Robust representation and recognition of facial emotions

Shojaeilangari, Seyedehsamaneh

2014

Shojaeilangari, S. (2014). Robust representation and recognition of facial emotions.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62922>

<https://doi.org/10.32657/10356/62922>

ROBUST REPRESENTATION AND RECOGNITION OF FACIAL EMOTIONS

SEYEDEHSAMANEH SHOJAEILANGARI

SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING

2014

Robust Representation and Recognition of Facial Emotions

Seyedehsamaneh Shojaeilangari

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirements for the Degree of
Doctor of Philosophy

2014

*In the name of God the
compassionate the merciful*

*To my dear parents and lovely
husband*

Acknowledgements

First of all, I would like to thank my supervisor Dr. Yau Wei-Yun, currently the program manager at the Institute for Infocomm Research (I²R), A*STAR. I greatly acknowledge him for his wise advice, ideas and his involvement in our research.

I would also like to express my sincere gratitude to my co-supervisor, Associate Professor Teoh Eam Khwang for providing me with guidance and enthusiastic encouragement throughout the duration of this research work. His technical insights, literary and organizational expertise were crucial to the contents of this research work.

I thank all my friends at Internet of Things Lab at the School of Electrical and Electronic Engineering, Nanyang Technological University, especially Zhou Zhi, Peh Jin Rui Raymond, and Tan Yikai who helped me to collect a database. I would also like to acknowledge my friends at the Institute for Infocomm Research, especially Dr. Karthik Nandakumar and Dr. Li Jun, for their wise advises and helps.

I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of science and supported me in all my pursuits, I am forever indebted to them.

My most important thanks here go to my husband, Dr. Abdollah Allahverdi, who always takes my problems as his own, help me to overcome them, and encourage me to achieve at the highest level. Thanks to him for his faithful love and endless help. Without him, this work would not exist.

Finally, I would like to acknowledge the School of Electrical and Electronic Engineering, Nanyang Technological University for providing the research facilities for my work in NTU, and also A*STAR for providing the award of a scholarship.

Summary

Facial Emotion detection under natural conditions is an interesting topic with a wide range of potential applications like human-computer interaction. Although there is significant research progress in this field, there are still challenges related to real-world unconstrained situations. One essential challenge is to find pose invariant spatio-temporal volumetric features to analyze the video sequence efficiently. Another important issue is how to deal with noisy and imperfect data recorded in uncontrolled environments such as illumination variations, partial occlusion, and head movements.

The focus of this research is to develop a robust system for facial expression recognition as a dynamic event in natural situations. Two strategies have been proposed in this research to address the uncontrolled environments related challenges:

- **Robust representation framework:** we propose a novel spatio-temporal descriptor based on Optical Flow (OF) components which is very distinctive and also pose-invariant.
- **Robust recognition framework:** we explored the effectiveness of sparse representation obtained by supervised learning a set of basis (dictionary). Extreme Sparse Learning (ESL) is proposed to jointly learn a dictionary and a nonlinear classification model to robustly detect the facial expression in real-world natural situations. The proposed approach combines the discriminative power of the Extreme Learning Machine (ELM) with the reconstruction property of the sparse representation to deal with noisy signal and imperfect data recorded in natural settings.

Since the facial feature extraction performance is highly dependent on facial pose, we propose a novel spatio-temporal descriptor which is robust to facial pose variations. However, the feature encoding may fail in the presence of extreme head pose variations, where some parts of the face are not visible in the recorded images. To address this problem and also dealing with illumination variations and occlusion, we suggested following the idea of sparse representation where the noisy data can be reconstructed from the clean data provided by the dictionary of the sparse representation.

While the sparse representation approach has the ability to enhance noisy data using a dictionary learned from clean data, it is not sufficient because the end goal is to correctly recognize the facial expression. In a sparse-representation-based classification task, the desired dictionary should have both representational ability and discriminative power. Since separating the classification training from dictionary learning may cause the learned dictionary

---

to be sub-optimal for the classification task, we propose to jointly learn a dictionary and classification model. In other words, in contrast with most existing schemes that attempt to update the dictionary and classifier parameters alternately by iteratively solving each sub-problem, we propose to solve them simultaneously. This joint dictionary learning and classifier training can be expected to result in a dictionary that is both reconstructive and discriminative for a robust recognition system.

To the best of our knowledge, this is the only work that attempts to simultaneously learn the sparse representation of the signal and train a *nonlinear* classifier to be discriminative for sparse codes. The proposed method jointly learns a single dictionary and also an optimal nonlinear classifier.

We have performed extensive experiments on both acted and spontaneous emotion databases to evaluate the effectiveness of the proposed feature extraction and classification schemes under different scenarios. Our results clearly demonstrate the robustness of the proposed emotion recognition framework, especially in challenging scenarios that involve illumination changes, occlusion, and head pose variations.

Contents

Acknowledgements.....	I
Summary.....	II
List of Figures.....	VII
List of Tables	X
List of Abbreviations	XII
1 Introduction.....	1
1.1 Motivation.....	2
1.2 Challenges.....	3
1.3 Our Contributions	4
1.4 Organization of the Thesis	5
2 A Literature Review.....	7
2.1 Introduction.....	7
2.2 Emotion Models.....	7
2.2.1 Categorical Model.....	8
2.2.2 Dimensional Model.....	8
2.3 Visual-based Emotion Recognition.....	10
2.3.1 Static Analysis of Emotion (Target-based Approaches)	12
2.3.2 Dynamic Analysis of Emotion (Gesture-oriented Approaches).....	13
2.4 Pose Invariant Feature Extraction	16
2.5 Sparse Representation Based Classification	17
2.6 Facial Expression Databases	19
2.6.1 CK ⁺ Database	20
2.6.2 Extended CK ⁺ Database	21
2.6.3 AVEC 2011 Database	22
2.6.4 EmotiW Database.....	24
3 Dynamic Feature Extraction for Emotion Recognition.....	27
3.1 Introduction.....	27
3.2 Method I: Extended Spatio-Temporal Histogram of Oriented Gradients	28
3.2.1 Feature Extraction	29
3.2.2 Spatio-Temporal Pyramid Decomposition.....	32
3.2.3 Feature Selection.....	33
3.2.4 Experimental Results	34

3.2.5	Discussion on Method I	37
3.3	Method II: Multi-Scale Analysis of Local Phase and Local Orientation	38
3.3.1	Methodology	40
3.3.2	Multi-Scale Analysis	42
3.3.3	Experimental Results	43
3.3.4	Discussion on Method II	45
3.4	Method III: Histogram of Dominant Phase Congruency	45
3.4.1	Spatio-Temporal PC Calculation	47
3.4.2	HDPC Feature	48
3.4.3	Experimental Results	49
3.4.4	Discussion on Method III	52
3.5	Concluding Remarks	52
4	Pose Invariant Feature Extraction for Emotion Recognition	54
4.1	Introducion	54
4.2	Pose-Invariant Feature Extraction	54
4.2.1	Optical Flow Calculation	55
4.2.2	Optical Flow Correction	58
4.2.3	Spatio-Temporal Descriptor	60
4.2.4	Spatio-Temporal Descriptor Construction	62
4.3	Experimental Results	63
4.3.1	Parameter Setting	63
4.3.2	Evaluation of the Proposed Pose-Invariant Descriptor	64
4.3.3	Comparison to Other Approaches	67
4.3.4	Robustness of the Descriptor to Pose Variations	68
4.4	Concluding Remarks	69
5	Extreme Sparse Learning for Robust Emotion Classification	71
5.1	Introduction	71
5.2	Extreme Learning Machine (ELM)	72
5.2.1	Random Feature Mappings	73
5.2.2	Kernels	73
5.3	Sparse Representation and Dictionary Learning	74
5.4	Classification Models based on Sparse Representation	76
5.4.1	Classification based on Reconstructive Dictionary	77
5.4.2	Classification based on Discriminative Dictionary	78
5.5	Proposed Methodology: Extreme Sparse Learning	80
5.5.1	Supervised Sparse Coding	83
5.5.2	Comparison of Two Methods IHT and CSMP for Supervised Sparse Coding	86
5.5.3	Dictionary Update Stage	86

5.6	Results and Discussion.....	87
5.6.1	Pre-processing.....	87
5.6.2	Parameter Setting and Initialization	88
5.6.3	Performance of ESL and KESL	90
5.6.4	Comparison to Other Results	91
5.6.5	Advantages and Limitations of the Proposed Classifiers	95
5.7	Analysis of Failure Cases.....	97
5.7.1	Failure of Face Detector.....	97
5.7.2	Failure of Optical Flow	99
5.7.3	Failure of Reference Point Detection	101
5.8	Concluding Remarks.....	101
6	Conclusion and Recommendations.....	102
6.1	Conclusion	102
6.2	Recommendations for Future Work.....	103
	Appendix A	106
	Appendix B	108
	Appendix C	110
	Author's Publications	111
	Bibliography	113

List of Figures

Figure 2-1: Illustration of Arousal-Valence model (Extracted from [23]).....	9
Figure 2-2: Schematic diagram of the dimensional models of emotions with common basic emotion categories overlaid. (Extracted from [27])	10
Figure 2-3: Upper face Action Units and some combinations. (Extracted from [29]).....	12
Figure 2-4: Example of basic emotions from CK ⁺ database. Each image is the last frame of a video clip that shows the most expressive face.....	21
Figure 2-5: Sample frames of our own collected data. The left column shows some examples of pose variation, the middle column depicts occlusion examples, and the right column includes illumination variations.	22
Figure 2-6: Sample video frames from AVEC2011 database. (Extracted from [85])	23
Figure 2-7: Binary labelling of the four affective dimensions (activation, expectation, power and valence) in a sample of AVEC2011 video database. (Extracted from [85]) ..	24
Figure 2-8: Sample frames with more than one human subject from EmotiW database. Multiple faces are detected by the face detector.	25
Figure 2-9: Sample frame sequences with provided labels from EmotiW database.....	26
Figure 3-1: Block-diagram of the proposed approach.	29
Figure 3-2: Descriptor computation; (a) The volume data is divided into a number of 3D grids. Each grid is denoted by a block (Bi). The final descriptor (F) consists of the block's feature; (b) Each block is divided into number of 3D cells (Cj). The block feature consists of the cell's histograms; (C) Each cell includes a number of points (Pk) which are characterized by a 3D gradient vector. Each gradient orientation and its variation in time domain (z axis) are quantized to compute the histogram of a cell.....	30
Figure 3-3: Overview of the face pre-processing; (a) The original face sample; (b) Normalized face based on constant distance between two eyes; (c) Rotated face based on eye coordinates.	35
Figure 3-4: Illustration of the monogenic components for one sample image of a video sequence; (a) Original sample frame; (b) First monogenic component (f_1); (c) Second monogenic component (f_2); (d) Third monogenic component (f_3); (e) Forth monogenic component (f_4).	40
Figure 3-5: Illustration of the components used to vote the phase and orientation bins for one sample image of a video sequence; (a) Energy signal for voting the phase bins; (b) $Mag\theta$ for voting the bins related to θ ; (c) $Mag\phi$ for voting the bins corresponding to ϕ	42
Figure 3-6: Illustration of the energy component at different scales. The scale of the bandpass filter is increased from (a) to (d).	42

Figure 3-7: Comparison of methods for line detection; (a) sharp line; (b) line detection based on Phase Congruency; (c) line detection based on Canny; (d) line detection based on Sobel; (e) Gradual line with intensity range of [0 3]; (f) Line detection based on Phase Congruency; (g) Line detection based on Canny; (h) Line detection based on Sobel.	47
Figure 3-8: Descriptor computation; (a) the volume data is divided into a number of 3D grids. Each grid is denoted by a block (Bi). The final descriptor (F) consists of the block's feature; (b) each block is divided into number of 3D cells (Cj). The block feature consists of the cell's histograms; (C) each cell includes a number of points (Pk) which are characterized by several oriented PC components; (D) a PC component with dominant orientation is selected for each pixel and then used to compute the histogram of a cell.	49
Figure 4-1: Reference vector and face coordinate system. The nose point is considered as the origin of the new coordinate system. The reference vector connects the nose point to the line joining the centre of the two eyes.	55
Figure 4-2: Optical flow correction for head movement; (a)-(b) two consecutive frames; (c) total optical flow (U_{tot}) is illustrated in blue; head movement optical flow (U_{head}) is shown in green; (d) expression related optical flow (U_{exp}) is illustrated in red; (e) U_{tot} of mouth region; (f) U_{exp} of mouth region.	59
Figure 4-3: Illustration of Projection descriptor for sad and happy expression.	61
Figure 4-4: Illustration of Rotation descriptor for happy and anger expressions.	61
Figure 4-5: Spatio-temporal descriptor construction; (a) volume data is divided into a number of 3D blocks (Bi). The final descriptor (F) is a concatenation of features from all the blocks; (b) each block is further divided into number of 3D cells (Cj). The feature vector of each block (f_{Bi}) is a concatenation of the cell histograms; (c) weighted and un-weighted histograms are calculated for each cell based on the four spatio-temporal features and concatenated to obtain the cell histogram.	63
Figure 4-6: Pose-invariant descriptor for surprise emotion; (a) features extracted from the lip segment from frontal face; (b) features extracted from the lip segment from non-frontal face.	68
Figure 4-7: Pose-invariant descriptor for surprise emotion; (a) features extracted from the lip segment from frontal face; (b) features extracted from the lip segment from non-frontal face.	69
Figure 5-1: Proposed method for recognition framework.	82
Figure 5-2: Recognition rate of ESL versus number of dictionary atoms.	90
Figure 5-3: Comparison of the recognition results of test set for AVEC2011.	93
Figure 5-4: Comparison of the results for EmotiW2013 database for test set.	95
Figure 5-5: Comparison the robustness of the methods to occlusion, head posed variations, and illumination changes.	97
Figure 5-6: Non-faces samples.	98
Figure 5-7: Failure of face detector to detect the whole face.	98
Figure 5-8: All faces detected from a video clips showing disgust.	99

Figure 5-9: Failure of optical flow. (a)-(b) two consequent frames; (c) estimated optical flow; (d) illustration of optical flow for nose region.	100
Figure 5-10: Wrong reference point localization	101

List of Tables

Table 2-1: Common used facial emotion databases. NA stands for Not Available.....	20
Table 3-1: Effect of number of blocks and number of cells on detection performance of ESTHOG feature.....	35
Table 3-2: Comparison the results of different features including STHOG, STHOV, and ESTHOG. The results are based on 75 blocks ($5 \times 5 \times 3$), and 9 cells ($3 \times 3 \times 1$).....	36
Table 3-3: Comparison the results of different pyramid levels. The first row is the result of 2spatial+ 1temporal levels of pyramid decomposition. Second row is for 2spatial+2temporal levels of pyramid decomposition.	37
Table 3-4: GA parameters for block selection	37
Table 3-5: Comparison the results of different methods. The first row is the results of ESTHOG descriptor without any pyramid decomposition. Second row is the results after 4 level pyramid decompositions. Third row shows the results of 4 level pyramid of the descriptor after feature selection.....	37
Table 3-6: Effect of log-Gabor bandwidth on classification accuracy for CK ⁺ database. The results are based on 50 blocks ($5 \times 5 \times 2$), and 9 cells ($3 \times 3 \times 1$).....	43
Table 3-7: Effect of log-Gabor scales on classification accuracy for CK ⁺ database. The results are based on 50 blocks ($5 \times 5 \times 2$), and 9 cells ($3 \times 3 \times 1$).....	43
Table 3-8: Comparison on number of blocks and number of cells on classification accuracy.	44
Table 3-9: Comparison of STHLP, STHLO, and combined features.	44
Table 3-10: Comparison of the detection rate for the AVEC 2011 database. “A” stands for activation, “E” for expectancy, “P” for power, and “V” for valence.	45
Table 3-11: Effect of number of blocks and number of cells on detection performance of HDPC feature for CK ⁺ database.....	51
Table 3-12: Effect of log-Gabor bandwidth on classification accuracy for CK ⁺ database. The results are based on 75 blocks ($5 \times 5 \times 3$), and 9 cells ($3 \times 3 \times 1$).....	51
Table 3-13: Effect of log-Gabor scales on classification accuracy for CK ⁺ database. The results are based on 75 blocks ($5 \times 5 \times 3$), and 9 cells ($3 \times 3 \times 1$).....	51
Table 3-14: Comparison of spatial HDPC and spatio-temporal HDPC on CK ⁺ database.	51
Table 3-15: Comparison of the detection rate for the AVEC 2011 database. “A” stands for activation, “E” for expectancy, “P” for power, and “V” for valence.	52

Table 4-1: Effect of number of blocks and cells on recognition rate for CK ⁺ database. We used SVM as the classifier.	64
Table 4-2: Earth Mover Distance of features for lip segment.	65
Table 4-3: Earth Mover Distance of features for nose segment.	66
Table 4-4: Confusion matrix of the result.	66
Table 4-5: The classification performance on each feature individually. We used SVM classifier on CK ⁺ database.	66
Table 4-6: Assessment the kind of histogram in classification performance. The results are based on SVM classifier on CK ⁺ database.	66
Table 4-7: Comparison of the proposed descriptor to other methods for the original CK ⁺ database. We used SVM as a classifier for all methods.	67
Table 4-8: Comparison the robustness of the methods to pose variations. We used SVM as a classifier for all methods.	69
Table 5-1: Comparison of two methods for supervised sparse coding applied for CK ⁺ database.	86
Table 5-2: Parameter setting for ESL and KESL.	88
Table 5-3: The sensitivity of the ESL to regularization parameters for ECK ⁺ database.	89
Table 5-4: Comparison of the proposed pose-invariant descriptor to other classifiers for all databases.	91
Table 5-5: Comparison of the proposed classification results to other approaches with the same descriptor on ECK ⁺	92
Table 5-6: Comparison of the detection rate for the AVEC 2011 database to check the efficiency of the proposed approach. “A” stands for activation, “E” for expectancy, “P” for power, and “V” for valence.	92
Table 5-7: Comparison of our proposed methods for EmotiW2013 database to the baseline results.	94
Table 5-8: Confusion matrix for seven emotions for EmotiW2013 database.	94

List of Abbreviations

AAM	Active Appearance Model
AdaBoost	Adaptive Boosting
AU	Action Unit
DBN	Dynamic Bayesian Network
DKSVD	Discriminative K-Singular Value Decomposition
DL	Dictionary Learning
ELM	Extreme Learning Machine
ESL	Extreme Sparse Learning
ESTHOG	Extended Spatio Temporal Histogram of Oriented Gradient
FACS	Facial Action Coding System
FDDL	Fisher Discriminant Dictionary Learning
GA	Genetic Algorithm
HCI	Human-Computer Interaction
HDPC	Histogram of Dominant Phase Congruency
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradient
IPTV	Internet Protocol Television
KELM	Kernel Extreme Learning Machine
KESL	Kernel Extreme Sparse Learning
K-SVD	K-Singular Value Decomposition
LBP	Local Binary Pattern
LBP-TOP	Local Binary Pattern on Three Orthogonal Planes
LCKSVD	Label Consistent K-Singular Value Decomposition

LPLO	Local Phase Local Orientation
LPQ-TOP	Local Phase Quantization on Three Orthogonal Planes
OF	Optical Flow
NN	Neural Network
PC	Phase Congruency
RIP	Restricted Isometry Property
RNN	Recurrent Neural Network
SELM	Sparse Extreme Learning Machine
SIFT	Scale Invariant Feature Transform
SLFN	Single-hidden Layer Feed-forward Network
SRC	Sparse Representation based Classifier
STHLO	Spatio Temporal Histogram of Local Orientation
STHLP	Spatio Temporal Histogram of Local Phase
STHOG	Spatio Temporal Histogram of Oriented Gradient
STHOV	Spatio Temporal Histogram of Orientation Variation
SVM	Support Vector Machine

Chapter 1

Introduction

Recent advances in technology have enabled human users to interact with computers in more efficient ways such as voice and gesture. However, one essential factor for natural interaction is still missing, and that is the emotion. Emotion plays an important role in human-human communication and interaction; allowing people to express themselves beyond the verbal domain.

Changes of person's affective state usually emphasise the transmission of a message in human-human interaction. People are able to sense these changes and used them to improve their communication. This fact has motivated a huge number of researches to enable machines to recognize human emotions.

There are several applications of human affective computing to facilitate human-computer interaction [1]. For example, a computer may become a more effective tutor if it can sense the user's affective state. On the other hand, emotional behaviour of the user can be used as feedback on their teaching and how well the students understand them. Another application is to warn drivers by monitoring their emotional condition. For instance, in Japan, there are cars equipped with a camera installed in the dashboard to detect whether the driver is angry, sleepy and generally whether he/she is in dangerous emotional situation or not [2]. Another application example is to use computer agents that could understand the user's preference through the affect-sensitive indicator to support advertisements that would target only things that the specific audience has shown interest in and not any generic product to any audience.

Essentially, audio and visual information are considered as the most important cues to assess human affective behaviour. However, the complementary relations of these two cues lead researchers to integrate audio and visual data for better performance. Physiological measurement of emotional state is also considered a reliable representation of human inner

feeling. In this research, we have focused on facial-based emotion recognition as a key part of the way that humans communicate with each other.

The goal of facial expression recognition system is to robustly determine the emotional state (e.g., happiness, sadness, surprise, neutral, anger, fear, and disgust) of a human being based on the facial images regardless of the identity of the face. In spite of various advancements reported in this field, there are still challenges remaining. The majority of the previous work tried to recognize facial expressions via static images and ignore the temporal information of an emotion in time domain due to expensive computational time or complicated temporal mode [3, 4]. However, temporal analysis of a dynamic process can strongly improve the performance of an affect analysis system in real world application.

Traditionally, emotion recognition has been performed on laboratory controlled data which is not representative of the environment faced in the real-world applications. Facial emotion recognition is very challenging due to the complexity of the expression under uncontrolled environments such as illumination variation, occlusion, head movement, etc. The accuracy of an emotion recognition system generally depends on two critical factors: 1) how to robustly represent the facial features in such a way that they are robust under intra-class variations (e.g. pose variations), but are distinctive for various emotions and 2) how to design a classifier that is capable of distinguishing different facial emotions based on noisy and imperfect data (e.g., illumination changes and occlusion).

For the representation part, we propose a novel spatio-temporal descriptor based on Optical Flow (OF) components, which is very distinctive and also pose-invariant. For the recognition part, the idea of sparse representation is combined with Extreme Learning Machine (ELM) to learn an efficient classifier that can handle noisy and imperfect data.

The main objective of the present work is to develop a facial-based emotion recognition system that is able to handle variations in facial pose, illumination and partial occlusion. In other words, the system aims to robustly represent and recognize the facial expressions in real-life situations.

1.1 Motivation

Automating facial emotion analysis offers a new modality for Human-Computer Interaction (HCI) to make the interaction responsive to human emotion. Nevertheless, a natural facial emotion analysis still poses many significant challenges to the pattern recognition and human-computer interface research community.

In this research, we are targeting automatic analysis of human emotional states in real world situations in order to facilitate HCI. Applications of such technology is vast, especially

in areas that require understanding of human preference such as interactive TV technology to analyze user's enjoyment or interest of the TV content especially advertisement in the context of electronic billboard. Indeed, Interactive TV can transform advertising from a passive consumer experience into an on-demand action to engage the user with brands and ultimately deriving closer personalization experience to the user [5].

1.2 Challenges

Traditionally, emotion recognition has been performed on laboratory controlled data which poorly represents the environment faced in real-world situations [6, 7]. As such, many of the existing methods are applicable only for laboratory-controlled data and not able to deal with natural settings such as illumination changes, occlusion, and head pose variations.

Additionally, the majority of existing work focused on facial expression processing via static images and ignored the temporal information of such a dynamic event [8-12]. However, research on the human visual system has demonstrated that better judgment of the facial expression is achieved when its temporal information is taken into account [13, 14].

Most of the reported work on automatic emotion detection are based on deliberate facial expressions while it is confirmed by several studies that spontaneous facial expressions are different from posed behaviour in term of the facial muscles used and their dynamics [15, 16]. Additionally, the expressions are very subtle when it is spontaneous compared to exaggerated or acted expressions. Thus many of the existing methods are unable to detect such subtle changes. Indeed automatic recognition of spontaneous emotional expression is a challenging problem which needs more research efforts.

In addition, most of the existing researches have focused on the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) or a subset of them [3, 4, 6]. However, the emotional state of humans in real world interaction is not actually restricted to these basic emotions but a mixture of them.

Also, there are vast studies on automatic affect analysis which use only single channel to collect the emotion information. Most efforts have been carried out in either the face or speech channel. However, it was established by several experiments that integration the information of different modalities can help to improve affect recognition rate [17, 18].

Although all previous studies confirmed that data fusion from different modalities is more accurate and reliable for human affect analysis, it is an open issue how to integrate various data stream more effectively. Some researchers believe that feature-level data fusion is superior to decision-level data fusion [19]. However, constructing a suitable joint feature vector including features from different channel is a challenging problem. Some other studies

---

have demonstrated the advantages of using decision-level fusion methods which consider multimodal signals mutually independent and combine the classification results of different single modality at the end. Many different methods for classifier fusion have been suggested, but there is no agreement on the optimal method. To get advantages of both strategies, hybrid fusion may be a good choice for the fusion problem [19].

Another important challenging issue is that interpretation of human affect state should be done with respect to the context which the expression was observed [15, 16]. For instance, having knowledge about the expresser's gender and age, his/her current task, environment or context the expresser is in and the audience of his/her expression will help to precisely analyse human expression behaviour. This is largely an unexplored field of studies in automated affect analysis.

1.3 Our Contributions

Robust representation and recognition of dynamic facial expressions from video sequences are our main focus in this research. While the majority of the previous work recognizes the facial emotions from frontal or near-frontal static images in laboratory controlled situations, our work focuses on the following three main challenges:

- Incorporating temporal information of an expression in feature extraction
- Overcoming head pose variation by proposing a robust feature coding
- Dealing with natural situations including illumination changes and partial occlusion.

We propose a novel spatio-temporal descriptor based on relative movements of the different facial regions which is robust to facial pose variations. The feature encoding may still fail when dealing with extreme head pose variations where some parts of the face are not visible from the recorded images. To address this problem and also dealing with illumination variations and occlusion, we incorporated the idea of sparse representation where noisy data can be reconstructed from clean data in the form of designed dictionary. Consequently, we propose to construct a dictionary to be both reconstructive and discriminative for a robust recognition system. The idea of joint dictionary learning and classifier training helped us to achieve the goal.

In contrast with most of the existing research work that updates dictionary and classifier alternately via iteratively solving each sub-problem, we propose to directly solve them jointly. To the best of our knowledge, this is the only work that attempts to simultaneously learn the sparse representation of the signal and train a *nonlinear* classifier to be discriminative for sparse codes. The proposed method jointly learns a single dictionary (which may not be necessarily over-complete) and also an optimal nonlinear classifier.

