FACIAL AFFECT RECOGNITION: FROM FEATURE ENGINEERING TO DEEP LEARNING

JUNKAI CHEN

Ph.D

The Hong Kong Polytechnic University

2017

The Hong Kong Polytechnic University

Department of Electronic and Information Engineering

Facial Affect Recognition: from Feature Engineering to Deep Learning

Junkai Chen

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

November 2016

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Junkai Chen (陈军凯) (Name of student)

Abstract

Facial expression recognition has been a long standing problem and attracted growing interest from the affective computing community. This thesis presents the research I conducted for facial affect recognition with novel hand-crafted features and deep learning. Three main contributions are reported in this thesis. They include: (1) an effective approach with novel features for facial expression recognition in video; (2) a framework with multiple tasks for detecting and locating pain events in video; and (3) an effective method with a deep convolutional neural network for smile detection in the wild.

In the first investigation, I propose novel features and an application of multi-kernel learning to combine multiple features for facial expression recognition in video. A new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) is proposed to characterize facial appearance changes. A new effective geometric feature is also proposed to capture facial configuration changes. The role of audio modality on affect recognition is also explored. Multiple feature fusion is used to combine different features optimally. Experimental results show that our approach is robust in dealing with video-based facial expression recognition problems under lab-controlled environment and in the wild compared with the other state-of-the-art methods.

In the second investigation, I propose an effective framework with multiple tasks for pain event detection and locating. Histogram of Oriented Gradients (HOG) of fiducial points (P-HOG) and HOG-TOP are used to characterize spatial features and dynamic textures from video frames and video segments. Both frame-level and segment-level detections are based on trained Support Vector Machines (SVMs). Max pooling strategy is further used to obtain the global P-HOG and global HOG-TOP, and an SVM with multiple kernels is trained for pain event detection. Finally, an effective probabilistic fusion method is proposed to integrate the three different tasks (frame, segment and sequence) to locate pain events in video. Experimental results show that the proposed method outperforms other state-of-the-art methods both in pain event detection and pain event locating in video.

In the third investigation, I propose an effective approach for smile detection in the wild with deep learning. Deep learning can effectively combine feature learning and classification into a single model. In this study, a deep convolutional network called Smile-CNN is used to perform feature learning and smile detection simultaneously. I also discuss the discriminative power of the learned features from the Smile-CNN model. By feeding the learned features to train an SVM or AdaBoost classifier, I show that the learned features have impressive discriminative power. Experimental results show that the proposed approach can achieve a promising

performance in smile detection.

List of Publications

Journal papers:

1. **J. Chen**, Z. Chen, Z. Chi, and H. Fu, "Facial Expression Recognition in Video with Multiple Feature Fusion," *IEEE Transactions on Affective Computing* (accepted for publication).

2. **J. Chen**, Z. Chi, and H. Fu, "A New Framework with Multiple Tasks for Detecting and Locating Pain Events in Video," *Computer Vision and Image Understanding* (accepted for publication).

3. **J. Chen**, Q. Ou, Z. Chi, and H. Fu, "Smile Detection in the Wild with Deep Convolutional Neural Networks," *Machine Vision and Applications* (accepted for publication).

Conference Papers:

4. **J. Chen**, Z. Chen, Z. Chi and H. Fu, “Facial Expression Recognition Based on Facial Components Detection and HOG Features”, *International Workshops on Electrical and Computer Engineering Subfields*, pp 884-888, 22-23 August 2014, Istanbul, Turkey.

5. **J. Chen**, Z. Chen, Z. Chi, and H. Fu, "Recognition of Facial Action Units with Action Unit Classifiers and an Association Network," pp 672-683, *in Workshop on Computer Vision for Affective Computing (CV4AC 2014, in conjunction with ACCV 2014)*, Singapore, Nov. 1-5, 2014.

6. **J. Chen**, Z. Chen, Z. Chi, and H. Fu, "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning," in *ACM International Conference on Multimodal Interaction (ICMI 2014)*, pp. 508-513, Nov. 12-16, 2014, Istanbul, Turkey. **(Second runner-up Award in EmotiW2014 Challenge)**

7. **J. Chen**, Z. Chen, Z. Chi, and H. Fu, "Dynamic Texture and Geometric Features for Facial Expression Recognition in Video," *IEEE the International Conference on Image Processing (ICIP 2015)*, pp. 4967-4971, Quebec City, Canada, 27-30 September 2015.

8. **J. Chen**, Z. Chi, and H. Fu, “A New Approach for Pain Event Detection in Video”, *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII2015)*, pp. 250-254, Xi’an, China, 21-24 September 2015.

9. J. Li, **J. Chen** and Z. Chi, “Smile Detection in the Wild with Hierarchical Visual Feature”, *IEEE the International Conference on Image Processing (ICIP 2016)*, pp. 639-643, 25-28 September 2016, Phoenix, USA.

10. **J. Chen**, Z. Chi, and H. Fu, “Facial Expression Recognition with Dynamic Gabor Volume Feature”, *IEEE Workshop on Multimedia Signal Processing (MMSP 2016)*, Paper 64, 21-23 September 2016, Montreal, Canada.

Acknowledgements

I would like to express my gratitude to a lot of people who directly or indirectly contributed to my PhD research and this thesis.

Most of all, I would like to express my most sincere gratitude to Dr Zheru Chi, my supervisor, for his excellent and outstanding guidance throughout my PhD study. It has been a privilege and truly an honor to have him as a mentor. During my PhD study, I benefited a lot from his rich academic experiences and nice personalities.

I also appreciate my colleagues, Dr Hong Fu, Dr Zenghai Chen, Mr. Yu Hu and Mrs. Hui Zhang, for their constructive suggestions and warm encourages.

I would also like to express my thanks to the staff at Department of Electronic and Information Engineering for their patience help and the postgraduate scholarship from The Hong Kong Polytechnic University for its financial support.

Finally, I would like to express my great gratitude and love to my parents and other family members. Without their self-giving and constant support, this thesis would not have been possible.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAM:	Active Appearance Model
BoW:	Bag of Words
BP:	Back-Propagation
CDA:	Contractive Discriminative Analysis
CNN:	Convolutional Neural Network
CCNET:	Contractive Convolutional Network
DBN:	Deep Belief Network
ELM:	Extreme Learning Machine
FACS:	Facial Action Coding System
HOG:	Histogram of Oriented Gradients
HOG-TOP:	Histogram of Oriented Gradients from Three Orthogonal Planes
ICA:	Independent Component Analysis
LBP:	Local Binary Patterns
LBP-TOP:	Local binary patterns from three orthogonal planes
LDN:	Local Directional Number Pattern
LGBP-TOP:	Local Gabor Binary Patterns from Three Orthogonal Planes
LLD:	Low Level Descriptors
MLP:	Multilayer Perceptron

MS-MIL:	Multiple Segments and Multiple Instance Learning
PCA:	Principal Component Analysis
RVR:	Relevance Vector Regression
SIFT:	Scale-Invariant Feature Transform
STLMBP:	Spatial Temporal Local Monogenic Binary Pattern
SVM:	Support Vector Machine
VAS:	Visual Analog Scale

Chapter 1 Introduction

1.1 Motivation

Two channels have been developed for human beings to communicate in social life: auditory channel and visual channel. Auditory channel carries speech and vocal language, visual channel carries facial expressions and body gestures. Facial expressions, as a powerful visual channel, play a vital role for human beings to convey emotions and transit messages. Facial expressions, together with voice, language and body gestures, constitute the principal communication system in social context (Corneanu et al., 2016). Automatic facial affect analysis system aims to interpret and understand human psychological activities by analyzing facial expressions. This technique can be widely used in many fields like security (e.g. lie detection), medicine (e.g. pain monitoring) and human computer interaction (e.g. interactive games) (Zeng et al., 2009). Automatic facial expression analysis has been an active research field in the past two decades with the development of psychology, computer science and artificial intelligence.

