To describe a facial action in the same standard to psychological research, the muscle-based coding is proposed by Ekman and Friesen. They mapped changes on human face with the actual muscles activations by Facial Action Coding System (FACS)[23][2].[1]. The facial expression is coded its underlying facial muscle movements in term of Action Units (AUs, see Table. 1.1). Each AUs involves the changes of one or more facial muscles and can be detected independently. FACS is considerably famous for its objective judgment of the facial muscle activation and can describe much more variation of the facial expressions. Instead of using anatomy terms, FACS is marked by numbers as standard definitions, and is easier for those who are not familiar with the biological or psychological terms. However, it takes an enormous effort and experiences to learn and use FACS since there are so many details for describing each AU. Therefore, this FACS has gained much attention of recent facial expression recognition research due to its possibility to implementing in computer vision based system. By introducing AUs as the mid-level feature representation, it has been believed by much research that the system would be an objective tool comparing to direct interpretation from facial expressions to emotion classes.

1.5.4 Beyond basic emotion categorization

The recent applications of affective computing have found the crucial flaws of categorical model in real-time application. There are several conjectures on the definition of emotional classes, for instance, what types of emotion labels we should define, and how many of them are supposed to be. The simplest way is to define the classes as the prototypical categories such as positive or negative emotions. The more generalized, and most famous categorization is correspond to Ekman's basic emotion categorization along with its conversion from facial muscles activations [21][2]. Much facial expression recognition research in computer vision utilized this basic emotion category due to its simplicity. Although this categorization intuitively captures the key emotional states of human, but they are rarely seen in daily life as our facial expression are not confined by only these labels. Recent affective computing trend shifted the focus on basic emotion to complex and spontaneous emotional expression.

One can model non-basic emotions as a mixture of basic emotions. In this direction, researchers treat 6 basic emotions as the atomic units. Many times we can see an ambiguous facial expression of different basic emotions mixtures. For example, smile while showing the small sadness or frustration expression. To implement a mixture of emotion, we need to define how to represent such a model. Du et al. simplified the mixture of emotion by observing the co-occurrence of Action Units (AUs) that has been converted

[1]Regarding the validity of Facial Action Coding System (FACS), we also discussed about this issue in Appendix C.

from a particular emotion. For example, the mixture happily surprised (indicated by AU1, AU2, AU12 ,AU25) is a mixture of happiness (AU12, AU25) and surprise (AU1, AU2, AU25, AU26) plus their variant AUs [42]. As a result, they expanded 6 classes of basic emotions into 21 classes including the original six basic emotions and their mixtures. However, this approach is often questioned whether Ekman's basic emotion or other similar discrete models are the legit atomic units since each of emotional labels may have its own variations. For example, sadness can be further divided into crying and non-crying case, which result in different reactions in autonomous nervous system [24]. Furthermore, this also raises the question that how many emotions should be considered at the same time.

An alternative trend to model non-basic emotion is dimensional emotion model, where the emotional states are parameterized in terms of dimensions. Each dimension represents the atomic elements of emotion. The most famous one is two dimension model composed by *Valence-Arousal (V-A)* model. Valence (or Evaluation) dimension represents the how much positive or negative of our feels. Arousal (or Activation) measures the activeness/passiveness. Other dimensional models may expand this 2D model by including other parameters such as Power, Expectation, and Intensity. The dimensional model has advantages in the case that discrete categorization cannot explain since this approach cover wider range of emotions. Nevertheless, this approach can be viewed as the projection of our high dimensional complex emotion into lower dimensional space, and causing a loss in translation of these parameters to a specific emotion.

Beside the above emotions, there are attempts to interpret other mental states from facial expressions such as fatigue [43], frustration [44], pain [45], agreement, confusion, concentration [46], and interesting levels [47].

1.6 Datasets

In this research, we use three datasets JAFFE [3], MyFace, and CK+ [4]. The number of training and testing set is provided in Table. 1.3-1.5. For each dataset, we choose 70% of samples in database as the training set, and the rest of them as the testing set. The emotion classes are labeled according to the basic emotion model from Ekman [21][22]. All expression is deliberately posed with near frontal view, head movement is kept to minimum, and the backgrounds are plain. The face region is segmented by Haar-like feature-based cascade face detector [48], and use Kanade-Lucus-Tomasi feature tracker (KLT) [49][50] to stabilize the transition of facial expression across the temporal axis.

Table 1.3: The number of samples in JAFFE dataset [3]

Class	an	di	fe	ne	ha	sa	su	Total
Training	21	20	22	22	21	22	21	149
Testing	9	9	10	9	9	9	9	64

Japanese Female Facial Expression (JAFFE) dataset contains 213 still images posed by 10 Japanese females in 7 emotion classes: neutral, happiness, angry, fear, disgusting, sadness, and surprise. This dataset is used only for evaluating the classification methods on intensities features of the still images (Fig. 1.4).

Table 1.4: The number of samples in MyFace dataset

Class	an	di	fe	ne	ha	sa	su	Total
Training	7	7	7	7	7	7	7	49
Testing	3	3	3	3	3	3	3	21

Table 1.5: The number of samples in CK+ dataset [4]

Class	an	co	di	fe	ha	sa	su	Total
Training	31	13	41	17	48	20	58	228
Testing	14	5	18	8	21	8	25	99

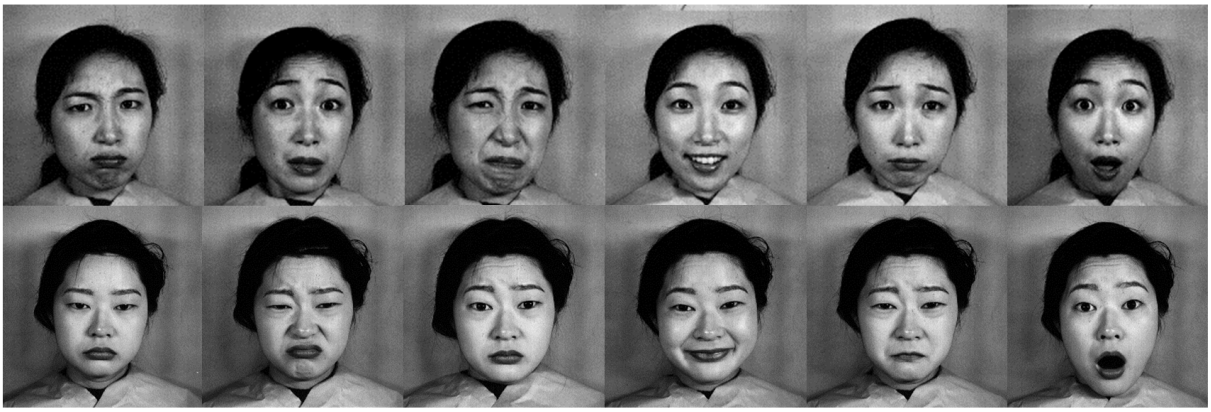


Figure 1.4: Example of facial expression images of two persons in JAFFE dataset [3] (first row and second row). The expressions from left to right are labeled as Angry, Disgust, fear, happiness, sadness, and surprise emotional states

MyFace dataset is the original dataset of this research (see Fig. 1.5). The dataset is aimed to evaluate both still and temporal feature. The emotion class is categorized into 7 classes in the same categorization as JAFFE dataset. Each emotion classes have 10 complete expression sequences with the different timing and speed of facial expressions. Duration of each sequences vary from 3 to 40 seconds including onset, apex, and offset of facial expressions. For the still image set, we choose one image out of each expression sequences with the highest emotion intensity. All image sequences in this dataset are recorded from only 1 subject by self-elicited method. Comparing to other datasets in this research, MyFace dataset has the weakest intensity expression i.e. subtle expression. Since every individual has a unique way to express the emotional messages, we are interested in developing a household robot that can interact with a particular owner, not with the average individual. In addition, we can determine only the discriminative of the feature without concerning the interpersonal issue.

Extended Cohn-Kanade (CK+) dataset. This dataset has the highest variation due to the highest number of subjects (123 subjects) and differences in nationality. The emotion labels are evaluated by the experts using Facial Action Coding System (FACS). The expression sequences are started from the neutral and stop at the peak of the expression. The dataset are labeled in different categorization by including the contempt class and excluding neutral class. We took the last frame of the sequences for evaluating the still

Figure 1.5: Example of image sequences in MyFace dataset. (First row) shows the samples from an angry expression, (Second row) shows the samples from a surprise expression, and (Third row) shows the samples from a happy expression

image feature, since all expressions are labeled from the last frame of the sequences.

1.7 Purpose of study

In this research, we estimate the state of human emotion by facial expression classification using temporal facial image features under the context of human-machine interaction. The principle of this research is founded by a combination of the multiple disciplines such as psychology, computer vision, and machine learning.

Although, facial expression and emotion are widely known concepts in psychology and cognitive science, the implementations of facial expression analysis system in the computer vision domain are still in an infant state. The state of the art research mainly focused on classification of the emotion expression from a still facial image. The usage of a still image in classification often mistakenly recognizes the input image features as another emotion expression, since the facial expression of any emotion is composed of various appearances. Furthermore, this assumption also contradicts to the emotion theorems in psychology domain, which are originally proposed by observing the changes of facial muscle over a period of time. In addition, the existing works also limited the number emotion categories into 6 or 7 classes, which cannot describe the complex and wider variation of facial expression in practice.

Thus, the goal of this research consists of

1. Improve the performance of facial expression classification system by analyzing the dynamic of facial expression from image sequences instead of using a still image.

Figure 1.6: Example of image sequences in CK+ dataset [4] of different subjects. (First row) shows the samples from an surprise expression, (Second row) shows the samples from a happy expression, and (Third row) shows the samples from a sadness expression

We model the representations of temporal facial image features, and improve the classification methods that suitable for such representations.

2. Expand the simple emotion model into more complex emotion model and create the new feature spaces/mathematical models by using the temporal variation of facial image features.

3. Combine and utilize the above task to construct the facial expression classification framework for interaction between human and intelligent machine.

1.8 Dissertation organization

The organization of this dissertation begins with the introduction in this chapter. Chapter 2 explains the overview of facial expression recognition systems in human-machine interaction. The computer vision methodology is composed of three main components: *preprocessing*, *feature extraction*, and *classification*. The related works, important issues, and our implementations of all of the above components are discussed. Chapter 3 introduces a novel robust temporal feature by the change of facial expression. The proposed method overcomes the problems in conventional features and can describe the actions of facial surface. In chapter 4, we further developed the action unit detectors by using a temporal feature from previous chapter. Instead of using a direct interpretation from face images to a set of specific emotion categories, facial expression can be described objectively via action units regardless to the prior assumption of emotions. The conclusion and

future works will be given at the end of dissertation. In addition, we also investigates a new aspect of complex emotion study in computer vision Appendix A. In order to prevent the misleading of research aspect in this dissertation, we separated some part of our works out of main content and place in the appendices.

1.9 Summary of publications

Publication I. proposed a new feature descriptor that represent the dynamic of facial expression and can distinguish the same action in different duration of time. In addition, we utilize both 2D and 3D depth sequences and explore the suitable primitive patterns for determining the dynamic of facial expression by classification methods. Experimental results show that the proposed feature extraction method can preserve the temporal information in a fix length and can be applied in conventional classification methods. We named this feature descriptor as Cumulative Change of Feature (CCF).

In *Publication II.*, we expanded our study by studying the state-of-the-art feature descriptor, Motion History Image (MHI). The realization of this paper found the differentiation of our proposed CCF feature which can be implied as the facial muscle activation levels, while MHI describes the order of an action. In this paper, we also determine the effectiveness of linear and non-linear classifiers on the standard datasets.

Publication III., is the sequel work aims to solve the problem we found in *Publication I. and II.* The quality of the proposed feature is suffered from various types of interference such as translation, scaling, noise, blurriness, and varying illumination. To cope with the problems, we derive a novel feature descriptor by expanding 2D Gabor features for time series data. This feature is named as Cumulative Differential Gabor feature (CDG). Then, we decompose the features onto discriminative subspace for estimating the emotion class. As a consequence, our method gains the advantages of the original Gabor feature, which utilizes both spatial and frequency components. The comprehensive evaluations of the proposed CDG show the potential and robustness to the underlying conditions. In addition, we also focus on the classification by subspace methods especially we are interested in Independent Component Analysis (ICA) concept. As the original ICA was not designed for the class separation problem, we adopt the idea in [7] and its prequel work.

Publication IV. determined the categorization of facial expression of a person by using topic model technique. Conventionally, the learning of facial expressions is supervised by a well-labeled dataset. The trained features are separated into a limited number of classes. However, there are much wider range of facial expression that human can produce or perceive in practice. By this limitation, we questioned about the possible number of facial expressions expressed. In contrast to the *Publication I.-III.*, the resultant classes (topics) are defined by determining the occurrences of features, which are represented in the bag-of-words form. This study pointed out a significant different between typical face clustering (which tends to categorize the person identity), and facial expression clustering (which contains less variation of the feature distribution). It also showed us the drawback of using texture based feature from a still image to analyze subtle facial expressions.

Publication V. is an extension of *Publication III.*, we introduced a new implementation by a superposition of all CDG responses in each frequencies and orientations. The new implementation boosted up the performance of features in inter-personal case significantly.

Furthermore, we also investigate the robustness of the CDG under different conditions.

Publication VI. proposed a novel action unit (AU) detector following the Ekman's Facial Action Coding System (FACS). Our AU detection system utilized the robust temporal features CDG proposed in *Publication V.* and a new architecture of classification methods based on discriminative ICA to detect subtle changes in facial expressions. Therefore we can objectively describe the subtle and complex facial expressions in the same standard in psychology studies. The experimental results show the higher performance of our proposed system comparing to our previous classification methods in the standard dataset.

Publication VII. used our robust AU detector in *Publication VI.* which can capture the subtle actions and overcome several outliers, and then we used SVR to model the predictor of dimensional emotion parameters. The preliminary experiment shows the prominent potential of the proposed method for estimating dimensional emotion parameters.

Chapter 2

Theoretical framework

Suppose, we have two agents; human and machine. The input data can be collected in form of facial activities, bodily gestures, speech, or bio-potential signal. In this research, only visual information is considered, especially facial expression of emotion. The overview of facial expression analysis system in human-machine interaction is illustrated in Fig. 2.1. Our primary task in this dissertation is bounded in image processing module. There is three important steps; data preprocessing, feature extraction, and classification. Then, the machine will response back to human based on estimated emotion parameters.

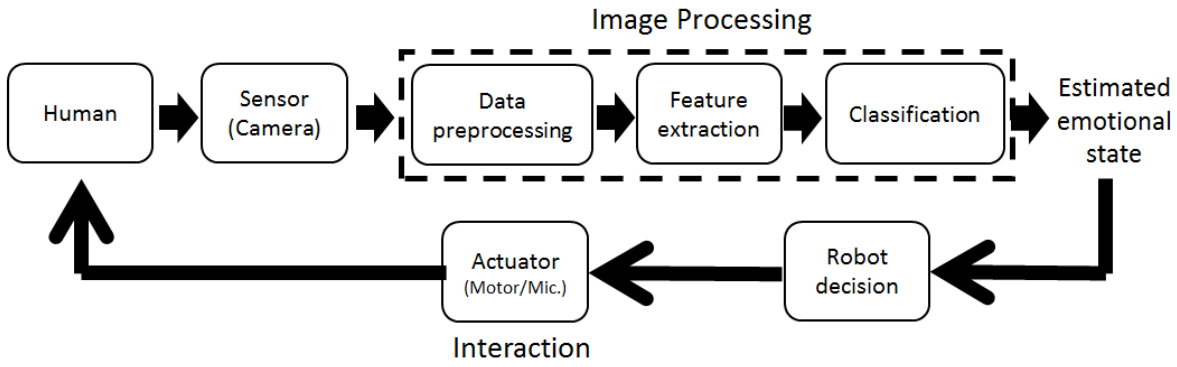


Figure 2.1: Overview of facial expression analysis system in human-machine interaction by using visual information

2.1 Data preprocessing

Data preprocessing is a compulsory step of the facial expression recognition system in prior to other steps. The quality of preprocessed output can affect the performance of overall system. The objective of data preprocessing is to eliminate the outliers such as head pose, position, size, illumination changes. As shown in Fig. 2.2, there are two main steps: face localization and face normalization.

2.1.1 Face localization

Face localization can be archived by *face detection* and *face tracking* methods. Face detection locates face position and area in an input image. Normally, the detected positions

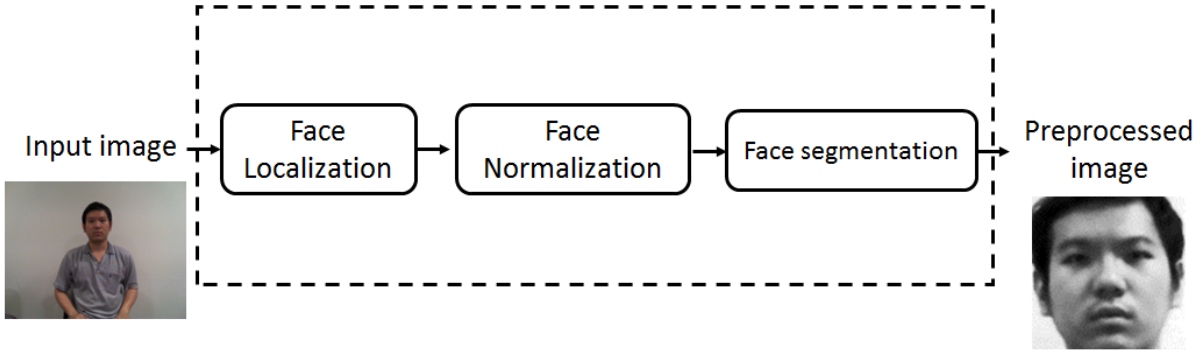


Figure 2.2: Data preprocessing procedure

and areas often present alignment errors. Although these errors can be ignored in the face detection of a still image application, but in the temporal analysis, the presence of alignment errors is crucial. The rapid change of size and position of face cause a glitch in temporal analysis. To smoothing the transition of facial expressions over an image sequence, face tracking is utilized to track the face positions or landmark points across the image sequence, and stabilize their trajectories by a smoothing filter i.e. face tracking is a continuous face detection. However, there is a tendency that the tracked positions may gradually drift away from the expected position in the long run. So, the tracking schema is not applied in some studies.

The vast amount of the recent face detection methods are based on Viola and Jones's object detector algorithm [48]. Their method basically use the Haar-like feature created by integral images, and then train cascade of classifiers with Adaboost technique. This method outperforms the previous works and becomes the standard face detector in academic and commercial usages due to its fast computation, good detection rate with toleration toward scale and translation, and the implementation tools are widely available.

After we can detect the face area, individual tracking of feature points can be archived by particle filter [51], or Kanade-Lucus-Tomasi feature tracker (KLT) [49][50]. These methods yield very stable transitions of facial expressions in frontal pose. Mach research integrated face detection and face tracking into the same process. The most well-known methods are Active Shape Models (ASM) [52], Active Appearance Models (AAM) [53] and its variant such as Constrained Local Model (CLM) [54]. These methods define the structure of face by shape and appearance parameters particularly proposed for 2D face tracking. The 3D based methods were also proposed in Candide model [55], Piecewise Bezier volume deformation (PBVD) [56][57]. Even though these methods can be considered as the face detection methods, but the searching procedure is suffered by the expensive computation cost and often stuck to the local optimum. So, the initialization of searching by face detection method is highly recommended to limit the search area.

Face localization is still a challenge topic in computer vision since the changes of face appearances are non-rigid transformation. It is difficult to define a generic representation of face that can fit the face in any environment. Most of the current face localization methods aim to detect face in severe conditions such as crowded environment where multiple faces can be found with different individual head pose variations, occlusion, and underexposed lighting conditions. Although recent state of the art methods claimed their successful results in the wild environment dataset [58][59][60], but the presences of mis-

alignment in their approximations are still need to be refined to apply on a complicated tasks such as facial expression recognition.

The uniqueness of face comparing with other typical moving objects is the non-rigidity property. Face feature can be varied its region of interest (ROI), shape of face, composition of face components. The changes can be made by many factors. The most influent one is the translation and rotation of head. It can be considered as a moving object, which dramatically change the image features as a whole. Beside that, facial expression, which is a product of muscles' contraction and relaxation, changes the texture of facial image. These actions produce not just the expression inside face region but sometimes change the shape of face silhouette, especially the movement of lower part of the face (e.g. jaw). Therefore, the detection and tracking methods need to handle the these appearance changes. In addition, to analyze face in temporal, we also need a high consistency of facial features across time since we intend to observe the transition of facial expression.

However, comparing with higher degree of non-rigidity cases such as tracking of flying birds, or in the extreme one such as cell division in bacteria, the degree of the non-rigidity of face is much less and mostly limited by the human anatomy. There are several facial components which remain their form and positions regardless to the facial expressions such as the positions of eye or nose. The structural composition of face is stationary for example nose is a vertical component between eyes and lie above mouth. By these known assumption we can define constrains in face detection and tracking.

Figure 2.3: The positions of landmark points detected by Asthana et al. method [5].

In this research, we employed face detection only in first frame and face tracking in the later frames of an image sequence. The initial position is found by the Viola and Jones method as a coarse estimation, then we used Asthana et al. method [5] to calibrate the fine positions of stationary face components such as eyes, and nose (landmark points numbers 23, 26, and 17 respectively in Fig. 2.3). For the following frames in each sequence, we apply the KLT feature tracking schema instead of applying face detection for each frame.

By tracking schema, we can maximize the smoothness of facial action transitions along the image sequences. However, this method can handle only in-plane rotation. So the out-of-plane rotations should be avoided in order to prevent a drifting problem.

2.1.2 Face normalization

Frontalization

Facial expression recognition systems usually have a mutual flaw when they encountered the non-frontal pose issue. Typical problem in facial expression analysis is the assumption of representing the expression in the frontal view. We refer this problem as the head pose problem. Based on the head pose parameters we solved in the previous localization step, frontal face can be recovered by rotation transformation or affine transform in case of in-plane rotation (roll direction, see Fig. 2.4). However, changing head pose in out-of-plane rotation (pitch and yaw direction) always produces a damage to face feature due to losing of a part of facial information, and respectively is recognized as a wrong emotion class.

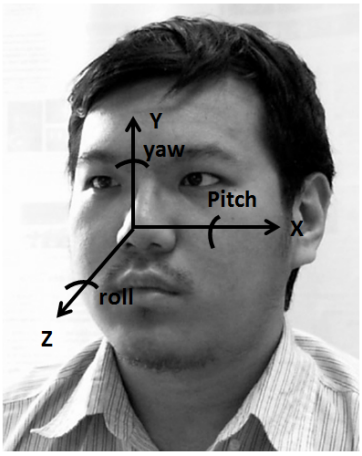


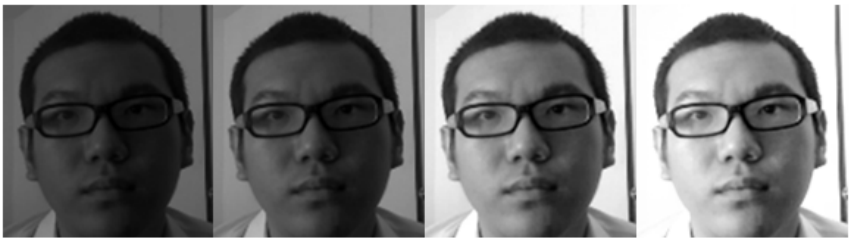
Figure 2.4: Head rotation and denoted axes. In-plane rotation refers to the rotation around z axis (roll). Out-of-plane rotation refers to the rotation around x axis (pitch) and y axis (yaw).

The procedure to recover the frontal face is called as Frontalization. Frontalization in facial expression recognition considers different elements comparing to the one in face recognition system. Typical face recognition system uses a non-rigid transformation to neutralize outliers that unrelated to person identification such as head poses, face shapes, and changing facial expressions. In the opposite direction, Frontalization in facial expression study requires a rigid transformation to align an input face with stationary points such as eyes and nose, the rest of face variance is considered as the changing facial expressions.

Frontalization process is a challenging topic, which is often stated in many recent studies, and need a lot of attention on the problem. In the state of the art frontalization methods, in DeepFace [59] and Hassner et al. method [61], they approximated the frontal face by projecting 2D feature points on the 3D models and estimate the transformation parameters, Then, they reconstructed the texture image back to frontal image coordinates and approximated the pixel values in the missing area by interpolation and symmetric

mirroring. Although it seems that these methods can frontalize the face, but our attempt to replicate Hassner et al. method [61] indicated that the facial expression information is often distorted during the interpolation process, and the method do not guarantee the same frontalize output even in the same pose from the same person. Therefore, these frontalization methods are still not mature enough to measure the transition of facial expressions. Therefore, in this research we do not focus much on the out-of-plane rotation issue, and consider only the frontalization for in-plane rotation by applying a two dimensional affine transform derived from tracked features as we mentioned above.

(a) Local changes



(b) Global changes

Figure 2.5: Illumination changes in form of (a) locally change from different position of light source, (b) globally change in accordance to various brightness levels

Illumination changes

Another problem is an illumination variation. Conventional image textures including intensity and motion based features are sensitive to the illumination changes of face. To perform a pattern matching between textures, globally or locally changes of lighting conditions dramatically change the structure of patterns and cause a large dissimilarity errors in matching process. Generally, global illumination changes issue can be handled by histogram equalization or gamma correction [62], which are suitable for global illumination changes. In many cases, the illumination change problem is considered in the robust feature extraction instead of normalize them in preprocessing step.

Notice that histogram equalization may produce the artifacts and overexposed condition from its incomplete quantization. In this research, we deal with the illumination problem by histogram equalization in the preprocessing step only for improving the efficiency of face localization, and the problem would be engaged later in feature extraction step by the robust feature descriptor.

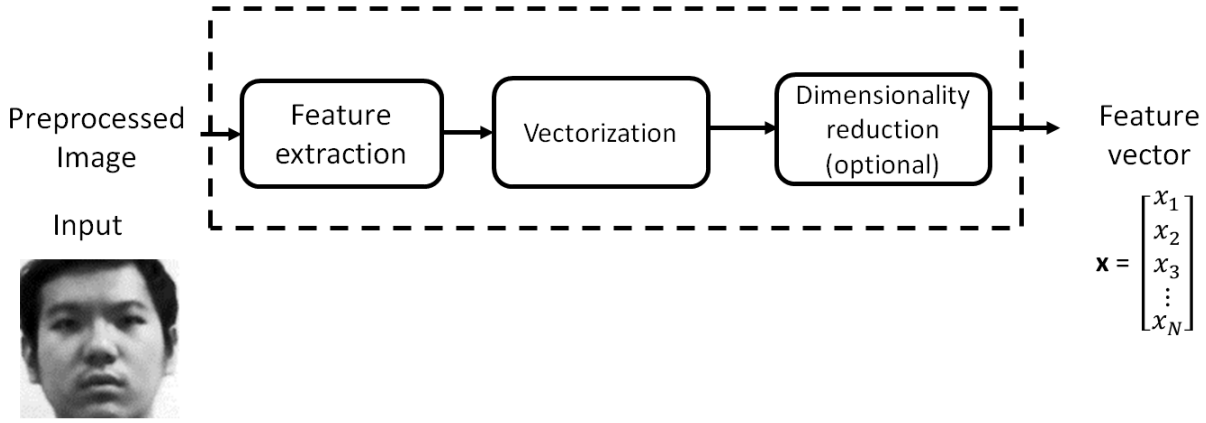


Figure 2.6: Feature extraction procedure

2.2 Feature extraction

2.2.1 Feature modeling

To determine how our face appearance changes when we display the particular facial expression of emotions, most of the existing facial expression recognition system uses a spatial feature of a still image. Typically, we can categorize them into two types of feature: *Geometric feature*, and *texture feature*.

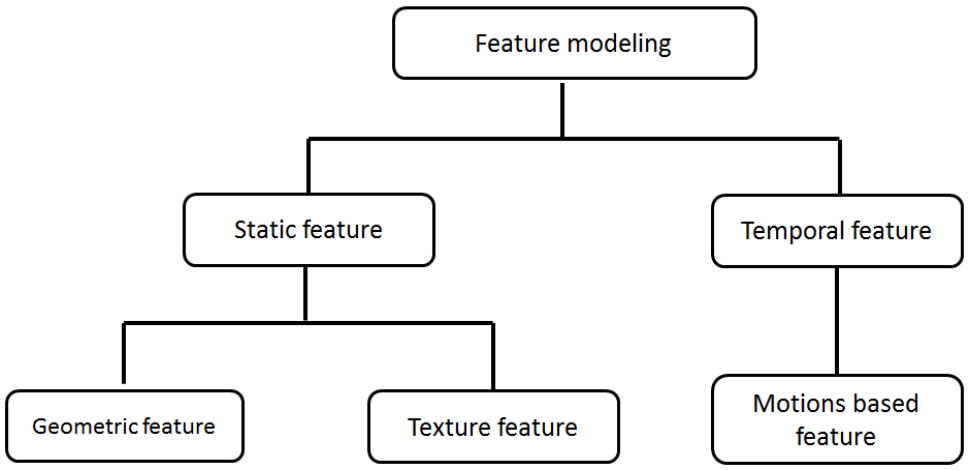


Figure 2.7: Feature modeling procedure

Geometric feature represents the changes of faces in form of shape parameters, landmark points of facial components (eyes, eye brows, mouth, noise), or their distances between these components. Normally, geometric features are by-products of the landmarking based face detection and face tracking processes such as Active Appearance Models (AAM) [63][64][65]. Tian et al. [66] designed independent feature descriptors for underlying facial components separately (lip, eye, brow, cheek furrow). Similar idea was applied by Tsalakanidou and Malassiotis [67] with the combination with texture pattern to detect wrinkles. Kotsia and Pitas [68] used the deformation of selected Candide [55] nodes to estimate emotions. Recent works were interested in the Facial Animation

Parameters (FAPs) [69][70], which are defined in the MPEG-4 standard to control the actions of synthetic facial animation.

Texture feature or appearance based feature makes use of a whole facial texture, or local areas pattern as the inputs for classification. As we can find the spatial relation of each elements, texture feature provide richer information than geometric features. However, texture features are highly sensitive to the errors caused by localization of faces, and cannot handle a large appearance changes. Therefore, it is required to normalize the input image before extracting the features. Alternatively, instead of using the raw intensity texture, many previous works adopt the robust feature descriptors in their implementation such as Local Binary Pattern (LBP) [71][72], Scale Invariant Feature Transform (SIFT) [73][72], Gabor feature [74][75][76][77], Histogram of Oriented Gradient [72], Haar-like feature [78]. Generally, the dimensions or sizes of feature are exponentially larger than geometric feature. Turk and Pentland [79] introduced the Principle Component Analysis (PCA) to project the face data on lower dimensional subspace known as Eigenface. As a result, these subspace idea greatly reduces the dimension of texture feature to only a few significant components.

Geometric feature has been expected to provide a better tolerance toward pose and localization errors. Since information is greatly reduced from a raw image input into a few dimensional feature descriptors, it also loss several appearance changes such as wrinkles, while the texture feature can maintain these information. In subtle expression or weak intensity expression, the distance gaps between each landmark points become smaller, and unable to distinguish facial expressions. Similar issue has been found in texture feature. Although texture feature may provide more information, the problem here is the texture of some facial component such as the wrinkles and lines many not appears clearly in subtle expression comparing to the exaggerated one. These texture details are also different among various ages and genders. Therefore, the texture alone may not suitable in model the facial expression, and we suggest using a motion based feature to observe these actions instead.

In order to model the temporal facial expression, the temporal modeling is applied as the one of these two levels: *feature level* and *classification level*. Basically, classification level modeling is wrapped around the concept of Hidden Markov Model (HMM) and its similar graphical models. The input observations in the previous methods were a texture feature from a single frame [76], or a motion based feature between frames [80][47][81][56], where the vector quantization (clustering) performances of these features are severely limited in subtle expressions due to smaller dissimilarity among data. Therefore, we would suggest modeling the temporal entity on the feature level instead.

The temporal concept in feature level focused on motion based features (e.g. optical flow methods). One of the most widely known dense optical flow method is Lucus-Kanade algorithm (LK) [49][82]. The algorithm solves the least square problem over the image patches. There are two assumptions in LK method; brightness constancy and spatial coherence. The brightness constancy is the same as the original optical flow problem which require the strict control of global and local lighting conditions between consecutive frames, and it should remain the same even if their location is changed. The spatial coherence assumption states that the neighboring points have same motion as the given points. Thus, the LK method practically calculates the motion vector at position p_i by using the neighboring patch. Farneback algorithm similarly uses the quadratic polynomial expansion to approximate the neighborhood pixels [83]. Thus, the dense optical flow

provided by the algorithm is more robust to the noise. In the previous works related to facial expression recognition, Mase [84] introduced the optical flows of local areas between neutral and a peak expression image to model the muscle actions. Black and Yacoob applied the similar local optical flows method in video sequences and converting the motions to their mid-level representations [85]. Donato et al. [86] demonstrated the feasibility to model the facial expression dynamics by dense corresponding (full registration) optical flow between two images. However, the conventional optical flows methods are sensitive to the head motions and illumination changes and not suitable to detect a subtle expression. These problems inspired us to develop a robust temporal template in the next chapter.

2.2.2 Vectorization

The vectorization is the conversion process from the $I \times J$ matrix (number of rows and columns) into a D-dimensions vector. This process is required for converting the 2D data into 1D vector before feeding the feature into the classification module. Suppose C be the temporal template feature from section 3.1 or 3.2, we can write as:

$$C = \begin{bmatrix} C(1,1) & \cdots & C(J,1) \\ \vdots & \ddots & \vdots \\ C(1,I) & \cdots & C(J,I) \end{bmatrix}_{I \times J} \tag{2.1}$$

where I and J is the number of row and column of CDGs. The matrix is transformed into D-dimensions column vector as:

$$C \longrightarrow C' = \begin{bmatrix} C(1,1) \\ \vdots \\ C(J,I) \end{bmatrix}_{IJ \times 1} \tag{2.2}$$

Thus, we obtain the feature vector of size $D = (I \times J)$ dimensions. This feature vector x is applied to the classification module later.