The key contributions of the present work are as follows:

- Proposing a pose-invariant optical flow (OF)-based descriptor which is able to robustly detect the facial emotions even when there are some head movements while expressing an emotion.
- Introducing two kinds of histogram for condensing the extracted features (weighted and un-weighted histogram) to be able to detect both the level and dynamic of expressions.
- Proposing a new classification framework by adding the Extreme Learning Machine (ELM) error term into the objective function of the conventional sparse representation to learn a dictionary which is both discriminative and reconstructive.
- Solving the objective function of our proposed Extreme Sparse Learning (ESL) which contains both linear and no-linear terms, via a novel approach named Class Specific Matching Pursuit (CSMP).
- Extending the framework by introducing Kernel ESL (KESL).

1.4 Organization of the Thesis

We organized the rest of the chapters as follows:

Chapter 2 summarizes the literature review of the recent advances in the research on human affect recognition. We review the researches in five different parts. The first part explores a few models proposed for emotion. The second part surveys the vision based emotion recognition techniques including static and dynamic approaches. The third part is a review of pose invariant techniques, and the next part is a survey of sparse representation based classification. Finally, we review the facial emotion databases commonly used in this field and describe four databases utilized in the thesis.

Chapter 3 introduces the various feature extraction methods developed versus those state-of-the-arts. After explaining each method, the corresponding experimental results are presented. The advantages and disadvantages of each method are also discussed in this chapter.

Chapter 4 introduces our contributions related to robust representation of facial emotions. We first describe our main methodology for pose-invariant feature extraction and then present the experiments conducted to evaluate the proposed descriptor. The result of the proposed pose-invariant descriptor is also compared to our other feature extraction developed and the other state-of-the-art methods.

Chapter 5 presents the robust emotion classification framework. In this chapter, we first review the Extreme Learning Machine (ELM) classifier, then give the background information for sparse representation and Dictionary Learning (DL), and describe related techniques in this field. Finally, our proposed methodology for robust classification method that combines both

---

the ELM and sparse representation into a single optimization is presented. Then, the experimental results for all databases are presented and discussed.

Chapter 6 presents the conclusion of the thesis recommendations for future work.

Chapter 2

A Literature Review

2.1 Introduction

Extensive researches have been previously reported in automated emotion analysis and related fields. In this chapter we briefly review the literatures in 4 parts. The first part is related to emotion modeling. The rest is a survey on recent research work based on facial modality to access human emotional states. Reviews of some successful techniques for vision-based emotion recognition, pose invariant feature extraction, and sparse representation based classification approaches are presented in this chapter. Finally, we review the facial emotion databases and describe those used in this thesis for evaluation of the methods.

2.2 Emotion Models

Theoretical modelling of the structure of emotions is the first step to develop an emotion recognition system. To analyse emotion, two different views are reported in the literatures: categorical and dimensional affective analysis.

It is notable that affect, emotion, and feeling are three different concepts which are usually used interchangeably. Shouse [20] defined these concepts as “Feeling is a sensation that has been checked against previous experiences and labelled. It is personal and biographical because every person has a distinct set of previous sensations. Emotion is the projection/display of a feeling. Unlike feelings, the display of emotion can be either genuine or feigned. Affect is a key part of the process of an organism’s interaction with stimuli. It is the body’s way of preparing itself for action in a given circumstance by adding a quantitative

---

dimension of intensity to the quality of an experience.” However, affect and emotion refers to the same concept in this thesis.

2.2.1 Categorical Model

Categorical model classifies emotions into discrete categories such as anger, disgust, fear, happiness, sadness and surprise which are considered as six universal emotions. The universality of these emotions was initially investigated by Ekman [21]. Although there are a lot of researches reported on six basic emotions, the other emotions like interest, frustration, puzzlement and boredom have also been studied recently [15]. The main benefit of categorical scheme of emotions is that human usually use these categories to express their emotional states in real life. The problem with this view is that choosing only these categories of emotion is too restrictive because they cover only a small range of our daily emotional expressions, whereas real human - human interaction may contain blended emotions.

2.2.2 Dimensional Model

Dimensional representation considers multiple dimensions to describe an emotional state. The theory of dimensional modelling of emotion was mainly introduced by Russel [22]. Arousal and valence are two common dimensions that may construct a 2D emotion model to describe different emotional state. One example of Arousal-Valence model is shown in Figure 2-1 where the emotional states are distributed in a 2D plane: x-coordinate (valence) indicates the degree of pleasure which is extended in the range of “negative” to “positive” and y-coordinate (arousal) measures the intensity of affective experience and is characterized in the range of “passive” to “active” [23]. Most researches using 2D dimensional representation simplify the emotion recognition problem into four classes (quadrants of 2D space) or two classes (positive-negative or active-passive) [15].

Unlike the categorical model, dimensional representation is able to label a wide range of emotions [15]. Additionally, dimensional spaces can provide the emotional intensity which can be used for evaluating the level of emotional reactions to outside stimulus. However, the projection of high-dimensional emotional states into a rudimentary 2D space results in loss of information. For instance, as shown to Figure 2-2, some emotions become indistinguishable (e.g., fear and anger).

The majority of existing work on emotion recognition focus on dimensional emotions and less attention has been paid to dimensional model. Some recent studies attempt to detect the human emotional states in dimensional spaces [24, 25]. Based on psychological findings that suggest arousal and valence are two correlated spaces, Nicolaou et al. [24] proposed a

classification scheme for continuous affect prediction. Zhang et al. [25] presented an evaluation framework for fusion of texture and geometric descriptors for dimensional representation of spontaneous emotions.

A few researchers investigated different models to explore a better model to represent the emotional states but there is no clear conclusion. Lichtenstein et al. [26] compared two emotion models for affect recognition from physiological data. The authors designed their experiment with respect to both modelling of user's expression: categorization and dimensional models. They found the most important physiological parameters (such as "heart rate", "breathing rate", and "number of deep breaths") are related to different emotional state in both models.

Figure 2-1: Illustration of Arousal-Valence model (Extracted from [23])

Figure 2-2: Schematic diagram of the dimensional models of emotions with common basic emotion categories overlaid. (Extracted from [27])

2.3 Visual-based Emotion Recognition

As face plays an essential role in emotion expression, most of the vision-based affect analysis research focused on facial expression recognition. Nowadays, there are many reported prior researches on emotional facial analysis systems inspired by the Facial Action Coding System (FACS) which is the most comprehensive, widely used and best accepted system proposed by Ekman et al. [28]. FACS is based on the anatomic study that is able to measure all visually discernible facial movements called Action Units (AUs). The upper face AUs and their combinations are shown in Figure 2-3 [29]. For example, AU1 is the action of raising the inner brows while AU2 is the action of raising the outer brows.

AUs in this system can be used to detect not only six basic emotions, but also variety of other affective states as well as complex physiological behaviours such as depression. Indeed, FACS is suitable for studying the natural and spontaneous facial behaviour that composed of a variety of facial movements. The richness of the database as well as ability to code all facial movements is the key advantage of this system. However, frame by frame coding of facial movements to AUs is highly manual intensive [2].

Tian et al. [29] developed an automatic face analysis system which can categorize facial expression changes into AUs, instead of predefined basic emotions. Their approach has been

---

proposed for tracking and modelling the various facial features including both permanent facial features (eyes, brows, mouth and cheeks) and transient facial features (deepening of facial furrows).

Automated facial expression recognition systems can also be classified into two categories: target- and gesture-oriented [30]. The target-oriented systems select a single key frame involving the peak facial expression from a given sequence of images. Gesture-oriented approaches use temporal information in the sequence by tracking some specific facial feature points over all frames of a video sequence. Some researches on machine analysis of emotions conducted on single frame believe that this approach avoids the computational complexity required for temporal analysis while preserves the necessary information for training the system [17, 31]. However, the crucial aspect of target-oriented (image-based) approaches is finding the key frame from a video sequence. The fundamental idea of selecting the key frames in several research work is based on the general observation of human subjects which suggests that facial features will be exaggerated at large voice amplitude [31]. Malakia et al. [17] used a semi-supervised clustering technique to select the key visual frames. Although the later proposed method is computational expensive, it is able to extract key frames of a mute video clip and does not require the speech signal.

In the following subsections, we divided our literature survey based on either the static or dynamic analysis used for emotion interpretation.

<i>NEUTRAL</i>	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 2-3: Upper face Action Units and some combinations. (Extracted from [29])

2.3.1 Static Analysis of Emotion (Target-based Approaches)

Most of the existing facial expression analysis algorithms employed different pattern recognition approaches. The fundamental step for pattern recognition is the feature extraction process. In the case of target-based approaches, the process of extracting the facial expression information is addressed as localization of face and its features. The widely used facial features are geometric and appearance-based features. Geometric features are considered either the shape of facial components (eyes, eyebrows, nose and mouth) or the location of facial landmarks (corner of eyes, brows, lip etc.), while the appearance features represent the facial texture (wrinkles, furrows and bulges) [15].

Gabor wavelet, Haar-like wavelet and Local Binary Pattern (LBP) are three examples of appearance feature-based methods. The majority of appearance feature-based researches have focused on Gabor-wavelet representation. Filtering face images with a bank of Gabor filters is both time and memory intensive, but it is still widely used because of its effective performance on affect analysis. The study of Bartlett et al. explored the fully automatic spontaneous facial expression detection via Gabor features [32]. Their proposed system is able to operate in real-time by applying AdaBoost for feature selection following classification by Support Vector Machine (SVM). Then, the classification output of sequenced frame was considered as a

function of time which could be used for dynamic expression measurement. However, classification of each video frame independently is not able to take into account the temporal information of the expressions.

Whitehill and Olmin [33, 34] used some Haar-like features selected via boosting algorithm which was proposed by Viola and Jones at first for face detection [33, 34]. They evaluated this approach (Haar+AdaBoost) both in classification accuracy and computational time and then compared it to the standard Gabor+SVM approach and showed that both methods have about the same recognition rate, but the first approach is faster.

Kim et al. [35] extended Haar features for emotion recognition purpose, because the Viola-Jones rectangle features were initially suggested for face detection. In this work, novel rectangle features were proposed to distinguish between two similar images.

LBP is another example of appearance based features which was used by Shan et al. [8]. In this research, it was illustrated that LBP feature can be useful for low resolution facial expression analysis which is common in real world application. On the other hand, the most important advantage of LBP feature is its computationally simplicity and robustness against illumination changes [8].

Senechal et al. [36] applied the LBP operator on Gabor filtered image and named the new descriptor as Local Gabor Binary Pattern (LGBP). They claimed that this combination makes the method very robust to misalignment and illumination variations.

Local Directional Number (LDN) pattern is a novel appearance feature introduced by Rivera et al. [37] for face analysis and facial emotion recognition. LDN encodes the facial texture's structure in a compact way to produce a discriminative and robust feature against noise and illumination variations.

2.3.2 Dynamic Analysis of Emotion (Gesture-oriented Approaches)

The majority of existing researches in this field have focused on facial expression processing via static image data and ignored the temporal information of such a dynamic event [8-12]. However, the human visual system is demonstrated to have better judgment about an expression when its temporal information is taken into account [13]. Following this fact, some techniques have been developed to deal with dynamic expression recognition. In the case of an image sequence (gesture-oriented approaches), the problem is classified as tracking the face and its features [38]. The temporal information is considered in dynamic processing systems either in the feature extraction part or during the classification stage. Typical examples of such techniques are Hidden Markov Models (HMM) [28, 39], Dynamic Bayesian Networks (DBN) [40], dynamic texture descriptors [6], and geometrical displacement [41].

There were several reported attempts to track the facial expression over time for emotion recognition via Hidden Markov Models (HMM). A multilevel HMM is introduced by Cohen et al. [42] to automatically segment the video and perform emotion recognition. Their experimental results indicated that multilevel HMM have better performance than the one layered HMM. Cohen et al. introduced a new architecture of HMMs for automatic segmentation and recognition of human facial expression from live videos [43].

Dynamic Bayesian Networks (DBN) is another successful method for sequence-based expression analysis. Ko and Sim [44, 45] developed a facial expression recognition system based on combining the Active Appearance Model (AAM) for feature extraction and DBN for modelling and analysing the temporal phase of an expression. They claimed that their proposed approach is able to achieve robust categorization of missing and uncertain data and temporal evolution of the image sequences.

Optical Flow (OF) is also a widely used approach for facial features tracking and dynamic expression recognition. Cohn et al. [46] developed an OF based approach to automatically discriminate the subtle changes in facial expression. They considered sensitivity to subtle motion when designing the OF which is crucial for spontaneous emotion detection. Tariq et al. [47] used an ensemble of features including both appearance and motion features for FAUs detection. Their proposed OF based motion features were extracted for seven regions of interest by computing the mean and variance of the OF components for each region. The head motion was also captured at this work from the nose region OF.

Methods based on local features or interest points such as Scale Invariant Feature Transform (SIFT) have shown to perform well for object recognition and then extended to video analysis [47-49]. Camara-Chavez and Araujo [50] proposed a method for event detection in a video stream by combining Harris-SIFT with motion information in the context of human action recognition. They used the Harris corner detection for key-point extraction and the phase correlation method was used to measure the motion information.

Zhao and Pietikainen [6] presented a successful dynamic texture descriptor based on the Local Binary Pattern (LBP) operator and applied it on facial expression recognition as a specific dynamic event. Their proposed dynamic LBP descriptors were calculated on Three Orthogonal Planes (TOP) of the video volume, resulting in LBP-TOP descriptor. Local processing, simple computation and robustness to monogenic gray-scale changes are the advantages of their method. Following their idea, Almaev and Valster [51] developed the LGBP descriptor to spatio-temporal volumes to combine spatial and dynamic texture analysis of facial expressions. Similarly, Bishan et al. [52] proposed an extension of Local Phase Quantization (LPQ) to a dynamic texture descriptor for AU detection. All these work concluded that such kind of dynamic appearance descriptors outperform the static ones.

Dollar et al. [53] developed a general framework for dynamic behaviour detection from videos by proposing descriptors to encode the spatio-temporal cuboids surrounding the points of interest. Extracted cuboids are clustered to form a dictionary of cuboid prototypes and then the information of location and type of cuboid prototypes is kept for further processing. They argued that the proposed representation is robust to many data variations. Their experimental results on different databases including facial expression and human activity show that their method is applicable for these tasks.

Guha and Ward [54] explored the effectiveness of sparse representations in the context of facial expressions and human action recognition. They extracted a set of spatio-temporal descriptors named Local Motion Pattern (LMP) to obtain key points of video sequences. A compact and rich representation was then suggested by learning the over complete dictionary and its corresponding sparse model. Their work presented a new local spatiotemporal feature that is distinctive, scale invariant, and fast to compute.

The work of Pantic et al. [55] is an example of geometric-feature-based methods. In this work, 15 facial points around the eyes, eye-brows, nose, mouth and chin were tracked to interpret the observed expression. They used particle filtering and temporal rules to track facial features in the video input. The proposed method was invariant to occlusion such as facial hair or glasses and also robust to illumination changes. Their work is the only study on automatic affect analysis based on profile-view of face image sequences. However, variation in viewing angle of the face, which is common in spontaneous behaviour, was not taken into account.

Park and Kim [56, 57] proposed a method for accurate facial emotion detection using motion. They extracted 70 facial feature points in the sequenced facial images using Active Appearance Model (AAM) and then tracked them to estimate the motion vectors of 27 feature points. Their proposed approach mapped the subtle facial expressions into the exaggerated expressions by magnifying the estimated motion vectors. Then, SVM classifier was used to classify the exaggerated expressions.

The system proposed by Kotsia and Pitas [41] was also based on geometric deformation features. Their system was initialized using default grid nodes at the first frame of the sequence and then a tracking approach based on deformable model was used to track the facial expression over time. The geometrical feature displacement was then used as an input to SVM classifier.

Although all these methods solved the challenge of dynamic analysis of facial expression by considering the temporal information of the event, however none of them are designed to handle variation of the pose.

2.4 Pose Invariant Feature Extraction

Facial feature extraction is a vital step for robust emotion recognition. Although facial expression recognition has been extensively studied in the past, most of the existing approaches focus on the frontal facial images. Thus even small changes in the facial pose may reduce the effectiveness of the existing methods. Only a few researchers have attempted to solve the facial pose challenge.

A probabilistic method based on 2D geometrical features was proposed by Rudovic et al. [58] for pose invariant facial expression recognition. The locations of 39 landmarks were extracted from an expressed facial image with an arbitrary head pose. The coupled scaled Gaussian process regression model was then applied to normalize the facial pose. This approach was claimed to be better than the state-of-the-art methods for head pose normalization. Furthermore, it was also claimed to be the first work that is able to deal with -45 to +45 degrees head pan rotation and -30 to +30 degrees tilt rotation. Although the model was trained based on only a few discrete head poses, the method was able to deal with continuous head pose variation within the above-mentioned limits. But the method requires high-quality capture of the facial feature points which is very challenging for automatic emotion recognition.

Jeni et al. [59] proposed to use the facial shape information as a robust representation to pose variations. The 3D landmark positions were estimated on face images using constrained local models at the first stage. Then, the rigid transformation from the obtained 3D shape was removed. Finally, SVM classifier was applied to the projected shape into 2D space. Although the shape information was shown to be efficient for facial expression detection, the pose-invariant ability of the proposed system was evaluated by only two simple posed facial emotion databases.

Songfan and Bhanu [60] introduced a homogeneous reference face model called avatar reference to capture the nature of the whole dataset. Then, a video sequence with any length was condensed into a single image representation. This representation is able to aggregate the temporal facial emotion information and also compensate the large rigid head motion as well as removing the person-specific information.

A novel representation approach for facial images using the regional covariance matrix was proposed by Zheng et al. [61]. A dimensionality reduction step is then applied to the resulting features based on the theory of discriminant analysis. An effective approach was further proposed to find the optimal discriminant vectors. The key advantage of this method is that it does not require any facial alignment and feature point localization, which are both challenging tasks.

A simple method named variable-intensity template was proposed by Kumano et al. [62] to obtain a person specific model for describing various facial expressions. The variable intensity templates define how the intensity of multiple facial points varies for an observed expression. The method is able to detect the facial expression and estimate the pose simultaneously within the framework of a particle filter. Simple modelling and low computational cost are considered as the advantages of this method. However, the method is quite sensitive to errors in interest point localization and misalignment.

Since the dynamics of facial expression is crucial for a reliable facial emotion analysis, a variety of approaches focus on motion and OF based feature extraction [63, 64]. However, none of them is suited for pose-invariant purpose.

2.5 Sparse Representation Based Classification

Our proposed approach for classification is very similar to techniques that try to simultaneously solve the sparse representation problem along with other constraints like classification error minimization or discriminative dictionary learning.

While Dictionary Learning (DL) directly from the training data usually leads to a satisfactory reconstruction, adding a specific discriminative criterion to dictionary training can improve the discriminative ability of the method and lead to better classification results. Recently, several methods have been developed to train a classification oriented dictionary. These methods can be divided into three broad categories. The first category of methods directly forces the dictionary atoms to be discriminative and use the reconstruction error for the final classification [65, 66]. The second approach makes the sparse coefficients discriminative by incorporating the classification error term into the dictionary learning and indirectly propagates the discrimination power to the overall dictionary [67-72]. The third category includes methods that apply the discriminative criterion for coefficients, but the classifier is not necessarily trained along with DL. They either use the reconstruction error based classification or employ other classifiers on the resulting sparse representation [72, 73].

An example of the first approach is the scheme proposed by Ramirez et al. [65] for learning the class specific sub-dictionaries by incorporating a penalty term to make the sub-dictionaries incoherent. Since the incoherence term was directly applied on the dictionaries, the method is applicable both in the supervised and unsupervised settings.

Another example of the first category is the classification-oriented DL model proposed by Wang et al. [66], which learns a class-specific dictionary (named particularity) to capture the most discriminative information of each category, and also a common pattern dictionary (named commonality) that only contributes the essential representation for all the data. The

---

reconstruction approach is used for classification and the authors in [66] claim that their method is able to discover the shared patterns among different categories. Promising results have been reported in various applications including scene classification, face recognition, handwritten digit recognition and object classification.

Most of the approaches proposed in the literature for sparse representation based classification (including the one proposed in this thesis) fall under the second category, where the classifier is simultaneously trained along with DL.

Mairal et al. [67] introduced a supervised DL method by incorporating a logistic loss function to the classical reconstructive term to simultaneously learn a classifier. Later, this work was extended to produce a general formulation of supervised DL that can be applied to a wide variety of problems and an efficient algorithm was also proposed to solve the corresponding optimization criterion [68]. This approach was evaluated on a variety of tasks including handwritten digit classification, digital art identification, nonlinear inverse image problems, and compressed sensing to demonstrate the effectiveness of the approach in large-scale settings.

In work done by Zhang and Li [69], a discriminative dictionary-learning approach called Discriminative K-Singular Value Decomposition (D-KSVD) was proposed by introducing a discriminative term into the conventional objective function of K-SVD. The dictionary learned by this method is guaranteed to be both reconstructive and discriminative.

Similarly, a supervised sparse coding using Label Consistent K-SVD (LCKSVD) algorithm was proposed in [70, 71] to train a discriminative dictionary. The class label of the training data and the associate label information of each dictionary atom are utilized in the dictionary learning process. This algorithm incorporated sparse coding error and classification error criterion into a unified objective function, which was optimized using the K-SVD algorithm. Indeed, it learns an over-complete discriminative dictionary and a linear classifier simultaneously. Compared to the methods that learn the dictionary and classifier separately, the above algorithm efficiently learns a compact discriminative dictionary and a multiclass linear classifier. However, the method cannot be directly extended to learn a nonlinear classifier, which is required when the data is not linearly separable.

A good example of third category is the Fisher Discriminative Dictionary Learning (FDDL) proposed by Yang et al. [72]. In this method, dictionary learning based on Fisher discrimination criterion is used to improve the classification performance. The method aims to learn a structured dictionary, where the class labels are presented to the sub-dictionaries. The criterion imposed on sparse coding causes the sparse coefficients have small within-class scatter but large between-class variance. Indeed, this method improved the classification performance by using both discriminative reconstruction error and sparse coding coefficients.

Compared to other state-of-the-art approaches, FDDL was shown to have competitive performance in various pattern recognition tasks.

A theoretical scheme for signal classification by combining a reconstructive approach with a discriminative term using linear discriminant analysis and a predefined dictionary was proposed by Huang and Aviyente [73]. The results indicate that the method is superior to the standard classification approaches accompanied with standard sparse coding for classification of corrupted and noisy signals.

Recently, He et al. [74] presented a novel approach for fast face recognition using sparse coding and ELM. The basis function was first constructed based on common feature hypothesis from the randomly universal images, and then an ELM is trained to learn the corresponding sparse codes. The resulting sparse coefficients of all the face images were further fed into the next ELM for testing stage. The method is claimed to have good performance comparable to the classical sparse coding in terms of accuracy and time complexity. But it should be further explored for general applications and more challenging scenarios to release the advantages.

To the best of our knowledge, none of the existing methods can learn a non-linear classifier in the context of simultaneous sparse coding and classifier training. Learning such a non-linear classifier is not only an interesting research topic, but also very important in many real-world applications where the observations are not probably linearly separable. This work is the first research work that explores how to simultaneously learn the sparse representation of the signal and train a non-linear classifier to be discriminative for sparse codes.

2.6 Facial Expression Databases

The first requisite in designing an automatic emotion recognition system is collecting a rich labeled database. The genuine emotion expressions are very difficult to record because they are rare, subtle, temporary, and context dependent. Additionally, manual labeling of the spontaneous emotions is very time consuming task which may contain many errors. Due to these problems, most of the research work on facial emotion recognition is based on posed and deliberate expression of emotions. A survey of the facial emotion databases can be found in [75]. Table 2-1 summarizes several common used emotion databases which are publicly available.

In this section, we only describe the databases used in this thesis which are based on image sequences. Four databases have been used for the facial expression recognition; three of them are publicly available and one of them is our own collected data.

The first database used in our study is the CK⁺ [76] dataset consisting of acted emotional data under controlled environment.

The second database is obtained by adding our own collected data, which includes pose variations, illumination changes, and occlusion situations, to the original CK⁺ database. We refer to this expanded database as Extended CK⁺ (ECK⁺) and this database is used to evaluate the robustness of our proposed approach under difficult environment settings.

The third dataset used for evaluation is the Audio Visual Emotion Challenge (AVEC 2011) [77] database, which contains spontaneous emotional states in naturalistic situation.

The Emotion Recognition in the Wild (EmotiW) database [78] is the fourth database we used to evaluate our methods. This dataset contains realistic challenges such as pose variations, various illumination conditions, occlusion, and spontaneous emotional expressions.

Table 2-1: Common used facial emotion databases. NA stands for Not Available.

Database	Image/Video	# Emotions	# Subjects	Posed/Natural	Year
EmotiW [78]	Video	7	NA	Natural	2013
AVEC [77]	Video	4	NA	Natural	2011
BU-3DFE [79]	Image	7	100	Posed	2006
BU-4DFE [80]	Video	6	101	Posed	2008
CK ⁺ [76]	Video	7	123	Posed	2010
CK [81]	Video	6	97	Posed	2000
MMI [82]	both	6	52	Posed	2005
FERET [83]	Image	2	1199	Posed	1996
JAFFE [84]	Image	7	10	Posed	1998

2.6.1 CK⁺ Database

The Cohn-Kanade (CK⁺) dataset [76] contains 593 video sequences recorded from 123 university students ranging from 18 to 30 years old. In this database, the subjects expressed a series of 23 facial displays including single or combined action units. Six of the displays are

labelled as prototype basic emotions (joy, surprise, anger, fear, disgust, and sadness). In this work, we used all the 309 sequences from the dataset that have been labelled with at least one of the six basic emotions. Figure 2-4 shows sample frames of the database expressing basic emotions.

Figure 2-4: Example of basic emotions from CK⁺ database. Each image is the last frame of a video clip that shows the most expressive face.

2.6.2 Extended CK⁺ Database

Since the robustness of our proposed algorithms can be evaluated only on a challenging dataset, we recorded 75 video sequences in total under different head pose, illumination conditions and facial occlusion. Our own collected database includes 42 samples of head pose variations, 15 samples of illumination changes, and 18 samples of facial occlusion. Three subjects participating in our defined scenarios were asked to show one expression (anger, happiness, sadness, and surprise) from neutral to apex. Some sample frames from our collected database for one of the subjects are depicted in Figure 2-5. The labelling is based on the subject's impression of each 4 basic emotion categories and then checked by an observer. The locations of nose points and face cropping have been done manually.

Figure 2-5: Sample frames of our own collected data. The left column shows some examples of pose variation, the middle column depicts occlusion examples, and the right column includes illumination variations.