There are two main streams in the current research on automatic facial affect analysis: message and sign judgment (Cohn and Ekman, 2005). The aim of message judgment is to infer a kind of emotion behind a displayed facial expression, six

universal facial expressions are widely considered for message judgment: anger, disgust, fear, happiness, sadness and surprise, as shown in Figure 1-1. Sign judgment aims to describe fine-grained facial component and muscle movements. These atomic facial motions are called facial action units (AUs). The Facial Action Coding System (FACS) (Ekman et al., 2002) is the best known and commonly used tools developed to describe facial action units.

Figure 1-1. Six basic facial expressions from the database (Kanade et al., 2000)). 1, disgust; 2, fear; 3, joy; 4, surprise; 5, sadness; 6, anger.

Facial expressions are caused by facial muscle movements which are subtle and transient. To capture and represent these movements is a key issue to be addressed in facial expression analysis. Recent advances in computer vision and machine learning open up the possibility of automatic facial expression recognition. Many efforts have been made to handle this problem. The methodologies used are commonly categorized into appearance based methods and geometry based methods (S. Z. Li and Jain, 2011). Appearance based methods commonly apply feature descriptors to model facial texture changes created by wrinkles, bulges and muscle movements; geometry

based methods generally apply the geometric properties of a face such as the facial landmarks to describe face shape or configuration.

Conventional automatic facial expression analysis techniques require domain expertise to create a feature descriptor which can effectively transform the raw data (such as the pixel values of an image) into an effective representation (Y. LeCun et al., 2015). Recently, deep learning, which aims to discover and automatically learn good representations from raw data with a complex hierarchical model composed of multiple layers, has attracted significant attention. Various deep learning models such as Deep Belief Networks (DBNs) (G. E. Hinton, Osindero, S., Teh, Y. W, 2006), Convolutional Neural Networks (CNNs) (Yann LeCun et al., 1989) and Stacked Auto-Encoders (Bengio et al., 2007) etc. have been developed for many applications including computer vision, natural language processing and automatic speech recognition etc. CNNs are inspired by the visual system's structure, especially by the visual models proposed in (Hubel and Wiesel, 1962). Recent study (Serre et al., 2007) pointed out that the physiology of the visual system is consistent with the processing style found in convolutional neural networks. Due to its superiority of dealing with 2-D images, CNNs have become the first choice for a wide range of computer vision tasks including visual recognition, object detection and image classification.

Automatic facial expression analysis is important and desirable for various

applications. In this thesis, efforts are made to address the following several issues.

(1) An effective and robust facial expression recognition system needs to address two issues: feature extraction and multimodal fusion. Current research shows that designing a kind of robust and effective feature is still important and meaningful. I have made efforts to design robust features to effectively characterize the facial appearance and configuration changes caused by facial muscular activities. Multimodal fusion is another key aspect to the facial expression recognition system. Multimodality can provide more information and improve emotion inference. Nevertheless, how to mine useful representations from different modalities and how to integrate these different representations optimally remain very challenging problems, which need to be carefully addressed. The potentials of audiovisual modalities on facial affect recognition have also been explored and an effective framework which can combine the audiovisual modalities optimally has been developed to perform the facial expression recognition in video.

(2) Except for the six universal facial expressions (anger, disgust, fear, happiness, sadness and surprise), research on pain analysis has also been conducted. Automatically detecting and locating pain events in video is an important task in medical assessment. It is a challenging problem in facial expression analysis

due to spontaneous faces, head movements and pose variations. The role of facial information at various time scales (frame, segment and sequence) are explored and a new framework is proposed to address this issue.

(3) Smile or happiness is one of the most universal facial expressions in our daily life. Smile detection in the wild is an important and challenging problem, which has attracted a growing attention from affective computing community. Deep learning has recently demonstrated outstanding performance in image classification, speech recognition and natural language understanding. An attractive property of deep learning is representation learning. With multiple layers, deep learning methods transform the raw pixels into hierarchical abstract representations. Convolutional Neural Network (CNNs) is proposed for smile detection in this thesis. I also investigate what kinds of learned representations are useful for facial recognition.

1.2 Statements of Originality

This thesis presents my research on facial expression recognition with multiple features and deep learning. The work described in this thesis was carried out at the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, between September 2013 and September 2016, under the supervision of Dr Zheru Chi.

The thesis consists of six chapters. The work described in this thesis was originated by the author except where acknowledged and referenced, or where the results are widely known. The following states the original contributions:

(1) An effective approach for facial expression recognition in video with novel features and multi-kernel learning is the work of the author. In this investigation, a new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) was proposed to extract dynamic textures from video sequences to characterize facial appearance changes. A new effective geometric feature derived from the warp transformation of facial landmarks was proposed to capture facial configuration changes. Moreover, the role of audio modalities on recognition was also explored in my study. The multiple feature fusion with multi-kernel learning was applied to tackle the video-based facial expression recognition problems under lab-controlled environment and in the wild, respectively. More details are presented in Chapter 3.

(2) A new framework for pain event detection and locating in video is the work of the author. In this investigation, the role of facial information at various time scales (frame, segment and sequence) was investigated to address the problem of pain analysis in video. I propose a method fusing spatial feature and spatial-temporal feature for pain event detection and a multiple-task fusion method for locating pain

events, respectively. More details are described in Chapter 4.

(3) An effective approach for smile detection in the wild with a deep convolutional network (CNN) model is the work of the author. In this investigation, a deep convolutional network called Smile-CNN was constructed to perform feature learning and smile detection simultaneously. A study was also carried out to analyze the ability of the Smile-CNN model to tackle the nuisance factors such as pose variations and background. More details are given in Chapter 5.

1.3 Outline of the Thesis

The thesis consists of six chapters. The thesis is outlined as follows.

Chapter 2 introduces basic principles of facial expression recognition, deep learning, and information fusion at different levels. The chapter also reviews some recent important developments of video based facial expression recognition, pain recognition in video and smile detection in the wild.

Chapter 3 discusses our approach for facial expression recognition in video with novel features and multiple feature fusion. The potentials of visual modalities (face images) and audio modalities (voice) are explored. In addressing visual modalities, a new feature descriptor called HOG-TOP is proposed to characterize facial appearance changes. Moreover, an effective geometric warp feature derived from the warp transformation of facial landmarks is proposed to characterize facial configuration

changes. The role of audio modalities for affect recognition is also explored. A multiple feature fusion method with multi-kernel learning is further employed to deal with facial expression recognition under lab-controlled environment and in the wild, respectively. Experimental results conducted on several public databases are also reported in this chapter.

In Chapter 4, my work for pain event detection and locating in video is presented. A new framework with multiple tasks is proposed to tackle this problem. Considering that information with various time scales (frame, segment and sequence) can make different contributions, I propose to combine three different tasks, that is, frame, segment and sequence detection, to effectively detect and locate pain events in video. A method which combined spatial feature and spatial-temporal feature for pain event detection and a multiple-task fusion method for locating pain events are introduced, respectively. Experimental results conducted on a public shoulder pain database are reported in this chapter.

Chapter 5 presents my work on smile detection in the wild with CNN. A deep convolutional network model called Smile-CNN is proposed to perform feature learning and smile detection simultaneously. I further explore the discriminative power of the learned features which are taken from the neuron activations of the last hidden layer of the Smile-CNN model. Experimental results conducted on a public

“smile in the wild” database are reported in this chapter.

Chapter 6 concludes the research work presented in this thesis and discusses some potential directions for future research.

Chapter 2 Literature Review

In this chapter, basic principles of facial expression recognition and deep learning are introduced. Some recent important developments of video based facial expression recognition, pain recognition in video and smile detection in the wild are also reviewed.

2.1 Facial Expression Recognition in Video

Faces are powerful nonverbal tools for human beings to transmit message and communicate with each other (Cowie et al., 2001). Facial expressions provide valuable information and important cues to understand the emotions and intentions of human beings. Facial expression recognition has been an active research field for more than two decades. We have witnessed much progress that has been made for addressing this problem.