2.2.3 Dimensionality Reduction

Dimensionality reduction is the process of deriving the original features into lower dimension subspace. The reduced features are introduced by taking salient components of the original features. The term *dimension* refers to the attributes of feature that can indicate the characteristics of data. If the feature is image intensities, each dimension indicates the spatial locations and the total number of dimension is equal to the number of pixels in the image. Although providing more feature dimension can describe more elements of the data and suppose to increase the performance of learning, but some dimensions are less important than other dimensions. They should be eliminated to reduce the classification error and speed up the computation. The comparison of computation time is shown in Table. 2.1. The training times and recognition times is greatly reduced. This example showed that by using dimensionality reduction we can apply subspace classification method in real-time application.

Table 2.1: Comparison of computation time between a standalone classification (EMC) and a classification method with dimensionality reduction (PCA then EMC). The test samples is run on JAFFE dataset using the intensity features (10,000 dimensions). To reduce feature's dimension, we projected the features on 100 largest eigenvectors from PCA.

Method	EMC	PCA then EMC
Training time (s)	200	0.042
Recognition time (s)	3.6	0.0034

Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is one of the most famous dimensionality reduction techniques by projecting the data onto a lower dimension subspace, such that the variance of projected data is maximized. PCA is widely used in face recognition application known as Eigenface [79].

Given total M features, the feature vector x_m is a column vector with D dimensions. m is the feature index. The mean of all feature vectors $\bar{x}$ and correspondence covariance matrix S are given:

$$\bar{x} = \frac{1}{M}\sum_{m=1}^{M} x_m \tag{2.3}$$

$$S = \frac{1}{M}\sum_{m=1}^{M}(x_m - \bar{x})(x_m - \bar{x})^T \tag{2.4}$$

Then, we define the vector ϕ as the subspace of data. This vector is equal to the eigenvectors in the following problem:

$$S\phi = \lambda\phi \tag{2.5}$$

After solving eigenvalues problem, we obtain the set of eigenvectors ϕ and corresponding eigenvalues λ as:

$$\phi = \{\phi_1, \phi_2, \phi_3, ..., \phi_M\} \tag{2.6}$$
$$\lambda = \{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_M\} \tag{2.7}$$

From the above Eqs. (2.6)-(2.7), we can choose some of important components that represent the original feature vector set in lower dimension by determine the K biggest eigenvalues and taking corresponding K eigenvectors. Thus, the chosen eigenvectors become $\phi = \{\phi_1, \phi_2, \phi_3, ..., \phi_K\}$, where ϕ_m has D elements. By projecting the original data onto the chosen eigenvectors, the newly feature vector z_m of m^{th} feature index can be calculated as:

$$z_m = \begin{bmatrix} \phi_1^T(x_m - \bar{x}) \\ \phi_2^T(x_m - \bar{x}) \\ \phi_3^T(x_m - \bar{x}) \\ ... \\ \phi_K^T(x_m - \bar{x}) \end{bmatrix} \tag{2.8}$$

Thus, we can write a set of M features in new dimension K by

$$z = \{z_1, z_2, z_3, ..., z_M\}. \tag{2.9}$$

However, the calculation of the above principle component analysis is inefficient due to the enormous size of covariance matrix S which produce $D \times D$ elements. This would lead to the memory shortage and rise the calculation complexity of eigenvalues problem become $O(D^3)$. We can resolve this problem by an alternative deviation of the eigenvalues problems as following steps [18]. Firstly, we define the set of centered feature vectors $x_m - \bar{x}$ in the matrix form X:

$$X = \left[(x_m - \bar{x})\ (x_m - \bar{x})\ (x_m - \bar{x})\ ...\ (x_m - \bar{x})\ \right] \tag{2.10}$$

where X is $D \times M$ matrix.

The covariance matrix can be written as below equation:

$$S = \frac{1}{M} X^T X \tag{2.11}$$

Notice that the size of this covariance matrix is reduced to $M \times M$ instead of $D \times D$. By solving eigenvalues problem by this covariance matrix, the M eigenvectors with the size of M elements are produced. In the similar fashion to the original PCA method, the K eigenvectors are picked according the K largest eigenvalues.

In order to project the original feature vector x which has length of D onto the eigenvectors, the eigenvectors must be scaled from M to D by following equations:

$$\phi_k = \frac{1}{\sqrt{M\lambda_k}} X^T v_k \tag{2.12}$$

where v_k is the eigenvector with M dimensions, ϕ_k is the eigenvector with D dimensions and k is the index of eigenvectors. By projecting the original feature vector x_m to ϕ_k as in Eq. (2.8), we can obtains the same principle subspace as in the original PCA problem.

Note that the projection of one vector onto another is equivalent to the dot product in linear algebra. For instance, consider dot product in two dimensions, let $\vec{a} = \{a_1\vec{i} + a_2\vec{j}\}$ and $\vec{b} = \{b_1\vec{i} + b_2\vec{j}\}$, the dot product $\vec{a} \cdot \vec{b} = a_1 b_1 \vec{i} + a_2 b_2 \vec{j}$.

2.3 Classification

In order to estimate an emotion class of the extracted feature, we applied a subspace classification method. Basically, subspace methods aim to represent high dimensional data on the basis vectors. The formulation of basis vectors and their characteristics are varied due to the applications, for instance, principle component analysis (PCA) is utilized for dimensionality reduction and identity recognition [79]. For classification application, linear discriminant analysis (LDA)[18] has been extensively used. In this research, we particularly interest in the independent component analysis (ICA), since it can split the original data into uniquely sub-signals which statistically independent to each other [87][19][88]. Similar to the framework in [89], they introduced the combination of Gabor feature and ICA subspace for face recognition. However, the original ICA was

not designed for the class separation problem. Therefore, we adopt the idea of class separation by ICA in [7] and its prequel work.

Generally, the problem in the discriminant analysis estimates the matrix ϕ in a way that the significant components are exposed from the input signal or features. The matrix is known as *transformation matrix*, *eigenspace*, *principle component*, or *independent component* according to the techniques. In the document, we use the term *subspace* to avoid the further confusion. The general form of the projection is denoted by:

$$z = \phi x \tag{2.13}$$

Given, the input feature vector with D-dimensions $x = (x_1, x_2, x_3, ..., x_D)^T$. The feature vector is projected onto the matrix ϕ, which results as the output feature $z = (z_1, z_2, z_3, ..., z_N)^T$ with N-dimensions, where $N \leq D$. The subspace matrix could be considered as the new basis of the data distribution. In this research, we utilize Eigenspace Method based on Class feature (EMC), Kernel Eigenspace Method based on Class feature (KEMC), and Independent Component Analysis (ICA). The variant of ICA by using EMC as preprocessing step is considered in this research. The classification framework is illustrated in Fig. 2.8.

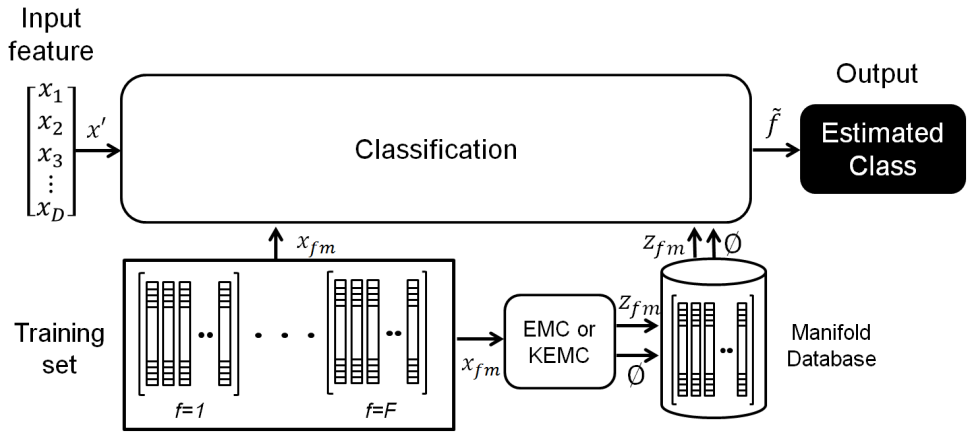


Figure 2.8: The classification framework

Training manifold vectors

The centered image $x_{fm} - \bar{x}$ is projected onto the eigenvectors and this creates the new manifold vector $z_f = \{z_{1f}, z_{2f}, z_{3f}, ..., z_{M_f f}\}$ of class $f \in F$. For each image m in class f, the manifold vector can be written by:

$$z_{fm} = \phi^T (x_{fm} - \bar{x}) \tag{2.14}$$

The manifold vector z_f can represent the salient components according to the class f with more robustness than original feature vector form.

Estimating an output class

In order to recognize the facial expression class of an unknown sample $\hat{x}$, the estimated class $\tilde{f}$ can be obtained by projecting the centered sample $\hat{z} = \hat{x} - \bar{x}$ onto the eigenvectors

of training sets. Then, we compute the corresponding Euclidean distance between the manifold vectors of each class and the projected input data as shown in Fig. 2.9

$$\tilde{f} = \underset{f \in F, m \in M_f}{argmin} \sqrt{\sum_{d}^{D} (z_{dfm} - \hat{z}_d)^2} \tag{2.15}$$

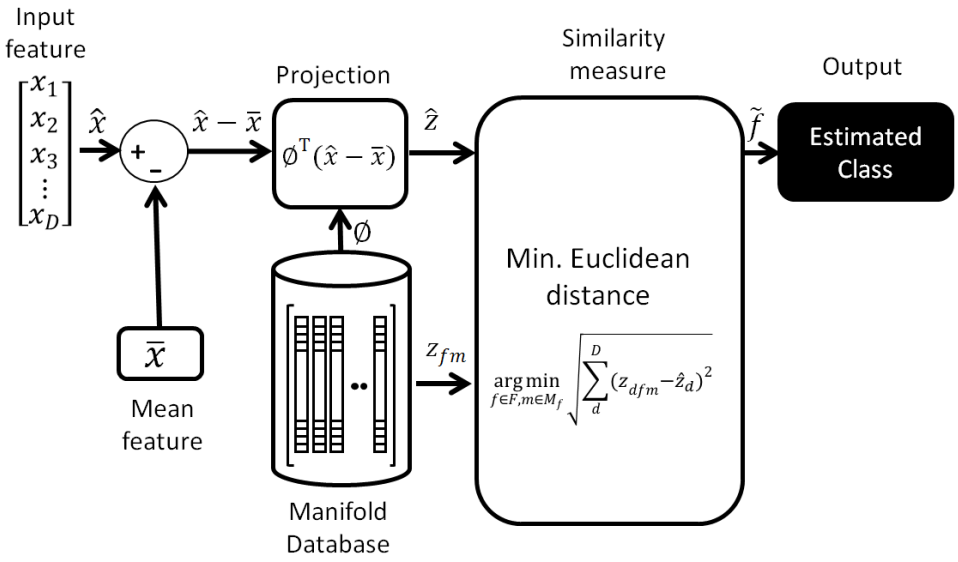


Figure 2.9: Recognition procedure by subspace method

2.3.1 Eigenspace Method based on Class feature (EMC)

Eigenspace Method based on Class features (EMC) is a linear classifier by finding the new subspace or linear combinations of features which can separates the classes of features. The method is proposed by Kurozumi et al. [6]. The training procedure is very similar to the Multiple Discriminant Analysis (MDA) (Fig. 2.10). However, the class separation problem is solved by maximizing the difference of between-class variance S_B and within-class variance S_W instead of maximizing their ratio. The comparison in [90] indicates that EMC subspace is more robust to the over-fitting problem than MDA subspace. Basically, the method establishes a new subspace by solving the eigenvalues problem by:

$$S_{emc}\phi = \lambda\phi \tag{2.16}$$

The covariance matrix can be defined as

$$S_{emc} = S_B - S_W \tag{2.17}$$

$$S_B = \frac{1}{M} \sum_{f=1}^{F} M_f (\bar{x}_f - \bar{x})(\bar{x}_f - \bar{x})^T \tag{2.18}$$

$$S_W = \frac{1}{M} \sum_{f=1}^{F} \sum_{m=1}^{M_f} (x_{fm} - \bar{x}_f)(x_{fm} - \bar{x}_f)^T \tag{2.19}$$

$$\bar{x} = \frac{1}{M} \sum_{f \in F} \sum_{m=1}^{M_f} x_{fm} \tag{2.20}$$

$$\bar{x}_f = \frac{1}{M_f} \sum_{m=1}^{M_f} x_{fm} \tag{2.21}$$

where f is the class index, F is the total class number, M_f is the number of feature vectors in the class f, M is the number of feature vectors or images which equal to $\sum_{f=1}^{F} M_f$, $\bar{x}$ is mean of all feature vectors, and $\bar{x}_f$ is mean of feature vectors in the class f, x_{fm} is a feature vector that belong to the class f.

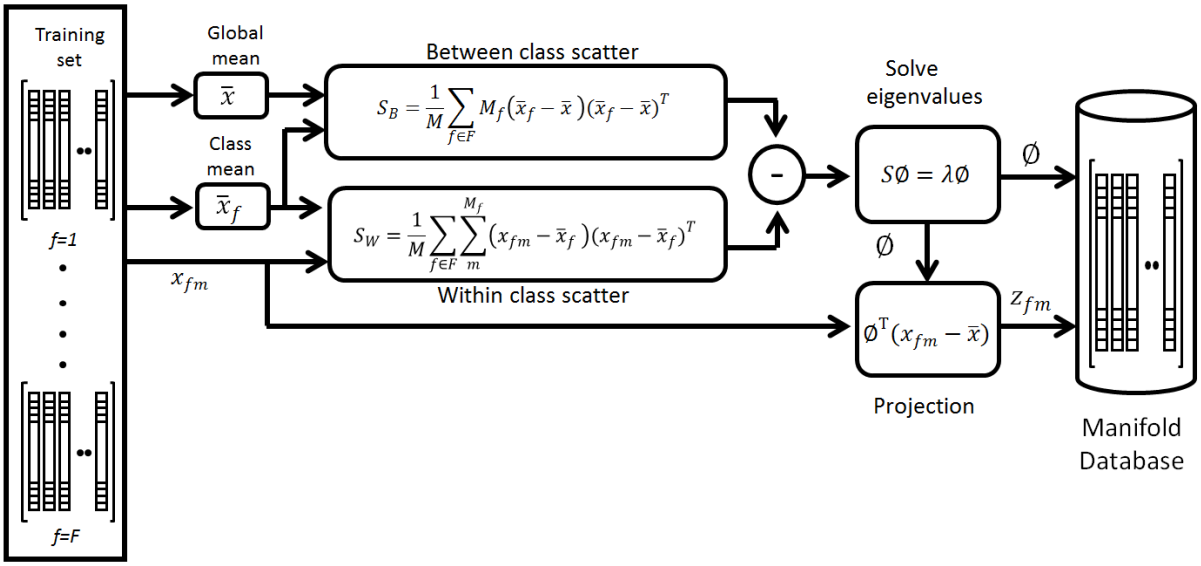


Figure 2.10: Subspace training procedure by EMC method

2.3.2 Kernel Eigenspace Method based on Class feature (KEMC)

The separations of the facial expression classes are usually non-linear; therefore the EMC which derived from the linear equation cannot create the best subspace for representing such a non-linear discrimination. In [20], Kosaka and Kotani proposed the Kernel Eigenspace Method based on Class feature (KEMC), which introduces the non-linear transformation by using the polynomial kernel function in Eq. (2.22) or radial basis function kernel (rbf) in Eq. (2.23)

$$k(x_1, x_2) = (x_1^T x_2)^p \tag{2.22}$$

$$k(x_1, x_2) = exp\left(\frac{||x_1 - x_2||^2}{2\sigma^2}\right) \tag{2.23}$$

where p is the polynomial order of the kernel function. To create the manifold vector z_{fm} We solve the eigenvalues problem by:

$$R\phi = \lambda\phi \tag{2.24}$$

$$R = \frac{1}{M}\sum_{f\in F} M_f m_f m_f^T - \frac{1}{M}\sum_{f\in F}\sum_{m=1}^{M_f}(\zeta - m_f)(\zeta - m_f)^T \tag{2.25}$$

$$m_f = \begin{bmatrix} \frac{1}{M_f}\sum_{m=1}^{M_f} k(x_{11}, x_m) \\ \frac{1}{M_f}\sum_{m=1}^{M_f} k(x_{12}, x_m) \\ \cdots \\ \frac{1}{M_f}\sum_{m=1}^{M_f} k(x_{21}, x_m) \\ \cdots \\ \frac{1}{M_f}\sum_{m=1}^{M_f} k(x_{fm}, x_m) \end{bmatrix} \tag{2.26}$$

$$\zeta_{fm} = [k(x_{11}, x_m), k(x_{12}, x_m), ..., k(x_{fm}, x_m)]^T \tag{2.27}$$

where ζ is the mapped feature on the kernel space. The manifold vectors z_{dfm} can be formulated by

$$z_{dfm} = \sum_{f\in F}\sum_{m=1}^{M_f} \phi_{dfm}^T k(x_{fm}, x) \tag{2.28}$$

To recognize the facial expression, let $\hat{x}$ be an input feature with D-dimensions. We project the input on the eigenvectors,

$$\hat{z} = \begin{bmatrix} \sum_{f\in F}\sum_{m=1}^{M_f} \phi_{1fm}^T k(x_{fm}, x') \\ \sum_{f\in F}\sum_{m=1}^{M_f} \phi_{2fm}^T k(x_{fm}, x') \\ \cdots \\ \sum_{f\in F}\sum_{m=1}^{M_f} \phi_{Dfm}^T k(x_{fm}, x') \end{bmatrix} \tag{2.29}$$

After obtain the projected feature, we estimate the class $\tilde{f}$ by measuring the similarity between projected input and the manifold vectors using Euclidean distance in similar fashion to the Eq. (2.15).

2.3.3 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) states that an arbitrary image feature can be decomposed into a linear combination of sub-components under the criteria that each of them are statistically independent to the other [19]. The procedure of ICA is illustrated in the Fig. 2.11.

Whitening process by PCA

In order to separate the independent components with the minimum mutual information, the whitening process is required as the preprocessing to decorrelate the input features.

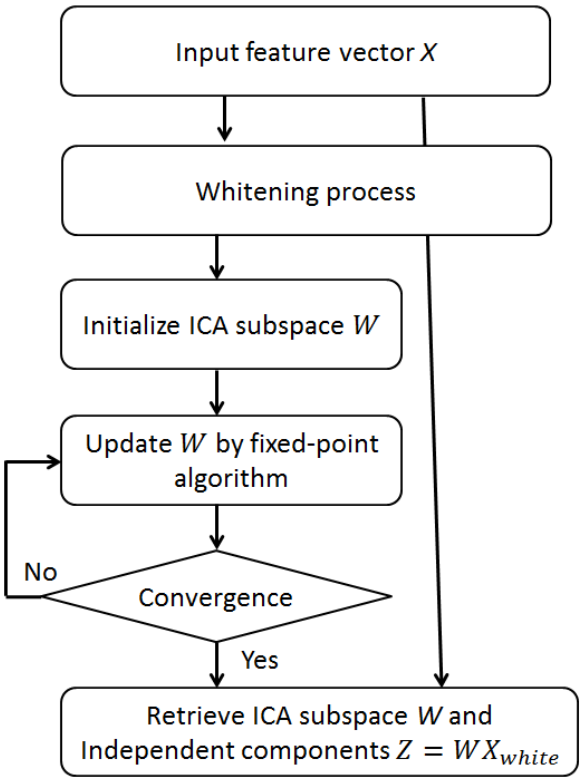


Figure 2.11: The flow chart of ICA method

The process normalizes the eigenvectors by the corresponding eigenvalues, which can be derived by the principle component analysis (see section 2.2.3). Therefore, the data distribution is normalized to unit variances.

Suppose ϕ is the collection of eigenvectors, $\phi = (\phi_1, \phi_2, ..., \phi_E)^T$ and $\lambda = (\lambda_1, \lambda_2, ..., \lambda_E)$ is the corresponding eigenvalues, E is the number of chosen eigenvectors such that $E \leq D$. The whitening process is the projection of an original data onto the whitened subspace:

$$x_{white,e} = \left(\frac{\phi_{pca,e}}{\sqrt{\lambda_{pca,e}}} \right)^T (x - \bar{x}) \tag{2.30}$$

where $x_{white,e}$ is the projected data from the e^{th} unit basis vector.

Whitening process by EMC

Despite of the standard whitening process by PCA, the variant of the ICA in [7] utilized EMC in section 2.3.1 as the whitening step. The advantage of the EMC subspace is the superiority in class separation over the PCA. We can rewrite the whitening process in Eq. (2.30) as:

$$x_{white,e} = \left(\frac{\phi_{emc,e}}{\sqrt{\lambda_{emc,e}}} \right)^T (x - \bar{x}) \tag{2.31}$$

Fixed-point algorithm

In this research, the independent component is defined as the minimization of mutual information. We adopt the fixed-point algorithm from Hyvarinen [88]. Suppose the number of independent component is Q. The linear transformation from original data to the independent component can be written as:

$$z_q = w_q^T x_{white} \tag{2.32}$$

where w_q is a vector that transform whitened data x_{white} to independent components z_q, q is the index of independent component.

The transformation vector w_q can be calculated by repeating the Eq. (2.33)-(2.36) until convergence. The convergence can be checked from the dot-product between $w_{old,q}$ and $w_{new,q}$ whether it is close to one. The update rule is:

$$w_{new,q} := E\left[x f(z_q)\right] - E\left[f'(z_q)\right] w_{old,q} \tag{2.33}$$

where the function f and its derivative are:

$$f(z_q) = -z_q \exp(-\frac{z_q^2}{2}) \tag{2.34}$$

$$f'(z_q) = z_q^2 \exp(-\frac{z_q^2}{2}) \tag{2.35}$$

Then, we normalize $w_{new,q}$ to unit vector by Eq. (2.36) and update the z_q by Eq. (2.32) for each iterations.

$$w_q := \frac{w_{new,q}}{\|w_{new,q}\|} \tag{2.36}$$

As the transformation vector w reaches convergence, we can calculate the independent components from Eq. (2.32). For simplicity in our implementation, we expand the equations as:

$$z_q = w^T \left(\frac{\phi_{pca}}{\sqrt{\lambda_{pca}}}\right)^T (x - \bar{x}),\ z_q = w^T \left(\frac{\phi_{emc}}{\sqrt{\lambda_{emc}}}\right)^T (x - \bar{x}) \tag{2.37}$$

where the left one is ICA with whitened PCA and the right one is ICA with whitened EMC. Thus, we can separate the subspace term ϕ_{ica} according to the whitening processes and rewrite as:

$$\phi_{ica} = w^T \left(\frac{\phi_{pca}}{\sqrt{\lambda_{pca}}}\right)^T, \phi_{ica} = w^T \left(\frac{\phi_{emc}}{\sqrt{\lambda_{emc}}}\right)^T \tag{2.38}$$

Then, we can estimate the class $\tilde{f}$ by Eq. (2.28)-(2.29) and (2.15) respectively.

2.3.4 Discussion

Performance of subspace methods on still image features

Although this dissertation mainly concerns the temporal analysis, we also determine the performance of the subspace methods on still image features. The average precision of each

dataset, features, and classification methods are given in Table. 2.2. They are defined by the mean of all class precisions. In this section, we implemented four subspace classifiers: EMC, KEMC, ICA with whitened PCA, and ICA with whitened EMC. Furthermore, K-Nearest Neighbor (KNN) is used as the baseline method due to its non-linear property and generalization. The method is efficient in typical cases, where the features should be modeled with less complexity and fewer dimensions. However the drawback of KNN is its brute-force characteristic which supposes to exhaustively compute the distances between a newly input to every instance in the dataset. So, it is used only as the baseline method, and is not suggested to apply in a real-time application.

To clarify the characteristic of the subspace methods we applied in this dissertation, the EMC and ICA methods are linear classifiers and the KEMC is non-linear classifier. In addition, we found the interpersonal issue in the CK+ dataset. Most of the subspace methods fail to classify the emotion classes in such circumstances except ICA with whitened EMC method.

Table 2.2: Average classification precision of EMC, KEMC, ICA with whitened PCA, and ICA with whitened EMC on still image feature of all datasets comparing to the baseline method (KNN).

Dataset	Method	Avg. Precision
JAFFE	Baseline method	79.84
	EMC [6]	74.92
	KEMC [20]	65.39
	ICA with whitened PCA [88]	79.69
	ICA with whitened EMC [7]	**84.38**
MyFace	Baseline method	**90.48**
	EMC [6]	42.86
	KEMC [20]	85.71
	ICA with whitened PCA [88]	85.71
	ICA with whitened EMC [7]	80.95
CK+	Baseline method	37.63
	EMC [6]	9.79
	KEMC [20]	18.81
	ICA with whitened PCA [88]	17.17
	ICA with whitened EMC [7]	**72.73**

Intra-personal cases (MyFace dataset) Among the subspace methods, both KEMC and ICA with whitened PCA yields the highest average precision (both 85.71%) in MyFace dataset. As we mentioned above that KNN is more generalized in the simple case, the results of the intra-personal dataset (MyFace) indicated that the performance of baseline method is higher than the subspace methods. This clearly shows that in the less complexity of the feature in intra-personal case.

Inter-personal cases (JAFFE, CK+ datasets) In contrast, the ICA with whitened EMC yields the better precision rates when the number of individualities is higher in JAFFE and CK+. The performance gap is overwhelmingly different in CK+ dataset.

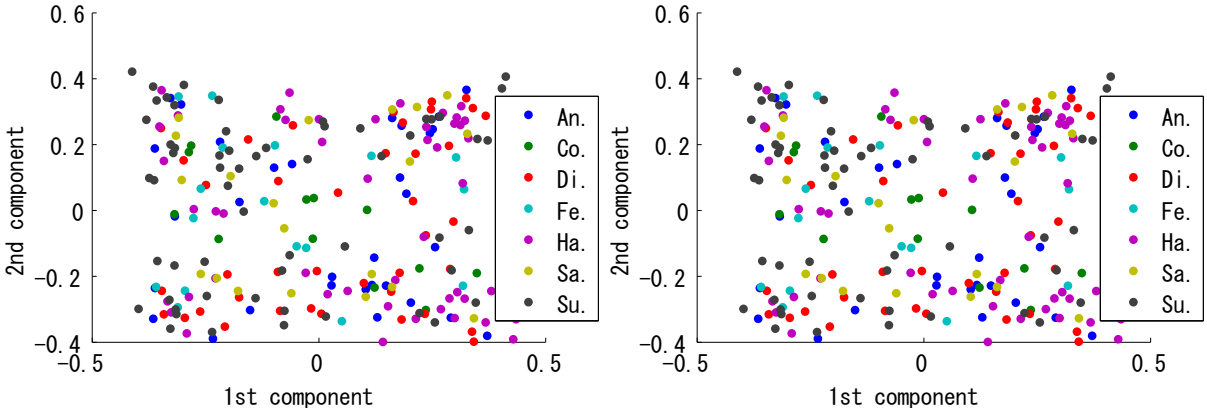
Most of the subspace methods fail to archive the correct classes in CK+ dataset (<20%), except the ICA with whitened EMC (72.73%). In addition, we found the EMC outperform KEMC in JAFFE dataset. The reason could be the over-fitting problem of the polynomial kernel in KEMC.

Notice that the evaluations of still image features and temporal features in Table. 2.2 and 3.3 do not imply that the temporal image is better. This section is established only to see the generalization of the subspace methods and to ensure the correction of subspace methods for the temporal features. The classification results of the still image feature are consistence to the experiment results with temporal features. Although we established and underlined the superiority of non-linear classifiers over the linear one in our earlier study, the compilation in our later studies showed the contradiction in this issue, especially ICA based methods which supposed to be linear classifiers surpassed all non-linear classifiers. These complementary results highly encourage the practical usages of the discriminative ICA method by EMC whitening process as the classification method.

Superiority of independent components

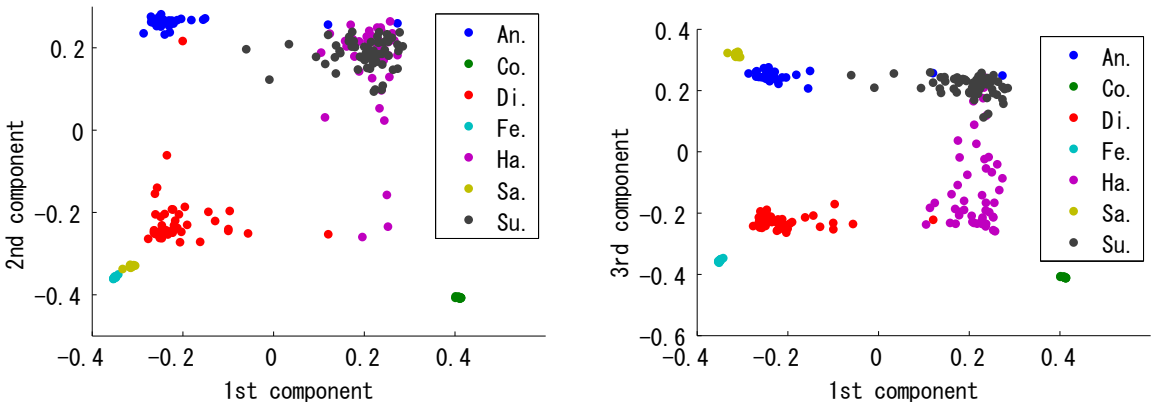
In this research, we determine the Independent Component Analysis (ICA) and its variant in [7]. The difference between them is the preprocessing step using whitened PCA or whitened EMC, which the latter has an ability to discriminate class better than PCA. The results of the above experiments confirm the superiority in class separation of the whitened EMC over the whitened PCA as it yields better classification performance. It is also important to note that both ICA with whitened PCA and whitened EMC yields the higher precision rates than the EMC method.

The example of scatter plot of the projected CCF features (see Chapter 3) on $1^{st} - 2^{nd}$ and $1^{st} - 3^{rd}$ independent components from CK+ dataset are shown in Fig. 2.12. Each feature is labeled its class by color as shown in the box on the right side of each plot. The distribution of the projected features on ICA with whitened EMC show the significant improvement of the class separation over the ICA with whitened PCA



(a) 1^{st} and 2^{nd} components of ICA with whitened PCA

(b) 1^{st} and 3^{nd} components of ICA with whitened PCA



(c) 1^{st} and 2^{nd} components of ICA with whitened EMC

(d) 1^{st} and 3^{nd} components of ICA with whitened EMC

Figure 2.12: Scatter of the projected CCF features on $1^{st} - 2^{nd}$ and $1^{st} - 3^{rd}$ independent components from CK+ dataset.

Chapter 3

Robust temporal features analysis

Facial expressions always act in animation. In the other word, it is the combination of many facial appearances or features in sequence. In previous researches, the interpretation of facial expression to features is frequently produced by using a still image. One of the remarkably works on the static analysis in computer vision is [75]. They employed the several feature selection and classifiers to recognize the action from a large number of Gabor features in the spontaneous facial expression dataset. Static analysis also appears in the recent works such as [91] which combining the geometric and appearance features.

However, analyzing facial expression from a single image cannot cover the whole diversity of our complex facial behaviors. For example in the case of smile which is supposed to represent the happiness emotion. But in fact people may smile during frustration to cover their true emotion [92]. Furthermore, the study in cognitive field also suggested that in order to fully understand the emotional messages from facial expression, the transition of facial expression should be perceptible in continuous temporal space [17].

In order to deal with the problem in the static analysis, the temporal analysis is taken place, where the change of facial expressions are observed in sequence and modeled as a feature for classification. In the early years of the studies about temporal analysis, the changes are modeled as the attribute of motion by finding full correspondence optical flow between frames [85]. Hidden Markov Model (HMM) is proposed to model the sequences of the optical flow for estimating the AUs [81]. Recent works of HMMs of optical flow is proposed in [47] to find the level of the interest, or exploit the multi-level architecture of Markov model layer and HMM layer [56]. Instead of model the temporal entity of facial expression in *classification level*, we suggested to model the temporal entity in *feature level* due to the quantization issue we discussed previously in section 2.2.1.

In this chapter we modeled temporal feature by temporal template approach. By assuming that a temporal transition of any action has a solid pattern, the temporal template packs the span of feature from several frames from a complete image sequence into a fixed size template. Thus, we can handle the time-varying feature in temporal axis, which may has different sizes due to the variable timing of actions. The temporal template based method is originally proposed by Jame et al. [14][15]. The method is widely known as the Motion History Image (MHI) and Motion Energy Image (MEI). Especially MHI is one of the most famous methods for modeling the temporal information of human action recognition. The application and variation of MHI is found in [93].

Notice that in this chapter, we evaluate the performances of temporal features by classification precision of basic emotions as their labels are provided in each datasets.