2.6.3 AVEC 2011 Database

The AVEC2011 database is a more challenging database collected under natural settings. This database consists of 95 videos recorded at 49.979 frames per second at a spatial resolution of 780×580 pixels. The binary labels along the four affective dimensions (activation, expectation, power and valence) are provided for each video frame. Activation

(arousal) measures the intensity of lethargy or dynamism. Valence indicates the degree of pleasure. Power dimension involves two related concepts; power and control. It is mainly characterized by vocal and action tendency reaction in relation to social experience of dominance. Expectation dimension also involves several concepts such as expecting, anticipating, and being taken by surprise.

The data is divided into 3 subsets: training, development, and testing. The training subset consists of 31 records, while the development subset (for validation of model parameters) consists of 32 sequences and the test subset consists of 11 video sequences. Figure 2-6 shows a few sample frames of this database.

The original labelling of the database is in the form of continuous affect binned in temporal units corresponding to each frame. Then, the binary labelling was provided by comparing each value to the mean value as a threshold. An example of binary labelling of a sample video is shown in Figure 2-7.

Figure 2-6: Sample video frames from AVEC2011 database. (Extracted from [85])

Figure 2-7: Binary labelling of the four affective dimensions (activation, expectation, power and valence) in a sample of AVEC2011 video database. (Extracted from [85])

2.6.4 EmotiW Database

EmotiW is an extended form of the Acted Facial Expressions in the Wild (AFEW) database [86]. The database is a collection of short video clips collected from some popular movies, where the actor is expressing one of seven emotions (anger, disgust, fear, happy, neutral, sad, and surprise) under near real-world conditions. EmotiW consists of three sets for training, validation, and testing including 380, 396, and 312 video clips respectively.

Many challenges have been encountered using this database:

- Since the range of pose variations is quite large, available face detectors like the Viola-Jones algorithm [34] often fail to detect the correct faces.
- Many video clips consist of more than one human subject, which makes it difficult to isolate them from the subject of interest. For example, Figure 2-8 shows two sample frames with more than one person. As shown, multiple faces are detected by face detector.
- There is a wide difference in the way that the same emotion is expressed by the various subjects. Some of expressions are very confusing and hard for even a human expert to detect the true class of emotions. Figure 2-9 shows a few sample frames with providing labelling from the database.

Figure 2-8: Sample frames with more than one human subject from EmotiW database. Multiple faces are detected by the face detector.

Figure 2-9: Sample frame sequences with provided labels from EmotiW database.

Chapter 3

Dynamic Feature Extraction for Emotion Recognition

3.1 Introduction

Deriving a proper facial representation from a sequence of images is crucial to the success of a facial expression recognition system, especially if the application requires continuous processing of the video stream such as in the case of human-computer interaction. Since facial expression is a dynamic event, it could benefit from the recent advances in the dynamic analysis of video related to human activities such as gesture recognition, sports-event analysis, abnormal action recognition, and video surveillance [87-90]. However, there are many challenges related to real-world unconstrained situations, such as finding suitable spatio-temporal volumetric features to efficiently analyze the video sequence.

In this chapter, we present some of our preliminary ideas for feature extraction which are able to solve a few of existing issues in this field. Indeed this chapter presents our initial exploration of features that showed promise but eventually not selected in our final experiments as none of them could deal with pose-invariant issue which is our main challenge to resolve in this thesis. We include them at this part as we will compare the performance of different approaches in the experiment section.

3.2 Method I: Extended Spatio-Temporal Histogram of Oriented Gradients

In this work, a novel local spatio-temporal descriptor is proposed for motion pattern detection in addition to appearance feature [91]. The proposed feature comprises histogram of 3D gradients and the gradients' variation over time to robustly describe the spatial and temporal information. It outperforms the basic HOG as it takes into account both spatial and temporal information. It also incorporates spatio-temporal pyramid structure to handle different resolution and frame rate. To reduce the dimension of the feature, we applied Genetic Algorithm (GA) for region-based feature selection.

The novelties of this framework are:

- 1) Extending the spatio-temporal histogram of oriented gradients to extract both static and dynamic information from a video sequence by considering the 3D oriented gradients instead of 2D local descriptor. We also extended the feature set by adding the information of gradient variations in time domain. The dynamics of gradient orientation are also informative for video processing and we showed that there exist similar patterns of gradient orientation variation for each group of facial expression.
- 2) Proposing the spatio-temporal pyramid decomposition using an edge preserving smoothing filter to extend the proposed method to deal with multi-resolutions or multiple sampling of frame rates. This will allow the proposed approach to be able to tolerate variation in scale and speed.
- 3) Selecting the more informative and discriminative facial regions by GA. Since some parts of the face are not crucial for emotion detection, the processing of the whole face is not required.

The framework of our proposed approach is shown in Figure 3-1. The first step is preprocessing the video sequences to be ready for feature extraction. Two sets of spatio-temporal local descriptors are then calculated for each video volume: Spatio-Temporal Histogram of Oriented Gradients (STHOG), and Spatio-Temporal Histogram of Orientation Variations (STHOV). STHOG descriptor encodes local edges of multiple orientations in 3D, while STHOV feature estimates the motion of a dynamic event in time domain. The final feature vector is obtained by concatenating both extracted features and is named Extended Spatio-Temporal Histogram of Oriented Gradients (ESTHOG). The third stage is incorporating the 3D pyramid concept into the framework to ensure that our method is able to withstand variations in the scale and speed by calculating multiple spatial and temporal resolution levels. Then feature selection using GA is utilized to reduce the dimension of the extracted feature.

The mentioned steps are described in details in the following subsections.

Figure 3-1: Block-diagram of the proposed approach.

3.2.1 Feature Extraction

In this section, local volumetric features are proposed to describe the local regions of sequenced images which is akin to 3D data. Figure 3-2 illustrates the outline of our proposed descriptors. The local regions are determined by dividing the volumetric data into predefined number of 3D blocks as shown in Figure 3-2a. To preserve the geometric information of descriptors, each block is divided into predefined number of 3D cells as illustrated in Figure 3-2b. Finally, 3D gradients are calculated for each pixel of a cell, and then a sub-histogram of orientation is computed by quantization of 3D gradients (Figure 3-2c).

A block-based approach is used to combine the extracted information from pixel-level, region-level, and volume-level. The video sequence can be partitioned using overlapping or non-overlapping blocks.

We propose two sets of features: STHOG, and STHOV. STHOG descriptor encodes the local edges in spatio-temporal domain. STHOV feature describes how the orientations change in a period of time. Thus the STHOV descriptor is expected to complement the descriptive property of STHOG. The final descriptor is the combination of two extracted features. Details of computation these two local descriptors are presented in the following subsections.

Figure 3-2: Descriptor computation; (a) The volume data is divided into a number of 3D grids. Each grid is denoted by a block (B_i). The final descriptor (F) consists of the block's feature; (b) Each block is divided into number of 3D cells (C_j). The block feature consists of the cell's histograms; (c) Each cell includes a number of points (P_k) which are characterized by a 3D gradient vector. Each gradient orientation and its variation in time domain (z axis) are quantized to compute the histogram of a cell.

A. Spatio-Temporal Histogram of Oriented Gradients (STHOG)

In order to compute the histogram of 3D gradient orientations, gradient operators are needed.

The 2D gradient magnitude and orientation is defined as follow:

$$M_{xy} = \sqrt{G_x^2 + G_y^2}, \quad \theta = \arctan\left(\frac{G_y}{G_x}\right) \quad (3-1)$$

where G_x and G_y are the result of convolving the images at each time with the horizontal and vertical 3×3 gradient filters as follows:

$$G_x = I(x, y) * \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad (3-2)$$

$$G_y = I(x, y) * \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \quad (3-3)$$

where $I(x, y)$ is the image pixel and $*$ is the convolution operator.

For 3D gradient magnitude and orientation, 3D gradient filter is convolved with the sequenced images as follows:

$$M_{xyz} = \sqrt{G_x^2 + G_y^2 + G_z^2}, \quad \varphi = \arctan\left(\frac{M_{xy}}{G_z}\right) \quad (3-4)$$

$$G_z = I(x, y, t) * S_z \quad (3-5)$$

where S_z is a 3D gradient kernel in z direction.

$$S_z(:, :, -1) = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, S_z(:, :, 0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, S_z(:, :, 1) = \begin{bmatrix} -1 & -2 & -1 \\ -2 & -4 & -2 \\ -1 & -2 & -1 \end{bmatrix}, \quad (3-6)$$

Each pixel of a sequenced data is encoded by two angles θ and φ which represent the orientation of 3D gradients. To construct a weighted histogram of orientations similar to conventional HOG [92], θ and φ are quantized into equally sized bins.

For each pixel, two weighted votes are calculated separately to construct two histograms based on the orientation of the gradient elements centred on it. The votes which are weighted by 2D and 3D magnitudes (M_{xy} and M_{xyz}) are accumulated into orientation bins over local spatio-temporal regions called cells. Cells can be either rectangular or radial. The orientation bins are spaced equally over 0° - 180° (the gradient's sign is ignored).

The next step is grouping cells into larger spatial segments named blocks. Each block descriptor comprises the cell histograms which are concatenated. The block histogram contrast normalization is then applied to achieve illumination invariance. The final descriptor is then the concatenated form of all normalized block histograms.

B. Spatio-Temporal Histogram of Orientation Variation (STHOV)

In addition to orientation information, the changes of orientation over time are useful for temporal analysis of an event. For analysis of facial expression, we observed that there exists similar pattern of changes of orientation over time for each class of emotion. This fact motivated us to use the histogram of orientation variations as a new feature.

The 3D orientation variations are defined as follows:

$$\delta\theta = |\theta(x, y, t) - \theta(x, y, t - 1)| \quad (3-7)$$

$$\delta\varphi = |\varphi(x, y, t) - \varphi(x, y, t - 1)| \quad (3-8)$$

where $\theta(x, y, t)$ and $\varphi(x, y, t)$ are the pixel spatio-temporal orientations in the current frame, and $\theta(x, y, t - 1)$ and $\varphi(x, y, t - 1)$ are the pixel spatial orientation in the previous frame.

To define a histogram for $\delta\theta$ and $\delta\varphi$, the magnitude of orientation variation is needed. δM_{xy} is defined by Eq. (3-9) to weight the votes of the bins corresponding to $\delta\theta$ and δM_{xyz} is defined as Eq. (3-12) to weight the votes of the bins related to $\delta\varphi$.

$$\delta M_{xy} = \sqrt{F_1^2 + F_2^2} \quad (3-9)$$

$$F_1(x, y, t) = G_y(x, y, t - 1)G_x(x, y, t) - G_x(x, y, t - 1)G_y(x, y, t) \quad (3-10)$$

$$F_2(x, y, t) = G_x(x, y, t - 1)G_x(x, y, t) + G_y(x, y, t - 1)G_y(x, y, t) \quad (3-11)$$

where G_x and G_y are defined by Eq. (3-2) and Eq. (3-3) respectively..

$$\delta M_{xyz} = \sqrt{E_1^2 + E_2^2} \quad (3-12)$$

$$E_1(x, y, t) = M_{xy}(x, y, t - 1)G_z(x, y, t) - G_z(x, y, t - 1)M_{xy}(x, y, t) \quad (3-13)$$

$$E_2(x, y, t) = G_z(x, y, t - 1)G_z(x, y, t) + M_{xy}(x, y, t - 1)M_{xy}(x, y, t) \quad (3-14)$$

where M_{xy} and G_z are defined by Eq. (3-1) and Eq. (3-5) respectively.

Similar to HOG, two angles, $\delta\theta$ and $\delta\varphi$, which represent the changes of orientation over time can describe a new feature set. To construct a weighted histogram of orientation changes, $\delta\theta$ and $\delta\varphi$ are quantized into equally sized bins. If θ and φ are in the range of 0° - 180° , then $\delta\theta$ and $\delta\varphi$ are defined in range of 0° - 90° considering only the smaller angle in the orientation difference.

3.2.2 Spatio-Temporal Pyramid Decomposition

Spatio-temporal pyramid transform is an effective way for multi-resolutions or multi-sampling rates analysis of an event. In spatial pyramid, each pixel in the low resolution is obtained by down sampling of its previous lowpass filtered image. For temporal pyramid, the same procedure is considered by lowpass filtering and subsampling of the original sequenced data in time domain. In other words, temporal pyramid contains multiple frame rates of the original video sequence.

The spatio-temporal pyramid generation can be constructed as follows:

$$G_l(x, y, t) = \begin{cases} f(x, y, l) & \text{for } l = 1 \\ \sum_m \sum_n \sum_p k(m, n, p) G_{l-1}(R_x x + m, R_y y + n, R_t t + p) & \text{for } l > 1 \end{cases} \quad (3-15)$$

where l represents the pyramid level, $f(x, y, l)$ denotes the original data. R_x , R_y , and R_t are the down sampling rates in x- y- and t-direction respectively while $k(x, y, t)$ is a spatio-temporal low-pass filter. Since our proposed feature is like edge-based descriptor, the filter should be chosen appropriately. Spatio-temporal bilateral filtering is a good choice that smooths the images in both spatial and time domains while preserving the edges. The method is based on combination of nearby pixel values by considering both the geometric distance and photometric similarity [93].

The bilateral filtering at a pixel location x of an image, $F(x)$, can be expressed as follows [94]:

$$F(x) = \frac{1}{C} \sum_{y \in N(x)} e^{-\frac{\|y-x\|^2}{2\sigma_d^2}} e^{-\frac{|I(y)-I(x)|^2}{2\sigma_r^2}} I(y) \quad (3-16)$$

where σ_d and σ_r are geometric and photometric spread parameters which control the amount of filtering in spatial and intensity domain. $N(x)$ is a spatial neighborhood around pixel $I(x)$ and C is the normalization factor defined as follow:

$$C = \sum_{y \in N(x)} e^{-\frac{\|y-x\|^2}{2\sigma_d^2}} e^{-\frac{|I(y)-I(x)|^2}{2\sigma_r^2}} \quad (3-17)$$

Note that the first term in Eq. (3-17) measures the geometric closeness between the neighboring pixels and center pixel and second term indicates the photometric similarity.

Spatio-temporal bilateral filter is an extension of the 2D bilateral filter that can be applied to video processing. It replaces each pixel value of a 3D data with an average of similar and nearby pixel values in a 3D volume.

3.2.3 Feature Selection

The dimension of the extracted features after 3D pyramid decomposition increases with the number of decomposition levels. Feature reduction is applied to sieve out the useful features. As the face images are divided into blocks, some of them may not be very relevant for facial

expression analysis and can be dropped. We used GA for feature selection as it is able to find the subsets of features that are collectively useful, instead of searching the effectiveness of a single feature individually [95].

GA starts by randomly generating a population of individuals which are candidate solutions, then tries to optimize the population of next generations by three processes named “reproduction”, “crossover”, and “mutation” [95]. It also requires a fitness function to provide a measure to evaluate the individuals of each population.

In this study, each individual which is a binary string, represents a subset of predefined video blocks. Each binary digit (gene) stands for the presence (1) or absence (0) of the blocks. The length of individuals should be set based on the total number of blocks. The detection rate of the classifier is used as the fitness function.

3.2.4 Experimental Results

This section presents our experimental results for the proposed dynamic descriptor on CK⁺ database. Note that all proposed methods in this thesis were implemented using Matlab 7.11.0 running on a Core i5 CPU (2.8 GHz with 16 GB RAM).

We used polynomial SVM as a classifier for all methods as the basic classifier for fair comparison. SVM has been originally proposed for binary categorization, and then developed for multi-class problems [96]. For CK⁺ database, we used one-against-all technique that constructs 6 binary SVM classifiers to categorize each emotion against all the others. Classification of a new instance is done by a winner-takes-all strategy, where the classifier with the highest output function assigns the class. Regarding the parameter selection of SVM, we carried out grid-search on the parameters as suggested in [8].

To evaluate our proposed approach, Leave-One subject-Out (LOO) cross validation was used. For the database with N subjects, we performed N experiments. For each experiment, the samples of $N-1$ subjects are used for training and the remaining samples for testing. Finally, the true detection rate is estimated as the average classification accuracy on test samples. Using this method, our experiments will be subject independent, because there is no information of the same subject in both training and test samples.

The pre-processing stage includes face alignment and segmenting the face region from the background. Since the facial landmarks’ locations are given in the used dataset, the images are aligned to a constant distance between the two eyes and rotated to align both eyes horizontally. Finally, the faces are cropped using a rectangle of size 100×100. Figure 3-3 illustrates the different steps for face pre-processing. The location of nose point, which is our reference point, is also provided in the database.

Figure 3-3: Overview of the face pre-processing; (a) The original face sample; (b) Normalized face based on constant distance between two eyes; (c) Rotated face based on eye coordinates.

Since the length of our proposed descriptors depends on the number of blocks and cells that we defined for partitioning the volume data, we did some experiments with different settings. Table 3-1 tabulates the feature length and detection rate of our experiments using different number of blocks and cells. The result shows that partitioning the data into 75 blocks ($5 \times 5 \times 3$) and 9 cells ($3 \times 3 \times 1$) gives the best result and is used in our subsequent experiments.

Table 3-1: Effect of number of blocks and number of cells on detection performance of ESTHOG feature.

# Blocks	# Cells	# Features	Detection rate (%)
$2 \times 2 \times 2$	$2 \times 2 \times 2$	2304	86.80
$2 \times 2 \times 2$	$3 \times 3 \times 1$	2592	91.32
$2 \times 2 \times 2$	$4 \times 4 \times 2$	9216	92.83
$4 \times 4 \times 2$	$2 \times 2 \times 2$	9216	90.47
$4 \times 4 \times 2$	$3 \times 3 \times 1$	10368	92.50
$4 \times 4 \times 2$	$4 \times 4 \times 2$	36864	92.42
$5 \times 5 \times 3$	$2 \times 2 \times 2$	21600	92.97
$5 \times 5 \times 3$	$3 \times 3 \times 1$	24300	94.52
$5 \times 5 \times 3$	$4 \times 4 \times 2$	86400	91.27

Another factor that affects the dimension of feature and classification performance is the number of bins used to quantize the gradient orientations and gradient orientation variations (θ , φ , $\delta\theta$, and $\delta\varphi$). We performed our experiment considering 9 bins for each angle and evaluate the effect of different number of blocks and cells on the detection rate. After experimenting with various block sizes, cell sizes, and overlapping ratios, the overlapping between adjacent blocks used is 30% of the original non-overlap block size for all subsequent experiments.

The block histogram normalization is an essential factor for achieving good results in our experiments. After evaluating different normalization schemes, L2-norm was selected to normalize each block descriptor [92].

The second experiment was conducted to evaluate the performance of the proposed descriptor individually and in combination to check whether the second descriptor (STHOV) improves the descriptive property of STHOG. As Table 3-2 shows, the detection rate of ESTHOG which is the concatenated histogram of STHOG and STHOV has higher accuracy compared to each individual descriptor.

Table 3-2: Comparison the results of different features including STHOG, STHOV, and ESTHOG. The results are based on 75 blocks ($5 \times 5 \times 3$), and 9 cells ($3 \times 3 \times 1$).

Descriptor	Detection rate (%)
STHOG	93.19
STHOV	87.57
ESTHOG	94.52

The next experiment is related to spatio-temporal pyramid representation. Table 3-3 tabulates the results of two experiments done on 4 and 5-level pyramid decompositions to compare the computational cost and classification detection rate. As we can see, the detection rate of 4-level pyramid decomposition is better. As the CK⁺ database contains subjects with only a few number of video frames (the minimum number of frames is 5); higher level of temporal pyramid is not used.

Feature reduction is the next step of our experiment. We applied GA with selected parameters as described in Table 3-4 for feature selection of 4-level pyramid ESTHOG. Since the proposed framework consist of many blocks to describe different facial components, the initial experiments illustrated that some blocks do not contribute to express the facial emotional states. The GA is designed to evaluate and select the blocks which have more contribution to facial expression recognition.

Regarding the generalization capability of the proposed method, the GA was applied only on the training data. In other words, we applied the GA for $N-1$ subjects (training data for LOO cross validation) to select the more discriminative blocks. Then the selected features were evaluated on the test data. Since the selected blocks for N trials are close to each other, we can treat the result of any trail as the final selected features.

The comparison results of the method before and after feature selection are given in Table 3-5. A 4-level pyramid ESTHOG as shown in this table consists of 128 blocks and 66816 features in total. The number of selected blocks using GA is reduced to 60, a reduction of

more than 3 times (21816 features). As a result, the classification accuracy increased from 93.44% to 96.10%.

The experimental results show that STHOG and STHOV are complementary descriptors that are able to combine the appearance and motion features. The inclusion of spatio-temporal pyramid analysis to describe the features at multiple resolutions and various facial expression speeds, though does not improve the performance, allow the proposed approach to be robust to natural variations. Furthermore, feature selection using GA is able to detect useful blocks and improve the recognition performance of face emotion recognition.

Table 3-3: Comparison the results of different pyramid levels. The first row is the result of 2spatial+1temporal levels of pyramid decomposition. Second row is for 2spatial+2temporal levels of pyramid decomposition.

Method	# Block	# Feature	Detection rate (%)
4 level Pyramid ESTHOG	128	66816	93.44
5 level Pyramid ESTHOG	160	152064	92.18

Table 3-4: GA parameters for block selection

Used parameters	value
Population size	10
No. of the genes	128
Crossover rate	0.7
Mutation rate	0.01

Table 3-5: Comparison the results of different methods. The first row is the results of ESTHOG descriptor without any pyramid decomposition. Second row is the results after 4 level pyramid decompositions. Third row shows the results of 4 level pyramid of the descriptor after feature selection.

Method	# Block	# Feature	Detection rate (%)
ESTHOG	32	36864	92.42
Pyramid ESTHOG	128	66816	93.44
Pyramid ESTHOG+GA	60	21816	96.10

3.2.5 Discussion on Method I

In this work, we introduced a novel method for analysis of dynamic events from video sequences. An extended spatio-temporal HOG was proposed to describe the motion features in addition to appearance features. The initial STHOG descriptor encoded the local edges in spatio-temporal domain to describe both appearance and motion features. STHOV descriptor

which analyzes the orientation variation was developed to complement the STHOG. The final descriptor, ESTHOG, is obtained by concatenating both descriptors. Subsequently, a 3D pyramid scheme using spatio-temporal bilateral filtering was proposed to ensure that the descriptor is able to handle variations in scale and speed of the video sequence. The method is very fast and simple to implement.

The limitation of the proposed method is that it is sensitive to misalignment error. As such if the subject moves his/her head while expressing an emotion, the frames should be re-aligned correctly. Failing to deal with head pose variations is another disadvantage of the method.

3.3 Method II: Multi-Scale Analysis of Local Phase and Local Orientation

This proposed descriptor is based on histograms of local phase and local orientation of gradients obtained from a sequence of face images to describe the spatial and temporal information of the face images [97]. The descriptor is able to effectively represent the temporal local information and its spatial locations which are important cues for facial expression recognition. This is further extended to multi-scale to achieve better performance in natural settings where the image resolution varies.

In this proposed approach, a novel phase-based descriptor is presented to process the local structures of sequenced images. We also extend our feature set by considering the complementary information of local orientations. The final descriptor is the concatenation of spatio-temporal histogram of local phase and spatio-temporal histogram of local orientation.

The novelty of our proposed method includes:

- (1) Extending the phase-based descriptor for spatio-temporal event analysis.
- (2) Formulating the 3D local orientation of features as additional information to represent all local structures.
- (3) Combining the spatio-temporal histogram of local phase and local orientation to extract both static and dynamic information as well as spatial information of a face from a video sequence.
- (4) Incorporating multi-scale analysis for better performance in natural settings with varying image resolution.

Local phase and local energy are two important concepts used to describe the local structural information of an image. Local phase is able to effectively depict useful image structures such as transitions or discontinuities, providing the type of structure and location information while preserving the image structures. Another advantage of phase-based feature

is that it is not sensitive to intensity variation. Local energy, as a complementary descriptor to local phase, contains the strength or sharpness of the feature [98].

The concept of local phase and energy was originally proposed for 1 dimension (1D) signal analysis. For 1D signal, orientation is trivial and does not contain other additional information. Therefore, the local structure is totally described by local phase and energy. However, for higher dimensional signal (3D in our case), local orientation is needed in addition to local phase and energy, to completely describe the features in the signal.

To extend the concept of local analysis to 3D, a quadrature pair of oriented bandpass filters is used. However, since the oriented filter bank is discrete, thus proper filter selection is required to cater to different orientation and scale present in an image. This motivate researchers to use the monogenic signal concept [98-99].

The monogenic signal provides an isotropic extension of the analytic signal to 3D by introducing a vector-valued odd filter (Riesz filter) with Fourier representation [98] as:

$$\begin{aligned} H_1(u, v, w) &= i \frac{u}{\sqrt{u^2 + v^2 + w^2}}, \\ H_2(u, v, w) &= i \frac{v}{\sqrt{u^2 + v^2 + w^2}}, \\ H_3(u, v, w) &= i \frac{w}{\sqrt{u^2 + v^2 + w^2}} \end{aligned} \quad (3-18)$$

where u , v , and w are the Fourier domain coordinates and i represents the imaginary part of the signal.

The monogenic signal is then represented by combining the original 3D image with the Riesz filtered components:

$$f_M(x, y, z) = [f_{M,1}(x, y, z), f_{M,2}(x, y, z), f_{M,3}(x, y, z), f_{M,4}(x, y, z)] \quad (3-19)$$

where $f_{M,1}$, $f_{M,2}$, $f_{M,3}$, and $f_{M,4}$ are defined as follows:

$$f_{M,1}(x, y, z) = f(x, y, z) * g(x, y, z) \quad (3-20)$$

$$f_{M,2}(x, y, z) = f(x, y, z) * g(x, y, z) * h_1(x, y, z) \quad (3-21)$$

$$f_{M,3}(x, y, z) = f(x, y, z) * g(x, y, z) * h_2(x, y, z) \quad (3-22)$$

$$f_{M,4}(x, y, z) = f(x, y, z) * g(x, y, z) * h_3(x, y, z) \quad (3-23)$$

$*$ is the convolution operation, and h_1 , h_2 and h_3 are the spatial domain representations of H_1 , H_2 and H_3 , respectively. The 3D image is first filtered using a bandpass filter g such as Gaussian, Gabor, and Log-Gabor filters. We used an isotropic (no orientation selectivity) log-Gabor filter as defined by Eq. (3-24) as such filter can be designed with arbitrary bandwidth and the bandwidth can be optimised to produce a filter with minimal spatial extent [100].

$$G(w) = \exp\left(-\frac{\left(\log\left(\frac{w}{w_0}\right)\right)^2}{2\left(\log\left(\frac{k}{w_0}\right)\right)^2}\right) \quad (3-24)$$

where w_0 is the filter's centre frequency, parameter k controls the bandwidth of the filter. Figure 3-4 illustrates the monogenic components of a sample video frame.