Previous works mainly focused on static and single face image based facial expression recognition (Liang et al., 2005; M. Lyons et al., 1998; M. J. Lyons et al., 1999; Shan et al., 2005; Wong and Cho, 2009). In those methods, Gabor filters (Feichtinger and Strohmer, 1998), Local Binary Patterns (LBP) (Ojala et al., 2002), Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and Scale-Invariant Feature Transform (SIFT) are commonly applied to extract

appearance features from face images. Recently, facial expression recognition in video has attracted great interest. Compared with a static image, a video sequence can provide not only spatial appearance information but also facial motions and accompanied speech. The methods of video based facial expression recognition can be categorized into static image based methods, dynamic texture based methods and audiovisual based methods. Figure 2-1 illustrates the differences of the three methods. I will review these methods in detail.

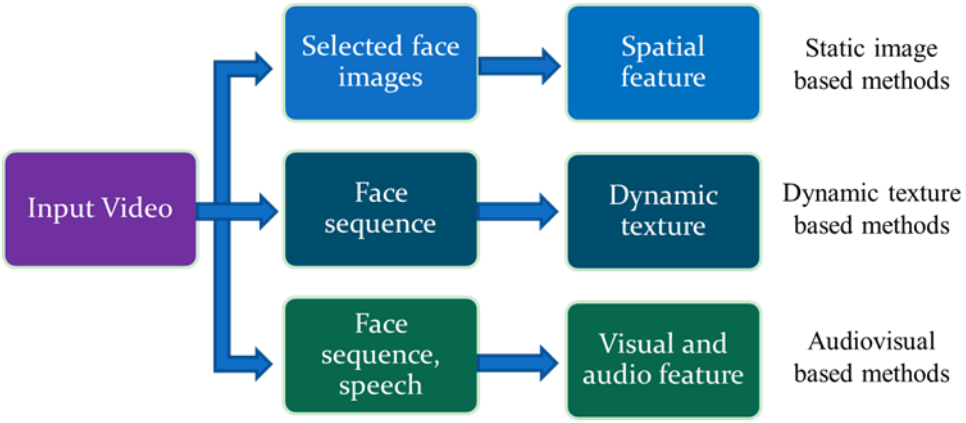


Figure 2-1. The diagram of the three kinds of methods.

2.1.1 Static Image Based Methods

Many researchers applied static image based models to handle the problem of video based facial expression recognition. In general, one or several peak face frames are first selected and feature descriptors are applied to extract geometric or appearance feature representations from these selected face images. For instance, the

methods reported in (Chew et al., 2011; Lucey et al., 2010; Taheri et al., 2011) applied the facial landmarks to characterize the whole face shape. And the method in (Saeed et al., 2012) measured the displacements of several selected candidate fiducial points as geometric features. Bag of Words (BoW) based on multi-scale dense SIFT features were applied to represent facial appearance textures in (Karan Sikka et al., 2012). A novel local feature descriptor called Local Directional Number pattern (LDN) was proposed to extract appearance features in (Ramirez Rivera et al., 2013). However, automatically distinguishing key frames from a video sequence is usually difficult. Some methods (Dahmane and Meunier, 2011; Dhall et al., 2011; Valstar et al., 2011) attempted to classify each individual frame first and adopted a voting scheme to label the video sequence. In these methods, it is necessary to extract features from each frame and to classify each frame, which is time consuming.

2.1.2 Dynamic Texture Based Methods

There exists a drawback for static image based methods: extracting feature from individual frame fails to consider spatial temporal information among frames, which is useful to describe facial muscle motions. Dynamic texture based methods were proposed to effectively tackle this problem. Dynamic texture based methods aim to simultaneously model the spatial appearance and dynamic motions in a video sequence. Local Binary Patterns from Three Orthogonal Planes (LBP-TOP), a

temporal extension of local binary patterns, proposed by Zhao et al. (Zhao and Pietikainen, 2007), has been widely used for video based facial expression recognition. Following LBP-TOP, a facial component LBP-TOP was proposed in (X. Huang et al., 2011). A Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) was proposed in (Senechal et al., 2011). A Spatial Temporal Local Monogenic Binary Pattern (STLMBP) feature descriptor was proposed in (X. Huang et al., 2014; X. Huang et al., 2012). In addition, Long et al. (Long et al., 2012) employed Independent Component Analysis (ICA) to learn spatiotemporal filters from videos, and then extracted dynamic textures using the learned filters. The method in (Chew et al., 2012) employed a sparse temporal representation to model the temporal dynamics of facial expressions in video. Li et al. (Y. Li et al., 2013) developed a dynamic Bayesian network to simultaneously and coherently represent the facial evolvement at different levels.

2.1.3 Audiovisual Based Methods

Static image based methods and dynamic texture based methods only rely on visual modalities. However, audio or speech also plays an important role for human beings to convey emotions and intentions. Audio modalities can provide complementary information to visual modalities. Recently, audiovisual based methods for affect recognition have attracted growing attention from the affective computing

community. A number of approaches have been proposed to combine audio and visual modalities for affect recognition (Chen et al., 2014; Kanou et al., 2013; Ringeval et al., 2014; Karan Sikka et al., 2013; S. Zhang et al., 2012). Audiovisual based methods generally couple acoustic features extracted from voice with visual features extracted from face images to tackle the problem. The method in (Ringeval et al., 2014) incorporated voice and lip activities to perform emotion recognition in video. Kim et al. (Kim et al., 2013) built a two layer deep belief network (DBN) to learn the feature representations from audiovisual modalities and the learned features were further used to perform emotion recognition. The methods in (Karan Sikka et al., 2013; Sun et al., 2014) extracted several different appearance features like HOG and SIFT etc. and these extracted features were further coupled with acoustic features to recognize emotions in the wild.

2.1.4 Motivation

The studies surveyed above demonstrate that feature extraction plays a central role on facial expression recognition in video. Designing robust and effective features is important and meaningful. LBP-TOP is widely used for modeling dynamic textures. However, there are two limitations of LBP-TOP. One is high dimensionality. The size of LBP-TOP coded in each block using a uniform pattern is 59×3. A video sequence is generally divided into many blocks, which will generate a very high dimensional

feature vector. Moreover, although LBP-TOP is robust to deal with illumination changes, it is insensitive to facial muscle deformations. It is necessary to design novel features which are more robust to characterize facial appearance changes with a more compact representation. In addition, configural and shape representations play an important role in human vision for the perception of facial expressions (Martinez and Du, 2012). Compared with appearance features, only small efforts have been made for facial expression recognition with geometric features. I believe that previous works have not yet fully exploited the potentials of configuration representations. Characterizing face shapes with facial landmarks (Chew et al., 2011; Lucey et al., 2010) or measuring displacements of fiducial points (Chen et al., 2015; Saeed et al., 2012) only are not sufficient to capture facial configuration changes, especially the subtle non-rigid changes. It is necessary to design novel geometric features to effectively capture locally subtle shape changes or deformations.

In this thesis (Chapter 3), I propose a novel feature called HOG-TOP, which is more compact and effective to characterize facial appearance changes, and introduce a more robust geometric feature to capture facial configuration changes.

2.2 Pain Analysis in Video

2.2.1 Background

Pain monitoring and measurement is an important task in medical assessment. Pain diagnosis can be used to identify many surgical diseases, like shoulder frozen, arthritis and ligament injury etc. (Lucey et al., 2012). A great challenge is how to effectively assess and measure the pain, since pain is a kind of subjective feeling. A widely used technique to evaluate pain is patient's self-reporting, which is convenient and easy to operate. It is typically measured by either through a clinical interview or by using a Visual Analog Scale (VAS) (Lynch et al., 2011). With VAS, a patient is asked to mark his pain on a linear scale with a range from 0 ("no pain") to 10 ("the worst pain"). It has become a very popular method due to its simplicity in implementation. However, this method has several limitations such as idiosyncratic use, subjective variations etc. (Williams et al., 2000). Therefore, it is not available for some important populations like young children, people who are deaf-and-dumb, and patients who require assisted breathing. Some researchers attempted to acquire a continuous measure of pain through analyzing tissue pathology, neurological "signatures" and so on (Turk and Melzack, 2001). These efforts are difficult because they are often inconsistent with other evidence of pain (Turk and Melzack, 2001). It is,

therefore, necessary to find a reliable and effective method for pain analysis.