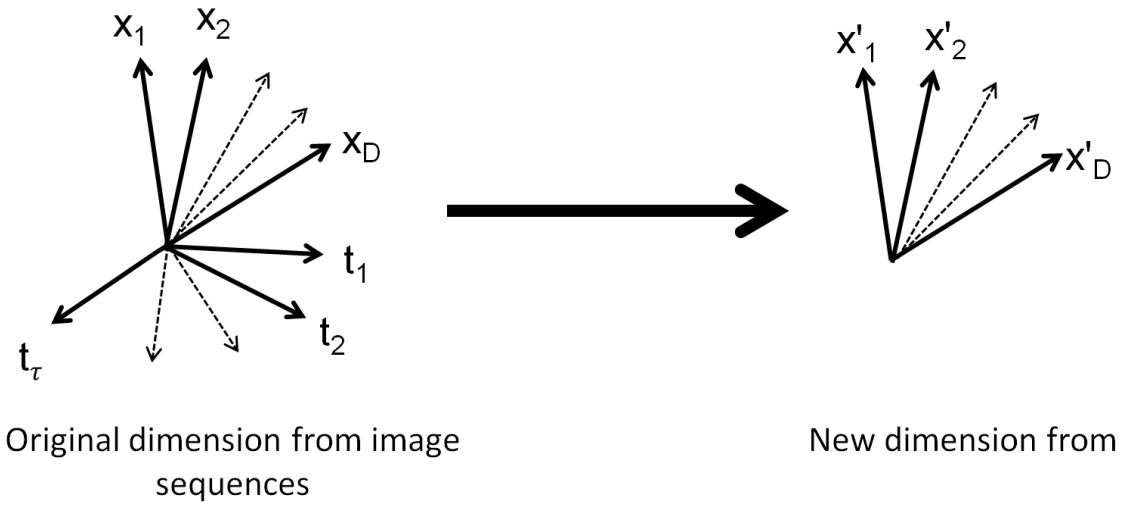


Figure 3.1: Temporal template describes the multidimensional feature of input sequence ($\tau \times D$ dimensions) by compressing the temporal variation into D-dimensions feature (τ is the number of observed frames, and D is the size of image feature)

3.1 Motion History Image (MHI)

The motion history image [14] at position (x, y) of time t is defined by:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \Psi(x,y,t) = 1 \\ max(0, \tilde{H}_\tau) & \text{otherwise} \end{cases} \tag{3.1}$$

where τ is the number of frame we observed. The updated $\tilde{H}_\tau$ is defined by:

$$\tilde{H}_\tau = H_\tau(x,y,t-1) - \delta \tag{3.2}$$

where δ is the decaying constant. The update rule $\Psi(x,y,t)$ can be expressed as:

$$\Psi(x,y,t) = \begin{cases} 1 & \text{if } D(x,y,t) \geq \xi \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

where ξ is the threshold, and $D(x,y,t)$ is the absolute differences between the intensities of frame t and $t \pm \Delta$:

$$D(x,y,t) = |I(x,y,t) - I(x,y,t \pm \Delta)| \tag{3.4}$$

As a result, MHI can indicate the order of the action over the image sequences. the direction and velocity of motion can be found if the frame rates or computation times of each frame are given. The direction of the motion can be found by extracting the gradient information in x and y directions using Sobel operator, and then we can find the direction of motion by the corresponding gradient orientations [94]. The example of MHI is shown in Fig. 3.5, where the colormap ranges from blue to yellow and red. The blue tone indicates the prior actions, while the red one shows the later actions. Even though the MHI is widely famous in human gesture recognition field, the implementation is rarely found in facial expression recognition field [95]. It is important to note that there is a significant difference between human gestures and facial expressions, where the human gestures are expressed with much larger motion than the facial expression that tend to be more subtle.

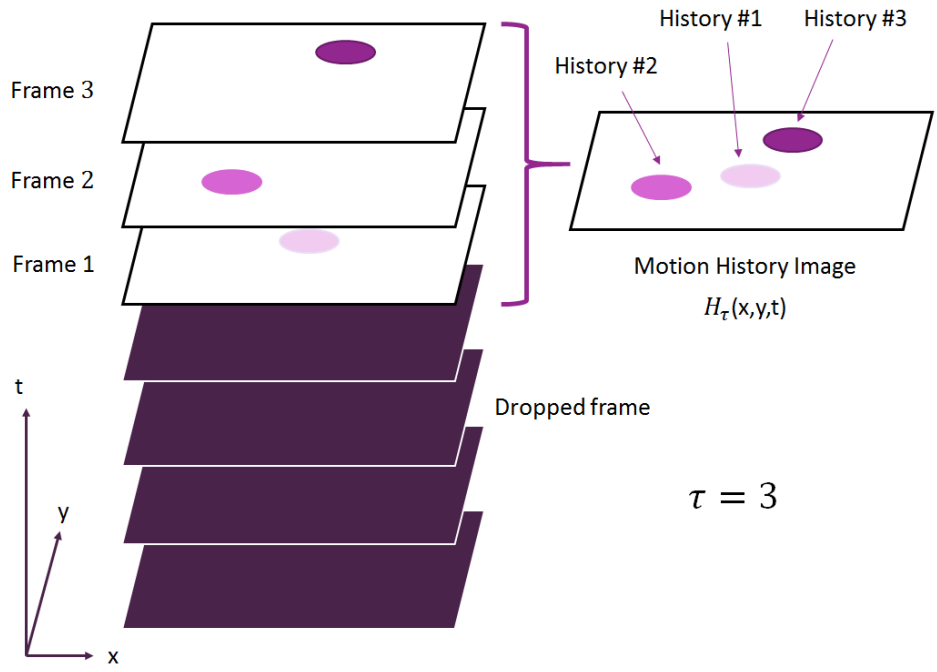


Figure 3.2: MHI procedure with τ=3

3.2 Cumulative Change of Feature (CCF)

Instead of updating the history of the last τ frame, we alternatively represent the temporal information in form of facial activation level by accumulating the change of facial expression over time. The new feature is called the Cumulative Changes of Feature (CCF). The procedure of CCF extraction method is illustrated in Fig. 3.3. The CCF feature is visualized by the colormap (Fig. 3.6), where the red tone indicates higher activeness of facial area, and the blue tone indicates inactiveness of facial area.

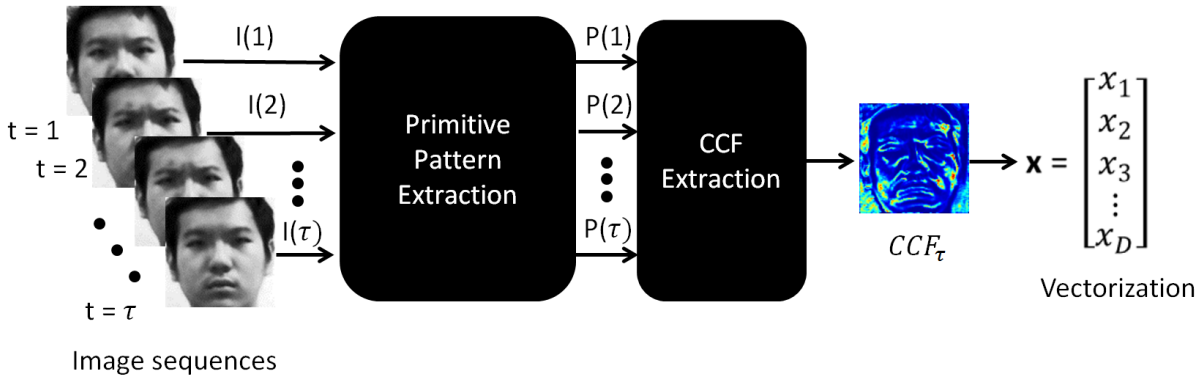


Figure 3.3: For each frame (1 to τ), the primitive pattern is extracted from its corresponding input intensity image.

3.2.1 Primitive patterns

In prior to the CCF extraction, the first step is started by converting the primitive pattern of each frame into more robust form. Typically, the intensity patterns can be utilized as the primitive pattern, however, they are sensitive to the variation of illumination. This susceptibility introduces the mis-calculated motion attributes, and noise. Therefore, this brightness inconsistency restriction highly influences the effectiveness of motion estimation techniques. In the early publications of this research, it is suggested to use the gradient orientation as the primitive pattern due to its robustness to the lighting change condition. The evidence of the robustness of gradient orientation has been supported in the literatures, author's previous works, and the experimental results in this dissertation.

Gradient Computation

A basic definition of the gradient is the first derivative of image intensities in spatial domain [62]. Let *f(x,y)* be a function of image intensities at pixel coordinate *(x,y)*. The gradient of a function f can be defined as Eq. (3.5)

$$\nabla f(x,y) = \begin{bmatrix} I_x(x,y) \\ I_y(x,y) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f(x,y)}{\partial x} \\ \dfrac{\partial f(x,y)}{\partial y} \end{bmatrix} \tag{3.5}$$

where I_x and I_y are the gradients in x and y directions. The gradient of image intensities is stronger at the region which the change of image intensities is higher. The direction of gradients points from the brighter pixels to darker neighborhood pixels. Because of this attribute, the gradient is typically used to represent the edge of an image or used for enhancing the quality of the images.

In order to compute the gradients of an arbitrary image, it can be approximated by convolution of a spatial filter to the image. The popular mask of gradient filter is known as the Sobel operators. The convolution of the Sobel masks to the image yields the gradient in x and y directions as shown in Eqs. (3.6)-(3.7)

$$I_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I(x,y) \tag{3.6}$$

$$I_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I(x,y) \tag{3.7}$$

The magnitude of the gradient vector at position *(x,y)* is given by Eq. (3.8)

$$||\nabla f(x,y)|| = \sqrt{I_x^2(x,y) + I_y^2(x,y)} \tag{3.8}$$

The magnitude of gradient vector at the position *(x,y)* can be computed from its vector in x and y directions. The gradient magnitude or length is typically referred to as a *Gradient* or *Edge* feature. It is one of the most well-known features used in several of object detection, tracking and recognition tasks. As previously mentioned, the gradient

of image is given by the strength of intensities change at a small region. The larger change of the intensities provide the bigger magnitude of the gradient vector, and smaller magnitude for the vice versa.

Next, the gradient orientation can be defined as Eq. (3.9)

$$\theta = atan2(I_x(x,y), I_y(x,y)) \tag{3.9}$$

Generally, the changing light conditions, locally or globally, result in different patterns of gradient magnitude, and this leads to fault detection or tracking, meanwhile the direction of gradient always remains the same. Gradient orientation is known to be insusceptible to the lighting condition changes over the image region. This advantage of gradient orientation is well considered in many robust feature extractions. Several feature descriptors utilize the gradient orientations by collecting the histogram [96], while some techniques use the gradient orientations directly in the spatial domain.

3.2.2 Cumulative Differential function

Suppose, a sequence of primitive patterns is described in vector form as:

$$P = \{P(t), P(t+1), P(t+2), ..., P(t+\tau)\} \tag{3.10}$$

For each primitive patterns at time t, it can be written in pixel-wise form as:

$$P(t) = \{P(1,1,t), ..., P(W,H,t)\} \tag{3.11}$$

where τ is the number of the observed frames, W and H is the width and height of image feature, and D is the total size of image feature given by $W \times H$. Then we compute the absolute difference between primitive pattern of two corresponding frames in the similar fashion to Eq. (3.4).

$$D(x,y,t) = |P(x,y,t) - P(x,y,t \pm \Delta)| \tag{3.12}$$

where $D(x,y,t)$ is the difference between two pattern at position (x,y) of frame t and $t-1$, $P(t)$ represents the primitive pattern at time t. Finally, we create a CCF by accumulating the difference from frame t to frame $t+\tau$ as in Eq. (3.13).

$$CCF_\tau(x,y) = \sum_{t=t}^{t+\tau} D(x,y,t) \tag{3.13}$$

The simplified visualization of the cumulative differential function is shown in Fig. 3.4

3.2.3 Experimental Results

Precision definition

The precision values are given by the average precisions P. They are calculated by:

$$P = \frac{1}{F} \sum_{f \in F} p_f \tag{3.14}$$

where p_f is the precision rate (%) of the individual class f, and F is the total number of classes.

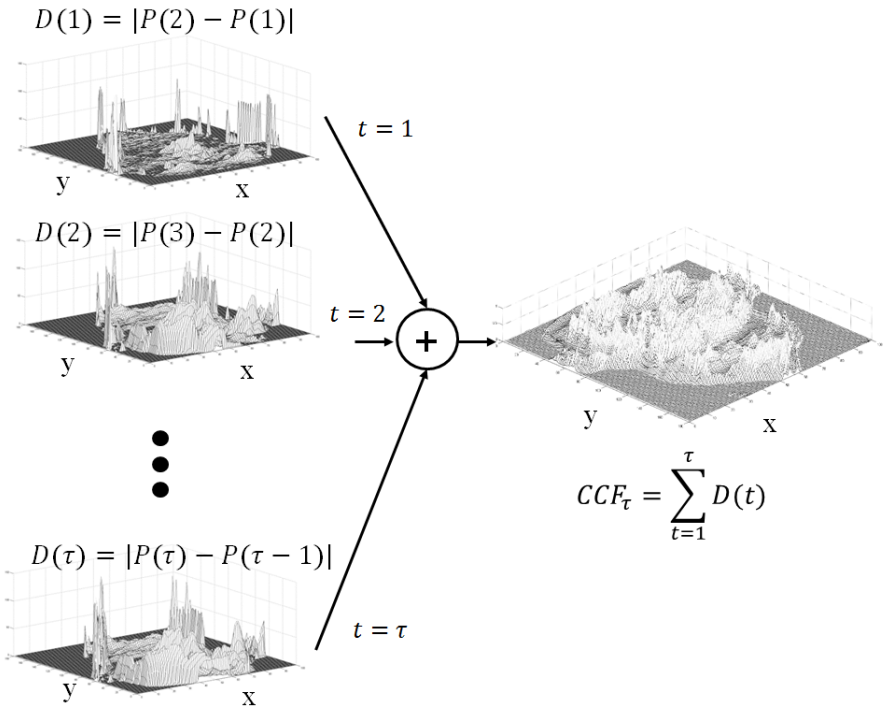


Figure 3.4: Simplified illustration of CCF extraction

Performance of subspace methods on temporal features

In this experiment, we consider both CCF and MHI features under intra-person case (MyFace) and inter-person case (CK+). In intra-person case, the degree of separable of CCF is averagely higher than MHI. The outstanding performance of ICA methods are demonstrated, which they yields 100% precision rates by both PCA and EMC whitening processes.

Despite the small variance of expressions in MyFace dataset, the CK+ dataset is also determined as the inter-person case. The experimental result yields the similar trend as in previous section. The ICA with whitened EMC method shows the remarkable improvement as the classification method across different persons. In spite of the poor

Table 3.1: Average classification precision of EMC, KEMC, ICA with whitened PCA, and ICA with whitened EMC on temporal template features of all datasets.

Dataset	Method	Avg. Precision	
		CCF	MHI [14]
MyFace	EMC [6]	85.72	42.86
	KEMC [20]	76.19	**66.67**
	ICA with whitened PCA [88]	**100**	23.81
	ICA with whitened EMC [7]	**100**	61.9
CK+	EMC [6]	2.95	9.57
	KEMC [20]	19.68	25.10
	ICA with whitened PCA [88]	13.13	6.06
	ICA with whitened EMC [7]	**43.43**	**58.59**

performance by most of the methods (<25%), the ICA with whitened EMC method yields 43.43% by CCFs, and 58.59% by MHIs.

Physical meaning of the features

As we mentioned previously, the implicit meaning of CCF and MHI are different. MHI presents the order of sequences (Fig. 3.5), and CCF can be interpreted as the facial surface activation levels (Fig. 3.6). In this experiment, we notice the severe problem by using MHI, which is the self-overwritten history. A newer motion can replace the previous history at the frequently active areas. Therefore, MHI in this experiment could be interpreting as the motions of face at a few last frames of the sequence rather than explaining the whole sequence. This resultant performance contradicts to the successfulness of MHI in gesture recognition due to the different size of the effective areas. The complexity of facial expression can be relatively higher than the gestures. A muscle action at the same spatial location can be occurred repeatedly over the observation time. In addition, MHI and CCF are both sensitive to head movement, noise, and mis-alignment from face tracker. In our empirical experiments, the presences of these outliers have an effect to the performance of temporal template feature especially MHI, which the outliers can overwrite almost of the previous history images. In this case, CCF has an advantage since the authentic facial motions are preserved, and the outliers will be a smaller as the accumulation proceeds. The activation images of CCF in Fig. 3.6 shows that CCF also enhances the head motion as a part of features. This indicates the blending of both facial motions, and head motions as a unified representation. The performance difference between MyFace and CK+ datasets suggests that the combination of head motion and facial expression can be a useful feature for person-dependent scenario.

Choices of primitive patterns

Furthermore, we also determine the optimal primitive patterns for creating CCF on MyFace dataset as shown in Table. 3.2. The gradient orientation is suggested as the primitive feature for creating CCF according to the experimental result in Table. 3.2.

Table 3.2: Average classification precision by finding the EMC subspace from CCF with different primitive patterns on MyFace dataset

Type of primitive pattern	Avg. precision (%)
Intensities	47.61
Gradient magnitude	66.66
Gradient orientation	85.72

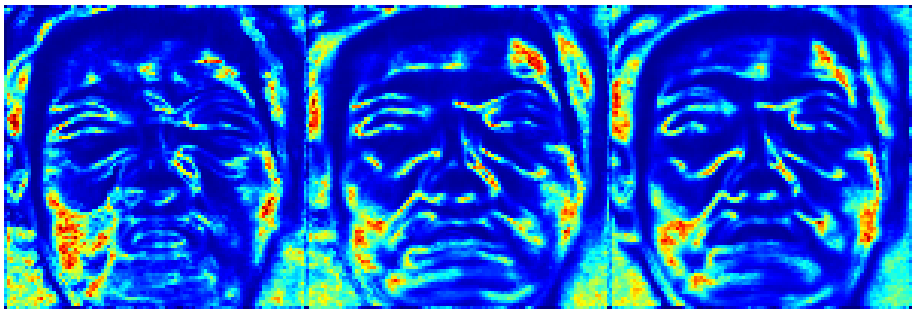
Interference from mis-alignment

Moreover, the quality of temporal template features is highly depended on the preprocessing process such as face tracking, segmentation, and head pose normalization. The current system is limited to only frontal face images. In order to show the importance of this issue, we exploited the different face alignment method by using color segmentation

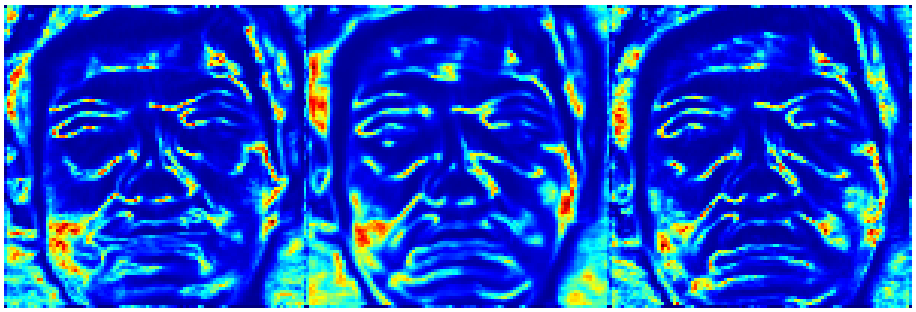
(a) Anger (b) Disgust (c) Fear

(d) Happiness (e) Sadness (f) Surprise

Figure 3.5: MHIs show the order of motions from different facial expressions. The orders of action are started from blue (first action) to red (last action).



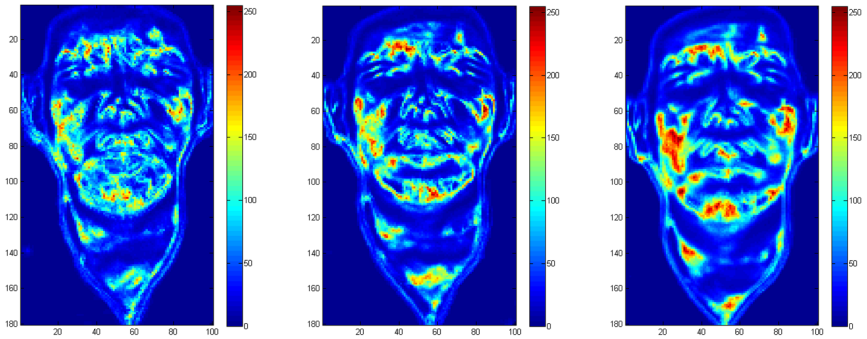
(a) Anger (b) Disgust (c) Fear



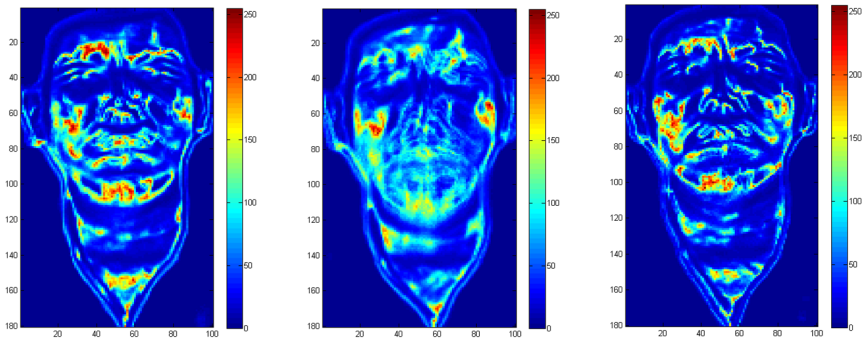
(d) Happiness (e) Sadness (f) Surprise

Figure 3.6: CCFs show the muscle activation area from different facial expressions, where the red color indicates the higher active areas and blue color show the less active areas

schema [97]. Under the strictly conditions of segmentation, head motions are almost completely neutralized. In comparison to Fig. 3.6, the CCFs under the well-aligned conditions shows the significant improvement of the CCFs quality (Fig. 3.7).



(a) Anger (b) Disgust (c) Fear



(d) Happiness (e) Sadness (f) Surprise

Figure 3.7: CCFs by using skin color segmentation schema as preprocessing method (*Publication I.*). Under the well-alignment, the quality of CCFs is much improved comparing to the results in Fig. 3.6

3.3 Cumulative Differential Gabor features (CDG)

This section introduces the novel temporal template feature for modeling the dynamic facial expression by the extended 2D Gabor features. This feature is named as Cumulative Differential Gabor feature (CDG). Then, the CDG feature is projected to the discriminative subspace for classification purpose. The overview methodology of this section can be described by the Fig. 3.8. The explanations of subspace classification methods are given in the previous chapter.

The concept of this section is driven by two biological characteristics. Firstly, the ability of recognizing a facial expression of an emotion is associated by the degree of perceptible motion or the change of facial expression [17]. Previously, much research neglected this part and estimated emotions from a still image [75], or the motions from a few consecutive frames [51]. We developed a feature descriptor, namely, Cumulative

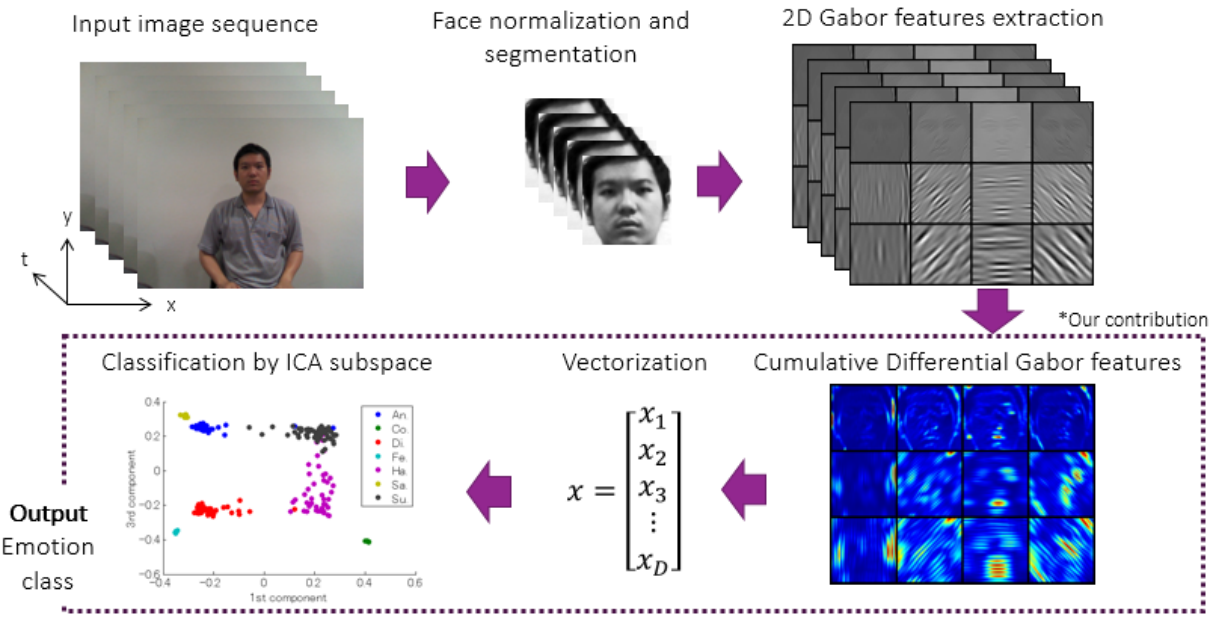


Figure 3.8: The overview methodology of *Publication III*. We introduce a novel feature modeled from dynamic facial expression by the extension of 2D Gabor features and independent component analysis.

Changes of Feature (CCF) [98], which model facial muscle activations by summarizing facial motions from an image sequence. Unfortunately, we found that CCFs are sensitive to interferences such as translation, scaling, blurriness, noise or varying illumination.

Secondly, the human visual system is discovered that its perceptual representation containing both spatial and frequency components. The mechanism is found also in the Gabor filters. Gabor features are often used as the robust patterns against the misalignment problems. One of the remarkable works that uses the Gabor filter to recognize facial expression is [75]. They employed the Gabor features with the several classifiers to discover Action Units (AUs). However, they use a still image to recognize facial expression which conflict with our previous statement. The temporal extension of the Gabor features in facial expression analysis is found in [76]. They extract the Gabor features from the several local areas in the face. Then, they use the Hierarchical Hidden Markov Model to recognize an emotion class. However, the generalization of the model is not assured due to the high dependency to the trained data and quantization methods. Although there is a work utilizing Gabor features for temporal analysis, the features are used as robust patterns for matching in tracking pursuit rather than analyzing the holistic representation [77].

Therefore, we combine the above two motivations and derive an extension of the 2D Gabor filter such that it can summarize the pattern of the facial expression over a time. We call this temporal feature as a Cumulative Differential Gabor features (CDG). Although there is a similar extension in [99] by adding another temporal filter on the top of the original 2D Gabor filters, their work utilized only a few frames for classification in either onset or offset states. Instead, we summarize a whole facial expression sequence by the temporal template representation in a similar fashion to the procedure of CCF extraction [98].

In addition to our *Publication III.*, we introduced a new implementation of the proposed feature by a superposition of all spatio-temporal responses in each frequencies and orientations. The new implementation boosted up the performance of features in interpersonal case significantly. We also investigate the robustness of the proposed feature under different conditions.

In order to describe the facial behavior in temporal, we expand the 2D Gabor features to a dynamic feature from an image sequence. Generally, the timing of a facial expression τ is not constant. The size of feature vector is proportional to the time τ and become $D \times \tau$ dimensions. Such an adaptable-length feature cannot be applied with the conventional classification. To represent the temporal feature as a fix-length vector, we employed a temporal template technique to wrap a sequence of spatial features into D-dimension feature [98]. The procedure of CDG extraction method is illustrated in Fig. 3.9.

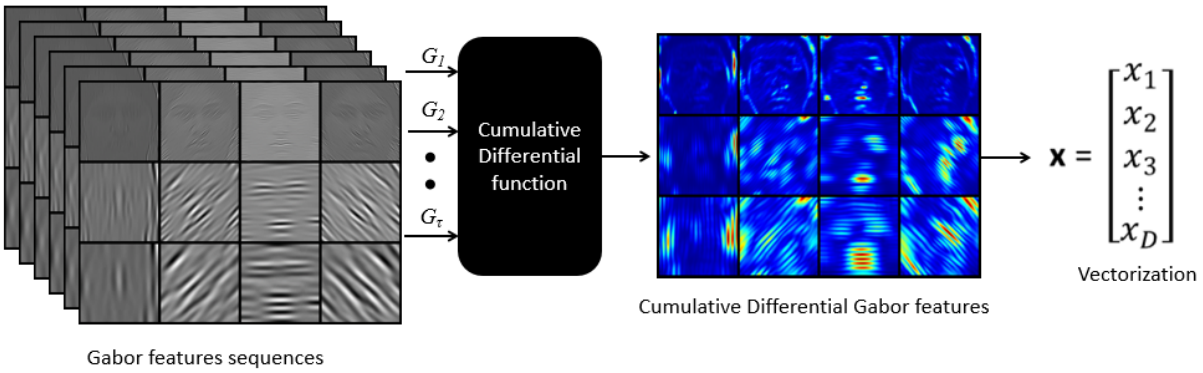


Figure 3.9: For each frame (1 to τ), the Gabor features pattern is extracted from its corresponding input intensity image. Then, we calculate the difference between two corresponding frames. The differential Gabor features are accumulated over τ frames creating the temporal template descriptor

3.3.1 Gabors features

Gabor filter is a linear filter named after Dennis Gabor [100], who proposed the mixture representation of the spatial and frequency domains of 1D signal, and later expanded into 2D representation by Daugman [16]. The filter has gained the attention from the research communities for a decade due to its similarity to the human's perception system.

Basically, the filter is defined by the product of the 2D Gaussian function and the complex exponential function. The kernel function is expressed by:

$$g(x, y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \tag{3.15}$$

where

$$x' = x\cos\theta + y\sin\theta \tag{3.16}$$

$$y' = -x\sin\theta + y\cos\theta \tag{3.17}$$

The notation θ represents the orientation in degrees, λ is the wavelength, ψ is the phase offsets, γ is the spatial aspect ratio to indicate the shape of Gabor filter, σ is the standard deviation of the Gaussian term.

According to the Euler's formula $e^{ix} = \cos x + i \sin x$, only the real part of the filters are utilized in this research. Thus, the expression becomes:

$$g(x, y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{3.18}$$

The Gabor features G can be extracted by the convolution of the input image I and the Gabor filter g with the specified wavelength λ and orientation θ as shown in Eq. (3.19).

$$G_{\lambda,\theta} = I * g_{\lambda,\theta} \tag{3.19}$$

The example of Gabor filters and their corresponding Gabor features from a particular face image are illustrated in Fig. 3.10. The advantages of Gabor feature are the invariants to small interferences from face localization misalignment (translation, rotation, scaling) [101]. Since Gabor features allow the simultaneous representation of both spatial and frequency, the features also gain the robustness to the blurriness, noises, and illumination variation [102].

3.3.2 Cumulative Differential function

A bag of Gabor features at time t is defined by a set of Gabor features according to different orientations θ_i and wavelengths λ_j:

$$G(t) = \{G_{\lambda_1,\theta_1}, G_{\lambda_1,\theta_2}, ..., G_{\lambda_U,\theta_V}\} \tag{3.20}$$

where U and V are the total numbers of wavelengths and orientations we used for creating the Gabor filters as we have discussed in the previous section. The sequence of Gabor features is:

$$G = \{G(t), G(t+1), G(t+2), ..., G(t+\tau)\} \tag{3.21}$$

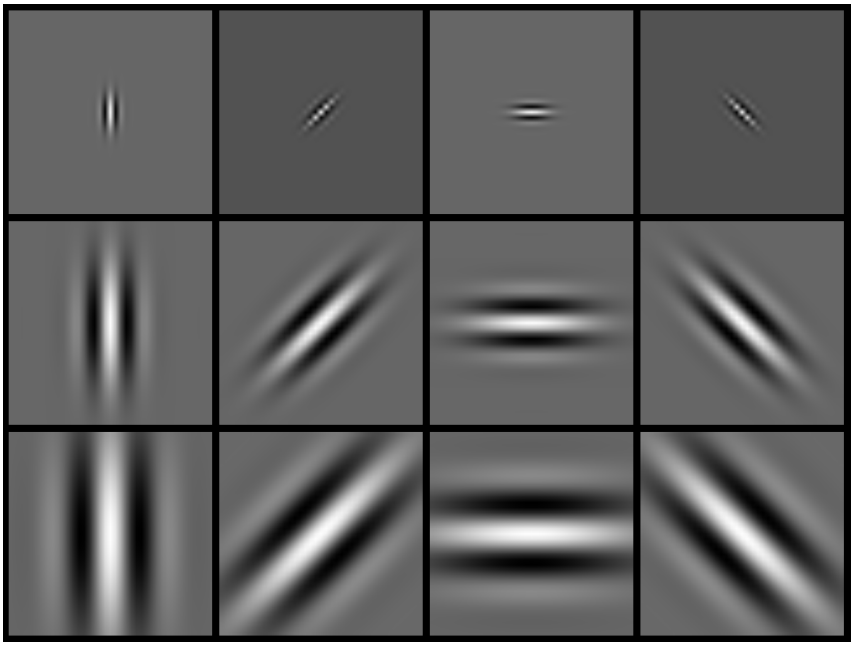
where τ is the number of observed frames. The Gabor feature $G(t)$ are normalized to 0 to 1. Then, we compute the cumulative change of the Gabor features from frame t to frame $t + \tau$. We obtain the Cumulative Differential Gabor features (CDG) from the sequence at positions (x, y) by the following equation:

$$C_{\tau,\lambda,\theta}(x, y) = \sum_{t=t}^{t+\tau} |G_{\lambda,\theta}(x, y, t) - G_{\lambda,\theta}(x, y, t \pm \Delta)| \tag{3.22}$$

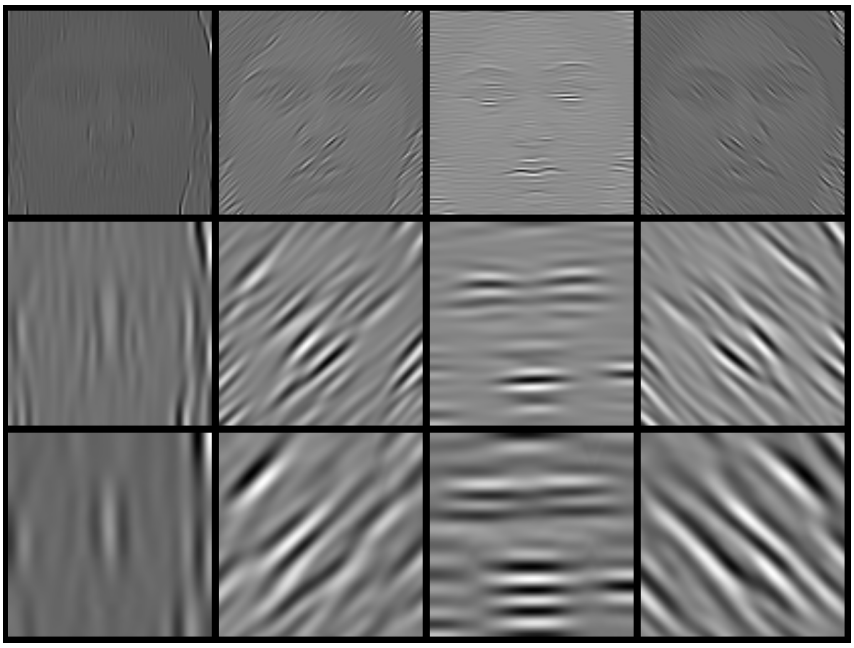
where $G_{\lambda,\theta}(x, y, t \pm \Delta)$ refer to the Gabor feature in the adjacent frame. The CDG feature of different orientations and wavelengths is shown in Fig. 3.11. The colormap is vary from blue (0 - inactive indicator) to red (1 - highest active indicator).