The monogenic signal can also be represented in the same form as 1D analytic signal:

$$f_{MG} = \text{even}_{MG}(x, y, z) + i \times \text{odd}_{MG}(x, y, z) \quad (3-25)$$

where the even and odd filter response are define as follows:

$$\text{even}_{MG}(x, y, z) = f_{M,1}(x, y, z) \quad (3-26)$$

$$\text{odd}_{MG}(x, y, z) = \sqrt{f_{M,2}(x, y, z)^2 + f_{M,3}(x, y, z)^2 + f_{M,4}(x, y, z)^2} \quad (3-27)$$

With the additions, the 3D monogenic signal can be used for feature extraction.

Figure 3-4: Illustration of the monogenic components for one sample image of a video sequence; (a) Original sample frame; (b) First monogenic component (f_1); (c) Second monogenic component (f_2); (d) Third monogenic component (f_3); (e) Forth monogenic component (f_4).

3.3.1 Methodology

To recognize facial expressions from video sequences, a set of features that best describe the facial muscle changes during an expression is required. Two complementary feature sets are proposed in this study: Spatio-Temporal Histogram of Local Phase (STHLP) and Spatio-Temporal Histogram of Local Orientation (STHLO). The final feature set is formed by concatenating both STHLP and STHLO which we named Local Phase-Local Orientation (LPLO).

A. Spatio-Temporal Histogram of Local Phase (STHLP)

The local energy and local phase of the 3D monogenic signal given by Eq. (3-25) is computed as follows:

$$E(x, y, z) = \sqrt{(\text{even}_{MG}(x, y, z))^2 + (\text{odd}_{MG}(x, y, z))^2} \quad (3-28)$$

$$Ph(x, y, z) = \text{atan}\left(\frac{\text{odd}_{MG}(x, y, z)}{\text{even}_{MG}(x, y, z)}\right) \quad (3-29)$$

where E and Ph denote local energy and phase respectively. even_{MG} and odd_{MG} are even and odd parts of the monogenic signal.

In this work, a novel local volumetric feature is proposed to describe the local regions of a video sequence. Each pixel of a video frame is encoded by a phase angle. To construct a weighted histogram of local phase, the phase is quantized into equal sized bins. Energy of each pixel is then used as a vote to the corresponding phase bin. Figure 3-5a shows the image of energy of a sample video frame. The votes belonging to the same bin are accumulated over local spatio-temporal regions that we call cells. To reduce the noise, only the pixels that have energy more than a pre-defined threshold can participate in the voting. The next step is grouping cells into larger spatial segments named blocks. Each block descriptor is composed by concatenating all the cell histograms within that block. The block histogram contrast normalization is then applied to get a coherent description. The final descriptor of the frame is then obtained by concatenating all the normalized block histograms of that frame.

It is worth noting that our proposed spatio-temporal descriptor is able to handle different length of the video sequence. It can also support video segments of varying lengths.

B. Spatio-Temporal Histogram of Local Orientation (STHLO)

We consider the local orientation as complementary information to describe the local structures of the video sequence. Each pixel orientation of the sequenced images is described by two angles as follows:

$$\theta(x, y, z) = \text{atan}\left(\frac{f_{M,2}(x, y, z)}{f_{M,3}(x, y, z)}\right) \quad (3-30)$$

$$\varphi(x, y, z) = \text{atan}\left(\frac{\sqrt{(f_{M,2}(x, y, z))^2 + (f_{M,3}(x, y, z))^2}}{f_{M,4}(x, y, z)}\right) \quad (3-31)$$

To construct a weighted histogram of orientations similar to STHLP, θ and φ are quantized into equally sized bins. For the vote's weight, pixel contribution can be used as the magnitude of the orientation vector given by the following equations:

$$Mag_{\theta} = \sqrt{(f_{M,2}(x, y, z))^2 + (f_{M,3}(x, y, z))^2} \quad (3-32)$$

$$Mag_{\varphi} = odd_{MG}(x, y, z) \quad (3-33)$$

where Mag_{θ} is used to vote the bins related to θ and Mag_{φ} for voting the bins corresponding to φ . The voting components (Mag_{θ} and Mag_{φ}) are depicted in Figure 3-5b and Figure 3-5c respectively for a sample video frame. Such proposed method is efficient as it utilizes the magnitude information already computed previously. This is different from the common approach used is to compute the local orientation, based on the output of ensembles of oriented filters like Prewitt, Sobel, and Laplace at differing orientation [92, 101-103].

Figure 3-5: Illustration of the components used to vote the phase and orientation bins for one sample image of a video sequence; (a) Energy signal for voting the phase bins; (b) Mag_{θ} for voting the bins related to θ ; (c) Mag_{φ} for voting the bins corresponding to φ .

3.3.2 Multi-Scale Analysis

Multi-scale or multi-resolution analysis of facial features has been used in many researches [104, 105]. Since some facial features are detectable at a certain scale and may not be as distinguishable at other scales, it is more reliable to extract the features at various scales. The multi-resolution representation can be achieved by varying the wavelengths of the bandpass filter in Eq. (3-24). Finally, the LPLO is obtained from the combination of the features at different scales. Figure 3-6 illustrates how the energy component defined by Eq. (3-28) varies by changing the filter scale.

Figure 3-6: Illustration of the energy component at different scales. The scale of the bandpass filter is increased from (a) to (d).

3.3.3 Experimental Results

This section presents our experimental results for the proposed dynamic descriptor on two databases; CK⁺ and AVEC2011.

A. Results for CK⁺ database

The first experiment in this section was carried out using different log-Gabor parameters to check the effect of filter wavelength and bandwidth on the classification performance. We again used polynomial SVM classifier in our experiments and report the results based on LOO cross validation. Our experimental results show that the bandwidth of 0.75 and 3 scales of the bandpass filter with wavelengths of {4, 8, 12} are superior to other settings in term of detection rate. The results are shown in Table 3-6 and Table 3-7 respectively. These parameters are then fixed in our subsequent experiments.

The next experiment was conducted on different number of blocks and cells to compare the length of features and the detection rate. Table 3-8 tabulates the results obtained. Based on this experiment, partitioning the data into 32 blocks (4×4×2) and 9 cells (3×3×1) outperforms the other settings in term of classification accuracy.

We also evaluated both feature sets (STHLP and STHLO) individually and in combination to validate whether they are indeed complementary descriptors. The combined feature is named LPLO in this table. The results of our evaluation are summarized in Table 3-9. The detection rate of the combined features is better than each feature set individually.

Table 3-6: Effect of log-Gabor bandwidth on classification accuracy for CK⁺ database. The results are based on 50 blocks (5×5×2), and 9 cells (3×3×1).

$\frac{k}{w_0}$	Detection rate (%)
0.55	91.94
0.65	92.24
0.75	94.58
0.85	93.47

Table 3-7: Effect of log-Gabor scales on classification accuracy for CK⁺ database. The results are based on 50 blocks (5×5×2), and 9 cells (3×3×1).

# Scales	Wavelength	Detection rate (%)
2	{4, 8}	91.24
2	{8, 16}	92.47
3	{4, 8, 12}	94.58
3	{8, 12, 16}	93.22
4	{2, 8, 12, 16}	92.86
4	{8, 12, 16, 32}	91.43

Table 3-8: Comparison on number of blocks and number of cells on classification accuracy.

# Blocks	# Cells	#Features	Detection rate (%)
3×3×2	2×2×1	5184	87.41
3×3×2	3×3×1	11664	80.64
4×4×2	2×2×1	9216	94.16
4×4×2	3×3×1	20736	94.58
5×5×2	2×2×1	14400	93.28
5×5×2	3×3×1	32400	93.57

Table 3-9: Comparison of STHLP, STHLO, and combined features.

Descriptor	Detection rate (%)
STHLP	91.23
STHLO	89.12
Combined feature(LPLO)	94.58

B. Results for AVEC2011 Database

We also evaluated our proposed approach using the AVEC 2011 database which is more challenging as it is captured in a natural setting. The information describing the position of the face and eyes are provided in the database. The pre-processing stage includes only normalizing the faces to have a constant distance between the two eyes.

As indicated in the baseline results of this challenge [77], the dataset contains a large amount of data (more than 1.3 million frames). Due to processor and memory constraints, we sample the videos at a constant sampling rate. We partition each video into segments containing 60 frames with 20% overlap between the segments. Each segment is then down sampled at a rate of 6. Thus, each volume data includes only 10 frames. We process only 1550 frames of each video for the training and development subsets (total of 48050 frames for training and 49600 frames for development).

Table 3-10 shows the recognition results of our approach compared to the other reported results. We just reported the weighted accuracy of the methods which is the correctly detected samples divided by the total number of samples. The average results obtained by our method are above the baseline results and also [34], [35] for the development subset. For test subset, we achieved the best average accuracy among all competitors. This means that the proposed descriptor can be effective for natural and spontaneous emotion detection in natural setting.

Table 3-10: Comparison of the detection rate for the AVEC 2011 database. “A” stands for activation, “E” for expectancy, “P” for power, and “V” for valence.

Reference	Development					Test				
	A	E	P	V	Average	A	E	P	V	Average
Baseline [77]	60.2	58.3	56.3	63.6	59.6	42.2	53.6	36.4	52.5	46.2
[106]	81.7	73.1	73.3	73.5	75.4	65.5	61.7	47.1	69.8	61.0
[107]	69.3	65.6	59.9	67.8	65.6	56.5	59.7	48.5	59.2	55.9
[108]	58.2	53.5	53.7	53.2	54.65	56.9	47.5	47.3	55.5	51.8
LPLO+SVM	65.3	68.4	55.5	55.5	61.2	75.6	59.8	56.9	56.9	62.3

3.3.4 Discussion on Method II

In this work, we proposed a novel local descriptor to analyze dynamic facial expression from video sequences. Our novel descriptor composed of two feature sets, STHLP and STHLO, to describe the local phase and orientation information of the structures in the images. The good performance of the method is due to:

1. Encoding the local structural information using local phase concept.
2. Formulating the 3D local orientation of features as additional information to represent all local structures.
3. Combining the spatio-temporal histogram of local phase and local orientation to extract both static and dynamic information as well as spatial information of a face from a video sequence.
4. Incorporating multi-scale analysis for better performance in natural settings with varying image resolution.

Our proposed phase-based descriptor provides a measure that is independent to the signal magnitude, making it robust to illumination variations. However, it is not robust to head pose variations and facial misalignment error.

3.4 Method III: Histogram of Dominant Phase Congruency

We also propose a novel spatio-temporal descriptor based on Phase Congruency (PC) concept and applied it to recognize facial expression from video sequences [109]. The proposed descriptor named HDPC comprises histograms of dominant PC over multiple 3D orientations to describe both spatial and temporal information of a dynamic event.

There are some advantages for PC-based feature extraction approaches over gradient-based techniques [110]. The gradient operators such as Prewitt, Sobel, Laplace, and Canny edge detector may fail to precisely identify and localize all image features, especially in region affected by illumination changes. Unlike the gradient-based approaches which look for sharp changes of image intensity, PC is a dimensionless quantity which is robust to image contrast and illumination changes.

In Figure 3-7, we show the advantages of PC-based line detection over Canny and Sobel methods. This figure illustrates that PC is able to localize the sharp line similar to the gradient operators. However, for features that are not sharp (gradual intensity variation), PC is able to detect such feature better than the traditional gradient operators as shown in Figure 3-7. Indeed, PC captures the discontinuities even at small intensity differences which might be missed by the typical image gradient-based edge descriptors. It can thus be useful for facial features detection including skin folds due to aging and expression.

To construct HDPC descriptor, the spatio-temporal PC values are calculated for multiple orientations. Therefore, each pixel of a video is characterized by multiple oriented PC values. Thus the PC values are able to encode various features at different scales and orientations for both spatial and time domains. After calculating the oriented PC values, the next step is to find the maximum PC for each pixel while preserving the dominant orientation. In other words, each pixel is represented by a vector where its length is equal to maximum PC, and its direction is determined by the dominant orientation. Keeping the dominant PC and its orientation information will preserve the key feature contributing to a dynamic event. The final step of our novel descriptor is building a local histogram of PC directions over all pixels over a spatio-temporal patch.

More precisely, the proposed approach involves two main steps: (1) calculating the spatio-temporal PC at various orientations, and (2) encoding the pixel's dominant oriented PC. We describe the details of each stage in the following subsections.

Figure 3-7: Comparison of methods for line detection; (a) sharp line; (b) line detection based on Phase Congruency; (c) line detection based on Canny; (d) line detection based on Sobel; (e) Gradual line with intensity range of [0 3]; (f) Line detection based on Phase Congruency; (g) Line detection based on Canny; (h) Line detection based on Sobel.

3.4.1 Spatio-Temporal PC Calculation

The monogenic signal framework provides the extended form of analytic signal to 3D by using a vector-valued odd filter (Riesz filter) which is represented in Fourier domain as defined by Eq. (3-18). The monogenic signal f_M is then calculated as follow:

$$f_M(x, y, z) = [f(x, y, z) * g(x, y, z), f(x, y, z) * g(x, y, z) * h_1(x, y, z), f(x, y, z) * g(x, y, z) * h_2(x, y, z), f(x, y, z) * g(x, y, z) * h_3(x, y, z)] \quad (3-34)$$

where f is the original signal, h_1 , h_2 and h_3 are the spatial domain representations of Riesz filter components, g is a bandpass filter, and $*$ denotes convolution operation. Indeed, the 3D image is first filtered using a bandpass filter such as log-Gabor filter. An oriented 3D log-Gabor filter is defined by the following:

$$G(w, \varphi, \theta) = \exp\left(-\frac{\left(\log\left(\frac{w}{w_0}\right)\right)^2}{2\left(\log\left(\frac{k}{w_0}\right)\right)^2} + \frac{(\varphi - \varphi_0)^2}{2\sigma_\varphi^2} + \frac{(\theta - \theta_0)^2}{2\sigma_\theta^2}\right) \quad (3-35)$$

where w_0 is the filter's centre frequency, parameter k controls the bandwidth of the filter, φ_0 and θ_0 denote the filter angles, and σ_φ and σ_θ are the respective filter spread.

Now, based on the original PC formulation proposed by Kovessi [110], the phase congruency for multiple orientations and scales of log-Gabor filter is defined as:

$$PC_o(x, y, z) = \frac{|E_o(x, y, z) - T_o|}{\sum_{sc} \sqrt{(even_o^{sc}(x, y, z))^2 + (odd_o^{sc}(x, y, z))^2} + \varepsilon} \quad (3-36)$$

where o and sc represent the orientation and scale variables respectively. PC_o is the phase congruency for a specific orientation and symbol $[\ ]$ denotes that the enclosed quantity is not allowed to be negative. $even_o^{sc}(x, y, z)$ and $odd_o^{sc}(x, y, z)$ are the even and odd components of the monogenic signal (defined by Eq. (3-26) and Eq. (3-27)) for a specific orientation and scale. E_o and T_o are the orientation specific energy and noise threshold respectively which can be calculated by Eq. (3-37) and Eq. (3-38), and ε is a small offset to avoid division by zero.

$$T_o = \sum_{sc} \exp^{mean[\log(\sqrt{(even_o^{sc}(x, y, z))^2 + (odd_o^{sc}(x, y, z))^2})]} \quad (3-37)$$

$$E_o(x, y, z) = \sqrt{(\sum_{sc} even_o^{sc}(x, y, z))^2 + (\sum_{sc} odd_o^{sc}(x, y, z))^2} \quad (3-38)$$

To detect the image features at various orientations, a bank of oriented log-Gabor filters are designed. However, the multi-orientation representation for the PC calculation may cause high dimensionality and expensive computational time. Therefore, we selected the dominant orientation where the PC value is the maximum and then designed a proper feature representation.

3.4.2 HDPC Feature

This section describes the proposed local volumetric feature. The outline of the proposed descriptor is illustrated in Figure 3-8. The first step for HDPC feature extraction is the partitioning of volumetric data into local regions. The local regions are determined by dividing the data into predefined number of 3D blocks as shown in Figure 3-8a. To preserve the geometric information of the descriptors, each block is divided into predefined number of 3D grids named cells as illustrated in Figure 3-8b. The block-based approach is used to combine the extracted information from pixel-level, region-level, and volume-level. The video sequence can be partitioned using overlapping or non-overlapping blocks. Then, for each pixel

of a cell, multiple 3D PCs at various orientations are calculated using Eq. (3-36). As Figure 3-8c shows, each pixel is characterized by several oriented PC components.

The next step is the selection of dominant orientation where the PC component is maximum. Therefore, each pixel of the 3D data is represented by an orientation and a maximum PC as shown in Figure 3-8d. Since we construct our descriptor based on the dominant PC for each cell, it will be less sensitive to noise.

Finally, the sub-histogram for predefined number of orientations (bin numbers) can be constructed. Each bin will be weighted by its dominant PC. On the other hand, for each orientation which is defined for log-Gabor filter, we accumulate the dominant PCs over each cell. The block feature consists of the cell's histograms.

The final HDPC descriptor is the concatenation of all the block's features. Therefore, the proposed spatiotemporal descriptor is able to handle different sequence length, allowing the use of variable length video segments which is common in real applications.

Figure 3-8: Descriptor computation; (a) the volume data is divided into a number of 3D grids. Each grid is denoted by a block (B_i). The final descriptor (F) consists of the block's feature; (b) each block is divided into number of 3D cells (C_j). The block feature consists of the cell's histograms; (c) each cell includes a number of points (P_k) which are characterized by several oriented PC components; (d) a PC component with dominant orientation is selected for each pixel and then used to compute the histogram of a cell.

3.4.3 Experimental Results

We evaluated our proposed method on two databases for facial expression recognition task; CK⁺ and AVEC2011. For classification, SVM with polynomial kernel function has been again used in the experiments.

A. Results for CK⁺ Database

We carried out the first experiment on different settings including various numbers of blocks and cells to evaluate its effect on feature dimension and classification detection rate. The results are tabulated in Table 3-11. Based on the results, partitioning the volume data into 75 blocks (5×5×3) and 9 cells (3×3×1) outperforms the other settings in term of classification detection rate.

For bandpass filtering with log-Gabor filter defined by Eq. (3-35), proper parameter setting is important to have acceptable results. A 16-oriented filter bank (4 values for θ ranged over 0^0 - 180^0 , and 4 values for φ ranged over 0^0 - 180^0) are found to be suitable for our experiment. The effect of parameter $\frac{k}{w_0}$ on the classification performance is also important.

Table 3-12 gives the effect of parameter $\frac{k}{w_0}$ on the classification performance. Based on this table, the value of 0.85 produced the best result. We also did an experiment to evaluate the log-Gabor wavelength on classification detection rate. As Table 3-13 shows, setting 3 scales of the bandpass filter with wavelengths of {8, 12, 16} is superior to the other settings in term of detection rate.

To show the importance of temporal information for facial emotion recognition, we compared the proposed spatio-temporal descriptor with the spatial HDPC applied on last frame of each sequence which is the most expressive face. For the spatial HDPC, we used 2D instead of 3D Gabor filter and then find the dominant PC over 2D cells of the image. In this experiment, 6 orientations are set for 2D Gabor filter. As Table 3-14 shows, the result of spatio-temporal HDPC is around 10.41% better than spatial HDPC. This shows that the role of temporal information of the emotion detection is useful.

We also tested the ability of the proposed descriptor for emotion detection from small scale and low resolution videos. We down sampled a recorded signal of happy expression from own collected data with 5 sampling rate (1/2, 1/4, 1/6, 1/8, and 1/10). We again used the down sampled sequences to evaluate our classifier trained using the original CK⁺ database. We observed that the classifier could detect the true label of all down sampled sequences (100% accuracy). The result shows the ability of the method to be used for low resolution video analysis.

Table 3-11: Effect of number of blocks and number of cells on detection performance of HDPC feature for CK⁺ database.

# Blocks	# Cells	Features	Detection rate (%)
2×2×2	2×2×2	1024	84.73
2×2×2	3×3×1	1152	91.84
2×2×2	4×4×2	4096	91.92
4×4×2	2×2×2	4096	91.64
4×4×2	3×3×1	4608	92.25
4×4×2	4×4×2	16384	93.55
5×5×2	3×3×1	7200	93.99
5×5×3	3×3×1	10800	95.44
5×5×3	4×4×2	38400	95.13

Table 3-12: Effect of log-Gabor bandwidth on classification accuracy for CK⁺ database. The results are based on 75 blocks (5×5×3), and 9 cells (3×3×1).

$\frac{k}{w_0}$	Detection rate (%)
0.55	93.00
0.65	93.99
0.75	94.10
0.85	95.44

Table 3-13: Effect of log-Gabor scales on classification accuracy for CK⁺ database. The results are based on 75 blocks (5×5×3), and 9 cells (3×3×1).

# Scales	Wavelength	Detection rate (%)
2	{4, 8}	95.20
3	{8, 12, 16}	95.44
4	{2, 8, 12, 16}	95.20
5	{2, 8, 12, 16, 32}	91.43

Table 3-14: Comparison of spatial HDPC and spatio-temporal HDPC on CK⁺ database.

Method	Detection Rate (%)
Spatial HDPC	85.03
Spatio-temporal HDPC	95.44

B. Results for AVEC2011 Database

Table 4 shows the recognition results of our approach in comparison to other reported results. We only report the weighted accuracy of the methods which is the correctly detected samples divided by the total number of samples. The results obtained by our method are above

the baseline results and ranked only second compared to [Ramirez et al. 2011] in the test subset. This means that the proposed descriptor can be effective for natural and spontaneous emotion detection in natural setting.

Table 3-15: Comparison of the detection rate for the AVEC 2011 database. “A” stands for activation, “E” for expectancy, “P” for power, and “V” for valence.

Reference	Development					Test				
	A	E	P	V	Average	A	E	P	V	Average
Baseline [77]	60.2	58.3	56.3	63.6	59.6	42.2	53.6	36.4	52.5	46.2
[106]	81.7	73.1	73.3	73.5	75.4	65.5	61.7	47.1	69.8	61.0
[107]	69.3	65.6	59.9	67.8	65.6	56.5	59.7	48.5	59.2	55.9
[108]	58.2	53.5	53.7	53.2	54.65	56.9	47.5	47.3	55.5	51.8
HDPC+SVM	63.1	69.2	57.4	57.4	61.8	57.8	53.5	61.6	61.5	58.6

3.4.4 Discussion on Method III

In this work, we proposed a novel descriptor for dynamic visual event analysis which has several desirable properties. HDPC is a spatio-temporal descriptor which is able to describe the motion features in addition to appearance features by extending the PC to 3D and incorporating histogram binning. As such, it is also able to detect the features at different orientations and scales as well as robustness to illumination variation. Although the method is very accurate, it is sensitive to facial misalignments and cannot tolerate the head pose variations.

3.5 Concluding Remarks

In this chapter, we proposed our preliminary feature extraction methods which are able to solve some existing challenges in this field. The framework proposed in MethodI is able to handle variations in scale and speed of the video sequence. The method is very fast and simple to implement. Our proposed phase-based descriptor described as MethodII provides a measure that is independent to the signal magnitude, making it robust to illumination variations. The proposed descriptor presented as MethodIII is also able to detect the features at different

---

orientations and scales with high accuracy as well as robustness to illumination variation. However all proposed descriptors in this chapter are sensitive to head pose variations.

Chapter 4

Pose Invariant Feature Extraction for Emotion Recognition

4.1 Introduction

Facial emotion recognition is challenging not only due to the complexity of the expression which differs from one subject to another, but also due to the uncontrolled environments such as illumination variation, occlusion, and head movement. The accuracy of an emotion recognition system generally depends on two critical factors: 1) how to robustly represent the facial features in such a way that they are robust under intra-class variations (e.g. pose variations), but are distinctive for various emotions and 2) how to design a classifier that is capable of distinguish different facial emotions based on noisy and imperfect data. In this work, imperfect data refers to the cases with pose variations, occlusion, and illumination changes.

This chapter presents our main methodology for robust representation of facial emotions. We propose a novel spatio-temporal descriptor based on Optical Flow (OF) components which is very distinctive and also pose-invariant. The comparison of the pose-invariant descriptor and our proposed preliminary feature extraction methods is also made at the end of this chapter to present the advantages and disadvantages of the proposed methods.

4.2 Pose-Invariant Feature Extraction

Having understood the advantages and disadvantages of the various approaches above based on gradient, orientation, intensity and phase information of the image, we further propose a set

of pose-invariant features derived from the OF in videos. In order to compute the dynamic features, we start by computing the OF of a given video based on both brightness and gradient constancy assumption, combined with a discontinuity-preserving spatio-temporal smoothness constraint as described in the following subsection [111].

Let $U(P; t_i)$ represent the flow vector $(u; v)$ at pixel location $P = (x; y)$ at time t_i . Each pixel location is calculated with respect to the new coordinate system constructed using the nose vector as shown in Figure 4-1. The nose point is considered as the origin of this coordinate system. . Indeed, to construct the pose-invariant features, we require a stable reference point that does not move or change when an emotion is expressed. Thus the nose point is chosen. The algorithm¹ proposed in [112] is used for nose point detection. From this single point, we can derive the reference coordinate system as the vertical vector connecting the nose tip to the midpoint between the centres of the two eyes can be obtained. This is considered as the positive y-axis. Then the perpendicular vector pointing to the left is the positive x-axis. We used this reference coordinate to calculate the features and portioning the face into segments as will be described in the following sections. Actually other points can also be chosen if they are stable as well as can be used to derive the reference coordinate easily.

Figure 4-1: Reference vector and face coordinate system. The nose point is considered as the origin of the new coordinate system. The reference vector connects the nose point to the line joining the centre of the two eyes.

4.2.1 Optical Flow Calculation

OF is the pattern of apparently visual motion of objects caused by the relative motion between an observer and the scene. OF methods try to estimate the motion between two images taken at time t and $t + \Delta t$. For each pixel location (x, y, t) of an image, it is assumed

¹ The source code for nose point detection is available at:
<http://humansensing.cs.cmu.edu/intraface/download.html>.

that its intensity is constant over time, which can be expressed as following brightness constancy constraint:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \quad (4-1)$$

If assume the movement is small enough, the Eq. (4-1) can be developed in Taylor series as follows:

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\partial I}{\partial x} \Delta x + \frac{\partial I}{\partial y} \Delta y + \frac{\partial I}{\partial t} \Delta t \quad (4-2)$$

where higher order terms are ignored in Taylor series. From Eq. (4-1) and Eq. (4-2), it is concluded that $\frac{\partial I}{\partial x} \Delta x + \frac{\partial I}{\partial y} \Delta y + \frac{\partial I}{\partial t} \Delta t = 0$, which results in:

$$\frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0 \quad (4-3)$$

where V_x and V_y represent the OF components in X and Y directions. The Eq. (4-3) can be re-written as follow:

$$I_x V_x + I_y V_y = -I_t \quad (4-4)$$

where I_x , I_y , and I_t stand for image derivatives at x , y , and t respectively.