A potential solution for pain detection is to analyze facial expression. Facial expression is a powerful nonverbal way for human beings to transmit messages and reveal emotions. Pain as a kind of emotion or affection, can be displayed by facial expression. With the advancement of techniques for facial expression recognition, recently, automatic pain detection through analyzing facial expressions has become an evolving research area and has attracted a growing interest from affective computing community.

2.2.2 Current Research for Pain Analysis in Video

A significant contribution to the research on pain analysis was the introduction of the UNBC-McMaster shoulder pain dataset (Lucey, Cohn, Matthews, et al., 2011), which recorded videos of faces of adult subjects with shoulder injuries. All the videos in the dataset were provided with two levels of annotation for measuring pain: frame level and sequence level. For the frame annotation, it followed the description proposed in (Prkachin and Solomon, 2008) which defined pain as the sum of the intensities of certain facial action units including brow lowering, orbital tightening and eye closure and then employed Facial Action Coding System (FACS) (Ekman et al., 2002; Essa and Pentland, 1997) to code each frame. Figure 2-2 shows some image sequences selected from this database.

Figure 2-2. The image sequences from the UNBC database.

Pain detection can be regarded as a spontaneous facial expression recognition problem. An early research on automatic pain recognition was done by Ashraf et al. by developing a "pain-no pain" detection system (Ashraf et al., 2009). In their framework, Active Appearance Models (AAMs) were used to extract shape and appearance features from face images. An SVM was trained for classification. Both frame-level detection and sequence-level detection results were reported in their study. After that, different approaches have been proposed to address this problem. Lucey et al. (Lucey et al., 2008) pointed out that temporal information played a vital role in pain recognition and experiments showed that by compressing the spatial signal instead of the temporal signal, a better pain recognition performance could be achieved. In addition, some researchers considered utilizing the relationship between facial expressions and facial action units defined in Facial Action Coding System (FACS) to recognize pain. Pain could be defined as a combination of several action units (Prkachin, 1992). The works of automatically recognizing pain in video via

facial action units were reported in (Lucey, Cohn, Matthews, et al., 2011; Lucey et al., 2012; Lucey, Cohn, Prkachin, et al., 2011). Hammal et al. (Hammal and Cohn, 2012) proposed a method based on Log-Normal filters and SVMs for four-level pain intensity estimation. Sebastian et al. also worked on pain intensity estimation (Kaltwang et al., 2012) in which they extracted shape and appearance features from face images and Relevance Vector Regression (RVR) was trained to predict pain intensity levels.

2.2.3 Motivation

The works investigated above mostly focused on frame-level detection, i.e. detecting the pain occurrence or intensity in each video frame. Some methods also provided sequence-level detection results (Ashraf et al., 2009; Lucey et al., 2012). In these methods, a video sequence under test is predicted as pain if the average score of its member frames exceeds a threshold. In other words, these methods are still based on a frame level detection mechanism. When a video sequence is long, it is computationally very time consuming. A better way is to apply a global feature vector to characterize the whole video sequence. In addition, performing the sequence-level detection can reduce the number of training instances and therefore the training would be more efficient. The works reported in (K. Sikka et al., 2014; Wang et al., 2012) attempted to recognize pain at sequence level. In (Wang et al., 2012), a video

sequence was first transformed into a feature vector of fixed length by applying Bag-of-Words (BoW), and a SVM was trained to classify the sequence level feature. In (K. Sikka et al., 2014), a method called Multiple Segments and Multiple Instance Learning (MS-MIL) was proposed to jointly detect and locate pain events in video.

Most previous works surveyed above only used spatial information and ignored spatial temporal information. I believe that information at various time scales (frame, segment and sequence) plays different roles. All these information can provide complementary cues. So a new framework which combines three different tasks (frame-level, segment-level and sequence-level detection) has been developed to effectively detect and locate pain events in video. In this study (Chapter 4), experimental results show that multiple features with kernel fusion is efficient for pain event detection and combining three tasks (frame-level, segment-level and sequence-level detection) for locating pain events is more robust than carrying out any individual task alone.

2.3 Smile Detection in the Wild

Smile or happiness is one of the most universal facial expressions. Although some previous works which handled a facial expression recognition problem also included happiness recognition, they mainly focused on facial expression recognition under lab-controlled environment, in which people depicted in these images exhibit in

a nearly front view with clean background and similar scale. The facial expressions displayed in the wild are more natural than those demonstrated under lab-controlled environment. Figure 2-3 shows some smile face images under lab-controlled environment and in the wild, respectively. The face images shown in the top row and middle row are from CK+ (Lucey et al., 2010) and JAFFE (M. Lyons et al., 1998) databases, respectively. And the face images shown in the bottom row are from the GENKI4K (Whitehill et al., 2009) database. It can be seen that smile faces under lab-controlled environment are in the front view with clean backgrounds and in similar scale. On the other hand, smile faces in the wild have various poses, illuminations and scales.

Figure 2-3. Smile face images under lab-controlled (top and middle rows) and in the wild (bottom row).

Emotions or intentions revealed by facial expressions in the wild are closer to the real inner psychological activities of human beings. Facial expression analysis in the

wild is important and meaningful. However, various limitations such as imprecise face detection and alignment, variation of illumination, pose changes and complex background increase the difficulty of facial expression recognition in the wild. It remains a significant research challenge. To tackle this problem, a reliable and extensive database is necessary. Jacob et al. (Whitehill et al., 2009) collected pictures which were photographed in real word and built the GENKI4K database. With this database, they explored necessary characteristics of the training set, image registration, feature representation, and machine learning algorithms for smile detection in the wild. After that, Shan (Shan, 2012) proposed to use pixel intensity differences to extract features and AdaBoost was trained to perform smile detection in the wild. Liu et al. (M. Liu et al., 2012) argued that unlabeled reference data could enhance the performance of facial expression recognition in the wild, they further combined labeled data and unlabeled reference data to deal with smile detection in the wild. Jain et al. (Jain et al., 2014) combined Gaussian derivatives with LBP to provide a robust descriptor which could effectively extract texture patterns from facial images and they applied this feature descriptor for smile detection. In (An et al., 2015), Le et al. first employed LBP and HOG to extract appearance features from face images, and Principal Component Analysis (PCA) was applied to reduce the dimensionality of the features. Finally, Extreme Learning Machine (ELM) was trained to perform the

classification.

From the research works mentioned above, I observe that hand-crafted features coupled with a supervised learning method are widely used. Some works (Devries et al., 2014; Ijjina and Mohan, 2014; Lawrence et al., 1997; Matsugu et al., 2003) have highlighted the effectiveness of deep learning methods for facial expression recognition. A rule-based algorithm for robust facial expression recognition combined with robust face detection using a convolutional neural network was proposed in (Matsugu et al., 2003). It tried to address the problem of subject independence as well as translation, rotation, and scale invariance in the recognition of facial expression. In (Devries et al., 2014), they introduced a multi-task convolutional neural network that simultaneously predicted facial landmarks and facial expression. Earnest et al. (Ijjina and Mohan, 2014) proposed an approach for facial expression recognition using deep convolutional neural networks (CNNs) based on features generated from depth information only. Glauner (Glauner, 2015) applied different CNNs to both the entire face and mouth region for smile recognition on the DISFA (Mavadati et al., 2013) database. Glauner's work focused on smile detection under lab-controlled environment. Zhang et al. (K. Zhang et al., 2015) proposed a CNN that used both recognition and verification signals as supervision to learn expression features for smile detection in the wild. They applied CNN to perform the classification directly

and did not investigate the representation learning of CNN. In their study, there was also no exploration on how the nuisance factors like pose variations, background and scales would affect the recognition ability of the CNN.