3.3.3 Superposition representation of CDGs

In our previous implementation [103], we represented a CDG as a standalone response in accordance to a particular frequency and orientation. Then we concatenated all vectorized CDGs into a single feature vector. As a result, we obtained a very huge feature



(a) Gabor filters



(b) Gabor features

Figure 3.10: The variation of Gabor filter kernels and the corresponding features by parameterizing 3 wavelengths ($\lambda = \{3, 8, 13\}$), and 4 orientations ($\theta = \left\{\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\right\}$)

vector, and the computation complexity of classification step became intractable. In this research, we newly implemented a superposition representation by simply combine all CDGs from every Gabor parameters (see Fig. 3.12). Therefore, we could obtain a compact representation and reduce the feature's dimension considerably. The superposition representation increases the performance, and allows us to have a better visualization of facial activation levels in spatial positions. The superposition formulation can be written

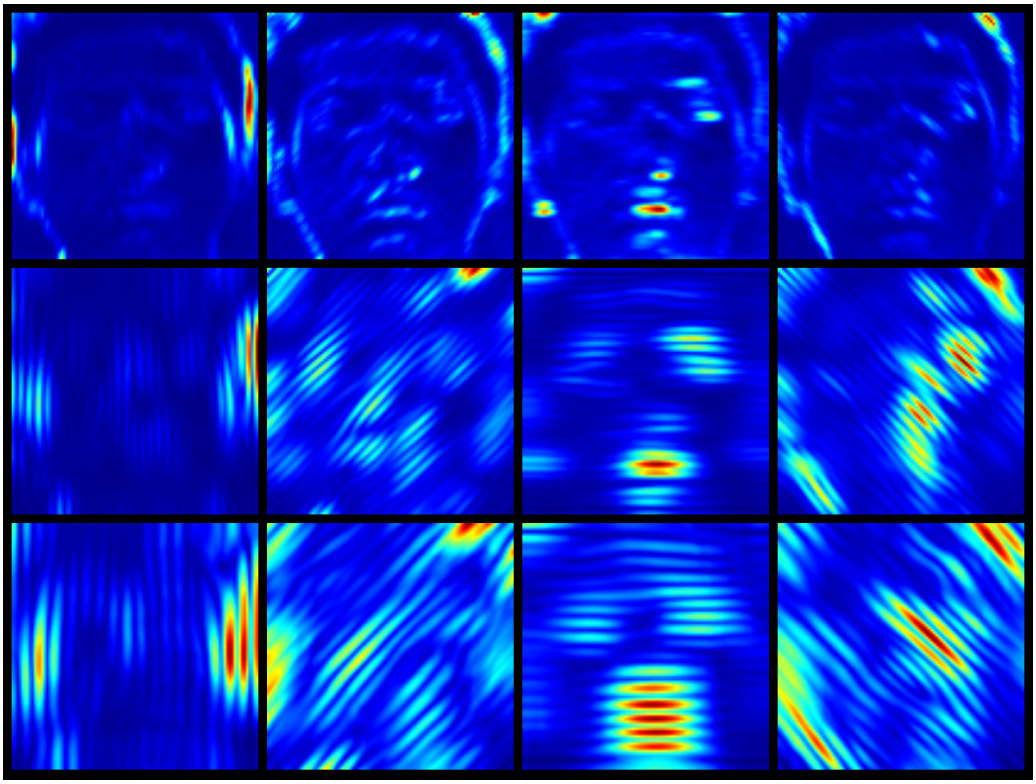


Figure 3.11: The Cumulative Differential Gabor features (CDG) produced from the Gabor filters with 3 wavelengths ($\lambda = \{3, 8, 13\}$), and 4 orientations ($\theta = \left\{\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\right\}$). The blue-to-red colormap represents the active levels from low to high respectively.

as:

$$C_\tau(x, y) = \sum_{\lambda,\theta} C_{\tau,\lambda,\theta}(x, y) \tag{3.23}$$

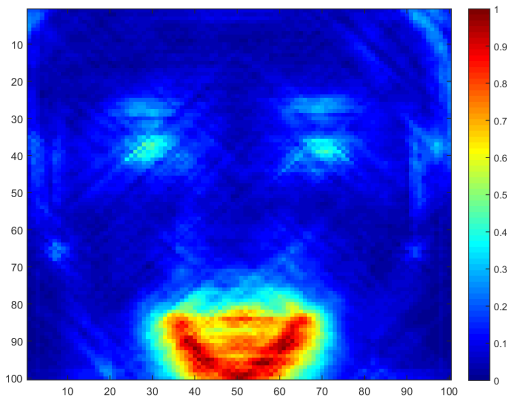


Figure 3.12: An example of superposition representation of CDGs from a happiness sequence in CK+ dataset

3.3.4 Vectorization

Concatenation representation

The vectorization process of CDG feature is slightly different to the conventional one, since it is a collection of responses rather than a single image. Firstly, we vectorize all CDGs in each orientations and wavelengths into column vector and then create a D-dimensions vector according to Eqs. (3.24)-(3.26). Given the Gabor feature with parameters λ_u and θ_v as:

$$C_{\tau,\lambda_u,\theta_v} = \begin{bmatrix} C(1,1) & \cdots & C(J,1) \\ \vdots & \ddots & \vdots \\ C(1,I) & \cdots & C(J,I) \end{bmatrix}_{I \times J} \tag{3.24}$$

where I and J are the numbers of row and column of CDGs. The matrix is transformed into D-dimensions column vector as:

$$C_{\tau,\lambda_u,\theta_v} \longrightarrow C'_{\tau,\lambda_u,\theta_v} = \begin{bmatrix} C(1,1) \\ \vdots \\ C(J,I) \end{bmatrix}_{IJ \times 1} \tag{3.25}$$

Then we append the feature vectors as follow:

$$x = \left(\begin{bmatrix} \vdots \\ C'_{\tau,\lambda_1,\theta_1} \\ \vdots \end{bmatrix}^T , \begin{bmatrix} \vdots \\ C'_{\tau,\lambda_1,\theta_2} \\ \vdots \end{bmatrix}^T ..., \begin{bmatrix} \vdots \\ C'_{\tau,\lambda_U,\theta_V} \\ \vdots \end{bmatrix}^T \right)^T_D \tag{3.26}$$

where the notation T is the transpose operator. Thus, we obtain the feature vector of size $D = (I \times J \times U \times V)$ dimensions. This feature vector x is applied to the classification module.

Superposition representation

In order to apply an extracted feature into the classification module, a superposition of CDGs is vectorized into a column vector. Suppose C is a superposition of CDGs from Eq. (3.23), we can create a D-dimensions vector according to Eqs. (2.1)-(2.2).

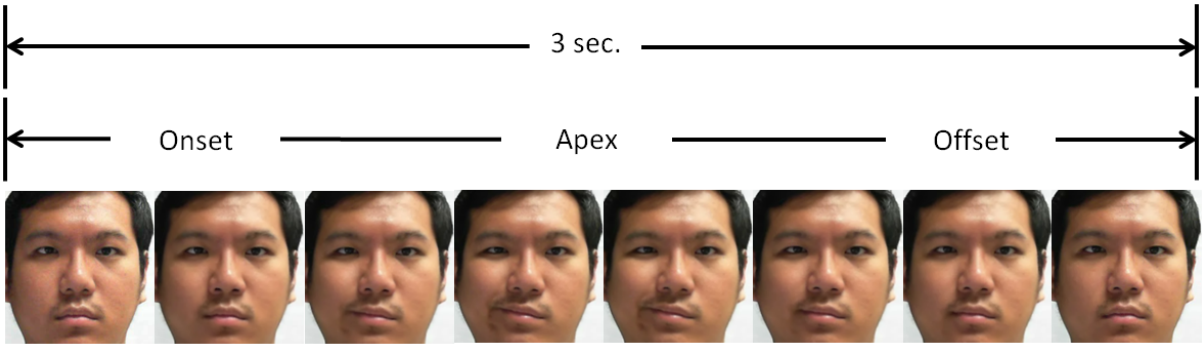
3.3.5 Experimental Results

Performance of features

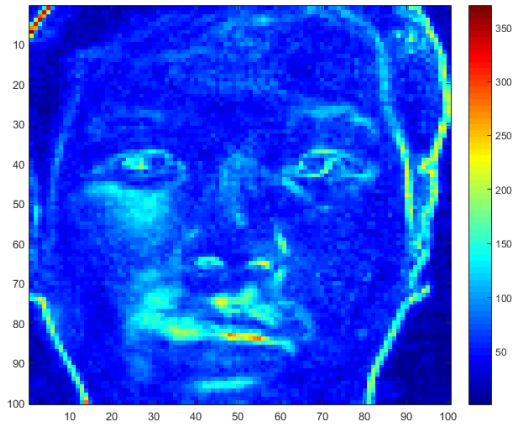
In the experiment, we observed the performance of the proposed Cumulative Differential Gabor features (CDG) compared with the feature descriptors in our previous work [98]. The notations Cumulative Differential Intensity (CDI) and Cumulative Differential Orientation (CDO) refer to the cumulative change of the features deriving by grey level intensities and gradient orientations respectively. In this experiment, we evaluated the performance of features by using the following classification methods: Eigenspace Method based on Class features (EMC) [6], and ICA with whitened PCA [88], and ICA

with whitened EMC [7]. The precision values in this experiment are given by the average precisions in Eq. 3.14.

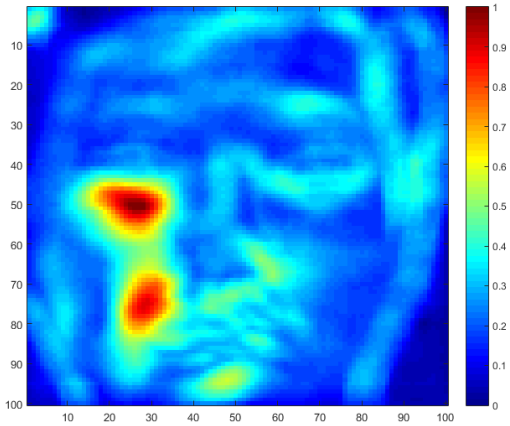
As we stated in the previous section, the major problem of the CDI and CDO are their sensitivities to head motions and misalignment. In the worst case scenario, CDI and CDO may enhance head motion rather than facial expression. As shown in Fig. 3.13, it shows a simple smirking of the right cheek without moving head. The resultant CDI has an unexpected peak at the edge, whereas the proposed CDG (superposition representation) can detect the muscle activations properly.



(a) Image sequence



(b) CDI



(c) Superposition of CDG

Figure 3.13: A simple facial expression by smirking right cheek. The other parts of face are inactive, and the head motions are minimized as much as possible. In order to illustrate the muscle activations, the superposition of CDG are calculated by summation of the CDGs in every θ and λ. The resultant superposition of CDG shows a clear activation of muscle activation comparing to the CDI.

Concatenation and superposition representations

In the first experiment, we parametrized the features by 3 wavelengths $\lambda = \{3, 8, 13\}$, and 4 orientations $\theta = \left\{\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\right\}$. The other parameters were left constant, where $\gamma = 0.5$, $\sigma = 0.56\lambda$, and $\psi = 0$ according to the setting in [104][105][106]. In this section, we studied two CDG representations: concatenation, and superposition (see section 3.3.3).

Concatenation representation The composition by concatenation of CDGs was applied in our previous study [103]. The experimental results are shown in Table. 3.3. As a result, this representation yields an insufficient performance, especially in the inter-personal scenario.

In the intra-personal scenario (MyFace), both CDO and CDG yield the best classification results by ICA with whitened EMC, where the CDO has a slightly higher average precision. The lower precision of CDG rate was caused by a drop of performance in disgust class (67%) as shown in the confusion matrix Table. 3.4.

In the inter-personal scenario (CK+ dataset), most the classification methods produce poor performance due to the higher interpersonal variances. However, the CDG overcomes the CDO when classified by ICA with whitened EMC (52.53%). The corresponding confusion matrix is shown in Table. 3.5. The classification result is especially high in the surprise class (90%) and happiness class (84%), but it fails to recognize the contempt, fear, and sadness class, because they partially share the overlapping areas of activation. This confusion is similarly found in both static analysis [107] and temporal analysis [56].

Table 3.3: Average classification precision on the changes of gradient orientations (CDO) and the changes of Gabor features (CDG) by using EMC, ICA with whitened PCA, and ICA with whitened EMC as the preprocessing step

Dataset	Method	CDO [98]	CDG (Concatenation)
MyFace (1 person)	EMC	85.72	38.10
	ICA with whitened PCA	100	80.95
	ICA with whitened EMC	100	95.24
CK+ (123 persons)	EMC	2.95	3.03
	ICA with whitened PCA	13.13	5.05
	ICA with whitened EMC	43.43	52.53

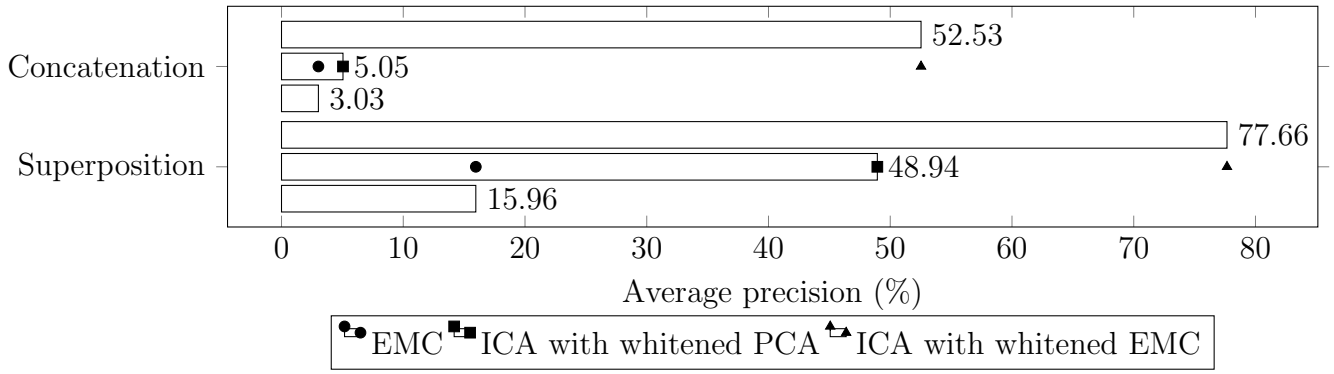


Figure 3.14: Average classification precision of the proposed CDG composed by using superposition and concatenation representations on the inter-personal scenario (CK+ dataset).

Superposition representation The superposition representation combines all the CDGs in every frequencies and orientations as in Eq.(3.23). Fig. 3.14 shows the comparison between concatenation and superposition representation in the interpersonal scenario. We

Table 3.4: Confusion matrix of the classification results of the CDG feature using ICA with whitened EMC (MyFace dataset)

		Predicted class						
		An.	Di.	Fe.	Ne.	Ha.	Sa.	Su.
Actual class	An.	**1**	0	0	0	0	0	0
	Di.	0	**0.67**	0.33	0	0	0	0
	Fe.	0	0	**1**	0	0	0	0
	Ne.	0	0	0	**1**	0	0	0
	Ha.	0	0	0	0	**1**	0	0
	Sa.	0	0	0	0	0	**1**	0
	Su.	0	0	0	0	0	0	**1**

Table 3.5: Confusion matrix of the classification results of the CDG feature (concatenation representation) using ICA with whitened EMC (CK+ dataset)

		Predicted class						
		An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
Actual class	An.	**0.21**	0	0.28	0	0.21	0.21	0.07
	Co.	0	**0**	0	0	0.4	0.2	0.4
	Di.	0.22	0	**0.5**	0	0	0.11	0.16
	Fe.	0.25	0	0.25	**0**	0.12	0	0.37
	Ha.	0	0	0.04	0	**0.90**	0.04	0
	Sa.	0.5	0	0.12	0	0	**0**	0.37
	Su.	0.04	0	0	0.04	0	0.08	**0.84**

Table 3.6: Confusion matrix of the classification results of the CDG feature (superposition representation) using ICA with whitened EMC (CK+ dataset)

		Predicted class						
		An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
Actual class	An.	**0.39**	0.07	0.15	0.15	0	0.23	0
	Co.	0.2	**0.8**	0	0	0	0	0
	Di.	0	0	**0.94**	0	0	0.06	0
	Fe.	0.12	0	0	**0.38**	0.12	0.25	0.12
	Ha.	0	0	0	0	**0.95**	0.05	0
	Sa.	0	0.12	0	0.12	0	**0.63**	0.12
	Su.	0.17	0.09	0.04	0.04	0	0.04	**0.61**

can see a clear gap of the performances between two representations. The superposition representation increased the classification precision up to 77.66% by using ICA with whitened EMC. The confusion matrix is shown in Table. 3.6. We found the promising improvements in almost every class especially contempt, disgust, and sadness.

Parameters evaluation

Theoretically, a low-frequency Gabor filter acts as a low-pass filter and reduces the influences from head motions and alignment errors. A high-frequency filter, however, is

required to compensate the loss information during smoothing process. By the multi-resolution schema, we use a set of Gabor features parametrized from various frequencies and orientations. As a result, we can reduce the effects of redundant motions around the high gradient areas, for example, a boundary of face, or eyes.

In this section, we performed an empirical experiment to investigate the effect of wavelength toward the CDG quality, and find the optimal frequency ranges. With the advantage of superposition, we can run an experiment on the parameter tuning faster than the concatenation one due to a smaller features size.

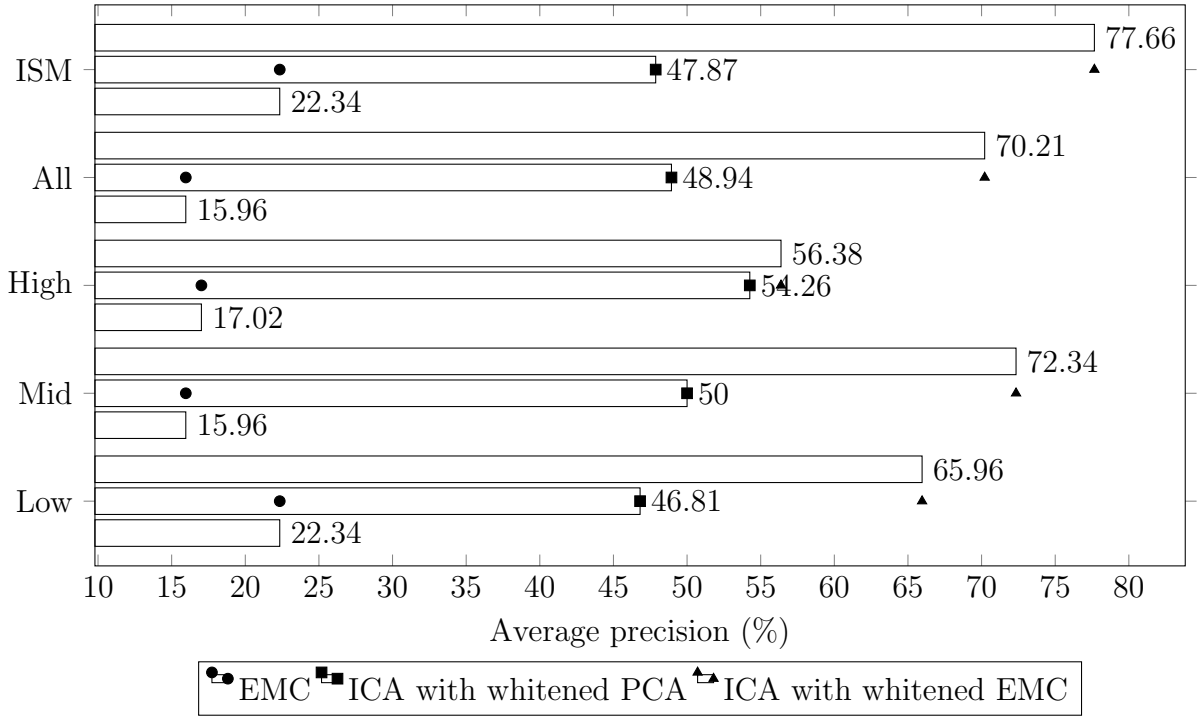


Figure 3.15: A parameter evaluation of the wavelength λ ranges.

In this experiments, we separated the wavelengths into 5 ranges as follows: *Low* range ($\lambda = \{3, 5, 7, 9, 11\}$), *Mid* range ($\lambda = \{10, 12, 14, 16, 18, 20\}$), *High* range ($\lambda = \{20, 22, 24, 26, 28, 30\}$), *All* range ($\lambda = \{3, 8, 13, 23, 28\}$), *ISM* range ($\lambda = \{3, 8, 13\}$). The *All* range is designed to covered *Low*, *Mid*, and *High* ranges. The *ISM* range is the wavelength ranges we used in [103] which contained both *Low* and *Mid* ranges. As you can see from the Fig. 3.15, the *Mid* range parameters tends to offer a better results than *Low* or *High* ranges. In our experiments, we found that the combination between the *Low*, and *Mid* ranges (*ISM* range) yields the best average classification precision (77.66%). It is interesting that our optimal parameters contain a fewer number comparing to the existing studies, which usually employed up to 7-8 orientations and 4-5 wavelengths.

Robustness evaluation

Since we want to use only motions related to facial expressions, the other interferences should be minimized. In this section, we evaluated the robustness of the features under various conditions. Only a part of MyFace dataset is applied to avoid the inter-personal

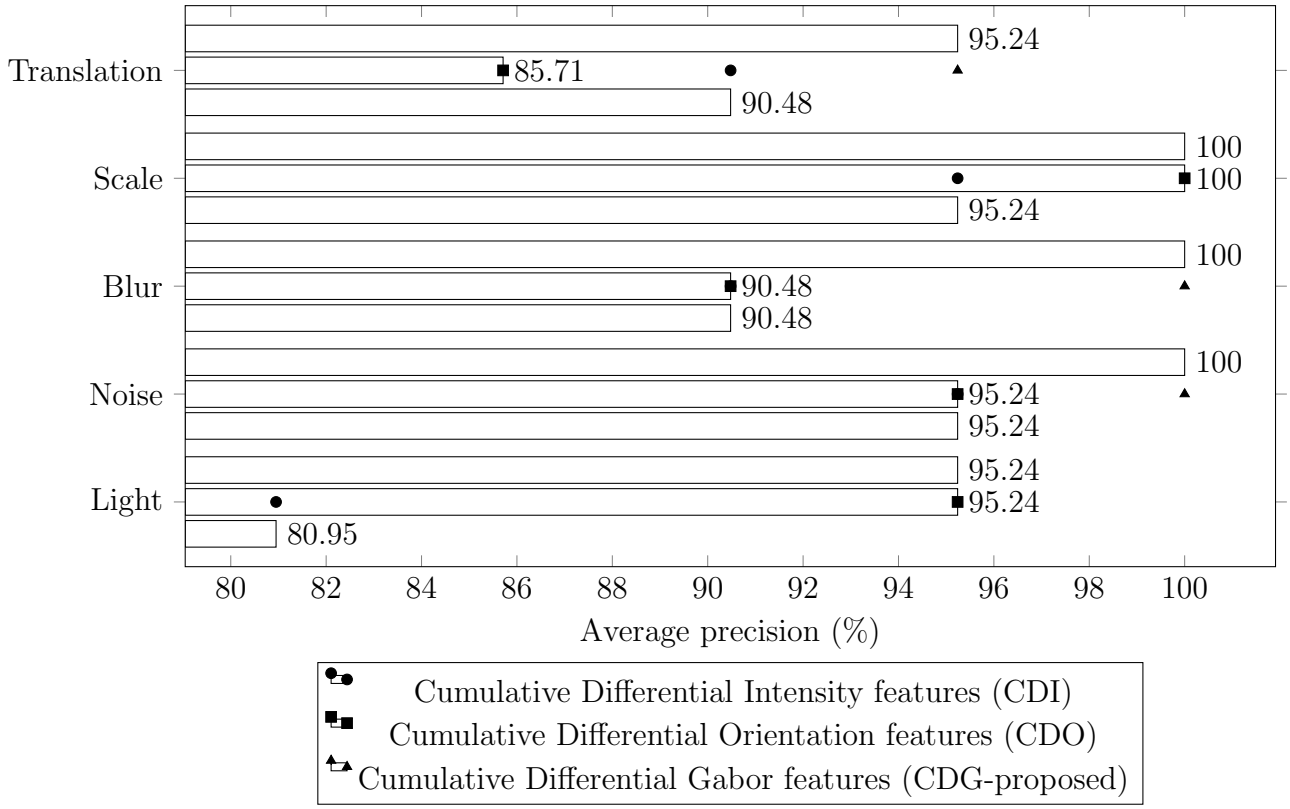


Figure 3.16: Comparison between CDI, CDO, and CDG (proposed feature) in all of the robustness evaluations

issue. For each frames in a sequence, we simulated following conditions: translation, scaling, blurriness, noise, and varying illumination conditions. Comparison between CDI, CDO, and CDG (proposed feature) in all of the robustness evaluations is illustrated in Fig. 3.16.

The translation and scaling conditions simulate the features with presence of localization errors. According to our previous study [98], the features quality is highly proportional to the accuracy of the face localization methods. In the given datasets (MyFace, CK+), the current face detection method is sufficient to extract a good temporal template features. However, there are limitations of efficiency of face detector in the real world. To simulate a translation condition, we deliberately shift the segmented face area away from the original tracked position. The range of shifting is randomly chosen from -5 to +5 pixels in both horizontal axis and vertical axis. The CDG feature yields the highest result (95.24%) in this translation evaluation. Notice that the performance of CDO is significantly lower than the others due to the dependency to edges of an image. For the scaling problem, we had a similar fashion to the translation condition by expanding or shrinking the segmentation area around the original one by 5 pixels. In this case, both of the CDG and CDO can perform equivalently.

In addition, we simulated blurriness and noise conditions which explicitly refer to the changes in frequency components. Blurriness eliminates the higher frequency components. In the opposite way, noise is an addition of high frequency components. These conditions were simulated by varying sizes of low pass filters, and 40dB additive Gaussian noise with

(a) Randomly shift the cropped area within range from -5 to +5 pixels in both x and y direction.

(b) Randomly expanding or shrinking the cropped area

Figure 3.17: Misalignment simulation. The original tracked position of face is marked by white bounding box.

random positions for each frames. Both blurriness and noisy conditions is surpassed by CDG.

Figure 3.18: Illumination variation by shifting intensity levels

Lastly, robustness to a changing light condition is tested by randomly shifting the global intensity levels of each frame. The brightness consistency is known as the fundamental drawback of the intensity based motion estimation methods. We can see the large gap in performance between CDG/CDO and CDI features. This result is consistent to the finding in earlier studies. This supports a usage of gradient orientations and Gabor features as the robust patterns against the illumination variations.

3.3.6 Discussion

In this chapter, we introduced a novel feature that describes a temporal pattern of a facial expression motivated by the biological characteristics. The proposed Cumulative Differential Gabor features (CDG) summarizes the changes of 2D Gabor features over a period of time. Furthermore, this chapter also considers the subspace approaches for estimating the facial expression, especially by using ICA based methods. We have known

that the facial expression problem is a class separation problem. The subspace methods are formed such that they can create the best discrimination subspace.

The experimental section in this chapter consists of the evaluation of different feature representations. The superposition of CDGs demonstrates clear facial activation, and yields significantly higher precision. We also investigated the effect of tuning frequency and orientation parameters. The lower frequency components are insusceptible to the head motions and alignment factors. However, high frequency components are also required. According to our experiment, the combination of low and high frequency responses are suggested in the multi-resolution schema. In contrast, we did not find any significant effects or trends of the orientation parameters toward the classification performance.

Lastly, the experimental results show the feasibility of proposed CDG and its robustness against misalignment (translation, scaling), and other types of interference (noise, blur, illumination varying). The experimental results are consistent with our assumption about the characteristics of CDG features. Moreover, ICA subspace method with EMC whitening process yields the highest precisions under the severe conditions.

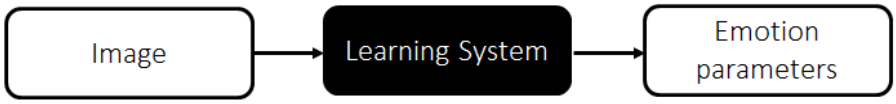
Nevertheless, the current extraction method is limited to offline simulation. The expression time can be varied, so the method requires a complete expression sequence, which starts from an onset to offset state of facial expression. In order to apply our proposed feature in the real-time environment, we suggest applying a temporal segmentation before the feature extraction.

Chapter 4

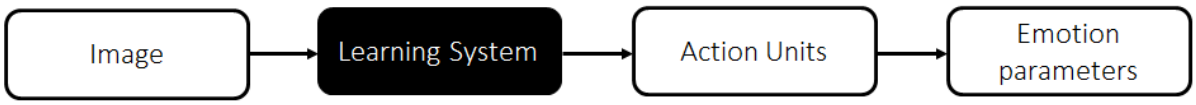
Facial action units detection system

In previous chapter, we proposed a new temporal feature Cumulative Differential Gabor features (CDGs). The proposed method overcomes the problems found in the conventional features and can robustly represents facial activations on the spatial area of frontal face. As we mentioned in the introduction chapter, a facial expression that we can see is practically a by-product of muscles contraction and relaxation, and those actions consequently draw fatty layer and skin surface. In this chapter we discussed how researchers can interpret our facial information to emotion by a mid-level feature Action Units (AUs) coded in Facial Action Coding System (FACS). Then we investigate the previous action unit detection methods the existing works, and proposed our system based on CDG feature, and multi-class discriminative ICA classification method.

4.1 Emotion translation from face



(a) Direct interpretation from face image to emotion parameters



(b) Indirect interpreation from face image to emotion parameters via action units (AUs)

Figure 4.1: Emotion translation from face

In computer vision studies, there are typically two approaches to recognize emotion from face image (Fig. 4.1). The first approach is to directly interpret the low level to emotion parameters. The successful of this approach has been well considered in the previous research. The second approach is to convert the low level to action units (AUs) in Facial Action Coding System (FACS)[2], and then use AUs to recognize emotions. An AU can be seen as the mid-level feature representation. By using the conversion table from AUs to basic emotions (Table. 1.2), we can approximate the emotional states as a rule based classifier. In addition, FACS is a standard tool to evaluate facial expression in

many field such as psychology especially emotion study, and also in medical studies (e.g. pain, stress).

In order to recognize more complex facial expressions, the second approach is investigated in this chapter. Although there are no evidences indicate that which approaches are more preferable in recognition of human emotion, the interpretation via AUs allows us to investigate more variation of facial expression. Moreover, the coding is independent of prior assumption of basic emotion expression.

4.2 Related works

In previous studies, action unit detection systems were implemented by both static and dynamic approaches. They are represented by either texture features or geometric features. The list of previous studies using AUs framework is shown in Appendix D.

In the early works, geometric based feature for recognizing AUs were done by Tian et al. They defined AUs by measuring activities of important face components such as eye, lip, brow, cheek [66] and apply to neuron networks, which were tuned for each single AUs. By similar idea, but with faster and more reliable ASM face detector, Tsalakanidou and Malassiotis [67] represented distance between landmark points and define several block local area on the face such as forehead, root of nose, infraorbital furrow, nasolabial furrow to determine wrinkles from texture feature, and use rule-based classifier to measure AUs.

In texture based studies, Donato et al. evaluated the feasibility to measure AU from motion between neutral and peak expression, and texture of a whole face (PCA, LDA, ICA) or local area (by Gabor features, and PCA jets). Mahoor et al. [108] recognized the combinations of AUs which represent basic emotions, and focused on the sparse representation to speed up the computation. The remarkable AU detector by texture was proposed by Bartlett et al. [109], they suggested to select an optimal set of Gabor features using AdaBoost and then use support vector machines (SVM) to train each AU detectors. They further developed the first commercial emotion recognition product under the name Emotient[1]. However the implementation used in their commercial product is not revealed to academic yet, but we suspected that its method might be a similar approach in their previous publications.

As you can see, the majority of AU detection methods in the previous works assumed that facial expressions can be measured from a single image by either geometric feature or texture feature. However geometric feature cannot handle subtle face actions since the distances between landmark points are too small to be recognized as a particular action, and the texture patterns may different from person to person due to ages or genders. For example a crow's feet is appeared more clearly in mid-age individuals and the resultant actions around this area yield different texture pattern. Therefore, we suggest to model the feature after the transition of facial expression in temporal instead of the conventional still image features.

There are a few numbers of temporal implementation of AU detectors in either feature level (Yang et al. [78]), or classification level (Tong et al. [110]). However we doubt if their methods are able to capture the subtle expression. In this chapter we introduced a novel action unit detection method by using our proposed temporal feature in Chapter 3 and new

[1]http://www.emotient.com/

implementation of discriminative independent component analysis as the classification method to recognize each AU individually.