Since the Eq. (4-4) includes two unknowns (known as aperture problem), some additional constraint are needed to solve the problem. Many ideas have been proposed for OF estimation: starting from Lucas-Kanade approach [113] as well as Horn-Schunck method [114], researches have been dramatically increased to deal with the drawbacks of previous methods [115-118].

Several techniques have been assessed by Barron et al. [119] on several standard image sequences for both synthetic (with known ground truth motion fields) and real samples. The evaluated techniques are categorized as differential-based, region-based, energy-based, and phase-based approaches. They concluded that the phase-based method of Fleet and Jepson [120], and differential technique of Lucas-Kanade perform superior than others in overall. Another comparative study was also done by Galvin et al. [121] conducted on more synthetic sequences. The best obtained results in their study belonged to Lucas-Kanade technique. Although these comparative studies are worth, they were limited to only several basic methods and there is no comprehensive study to compare the recent techniques.

Here we introduce a successful robust method which used in our experiments

4.2.1.1 Brox Method

Brox et al. [111] proposed a method based on both brightness and gradient constancy assumption, combined with a discontinuity-preserving spatio-temporal smoothness constraint².

The energy function based on grey value constancy assumption and the gradient constancy assumption is measured by:

$$E_1 = \int (|I(X + w) - I(X)|^2 + \alpha |\nabla I(X + w) - \nabla I(X)|^2) dX \quad (4-5)$$

where α is a weigh between two assumptions. To minimize the effect of outliers on the estimation, new formulation was suggested as follow:

$$E_1 = \int \varphi(|I(X + w) - I(X)|^2 + \alpha |\nabla I(X + w) - \nabla I(X)|^2) dX \quad (4-6)$$

where $\varphi(s^2) = \sqrt{s^2 + \epsilon^2}$, for ϵ as a small positive constant for numerical reasons.

For the smoothness assumption, the second energy function is defined as:

$$E_2 = \int \varphi(|\nabla_3 u|^2 + |\nabla_3 v|^2) dX \quad (4-7)$$

where ∇_3 is the spatio temporal gradient operator and φ is the same function as defined in Eq. (4-6).

The total energy was defined as:

$$E = E_1 + \beta E_2 \quad (4-8)$$

where β is a regularization parameter. Now the goal is to find the flow components u and v such a way that the energy function E is minimized. This functional was suggested to be minimized by solving the associated Euler–Lagrange equations with help of numerical approximations.

Coarse-to-fine warping technique was also applied to improve the performance of estimated OF dealing with large displacements.

We found this method suitable for our application due to its robustness to grey value changes as it has added the gradient constancy assumption, is less sensitive to parameter setting, and is robust to noise. Additionally, the multi-scale approach used in this technique increases the performance of the flow field estimation in the presence of large displacements.

² The source code for OF extraction is available at:

<http://www.mathworks.com/matlabcentral/fileexchange/17500-high-accuracy-optical-flow>

4.2.2 Optical Flow Correction

Since we are only interested in the local motion of facial components resulting from the act of expressing an emotion, the global motion of the head is subtracted from flow vector.

$$OF_{exp} = OF_{tot} - OF_{head} \quad (4-9)$$

where U_{exp} is the expression-related OF that we aim to measure, U_{tot} is the overall OF, and U_{head} is the OF representing the global head movement. To measure U_{head} , we divide the face into a few regions and compute the average flow vector in each region. If the angle difference between the flow vector at individual pixels and the corresponding average flow vector is less than a threshold for a majority of the pixels in the region, the average flow vector is considered as U_{head} . Otherwise, U_{head} is set to zero for that region. Note that in all the subsequent processing steps, $U(P; t_i)$ indicates only U_{exp} and not the overall OF.

Figure 4-2 shows an example of head movement correction using the above method. As shown in Figure 4-2a and Figure 4-2b, the expression does not change between the two successive frames, but there is a slight head movement. Figure 4-2c shows the region-wise estimate for U_{head} . For some regions, U_{head} is not shown because majority of the movements in these regions are not coherent ($U_{head}=0$). For better illustration, we zoomed out the OF of the mouth region before and after correction in Figure 4-2 (e-f). It shows that the emotion related OF (U_{exp}) is almost zero in the mouth region.

Figure 4-2: Optical flow correction for head movement; (a)-(b) two consecutive frames; (c) total optical flow (U_{tot}) is illustrated in blue; head movement optical flow (U_{head}) is shown in green; (d) expression related optical flow (U_{exp}) is illustrated in red; (e) U_{tot} of mouth region; (f) U_{exp} of mouth region.

4.2.3 Spatio-Temporal Descriptor

Four pose-invariant features are proposed for encoding the motion information of facial components. The first descriptor is the divergence of the flow field that measures the amount of local expansion or contraction of the facial muscles.

$$Div(P, t_i) = \frac{\partial u(P, t_i)}{\partial x} + \frac{\partial v(P, t_i)}{\partial y} \quad (4-10)$$

where $\frac{\partial u(P, t_i)}{\partial x}$ and $\frac{\partial v(P, t_i)}{\partial y}$ are the partial derivative of the u and v component of the OF respectively.

The second descriptor that captures the local spin around the axis perpendicular to the OF plane is named *Curl*. It is useful to measure the dynamics of the local circular motion of the facial components. This feature is a vector in direction of $\hat{Z}$, where $\hat{Z}$ is a unit vector perpendicular to the image plane.

$$Curl(P, t_i) = \left(\frac{\partial v(P, t_i)}{\partial x} - \frac{\partial u(P, t_i)}{\partial y} \right) \hat{Z} \quad (4-11)$$

The third descriptor as defined in Eq. (4-12) is the scalar projection of the OF vector $\vec{U}$ in the direction of $\hat{P}$, where $\hat{P}$ is the unit position vector ($\hat{P} = \frac{\vec{P}}{|\vec{P}|}$) originating from the nose point in the coordinate system shown in Figure 4-1.

$$Proj(P, t_i) = \vec{U} \cdot \hat{P} \quad (4-12)$$

This *Proj* feature captures the amount of expansion or contraction of each point with respect to the nose point. For example, the “happy” and “sad” expressions can be distinguished by this feature clearly. This phenomenon is shown by Figure 4-3. The figure shows how the sign and length of *Proj* feature changes for a sample lip point (the length of vectors may be exaggerated for better illustration) depending on the facial expression.

The final descriptor called *rotation* is the defined by the cross product of the unit position vector $\hat{P}$ and OF vector $\vec{U}$ as follow:

$$Rot(P, t_i) = \hat{P} \times \vec{U} \quad (4-13)$$

This feature is a vector perpendicular to the plane constructed by $\hat{P}$ and $\vec{U}$ that measures the amount of clockwise or anti-clockwise rotation of each facial point movement with respect to the position vector.

Figure 4-4 illustrates this phenomenon for “happy” and “anger” expression. As shown in this figure, the sign and probably the length of *Rot* feature are different for a sample lip point (the length of vectors may be exaggerated for better illustration) depending on the facial expression.

Figure 4-3: Illustration of Projection descriptor for sad and happy expression.

Figure 4-4: Illustration of Rotation descriptor for happy and anger expressions.

4.2.4 Spatio-Temporal Descriptor Construction

A spatio-temporal descriptor is obtained by accumulating the spatio-temporal features extracted at each pixel. Figure 4-5 illustrates the construction of the spatio-temporal descriptor. The local regions are determined by dividing the volumetric data into predefined number of 3D blocks as shown in Figure 4-5a. The video sequence can be partitioned using overlapping or non-overlapping blocks. To preserve the geometric information of descriptors, each block is further divided into predefined number of 3D cells as illustrated in Figure 4-5b.

Two types of histograms are used to accumulate the features in each cell. A Weighted Histogram (WH) is used to characterize the magnitude of emotion. It consists of two bins - positive and negative bins, and the magnitude of the associated features is used to vote for each bin. The Un-Weighted Histogram (UWH) ignores the magnitude of the emotion and attempts to characterize its dynamics. It involves three bins related to positive, negative, and zero features. Equal vote is assigned for each bin, which means that the total number of positive, negative, and zero features are counted. Since the magnitude of emotion is ignored in UWH, it is able to handle variations in the emotion speed.

WH and UWH are computed for each cell based on the four spatio-temporal features (*Div*, *Curl*, *Proj*, and *Rot*) described earlier. The concatenation of both these histograms is considered as the final descriptor of the corresponding cell. The dimensionality of each cell descriptor is 20 as shown in Figure 4-5c. The concatenation of all the cell descriptors results in the final spatio-temporal descriptor representing the given video sequence.

Figure 4-5: Spatio-temporal descriptor construction; (a) volume data is divided into a number of 3D blocks (B_i). The final descriptor (F) is a concatenation of features from all the blocks; (b) each block is further divided into number of 3D cells (C_j). The feature vector of each block (f_{B_i}) is a concatenation of the cell histograms; (c) weighted and un-weighted histograms are calculated for each cell based on the four spatio-temporal features and concatenated to obtain the cell histogram.

4.3 Experimental Results

This section presents our experimental results for the proposed OF-based pose invariant descriptor and also the comparison, using CK⁺ database, to several state-of-the-art feature extraction methods.

Note that the proposed algorithms were implemented using Matlab 7.11.0 running on a Core i5 CPU (2.8 GHz with 16 GB RAM).

To evaluate our proposed approaches, 5-fold Cross Validation (CV) was used. We used polynomial SVM as a classifier for all methods as the basic classifier for fair comparison. Regarding the parameter selection of SVM, we carried out grid-search on the parameters as suggested in [8]. The parameters producing the best result are chosen.

4.3.1 Parameter Setting

Our proposed method requires selection of a few parameters for feature extraction. For OF extraction, we followed the default setting of parameters used in [111] as it was claimed that the algorithm is fairly insensitive to parameter variations. We also carried out a preliminary experiment on different settings including various numbers of blocks and cells to evaluate its

effect on feature dimension and classification detection rate. The results of the experiment are tabulated in Table 4-1. Based on the results, partitioning the volume data into 100 blocks ($10 \times 10 \times 1$) and 4 cells ($2 \times 2 \times 1$) outperforms the other settings in term of classification detection rate.

Table 4-1: Effect of number of blocks and cells on recognition rate for CK⁺ database. We used SVM as the classifier.

# Blocks	# Cells	# Features	Detection rate (%)
4×4×1	2×2×1	1280	91.85
4×4×1	2×2×2	2560	93.94
4×4×2	2×2×1	2560	92.21
5×5×1	2×2×1	2000	93.99
5×5×1	2×2×2	4000	92.58
5×5×2	2×2×1	4000	94.21
10×10×1	2×2×1	8000	94.48
10×10×1	2×2×2	16000	93.58
10×10×2	2×2×1	16000	93.03

4.3.2 Evaluation of the Proposed Pose-Invariant Descriptor

To evaluate the discriminant property of the proposed feature, we measured the Earth Mover Distance (EMD) of the features between any two expressions for lip segment of a subject. EMD is a similarity metric between two distributions that has been shown to be an effective measure for comparing dissimilarity between image features [122]. We computed the total histogram features (WH+UWH) for the lip segment of size $10 \times 10 \times 10$ for all emotions expressed by a subject and then measured the EMD of the calculated features between every two emotions. The experimental results are tabulated in Table 4-2. To check whether these values are good enough for the proposed descriptor to discriminate the emotions, EMD is applied for nose segment (the same size as the lip segment) of the same subject where no expression related movement is expected. Table 4-3 shows the results of EMD measurement for the nose segment. Comparison of these two tables confirms the discriminative ability of the proposed feature set. As shown, the EMD of the proposed feature between every two different emotions for lip segment is bigger than those in nose segment.

The details of the results for our method are presented in Table 4-4 as a confusion matrix. From this, it is seen that the detection rate of “sadness” and “fear” are less than other emotions. It is due to the disability of some subjects showing the mentioned expressions in this

database, not the deficiency of the pose-invariant feature, since we obtained similar results for our preliminary proposed descriptors [97]. For example, the way some subjects express “sadness” is very similar to “anger”.

We have evaluated the performance of each feature individually for CK^+ . Table 4-5 shows the results. As shown, the features are complementary and can improve the classification performance when combined together. In other words, an ensemble of all these features gives better classification performance compared to any subset of four features. However, the performance difference is not significant. This is because CK^+ is less challenging, with minimal pose variations.

We have also done an experiment on CK^+ to evaluate the performance of each histogram individually. Table 4-6 shows the results. Indeed, we found that WH and UWH have complementary information that can improve the performance. Indeed, WH is useful to characterize the magnitude of emotion and UWH attempts to characterize its dynamics to be able to handle variations in the emotion speed. The term “emotion speed” is inversely proportional to time lapse (or number of frames) between the start of the expression and the peak expression. While emotion speed does not provide any discriminatory information, it affects the optical flow in the following way. When the emotion speed is slow, the magnitude of the flow vectors tends to be small and vice versa. This further affects the magnitude of the *Div*, *Curl*, *Proj*, and *Rot* features (as defined in section 4.2.3) derived from the optical flow. The UWH minimizes the effect of changes in the emotion speed by considering only the sign of the features (positive, negative, or zero) and ignoring their magnitude. However, the magnitude of the *Div*, *Curl*, *Proj*, and *Rot* features cannot be completely ignored because it does contain some useful information to differentiate between a subtle emotion and an exaggerated emotion. Therefore, the WH has been used to capture the magnitude information.

Table 4-2: Earth Mover Distance of features for lip segment.

Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	4.78×10^{-13}	6.76	4.52	6.81	3.68	4.44
Disgust	6.76	3.78×10^{-13}	4.28	2.46	4.88	5.65
Fear	4.52	4.28	4.60×10^{-7}	4.75	4.63	1.84
Happiness	6.81	2.46	4.75	3.56×10^{-10}	4.61	6.12
Sadness	3.68	4.88	4.63	4.61	5.26×10^{-9}	4.84
Surprise	4.44	5.62	1.84	6.12	4.84	3.23×10^{-9}

Table 4-3: Earth Mover Distance of features for nose segment.

Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	4.58×10^{-8}	0.93	1.21	0.79	1.07	0.73
Disgust	0.93	1.40×10^{-11}	0.97	0.87	0.80	0.85
Fear	1.21	0.97	2.33×10^{-7}	1.34	0.68	1.50
Happiness	0.79	0.87	1.34	7.50×10^{-7}	1.03	0.61
Sadness	1.07	0.80	0.68	1.03	4.73×10^{-13}	1.19
Surprise	0.73	0.85	1.50	0.61	1.19	5.62×10^{-12}

Table 4-4: Confusion matrix of the result.

Actual \ Predict	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	95.45	1.69	0	0	10.71	0
Disgust	2.27	98.31	0	0	0	0
Fear	2.27	0	87.50	0	0	0
Happiness	0	0	4.17	100	0	1.20
Sadness	2.27	0	8.33	0	89.29	0
Surprise	0	0	4.17	0	0	98.80

Table 4-5: The classification performance on each feature individually. We used SVM classifier on CK+ database.

Feature	Detection Rate (%)
Div	92.42
Curl	91.05
Proj	84.75
Rot	88.57
Div+Curl+Proj	94.83
Div+Curl+Rot	93.38
Div+Proj+Rot	92.37
Curl+Proj+Rot	92.95
All	95.33

Table 4-6: Assessment the kind of histogram in classification performance. The results are based on SVM classifier on CK+ database.

Method	Detection Rate (%)
WH	90.66
UWH	91.91
WH+UWH	95.14

4.3.3 Comparison to Other Approaches

The next experiment was conducted to quantify the efficiency of the proposed dynamic descriptor. A lot of researches have used the CK+ database as the benchmark database and reported their experimental results on it. However the results cannot be directly compared due to different experimental setups (such as pre-processing approaches, number of sequences used and evaluation methods). Therefore, we have re-run the better approaches ourselves using the same experimental setup.

Our method is compared to two successful dynamic approaches in this field; LBP-TOP, LPQ-TOP, and also our proposed features described in Chapter 3 (STHOG, LPLO, and HDPC) for CK⁺ database [6, 123]. The source codes of LBP-TOP and LPQ-TOP are publicly available*. We varied the parameter settings to get the best result which is reported.

The execution time was also measured for the extraction of descriptors from a volume data of size 100×100×10.

The results are reported in Table 4-7. As shown, the detection rate of our proposed descriptor is significantly better than the other descriptors, but at the cost of increased execution time. The dimensionality of the proposed descriptor is also less compared to the other methods.

It is noted that the results shown in this table for LBP-TOP, LPQ-TOP, and our preliminary proposed descriptors are different from one reported in the original papers due to different experimental setup such as pre-processing, evaluation measurement, segmentation processing, number of sequences used, and etc.

Table 4-7: Comparison of the proposed descriptor to other methods for the original CK⁺ database. We used SVM as a classifier for all methods.

Method	# Feature	Detection rate	Time complexity (Sec)
LBP-TOP [6]	17700	89.31	5.22
LPQ-TOP [123]	76800	89.17	2.19
STHOG [91]	24300	90.76	0.85
LPLO [97]	20736	91.87	0.5
HDPC [109]	10800	92.88	4.20
OF-based Pose Invariant Descriptor	8000	94.48	13.12

* http://www.ee.oulu.fi/~gyzhao/LBP_Book.htm

4.3.4 Robustness of the Descriptor to Pose Variations

To check the robustness of the descriptor to facial pose variations, we have compared the features extracted from a lip segment for both frontal and non-frontal faces for happy and surprise emotions as shown in Figure 4-6 and Figure 4-7 respectively. As illustrated in each figure, the histograms of features look similar. For instance, if we compare the WH of Figure 4-6, we can see for *Curl* feature, positive bin is bigger than negative one for both frontal face and non-frontal one. This phenomenon is also happening for other features. But if we compare the *Curl* feature of happy emotion in Figure 4-6 to the surprise emotion in Figure 4-7, the result is totally different (negative bin is bigger than positive bin for surprise). Indeed, the extracted features are view invariant, but sensitive to various emotions.

We also evaluated the pose-invariant property of the descriptor by training the SVM classifier on the original CK⁺ and then test it for own collected samples (42 samples with pose variations). As a comparison, the same experiment has been performed on LBPTOP descriptor. Table 4-8 tabulates the results of this experiment. The accuracy of our proposed descriptor is quite superior to LBPTOP in term of detection rate.

Figure 4-6: Pose-invariant descriptor for surprise emotion; (a) features extracted from the lip segment from frontal face; (b) features extracted from the lip segment from non-frontal face.

Figure 4-7: Pose-invariant descriptor for surprise emotion; (a) features extracted from the lip segment from frontal face; (b) features extracted from the lip segment from non-frontal face.

Table 4-8: Comparison the robustness of the methods to pose variations. We used SVM as a classifier for all methods.

Method	Detection rate
LBP-TOP [6]	23.81
OF-based Pose Invariant Descriptor	80.95

4.4 Concluding Remarks

In this chapter, we proposed our novel feature extraction method based on OF concept which can deal with the uncontrolled situations such as pose variation. Our proposed dynamic descriptor is different from the existence OF-based representations in three aspects. Firstly, we propose a new set of spatio-temporal descriptors to capture the dynamic information hidden in a flow field. Secondly, the extracted features are encoded effectively for pose-invariant purpose through relative measure from the reference points derived from the face image.

Finally, only the statistics of the extracted features is retained as discriminative information for further processing. In other words, we introduced two kinds of histogram for condensing the extracted features (weighted and un-weighted histogram) to be able to detect both the level and dynamic of expressions.

The proposed feature set is able to handle the emotion speed variations and also the head pose variations even it slightly happens during expressing an emotion. Using a common SVM classifier, we showed that the performance of our proposed approach gives the best performance compared to the other feature extraction methods on the standard CK⁺ database with 94.48% accuracy in detecting six basic emotions. We also evaluated the robustness of the descriptor to facial pose variations using own collected database.

Chapter 5

Extreme Sparse Learning for Robust Emotion Classification

5.1 Introduction

In this Chapter, we propose a sparse representation based classification method called ESL to accurately recognize facial expressions in real-world natural situations. The proposed approach combines the discriminative power of ELM with the reconstruction capability of sparse representation. This enables the ESL approach to deal with noisy signals and imperfect data recorded in natural settings.

While the sparse representation approach has the ability to enhance noisy data using a dictionary learned from clean data, it is not sufficient because the end goal is to correctly recognize the facial expression. In a sparse-representation-based classification task, the desired dictionary should have both representational ability and discriminative power. Since separating the classifier training from dictionary learning may cause the learned dictionary to be sub-optimal for the classification task, we propose to jointly learn a dictionary and classification model. In other words, in contrast with most existing schemes that attempt to update the dictionary and classifier parameters alternately by iteratively solving sub-problems [73], we propose to solve them simultaneously. This joint dictionary learning and classifier training can be expected to result in a dictionary that is both reconstructive and discriminative for a robust recognition system.

Before introducing the ESL, we briefly present the ELM as well as the concepts underlying sparse representation and Dictionary Learning (DL) in the following sub-sections. We also

review several classification techniques based on sparse representation which we used for comparison to our proposed ESL.

5.2 Extreme Learning Machine (ELM)

Since ELM has a successful performance especially for multi-class classification problems, it is a good choice for our problem to be jointed with sparse representation and dictionary learning for further improvement. Additionally, ELM requires fewer optimization constrains in compare to SVM which results in simple implementation, fast learning, and better generalization performance [124].

ELM was initially proposed for the single-hidden layer feed-forward Neural Networks (NNs) and then developed to the generalized single-hidden layer feed-forward networks (SLFNs) which not need to be tuned in hidden layer [124]. Indeed, unlike the conventional NN learning algorithm, the parameters of ELM hidden nodes are randomly selected. ELM learning algorithm is trying to minimize the training error and the norm of output weights as well. The objective function of the ELM is summarized as follow:

$$\text{objective: } \min_{\beta} \{ \|H(X)\beta - T\|_2^2 + \|\beta\|_2^2 \} \quad (5-1)$$

where X denotes the training samples, β is the output weight, and T is the target vector. H is the hidden layer output (with L nodes) matrix:

$$H(X) = \begin{bmatrix} h(X_1) \\ \vdots \\ h(X_N) \end{bmatrix} = \begin{bmatrix} h_1(X_1) & \dots & h_L(X_1) \\ \vdots & \vdots & \vdots \\ h_1(X_N) & \dots & h_L(X_N) \end{bmatrix} \quad (5-2)$$

The minimal norm least square method can find the solution as follows:

$$\begin{aligned} \text{when } N < L : \beta &= H^\dagger T = H^T \left(\frac{I}{C} + HH^T \right)^{-1} T \\ \text{when } N > L : \beta &= H^\dagger T = \left(\frac{I}{C} + HH^T \right)^{-1} H^T T \end{aligned} \quad (5-3)$$

where N denotes the total number of samples, L is the number of nodes in the hidden layer, $H^\dagger$ is the Moore–Penrose generalized inverse of matrix H , and C is the user-specified parameter added to the formulation for better generalization performance [124, 125]. For ELM parameter setting, the same strategy as mentioned in previous section can be applied.

The output function of ELM is defined as follow:

$$f(x) = h(x)\beta = h(x)H^T(\frac{I}{C} + HH^T)^{-1}T \quad (5-4)$$

Finally, the sign of ELM output function ($sign(f(x))$) will indicate the classifier decision for binary problem and the highest output value of $f(x)$ will give the predicted label for multiclass cases:

$$\begin{aligned} label(x) &= sign(f(x)) && \text{for binary classification} \\ label(x) &= argmax(f_i(x)) && \text{for multi - class problem} \end{aligned} \quad (5-5)$$

It was shown that a wide type of feature mappings including random hidden nodes and kernels can be utilized for ELM [124].

5.2.1 Random Feature Mappings

Various nonlinear feature mappings in form of $G(a, b, X)$ can be utilized in ELM such as:

$$\text{Gaussian function:} \quad G(a, b, X) = \exp(-b\|X - a\|^2) \quad (5-6)$$

$$\text{Sigmoid function:} \quad G(a, b, X) = (1 + \exp(-(aX + b)))^{-1} \quad (5-7)$$

$$\text{Multi-quadric function:} \quad G(a, b, X) = (\|X - a\|^2 + b^2)^{0.5} \quad (5-8)$$

Then the hidden layer output vector is written as:

$$h(X) = [G(a_1, b_1, X), \dots, G(a_L, b_L, X)] \quad (5-9)$$

where $\{(a_i, b_i)\}_{i=1}^L$ are randomly generated based on any probability distribution.

5.2.2 Kernels

For unknown feature mappings $h(X)$, kernels can be applied interchangeably. The kernel matrix for ELM is defined as:

$$\Omega_{ELM} = HH' \quad (5-10)$$

$$\Omega_{ELMi,j} = h(X_i)h(X_j) = K(X_i, X_j) \quad (5-11)$$

Then, the output weight β and output function can be re-written as:

$$\beta = H' \left(\frac{I}{C} + \Omega_{ELM} \right)^{-1} T \quad (5-12)$$

$$f(X) = h(X)\beta = h(X)H^T \left(\frac{I}{C} + \Omega_{ELM} \right)^{-1} T = \begin{bmatrix} K(X, X_1) \\ \vdots \\ K(X, X_N) \end{bmatrix} \left(\frac{I}{C} + \Omega_{ELM} \right)^{-1} T \quad (5-13)$$

Some commonly used kernel functions are described as follows:

Polynomial (homogeneous): $k(x_i, x_j) = (x_i \cdot x_j)^d \quad (5-14)$

Polynomial (inhomogeneous): $k(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (5-15)$

Gaussian radial basis function: $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \text{ for } \gamma > 0 \quad (5-16)$

Hyperbolic tangent: $k(x_i, x_j) = \tan h(\beta_0 x_i \cdot x_j + \beta_1), \text{ for some } \beta_0 > 0 \text{ and } \beta_1 < 0 \quad (5-17)$

5.3 Sparse Representation and Dictionary Learning

In recent years, the concepts of sparse representation and dictionary learning have attracted much attention in signal/image processing and computer vision fields. It is a powerful tool for reconstruction, representation, and compression of high-dimensional data. The ability of sparse representation to uncover the important information of data from the base elements or dictionary atoms makes it a successful technique to represent and compress the high-dimensional noisy signals. Since the images and videos (and their features) are naturally very high dimensional, sparse coding is a useful technique to encode the essential information in a compact representation.

The basic model is based on the fact that the natural signals or images can be efficiently approximated by linear combination of a few elements (so called atoms) of a dictionary. Given Y as the set of training samples, and D as the dictionary, the sparsity model is described by the following:

$$\min_x \{ \|Y - DX\|_2^2 \} \quad s.t. \ \|x\|_0 \leq L \quad (5-18)$$

where x represents the sparse code of the input signal y , X denotes the coefficient matrix, L is sparsity threshold constrain, and $\|\cdot\|_0$ is the l^0 pseudo norm that counts the total non-zero elements. The sparse approximation problem, which is an NP-hard problem has been already

solved using several techniques like Orthogonal Matching Pursuit (OMP) [126], Basis Pursuit (BP) [127], Iterative Hard Thresholding (IHT) [128], etc.