2.4 Deep Neural Networks and Deep Learning

Neural networks are inspired by biological neural networks, especially the central nervous systems of the brain. Neural networks provide an efficient way to estimate or approximate complex non-linear functions. Feed forward network is one of the most widely used neural networks. It generally consists of many layers, including an input layer, one or more hidden layers and an output layer. Figure 2-4 illustrates a typical feed forward network with one hidden layer. It transforms D inputs to K outputs. The network parameters W can be learned by using the technique of *error propagation* with gradient descent.

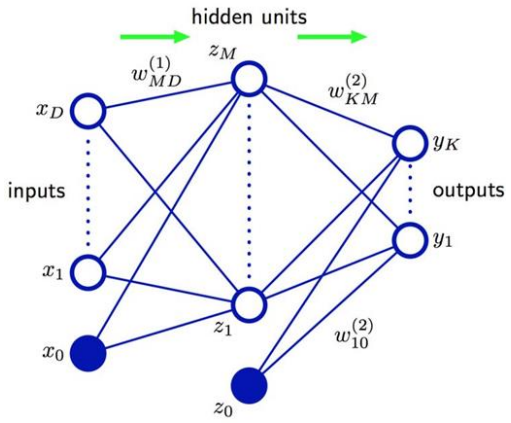


Figure 2-4. A feedforward network with one input layer, one hidden layer and one output layer (Bishop, 2007).

The neural networks with one or two hidden layers are called *shallow* networks. On the other hand, the *deep* networks often include more hidden layers. Deep networks can represent a function more compactly than shallow networks (Bengio, 2009). Depth plays a key role for feature learning or representation learning.

Conventional computer vision and machine learning systems often consist of three components: preprocessing, feature extraction and classification. Feature extraction often plays a central role. In order to design robust and effective features, careful engineering and considerable domain expertise are needed to transform raw data into high-level representations, which amplifies aspects of raw data that are important for discrimination and suppress irrelevant variations (Y. LeCun et al., 2015).

Different from hand designed features, feature learning or representation learning aims at training models with raw data to make the models automatically discover the representations which are useful for classification or prediction (Bengio et al., 2013). Deep learning methods are representation learning methods with multiple levels of representations, which are obtained by constituting non-linear modules layer by layer.

Deep learning methods have attracted growing interest since 2006. A group of researchers (Bengio et al., 2007; G. E. Hinton, Osindero, S., Teh, Y. W, 2006; Poultney et al., 2007) introduced an efficient training algorithm namely *greedy layer*

wise unsupervised pre-training to learn a hierarchical architecture with multiple layers. The core idea of this training algorithm is that each layer is pre-trained with an unsupervised learning algorithm, one layer after the other; after having thus initialized a number of layers, the whole neural network can be fine-tuned with respect to a supervised training criterion (Bengio, 2009). Several deep learning models like Deep Belief Network (DBN) (G. E. Hinton and Salakhutdinov, 2006) and Deep Boltzmann Machine (DBM) (Salakhutdinov and Hinton, 2009) applied this algorithm to train deep networks and achieved promising performance for prediction or generalization.

Although training is a big challenging issue for most deep networks, there is one particular type of deep network that is possible to train: Convolutional Neural Network (CNN). CNNs are inspired by the visual system's structure, especially by the visual models proposed in (Hubel and Wiesel, 1962). Fukushima (Fukushima, 1980) first proposed computational models based on local connectivity between neurons and hierarchical layers. LeCun et al. further developed CNN for pattern recognition tasks and achieved promising performance (Yann LeCun et al., 1989; Yann LeCun et al., 1998). Recent study (Serre et al., 2007) pointed out that the physiology of the visual system is consistent with the processing style found in CNNs. Due to its superiority of dealing with 2-D images, CNNs have become the first choice for a wide range of computer vision tasks including visual recognition (Kavukcuoglu et al., 2010), object

detection (Eslami and Williams, 2012) and image classification (Krizhevsky et al., 2012).

Some research works on using CNNs for facial expression recognition have also been reported (M. Liu et al., 2013; Tang, 2013). CNNs can be used for feature learning. It can be first trained using a supervised/unsupervised method to learn abstract representations from raw data and the abstract representations (generally corresponding to the higher level hidden latent variables) are treated as learned features and fed to train other classifiers like a SVM (Bengio, 2009). In (Rifai et al., 2012), consecutively designed a multi-scale Contractive Convolutional Network (CCNET) and a Contractive Discriminative Analysis (CDA) to learn the features which are robust to handle the illumination and pose changes for emotion recognition. Kim et al. (Kim et al., 2013) utilized DBNs to learn audio-visual features for emotion classification. The experimental results demonstrated that DBNs could generate representative features effectively. Liu et al. (P. Liu et al., 2014) proposed a novel Boosted Deep Belief Network (BDBN) for performing feature learning, feature selection and classifying in a unified loopy framework. Samira et al. (Ebrahimi Kahou et al., 2015) employed a hybrid architecture which combined a CNN with a recurrent neural network (RNN) for emotion recognition in video. With extra training data, Yu and Zhang (Yu and Zhang, 2015) built a learning model with the ensemble multiple

CNNs for static image based facial expression recognition. Ng et al. (Ng et al., 2015) applied a pre-trained CNN and transfer learning for facial expression recognition on small databases. Joy et al. proposed to fuse deep learned and hand-crafted features for automatic pain estimation (Egede et al., 2017).

2.5 Fusion Methods

Fusion methods are widely used and play a vital role in many computer vision applications. In general, fusion can be performed at four different levels: sensor level, feature level, matching score level and decision level (Nandakumar et al., 2008). For sensor level fusion, raw data obtained from different sensors are combined to produce a new raw data, which is more informative than the inputs. For feature level fusion, the raw data are transformed to representative features by applying hand-crafted feature descriptors or pre-trained deep neural networks, and these extracted features are further fused to generate a new feature vector. Feature fusion aims to remove redundant or irrelevant information and to improve the discriminative ability of multiple features. Feature combination or weighted combination is a simple type of feature fusion (Mangai et al., 2010). For matching score level fusion, multiple modalities or instances are first compared to templates to compute the similarity scores and the scores are integrated to generate a single fused score (Hicklin et al., 2006). Three score fusion techniques are widely used: transformation-based score

level fusion, classifier-based score level fusion and density-based score level fusion (He et al., 2010; Nandakumar et al., 2008). For decision level fusion, the predicting results of a set of classifiers are transformed to a single final output, which is expected to outperform each individual classifier.

In general, feature fusion can be performed before the training process; score fusion can be used after matching; and decision fusion can be employed after classification. Fusion methods attempt to explore different features, matching scores or decision values to achieve a better overall performance. In order to further enhance the recognition ability of an automatic facial expression recognition system, different fusion methods can be explored. For example, Zhang et al. (Y. Zhang and Ji, 2005) applied multisensory information fusion technique and dynamic Bayesian networks (DBN) to model and understand the temporal behaviors of facial expressions in image sequences. Thiago et al. (Zavaschi et al., 2013) proposed a novel method which employed a combination of two different feature sets in an ensemble approach. The feature fusion and decision fusion were both used in their method. Turan et al. (Turan and Lam, 2014) proposed a feature fusion method based on Canonical Correlation Analysis (CCA) for facial expression recognition. Liu et al. (W. Liu and Wang, 2006) proposed a method for facial expression recognition based on the fusion of multiple Gabor feature sets. In (Sun et al., 2014), Sun et al. explored the feature fusion with

multiple kernel fusion (A proof of multiple kernel fusion with linear kernels being equivalent to a weighted feature fusion is given in Section 3.2.4) and the classifier fusion with a hierarchical fusion strategy. In (Z. Liu et al., 2013), Liu et al. proposed the use of a new image representation and multiple feature fusion to handle the problem of facial expression recognition. They also demonstrated that combining the classification results of multiple features at score level could further improve recognition performance.