4.3 Proposed action unit detection method

4.3.1 Temporal feature extraction

In order to learn a temporal feature from image sequences, we apply Cumulative Different Gabor features (CDG) (section 3.3). The feature has been proved its robustness to various outliers, particularly misalignment and illumination variation. The hyper-parameters of CDG in the proposed AU detector are set follow the optimal values given in the previous investigation i.e. the frequency $\lambda = \{3, 8, 13\}$, $\theta = \left\{\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\right\}$. The input feature vectors for classification are represented by the vectorization of CDG from Eq. (3.23).

4.3.2 Classification architecture

In machine learning aspect, the problem of action units detection is the classification issue in the same way as emotion recognition system in previous chapter. However the number of classes in action units detection is higher than recognizing basic emotions in previous system. Consequently, the separation problem becomes more complicated. Thus, we proposed to separate the classification problem into several classifiers instead of use one subspace to classify 21 classes. This new architecture is commonly known as *one-versus-all*. The one-versus-all architecture simplifies the multi-class classification problem into binary decision between positive and negative samples. Therefore, we created 21 binary classifiers separately.

Training manifold vectors

In order to train AU detectors, we used facial expressions with their labeled AUs in the extended Cohn-Kanade dataset (CK+). The number of training samples and classes are limited by available AUs in CK+ dataset. Despite the 30 AUs are provided in 593 image sequences, we constructed only 21 AUs by choosing the frequently used action units that has a sufficient number of samples for training (see Table. 4.1).

Table 4.1: Number of AUs coded in CK+ dataset. The chosen AUs in our experiment are highlighted by bold font character

AU	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU11	AU12
N	**173**	**116**	**191**	**102**	**122**	**119**	**74**	**21**	**33**	**111**
AU	AU13	AU14	AU15	AU16	AU17	AU18	AU20	AU21	AU23	AU24
N	2	**29**	**89**	**24**	**196**	**9**	**77**	3	**59**	**57**
AU	AU25	AU26	AU27	AU28	AU29	AU31	AU34	AU38	AU39	AU43
N	**287**	**48**	**81**	1	1	3	1	29	16	9

The training process is visualized in Fig. 4.2. The mathematical formulation of the process can be written as follow.

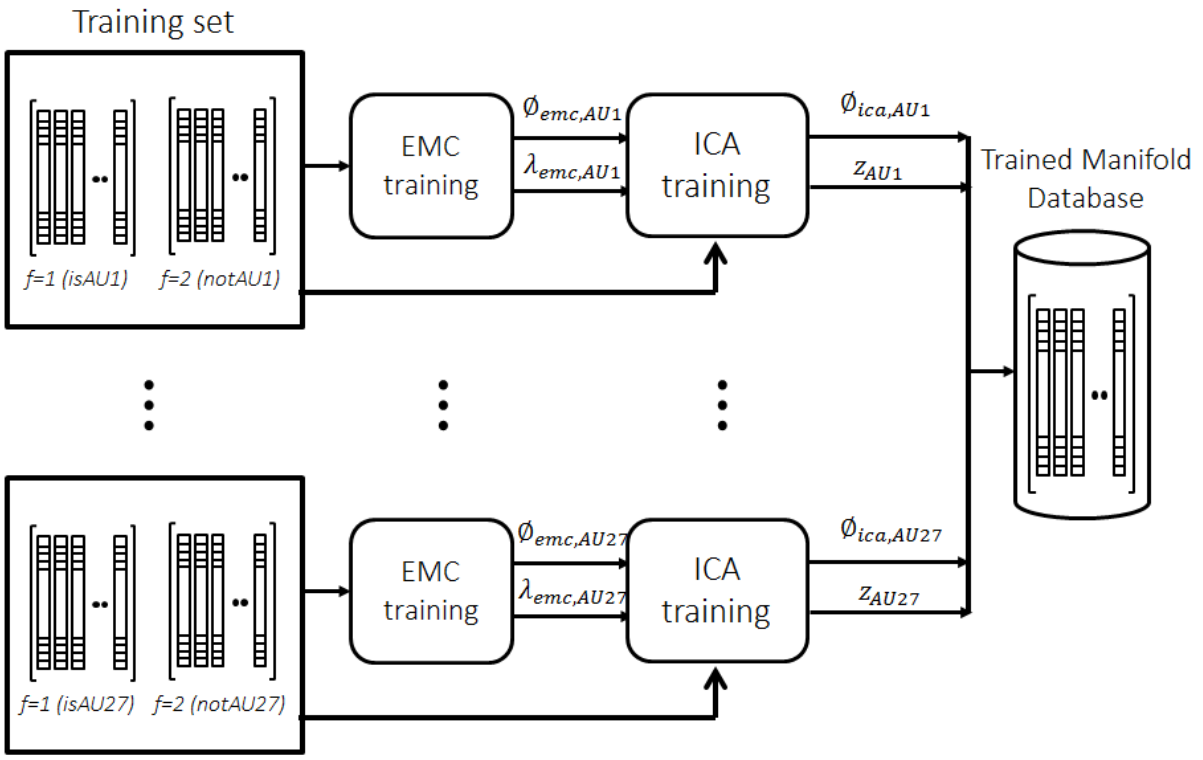


Figure 4.2: Training ICA subspace for each AUs

Suppose A is the total number of AUs, the ICA subspace $\phi_{ica,a}$ of AU index a is defined by:

$$\phi_{ica,a} = w_q^T \left(\frac{\phi_{emc,a}}{\sqrt{\lambda_{emc,a}}} \right)^T \tag{4.1}$$

where the term $\left(\frac{\phi_{emc,a}}{\sqrt{\lambda_{emc,a}}} \right)$ is the whitened EMC subspace. w is the transformation from whitened vectors on independent components from solving optimization problem following Eq. (2.33)-(2.36).

In order to create manifold vector $z_{fa} = \{z_{1fa}, z_{2fa}, z_{3fa}, ..., z_{M_f fa}\}$ of AU a^{th} classifier. For each feature vector m in class f, the manifold vector can be written by:

$$z_{fma} = \phi_{ica,a}^T (x_{fma} - \bar{x_a}) \tag{4.2}$$

where x_{fma} is an arbitrary feature vector, x_a is a mean of all feature vectors in AU a. In this system, the total number of class F is 2 ($isAU$ or $notAU$).

Recognizing Action Units

To recognize the occurrences of 21 AUs from an image sequence. The input feature vector $\hat{x}$ is projected on 21 subspaces:

$$\hat{z}_a = \phi_{ica,a}^T (\hat{x} - \bar{x_a}) \tag{4.3}$$

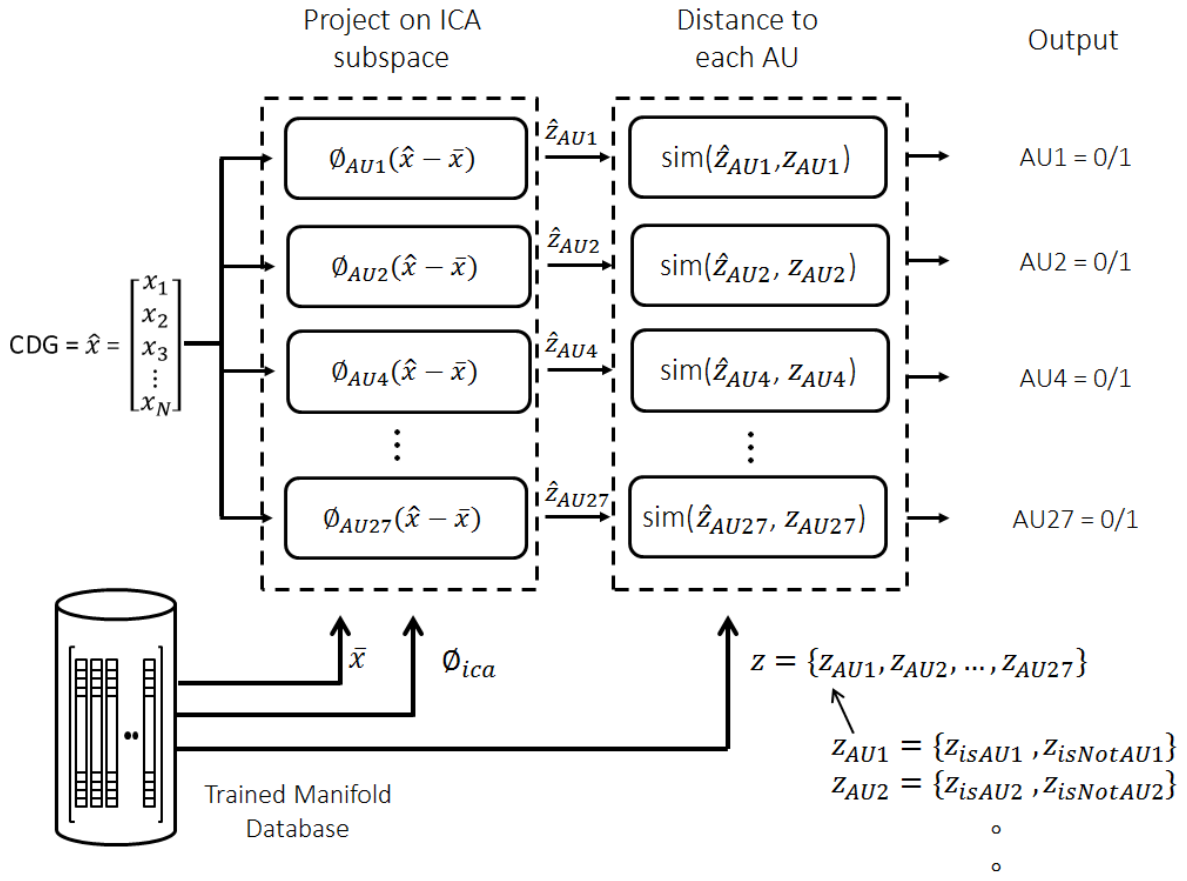


Figure 4.3: Block diagram of AU detection of a new sample by one-versus-all architecture

Then, for each classifiers, the input vector $\hat{z}_a$ is compared its similarity to manifold vectors in that AU's dictionary. The set of manifold vector is written by $z_a = \{z_{1,1,a}, z_{1,2,a}, \dots, z_{1,M_1,a}, z_{F,1,a}, \dots, z_{F,M_F,a}\}$

$$\tilde{f}_a = sim\left(\hat{z}_a, z_a\right) \tag{4.4}$$

$$sim\left(\hat{z}_a, z_a\right) = \underset{f\in F, m\in M_f}{argmin} \sqrt{\sum_{d}^{D} \left(z_{dfma} - \hat{z}_{da}\right)^2} \tag{4.5}$$

where M_f is total number of feature vectors in class f. The above AU detection procedure can be illustrated in Fig. 4.3.

4.4 Experimental results

To test the performance of our proposed method, we evaluated our proposed AU detector on CK+ dataset [4] by closed-data evaluation (Fig. 4.4) and random subsampling to split training and testing set (Fig. 4.5). Notice that closed-data evaluation indicates that all samples in the test set are also included in the subspace training process. Therefore, we would expect the perfect recognition rates in the case that the subspaces are modeled

properly. In addition, the result on closed-data can imply the fitting degree of subspaces to all samples available in the dataset.

The random subsampling evaluation can be conducted by splitting the training and testing set with a certain ratio (in this research, we split the dataset into 9:1). The experiments were repeated for 10 times for each AU detectors. Then, we estimate the mean and standard deviations of accuracy rates. By this approach, the experimental result shows less dependency of the trained subspace to the dataset.

The experimental results in CK+ dataset are shown in Fig. 4.4 and Fig. 4.5. The bar graph shows the average accuracy of each single action unit detectors. In our evaluation, the previous architecture of EMC [6] and ICA with whitened EMC [7] are applied as the baseline methods. The new architecture of subspace methods derived by multiple binary classifiers are denoted by M-ICA and M-EMC respectively.

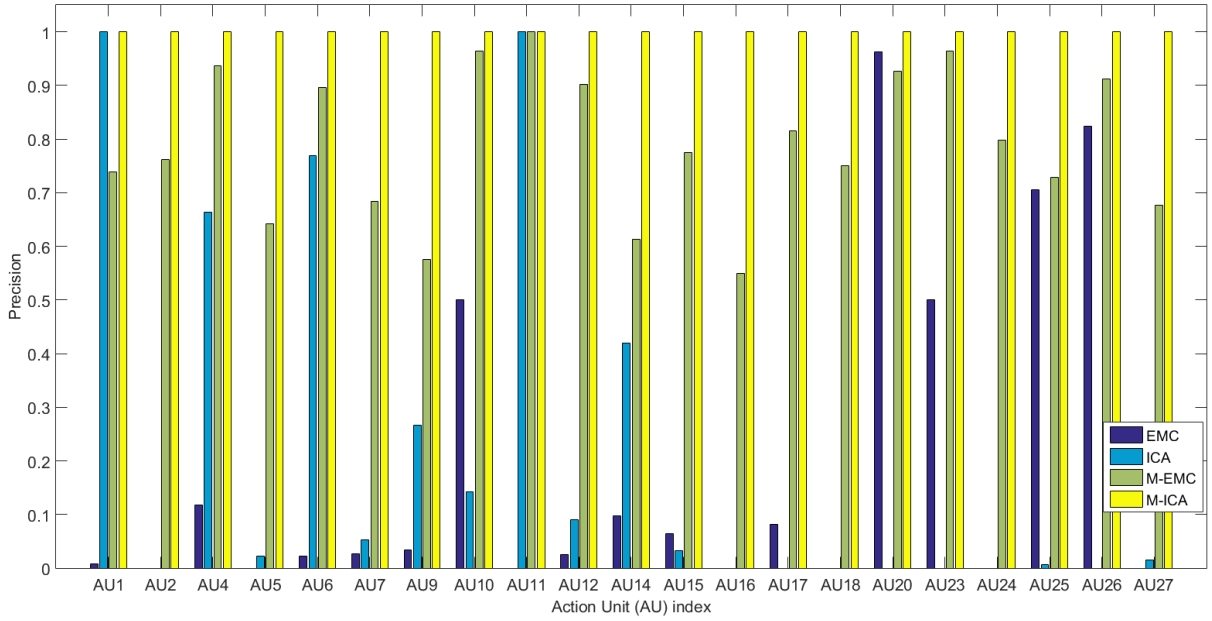


Figure 4.4: Performance of action unit (AU) detectors by using the one-versus-all architecture of multiple discriminative ICA (M-ICA) and multiple EMC (M-EMC), and the original implementation of EMC in [6] and discriminative ICA in [7]. The performances of all action unit detectors are evaluated by closed-data (Samples in test set are also included in the subspace training process).

As you can see from Fig. 4.5 (see numerical details in Table. 4.2), there are distinguishable gaps between previous architecture (EMC, ICA) and the new architecture (M-EMC, M-ICA). In the previous architecture of subspace methods, we measure distances from an input sample to every manifold vectors. Then we select the most matching class from shortest distance. In the other word, we have only 1 classifier to measure distances from a new input to the all 21 AUs classes. Since we prior assumed that the subspace methods can maximize the separability between each class. This assumption may be applicable for a fewer number of classes. But as we increased the class number (6 classes to 21 classes), the performance of previous methods are severely dropped especially in EMC method, which were fail in most AUs. We suspect that the performance of previous methods is influenced by co-occurrences of AUs. Some AU often appears in combination with another

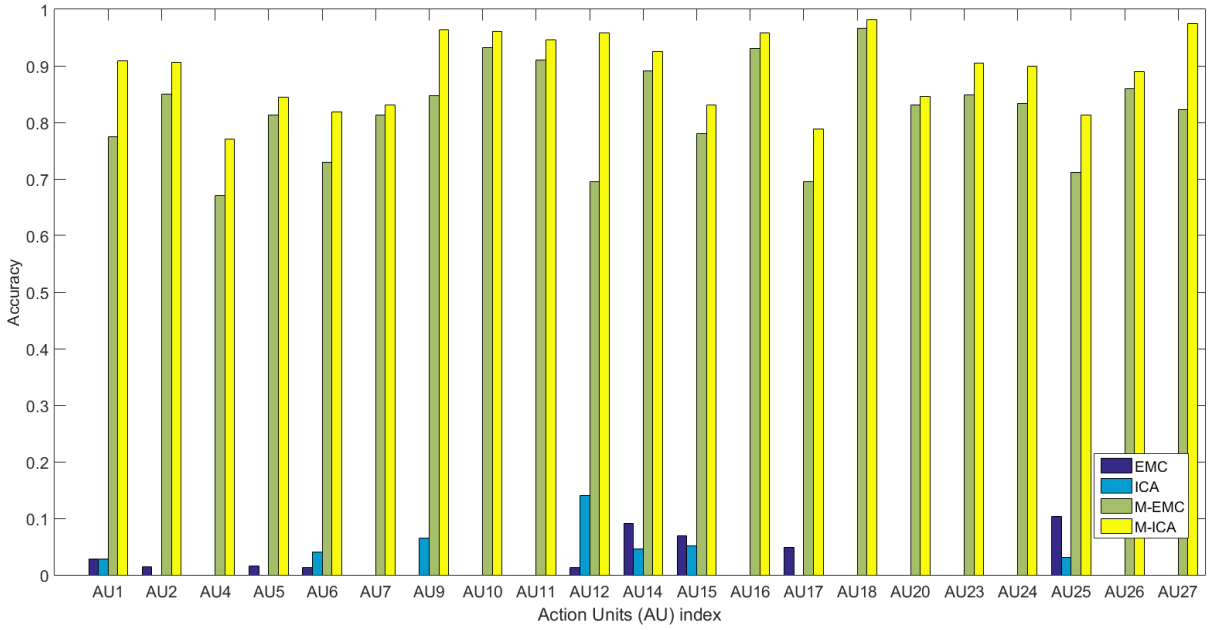


Figure 4.5: Performance of action unit (AU) detectors by using the one-versus-all architecture of multiple discriminative ICA (M-ICA) and multiple EMC (M-EMC), and the original implementation of EMC in [6] and discriminative ICA in [7]. The performances of all action unit detectors are evaluated by randomly spiting training and testing set (see details of M-EMC and M-ICA in Table.4.2)

AU. So the vectors that contain either of those AUs may project on the same position on the subspace.

On the contrary, the one-versus-all architecture we introduced in this paper measures a distance from input feature vector toward positive and negative manifold features of one specific AU. In the other word, in this case we have 21 binary classifiers for 21 AUs. By this architecture we can maximize the discriminative elements between relevant and irrelevant features. As a result, our one-versus-all architecture dramatically increased the performance of the existing subspace methods. The outstanding classification performance was found by M-ICA method. The resultant recognition accuracy rates of M-ICA are generally higher than M-EMC for every AUs. The overall accuracy of action unit detectors derived by M-ICA is 89.13 % with standard deviation 2.741. The least performance AU detector in this paper was found in AU4 detector (brow lower) with 77.12% average accuracy. We noticed that the action has been co-occurred with many other AUs. So, there are large overlaps of active areas. In this case, class separation problem becomes excessively complex and produce less generalized subspace. Similar issue has been found in AU17 (chin raiser) with 78.82% average accuracy.

The superiority in class separation is shown in Fig. 4.6. For a single AU detector, the distances are calculated by the similarity function between an input manifold vector $\hat{z}_a$ to all manifold vectors z_{fma} in the training set (Eq. 4.5). We can see a clear separation of the samples that relevant and irrelevant to the target AU (e.g. the minimum distances indicated that the AU5 is activated in this figure). For visualization, the relevant samples that have AU are sorted on the left side and the irrelevant samples on the other side.

Furthermore, since our system separate the decision process into 21 independent bi-

Table 4.2: Performance of action unit (AU) detectors (in %). The performances of all action unit detectors are evaluated by random subsampling to split training set and test set.

AU	M-EMC		M-ICA	
	Mean	SD.	Mean	SD.
1	77.50	1.18	90.84	5.9
2	85.00	3.17	90.67	3.52
4	67.12	3.73	77.12	6.00
5	81.36	9.59	84.41	4.54
6	72.89	5.31	81.87	2.85
7	81.36	2.40	83.06	2.40
9	84.67	2.67	96.34	1.46
10	93.23	3.04	96.11	0.78
11	91.02	1.71	94.58	1.48
12	69.5	4.80	95.77	1.2
14	89.17	1.18	92.50	1.18
15	77.97	4.80	83.06	4.80
16	93.06	1.77	95.77	0.85
17	69.50	3.77	78.82	2.45
18	96.62	2.40	98.14	0.51
20	83.06	4.02	84.58	3.52
23	84.83	3.07	90.44	3.43
24	83.34	4.72	90.00	2.36
25	71.19	2.40	81.36	3.94
26	85.94	4.43	88.99	2.05
27	82.21	3.60	97.46	1.20
Avg.	81.93	3.90	89.13	2.741

nary classifiers, so we can detect a combination of AUs that occur simultaneously at the moment. The example of multi-AUs detection (AU6+AU12+AU16+AU25) is shown in Fig. 4.7. We can see that for each detected AU the distances of on the left hand side (relevant to the specific AU) is lower than the irrelevant samples.

Although the proposed system is successful in the provided dataset, however, the major disadvantage of the proposed architecture is the higher consumption of resources comparing to the previous architecture in proportional to the number of classifiers.

4.5 Discussion

With a limitation of basic emotion implementation found in real life, we expanded the concept of interpretation between face image and emotions though Facial Action Coding System (FACS). Although we conjectured that either interpretation approaches (direct interpretation from face to emotion, or interpretation via AUs) may has the same implication in the aspect of pattern recognition regardless to the different procedures, FACS has the potential to objectively explain more complex facial expressions and can use to assist the further studies in psychology field since FACS is a standard tool for study face in the field.

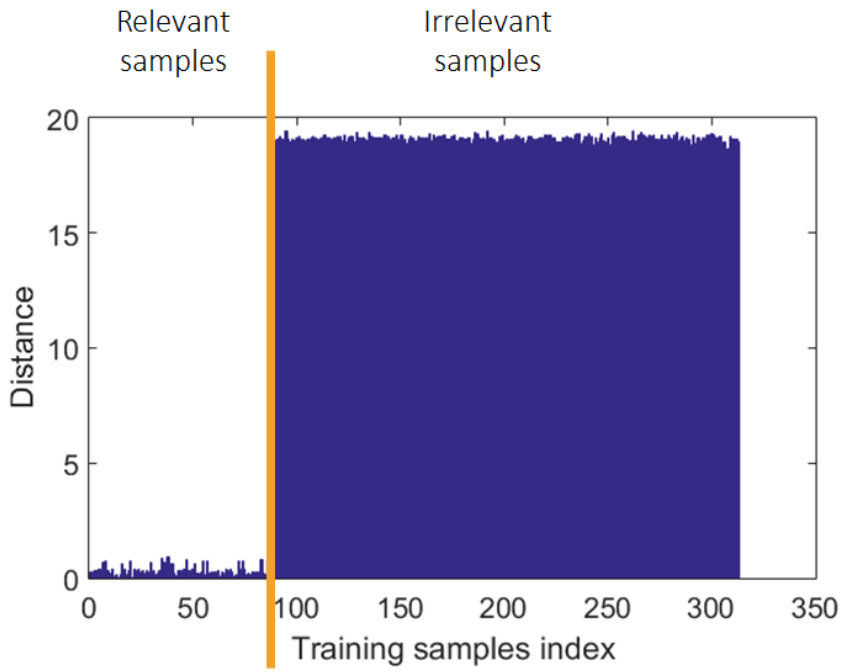
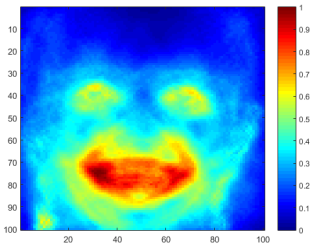


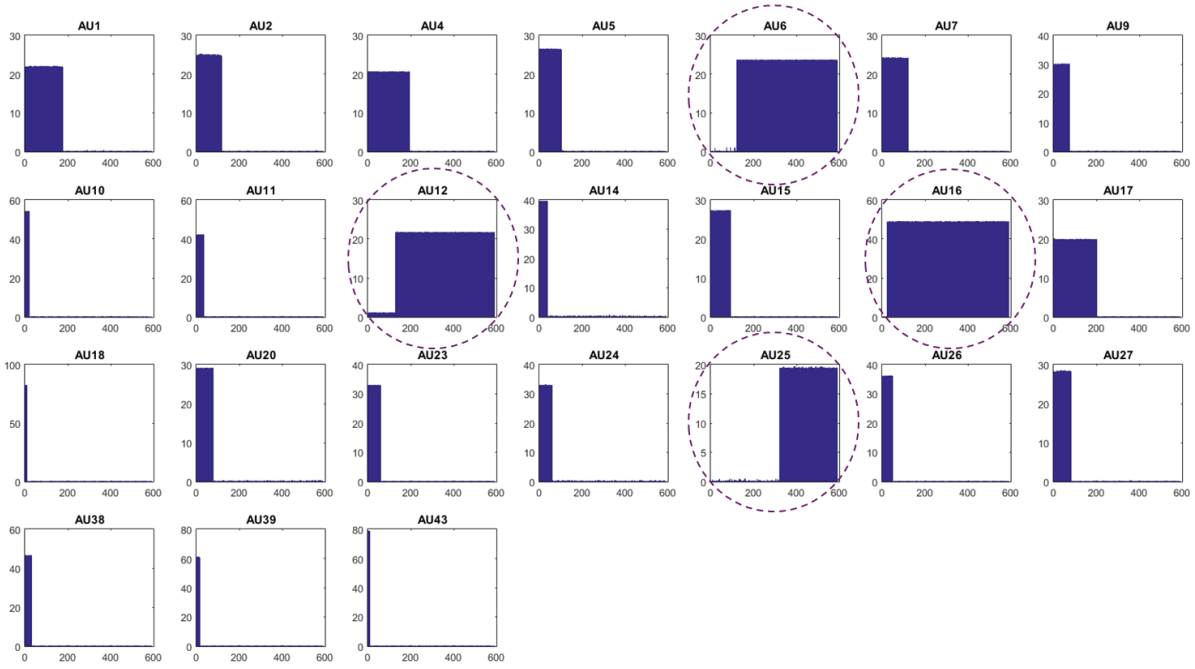
Figure 4.6: The result distances between an input $\hat{z}$ to all manifold vector z_{fma} in an single AU classifier shows a clear separation between distances from an input sample to relevant and irrelevant samples. The input sample is classified as relevant to AU5

Action unit detection is also the classification problem similar to the realization of basic emotions in our previous work. In this paper, we proposed a novel action unit detector that utilized a robust temporal feature that can preserve subtle expression, and we introduced an improved architecture of subspace classification methods. Instead of using only a single set of discriminative subspace to separate all action unit classes, we improved our previous classification methods by implementing one-versus-all architecture. Therefore, we can determine if the sample is relevant to a specific AU or not i.e. each detector is considered as a binary classifier. The improved method is denoted by multiple discriminative ICA (M-ICA). Experimental results indicated that the method is effective to train and recognize AU in the interpersonal dataset. However we conjectured that the method might be overfit to the tested dataset and need further investigation in open world environments. To the end we viewed action units (AUs) as the mid-level features, the estimation of output emotion from detected AUs can be archived by using Ekman's conversion [9] or other rules that might be available in the future psychology research.

(a) A part of image sequence with action units AU6+AU12+AU16+AU25 [4]



(b) CDG



(c)

Figure 4.7: The result distances of the successful multiple AU detectors. They are calculated by euclidean distances between the projected data (extracted from a sequence shown in sub-figure (a)) on the subspace and each manifold vectors in the dictionary trained from CK+ dataset [4].

Chapter 5

Conclusion and Future works

5.1 Conclusion

In the generation that robots are entering into mainstream industries. It has been a dream to build an intelligent machine that can understand human emotional states. However it brings us to the greater question of how can we teach a machine to understand our emotional expression?

Of all the human body, the face is one the most expressive and readable area in nonverbal communications. In addition, it is non-invasive perceptible information, which makes a face become the most promising cue to observe the emotional messages in the intellectual machine. By the simplification of the system and all of the issues we mentioned in this dissertation, we limited the scope of this research to the estimation from facial expression as the initial phase of complex system development. We believe that there is a way to find a relationship between complex emotion and complex facial expression. With this anticipation, we emphasized the contents in this research on the analysis of complex facial expression.

In this research, we particularly study the transition of facial expressions indicating an emotional state. The research has been studied under the context of human-machine interaction due to its obvious cue for displaying the emotional states for communication and can be perceived by a robot or machine in non-invasive environment. The problems, trends, and previous methodology were explained in the Chapter 1 and Chapter 2

Among the components of facial expression recognition system: *data preprocessing*, *feature extraction*, and *classification*, we especially interested on the feature extraction as it is the key to understanding our face behaviors. We believe that the facial behaviors should be observed in form of changes, transitions, or motions. Although the point of using motions had been interested in the early works, but most of the recent studies avoided this crucial points due to the sensitivity of motion based methods toward face alignment issue. In addition, many recent studies often mentioned and claimed about the *spontaneous* facial expression, however they did not truly consider the *subtle* element of facial expression. By introducing a new compilation of robust patterns influenced by biological characteristics and accumulative procedure in Chapter 3, we proposed a novel robust temporal feature that can explain the *subtle* changes of facial expression and can overcome several problems in the previous works, including misalignment and illumination variation. Our proposed temporal feature can be seen as a coarse approximation to illustrate what is going on over our face and which spatial locations are active at the observed moment.

One may notice that our proposed feature has a similar intuition to FACS system, in which observing the activation of muscles. However, according to the previous studies, the most imprecise assumption of these AUs detectors is the usage of single frame's texture rather than the collection of motion information. Thus, we use the proposed robust temporal feature and discriminative subspace to recognize AUs in Chapter 4. Therefore, we can explain the complex and subtle facial expression in the same standard with psychology and medical research, which is very useful as its usage can be applied in the further development of theories and applications in other research fields.

Several systems we proposed in this research has been noted as the *recognition* system. We adopted the term *recognition* in the psychology research to describe the process of facial expression understanding. Although, it can be said that people *recognizes* an emotional state from other's face, but in engineering aspect it is more appropriate to refer to the process of analyzing facial expression as the *estimation* process. Emotion study is one of the disciplinary with uncertainty in its theorem and need to be refined in the future. Regarding our study of complex facial expression by analyzing the transition of face in image sequences, we have guided the new foundation the future complex emotional estimation system.

The contributions of our research are delivered under the fields of facial expression analysis, computer vision, and machine learning. To the end of this dissertation, we summarized the contribution as follow:

- This dissertation presented a novel robust temporal feature modeled after the transition of changes in facial expression. The robust temporal feature can handle the small interferences such as alignment errors, illumination variation, noise, and blurriness.
- Subtle expressions, which were unable to be detected by using texture and geometric features, has been recognized by our system.
- This dissertation presented a novel Action Units (AUs) detector based on our proposed temporal feature and discriminative subspace methods.

5.2 Future works

5.2.1 Short term

Spontaneous facial expression In this research, we already determined the natural facial expression by taking the subtle expression issue into an account. However, there are other elements of spontaneous facial expression that we did not yet considered in this research such as duration, trajectory, symmetry, and co-occurrence that can be considered to distinguish a deliberately expressions (voluntary action) and a spontaneous expression (non-voluntary or reflex-like action) [111].

Personalized facial expression Most of the content in this dissertation assumed that facial expression of emotions can be recognized automatically by using statistical methods to form an average expression or common actions of a particular emotion. However each individual has their own way of expressing their feelings. In order to develop a system

that can interact with an individual person, the personalized system should be considered in the future works.

5.2.2 Long term

Pose and occlusion invariant system Head pose variation is considered as one of the most important issue in the modern face study in computer vision. Both face recognition system and facial expression recognition system requires a good frontalization technique. There is a trade-off between computational complexity and its performance. Although the current face detection methods provided a good accuracy for matching face in a clutter scene [59][60][112], but they are coarse approximations which are not enough to keep the consistency of face trajectory without making any self-occlusion. In order to investigate the feature transition across the facial expression sequence, the precise face detection system is inevitably in need. Up to our knowledge, there is no available frontalization that can fully handle out-of-plane rotation without making unwanted distortion. This distortion makes the system become malpractice in temporal analysis. Alternatively, the pose problem can also be dealt by multi-view approach in similar way to the methods in [113]. This approach practically utilizes a collection of all possible facial appearances in many angles, and then use them for training the classifier. However this approach requires an extensive labor to labels all features.

In addition, the occlusions found in typical uses such as hairs, glasses, hands, etc. damage the facial features. This problem has not been received much attention in the existing studies.

Three dimensional temporal analysis 3D facial expression recognition has been a trend in these recent years. To study action unit detection from 3D face, Tsalakanidou et al. attempted to implement automatic 3D facial recognition system for the first time. They combined 2D and 3D information by using geometric distances, edge texture, curvature from depth surface [67]. Similarly, Savran et al. also extracted multiple texture features of 2D and 3D images and determine the presence of a specific AU [114].

In order to study temporal facial expression, the 3D acquisition is required a precision in millimeter resolution and also is worked in real-time. Such a system needs extreme control in laboratory setting [115][116], and not yet available in consumer 3D acquisition sensors. Comparing to the studies in 2D information, the temporal analysis by using 3D information is hardly found due to the lack of such a precise dynamic 3D database and the dimension of feature is exponentially increased according to the enormous amount of data from 3D surface, textures, and timing of observed expressions.

Temporal segmentation The common problem of temporal analysis methods is they require the complete facial expressions, which start and end with neutral expressions (onset, apex, offset states). We may refer this issue as the temporal segmentation problem. Most of the facial expression research in computer vision neglects the temporal segmentation processes. One may avoid this issue by fix a certain period of observed windows as usually did in the body gesture recognitions. However, facial expression is an accumulation of actions that we can interpret as the emotions. The shorter periods of facial expression should be rather interpreted as the muscle activations instead of emotions. The behaviors of facial expressions are significantly different to the human action, which

the timing of expressions can be changed in larger variation comparing to simple bodily gesture. In addition, tuning the windows size also requires greatly effort to tune the optimal size of windows.