One important issue in the mentioned sparse model is the choice of the dictionary. It can be either based on a predefined transforms on data like Fourier transforms, or can be directly learned from training data. It has been observed that the latter case leads to better representation. The K-Singular Value Decomposition (K-SVD) algorithm is an efficient technique for training a dictionary from training data [129]. The algorithm aims to iteratively improve the dictionary to achieve a satisfactory sparse representation of the signals, by solving the following optimization formula:

$$\min_{X,D} \{\|Y - DX\|_2^2\} \quad s.t. \quad \|x_i\|_0 \leq L \quad (5-19)$$

The KSVD algorithm consists of two steps: first, dictionary is considered to be known, and sparse codes of the signals are estimated using OMP. Second, given the current sparse coefficients, the dictionary atoms are updated. OMP and KSVD algorithms are described in detail in Algorithm1 and Algorithm2, respectively.

Algorithm1: OMP [130]

---

1. Input: signal set Y , dictionary D , number of iterations k , target sparsity L or target error ε .
  2. Initialize the residual matrix, $R_0 = Y$, the index set $\Lambda_0 = \emptyset$, iteration $t=1$
  3. Repeat until stopping criterion is met.
    - 3.1. Find the index λ_t of the dictionary atom d that best matches to the residual: $\lambda_t = \operatorname{argmax} |d_{\lambda_t}^T R_{t-1}|$, where $d_{\lambda_t}^T$ is the vector transpose of the atom d with the index λ_t .
    - 3.2. Update the index set: $\Lambda_t = \Lambda_{t-1} \cup \lambda_t$
    - 3.3. Determine the orthogonal projection onto the span of atoms indexed in Λ_t as: $X_{\Lambda_t} = (D_{\Lambda_t})^\dagger Y$, where $\dagger$ represents the Pseudo-inverse.
    - 3.4. Update the residual $R_t = Y - D_{\Lambda_t} X_{\Lambda_t}$
    - 3.5. $t = t + 1$.
  5. Output: Sparse matrix X_{Λ_k}
-

Algorithm2: KSVD [129]

-
1. Input: signal set Y , target sparsity L
  2. Initialization: dictionary D_0 with K number of atoms, iteration $t=1$
  3. Repeat until convergence (stopping rule)
    - 3.1. Sparse coding stage: apply any pursuit algorithm like OMP to compute the sparse coefficients of the input signals by approximating the solution of $\min_X \{\|Y - D_{t-1}X\|_2^2\}$ s.t. $\|x\|_0 \leq L$
    - 3.2. Codebook update stage: for each dictionary atom ($k=1,2,\dots,K$) update it by:
      - 3.2.1. Find the group of samples that use the current atom, $w_k = \{i | 1 \leq i \leq N, x^k(i) \neq 0\}$
      - 3.2.2. Compute the representation error matrix: $E_k = Y - \sum_{j \neq k} d_j x^j$
      - 3.2.3. Restrict E_k by choosing only the columns corresponding to w_k , and name it E_k^R .
      - 3.2.4. Apply SVD decomposition $E_k^R = U\Delta V^T$. Choose the updated dictionary atom to be the first column of U , and update the coefficient vector to be the first column of V multiplied by $\Delta(1,1)$.
      - 3.2.5. $t = t + 1$.
  4. Output: dictionary D , Sparse matrix X
-

5.4 Classification Models based on Sparse Representation

Sparse representation algorithms try to find a few number of representative patterns from a large set of given signals. Recent researches have shown that sparse coefficients of the patterns can be more discriminative to differentiate various pattern categories. Thus, it may provide higher performance for classification problems. DL concept is an important issue for sparse representation based classification that aims to design an effective dictionary to adapt the input data.

Some of previous work focused on learning a reconstructive dictionary for pattern recognition. However, it is observed that the conventional DL framework like KSVD is more suited for signal reconstruction and not adapted for classification purpose. To address this problem, recently, several techniques have been developed to learn a classification-oriented dictionary. In the following subsections, we review several DL based classification methods, which are used in our result section for performance comparison.

5.4.1 Classification based on Reconstructive Dictionary

In this section, we first introduce the sparse representation based classification which even though does not learn the dictionary; inspires later techniques in this field.

A. Sparse Representation based Classifier (SRC)

Wright et al. [131] introduced SRC for robust face detection and achieved successful results even in the presence of noise such as illumination changes and occlusion. They constructed the dictionary based on the training samples as follows:

$$D = [Y_1, \dots, Y_C] \quad (5-20)$$

where Y_i is a subset of training samples belongs to class i , with totally C classes of individual faces. Then, for the query facial image y , SRC should find the sparse coefficient via L^1 norm minimization as follow:

$$x = \operatorname{argmin}_x \{\|y - Dx\|_2^2 + \lambda \|x\|_1\} \quad (5-21)$$

where λ is a constant parameter which controls the sparsity level.

The decision function of SRC is based on the image reconstruction error as defined in follow:

$$\text{label} = \operatorname{argmin}_i \{\|y - Y_i \delta_i(x)\|_2^2\} \quad (5-22)$$

where $\delta_i(x)$ is a vector indicator function that find the corresponding elements to class i .

B. Sparse ELM (SELM)

Inspired by the method proposed by Guha and Ward [132] that applied SVM classifier on sparse coefficients, it is possible to compute the sparse representation of the signals by learning a reconstructive dictionary and then apply the ELM classifier. Thus the SELM can be formulated as follows:

$$\langle X, D \rangle = \operatorname{argmin}_{X,D} \{\|Y - DX\|_2^2 + \lambda \|X\|_1\} \quad (5-23)$$

where Y is the input pattern, X is the coefficient matrix, and λ denotes the regularization term to control the sparsity.

Then ELM is trained over sparse coefficients X using:

$$ELM \text{ training: } \min_{\beta} \{ \|H(X)\beta - T\|_2^2 + \|\beta\|_2^2 \} \quad (5-24)$$

For the recognition of the query signal, the sparse coefficient is obtained using the learned dictionary first, and then predicts the class label using ELM as described in section 5.2. The classification training is formulated separately from DL in the sparse representation of SELM, thus revealing the disadvantageous of this approach.

5.4.2 Classification based on Discriminative Dictionary

While DL directly from the training data usually leads to a satisfactory reconstruction, adding a specific discriminative criterion to dictionary training can improve the discriminative ability of the method and lead to better classification results. Recently, several methods have been developed to train a classification oriented dictionary.

A. Discriminative KSVD (DKSVD)

DKSVD was proposed by Zhang and Li [69] to jointly train a desired dictionary which has both representation and discrimination power, and an optimal linear classifier. They formulated DKSVD as following:

$$\langle X, D, W \rangle = \operatorname{argmin}_{X, D, W} \left\{ \|Y - DX\|_F^2 + \lambda_1 \|H - WX\|_F^2 + \lambda_2 \|W\|_F^2 \right\} \quad s. t. \|x\|_0 \leq L \quad (5-25)$$

$H = [h_1, \dots, h_N] \in R^{C \times N}$ denotes the label information of the training data such that $h_i = [0, \dots, 1, \dots, 0]$, where the position of nonzero elements indicates the class. W is the classifier parameter, λ_1 and λ_2 are the constants that control the tradeoff between corresponding terms $\| \cdot \|_F$ represents the Frobenius norm of the entry.

Ignoring the regularization penalty term $\|W\|_F^2$, the final formulation was simplified to:

$$\langle X, D, W \rangle = \operatorname{argmin}_{X, D, W} \left\| \left(\frac{Y}{\sqrt{\lambda_1}} \right) - \left(\frac{D}{\sqrt{\lambda_1}} \right) X \right\|_F^2 \quad s. t. \|x\|_0 \leq L \quad (5-26)$$

Now, the problem can be efficiently solved using original KSVD.

B. Label Consistent KSVD (LCKSVD)

Jiang et al. [71] proposed to incorporate a discriminative sparse coding criterion and a linear classification error criterion into a unified objective function.

The first proposed model named LCKSVD1 focuses on adding a label consistent regularization term to the basic objective function of DL as follow:

$$\langle X, D, A \rangle = \underset{X, D, W}{\operatorname{argmin}} \{ \|Y - DX\|_F^2 + \alpha \|Q - AX\|_F^2 \} \quad s. t. \|x\|_0 \leq L \quad (5-27)$$

where α is a constant that controls the contribution of the corresponding terms. $Q = [q_1, \dots, q_N] \in R^{K \times N}$ denotes the discriminative sparse matrix, in which $[q_i^1 \dots q_i^K]^T = [0 \dots 1, 1, \dots 0]^T$, where the position of non-zero elements are those that the input signal y_i and dictionary atom d_k share the same class label. A is a linear transformation matrix to transform the original sparse codes to be more discriminative.

The second proposed model named LCKSVD2 adds the label consistent regularization term and a joint linear classification criterion to the basic objective function of DL as follow:

$$\langle X, D, A, W \rangle = \underset{X, D, W}{\operatorname{argmin}} \left\{ \|Y - DX\|_F^2 + \alpha \|Q - AX\|_F^2 + \beta \|H - WX\|_F^2 \right\} \quad s. t. \|x\|_0 \leq L \quad (5-28)$$

where the last term represents the classification error same as Eq. (4-25).

Following the procedure used in DKSVD, the optimization problem of both above mentioned models can be efficiently solved by KSVD.

C. Fisher Discriminative Dictionary Learning (FDDL)

Yang et al. [72] proposed a novel DL method named FDDL based on the Fisher discriminative criterion imposed on the sparse coefficients to have small within-class variance and large between-class variance. FDDL tries to learn a structured dictionary $D = [D_1, \dots, D_c]$, in which D_i is the class-specific sub-dictionary corresponding to class label i , and c is the total number of classes. The training samples were denoted by $Y = [Y_1, \dots, Y_c]$, where Y_i is the sub-set of the training samples belong to class i .

The FDDL model was proposed as follow:

$$\langle D, X \rangle = \underset{D, X}{\operatorname{argmin}} (r(Y, D, X) + \lambda_1 \|X\|_1 + \lambda_2 f(X)) \quad (5-29)$$

where $r(Y, D, X)$ is the discriminative fidelity term, $f(X)$ is the discriminative term posed on sparse codes, and λ_1 and λ_2 controls the contribution of the corresponding terms.

Denote X_i as the coefficient matrix of Y_i over D , as $X_i = [X_i^1, \dots, X_i^j, \dots, X_i^c]$, where X_i^j is the sparse representation of Y_i over D_j . The discriminative fidelity term is defined as:

$$r(Y_i, D, X_i) = \|Y_i - DX_i\|_F^2 + \|Y_i - D_i X_i^i\|_F^2 + \sum_{j \neq i} \|D_j X_i^j\|_F^2 \quad (5-30)$$

whereby the first term implies that the dictionary D should be able to efficiently represent Y_i , the second and third term imply that Y_i should be efficiently represented by D_i but not by D_j , for $i \neq j$.

For making the dictionary D discriminative, the Fisher discriminative criterion was applied on coefficients as:

$$f(X) = \text{tr}(S_W(X)) - \text{tr}(S_B(X)) \quad (5-31)$$

where $S_W(X)$ and $S_B(X)$ are the within-class and between-class scatter of coefficient respectively and defined as:

$$S_W(X) = \sum_{i=1}^c \sum_{x_k \in X_i} (x_k - m_i)(x_k - m_i)^T \quad (5-32)$$

$$S_B(X) = \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T \quad (5-33)$$

where m_i and m represent the mean vector of X_i and X respectively, and n_i denotes the total number of samples belong to class i .

To solve the FDDL optimization problem, Iterative Soft Thresholding (IST) technique was utilized. Finally, for classification, they used the reconstruction error based method.

5.5 Proposed Methodology: Extreme Sparse Learning

Separating the classification training from dictionary learning may lead to a scenario where the learned dictionary is not optimal for classification. We propose to jointly learn a dictionary and classification model for better performance.

Learning a discriminative dictionary for sparse representation of Y can be accomplished by solving the following problem:

$$\min_{X,D,\beta} \{\|Y - DX\|_2^2 + \gamma_1 (\|H(X)\beta - T\|_2^2 + \|\beta\|_2^2) + \gamma_2 \|X\|_1\} \quad (5-34)$$

where D is the learned dictionary, X denotes the sparse codes of the input signals, and $\|\cdot\|_1$ is the l^1 norm that simply sums up the absolute value of the elements. The first term $\|Y - DX\|_2^2$ denotes the reconstruction error, second term $(\|H(X)\beta - T\|_2^2 + \|\beta\|_2^2)$ is related to ELM optimization constraints, and third term is related to the sparsity criterion. The framework in Eq. (5-34) is referred to as ESL. When kernels are incorporated in the above framework, we refer to it as Kernel ESL (KESL). In other words, the framework formulated as follow is referred as KESL:

$$\min_{X,D,\beta} \{\|Y - DX\|_2^2 + \gamma_1 (\|K(X)\beta - T\|_2^2 + \|\beta\|_2^2) + \gamma_2 \|X\|_1\} \quad (5-35)$$

where $H(X)$ in Eq. (5-34) is replaced by $K(X)$.

Figure 5-1 shows the high-level outline of the proposed recognition framework. Training and testing algorithms are also described in details as Algorithm3 and Algorithm4 respectively.

As shown in Figure 5-1, there are three main steps involved in ESL training: supervised sparse coding, ELM output weight update, and dictionary optimization. The first step estimates the sparse coefficients X of the input signals Y based on initial D_0 and β_0 . Based on estimated sparse codes, the ELM output weight β is updated in the second step. Then this process is repeated until the stop criterion is met. The stop criterion could be defined based on the total error $(\|Y - DX\|_2^2 + \gamma_1 (\|H(X)\beta - T\|_2^2 + \|\beta\|_2^2) + \gamma_2 \|X\|_1)$ or certain limit of iterations. Finally, the dictionary atoms are updated based on the estimated sparse matrix. Then all the steps are iteratively repeated to achieve the final D and β . For the testing phase, the sparse coefficients of the testing samples are estimated using known D , and then classified by ELM with known β . These steps are explained in detail in the following subsections. We propose a novel approach for supervised sparse coding, which is a critical component of the proposed recognition framework.

Figure 5-1: Proposed method for recognition framework.

Algorithm3: Training Algorithm for ESL

1. Input: Signal set Y , class labels (T) of Y , regularization terms γ_1 and γ_2 , stopping criterion for outer and inner loops ($\epsilon_1, \eta_1, \epsilon_2, \eta_2$)
2. Initialize the dictionary $D_{(0)}$ and the ELM output weight vector $\beta_{(0)}$
3. Repeat until the stopping criterion defined by (ϵ_2, η_2) is met
  - 3.1. Repeat until the stopping criterion defined by (ϵ_1, η_1) is met
    - 3.1.1. Supervised sparse coding: find the sparse matrix $X_{(i)}$ by approximating the solution of $\min_X \{\|Y - DX\|_2^2 + \gamma_1 (\|H(X)\beta - T\|_2^2) + \gamma_2 \|X\|_1\}$
    - 3.1.2. ELM output weight optimization: find the ELM output weight $\beta_{(i)}$ by approximating the solution of $\min_{\beta} \{\|H(X)\beta - T\|_2^2 + \|\beta\|_2^2\}$
  - 3.2. Dictionary update: find D by approximating the solution of $\min_D \{\|Y - DX\|_2^2\}$
4. Output: D, β

Algorithm4: Classification Algorithm for ESL

-
1. Input: test signal y , learned dictionary D and the ELM output weight β , regularization term γ
  2. Find the sparse code of test data by approximating the solution of $\min_x \{\|y - Dx\|_2^2 + \gamma\|x\|_1\}$
  3. Find the output function of ELM: $f(x) = h(x) \beta$
  4. Estimate the class label of the test data: $label(x) = \operatorname{argmax}(f_i(x))$
  5. Output: $label(x)$
-

5.5.1 Supervised Sparse Coding

The main objective of this step is to solve the sparse approximation problem while simultaneously minimizing the classification error as shown in the following equation:

$$\text{objective: } \min_x \{\|Y - DX\|_2^2 + \gamma_1(\|H(X)\beta - T\|_2^2 + \|\beta\|_2^2) + \gamma_2\|X\|_1\} \quad (5-36)$$

where Y is the extracted features, D is the learned dictionary, β is the output weight of ELM classifier, and T denotes the class label of training data. γ_1 and γ_2 controls the contribution of reconstruction criterion, classification error, and sparsity. The objective function can be formulated either by $\|X\|_1$ or $\|X\|_0$. We chose $\|X\|_1$ as we obtained better performance in our preliminary experiments.

Although different methods have been suggested in the literature for sparse coding [128, 133-136], none of them can be applied to Eq. (5-36) directly due to the presence of the non-linear term $H(X)$. Moreover, we cannot simplify the above optimization problem to the basic sparse representation form as applied in [69-71]. Inspired by existing algorithms for sparse coding, we proposed two following methods to solve Eq. (5-36).

A. Method I: “Iterative Hard Thresholding”

Iterative Hard Thresholding (IHT) is a projected gradient descent algorithm which is very simple but effective algorithm for sparse representation problems and compressed sensing [128, 137, 138]. We developed the algorithm for solving Eq. (5-36) as follows:

$$E = \|Y - DX\|_2^2 + \gamma(\|H(X)\beta - T\|_2^2) \quad (5-37)$$

$$X = P_k(X - \tau \nabla E) \quad (5-38)$$

$$\nabla E = 2D'(DX - Y) + 2\gamma \frac{dH(X)}{dX} \beta(H(X)\beta - T) \quad (5-39)$$

where τ is the step size, and $P_k(x)$ is the non-linear operator that sets all but the largest (in magnitude) k elements of x to zero. D' denotes the transpose of matrix D .

The procedure of finding $\frac{dH(X)}{dX}$ for KELM ($H(x)$ is unknown) is explained in appendix A.

The standard IHT implementation faces several challenges when applied to practical problems. The step-size and sparsity parameters have to be chosen appropriately.

It was shown that under conditions similar to those required in the linear setting, the IHT algorithm can be applied to nonlinear cases [139]. The Restricted Isometry Property (RIP) condition is a sufficient condition for the recovery of sparse signals from linear or non-linear observations. Since Eq. (5-36) consists of two terms, the RIP condition should be satisfied for both terms to guarantee the convergence of the algorithm to a local minimum of our original cost function defined by Eq. (5-36). The convergence proof of the first term of the function related to the reconstruction error is already investigated in the literature [140]. All we require here is to show that the second term of Eq. (5-36) will satisfy the RIP. It is investigated for sigmoid and kernel ELM in Appendix B and Appendix C respectively. Then, it is guaranteed that the IHT algorithm can find a sparse vector that is close to the true minimizer of the cost function including both linear and non-linear observations.

B. Method II: Class Specific Matching Pursuit (CSMP)

In this section, we have developed an algorithm for supervised sparse coding inspired by the simultaneous sparse approximation algorithm described by Huang and Aviyente [73]. The method is similar to Simultaneous Orthogonal Matching Pursuit (S-OMP), which seeks to find a set of dictionary atoms that best represent all the signals [133, 134]. In contrast to S-OMP, the proposed method attempts to find a fixed set of atoms that can optimize the objective function in Eq. (5-36) for all signals that belong to the same class. Therefore, the proposed method is referred to as Class Specific Matching Pursuit (CSMP).

Denote the training samples by $Y = [Y_1, \dots, Y_{cl}]$, where Y_i , is the sub-set of the training samples belong to class i , and c is the total number of classes. Similarly, the sparse coefficient matrix X can be written as $X = [X_1, \dots, X_{cl}]$, where X_i is the coefficient matrix of Y_i . Given dictionary D , CSMP applies simultaneous matching pursuit for each Y_i as shown in Algorithm5.

Algorithm5: Class Specific Matching Pursuit (CSMP)

-
1. Input: Signal set Y , class labels (T) of Y , total number of classes ω , number of dictionary atoms K , regularization terms γ_1 and γ_2 , stopping threshold ε_3 , dictionary D from previous iteration, sparse code matrix X_{prev} from previous iteration to train the classifier, ELM output weight β from previous iteration
  2. for $i = 1, 2, \dots, \omega$ do
    - 2.1. Find all signals that belong to class i and form the matrix Y_i
    - 2.2. Initialize the temporary sparse matrix $X^{temp} = 0$ (same size of X_i), selected index set $\Lambda_0 = \emptyset$, unselected index set $\Omega_1 = \{1, 2, \dots, K\}$, $j=1$
    - 2.3. While $j \leq K$ do
      - 2.3.1. Let m be the number of elements in the set Ω_j
      - 2.3.2. Let $\Delta \in R^{N \times m}$ be the sub-matrix of D containing only the atoms (columns) in the dictionary indexed by Ω_j
      - 2.3.3. Compute coefficient matrix $\Gamma = \Delta^\dagger Y_i$, where $\dagger$ is the pseudo-inverse operator, $\Gamma \in R^{m \times S_i}$
      - 2.3.4. For $k = 1, 2, \dots, m$ do
        - 2.3.4.1. Let κ be the k^{th} element of Ω_j : $\kappa = \Omega_j(k)$
        - 2.3.4.2. Replace the κ^{th} row of X^{temp} with k^{th} row of coefficient matrix Γ : $X_{\kappa,}^{temp} = \Gamma_{k,}$
        - 2.3.4.3. Calculate the residual: $RE = \|Y_i - \Delta_{:,k} \Gamma_{k,}\|_2^2$ (where $\Delta_{:,k}$ is the k^{th} column of Δ)
        - 2.3.4.4. Find the classification error of signal X^{temp} using ELM and denote it as CE
        - 2.3.4.5. Calculate the sparsity: $SE = \|\Gamma_{k,}\|_1$
        - 2.3.4.6. Find the total error: $E_k = RE + \gamma_1 CE + \gamma_2 SE$
      - 2.3.5. end for
      - 2.3.6. Find the index λ_j of the atom dictionary that minimizes the total error: $\lambda_j = \Omega_j(\text{argmin}_k(E_k))$
      - 2.3.7. Update the selected index set: $\Lambda_j = \Lambda_{j-1} \cup \lambda_j$
      - 2.3.8. Update the unselected index set: $\Omega_{j+1} = \Omega_j \setminus \lambda_j$
      - 2.3.9. Let D_{Λ_j} be the sub-matrix of D containing only the selected atoms (columns) in the dictionary indexed by Λ_j . Determine the orthogonal projection onto the span of selected atoms as: $X_{\Lambda_j} = D_{\Lambda_j}^\dagger Y_i$
      - 2.3.10. Update the temporary sparse matrix: $X^{temp} = X_{\Lambda_j}$
      - 2.3.11. Find the total error: $E = \{\|Y_i - DX^{temp}\|_2^2 + \gamma_1(\|H(X^{temp})\beta - T\|_2^2 + \|\beta\|_2^2) + \gamma_2\|X^{temp}\|_1\}$
      - 2.3.12. If $(E < \varepsilon_3)$ break while loop
      - 2.3.13. $j = j+1$
    - 2.4. end while
    - 2.5. Output: $X_i = X^{temp}$
  3. end for
  4. Output: Sparse code matrix $X = [X_1, X_2, \dots, X_\omega]$
-

5.5.2 Comparison of Two Methods IHT and CSMP for Supervised Sparse Coding

As we mentioned in section 5.5.1, two methods have been suggested to solve the simultaneous optimization problems. To compare the efficiency of these two methods, we designed a preliminary experiment on CK⁺ database. It is notable that IHT is based on l^0 norm of the sparse signal. So, for the fair comparison in this experiment, we applied both methods on the same formulation of the optimization problem defined by $\|X\|_0$. For both methods, we used the sigmoid ELM with the same value for C parameter (refer to Eq. (5-3)). The stop criterion for both algorithms was triggered when the value of the objective function defined in Eq. (5-36) fell below a threshold. The regularization parameter and sparsity threshold were also set the same.

Table 5-1 compares IHT and CSMP for supervised sparse coding of ESL in terms of time complexity and recognition rate for classification of CK⁺ database. Although IHT is faster than CSMP, the convergence of the algorithm to a local minimum is dependent how to set the step-size. However, the recognition rates of both methods are similar. For KESL, due to the convergence problem of IHT and difficulties of parameter setting, we preferred to use CSMP method. Thus, the following experiments in this chapter are based on CSMP algorithm.

Table 5-1: Comparison of two methods for supervised sparse coding applied for CK⁺ database.

Method	Time-complexity (for training)	Detection rate
IHT	110	90.36
CSMP	1404	90.21

5.5.3 Dictionary Update Stage

The objective function is to minimize the reconstruction error of signal Y estimated by sparse signal X as follow:

$$\text{objective: } \min_D \{\|Y - DX\|_2^2\} \quad (5-40)$$

We used the classical “Projected Gradient Descent” [141] to solve this optimization problem. Given the initial dictionary D_0 , the algorithm updates it at each iteration t as follow:

$$D_t = H_D (D_{t-1} - \tau \nabla E) \quad (5-41)$$

where the function H_D is a simple normalizing function that force each dictionary atom to be unit norm. τ is the step size, $E = \|Y - DX\|_2^2$, and $\nabla E = 2(DX - Y)X^t$. X^t denotes the transpose of matrix X .

We randomly selected several training samples from all data classes to initialize the dictionary. Some techniques have been suggested in the literature for finding the step size effectively [142-144]. We followed the Barzilai–Borwein technique [142].

It is notable that we have not used the over-complete dictionary because perfect reconstruction is not our aim. Indeed, for the classification task, the over-completeness is not always necessary as long as discriminative features are captured in sparse coding procedure [68].

5.6 Results and Discussion

We systematically evaluate each component of the proposed facial expression recognition framework. Note that the proposed algorithms were implemented using Matlab 7.11.0 running on a Core i5 CPU (2.8 GHz with 16 GB RAM).

We have evaluated our proposed ESL algorithm on all four databases described in section 2.6 for the facial expression recognition task.

5.6.1 Pre-processing

The pre-processing stage includes face alignment and segmenting the face region from the background.

For CK⁺ database, since the facial landmarks' locations are given in the used dataset, the images are aligned to a constant distance between the two eyes and rotated to align both eyes horizontally. Finally, the faces are cropped using a rectangle of size 100×100.

For own collected database, the same procedure have been done for pre-processing.

For AVEC2011 database, as indicated in the baseline results of this challenge [77], the dataset contains a large amount of data (more than 1.3 million frames). Due to processor and memory constraints, we sample the videos at a constant sampling rate. We partition each video into segments containing 60 frames with 20% overlap between the segments. Each segment is then down sampled at a rate of 6. Thus, each volume data includes only 10 frames. We process only 1550 frames of each video for the training and development subsets (total of 48050 frames for training and 49600 frames for development). Information about the position of the

face and eyes are provided in the database. Thus, similar to CK^+ , the pre-processing stage includes normalization of the face to achieve a constant distance between the two eyes.