Chapter 3 Facial Expression Recognition in Video with Multiple Feature Fusion

3.1 Introduction

In this chapter, the proposed effective framework based on multiple feature fusion for facial expression recognition in video is presented. In this study, both the potentials of visual modalities (face images) and audio modalities (speech) are explored. In addressing visual modalities, I extend the Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) to temporal Three Orthogonal Planes (TOP), inspired by a temporal extension of local binary patterns, LBP-TOP (Zhao and Pietikainen, 2007). The proposed HOG-TOP is used to characterize facial appearance changes. Experimental results show that HOG-TOP performs as well as LBP-TOP for facial expression recognition. In addition, compared with LBP-TOP, HOG-TOP is more compact and computationally efficient. Moreover, an effective geometric warp feature derived from the warp transformation of facial landmarks is proposed to characterize facial configuration changes. The proposed geometric warp feature is more effective compared with other previously proposed geometric features (Chen et al., 2015; Chew et al., 2011; Lucey et al., 2010). The role of audio modalities on affect recognition is also explored in this work. The audio modalities can provide

some complementary information, especially for facial expression recognition in the wild. Finally, a multiple-feature fusion method is developed to deal with facial expression recognition under lab-controlled environment and in the wild, respectively. The diagram of our proposed framework is shown in Figure 3-1. Geometric features coupled with dynamic textures (HOG-TOP) are used to deal with lab-controlled facial expression recognition. Acoustic features and dynamic textures (HOG-TOP) are fused to tackle facial expression recognition in the wild.

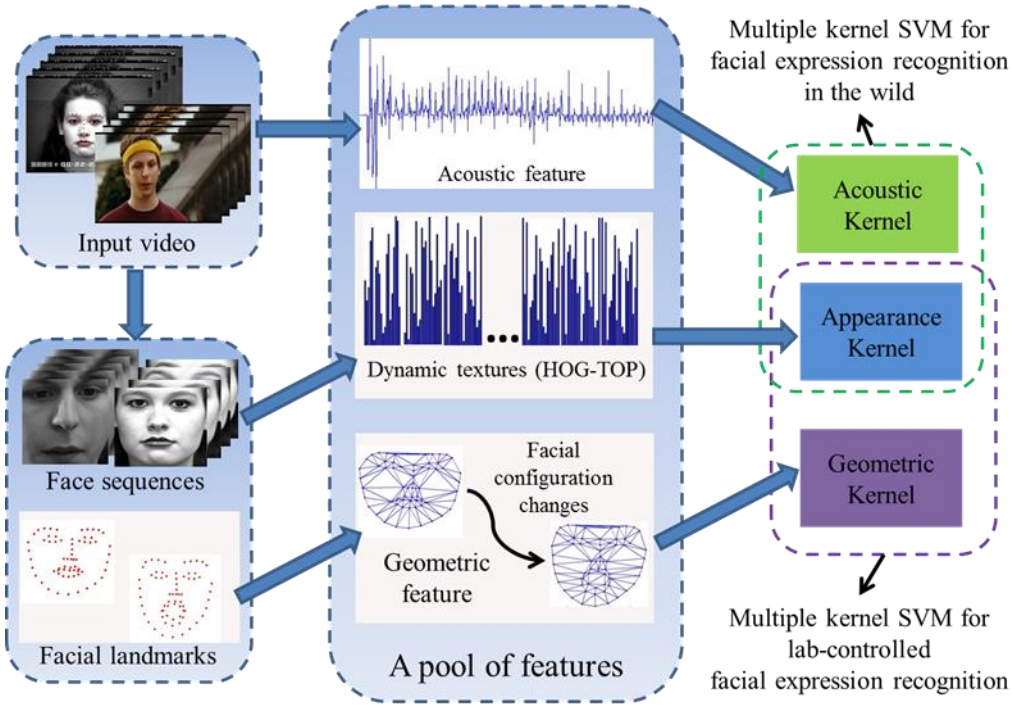


Figure 3-1. Block diagram of our proposed framework. Geometric features coupled with dynamic textures HOG-TOP are used to deal with lab-controlled facial expression recognition; acoustic features and dynamic textures HOG-TOP are fused to tackle facial expression recognition in the wild.

3.2 Methodology

3.2.1 Histograms of Oriented Gradients from Three Orthogonal Planes

Histograms of oriented gradients (HOG) were first proposed for human detection. The basic idea of HOG is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions (Dalal and Triggs, 2005). HOG is sensitive to object deformations. Facial expressions are caused by facial muscle movements. For example, mouth opening and raised eyebrows will generate a “surprise” facial expression. These movements could be regarded as types of deformations. HOG can effectively capture and represent these deformations (Orrite et al., 2009). However, the original HOG is limited to deal with a static image. In order to model dynamic textures from a video sequence with HOG, I extend HOG to 3-D to compute the oriented gradients on Three Orthogonal Planes XY, XT, and YT (TOP), i.e. HOG-TOP. The proposed HOG-TOP is able to characterize facial appearance changes and facial muscular motions.

Figure 3-2. The textures in XY, XT and YT planes.

A video sequence includes three orthogonal directions, i.e. X, Y, and T (time) directions. Figure 3-2 illustrates the textures extracted from the three orthogonal planes. The X-Y plane provides spatial appearance, and X-T and Y-T planes record temporal or motion information along the time. In this study, I propose to compute the distributions of oriented gradients of each plane to obtain HOG features, namely HOG-XY, HOG-XT and HOG-YT, as shown in Figure 3-3.

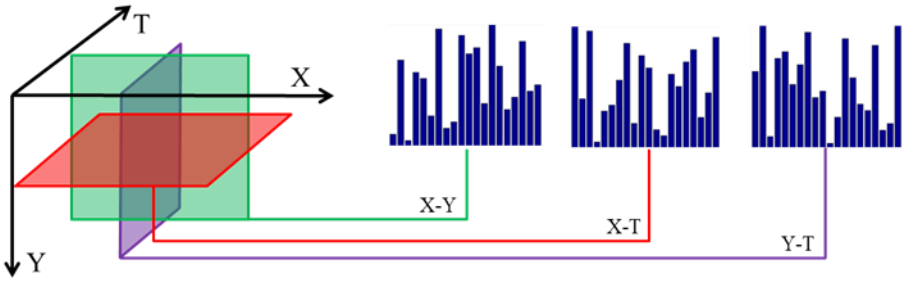


Figure 3-3. The HOG from Three Orthogonal Planes (TOPs).

Each point in a video sequence includes three orthogonal neighborhoods lying on X-Y, X-T and Y-T planes, respectively. The gradients along X, Y and T directions are first computed with a 3×3 Sobel mask. The gradient orientations are further defined as $\theta_{XY} = \tan^{-1}(G_Y/G_X)$, $\theta_{XT} = \tan^{-1}(G_T/G_X)$, $\theta_{YT} = \tan^{-1}(G_T/G_Y)$, where G_X, G_Y,

and G_T are the gradients along the X, Y and T directions, respectively. These angles are further quantized into K bins in a range of $0° - 180°$ or $0° - 360°$.

Enumerating the appearance of these gradient orientations can obtain a histogram in each plane. The three histograms are concatenated to form a global description with the spatial and temporal features. Figure 3-3 shows that the three histograms from the three planes are combined into a single one. The HOG-TOP computation algorithm is shown in Algorithm 1.

Algorithm 1: HOG-TOP

Input:

Video sequence V, which contains N frames with the same width and height.

Output:

The histograms of oriented gradients from three orthogonal plans (HOG-TOP).

Algorithm:

Get the number of frames N, frame width and height.

for t=2:N-1

for x=2:width-1

for y=2:height-1

get the local patch in X-Y, X-T, and Y-T planes.

$\boldsymbol{P_{xy}}$=V(x-1:x+1, y-1:y+1, t);

$\boldsymbol{P_{xt}}$=V(x-1:x+1, y, t-1:t+1);

$\boldsymbol{P_{yt}}$=V(x, y-1:y+1, t-1:t+1);

Compute the gradients $\boldsymbol{G_X}$, $\boldsymbol{G_Y}$ and $\boldsymbol{G_T}$

Compute gradient orientations $\boldsymbol{\theta_{XY}, \theta_{XT}, \theta_{YT}}$

Quantize the orientations $\boldsymbol{\theta_{XY}, \theta_{XT}, \theta_{YT}}$ into one of 9 bins for each plane.