Temporal segmentation related works can be found in [117] which they perform binary classification (onset and offset states) by using SVM. Then, they later introduced the HMM in their system to analyze the features [118][119][120]. Alternate research in [121] explores the temporal segmentation of the lower face area by clustering the neighborhood frame which contain similar shapes and appearances features, then they iteratively group those clusters into the temporal gestures. However, the current studies of this issue are still in an infant stage. Due to the complexity of the process, most of the facial expressions in this field are preselected as a standalone image sequence with a single label of one particular emotion class.

Fusion of affective information from other sensors As we have mentioned this issue at the beginning of this research. Emotional estimation can be inferred from various cues. In order to improve and calibrate the estimation of emotional parameters, the fusion of affective information from other sensors is suggested.

Beyond basic emotion To the end of this dissertation, since temporal feature is generally superior over the still image usage in term of richer information, we can apply our proposed temporal feature to investigate the facial expression beyond the basic emotion model by introducing the mixtures of emotion expression in physical form of facial muscles, and the interpretation of emotion in dimensional model. The discussion is provided in the Appendix A.

Appendix A

Beyond basic emotions

In this chapter, we further discuss about an insight of the complex emotions issue in computer vision. We define the term complex emotion as the affective states that do not belong to the Ekman's basic emotion sense. To emphasize on this point, we can show the emotion representation and its implication in Fig. A.1.

A.1 Emotion Representation

To understand the nature of human emotion, scientists have been studied emotion topic over hundred years. It appears that there are two main fractions of the emotion models; discrete model and dimensional model. The conflict of these two emotion theories is still under debates in the psychology field .

The theorists in the first group believed that emotion model are universally understandable and can be discretely categorized. The theorem is based on the idea of universality i.e. everyone can express and understand facial expression of the basic emotions in the same way. Led by Ekman et al.'s works, they categorized the key emotions into what so-called *basic emotion*. The definition of *basic emotion* in discrete model refers to the significant emotional states that play an important role in human life. However, there is a problem of how to define emotion categories by linguistic approach. Since most of the them are theorized by a set of English words, an emotion word from one language may not available in another language. In addition, language and perception of emotion may different for each person. Lindquist et al. demonstrated that language can influence the perception of emotion that a person may see from the face of another [33].

Dimensional model is offered to fill the loophole in discrete model by parameterizing emotional states in term of dimensions. This model represents the *basic emotions* as atomic units of emotional states and cannot be divisible into smaller elements. Barrett urged that the valence is a basic emotions of life [122]. The most well-known dimensional models is Circumplex model, which defined emotions into two dimensions of Valence and Arousal. Valence (or Evaluation) dimension represents how much positive or negative of our feels. Arousal (or Activation) measures the activeness/passiveness. Other dimensional models may expand this 2D model by including other parameters such as Power, Expectation, and Intensity.

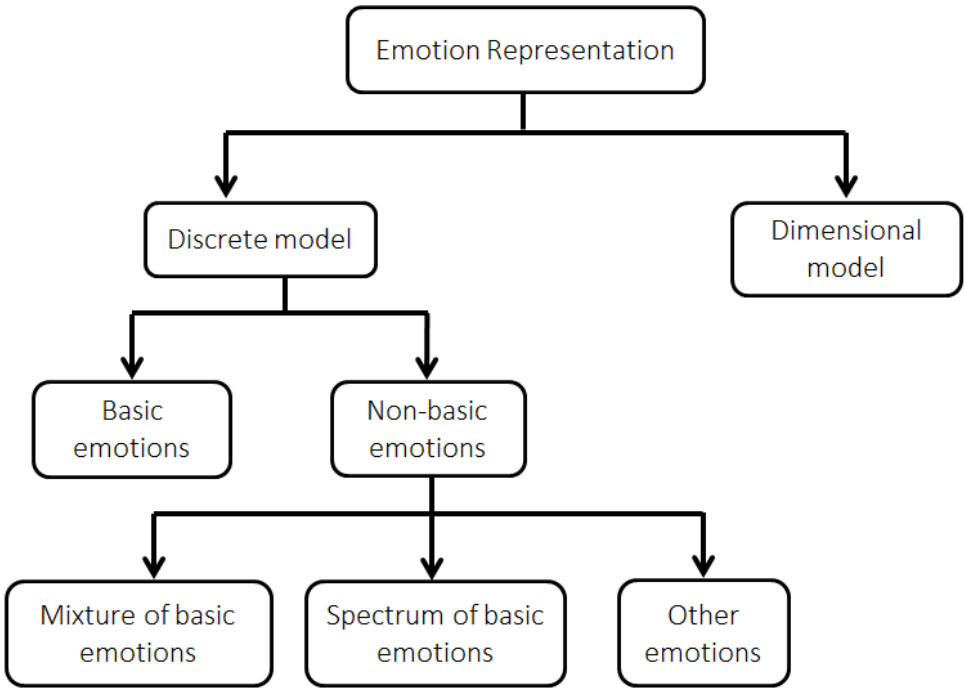


Figure A.1: Emotion representation

A.2 Modeling of complex emotion by discrete models

In order to determine the complex emotion based on discrete model assumption, it can be represented as the *mixture* of two or more basic emotions, *spectrum* or sub-categories of a particular basic emotion, and *other emotions* beside the Ekman's categorization.

However, the explanation on the *spectrum* approach has not been supported by the existing evidences in the affective computing field. Similar problem has been also found in the *other emotions* approach too. In this case, Jack et al. suggested a fewer numbers by 4 basic emotions (anger, fear, happiness and sadness) [123], Plutchik's psychoevolutionary of basic emotions proposed eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy [124]. Although there are applications of other non-basic emotion such as pride [125], tension, relaxation, exciting [126], elation, hot anger (rage), cold anger (irritation), despair, pleasure, relief, interest, anxiety [127]. However, all of the above affective states was considered from other modalities but not face, such as bodily gestures and sound. It is difficult to define how many discrete emotions of human can possibly have, and which one is measurable by face. Thus, we suggest to consider the complex emotion based on discrete model by only mixture of basic emotions in this research.

A.2.1 Existing of emotion mixtures

Regarding the consideration of mixed emotions, the first question is Can the human has two emotions at the same time? Russell and Barrett [128] urged that happiness and sadness are mutually exclusive (cannot happen at the same time). Since in Russell's circumflex model placed the happiness and sadness on the extreme opposite side of valence axis. However their model was created by asking their subjects to group and sort similar

emotional words together. So it is normal that subjects did not group happiness and sadness in the same group. Lasen and McGraw [129] debated this issue through a series of their studies and concluded that the mixture of happiness and sadness do exist.

By given an assumption that basic discrete emotion is valid, Plutchik postulated that other emotions beside his primary emotions can be mixture of these emotions. For instance, anxiety is considered as a mixture of anticipation and fear, shame is a mixture of fear and disgust [124]. Picard suggested that the categories of emotions can be fuzzy in the sense that one emotion can be divided into other emotion categories, or it can belong to more than emotion at a time. For example, sadness can be explained in form of both grief and melancholy [130][1].

A.2.2 Mixture of emotions via facial expression

In order to consider the mixture of basic emotions by facial expression, we suppose emotion expressions as message signals in communication. The mixed emotions phenomenon can occur when peoples try to cover their real emotion, or intentionally sending a mixed emotion signals in the social context. For example when your friend tells you a dirty joke, one may response with a smile with wrinkles on the root of nose indicating the disgust expression at the same time. To this approach, we can define mixture of emotions as a mixed expression between two or more different set of facial action units that reflects Ekman's basic emotions.

For the this approach, it has been shown in the Du et al. work that the mixture of basic emotions can be represented by the mutual occurrences of the AUs corresponding to two different emotion expressions [42]. However their system cannot recognize the subtle expression of mixed emotions due to the usages of geometric feature from a still image. As we have mentioned earlier, temporal features are superior over the still image usage in term of richer information. It is suggested to investigate the mixed facial expression based on our proposed system in the previous chapters.

A.3 Modeling of complex emotion by dimensional models

The literature review about the implementation of dimensional models in their automatic emotional recognition system is given in Table. A.1.

[1]Grief is a response to a loss of someone or something. Melancholy a feeling of long-term sadness with no obvious cause.(Oxford Advanced Learner's Dictionary)

Table A.1: The review of the previous works that used dimensional model in their system. Notations of emotion cues are marked by: F = Face, H = Head, S = Shoulder, Au = Audio, G = Gestures. Notation for dimensional elements is marked by V = Valence/Pleasure, A = Arousal/Activation, E = Expectation, P = Power/Dominant, I = Intensity. The hyphen sign '-' indicates that the information is unavailable, or unable to determine.

Name	**Conversion methods**	**Continuity**	**Cue**	**Face Image Features**	**Dimensions**	**Dataset**
Du et al. [131]	Least square method	Yes	F	SHP, intensity, LBP	V, A	JAFFE, CK, TV series
Hakim et al. [132]	Distance to basic emotion + Gaussian mixture	Yes	F	feature points	V,A	IEMOCAP
McDuff et al. [133]	SVM, CRF, LDCRF (latent dynamic conditional random field), HMM	-	F	AUs	V	their dataset
Nicolle et al. [134]	Nadaraya-Watson estimator (Kernel regression)	Yes	F, Au	Shape parameter, Whole face, Local area	V, A, P, E	AVEC2012 (SEMAINE)
Zhang et al. [135]	SVR	Yes	F	Gabor, LBP, FAPs, combine	V, A	NVIE, FEEDTUM, SEMAINE
Nicolaou et al. [136]	HMM, HMM+SVM	No	F, S, Au		V	SAL
Nicolaou et al. [137]	SVR,BLSTM-NN	Yes	F, S, Au	Feature points,	V, A	SAL
Nicolaou et al. [138]	Output Associative RVM	Yes	F, S, Au	Feature points	V, A	SAL
Nicolaou et al. [139]	RVM	Yes	F, Au	Feature points	V, A, P, E, I	SEMAINE
Baltrusaitis et al. [140]	CCRF	Yes	F, Au	Feature points, temporal LBP	V, A, P, E	AVEC2012 (SEMAINE)
Eyben et al. [141]	SVR	Yes	H, F, Au	motion	V, A, E, P, I	SEMAINE

Continued on next page

Name	Conversion methods	Continuity	Cue	Face Image Features	Dimensions	Dataset
Sanchez-Lozano et al. [142]	SVR	Yes	F, Au	LBP, Gabor	V, A	AVEC2013 (SEMAINE)
Wollmer et al. [143]	LSTM-RNN, CRF, SVR	Yes/No	Au	-	V, A	SAL
Wollmer et al. [144]	LSTM, BLSTM	No	F, Au	segmented optical flow, head tilt	V, A, P, E	AVEC2011 (SEMAINE)
Grimm et al. [145]	neural network + fuzzy logic	Yes/No	F	Gabor (Eye and Mouth area)	V, A, P	VAM
Grimm & Kroschel [146]	SVR, Fuzzy K-Nearest, Rule-based Fuzzy Logic estimator	Yes	Au	-	V, A, P	VAM
Kanluan & Grimm [147]	SVR	Yes	F, Au	DCT	V, A, P	VAM
Caridakis et al. [148]	modified Recursive Neural Network	No	F, Au	FAPs,	V, A	SAL
Chi-Chun Lee et al. [149]	DBN	No	Au	-	V, A	IEMOCAP
Glowinski et al. [150]	Rule based	No	G	-	V, A	GEMEP
Glowinski et al, 2011 [151]	Bootstrapping classification	No	G	-	V, A	GEMEP
Kipp & Martin [152]	Correlation	-	G	-	V, A, P	movie
Fewzee & Karray [153]	CC, HSIC	Yes/No	Au	-	V, A, P, E	AVEC2012 (SEMAINE)

A.3.1 Dimensional parameters

It is obvious that the common dimensions that have been used in most studies are *V-A*: *Valence/Pleasure* and *Arousal/Activation* elements. The simpler model by using only valence dimensions is also found in [133][136]. This V-A model has contributed as the most fundamental components of human emotion in psychology study for a long time.

The formation of two dimensional model is usually illustrated their relation in Cartesian coordinates. Circumplex model has been formed and discussed by many scientist such as Plutchik [154], Russell [155], Watson and Tellegen [156]. Beside the famous V-A model, the combination of V-A-P (Valence, Arousal, Power) or being called as the PAD (Pleasure, Activation, Dominance) [157] is also often used. The additional dimension is Power/Dominance which defines the controlling of dominance or submissive feeling. Moreover most of the works on dimensional models neglect another two dimensions: Expectation and Intensity. In this research, we also consider these dimensions in our recognition system.

A.3.2 Implementation methods in machine learning

In order to develop an automatic facial expression recognition system for recognizing dimensional emotions, it is required to determine the conversion methods of image features to dimensions available in the previous studies. Typically, the conversion from input feature to dimensional emotion parameters are divided into two approaches: *classification* and *regression*.

In these early studies, they used the *classification* approach to represent the conversion from face (or other modalities such as audio or bodily gesture) into dimensional emotion classes. The outputs are simply discrete values and mostly they are binary outputs (e.g. 0/1 for Positive or Negative, Active or Passive). In this classification approach, the learning problem focuses on how to optimize the boundary decision or subspace that best *separate* all input data into different classes.

Instead, the recent works focused more on conversion of face onto continuous dimensional axis due to more convenient usages in real-world application. In this case, the learning problem becomes *regression* approach, which attempts to optimize a function to *fitting* all input data to continuous dimensional parameters. The famous regression methods are Support Vector Regression (SVR) [135]-[143], various types of neuron networks [145]-[144]. Recently a series of studies from Nicolaou et al. introduced Relevant Vector Machines (RVM) in their system [137]-[139].

According to the previous works, we noticed that most of the facial expression recognition systems from dimensional model have the flaws on the following points:

- Almost of the previous works extracted either geometric or texture features from a still image. The drawback of using a still image feature has been stated throughout the contents of this dissertation.

- The usages of Action Units (AUs) in the implementations of dimensional model are limited and need further investigation on this issue.

Thus, we proposed a new facial expression recognition of dimensional parameters by using our action units detector to recognize AUs (*Publication. VI.*), and then we use Support Vector Regression (SVR) to estimate an emotion vector in dimensional emotion space (Fig. A.2).

In our preliminary experiment, we utilized a portion of SEMAINE dataset [158] (3,836 frames) with the given the emotion labels. We recognized AUs and use as the observation data. Then we used the Support Vector Regression (SVR) with RBF kernel for the regression. Hyper parameters of SVR has been tuned by 5 fold cross validation. For evaluation, we split half of data for training set and another half for test set.

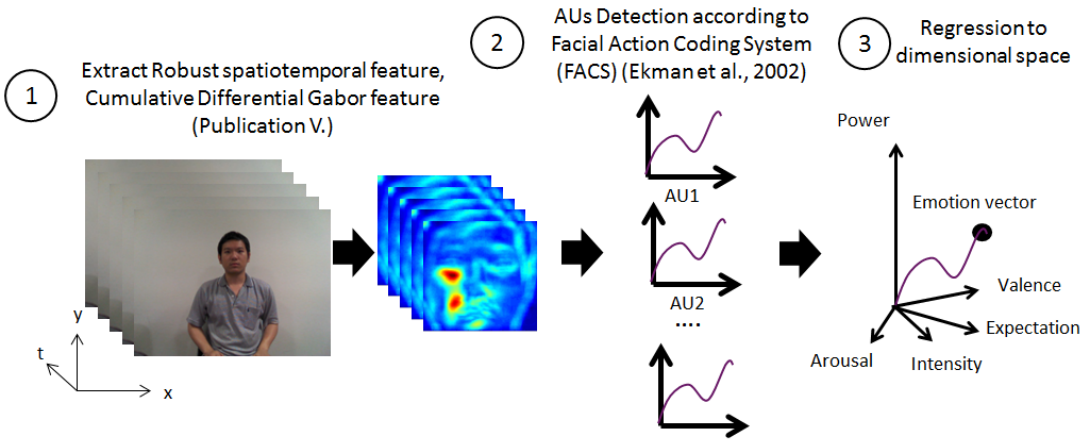


Figure A.2: Proposed dimensional emotion estimation framework

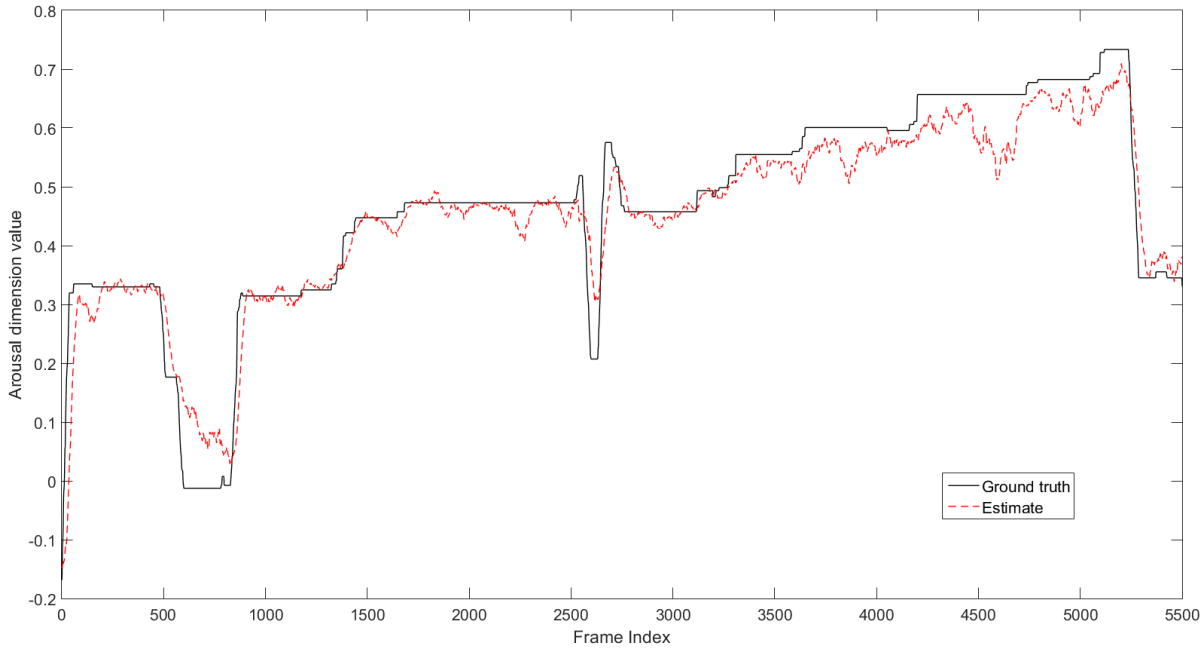


Figure A.3: Estimation of arousal parameter by using recognized AUs and SVR comparing to ground truth.

The mean square error (MSE) of the emotion estimators in each dimension has been shown in Table. A.2. The example of estimation is shown in Fig. A.3.

Table A.2: Mean square error (MSE) of testing set comparing to ground truth

	A	V	P	E	I
MSE	0.0102	0.0107	0.0110	243.5876	0.0017

Dimensional model approach can be viewed as the projection of our high dimensional complex emotion into lower dimensional space. Since most of the emotional studies in computer vision is dominated by the Ekman's basic emotions category, It is interested to see if it is possible to archive an automatic dimensional emotion recognition system

based on facial action units (AUs). In above preliminary experiment, we used our robust AU detector which can capture the subtle actions and overcome several outliers, and then we used SVR to model the predictor of dimensional emotion parameters. The preliminary experiment shows the prominent potential of the proposed method for estimating dimensional emotion parameters. However the result of Expectation dimensions shows distinctively high error, it may require a further investigation whether the result in this paper indicated that facial expression is not a suitable cue to measure the Expectation dimension, or there are insufficient samples to model a predictor.

Appendix B

Supplementary experiments

This appendix chapter provides the supplementary studies on the parameter tuning of the classification's hyper-parameters

B.1 Number of independent components

The number of independent components is evaluated in Fig. B.1. We varied the number of independent components and measured the average precision of the classification by ICA with whiten EMC. The wavelength ranges are set at the *All* range. The classification performance reached the convergence after rising rapidly at a few iterations.

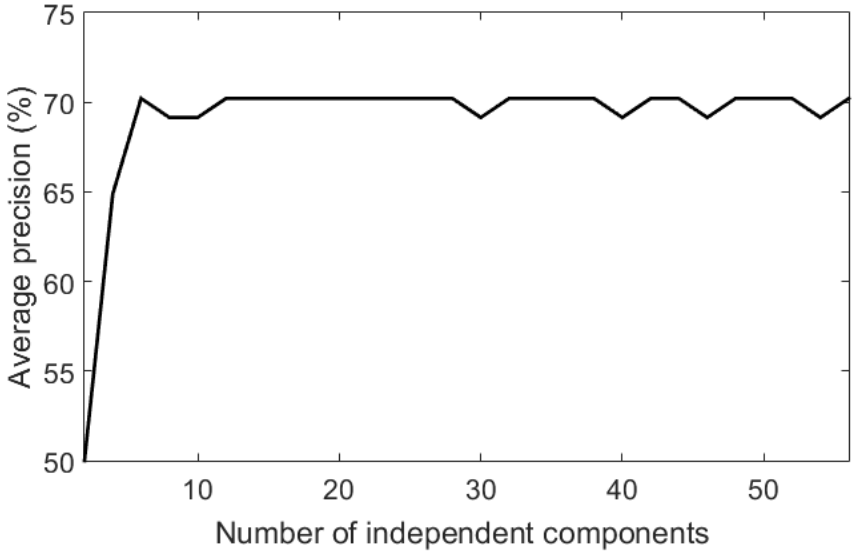


Figure B.1: The varying number of independent components versus the average precisions of the proposed CDG (superposition representation).

B.2 Hyper-parameter tuning

The polynomial degree of the kernel p (KEMC) and the number of nearest neighbors k (KNN) are chosen by empirical approach (see Figs. B.2-B.4 for the references).

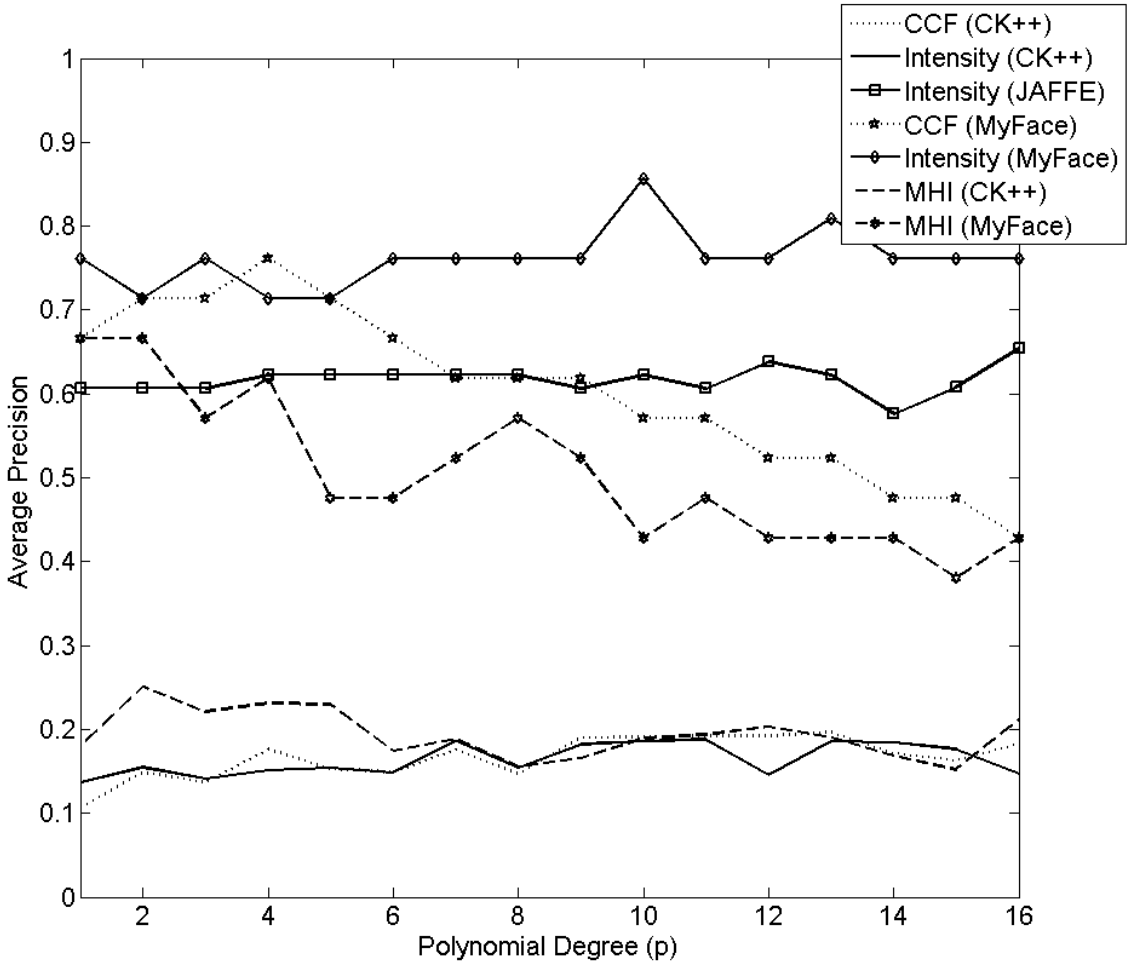


Figure B.2: Average precision of KEMC by different polynomial degree p of the kernel function

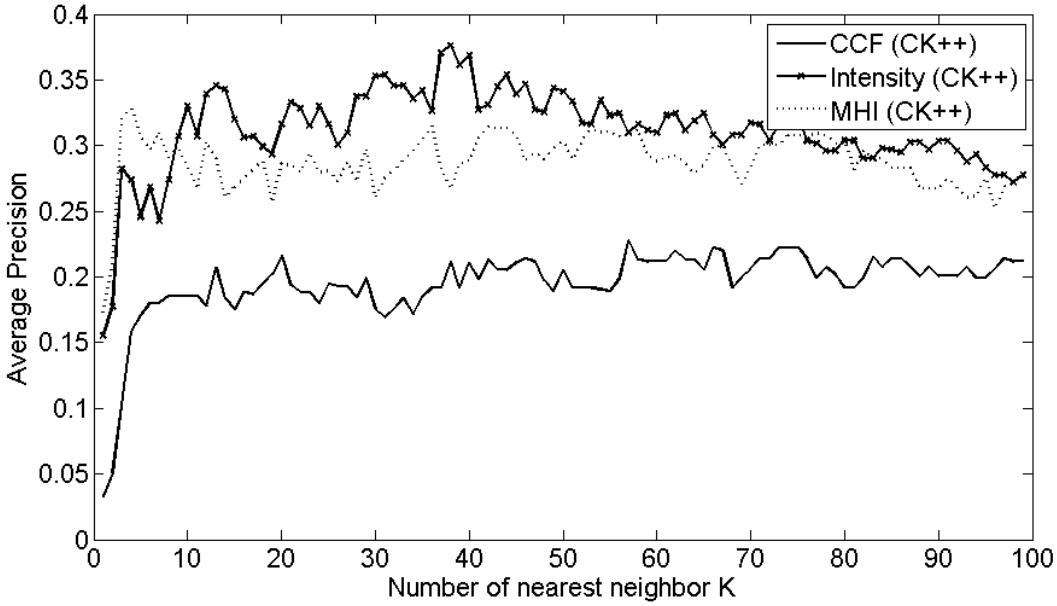


Figure B.3: Average precision of KNN in CK+ dataset versus the number of nearest neighbor k

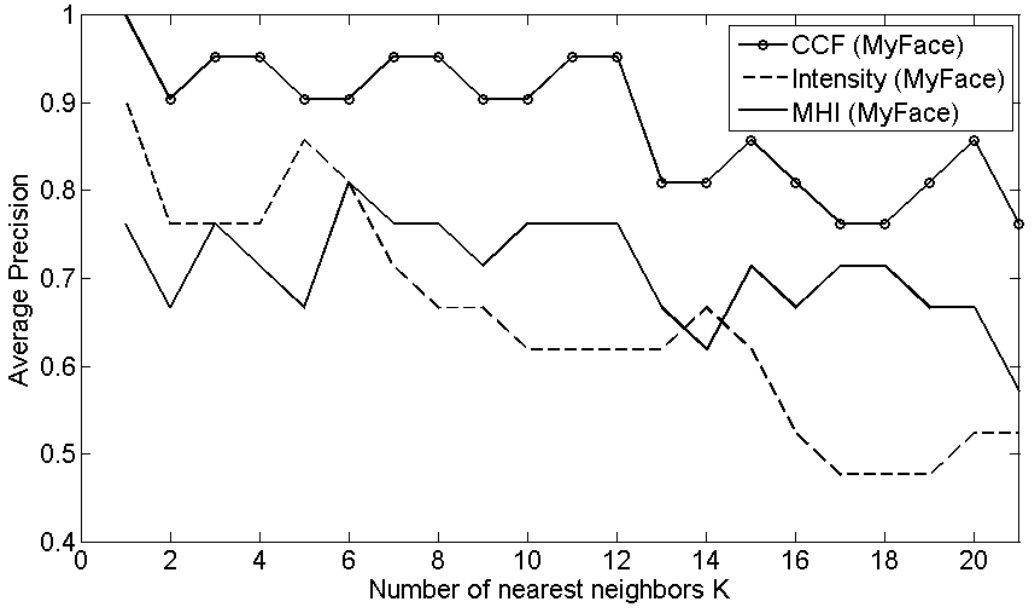


Figure B.4: Average precision of KNN in MyFace dataset versus the number of nearest neighbor k

Appendix C

Validity of Facial Action Coding System (FACS)

Regarding the validity of Facial Action Coding System (FACS), it is a curiosity of how Ekman et al. created FACS, and how much credibility of their system. Therefore, I will discuss these issues here.

Development of FACS To derive the minimal action units, Ekman et al. used two methods:

1. *Self-observation*: Ekman et al. follow the method from Hjortsjo [159] by observing their own face as they mentioned in the investigator guide, "...Following Hjorstjo's lead, we spent the better part of a year with a mirror, anatomy texts, and cameras. We learned to fire separately the muscles in our own faces ... By feeling the surface of our faces we could usually determine whether the intended muscle was contracting...." ([9], p.5)

2. *Direct electrical stimulation*: Ekman and Friesen asked a doctor to physically intrude the needle directly to a specified muscle. Then, they electrically stimulate the muscle and see the changes on facial surface. This experiment is the replication from Duchenne's method [160]. The verification is run by voluntary activating the target muscle and measured by EMG (both invasive needle and surface EMG). This procedure was mentioned in FACS investigator guide, "... Ekman and Friesen also resurrected Duchenne's (1862) technique of determining how muscles change appearance by inserting a needle into and electrically stimulating muscles.... " ([9] p.106), and "... Ekman and Friesen's use of fine-wire EMG to stimulate and record facial movement in order to discover how the muscles work to change appearance... ([9], p.123).

As they can map the muscles activations to the changes on facial surface, they learned to voluntary activate specific muscles, and capture images and videos of facial actions, and use as the material to train other observers.

Evidences on the validity of FACS To prove that FACS really measures the actual muscle activation. Ekman et al. claimed two approaches.

1. *Ask the people* who was trained to use FACS to watch images/videos, and score Action Units (AUs) without knowledge of the performed actions. This is the indirect verification to show to consistency of the method. For the highly trained people, they can archive very accurate read of action units from an unknown subject and his/her actions.

2. *Measurement by EMG.* The more precise verification of the association between face appearances and muscle activation corresponding to FACS is EMG. The numerical result on relation between FACS and EMG provided in the FACS investigator guide , " ...FACS scoring was later found to be highly correlated with the EMG readings (Pearson r = 0.85)... ", ([9], p.123).

Unfortunately, the further detail of experimental result about the relation between EMG and FACS is not publicly available. We inevitably have to refer to the Pearson's r = 0.85 for primary evidence as Ekman mentioned many times in his later papers [161].

Beside the above evidences, the reliability for FACS coding from different observers has been tested across several laboratories [162][163][164]. These works could be considered as the secondary evidences.