For EmotiW2013 database, two methods for face detection and alignment have been used for the baseline results of the challenge: a simple eye based alignment [145] and a fiducial points based warping [146]. We also used the face detected by the first method as provided by the organizers of the EmotiW2013 challenge. Since our proposed feature is not sensitive to the face alignment, the second option can also be used.

5.6.2 Parameter Setting and Initialization

Our proposed classifiers (ESL and KESL) do require selection of a few parameters. We performed a greedy search to select the parameters including the ELM parameter denoted as C (refer to Eq. (5-3)), kernel parameters, and regularization parameters γ_1 and γ_2 (refer to Eq. (5-34)). We used the polynomial kernel ($k(x, y) = (x \cdot y + d)^n$) for KESL after testing the effectiveness of different kernels in our preliminary experiments. The same kernel has been used for SVM and KELM in all experiments. The details of parameter setting for all databases are summarized in Table 5-2.

Table 5-2: Parameter setting for ESL and KESL.

Method	Database					
	ECK ⁺		Avec2011		EmotiW	
	ESL	KESL	ESL	KESL	ESL	KESL
No. of atoms	100	100	150	150	200	200
γ_1	2	1.5	1	2	2	1
γ_2	0.1	0.05	0.05	0.5	0.05	0.01
C	16	1	100	8	16	8
d	-	0.5	-	0.25	-	0.125
n	-	8	-	8	-	6

The sensitivity of the proposed algorithm with respect to the regularization parameters γ_1 and γ_2 is shown in Table 5-3. These results are based on the ECK⁺ database. Our experiments indicate that the best performance is obtained by setting $\gamma_1 > 1 > \gamma_2$. In other words, the best performance is obtained when the classification error term (with weight γ_1) is assigned the highest priority, the reconstruction error term is assigned the second priority, and the sparsity constraint (with weight γ_2) is assigned the least priority. This result is intuitive because we are primarily interested in classification accuracy for this application. However, for other

applications such as noise reduction, the priorities may change. Therefore, the regularization terms should be set according to the application.

We need to initialize the dictionary D and ELM output weight β for proposed ESL classifier. D is initialized by randomly selecting the input samples Y from all classes. To initialize β , we first compute the initial sparse matrix $X = D^\dagger Y$, where $\dagger$ represents the pseudo-inverse, and D is initialized dictionary. Then we apply Eq. (5-3) to get the initial ELM output weight.

One important parameter is the number of dictionary atoms. If this parameter is too small, the dictionary will not be very representative. On the other hand, the execution time would be prohibitively large for a large number of dictionary atoms. In fact, the number of dictionary atoms should be set depending on the characteristics of the database. If the same emotion is expressed differently by different subjects (as in the EmotiW database), the number of dictionary atoms would be on the higher side. However, if the subjects show the same emotion in a similar fashion (as in the CK⁺ database), the number of dictionary atoms need not be large. Figure 5-2 plots the recognition rates of ESL for different number of dictionary atoms for different databases. We observe that the recognition rate increases with the number of dictionary atoms only up to a point (100 atoms for ECK+, 150 atoms for AVEC2011, and 200 atoms for EmotiW in Figure 5-2).

Table 5-3: The sensitivity of the ESL to regularization parameters for ECK⁺ database.

$\gamma_2 \backslash \gamma_1$	0.5	1	1.5	2
0.01	82.69	86.22	82.03	86.82
0.05	88.32	84.38	88.07	87.83
0.1	87.69	86.56	86.34	90.36
0.5	80.59	83.04	89.81	87.07
1	82.87	84.57	83.73	84.61
1.5	78.97	86.89	80.55	85.47

Figure 5-2: Recognition rate of ESL versus number of dictionary atoms.

5.6.3 Performance of ESL and KESL

Table 5-4 compares our proposed methods for classification (ESL and KESL) to other methods in term of final feature dimension, detection rate, and time complexity of training and testing phases individually. If we have sparse coding and classification individually in the training phase, the learned dictionary and training data will be totally different from our proposed approach. We also implemented Sparse ELM (SELM) to check the efficiency of joint optimization problem defined in Eq. (5-34) versus separate optimization. The results show that ESL is superior to ELM and SELM in term of detection rate. KESL has the best accuracy among all other methods but at the cost of time complexity. Although this experiment shows the efficiency of the proposed method, the accuracy improvement is not considerable. The reason is that the total number of own collected samples (75 sequences) is not comparable to the original CK⁺ data (309 sequence). The efficiency of our proposed methods will be more considerable if we train the classifier on the clean data and then test it for own collected samples. We will prove our claim in section 5.6.5.

Table 5-4: Comparison of the proposed pose-invariant descriptor to other classifiers for all databases.

Method	Detection rate					Time complexity (Sec)	
	ECK+	Avec2011		EmotiW		Training	Testing
		Val	Test	Val	Test		
SVM	90.60	62.61	56.52	27.42	22.75	0.96	0.14
ELM	81.25	59.04	52.48	23.18	19.23	0.26	0.04
SELM	86.19	60.14	54.18	24.64	26.93	110.55	7.42
ESL	90.36	66.73	59.21	29.85	27.88	1403.51	5.06
KELM	90.12	60.79	55.19	27.88	23.72	0.06	0.04
KESL	92.74	65.92	61.82	31.34	29.81	1629.20	5.02

5.6.4 Comparison to Other Results

While a number of researchers have reported the performance of their facial emotion recognition algorithms on the CK⁺ benchmark database, these results are not directly comparable to those reported in Table 5-4. This is due to the large differences in the experimental setup (e.g., pre-processing steps, feature extraction method, number of sequences used for training and evaluation, etc.). Therefore, to obtain a meaningful comparison of the proposed ESL classifier with other state-of-the-art classifiers that involve sparse coding, we have evaluated some of the successful techniques reported in the literature using a common experimental setup. In other words, we compared the performance of ESL and KESL to some similar work in this field. All these methods try to classify the sparse code of the original signal for different applications. The results are tabulated in Table 5-5. The source codes of these methods are publicly available*. Since these methods were written using different languages, we did not compare the computational cost of the methods, and limit the comparison to just the detection rate. We optimized the parameters of each method by greedy search and evaluation is conducted with 5-fold CV. From Table 5-5, we observe that the recognition rates of the proposed method are quite comparable to the state-of-the-art methods.

Table 5-6 tabulates the comparison of our results to the baseline and the other reported results for AVEC2011 databases. Comparison using the development set is not a fair measure because the total frames were not included in the experiments. However, for the test set, the proposed descriptor outperforms the one used in the baseline. It also shows that the proposed

*SRC: <http://www.mathworks.com/matlabcentral/fileexchange/35813-sparse-representations-classifier>
 DKSVD, LCKSVD: <http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html>
 FDDL: <http://www4.comp.polyu.edu.hk/~cslzhang/code.htm>

ESL and KESL improve the results of ELM and KELM. Figure 5-3 shows a better comparison between all the reported results and our best results obtained using KESL.

Table 5-5: Comparison of the proposed classification results to other approaches with the same descriptor on ECK⁺.

Method	SRC [147]	DKSVD [69]	LCKSVD [70]	FDDL [72]	ESL	KESL
Detection Rate	89.00	90.11	90.38	91.93	90.36	92.74

Table 5-6: Comparison of the detection rate for the AVEC 2011 database to check the efficiency of the proposed approach. “A” stands for activation, “E” for expectancy, “P” for power, and “V” for valence.

Reference	Development					Test				
	A	E	P	V	Average	A	E	P	V	Average
Baseline [77]	60.2	58.3	56.3	63.6	59.6	42.2	53.6	36.4	52.5	46.2
[106]	81.7	73.1	73.3	73.5	75.4	65.5	61.7	47.1	69.8	61.0
[148]	69.3	65.6	59.9	67.8	65.6	56.5	59.7	48.5	59.2	55.9
[108]	58.2	53.5	53.7	53.2	54.65	56.9	47.5	47.3	55.5	51.8
Pose Invariant Descriptor + SVM	62.6	66.4	60.7	60.7	62.6	57.7	60.0	48.4	60.0	56.5
Pose Invariant Descriptor + ELM	61.4	57.3	59.5	57.9	59.0	58.0	51.9	47.9	52.2	52.5
Pose Invariant Descriptor + ESL	64.6	65.0	65.2	66.7	65.4	62.4	62.4	52.4	59.7	59.2
Pose Invariant Descriptor + KELM	62.6	59.5	61.6	59.5	60.8	55.7	56.4	52.4	56.4	55.2
Pose Invariant Descriptor + KESL	65.2	65.8	66.3	65.9	65.8	63.2	63.0	58.1	63.0	61.8

Figure 5-3: Comparison of the recognition results of test set for AVEC2011.

We have done the same experiments using the EmotiW database and the result is tabulated in Table 5-7. The accuracy of the proposed descriptor with SVM classifier is around the baseline results for both validation and test sets. However, the results increased by changing the classifier to ESL and KESL. We also showed the details of the best result obtained by KESL in form of confusion matrix in Table 5-8. Our method is better in identifying “anger”, “happiness”, and “neutral” with accuracies of 55.55%, 54.00%, and 56.25% respectively, but not very successful in identifying “disgust”, “fear”, “sadness”, and “surprize”. It can also be seen that a high percentage of emotions are wrongly classified as “neutral” which means that our method is still not very effective to detect subtle emotions. The comparison on the test set with the other methods is depicted in Figure 5-4.

It is notable that we just included the results of the vision part and ignored the audio-based or the audio-video-based results reported by the researchers in competitions. As shown, our result is comparable to the best archived in the competition for both databases.

Table 5-7: Comparison of our proposed methods for EmotiW2013 database to the baseline results.

Method	set	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
Baseline [78]	Val	44.00	20.00	14.81	43.55	34.55	20.31	9.62	27.27
	Test	50.00	12.24	0.00	48.00	18.75	6.97	5.71	22.75
Pose Invariant Descriptor +SVM	Val	32.58	6.38	29.15	58.28	48.93	11.75	4.85	27.42
	Test	35.18	0.00	3.03	56.00	43.75	4.65	0.00	22.75
Pose Invariant Descriptor +ELM	Val	42.38	10.96	13.25	42.38	50.39	2.85	0.00	23.17
	Test	42.59	18.36	6.06	24.00	16.66	4.65	11.42	19.23
Pose Invariant Descriptor +KELM	Val	33.33	0.00	15.15	66.00	58.33	4.65	2.85	27.88
	Test	55.55	0.00	0.00	62.00	27.08	0.00	0.00	23.72
Pose Invariant Descriptor +ESL	Val	36.36	12.00	21.43	56.36	38.78	17.65	22.73	29.85
	Test	33.33	0.00	15.15	66.00	58.33	4.65	2.86	27.88
Pose Invariant Descriptor +KESL	Val	20.45	18.00	28.57	58.18	38.65	15.69	22.73	31.34
	Test	55.55	4.08	3.03	54.00	56.25	9.30	5.71	29.81

Table 5-8: Confusion matrix for seven emotions for EmotiW2013 database.

Predict \ Actual	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	30	5	0	7	7	3	2
Disgust	17	2	3	13	8	5	1
Fear	8	2	1	1	12	4	5
Happiness	6	3	0	27	10	3	1
Neutral	7	3	0	6	27	4	1
Sadness	17	4	0	7	10	4	1
Surprise	8	1	0	8	10	6	2

Figure 5-4: Comparison of the results for EmotiW2013 database for test set.

Indeed, unlike the CK+ database, the AVEC 2011 and EmotiW databases are quite representative of real-world applications because they contain samples with natural or spontaneous emotions that do not exhibit sharp facial changes from the start to apex of an expression. This is one of the main reasons for the huge difference between the recognition rates reported in the fourth (AVEC 2011 - Test) and sixth (EmotiW - Test) columns of Table 5-4 compared to those reported in the second column (ECK+) of Table 5-4. However, the proposed system is able to achieve recognition rates that are comparable to the state-of-the-art performance reported on these two databases (as indicated by the results in Figure 5-3 and Figure 5-4). This shows that the proposed emotion recognition system is indeed capable of recognizing natural emotions with subtle changes in facial expression, although there is a scope for significant improvement in the recognition accuracy in such scenarios.

5.6.5 Advantages and Limitations of the Proposed Classifiers

A possible limitation of the proposed emotion recognition system is the need for database-specific tuning of parameters as described in section 5.6.2. However, it must be emphasized that this phenomenon is not unique to the proposed system, but is common to most pattern recognition systems. In fact, we have performed database-specific parameter tuning via greedy search for all the competing systems reported in Tables 5-2, 5-4, and 5-5. In the case of AVEC 2011 and EmotiW databases, the primary purpose of having a validation or development subset is to allow tuning of parameters before the model is evaluated on the test set. Thus, the

---

comparison of results in Tables 5-6 and 5-7 is fair, with all the competing approaches being allowed the luxury of parameter tuning.

One approach to measure the sensitivity of pattern recognition systems to various parameters is to evaluate the generalization performance on unseen test data. While the results in sections 5.6.2 to 5.6.4 demonstrate that the proposed ESL and KESL algorithms have good generalization performance, we observe that the accuracy improvement is not very significant in most cases. For example on the ECK+ database, the accuracy of KESL is only $\sim 2\%$ better than the competing approaches. However, we believe that the advantage of the proposed algorithms can be readily observed if we train the classifier on clean data and test it using samples with large intra-class variations (i.e., the training set is no longer representative of the test set). To investigate this claim, we train the ESL and KESL classifiers on the original CK+ data and then test it using our own collected samples (occlusion, pose variation, and illumination changes). Figure 5-5 illustrates the results of this experiment, which clearly demonstrates the advantage of the ESL and KESL methods over other classifiers in terms of generalization performance. The reason for this better generalization performance is that ESL algorithms do not directly use the noisy input samples, but only the sparse coefficients based on a learned dictionary. Thus, ESL algorithms are indeed more robust to noisy and imperfect test data, but this comes at the cost of longer execution times during training and testing.

Figure 5-5: Comparison the robustness of the methods to occlusion, head posed variations, and illumination changes.

5.7 Analysis of Failure Cases

In this section, we investigated several sources of error for feature extraction including failure of face detector, face alignment, OF computation, and reference point localization. We have done some analysis on EmotiW database to check the source of failures as follow:

5.7.1 Failure of Face Detector

The first source of error is due to the failure of face detector to correctly detect the faces. For example, in the validation set, no face was detected for 60 sequences out of 389 sequences which caused 15.42% error in the final recognition rate.

False positive faces detected by the face detector is another source of error. Many sequences consist of non-face frames as displayed in Figure 5-6. To overcome this problem, we could filter out the non-face samples using the method used in [149].

There are also many cases where the face detector detected only a part of the face instead of the whole face. Figure 5-7 shows a few samples of this case.

Another source of error is due to the lack of face variation showing an expression. For example, Figure 5-8 shows the faces detected in a whole sequence. As shown, all faces represent the apex and there is no facial movement due to expression from neutral to apex. In many cases in this database, the face detector was failed to detect the right faces from all video

frames showing the dynamics of the expression. Since the proposed descriptor only encodes the dynamic information (motion), it fails to accurately extract our target features. So the proposed system incorrectly predicts the “neutral” label for them. Incorporating the static features may be helpful to overcome this problem.

Figure 5-6: Non-faces samples.

Figure 5-7: Failure of face detector to detect the whole face.

Figure 5-8: All faces detected from a video clips showing disgust.

5.7.2 Failure of Optical Flow

When the range of head movement is very large, the OF algorithm fails to correctly compute the facial movement caused by the emotion even after applying the OF correction technique explain in section 4.2.2.

Figure 5-9 illustrates an example of OF failure due to large head movement. As shown in this figure, the head turned left, but the estimated OF dose not capture this information. We zoomed out the nose OF for better illustration in Figure 5-9d.

Figure 5-9: Failure of optical flow. (a)-(b) two consequent frames; (c) estimated optical flow; (d) illustration of optical flow for nose region.

5.7.3 Failure of Reference Point Detection

The error of the nose point localization affects the feature extraction process. Figure 5-10 shows a few sample faces with wrong reference point localization.

Figure 5-10: Wrong reference point localization.

5.8 Concluding Remarks

In this chapter, we proposed our second strategy dealing to uncontrolled situations facing in real world application: we proposed ESL and KESL as robust classifiers by incorporating a sparse coding criterion and a nonlinear classification error criterion into a unified objective function. The proposed approach combines the discriminative power of the ELM with the reconstruction property of the sparse representation to enable to deal with noisy signals and imperfect data recorded in natural settings. To solve the objective function of our proposed ESL including both linear and no-linear terms, we proposed two different methods: IHT and CSMP. We have shown that IHT method is superior to CSMP in term of time complexity; however, it may not converge to an optimal solution.

We presented the results of our extensive experiments at the end of this chapter. We evaluated the efficiency of the proposed descriptor and recognition frameworks especially in the presence of head pose variations, occlusion and illumination variations. The proposed approach is able to perform well even if the classifier is trained with frontal or near-frontal faces but tested on non-frontal faces. The performance of ESL and KESL is highly dependent on the approach utilized for supervised sparse representation. The accuracy obtained for all databases proved the efficiency of ESL and KESL over ELM and KELM.

Chapter 6

Conclusion and Recommendations for Future Work

6.1 Conclusion

In this thesis, we have presented an effective approach for facial emotion recognition that can handle challenges such as illumination variation, expression speed variations, different image resolution, head pose changes, and partial occlusion. The proposed approach has novelty in both the feature extraction and recognition. We have performed extensive experiments on both acted and spontaneous emotion databases to evaluate the effectiveness of the proposed feature extraction and recognition schemes under different scenarios. The results clearly demonstrate the robustness of the proposed emotion recognition framework, especially in challenging scenarios that involve illumination changes, occlusion, and head pose variations. However, possible limitations include higher computational cost for both feature extraction and classification and parameter optimization.

In this final chapter, we summarize the contributions of this thesis in the followings and then suggest some possible future improvements.

1. **Robust Representation:** we proposed a novel OF-based spatio-temporal descriptor for pose invariant facial emotion detection. Our contributions in this part are:
  - A. We proposed a new set of spatio-temporal descriptors to capture the dynamic information hidden in a flow field. Divergence and curl features were used before by other researches, but projection and rotation descriptors are new and indicate the movement characteristics of each facial point with respect to a reference point not affected by expression changes.

- B. The extracted features are encoded for pose-invariant purpose. It is not necessary to know the details of all movement information to discriminate different emotions. As such, we encoded only the sign information of each feature which is not sensitive to head pose variations.
- C. Only the statistics of the extracted features is retained as discriminative information for further processing. Here, we introduced two types of histogram to condense the extracted features, namely the weighted and un-weighted histogram, to be able to detect both the level and dynamic of expressions.

The proposed feature set is able to handle the emotion speed variations and also the head pose variations even for natural expression. It can also handle continuous head pose variation even if the classifier is trained using only the frontal and near-frontal faces. However the performance of the proposed features is dependent on the accuracy of the computed OF.

2. **Robust Recognition:** a novel classification scheme called ESL is proposed. ESL incorporates the reconstruction properties of sparse representation and discrimination power of a nonlinear ELM for robust classification. Our contributions in this part are:
  - A. We formulated the classification model by adding the ELM error term into the objective function of the conventional sparse representation to learn a dictionary which is both discriminative and reconstructive.
  - B. We proposed two methods for solving the objective function of our proposed ESL (containing both linear and no-linear terms). The first method was IHT and we developed it for our problem. The second named CSMP was our novel method that can be applied for any similar objective function. We have shown that IHT method is superior to CSMP in term of time complexity; however, it may not converge to an optimal solution.
  - C. We extended the recognition framework by introducing the kernel function to the ESL classifier.

The proposed classifier does not directly use the imperfect signal for classification, but it is able to reconstruct it from clean data based on the learned dictionary. Therefore, the method can tolerate the errors caused by partial occlusion, illumination and pose variations.

6.2 Recommendations for Future Work

Although the efficiency of the proposed methods has been proved in chapter 5, there are some limitations that can be further enhanced. These are:

---

1. Minimizing the error caused by face detector.

We have shown in section 5.7.1 that a high percentage of the error encountered is due to the face detector failure. We propose to minimize the error caused by face detector by using more robust face detector and to eliminate false or non-face frames. Applying some recent accurate techniques for face detection may be helpful [154]. Additionally, an automatic post-processing technique is needed to remove falsely detected faces [149].

2. Improving the feature extraction method.

We suggest several ideas to enhance the performance of the extracted features:

- A. Using more accurate technique for OF estimation. Since facial patterns include many soft and non-textured regions, facial movement estimation is challenging for dense OF estimation and more accurate OF technique is needed.
- B. Applying a better technique for head-movement estimation and OF correction. As we mentioned in section 4-2-2, only the movement related to facial expression is of interest and the global motion of the head should be subtracted from the flow vector. Our proposed method for OF correction is not able to handle very complex and extreme head motion. As our estimation is based on the projection of 3D head movement on a 2D plane, using the 3D techniques for head movement estimation may be more accurate for real world applications.
- C. Removing the variation due to a person speaking from facial emotion. Speaking effect is a crucial source of noise that degrades the recognition performance. In our proposed system, we assume that the proposed sparse representation classifier is able to handle this kind of noise. We could incorporate the approach in [155] to reduce this effect.
- D. Combining the appearance based features with the motion features. The result section illustrates the strengths of our proposed descriptor for face emotion recognition. However fusing both appearance and motion features may result in better performance [47]. In addition, the appearance based feature is more useful when the motion feature fails, such as due to subtle expression or OF failure.
- E. Incorporating the local phase and local energy information of the optical flow field to the feature set. We have shown the effectiveness of phase based approaches to represent the local structural information of the facial image in chapter 3. Utilizing the phase information either for OF estimation or feature encoding is a promising but less explored research topic.

3. Using motion exaggeration approach to detect the subtle facial emotions.

In the posed facial expression, facial muscle movement is quite clear and is easier to detect. However, for spontaneous and natural facial expression, detection of facial

---

component movements is not an easy task. Indeed, the facial descriptor should be able to describe every subtle movement. Motion exaggeration techniques can be helpful to solve this [156].

4. Improving the methods used for supervised sparse coding.

As we highlighted in section 5.5.1, there is no method for supervised sparse coding when there is a nonlinear term involved in the formulation. The first method we used (IHT) was fast but the convergence is not guaranteed. The second proposed method (CSMP) will converge but is has higher complexity. These methods can be improved to achieve better accuracy and speed.

Appendix A

Gradient calculation for KELM related error term

Referring to Eq. (5-36), we can rewrite the error term as:

$$E = E_1 + \gamma E_2 \quad (\text{A-1})$$

where $E_1 = \|Y - DX\|_2^2$ is the reconstruction error term and $E_2 = \|H(X)\beta - T\|_2^2$ is KELM related error term. Now, the problem is to find the gradient of E_2 , where $H(X)$ is unknown.

Referring to section 5.2.2, the kernel Matrix for ELM is defined as follow:

$$\Omega_{ELM} = HH' \quad (\text{A-2})$$

$$\Omega_{ELM i,j} = h(x_i)h(x_j) = K(x_i, x_j) \quad (\text{A-3})$$

Then, the output weight β can be re-written as:

$$\beta = H'(\frac{I}{C} + \Omega_{ELM})^{-1}T \quad (\text{A-4})$$

Then, for known β , $\frac{dE_2}{dx}$ can be calculated as follow:

$$\frac{dE_2}{dx} = 2 \frac{dH(x)}{dx} \beta (H(x)\beta - T) \quad (\text{A-5})$$

If we replace Eq. (A-4) into Eq. (A-5), we will have:

$$\frac{dE_2}{dx} = 2 \frac{dH(x)}{dx} H'(X_{tr}) \left(\frac{I}{C} + \Omega_{ELM} \right)^{-1} T (H(x)H'(X_{tr}) \left(\frac{I}{C} + \Omega_{ELM} \right)^{-1} T - T) \quad (\text{A-6})$$

where X_{tr} is the training data and $H(x)$ is a function of unknown x , but $H(X_{tr})$ is a function of known X_{tr} . Eq. (A-6) can be rewritten as:

$$\frac{dE_2}{dx} = 2 \frac{dH(x)}{dx} H'(X_{tr}) \alpha (H(x)H'(X_{tr}) \alpha - T) \quad (\text{A-7})$$

where $\alpha = (\frac{I}{C} + \Omega_{ELM})^{-1}T$.

Now, referring to Eq. (A-3), we use the kernel function for $H(x)H'(X_{tr})$ as follow:

$$\frac{dE_2}{dx} = 2 \frac{dK(x, X_{tr})}{dx} \alpha(K(x, X_{tr})\alpha - T) \quad (\text{A-8})$$

Appendix B

Restricted Isometry Property (RIP) for Sigmoid ELM

As we mentioned in Eq. (5-4), the output function of ELM is described by $f(x)$:

$$f(x) = H(x)\beta \quad (\text{B-1})$$

$f(x)$ is a nonlinear function applied to each element of the vector x . As suggested by [137], the Jacobian of $f(x)$ should satisfy the RIP to guarantee that the IHT algorithm can find a sparse vector that is close to the true minimizer of the cost function $\|f(x) - T\|^2$ (refer to Eq. (5-1)). The Jacobian of $f(x)$ can be written as:

$$f_{x^*} = H'(x)\beta \quad (\text{B-2})$$

Since the ELM training algorithm try to reach the minimum norm of output weights $\|\beta\|$, then we can assume $\|\beta\| < M$, with M as a threshold. Then for all x_1, x_2 belongs to the constraint set A , we have:

$$\|f_{x^*}(x_1 - x_2)\| \leq \|H'(x)(x_1 - x_2)\| \|\beta(x_1 - x_2)\| \quad (\text{B-3})$$

where $H(x) = [G(a_1, b_1, x), \dots, G(a_L, b_L, x)]$, and $G(a, b, X)$ is a sigmoid function defined by Eq. (B-4).

$$G(a, b, x) = \frac{1}{1 + \exp(-(ax + b))} \quad (\text{B-4})$$

To find the upper limit of $\|f_{x^*}(x_1 - x_1)\|$, it is enough to find the upper limit of $\|H'(x)\|$ as follows:

$$H'(x) = \frac{a \times \exp(-(ax + b))}{[1 + \exp(-(ax + b))]^2} \quad (\text{B-5})$$

$$\|H'(x)\| \leq \frac{a \times \exp(-b)}{[1 + \exp(-b)]^2} \quad (\text{B-6})$$

where a and b are randomly generated values according to any continuous probability distribution. If we assume $\frac{a \times \exp(-b)}{[1 + \exp(-b)]^2} < N$, then the RIP will be satisfied by following upper bound:

$$\|f_{x^*}(x_1 - x_1)\| \leq NM\|x_1 - x_2\| \quad (\text{B-7})$$

Appendix C

Restricted Isometry Property (RIP) for Kernel ELM

For KELM, $H(x)$ is unknown. We have shown in Appendix A that $f_{x^*} = \frac{dK(x, X_{tr})}{dx} \alpha$. Here, we discuss about RIP for linear and polynomial kernels.