Count the appearance of these quantized orientations and obtain a histogram in each plan, i.e. HOG-XY, HOG-XT and HOG-YT.

end

end

end

Normalize the HOG-XY, HOG-XT and HOG-YT respectively. Concatenate the three histograms into a long histogram.

LBP-TOP computes the difference of a pixel with respect to its neighborhood, making LBP-TOP robust in dealing with illumination changes. HOG-TOP computes the oriented gradients of a pixel, which is more effective to capture object deformations. Facial expressions are caused by facial muscle movements, which can be regarded as types of muscle deformations. HOG-TOP is therefore more effective to characterize facial appearance changes than LBP-TOP. Another advantage of HOG-TOP is the feature dimensionality. Compared with LBP-TOP, the size of HOG-TOP is much smaller than that of LBP-TOP. The size of LBP-TOP coded using a uniform pattern is 59×3, while the size of HOG-TOP quantized into 9 bins is 9×3, which is much more compact than that of LBP-TOP.

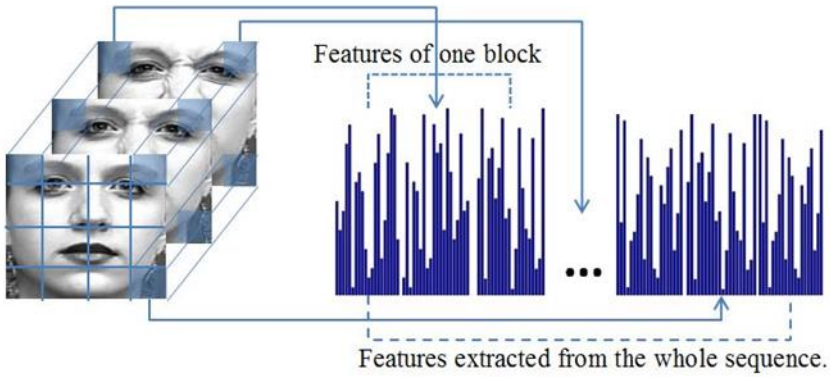


Figure 3-4. The HOG-TOP features extracted from each block are concatenated to represent the whole sequence.

Moreover, a block-based method is also introduced in this study, as shown in Figure 3-4. The image sequence can be divided into many blocks and HOG-TOP

features are extracted from each block. The HOG-TOP features of all the blocks can be concatenated to represent the whole sequence. In our experiments, the face image is first cropped from the original image and resized to 128×128. The face image is partitioned into 8×8 blocks with each block having a size of 16×16. The number of bins is set to 9 with an angle range of $0° - 180°$.

3.2.2 Geometric Warp Feature

In this section, a more robust geometric feature namely geometric warp feature, which is derived from the warp transform of the facial landmarks, is introduced. Facial expressions are caused by facial muscle movements. These movements result in the displacements of the facial landmarks. Suppose that each face image consists of a number of sub-regions. These sub-regions are triangles with their vertexes located at facial landmarks, as shown in Figure 3-5. The displacements of facial landmarks cause the deformations of these triangles. I propose to utilize the deformations to represent facial configuration changes.

Facial expression can be considered as a dynamic process including onset, peak and offset. There exist the displacements of the corresponding facial landmarks between onset (neutral face) and peak (expressive face). Given a set of facial landmarks $\mathrm{s} = (x_1, y_1, x_2, y_2, \ldots x_n, y_n)$, where (x_i, y_i) denote the coordinates of the i-th facial landmark. These facial landmarks make up the mesh of a face, as

shown in Figure 3-5.

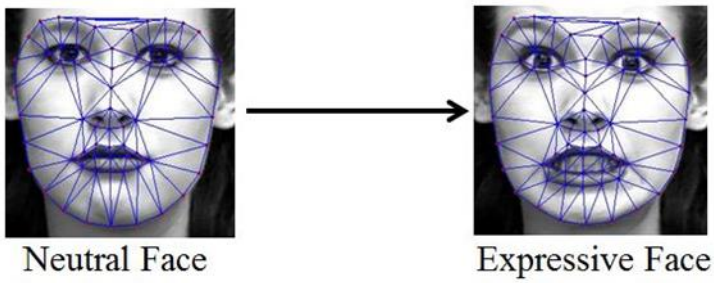


Figure 3-5. Facial landmarks characterizing the shape of a face.

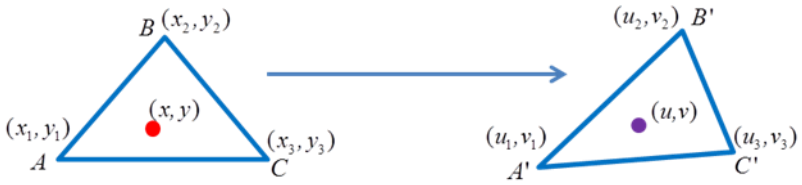


Figure 3-6. A pixel (x, y) lying in a triangle ΔABC of the neutral face is transformed to another pixel (u, v) lying in a triangle $\Delta A'B'C'$ of the expressive face.

As illustrated in Figure 3-5, there are many small triangles covering the face, and each triangle is determined by three facial landmarks. Facial muscle movements cause the deformations of the triangles when a neutral face transforms to an expressive face.

Suppose that a pixel (x, y) which lies in a triangle ΔABC belonging to the neutral face, and the corresponding pixel (u, v) lies in a triangle $\Delta A'B'C'$ on the expressive face, as shown in Figure 3-6. From (Matthews and Baker, 2004), the pixel (x, y) can be expressed with a linear combination of the three vertexes.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \lambda_1 \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix} + \lambda_2 \begin{bmatrix} x_3 - x_1 \\ y_3 - y_1 \end{bmatrix} \quad (3\text{-}1)$$

And the coefficients λ_1, λ_2 can be obtained as

$$\lambda_1 = \frac{(x-x_1)(y_3-y_1)-(y-y_1)(x_3-x_1)}{(x_2-x_1)(y_3-y_1)-(y_2-y_1)(x_3-x_1)} \quad (3\text{-}2)$$

$$\lambda_2 = \frac{(x_2-x_1)(y-y_1)-(y_2-y_1)(x-x_1)}{(x_2-x_1)(y_3-y_1)-(y_2-y_1)(x_3-x_1)} \quad (3\text{-}3)$$

The point (u, v) in the triangle of the expressive face can be defined with the three vertexs and λ_1, λ_2

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \lambda_1 \begin{bmatrix} u_2 - u_1 \\ v_2 - v_1 \end{bmatrix} + \lambda_2 \begin{bmatrix} u_3 - u_1 \\ v_3 - v_1 \end{bmatrix} \quad (3\text{-}4)$$

Combining Eq. (3-2) with Eq. (3-3), Eq. (3-4) can be rewritten as:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{bmatrix} \quad (3\text{-}5)$$

Each pair of triangles between the neutral face and the expressive face can define a unique affine transform and each transform is determined by 6 parameters $a_1, a_2, \ldots a_6$. The 6 parameters for each warp transform are computed and all the parameters are concatenated as a long global feature vector, which is used to characterize facial configuration changes. Experiments show that the proposed geometric warp feature is more effective than the other geometric features (Chen et al., 2015; Chew et al., 2011; Lucey et al., 2010).

3.2.3 Acoustic Feature

Visual modality (face images) and audio modality (speech) can both convey the emotions and intentions of human beings. Audio modality also provides some useful clues for affect recognition in video. For instance, with voice signal, Meudt and Schwenker (Meudt and Schwenker, 2014) proposed an enhanced autocorrelation

(EAC) feature for emotion recognition in video.

One successful acoustic feature extraction is to obtain the time series of multiple paralinguistic descriptors and then using pooling operations on each time series to extract feature vectors. Schuller et al. (Schuller et al., 2010) showed how to compute the acoustic features by taking 21 functionals of 38 low level descriptors and their first regression coefficients. The 38 low-level descriptors shown in Table 3-1 are first extracted and smoothed by simple moving average low-pass filtering. After that, 21 functionals are employed and 16 zero-information features are eliminated. Finally, two single features: the number of onsets (F0) and turn duration are added. A total of 1,582 acoustic features are extracted from each video. These acoustic features include energy/spectral Low Level Descriptors (LLD) (top 6 items in Table 3-1) and voice related LLD (bottom 4 items in Table 3-1).