Appendix D

List of existing works using FACS framework

Table D.1: List of existing works using AUs framework

Paper	Number of AUs	Emotion Labels	Feature Extraction method	Classification method	Geometric or Texture feature	3D or 2D	Temporal or static analysis	Precision rate
Donato et al., 1999 [165]	16 (1, 2, 4, 5 ,6, 7 ,17, 18 , 9 + 25, 10 + 25, 16 + 25,20 + 25)	Single camera	Optical flow, Eigen, PCA, LFA, Fisher, ICA, Gabor	Nearest Neighbor, Euclidean distance	Texture	2D	Static	>53%
Tian et al., 2001 [66]	16 (1, 2, 4, 5, 6, 7 , 9, 10, 12, 15, 17, 20, 25, 26, 27, 23 + 24)	-	Multistate facial component by geometric feature (angle, distance, state) of local parts using template matching and edge detector	Neural Network	Geometric	2D	Static	>88%
Bartlett et al., 1999 [166]	6 (1, 2, 4, 5, 6 ,7)	Single camera	Optical flow, Eigen, PCA, Fusion	Neuron network	Both	2D	Static	92%
Bartlett et al., 2005 [75].	17 (1, 2, 4, 5, 6, 7, 9 ,10 ,12, 15, 17, 20, 23, 24, 25, 26)	-	Gabor filter	SVM ,Adaboost, LDA	Texture	2D	Static	93%

Continued on next page

Paper	Number of AUs	Emotion Labels	Feature Extraction method	Classification method	Geometric or Texture feature	3D or 2D	Temporal or static analysis	Precision rate
Bartlett et al., 2006 [109]	19 (1, 2, 4 ,5 ,6 ,7, 9 ,10 11 ,12 ,14 ,1 5 ,16 , 17, 20 , 23 ,24 ,25 ,26)	-	Gabor filter	SVM ,Adaboost	Texture	2D	Static	>67%
Tong et al., 2007 [110]	14 (1, 2, 4, 5, 6, 7, 9 ,12, 15, 17, 23, 24, 25, 27)	-	Gabor wavelet	Adaboost, Dynamic Bayesian network	Texture	2D	Temporal	-
Mahoor et al., 2011 [108]	5 (1 + 2 + 5 + 27, 15 + 17 , 6 + 12 + 25 , 4 + 9 + 17 + 23, 20 + 25)	-	Gabor filter, Sparse representation	SVM, NN	Geometric	2D	Static	>85%
Yang et al., 2007 [78]	8 (1, 2, 4, 5, 10, 12 , 14, 20)	6 (Anger, disgust, fear, happiness, sadness ,surprise)	Haar-like feature, coded dynamic feature	Adaboost	Geometric	2D	both	-

Continued on next page

Paper	Number of AUs	Emotion Labels	Feature Extraction method	Classification method	Geometric or Texture feature	3D or 2D	Temporal or static analysis	Precision rate
Tsalakanidou et al., 2010 [67]	11 (1, 2, 4, 5, 7, 12, 15, 25, 26, 27)	4 (Disgust, Happy, Sad, Surprise)	Geometric feature (Euclidean distance between landmark)	Rule based classification	Geometric	Both	Static	>73
Sun et al., 2008 [116]	8 (1, 2, 4,5, 15, 20 , 27, 1+2)	6, (Anger, disgust, fear smile, sadness, surprise)	AAM tracking	6-state HMM	Geometric	3D	Temporal (multi-state motion)	>69%
Savran et al., 2012 [114]	25	-	Gabor Wavelet, ICA, NMFSC, Diff. ICA, Diff. NMFSC	AdaBoost, RBF SVMs. Nave Bayes	both	both	Static	>90%

Appendix E

Categorization of facial expression image features

In this appendix chapter[1], we determine the categorization of facial expression of a person in the frontal face dataset by using topic model technique. The work in this chapter shares the mutual objective of clustering analysis, which attempts to group the similar feature together in the same cluster. It is importance to note that there is significant differences between typical face clustering (which tends to categorize the person identity), and facial expression clustering, which contain less variation of the feature distribution.

Although a large amount of facial expression research has been conducted in computer vision domain, the problem is limited to recognize only 6-7 emotion classes according to Ekman's basic emotion categorization [22]. The facial expressions of the basic emotions are defined by the specific muscle movement according to the Facial Action Coding System (FACS) [2]. However, the facial expressions in our life basis are not confined to a limit set of basic emotions. There are much more variance of the facial expression which not follow the Ekman's rules. By this limitation, we are interested in categorization of facial expressions from video sequences to explore more categories of facial expression. This kind of research problem belongs to the clustering analysis domain, which will be discussed in the next section of this article.

E.1 Related Works

Facial expression clustering is a problem of grouping faces by feature similarity. Most literatures related to the clustering of facial features focused on the person identities issue, in which similar to a face recognition task.

Instead, this chapter focuses on the clustering of facial expression. Assume that we neglect the person identity issues by observing facial behaviors of only one person at a certain period. The challenge of this research is the smaller dissimilarities between each facial expression images comparing to the face clustering of different person. Therefore, this handicap yields less diversity of sample distribution, and becomes harder to clarify the cluster separation.

Since, there are massive amounts of works on the clustering analysis. The review in this article will only present the methods from the computer vision community that related

[1]This article is a part of minor research under the supervision of Prof. Ho Tu Bao *(Publication IV)*

to the face clustering issue. Based on the frequently used models, we can categorize the clustering type among the existing research as follows:

Centroid based clustering represents each cluster by its mean vector such as K-mean clustering algorithm or Fuzzy c-means algorithm. It is one of the most common method used in the computer vision community due to its simplicity in implementation. However, the limitation of K-mean algorithm and its family are the requirement to determine the proper number of clusters. Additionally, the algorithms cannot determine the non-convex and non-linear distribution of the cluster. To avoid the initialization of the cluster number, one may apply the mean shift algorithm [167] that has a similar intuitive to the K-mean clustering, but it does not assume anything about the number of clusters, or the shape of clusters. Moreover, an alternative centroid based is also found in [168] by applying the Expectation Maximization (EM) algorithm to reduce the computational complexity.

Another de facto clustering method besides centroid based clustering is hierarchical clustering method. This method creates a hierarchy of clusters that merge smaller clusters into a bigger cluster (Agglomerative clustering), or splitting one to smaller clusters (Divisive clustering) recursively until reach a certain threshold. The agglomerative clustering has been widely used in face clustering applications more than the divisive one e.g. the implementation in [169] that utilized the agglomerative hierarchical clustering form the clusters from the similarity between SIFT features. Similar approach has been also found in [170] with some constraints to verify a non-duplicated cluster in multi-person images. Furthermore, the method in natural language processing has been applied with the hierarchical clustering method by representing the visual information in the bag-of-words form [171].

In the recent works of face clustering, subspace clustering method are applied to overcome the curse of dimensionality in a higher dimensional space since it is difficult to analyze and visualize the cluster of the data in the original space. Subspace method reduces the dimension of data such that we can separate the cluster on the fewer subspace bases. One of the simplest subspaces is Principle Component Analysis (PCA). The generalized PCA that was utilized with a face application is found in [172]. The improved version from the same author was introduced by the hybrid between centroid based method and subspace method [173]. Another recent subspace clustering methods are based on spectral clustering technique [174][175]. The method utilizes the spectrum of the similarity matrix of data to reduce the dimension of data and cluster without any assumption about the shape of the clusters. Based on the spectral clustering method, we found the work in [176] measures the mutual information of face images by their entropy as the distance function. The sparse representation of the spectral clustering are presented in [177]. Furthermore, the combination of centroid based clustering and spectral clustering is found in [178].

Beside the above clustering methods, there are other implementations in face application which cannot be categorized by the above clustering types. To cluster face in the video sequences, we can utilize the spatio-temporal relation to improve the clustering such as smoothing the segmentation of the consecutive frames by affinity propagation [179], define as a probabilistic constraint in a generative graphical model by Hidden Markov Random Fields [180], or one may uses the combination of multiple-cue data and probabilistic maximum a posteriori (MAP) to estimate the clusters [181].

Alternatively, we can also improve the clustering performance by utilizing the context information such as clothes, hair [182], captions or names [183], people co-occurrence [184], relation grouping in social network [185], or combine with SIFT feature point to

capture an individuality characteristic [186]. To overcome the head pose problem, [187] perform a dual clustering by firstly performing the pose clustering, and then perform texture clustering to identify the individuality. However, these kinds of attribute are usually suitable only for person identification, which are not involved with facial expression analysis.

Many of the previous research established clusters by a deterministic model, which fail to generalize the clustering. Topic model is a probabilistic data associated model often use in text processing field to categorize a label of data in the collection e.g. suggest keywords of a book or document to a user. It is also often known as the collaborative filter in recommendation system which predicts the categorization of user based on their behavior information.

Basically, topic model represents the document by the topic distribution, which can also be seen as the feature vector in a reduced dimension space. It is an alternative way to exhibit the similarity of the documents in the same sense to traditional clustering. In addition, the topic model has the advantage over the other traditional clustering methods by not assuming that the data must be belonged to only one cluster, but instead, it is modeled as a mixture of topics. Therefore, there is more room for considering other choice of categories. Therefore, many computer vision research utilized topic model as the categorization technique for example the scene categorization [188], and object categorization [189][190]. The existing research related to facial expression has been found in [191]. They attempt to group the co-occurrence of Action Units (AUs) and create the meaningful categories of AUs. However, they did not derived any visual information. The only work that applied the topic model with the facial expression images was presented in [192]. Unfortunately, they utilized topic model as the classification method in similar way to support vector machine, which is supervised learning. In addition, their representation of visual information was defined as the motion of a few landmark points, which neutralized most of the spatial information.

In this research, we present a novel idea of utilizing the topic model technique for unsupervised clustering facial behavior. Up to our knowledge, there a limited number of the present literatures focused on the facial expression clustering. None of them apply the topic model technique in this application.

E.2 Topic model

One of the most famous topic modeling algorithm is the Latent Dirichlet Allocation (LDA)[2] proposed by Blei and his colleagues in 2003 [8]. The algorithm is derived from the probabilistic latent semantic indexing (pLSI) [193] by introducing the Dirichlet distribution to generalize the topic distribution of each documents and words. It is widely used in learning from texts and documents for finding the significant keywords that can describe a particular document. In this research, we implement LDA for unsupervised learning from facial expression. Since LDA has been originally proposed for processing the text data, there is no direct representation for visual information as we can archive on the text data by looking up on the dictionary. Therefore, we need to convert the image features into the words via some representation. This issue is discussed in the subsection

[2]The abbreviation of Latent Dirichlet Allocation (LDA) is used only in this appendix chapter. For other chapters, LDA refers to Linear Discriminant Analysis

E.2.1. In the subsection E.2.2, we describe the probabilistic model and inference of LDA. In this research, we utilize the Variational Bayes inference method to infer the hidden parameters. Then, we utilize this generative model to determine the topics of a particular document.

E.2.1 Representation of visual information

Basically, LDA defines four components:

- **word** is an atomic unit
- **document** is the collection of words
- **topic** is the prominent word that can represent the whole documents
- **dictionary** is the collection of all words in all documents

Although LDA is designed for the arbitrary discrete data, the characteristic of texts and images are significantly difference. Since we cannot define the image as the word or document directly in similar way to text document, we need to convert the image features into words.

In order to do that, the intermediate representations are required. Feifei and Perona represents the patches in the images as the words to recognize the context of an unknown scene [188]. The patch has the structure similar to the biological perception. However, their model is supervised by the labeled theme parameters for each word. In addition, it is question that how many number of image patches we should defined as the words in the dictionary, and how could we match an image patch to a particular word in the dictionary, whether they are the same or not. The similar representation in Wang and Grimson's work [189] is also suffered from such the issues.

Furthermore, the existing research related to face images has been found in Shang et al. paper [192]. They applied the topic modeling for recognizing the facial expression. However, their method utilized the action of landmark points as the representation for the words, which neutralized most of the spatial information. Such a representation is not sufficient to understand the complexity of facial expression. Therefore, we propose a new representation by using the quantized gradient orientations of an image in the next subsection.

Proposed representation

The procedure of drawing the topic from an input image is described in Fig. E.1. Given an arbitrary input image, we extract the gradient orientations, and then quantize the gradient orientation from 360 degrees into 60 bins. In this case, we can simply define the bins as the words, and the whole image as the documents. As the consequences, we count the number of occurrence of the bins in an image (see Fig. E.2). An example of gradient orientation in spatial domain is illustrated in Fig. E.3. Although this representation resembles the famous Histogram of Oriented Gradient (HOG) [96], we utilize the only a bin as the word's presentation instead of using the whole histogram. The histogram can be interpreted as the document. Then, we use this representation for discovering the structure of facial expression by LDA model.

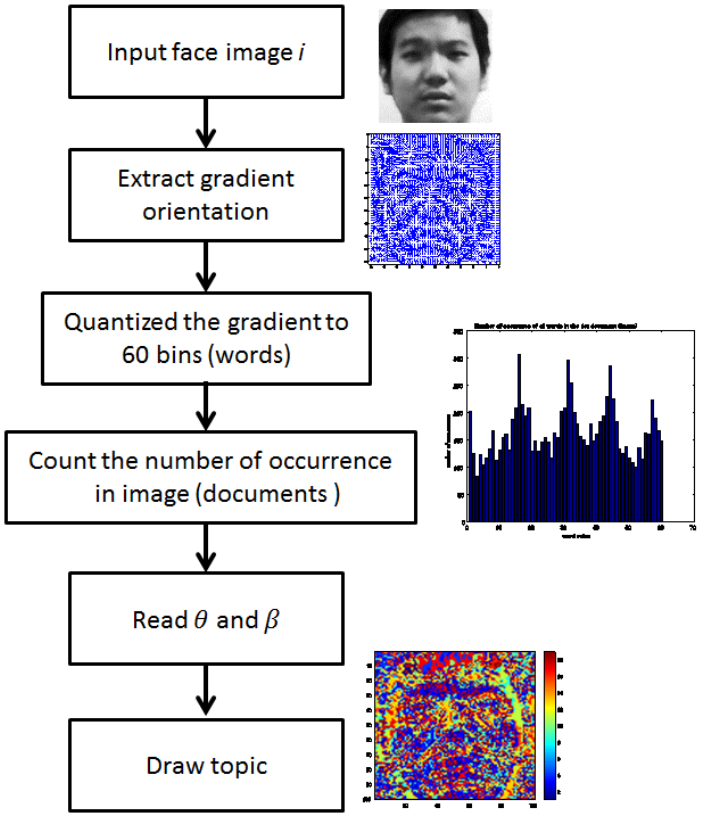


Figure E.1: The flowchart describes the topic modeling procedure from an input image

The motivation for using gradient orientation is due to the effect of the changing light conditions which result in different patterns of intensity and gradient magnitude. Gradient orientation is known to be insusceptible to the lighting condition changes. This advantage is well considered in many robust feature extractions. The gradient orientation extraction procedure is explained in the Chapter 3, section 3.9

Thus, we can redefine the four LDA components in this research as:

- **word** is a quantized gradient orientation at the position (x, y)
- **document** is an arbitrary image i.e. a collection of gradient orientations
- **topic** is a collection of the significant gradient orientation.
- **dictionary** is the look up table for the gradient orientation.

E.2.2 Latent Dirichlet Allocation (LDA)

Probabilistic model

The Latent Dirichlet Allocation (LDA) is a generative model for learning of the topics from a set of discrete data. LDA can be represented by the graphical model in Fig. E.4.

From the graphical model in Fig. E.4, we can write the joint probability of the all parameter as:

$$p(\beta, \theta, z, w) = \prod_{k=1}^{K} \beta_k \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_n|\theta_d) p(w_n|z_n, \beta) \right) \tag{E.1}$$

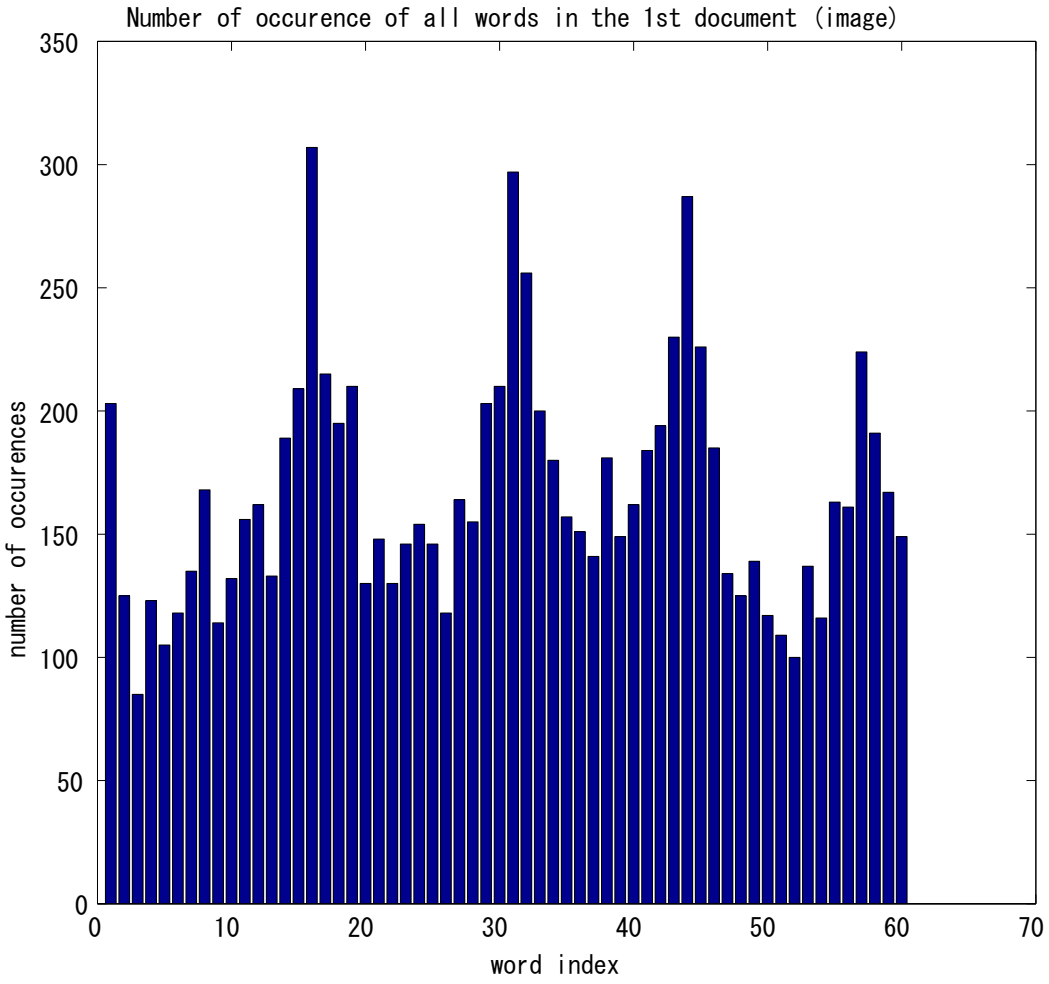
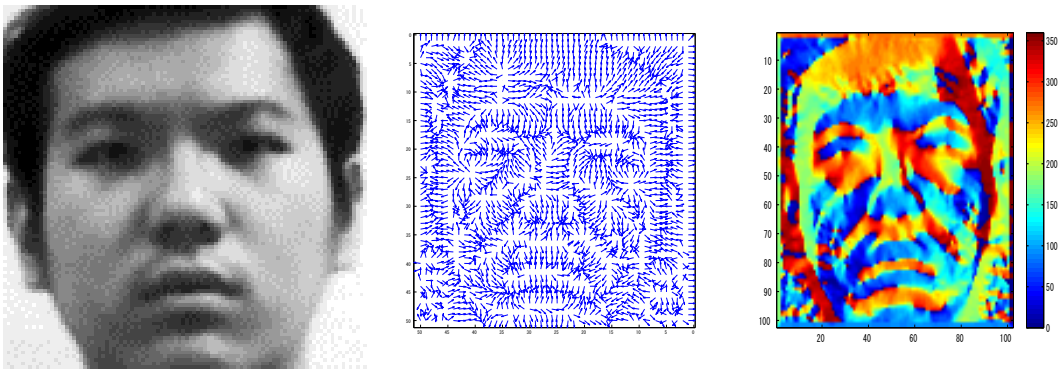


Figure E.2: Histogram of gradient orientations indicates the occurrences of the quantized gradient orientation in a particular bin. In order to simplify the dictionary, we defined the total number of bin equal to 60 bins. Each bins represents the words for LDA model.



(a) Reference image (b) Gradient vectors (c) Gradient orientation

Figure E.3: The gradient information of a (a) reference image can be visualized by (b) gradient in unit vectors form and (c) gradient orientations illustrated by colors from 0 (blue) to 360 (red) degrees.

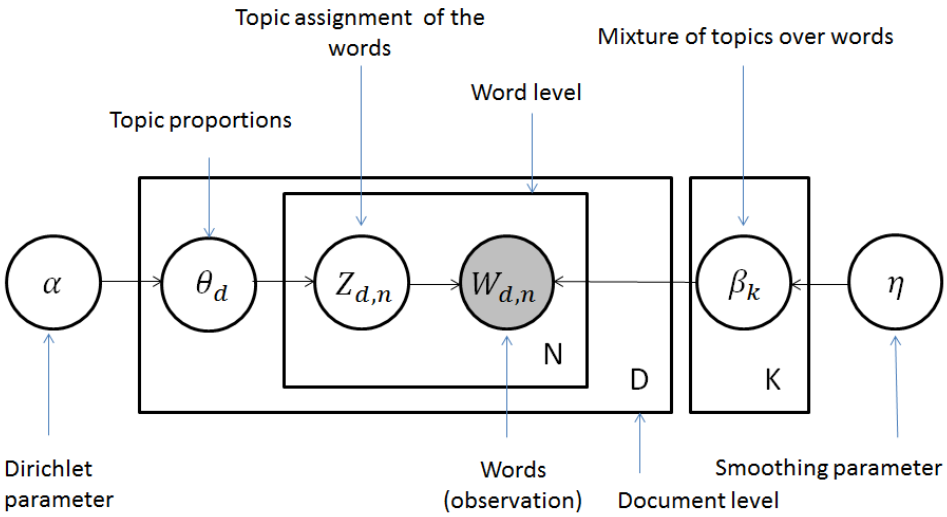
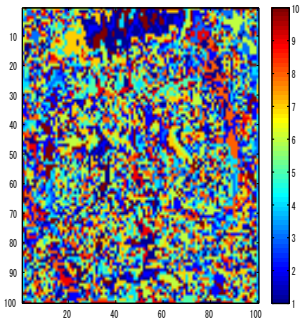


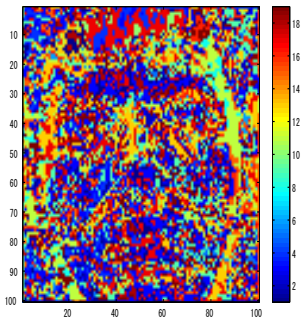
Figure E.4: The graphical model of Latent Dirichlet Allocation [8]. The observation is $W_{d,n}$ and the rest of them are latent parameters.

The notations of all parameters and the corresponding descriptions of the LDA model are shown in the Table. E.1.

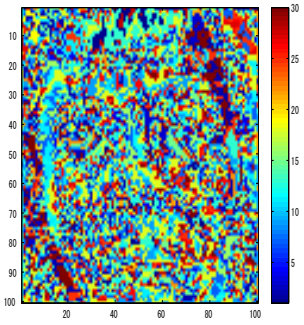
(a) Reference image



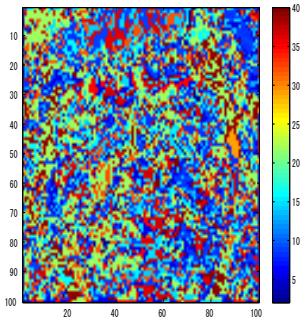
(b) 10 topics



(c) 20 topics



(d) 30 topics



(e) 40 topics

Figure E.5: Visualization of the topic assignments over a particular image (document) with the varying number of topics

Table E.1: The notations of the all parameters and the corresponding descriptions of LDA model

Notations	Description
d	document index
n	word index in a document
v	word index in the dictionary
k	topic index
D	Total number of document
N	Total number of words in a document
V	Total number of words in the dictionary
K	Total number of topic
$w_{d,n}$	the n^{th} word in the document d
W_d	A set of words in the document $W_d = \{w_{d,1}, w_{d,2}, ..., w_{d,n}\}$
$z_{d,n}$	topic assignment of the word $w_{d,n}$
Z_d	A set of topic assignments in the document $Z_d = \{z_{d,1}, z_{d,2}, ..., z_{d,n}\}$
θ	topic proportion (distribution) in the document d
α	Dirichlet parameter for modeling θ
β	Mixture of topics over words
η	Dirichlet parameter for modeling β

Suppose that the only observation data are the words $W = w_1, w_2, ..., w_N$ in documents $d \in D$, we can approximate the latent parameters θ and β by the inference methods in section E.2.2. Both of them is modeled by the by Dirichlet distribution as $\theta \sim Dir(\alpha)$ and $\beta \sim Dir(\eta)$. The Dirichlet distribution Dir($\boldsymbol{\alpha}$) has the probability density function (pdf) given by:

$$p(\theta_d \mid \alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1} \tag{E.2}$$

where the vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ determines the shape of the distribution over K topics.

The main problem of LDA is to approximate θ, β. and z by the following posterior distribution:

$$p(\theta, \beta, z|w) = \frac{p(\beta, \theta, z, w)}{p(w)} \tag{E.3}$$

However, the posterior distribution in Eq. (E.3) is intractable. Hence, we need an indirect method to infer the hidden variables. The inference method will be discussed in the section E.2.2.

In addition, the generative model can draw a words from the given model parameter α and β. The probability density function of the word w is the marginalization of the Eq. (E.1) over θ and z:

$$p(w|\theta,\beta) = \int \sum_{z} p(\theta, z, w|\alpha, \beta) d\theta \tag{E.4}$$

$$p(\theta, z, w|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_n|\theta_d) p(w_n|z_n, \beta) \right) \tag{E.5}$$

In opposite direction, from the mixture of topics over document θ_d and the mixture of the topics over a words $\beta_{k,v}$, we can assigned the topics to each words. Therefore, we can draw the topic assignments of the document as shown in Fig. E.5.

Variational Bayes inference method

Variational Bayes inference method is the inference method based on Bayesian method to infer the latent variables from the observed data. Although we cannot solve the intractable posterior in Eq. (E.3), we can approximate the lower bound of the marginal likelihood $\log p(w|\alpha, \beta)$ of the observations according to the Jensen's inequality [8]:

$$\log p(w|\alpha, \beta) \geq E_q[\log p(\theta, z, w|\alpha, \beta)] - E_q[\log q(\theta, z)] \tag{E.6}$$

Suppose $q(\theta, z, \beta|\gamma, \phi)$ is the approximate posterior probability noted by:

$$q(\theta, z, \beta|\gamma, \phi) = \prod_{i=1}^{K} q(\beta_k|\lambda_k) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}) \tag{E.7}$$

In this problem, we fit the approximate parameter q to the true posterior p by minimizing the Kullback-Leibler divergence. We can estimate the variational parameter $\hat{\gamma}$ and $\hat{\phi}$ by:

$$(\hat{\gamma}, \hat{\phi}) = \underset{\gamma,\lambda,\phi}{argmin} KL(q(\theta, z, \beta)||p(\theta, z, \beta)) \tag{E.8}$$

As we proceed the optimize problem for minimizing Eq. (E.8), we follow the implementation in [8]. The update functions for each iterations are shown in Algorithm 1,

where the digamma function $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$, $n_{d,v}$ is the number of the word v in document d. Then, we can obtain the $\beta_{k,v}$, $\theta_{d,k}$, and $z_{d,n,k}$ by:

$$\beta_{k,v} = \frac{\lambda_{k,v}}{\sum_{v=1}^{V} \lambda_{k,v}} \tag{E.12}$$

$$\theta_{k,v} = \frac{\gamma_{d,k}}{\sum_{k=1}^{k} \gamma_{d,k}} \tag{E.13}$$

$$z_{d,n,k} = \phi_{d,n,k} \tag{E.14}$$

Data: words, counts, dictionary
Result: variational parameter γ and ϕ
initialization;
while *until convergence* **do**
  for *topics* k *and the word in the dictionary* v **do**

$$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)} \tag{E.9}$$

  end
  for *document* d **do**

$$\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} n_{d,v}\phi_{d,n,k}^{(t)} \tag{E.10}$$

$$\phi_{d,n,k}^{(t+1)} \propto exp(\Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\sum_{v=1}^{V}\lambda_{k,v}^{(t+1)})) \tag{E.11}$$

  end
end

Algorithm 1: Optimization of by variational inference methods. The process from Eq. (E.9)-(E.11) is repeated until it reaches convergence at the local minimum. Note that the convergence can be checked by comparing the perplexity (see Eq. (E.15)) from the previous iteration and the current iteration's one

E.3 Experimental results

In this research, we employed a collection of face images from the MyFace dataset (See chapter 1, section 1.6). The dataset contains 35,644 face images with the size 100 × 100 pixels. The dictionary contains 60 words, which has been described in the section E.2.1. The dataset contain only one person such that we can analyze the facial features without concerning the interpersonal issue. The dataset is originally aimed to evaluate both still and temporal features with the corresponding emotion labels. In this chapter, however, we utilized only the still images in order to discover the new clusters of face appearances.

In order to evaluate the quality of the topics, we can measure the perplexity by the following equation:

$$perplexity = exp\left(-\frac{\sum_{d=1}^{D}\log p(w_d)}{\sum_{d=1}^{D} N_d}\right) \tag{E.15}$$

where N_d is the number of occurrences (word) in the document d. The perplexity versus number of topic in MyFace dataset is shown in Fig. E.7. The lower perplexity indicates the better performance. In our experiment, the best perplexity score we obtained is calculated from 10 topics. We can draw the topics of an input image by determine the β parameter, which indicates the probability of the topic mixture given by a word. Thus, we can estimate the topic from the words in an image as shown in Fig. E.8. We can see the mixture of topics in subfigure E.8c and the corresponding topic bands in subfigure E.8d-E.8h.

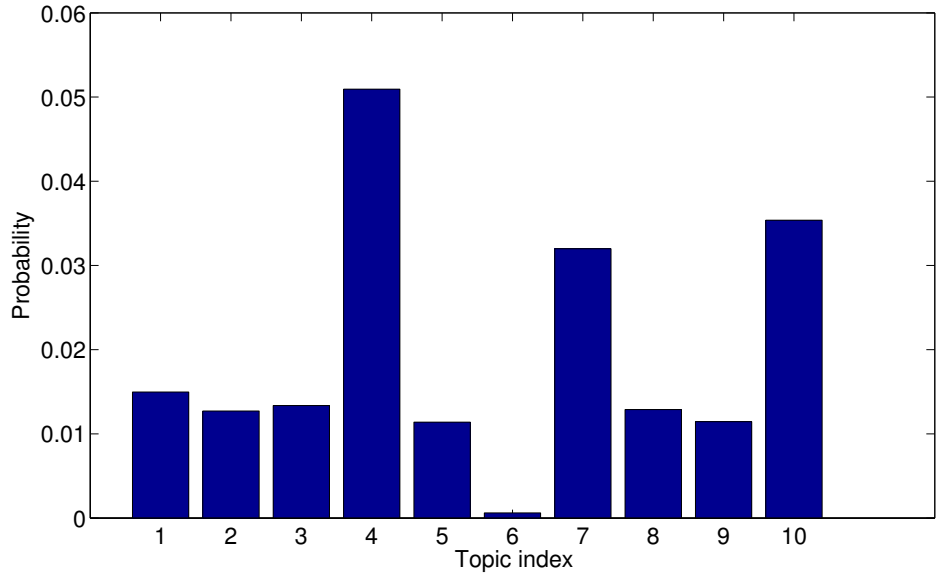


Figure E.6: The topic distribution of the 1^{st} word (from $\beta_{k,v}$ where $v = 1$). This example is trained by LDA with setting 10 topics. The horizontal axis shows the topic index, and the vertical axis indicates the probability.

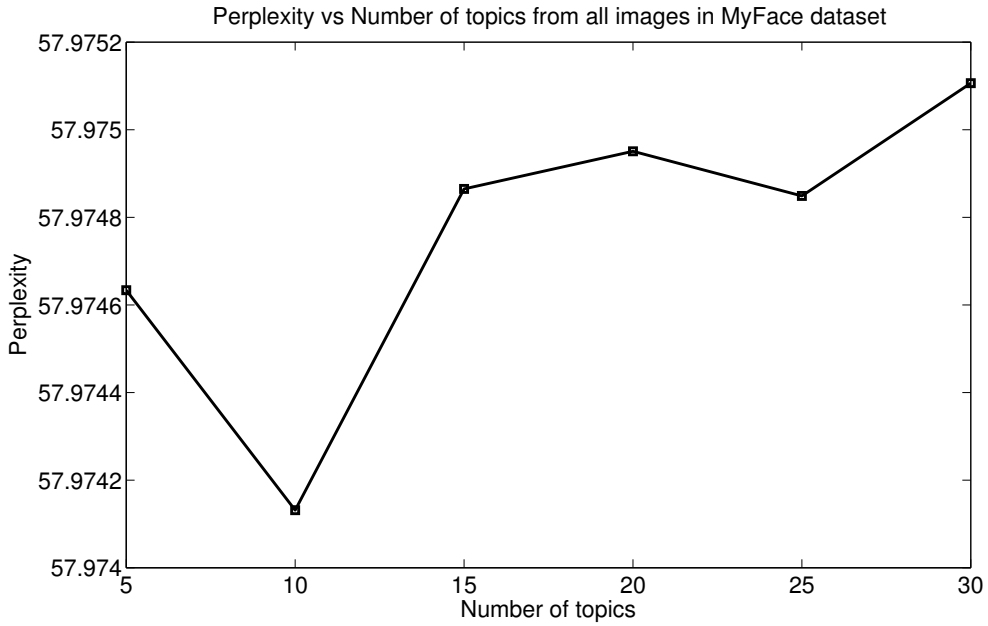
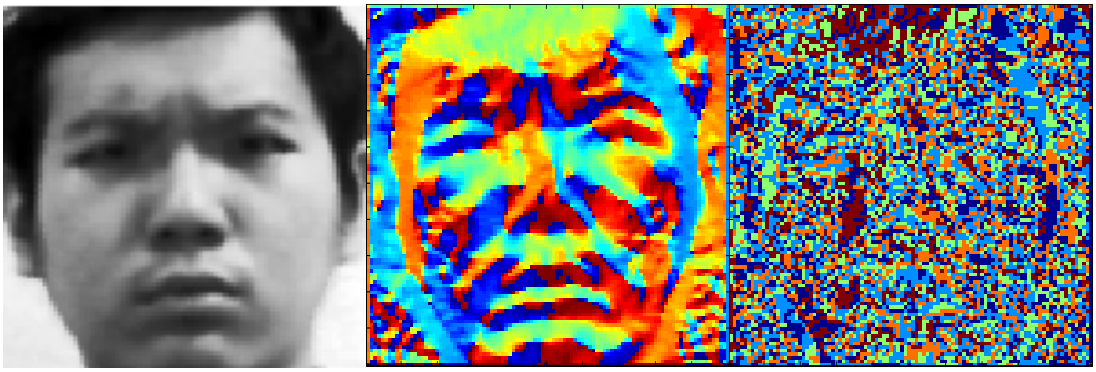


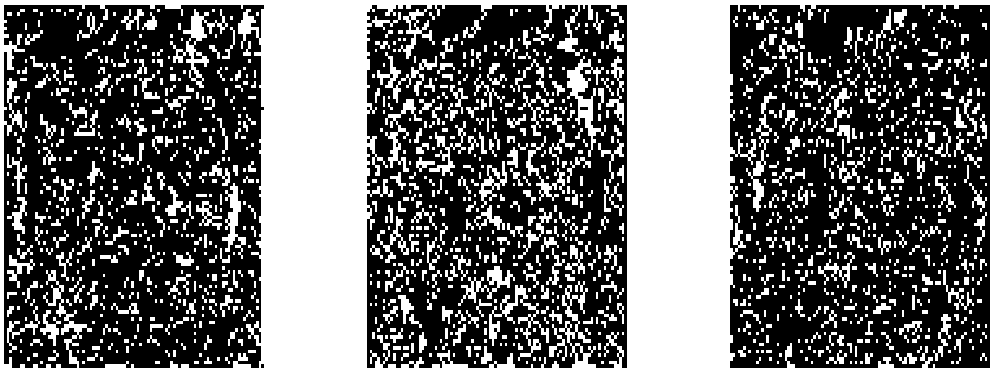
Figure E.7: Perplexity (vertical axis) versus number of topic (horizontal axis) in MyFace dataset

E.4 Conclusion

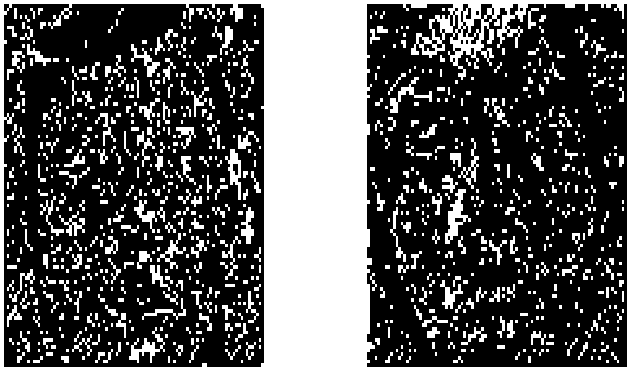
In this chapter, we determine the categorization of facial expression of a person in the frontal face dataset by using topic model technique. Instead of given a hard decision of cluster labeling as in traditional clustering methods, we define face images as the mixture of topics. Since the LDA was originally designed for the text processing, we cannot



(a) Reference image (b) gradient orientation (c) mixture of 5 topics



(d) 1^{st} topic (e) 2^{nd} topic (f) 3^{rd} topic



(g) 4^{th} topic (h) 5^{th} topic

Figure E.8: Mixture of five topics of (a) a sample image (at index 350). The topics are shown in the 5 separated bands (d)-(h)

apply LDA directly with image data. Thus, we essentially focus on the conversion from an image to the text format. There are four components we need to concern in the conversion: words, document, topic, and dictionary.