- Linear Kernel

For linear ELM, $K(x, y) = xy$, so we have:

$$f_{x^*} = \frac{dK(x, X_{tr})}{dx} \alpha = X_{tr} \alpha = \beta \quad (C-1)$$

Since the ELM training algorithm try to minimize $\|\beta\|$, then we can assume $\|\beta\| < M$. Then the RIP will be satisfied by following upper bound:

$$\|f_{x^*}(x_1 - x_2)\| \leq M \|x_1 - x_2\| \quad (C-2)$$

- Polynomial Kernel

For Polynomial Kernel, $K(x, y) = (xy + b)^a$

$$f_{x^*} = \frac{dK(x, X_{tr})}{dx} \alpha = a X_{tr} (x X_{tr} + b)^{a-1} \alpha \quad (C-3)$$

Since f_{x^*} depends on kernel parameters and also estimated sparse coefficients, we are not able to find the upper bound of Eq. (C-3) theoretically. Indeed, to satisfy the RIP, the kernel parameters cannot be arbitrarily chosen. In other words, the convergence of iterative hard thresholding to local minimizers of the nonlinear term $E_2 = \|H(X)\beta - T\|_2^2$ will achieve for certain limits of kernel parameters.

Author's Publications

Journal Papers:

1. S. Shojaeilangari, W. Y. Yau, J. Li, and E. K. Teoh., "*Dynamic facial expression analysis based on extended spatio-temporal histogram of gradients*", International Journal of Biometrics Vol. 6. No. 1, pp. 33-52, 2014.
2. S. Shojaeilangari, W. Y. Yau, J. Li, and E. K. Teoh., "*Multi-scale Analysis of Local Phase and Local Orientation for Dynamic Facial Expression Recognition*", invited paper, Journal of Multimedia Theory and Applications (JMTA), Vol. 2, No. 1, pp. 1-10, 2014.
3. S. Shojaeilangari, W. Y. Yau, and E. K. Teoh., "*A Novel Phase Congruency based Descriptor for Dynamic Facial Expression Analysis*", Pattern Recognition Letters, Vol. 49, pp. 55-61, 2014.
4. S. Shojaeilangari, W. Y. Yau, k. Nandakumar, J. Li, and E. K. Teoh., "*Robust Representation and Recognition of Facial Emotions Using Extreme Sparse Learning*", Accepted to publish to IEEE transactions on Image Processing, March 2015.

Conference Papers:

1. S. Shojaeilangari, W. Y. Yau, and E. K. Teoh., "*Person Independent Facial Expression Analysis using Gabor Features and Genetic Algorithm*", 8th International Conference on Information, Communications and Signal Processing (ICICS), Singapore, 2011, pp. 1-5.
2. S. Shojaeilangari, W. Y. Yau, J. Li, and E. K. Teoh., "*Feature Extraction through Binary Pattern of Phase Congruency for Facial Expression Recognition*", 12th International Conference on Control, Automation, Robotics and Vision (ICARCV 2012), China, May 2012, pp. 166-170.
3. S. Shojaeilangari, W. Y. Yau, J. Li, and E. K. Teoh., "*Dynamic facial expression analysis based on histogram of local phase and local orientation*", International Conference on Multimedia and Human Computer Interaction (MHCI 2013), Canada, Jul 2013, pp. 57_1-57_8.

-
4. Shojaeilangari, W. Y. Yau, and E. K. Teoh., “*Pose-invariant Descriptor for Facial Emotion Recognition*”, to be submitted to British Machine Vision Conference (BMCV 2015), Sep 2015.

Bibliography

- [1] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, pp. 116-134, 2007.
- [2] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. A. C. Hutcherson, C. Nass, and O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int. Journal of Human-Computer Studies*, vol. 66, pp. 303-317, 2008.
- [3] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," *IEEE Int. Conf. Image Processing (ICIP)*, pp. II-370-3, 2005.
- [4] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction," *Computer Vision and Pattern Recognition Workshop (CVPRW '03)*, pp. 53-53, 2003.
- [5] D. H. Shin, "Potential user factors driving adoption of IPTV. What are customers expecting from IPTV?," *Technological Forecasting and Social Change*, vol. 74, pp. 1446-1464, 2007.
- [6] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915-928, 2007.
- [7] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, pp. 607-619, 2011.
- [8] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, pp. 803-816, 2009.
- [9] D. H. Kim, S. U. Jung, and M. J. Chung, "Extension of cascaded simple feature based face detection to facial expression recognition," *Pattern Recognition Letters*, vol. 29, pp. 1621-1631, 2008.
- [10] S. Bashyal and G. K. Venayagamoorthy, "Recognition of facial expressions using Gabor wavelets and learning vector quantization," *Engineering Applications of Artificial Intelligence*, vol. 21, pp. 1056-1064, 2008.
- [11] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognition*, vol. 45, pp. 80-91, 2012.
- [12] X. Xie and K.-M. Lam, "Facial expression recognition based on shape and texture," *Pattern Recognition*, vol. 42, pp. 1003-1011, 2009.

-
- [13] T. Wehrle, S. Kaiser, S. Schmidt, and K. R. Scherer, "Studying the Dynamics of Emotional Expression Using Synthesized Facial Muscle Movements," *Journal of Personality and Social Psychology*, vol. 78, pp. 105-119, 2000.
  - [14] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling," *IEEE Trans. Cybernetics*, vol. 44, pp. 161-174, 2014.
  - [15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 39-58, 2009.
  - [16] J. Kim and E. André, "Four-Channel Biosignal Analysis and Feature Extraction for Automatic Emotion Recognition," *Biomedical Engineering Systems and Technologies*, vol. 25, pp. 265-277, 2009.
  - [17] M. Meghjani, F. Ferrie, and G. Dudek, "Bimodal information analysis for emotion recognition," *Applications of Computer Vision (WACV)*, pp. 1-6, 2009.
  - [18] H. Gunes and M. Piccardi, "Fusing face and body gesture for machine recognition of emotions," *Robot and Human Interactive Communication (ROMAN)*, pp. 306-311, 2005.
  - [19] Z. Zeng, M. Pantic, and T. S. Huang, "Emotion Recognition Based on Multimodal Information," *Affective Information Processing*, pp. 241-265, 2009.
  - [20] E. Shouse, "Feeling, Emotion, Affect," *M/C Journal*, vol. 8, 2005.
  - [21] P. Ekman, "Strong Evidence for Universals in Facial Expressions: a Reply to Russell's Mistaken Critique," *Psychological Bulletin*, vol. 115, pp. 268-287, 1994.
  - [22] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161-1178, 1980.
  - [23] Z. Shiliang, H. Qingming, J. Shuqiang, G. Wen, and T. Qi, "Affective Visualization and Retrieval for Music Video," *IEEE Trans. Multimedia*, vol. 12, pp. 510-522, 2010.
  - [24] M. A. Nicolaou, H. Gunes, and M. Pantic, "Designing frameworks for automatic affect prediction and classification in dimensional space," *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 20-26, 2011.
  - [25] L. Zhang, D. Tjondronegoro, and V. Chandran, "Evaluation of Texture and Geometry for Dimensional Facial Expression Recognition," *Digital Image Computing Techniques and Applications (DICTA)*, pp. 620-626, 2011.
  - [26] A. Lichtenstein, A. Oehme, S. Kupschick, and T. Jürgensohn, "Comparing Two Emotion Models for Deriving Affective States from Physiological Data," *Affect and Emotion in Human-Computer Interaction*. vol. 4868, pp. 35-50, 2008.
  - [27] T. Eerola, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music* vol. 39, pp. 18-49, 2011.
  - [28] J. F. Cohn, Z. Ambadar, and P. Ekman, "Observer-Based Measurement of Facial Expression with the Facial Action Coding System," *The handbook of emotion elicitation and assessment*, Oxford University Press Series in Affective Science, 2005.
  - [29] Y. I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 97-115, 2001.

-
- [30] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, pp. 32-80, 2001.
  - [31] Y. Wang and L. Guan, "Recognizing Human Emotional State From Audiovisual Signals*," *IEEE Trans. Multimedia*, vol. 10, pp. 936-946, 2008.
  - [32] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 568-573, 2005
  - [33] J. Whitehill and C. W. Omlin, "Haar features for FACS AU recognition," *Automatic Face and Gesture Recognition (FGR)*, pp. 97-101, 2006.
  - [34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol.1, pp. I-511-I-518, 2001.
  - [35] L. Kun and Z. Xin, "Facial Expression Recognition from Image Sequences Based on Feature Points and Canonical Correlations," *Artificial Intelligence and Computational Intelligence (AICI)*, pp. 219-223, 2010.
  - [36] T. Senechal, V. Rapp, H. Salam, R. Segulier, K. Bailly, and L. Prevost, "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, pp. 993-1005, 2012.
  - [37] A. R. Rivera, J. R. Castillo, and O. Chae, "Local Directional Number Pattern for Face Analysis: Face and Expression Recognition," *IEEE Trans. Image Processing*, vol. 22, 2013.
  - [38] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1424-1445, 2000.
  - [39] S. Mingli, B. Jiajun, and C. Chun, "Expression recognition from video using a coupled hidden Markov model," *TENCON*, Vol. 1, pp. 583-586, 2004.
  - [40] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 699-714, 2005.
  - [41] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," *IEEE Trans. Image Processing*, vol. 16, pp. 172-187, 2007.
  - [42] I. Cohen, A. Garg, and T. Huang, "Emotion recognition from facial expressions using multilevel HMM," *Neural Inform. Process. Syst.*, 2000.
  - [43] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding - Special issue on Face recognition*, vol. 91, pp. 160-187, 2003.
  - [44] K.-E. Ko and K.-B. Sim, "Development of a Facial Emotion Recognition Method Based on Combining AAM with DBN," *Cyberworlds (CW)*, pp. 87-91, 2010.
  - [45] K.-E. Ko and K.-B. Sim, "Facial emotion recognition using a combining AAM with DBN," *Control Automation and Systems (ICCAS)*, pp. 1436-1439, 2010.

-
- [46] J. F. Cohn, A. J. Zlochow, J. J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," *IEEE Conf. Automatic Face and Gesture Recognition*, pp. 396-401, 1998.
  - [47] U. Tariq, L. Kai-Hsiang, L. Zhen, Z. Xi, W. Zhaowen, L. Vuong, T. S. Huang, L. Xutao, and T. X. Han, "Recognizing Emotions From an Ensemble of Features," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, pp. 1017-1026, 2012.
  - [48] W. S. Chu, F. De la Torre, and Cohn J. F., "Selective Transfer Machine for Personalized Facial Action Unit Detection," *Computer Vision and Pattern Recognition (CVPR)*, pp. 3515-3522, 2013.
  - [49] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A Set of Selected SIFT Features for 3D Facial Expression Recognition," *Int. Conf. Pattern Recognition (ICPR)*, pp. 4125-4128, 2010.
  - [50] G. Camara-Chavez and A. A. Araujo, "Harris-SIFT Descriptor for Video Event Detection Based on a Machine Learning Approach," *IEEE Int. Sym. Multimedia (ISM)*, pp. 153-158, 2009.
  - [51] T. R. Almaev and M. F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition," *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356-361, 2013.
  - [52] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," *IEEE Conf. Automatic Face & Gesture Recognition and Workshops (FG)*, pp. 314-321, 2011.
  - [53] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72, 2005.
  - [54] T. Guha and R. K. Ward, "Learning Sparse Representations for Human Action Recognition," *IEEE Trans. Pattern Analysis And Machine Intelligence*, vol. 34, pp. 1576-1588, 2012.
  - [55] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, pp. 433-449, 2006.
  - [56] S. Park and D. Kim, "Subtle facial expression recognition using motion magnification," *Pattern Recognition Letters*, vol. 30, pp. 708-716, 2009.
  - [57] S. Park and D. Kim, "Spontaneous facial expression classification with facial motion vectors," *IEEE Conf. Automatic Face & Gesture Recognition (FG)*, pp. 1-6, 2008.
  - [58] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1357-1369, 2013.
  - [59] L. A. Jeni, A. Lorincz, T. Nagy, Z. Palotai, J. Sebok, Z. Szabo, and D. Takacs, "3D shape estimation in video sequences provides high precision evaluation of facial expressions," *Image and Vision Computing*, vol. 30, pp. 785-795, 2012.
  - [60] Y. Songfan and B. Bhanu, "Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, pp. 980-992, 2012.

-
- [61] W. Zheng, H. Tang, Z. Lin, and T. Huang, "Emotion Recognition from Arbitrary View Facial Images," *ECCV*, vol. 6316, pp. 490-503, 2010.
  - [62] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates," *Int. J. Computer Vision*, vol. 83, pp. 178-194, 2009.
  - [63] A. Sánchez, J. V. Ruiz, A. B. Moreno, A. S. Montemayor, J. Hernández, and J. J. Pantrigo, "Differential optical flow applied to automatic facial expression recognition," *Neurocomputing*, vol. 74, pp. 1272-1282, 2011.
  - [64] R. Niese, A. Al-Hamadi, A. Farag, H. Neumann, and B. Michaelis, "Facial expression recognition based on geometric and optical flow features in colour image sequences," *Computer Vision - IET*, vol. 6, pp. 79-89, 2012.
  - [65] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3501-3508, 2010.
  - [66] D. Wang and S. Kong, "A classification-oriented dictionary learning model: Explicitly learning the particularity and commonality across categories," *Pattern Recognition*, vol. 47, pp. 885-898, 2014.
  - [67] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised Dictionary Learning," *Advances in Neural Information Processing Systems 21 (NIPS'08)*, 2008.
  - [68] J. Mairal, F. Bach, and J. Ponce, "Task-Driven Dictionary Learning," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 34, pp. 791-804, 2012.
  - [69] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2691-2698, 2010.
  - [70] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1697-1704, 2011.
  - [71] Z. Jiang, Z. Lin, and L. S. Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2651-2664, 2013.
  - [72] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher Discrimination Dictionary Learning for sparse representation," *IEEE Int. Conf. Computer Vision (ICCV)*, pp. 543-550, 2011.
  - [73] K. Huang and S. Aviyente, "Sparse Representation for Signal Classification," *Advances in Neural Information Processing Systems*, 2006.
  - [74] B. He, D. Xu, R. Nian, M. Heeswijk, Q. Yu, Y. Miche, and A. Lendasse, "Fast Face Recognition Via Sparse Coding and Extreme Learning Machine," *Cognitive Computation*, vol. 6, pp. 264-277, 2014.
  - [75] C. Anitha, M. K. Venkatesha, and B. S. Adiga, "A survey on facial expression databases," *Int. J. Engineering Science and Technology*, vol. 2, pp. 5158-5174, 2010.
  - [76] 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," *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94-101.

-
- [77] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–The First International Audio/Visual Emotion Challenge," *Affective Computing and Intelligent Interaction*, vol. 6975, pp. 415-424, 2011.
  - [78] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion Recognition In The Wild Challenge and Workshop (EmotiW 2013)," *ACM International Conference on Multimodal Interaction (ICMI)*, 2013.
  - [79] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research," *7th Int. Conf. Automatic Face and Gesture Recognition (FG)*, 2006.
  - [80] C. X. Yin L, Sun Y, Worm T, Reale M, "A High-Resolution 3D Dynamic Facial Expression Database," *8th Int. Conf. Automatic Face and Gesture Recognition (FG)*, 2008.
  - [81] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial Action Coding System: Investigators guide," Salt Lake City, UT: Research Nexus, 2002.
  - [82] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based Database for Facial Expression Analysis," *IEEE Int. Conf. Multimedia and Expo (ICME)*, 2005.
  - [83] P. J. Phillips, H. Moon, S. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, 2000.
  - [84] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," *3rd Int. Conf. Automatic Face and Gesture Recognition (FG)*, 1998.
  - [85] H. Meng and N. Bianchi-Berthouze, "Affective State Level Recognition in Naturalistic Facial and Vocal Expressions," *IEEE Trans. Cybernetics* vol. 44, pp. 315-328, 2014.
  - [86] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies," *IEEE Trans. Multimedia*, vol. 19, pp. 34-41, 2012.
  - [87] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition," *ACM Multimedia*, pp. 357-360, 2007.
  - [88] H.-Y. Cheng and J.-N. Hwang, "Integrated video object tracking with applications in trajectory-based event detection," *J. Visual Communication and Image Representation*, vol. 22, pp. 673-685, 2011.
  - [89] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, pp. 323-333, 2011.
  - [90] C.-C. Lien, C.-L. Chiang, and C.-H. Lee, "Scene-based event detection for baseball videos," *J. Visual Communication and Image Representation*, vol. 18, pp. 1-14, 2007.
  - [91] S. Shojaeilangari, W. Y. Yau, J. Li, and E. K. Teoh, "Dynamic facial expression analysis based on extended spatio-temporal histogram of gradients," *Int. J. Biometrics* vol. 6, pp. 33-52, 2014.
  - [92] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, pp. 886-893, 2005.
  - [93] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Sixth Int. Conf. Computer Vision*, pp. 839-846, 1998.
-

-
- [94] M. Zhang and B. Gunturk, "A new image denoising method based on the bilateral filter," *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 929-932, 2008.
  - [95] R. Lior, "Genetic algorithm-based feature set partitioning for classification problems," *Pattern Recognition*, vol. 41, pp. 1676-1700, 2008.
  - [96] E. D. Übeyli, "Multiclass support vector machines for diagnosis of erythematous diseases," *Expert Systems with Applications*, vol. 35, pp. 1733-1740, 2008.
  - [97] S. Shojaeilangari, W. Y. Yau, J. Li, and E. K. Teoh, "Multi-scale Analysis of Local Phase and Local Orientation for Dynamic Facial Expression Recognition," *J. Multimedia Theory and Applications (JMTA)*, vol. 2, pp. 1-10, 2014.
  - [98] K. Rajpoot, V. Grau, and J. A. Noble, "Local-phase based 3D boundary detection using monogenic signal and its application to real-time 3-D echocardiography images," *IEEE Int. Sym. Biomedical Imaging: From Nano to Macro (ISBI)*, pp. 783-786, 2009.
  - [99] M. Flesberg, and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Processing*, vol. 49, pp. 3136-3144, 2001.
  - [100] Y. Peng, L. Jun, Y. Xueyi, Z. Zhenquan, and L. Bin, "Iris Recognition Algorithm Using Modified Log-Gabor Filters," *18th Int. Conf. Pattern Recognition (ICPR)*, pp. 461-464, 2006.
  - [101] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," *Int. Workshop on Automatic Face and Gesture Recognition*, 1995.
  - [102] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," *IEEE Conf. Automatic Face & Gesture Recognition and Workshops (FG)*, pp. 884-888, 2011.
  - [103] J. Wang, P. Liu, M. F. H. She, A. Kouzani, and S. Nahavandi, "Human action recognition based on Pyramid Histogram of Oriented Gradients," *IEEE Conf. Systems, Man, and Cybernetics (SMC)*, pp. 2449-2454, 2011.
  - [104] S. Xie, S. Shan, X. Chen, and J. Chen, "Fusing Local Patterns of Gabor Magnitude and Phase for Face Recognition," *IEEE Trans. Image Processing*, vol. 19, pp. 1349-1361, 2010.
  - [105] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor Phase Patterns (HGPP): A Novel Object Representation Approach for Face Recognition," *IEEE Trans. Image Processing*, vol. 16, pp. 57-68, 2007.
  - [106] G. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling Latent Discriminative Dynamic of Multi-dimensional Affective Signals," *Affective Computing and Intelligent Interaction*, vol. 6975, pp. 396-406, 2011.
  - [107] J. Kim, H. Rao, and M. Clements, "Investigating the Use of Formant Based Features for Detection of Affective Dimensions in Speech," *Affective Computing and Intelligent Interaction*, vol. 6975, pp. 369-377, 2011.
  - [108] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple Classifier Systems for the Classification of Audio-Visual Emotional States," *Affective Computing and Intelligent Interaction*, vol. 6975, pp. 359-368, 2011.
  - [109] S. Shojaeilangari, Y. W. Yun, and T. E. Khwang, "A Novel Phase Congruency based Descriptor for Dynamic Facial Expression Analysis," *Pattern Recognition Letters*, vol. 49, pp. 55-61, 2014.
-

-
- [110] P. Kovési, "Phase congruency: A low-level image invariant," *Psychological Research*, vol. 64, pp. 136-148, 2000.
  - [111] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High Accuracy Optical Flow Estimation Based on a Theory for Warping," *ECCV*, vol. 3024, pp. 25-36, 2004.
  - [112] X. Xiong and F. De La Torre, "Supervised Descent Method and Its Applications to Face Alignment," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 532-539, 2013.
  - [113] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Int. Joint conf. Artificial Intelligence*, 1981.
  - [114] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185-203, 1981.
  - [115] A. Fahad and T. Morris, "Multiple Combined Constraints for Optical Flow Estimation," *Advances in Visual Computing*, vol. 4842, pp. 11-20, 2007.
  - [116] T. Gautama and M. M. Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Trans. Neural Networks*, vol. 13, pp. 1127-1136, 2002.
  - [117] A. Rabcewicz, "CLG Method for Optical Flow Estimation Based on Gradient Constancy Assumption," *Image Processing Based on Partial Differential Equations*, pp. 57-66, 2007.
  - [118] M. Mohamed, H. Rashwan, B. Mertsching, M. A. Garcia, and D. Puig, "Illumination-Robust Optical Flow Using Local Directional Pattern," *IEEE Trans. Circuits and Systems for Video Technology*, vol. PP, pp. 1-1, 2014.
  - [119] J. L. Barron, D. J. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, pp. 43-77, 1994.
  - [120] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, pp. 77-104, 1990.
  - [121] B. Galvin, B. McCane, K. Novins, and S. Mills, "Recovering motion fields: An evaluation of eight optical flow algorithms," *Proc. British Machine Vision Conf.*, pp. 195-204, 1998.
  - [122] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's distance as a metric for image retrieval," *Int. J. Computer Vision*, vol. 40, pp. 99-121, 2000.
  - [123] J. Bihan, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," *IEEE Int. Conf. Automatic Face & Gesture Recognition and Workshops (FG)*, pp. 314-321, 2011.
  - [124] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, pp. 513-529, 2012.
  - [125] W. Zong, G. B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229-242, 2013.
  - [126] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," *Signals, Systems and Computers*, 1993.
  - [127] S. Chen, D. Donoho, and M. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Review*, vol. 43, pp. 129-159, 2001.
-

-
- [128] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, pp. 265-274, 2009.
  - [129] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. Signal Processing*, vol. 54, pp. 4311-4322, 2006.
  - [130] T. Blumensath and M. E. Davies, "Gradient Pursuits," *IEEE Trans. Signal Processing*, vol. 56, pp. 2370-2382, 2008.
  - [131] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210-227, 2009.
  - [132] T. Guha and R. Ward, K., "Learning Sparse Representations for Human Action Recognition," *IEEE Trans. Pattern Analysis And Machine Intelligence*, vol. 34, pp. 1576-1588, 2012.
  - [133] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *IEEE Trans. Signal Processing*, vol. 86, pp. 572-588, 2006.
  - [134] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Processing*, vol. 86, pp. 589-602,, 2006.
  - [135] J. A. Tropp and A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Trans. Information Theory*, vol. 53, pp. 4655-4666, 2007.
  - [136] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, 2009.
  - [137] T. Blumensath, "Compressed Sensing With Nonlinear Observations and Related Nonlinear Optimization Problems," *IEEE Trans. Information Theory*, vol. 59, pp. 3466-3474, 2013.
  - [138] T. Blumensath, "Accelerated iterative hard thresholding," *Signal Processing*, vol. 92, pp. 752-756, 2012.
  - [139] T. Blumensath, "Compressed Sensing With Nonlinear Observations and Related Nonlinear Optimization Problems," *IEEE Trans. Information Theory*, vol. 59, pp. 3466-3474, 2013.
  - [140] T. Blumensath, M. Yaghoobi, and M. E. Davies, "Iterative Hard Thresholding and L0 Regularisation," *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. III-877-III-880.
  - [141] Mairal. J., Bach. F., and J. Ponce, "Online dictionary learning for sparse coding," *ICML '09 Proc. 26th Annual International Conference on Machine Learning*, 2009.
  - [142] Z. Xie and S. Chen, "SCIHTBB: Sparsity constrained iterative hard thresholding with Barzilai-Borwein step size," *Neurocomputing*, vol. 74, pp. 3663-3676, 2011.
  - [143] E. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for l1-minimization: methodology and convergence," *IAM Journal on Optimization*, vol. 19, pp. 1107-1130, 2008.
  - [144] T. Blumensath and M. E. Davies, "Normalized Iterative Hard Thresholding: Guaranteed Stability and Performance," *Selected Topics in Signal Processing*, vol. 4, pp. 298-309, 2010.

-
- [145] T. Gehrig and H. K. Ekenel, "A common framework for real-time emotion recognition and facial action unit detection," *IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1-6, 2011.
  - [146] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2879-2886, 2012.
  - [147] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210-227, 2009.
  - [148] A. Cruz, B. Bhanu, and S. Yang, "A Psychologically-Inspired Match-Score Fusion Model for Video-Based Facial Expression Recognition," *Affective Computing and Intelligent Interaction*, vol. 6975, pp. 341-350, 2011.
  - [149] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, "Partial least squares regression on grassmannian manifold for emotion recognition," *15th ACM on Int. Conf. Multimodal Interaction*, Sydney, Australia, 2013.
  - [150] M. Day, "Emotion recognition with boosted tree classifiers," *15th ACM on Int. Conf. Multimodal Interaction*, Sydney, Australia, 2013.
  - [151] T. R. Almaev, A. Yuce, A. Ghitulescu, and M. F. Valstar, "Distribution-based iterative pairwise classification of emotions in the wild using LGBP-TOP," *15th ACM on Int. Conf. Multimodal Interaction*, pp. 535-542, 2013.
  - [152] S. E. Kahou et al., "Combining modality specific deep neural networks for emotion recognition in video," *15th ACM on Int. conf. Multimodal Interaction*, Sydney, Australia, pp. 543-550, 2013.
  - [153] T. Gehrig, H. K. Ekenel, "Why is facial expression analysis in the wild challenging?," *Emotion recognition in the wild challenge and workshop*, pp. 9-16, 2013.
  - [154] S. M. H. Avnar, W. Y. Yau, and E. K. Teoh, "Multi-View Face Detection and Registration Requiring Minimal Manual Intervention," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2484-2497, 2013.
  - [155] W. Chung-Hsien, W. Wen-Li, L. Jen-Chun, and L. Wei-Yu, "Speaking Effect Removal on Emotion Recognition From Facial Expressions Based on Eigenface Conversion," *IEEE Trans. Multimedia*, vol. 15, pp. 1732-1744, 2013.
  - [156] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graphics (TOG)*, vol. 31, 2012.