Firstly, the image-to-words conversion method is proposed in this chapter. We define the gradient orientations in pixel level as the words, and the whole image as the document. The word index in the dictionary is defined by the quantized bins of the gradient orientation. The motivation behind this setting is the robustness of the gradient orientation under the changing light conditions. To address the problem in the experiment, the issue on the image-to-words conversion method is the prominent factor for the successfulness

of the topic modeling on the face images. Unlike typical face clustering methods, which distinguish the person identity, the dissimilarity of facial appearances from a single person is much smaller. Therefore, we may found many similar trends of the word occurrences across different documents. In addition, the current image-to-words conversion method contains only a small number of vocabularies, which also effect to the perplexity evaluation performance. Thus, we would highly recommend focusing on this issue for the further development.

The second part in this article discussed about how to approximate the LDA parameter given by an image dataset. We employed the Variational Bayes inference to infer the parameters. As a result, we demonstrate the topics we extracted from an image (document), and evaluate the optimal number of topic for our dataset. Implicitly, the topics we discovered can tell us what is the importance pixels we should interest. However, the Variational Bayes inference method assumes that the hidden parameters are independent to the true posterior. Thus, this method is often yields the poor estimation. It is suggested to investigate more on the other inference methods such as Gibbs sampling [194] or expectation propagation.

In addition to the above issues, we realized the disadvantage of using texture from a still image in the case that the intensity of expression is very weak. We refer this problem as a subtle facial expression. As the majority of the dataset used in this experiment are subtle expressions which the differences cannot be recognized by both geometric and texture feature. Therefore, this is also one of our supportive reason to use observe facial expression by motion based feature.

Bibliography

[1] M. Schünke, L.M. Ross, E. Schulte, E.D. Lamperti, U. Schumacher, E. Taub, J. Rude, M. Voll, and K. Wesker, *Thieme Atlas of Anatomy: Head and Neuroanatomy*, Thieme Atlas of Anatomy. Thieme, 2007.

[2] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager, *Facial Action Coding System: The Manual on CD ROM*, A Human Face, Salt Lake City, 2002.

[3] Michael J. Lyons, Miyuki Kamachi, and Jiro Gyoba, "Japanese female facial expressions (jaffe)," 1997.

[4] 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2010, pp. 94–101.

[5] A. Asthana, S. Zafeiriou, Shiyang Cheng, and M. Pantic, "Robust Discriminative Response Map Fitting with Constrained Local Models," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3444–3451.

[6] T. Kurozumi, Y. Shinza, Y. Kenmochi, and K. Kotani, "Facial individuality and expression analysis by eigenspace method based on class features or multiple discriminant analysis," in *International Conference on Image Processing, 1999. ICIP 99*, 1999, vol. 1, pp. 648–652 vol.1.

[7] I. Eguchi and K. Kotani, "Facial expression analysis by generalized eigen-space method based on class-features (gemc)," in *IEEE International Conference on Image Processing, 2005. ICIP 2005*, Sept 2005, vol. 1, pp. I–293–6.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[9] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager, *Facial action coding system: Investigator Guide*, A Human Face, Salt Lake City, 2002.

[10] Albert Mehrabian and Morton Wiener, "Decoding of inconsistent communications," *Journal of Personality and Social Psychology*, vol. 6, no. 1, pp. 109–114, 1967.

[11] Albert Mehrabian and Susan R. Ferris, "Inference of attitudes from nonverbal communication in two channels," *Journal of Consulting Psychology*, vol. 31, no. 3, pp. 248–252, 1967.

[12] A. Mehrabian, *Nonverbal Communication*, Aldine, 1977.

[13] R. W. Picard, "Affective Computing," Tech. Rep. 321, M.I.T Media Laboratory Perceptual Computing Section, 1995.

[14] James Davis and Aaron Bobick, "The representation and recognition of action using temporal templates," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 928–934.

[15] Aaron F. Bobick and James W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[16] John G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, vol. 2, no. 7, pp. 1160–1169, Jul 1985.

[17] Zara Ambadar, J. Schooler, and Jeffrey Cohn, "Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions," *Psychological Science*, 2005.

[18] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[19] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287 – 314, 1994, Higher Order Statistics.

[20] Y. Kosaka and K. Kotani, "Facial expression analysis by kernel eigenspace method based on class features (kemc) using nonlinear basis for separation of expression-classes," in *International Conference on Image Processing, 2004. ICIP '04. 2004*, Oct 2004, vol. 2, pp. 1409–1412 Vol.2.

[21] Paul Ekman, "Universal and Cultural Differences in Facial Expression of Emotion," in *Nebraska Symposium on Motivation*. 1972, vol. 19, pp. 207–281, Lincoln University of Nebraska Press.

[22] Paul Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, pp. 45–60. John Wiley & Sons, 2005.

[23] Paul Ekman and Wallace V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.*, Consulting Psychologists Press, Palo Alto, 1978.

[24] Sylvia D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394 – 421, 2010, The biopsychology of emotion: Current theoretical and empirical perspectives.

[25] P. Ekman, R.W. Levenson, and W.V. Friesen, "Autonomic Nervous System Activity Distinguishes Among Emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[26] Paul Ekman, "Expression and the Nature of Emotion," in *Approaches to Emotion*, K. Scherer and P. Ekman, Eds., pp. 319–343. Lawrence Erlbaum, Hillsdale, NJ, 1984.

[27] Paul Ekman, "All Emotions Are Basic," in *The Nature of Emotion: Fundamental Questions*, P. Ekman and R. Davidson, Eds., pp. 15–19. Oxford University Press, New York, 1994.

[28] David Matsumoto and Bob Willingham, "Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals," *Journal of Personality and Social Psychology*, vol. 96, no. 1, pp. 1–10, Jan. 2009.

[29] Jen Wathan, Anne M. Burrows, Bridget M. Waller, and Karen McComb, "Equifacs: The equine facial action coding system," *PLoS ONE*, vol. 10, no. 8, pp. e0131738, 08 2015.

[30] Charles Darwin, *The expression of the emotions in man and animals*, John Murray, London, 1872.

[31] James A. Russell, "Culture and the categorization of emotions," *Psychological Bulletin*, vol. 110, no. 3, pp. 426–450, Nov. 1991.

[32] James A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, pp. 102–141, 1994.

[33] Kristen A. Lindquist, Lisa Feldman Barrett, Eliza Bliss-Moreau, and James A. Russell, "Language and the perception of emotion," *Emotion*, vol. 6, no. 1, pp. 125–138, 2006.

[34] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett, "Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture," *Emotion*, vol. 14, no. 2, pp. 251–262, 2014.

[35] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image Vision Comput.*, vol. 30, no. 10, pp. 683–697, Oct. 2012.

[36] Hatice Gunes and Maja Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotion*, vol. 1, no. 1, pp. 68–99, 2010.

[37] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S. Huanag, "Human-centred intelligent human computer interaction ; how far are we from attaining it?," *International Journal Autonomous and Adaptive Communications Systems*, vol. 1, no. 2, pp. 168–187, Aug. 2008.

[38] Ursula Hess and Robert E. Kleck, "Differentiating emotion elicited and deliberate emotional facial expressions," *European Journal of Social Psychology*, vol. 20, no. 5, pp. 369–385, 1990.

[39] Karen L. Schmidt, Jeffrey F. Cohn, and Yingli Tian, "Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles," 2003.

[40] Jeffrey Cohn and Karen Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, pp. 1 – 12, March 2004.

[41] Mohammed Hoque, Louis-Philippe Morency, and Rosalind W. Picard, "Are you friendly or just polite? - analysis of smiles in spontaneous face-to-face interactions," in *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part I*, Berlin, Heidelberg, 2011, ACII'11, pp. 135–144, Springer-Verlag.

[42] Shichuan Du, Yong Tao, and Aleix M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[43] Qiang Ji, P. Lan, and C. Looney, "A probabilistic framework for modeling and real-time monitoring human fatigue," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 36, no. 5, pp. 862–875, Sept 2006.

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

[45] Gwen C. Littlewort, Marian Stewart Bartlett, and Kang Lee, "Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, New York, NY, USA, 2007, ICMI '07, pp. 15–21, ACM.

[46] Alea Teeters, Rana El Kaliouby, and Rosalind Picard, "Self-cam: Feedback from what would be your social partner," in *ACM SIGGRAPH 2006 Research Posters*, New York, NY, USA, 2006, SIGGRAPH '06, ACM.

[47] M. Yeasin, B. Bullot, and R. Sharma, "From facial expression to level of interest: a spatio-temporal approach," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, June 2004, vol. 2, pp. II–922–II–927 Vol.2.

[48] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, 2001, vol. 1, pp. I–511–I–518 vol.1.

[49] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," 1981, pp. 674–679.

[50] Carlo Tomasi and Takeo Kanade, "Detection and tracking of point features," Tech. Rep., International Journal of Computer Vision, 1991.

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

[52] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[53] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.

[54] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proceedings of the British Machine Vision Conference*. 2006, pp. 95.1–95.10, BMVA Press, doi:10.5244/C.20.95.

[55] Jrgen Ahlberg, "Candide-3 - an updated parameterised face," Tech. Rep., 2001.

[56] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen, and Thomas S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 12, pp. 160 – 187, 2003, Special Issue on Face Recognition.

[57] Hai Tao and T.S. Huang, "Explanation-based facial motion tracking using a piecewise bezier volume deformation model," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999.*, 1999, vol. 1, pp. –617 Vol. 1.

[58] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[59] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

[60] Chaochao Lu and Xiaoou Tang, "Surpassing human-level face verification performance on lfw with gaussianface," in *AAAI Conference on Artificial Intelligence*, 2015.

[61] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar, "Effective face frontalization in unconstrained images," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA*, 2015, pp. 4295–4304.

[62] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 2001.

[63] C. Martin, U. Werner, and H.-M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *8th IEEE International Conference on Automatic Face Gesture Recognition, 2008. FG '08.*, Sept 2008, pp. 1–6.

[64] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M. Prkachin, and Patricia E. Solomon, "The painful face - pain expression recognition using active appearance models," *Image Vision Comput.*, vol. 27, no. 12, pp. 1788–1796, Nov. 2009.

[65] J. Sung and Daijin Kim, "Pose-robust facial expression recognition using view-based 2d + 3d aam," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 38, no. 4, pp. 852–866, July 2008.

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

[67] Filareti Tsalakanidou and Sotiris Malassiotis, "Real-time 2d+3d facial action and expression recognition," *Pattern Recogn.*, vol. 43, no. 5, pp. 1763–1775, May 2010.

[68] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, Jan 2007.

[69] P.S. Aleksic and A.K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream hmms," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 1, pp. 3–11, March 2006.

[70] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran, "Geometry vs. appearance for discriminating between posed and spontaneous emotions," in *ICONIP (3)*, 2011, pp. 431–440.

[71] Caifeng Shan, Shaogang Gong, and Peter W. McOwan, "Appearance manifold of facial expression.," in *ICCV-HCI*, Nicu Sebe, Michael S. Lew, and Thomas S. Huang, Eds. 2005, vol. 3766 of *Lecture Notes in Computer Science*, pp. 221–230, Springer.

[72] Yuxiao Hu, Z. Zeng, Lijun Yin, Xiaozhou Wei, Xi Zhou, and T.S. Huang, "Multi-view facial expression recognition," in *8th IEEE International Conference on Automatic Face Gesture Recognition, 2008. FG '08.*, Sept 2008, pp. 1–6.

[73] S. Berretti, A.D. Bimbo, P. Pala, B.B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *20th International Conference on Pattern Recognition (ICPR), 2010*, Aug 2010, pp. 4125–4128.

[74] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05).*, March 2005, vol. 2, pp. ii/1085–ii/1088 Vol. 2.

[75] M.S. Bartlett, Gwen Littlewort, M. Frank, C. Lainscsek, Ian Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2005*, 2005, vol. 2, pp. 568–573 vol. 2.

[76] Limin Ma, D. Chelberg, and M. Celenk, "Spatio-temporal modeling of facial expressions using gabor-wavelets and hierarchical hidden markov models," in *IEEE International Conference on Image Processing*, Sept 2005, vol. 2, pp. II–57–60.

[77] Ligang Zhang and D. Tjondronegoro, "Facial expression recognition using facial movement features," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 219–229, Oct 2011.

[78] Peng Yang, Qingshan Liu, and D.N. Metaxas, "Boosting coded dynamic features for facial action units and facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.*, 2007, pp. 1–6.

[79] M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[80] T. Otsuka and Jun Ohya, "Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences," in *International Conference on Image Processing, 1997*, Oct 1997, vol. 2, pp. 546–549 vol.2.

[81] Jenn-Jier James Lien, *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*, Ph.D. thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1998.

[82] Bruce David Lucas, *Generalized Image Matching by the Method of Differences*, Ph.D. thesis, Pittsburgh, PA, USA, 1985, AAI8601180.

[83] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proceedings of the 13th Scandinavian Conference on Image Analysis*, Berlin, Heidelberg, 2003, SCIA'03, pp. 363–370, Springer-Verlag.

[84] Kenji Mase, "Recognition of Facial Expression from Optical Flow," *IEICE Transactions on Information and Systems*, vol. E74-D, no. 10, pp. 3474–3483, Oct. 1991.

[85] Michael J. Black and Yaser Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *International Journal of Computer Vision*, vol. 25, pp. 23–48, 1997.

[86] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[87] M.S. Bartlett, Javier R. Movellan, and T.J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, Nov 2002.

[88] Aapo Hyvarinen, "Independent component analysis by minimization of mutual information," Tech. Rep. A46, Helsinki University of Technology, Espoo, 1997.

[89] Chengjun Liu and H. Wechsler, "Independent component analysis of gabor features for face recognition," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, July 2003.

[90] Fan Chen and Kazunori Kotani, "Comparison of mda and emc in robustness against over-fitting for facial expression recognition," *IEICE technical report. Image engineering*, vol. 107, no. 538, pp. 483–488, mar 2008.

[91] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran, "Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition," in *ICME*, 2012, pp. 1027–1032.

[92] Mohammed Ehsan Hoque, Daniel J. McDuff, and Rosalind W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 323–334, 2012.

[93] Md. Atiqur Rahman Ahad, J. K. Tan, H. Kim, and S. Ishikawa, "Motion history image: Its variants and applications," *Mach. Vision Appl.*, vol. 23, no. 2, pp. 255–281, Mar. 2012.

[94] J.W. Davis, "Hierarchical motion history images for recognizing human motion," in *IEEE Workshop on Detection and Recognition of Events in Video, 2001. Proceedings.*, 2001, pp. 39–46.

[95] M. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection in video," in *IEEE International Conference on Systems, Man and Cybernetics, 2004*, Oct 2004, vol. 1, pp. 635–640 vol.1.

[96] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, June 2005, vol. 1, pp. 886–893 vol. 1.

[97] Yu-Ting Pai, Shanq-Jang Ruan, Mon-Chau Shie, and Yi-Chi Liu, "A simple and accurate color face detection algorithm in complex background," in *IEEE International Conference on Multimedia and Expo, 2006*, July 2006, pp. 1545–1548.

[98] Prarinya. Siritanawan and Kazunori. Kotani, "Facial expression classification by temporal template features," in *SICE Annual Conference 2014, SICE 2014*, 2014, pp. 604–609.

[99] Tingfan Wu, M.S. Bartlett, and Javier R. Movellan, "Facial expression recognition using gabor motion energy filters," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2010, pp. 42–47.

[100] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, November 1946.

[101] Shiguang Shan, Wen Gao, Yizheng Chang, Bo Cao, and Pang Yang, "Review the strength of gabor features for face recognition from the angle of its robustness to mis-alignment," in *International Conference on Pattern Recognition, 2004. ICPR 2004*, Aug 2004, vol. 1, pp. 338–341 Vol.1.

[102] J.-K. Kamarainen, *Feature Extraction Using Gabor Filters*, Ph.D. thesis, Lappeenranta University of Technology, 2003.

[103] P. Siritanawan, K. Kotani, and Fan Chen, "Independent subspace of dynamic gabor features for facial expression classification," in *IEEE International Symposium on Multimedia (ISM2014)*, Dec 2014, pp. 47–54.

[104] Chengjun Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, Apr 2002.

[105] LinLin Shen, Li Bai, and Michael Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification," *Image and Vision Computing*, vol. 25, no. 5, pp. 553 – 563, 2007.

[106] S.E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on gabor filters," *IEEE Transactions on Image Processing*, vol. 11, no. 10, pp. 1160–1167, Oct 2002.

[107] Ying li Tian, "Evaluation of face resolution for expression analysis," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 5, pp. 82, 2004.

[108] Mohammad H. Mahoor, Mu Zhou, Kevin L. Veon, Seyed Mohammad Mavadati, and Jeffrey F. Cohn, "Facial action unit recognition with sparse representation," in *FG*. 2011, pp. 336–342, IEEE.

[109] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, and Javier R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[110] Yan Tong, Wenhui Liao, and Qiang Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[111] Michel F. Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F. Cohn, "Spontaneous vs. posed facial behavior: Automatic analysis of brow actions," in *Proc. ACM Int'l Conf. Multimodal Interfaces*, 2006, pp. 162–170.

[112] Xuehan Xiong and F. de la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*, June 2013, pp. 532–539.

[113] Khin Thu Zar Win, Fan Chen, J. Izawa, and K. Kotani, "Pose invariant robust facial expression analysis," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Sept 2010, pp. 3837–3840.

[114] Arman Savran, BüLent Sankur, and M. Taha Bilge, "Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units," *Pattern Recogn.*, vol. 45, no. 2, pp. 767–782, Feb. 2012.

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

[116] Yi Sun, M. Reale, and Lijun Yin, "Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition," in *8th IEEE International Conference on Automatic Face Gesture Recognition, 2008. FG '08.*, 2008, pp. 1–8.

[117] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Conference on Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06.*, June 2006, pp. 149–149.

[118] Bihan Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, Feb 2014.

[119] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940–1954, Nov 2010.

[120] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert, "Recognition of 3d facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762 – 773, 2012, 3D Facial Behaviour Analysis and Understanding.

[121] F. De la Torre, J. Campoy, Z. Ambadar, and J.F. Cohn, "Temporal segmentation of facial behavior," in *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.*, Oct 2007, pp. 1–8.

[122] Lisa Feldman Barrett, "Valence is a basic building block of emotional life," *Journal of Research in Personality*, vol. 40, no. 1, pp. 35–55, Feb. 2006.

[123] RachaelE. Jack, OliverG.B. Garrod, and PhilippeG. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," *Current Biology*, vol. 24, no. 2, pp. 187 – 192, 2014.

[124] Robert Plutchik and Henry Kellerman, Eds., *Biological foundations of emotion*, Number v. 3 in Emotion, theory, research, and experience. Academic Press, Orlando, 1986.

[125] M. Lewis and L. Canamero, "Are discrete emotions useful in human-robot interaction? feedback from motion capture analysis," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), 2013*, Sept 2013, pp. 97–102.

[126] Kai Sun, Junqing Yu, Yue Huang, and Xiaoqiang Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *IEEE International Conference on Multimedia and Expo, 2009. ICME 2009.*, June 2009, pp. 566–569.

[127] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 106–118, April 2011.

[128] James A. Russell and Lisa Feldman Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.," *Journal of personality and social psychology*, vol. 76, no. 5, pp. 805, 1999.

[129] Jeff T. Larsen and A. Peter McGraw, "The Case for Mixed Emotions," *Social and Personality Psychology Compass*, vol. 8, no. 6, pp. 263–274, June 2014.

[130] R.W. Picard, *Affective Computing*, MIT Press, 2000.

[131] Yangzhou Du, Wenyuan Bi, Tao Wang, Yimin Zhang, and Haizhou Ai, "Distributing Expressional Faces in 2-D Emotional Space," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, New York, NY, USA, 2007, CIVR '07, pp. 395–400, ACM.

[132] A. Hakim, S. Marsland, and H.W. Guesgen, "Statistical Modelling of Complex Emotions Using Mixture of Von Mises Distributions," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept. 2013, pp. 517–522.

[133] D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard, "Affect valence inference from facial action unit spectrograms," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2010, pp. 17–24.

[134] Jrmie Nicolle, Vincent Rapp, Kvin Bailly, Lionel Prevost, and Mohamed Chetouani, "Robust Continuous Prediction of Human Emotions Using Multiscale Dynamic Cues," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2012, ICMI '12, pp. 501–508, ACM.

[135] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran, "Representation of facial expression categories in continuous arousalvalence space: Feature and correlation," *Image and Vision Computing*, vol. 32, no. 12, pp. 1067–1079, Dec. 2014.

[136] M.A. Nicolaou, H. Gunes, and M. Pantic, "Audio-Visual Classification and Fusion of Spontaneous Affective Data in Likelihood Space," in *2010 20th International Conference on Pattern Recognition (ICPR)*, Aug. 2010, pp. 3695–3699.

[137] M.A. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[138] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, Mar. 2012.

[139] Mihalis A. Nicolaou, Stefanos Zafeiriou, and Maja Pantic, "Correlated-spaces Regression for Learning Continuous Emotion Dimensions," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM '13, pp. 773–776, ACM.

[140] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using Continuous Conditional Random Fields," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–8.

[141] F. Eyben, M. Wollmer, M.F. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, Mar. 2011, pp. 322–329.

[142] Enrique Snchez-Lozano, Paula Lopez-Otero, Laura Docio-Fernandez, Enrique Argones-Ra, and Jos Luis Alba-Castro, "Audiovisual Three-level Fusion for Continuous Estimation of Russell's Emotion Circumplex," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA, 2013, AVEC '13, pp. 31–40, ACM.

[143] Martin Wllmer, Florian Eyben, Stephan Reiter, Bjrn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies.," in *INTERSPEECH*, 2008, vol. 2008, pp. 597–600.

[144] Martin Wllmer, Moritz Kaiser, Florian Eyben, Bjrn Schuller, and Gerhard Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, Feb. 2013.

[145] Michael Grimm, D. Dastidar, and Kristian Kroschel, "Recognizing emotions in spontaneous facial expressions," in *Proceedings: International Conference on Intelligent Systems and Computing (ISYC)*, 2006.

[146] Michael Grimm, Kristian Kroschel, and others, *Emotion estimation in speech using a 3d emotion space concept*, Citeseer, 2007.

[147] Ittipan Kanluan, Michael Grimm, and Kristian Kroschel, "Audio-visual emotion recognition using an emotion space concept," in *16th European Signal Processing Conference, Lausanne, Switzerland*, 2008.

[148] George Caridakis, Kostas Karpouzis, Manolis Wallace, Loic Kessous, and Noam Amir, "Multimodal users affective state analysis in naturalistic interaction," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 49–66, Mar. 2010.

[149] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions.," in *INTERSPEECH*, 2009, pp. 1983–1986.

[150] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08*, June 2008, pp. 1–6.

[151] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a Minimal Representation of Affective Gestures," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 106–118, Apr. 2011.

[152] M. Kipp and J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009*, Sept. 2009, pp. 1–8.

[153] Pouria Fewzee and Fakhri Karray, "Continuous Emotion Recognition: Another Look at the Regression Problem," Sept. 2013, pp. 197–202, IEEE.

[154] R. Plutchik, *The Emotions: Facts and Theories, and a New Model*, Random House studies in psychology. Random House, 1962.

[155] James A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[156] David Watson and Auke Tellegen, "Toward a consensual structure of mood," *Psychological Bulletin*, vol. 98, no. 2, pp. 219–235, 1985.

[157] Albert Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, Dec. 1996.

[158] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan 2012.

[159] C.H. Hjortsjö, *Man's Face and Mimic Language*, Studen litteratur, 1969.

[160] G.B. Duchenne and R.A. Cuthbertson, *The Mechanism of Human Facial Expression*, Cambridge books online. Cambridge University Press, 1990.

[161] Jeffrey F. Cohn and Paul Ekman, "Measuring facial action," *The new handbook of methods in nonverbal behavior research*, pp. 9–64, 2005.

[162] Paul Ekman, Wallace V. Friesen, and Ronald C. Simons, "Is the startle reaction an emotion?," *Journal of Personality and Social Psychology*, vol. 49, no. 5, pp. 1416–1426, 1985.

[163] Nathan A. Fox and Richard J. Davidson, "Patterns of brain electrical activity during facial signs of emotion in 10-month old infants," *Developmental Psychology*, pp. 230–236, 1988.

[164] M.A. Sayette, Jeffrey Cohn, J.M. Wertz, M.A. Perrott, and D.J. Parrott, "A psychometric evaluation of the facial action coding system for assessing spontaneous expression," *Journal of Nonverbal Behavior*, vol. 25, pp. 167 – 186, 2001.

[165] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[166] Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, "Measuring facial expressions by computer image analysis," 1999.

[167] E. Ardizzone, M. La Cascia, and F. Vella, "Mean shift clustering for personal photo album organization," in *15th IEEE International Conference on Image Processing*, Oct 2008, pp. 85–88.

[168] Florent Perronnin and Jean-Luc Dugelay, "Clustering face images with application to image retrieval in large databases," in *Biometric Technology for Human Identification, SPIE International Symposium on Defense and Security, 28 March-1 April 2005, Orlando, USA / Also published in SPIE 5779, 256 (2005)*, Orlando, UNITED STATES, 03 2005.

[169] P. Antonopoulos, N. Nikolaidis, and I. Pitas, "Hierarchical face clustering using sift image features," in *IEEE Symposium on Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007.*, April 2007, pp. 325–329.

[170] Liexian Gu, Tong Zhang, and Xiaoqing Ding, "Clustering consumer photos based on face recognition," in *IEEE International Conference on Multimedia and Expo, 2007*, July 2007, pp. 1998–2001.

[171] Wei-Ta Chu, Ya-Lin Lee, and Jen-Yu Yu, "Visual language model for face clustering in consumer photos," in *Proceedings of the 17th ACM International Conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 625–628, ACM.

[172] R. Vidal, Yi Ma, and J. Piazzi, "A new gpca algorithm for clustering subspaces by fitting, differentiating and dividing polynomials," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, June 2004, vol. 1, pp. I–510–I–517 Vol.1.

[173] Le Lu and Ren Vidal, "Combined central and subspace clustering for computer vision applications," in *In Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 593–600.

[174] S. Foucher and L. Gagnon, "Automatic detection and clustering of actor faces based on spectral clustering techniques," in *Fourth Canadian Conference on Computer and Robot Vision, 2007. CRV '07.*, May 2007, pp. 113–122.

[175] Guangliang Chen and Gilad Lerman, "Spectral curvature clustering (scc)," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[176] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movie content analysis," *Image and Vision Computing*, vol. 29, no. 10, pp. 693 – 705, 2011.

[177] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, Nov 2013.

[178] A. Ekin, S. Pankanti, and A. Hampapur, "Initialization-independent spectral clustering with applications to automatic video analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, vol. 3, pp. iii–641–4 vol.3.

[179] Ji Tao and Yap-Peng Tan, "Efficient clustering of face sequences with application to character-based movie browsing," in *15th IEEE International Conference on Image Processing, 2008. ICIP 2008.*, Oct 2008, pp. 1708–1711.

[180] Baoyuan Wu, Yifan Zhang, Bao-Gang Hu, and Qiang Ji, "Constrained clustering and its application to face clustering in videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*, June 2013, pp. 3507–3514.

[181] Simeon Schwab, Thierry Chateau, Christophe Blanc, and Laurent Trassoudaine, "A multi-cue spatio-temporal framework for automatic frontal face clustering in video sequences," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, 2013.

[182] Wei Zhang, Tong Zhang, and D. Tretter, "Beyond face: Improving person clustering in consumer photos by exploring contextual information," in *IEEE International Conference on Multimedia and Expo (ICME), 2010*, July 2010, pp. 1540–1545.

[183] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D.A. Forsyth, "Names and faces in the news," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, June 2004, vol. 2, pp. II–848–II–854 Vol.2.

[184] Liyan Zhang, Dmitri V. Kalashnikov, and Sharad Mehrotra, "A unified framework for context assisted face clustering," in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, New York, NY, USA, 2013, ICMR '13, pp. 9–16, ACM.

[185] Peng Wu and Feng Tang, "Improving face clustering using social context," in *Proceedings of the International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 907–910, ACM.

[186] Wei-Ta Chu, Ya-Lin Lee, and Jen-Yu Yu, "Using context information and local feature points in face clustering for consumer photos," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, April 2009, pp. 1141–1144.

[187] Panpan Huang, Yunhong Wang, and Ming Shao, "A new method for multi-view face clustering in video sequence," in *IEEE International Conference on Data Mining Workshops, 2008. ICDMW '08.*, Dec 2008, pp. 869–873.

[188] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, June 2005, vol. 2, pp. 524–531 vol. 2.

[189] Xiaogang Wang and Eric Grimson, "Spatial latent dirichlet allocation," in *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, Eds., pp. 1577–1584. Curran Associates, Inc., 2008.

[190] M. Fritz and B. Schiele, "Decomposition, discovery and detection of visual categories using topic models," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, June 2008, pp. 1–8.

[191] Beat Fasel, Florent Monay, and Daniel Gatica-Perez, "Latent semantic analysis of facial action codes for automatic facial expression recognition," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2004, MIR '04, pp. 181–188, ACM.

[192] Lifeng Shang, Kwok-Ping Chan, and Guodong Pan, "Dttm: A discriminative temporal topic model for facial expression recognition," in *Advances in Visual Computing*, vol. 6938 of *Lecture Notes in Computer Science*, pp. 596–606. Springer Berlin Heidelberg, 2011.

[193] Thomas Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 1999, SIGIR '99, pp. 50–57, ACM.

[194] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

Publications

I. Prarinya Siritanawan, Kazunori Kotani and Fan Chen: "Feature extraction method for facial expression classification using cumulative change of 2D and 3D image sequences," PCSJ/IMPS2013, 2013.

II. Prarinya Siritanawan, Kazunori Kotani : "Facial expression classification by temporal template features," in proc. SICE Annual Conference (SICE), pp.604-609, Sapporo, 2014.

III. Prarinya Siritanawan, Kazunori Kotani and Fan Chen: "Independent subspace of dynamic Gabor features for facial expression classification," in proc. IEEE International Symposium on Multimedia (ISM2014) , Taichung, 2014.

IV. Prarinya Siritanawan, Bao Tu Ho, Kazunori Kotani: "Unsupervised learning from facial expression by Latent Dirichlet Allocation," IEICE Technical Report, BioX2014-37, 2014.

V. Prarinya Siritanawan, Kazunori Kotani, Fan Chen: "Cumulative Differential Gabor Features for Facial expression classification," International Journal of Semantic Computing, vol. 09, no. 02, pp. 193-213, 2015.

VI. Prarinya Siritanawan, Kazunori Kotani: "Facial action units detection by robust temporal features," in proc. International Conference on Soft Computing and Pattern Recognition (SoCPaR), Fukuoka, 2015.

VII. Prarinya Siritanawan, Kazunori Kotani: "Estimating dimensional emotion parameters from facial image features ," PCSJ/IMPS2015, 2015.