The representation ability of acoustic features for affect recognition is explored in this study. Experimental results show that audio modalities (speech) can provide useful complementary information in addition to visual modalities. The visual features coupled with acoustic features can achieve better performance for facial expression recognition in the wild.

Table 3-1. Acoustic features: 38 low level descriptor along with their first regression coefficients and 21 functionals ((Schuller et al., 2010)).

Descriptors	**Functionals**
PCM loudness	Position max./min.
MFCC (0-14)	Arithmetic mean
Log mel freq. band (0-7)	Skewness, kurtosis
LSP frequency (0-7)	Lin. regression coeff.
F0	Lin. regression error
F0 envelope	Quartile
Voicing prob.	Quartile range
Jitter local	Percentile
Jitter consec. frame pairs	Percentile range
Shimmer local	Up-level time

3.2.4 Multiple Feature Fusion

Features from different modalities can make different contributions. A traditional SVM concatenates different features into a single global feature vector and uses a single kernel for all these different features. However, constructing an individual kernel for each type of features and integrating these kernels optimally can enhance the discriminative power of these features. The study in (Gönen and Alpaydın, 2011) showed that using multiple kernels with different types of features can improve the recognition performance. A multiple kernel SVM was designed to learn both the decision boundaries between data from different classes and the kernel combination weights through a single optimization problem (Lanckriet et al., 2004).

Given a training set $S = \{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1,1\}\}_{i=1}^N$, a decision line is

obtained by solving the following primal optimization problem,

$$\begin{aligned} &\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ &s.t.\ \ y_i(\mathbf{w}\mathbf{x}_i + \mathrm{b}) \geq 1 - \xi_i, \\ &\quad \xi_i \geq 0,\ \ i = 1,2,\ldots \mathrm{N} \end{aligned} \tag{3-6}$$

In general, we solve the dual form of the primal optimization problem. The dual formulation of the traditional single kernel SVM optimization problem is given by

$$\max_{\alpha}\left[\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K_{ij}\right] \tag{3-7}$$

subject to $\sum_{i=1}^{N}\alpha_i y_i = 0$, $0 \leq \alpha_i \leq C$, where K_{ij} is the kernel matrix, and $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $k(\cdot,\cdot)$ is the kernel function and $\mathbf{x}_i, \mathbf{x}_j$ are the feature vectors. Multiple kernel fusion applies a linear combination of multiple kernels to substitute for the single kernel. In this study, I adopt the formulation proposed in (Rakotomamonjy et al., 2008), in which the kernel is actually a convex combination of basis kernels:

$$K_{ij} = \sum_{m=1}^{M}\beta_m k_m(\mathbf{x}_i, \mathbf{x}_j) \tag{3-8}$$

with $\beta_m \geq 0, \sum_{m=1}^{M}\beta_m = 1$.

A multiple kernel fusion framework is employed to deal with facial expression recognition under lab-controlled environment and in the wild, respectively, as shown in Figure 3-1. HOG-TOP and acoustic feature are optimally fused to handle the problem of facial expression recognition in the wild, while HOG-TOP and geometric warp feature are combined to tackle the problem of facial expression recognition under lab-controlled environment.

In the followings, I detail how to find an optimal combination of HOG-TOP and acoustic feature for facial expression recognition in the wild. It can be easily extended to the problem of facial expression recognition under lab-controlled environment.

The dynamic texture HOG-TOP is denoted as $\mathbf{x}$ and the acoustic feature as $\mathbf{z}$, then we have

$$K_{ij} = \beta k_1(\mathbf{x}_i, \mathbf{x}_j) + (1-\beta)k_2(\mathbf{z}_i, \mathbf{z}_j) \tag{3-9}$$

with $0 \leq \beta \leq 1$, where K is the kernel matrix, $k_1(\cdot,\cdot), k_2(\cdot,\cdot)$ are the basis kernels. The basis kernels could be linear kernel, radial basis function (RBF) kernel and polynomial kernel, etc. In order to find the decision boundary, it's necessary to learn the kernel weight β and coefficient α. In this work, a linear kernel is constructed for each type of feature and a two-step method is used to search for the optimal values of β and $\alpha's$. Two nested iterative loops are set to optimize both the classifier and kernel combination weights. In the outer loop, the grid search is adapted to find the kernel weight β. In the inner iteration, a solver of SVM (LIBSVM (Chang and Lin, 2011) is used in our work) is implemented by fixing the kernel weight β to find the coefficients α. Then given a sample which contains the visual feature HOG-TOP $\mathbf{x}$ and the acoustic feature $\mathbf{z}$, the predict label y can be obtained by

$$y = \text{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \left(\beta k_1(\mathbf{x}_i, \mathbf{x}) + (1-\beta)k_2(\mathbf{z}_i, \mathbf{z})\right) + b\right) \tag{3-10}$$

In this work, the one-vs-one method is employed to deal with the multiclass-SVM

problem and the max-win voting strategy is adapted to do the classification. Finally, the β value and α values with the highest overall classification accuracy in the validation data set are obtained as the optimal kernel weight and coefficients. The algorithm to compute the optimal kernel weight is shown in Algorithm 2.

Algorithm 2: Compute optimal kernel weight $\boldsymbol{\beta}^*$

Input: Training set, HOG-TOP and acoustic feature $\boldsymbol{S} = \{(\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$

Validation set, HOG-TOP and acoustic feature $\boldsymbol{T} = \{(\mathbf{x}_k, \mathbf{z}_k, \boldsymbol{y}_k)\}_{k=1}^{M}$

Output: Optimal kernel weight $\boldsymbol{\beta}^*$.

Algorithm:

Initialize highest classification accuracy $\boldsymbol{c}^* = \mathbf{0}$ and optimal kernel weight $\boldsymbol{\beta}^* = \mathbf{0}$.

for = $\boldsymbol{\beta}$:0.01:1

Apply Eq. (3-9) to compute the kernel matrix.

Solve Eq. (3-7) to get the coefficients $\boldsymbol{a}$ and bias $\boldsymbol{b}$.

With coefficients $\boldsymbol{a}$ and bias $\boldsymbol{b}$, apply Eq. (3-10) to test the SVM on the validation set and compute the classification accuracy $\boldsymbol{c}$.

if ($\boldsymbol{c}^* < \boldsymbol{c}$)

$\boldsymbol{c}^* = \boldsymbol{c}$,

$\boldsymbol{\beta}^* = \boldsymbol{\beta}$

end

end

Although I applied multiple kernel fusion to combine multiple features, it can be proved that multiple kernel fusion is equivalent to weighted feature combination. In my work, I applied two different feature sets $\mathbf{x}$ and $\mathbf{z}$. Based on the definition in (Mangai et al., 2010), feature fusion would produce a new combined feature vector as $(\mathbf{x}, \mathbf{z})$. If we consider a weighted combination, the new feature vector can be written as $(c_1\mathbf{x},\ c_2\mathbf{z})$ assuming that all the features in the same set has the same weight. When

we train an SVM to perform the classification, we first apply a kernel function to transform the new feature vector:

$$k([c_1\mathbf{x}_i, c_2\mathbf{z}_i], [c_1\mathbf{x}_j, c_2\mathbf{z}_j]) \tag{3-11}$$

For the linear kernel, it can be written as:

$$[c_1{\mathbf{x}_i}^T, c_2{\mathbf{z}_i}^T] \cdot [c_1\mathbf{x}_j, c_2\mathbf{z}_j] = {c_1}^2\left({\mathbf{x}_i}^T \cdot \mathbf{x}_j\right) + {c_2}^2({\mathbf{z}_i}^T \cdot \mathbf{z}_j) \tag{3-12}$$

The above equation can be rewritten as

$${c_1}^2\left({\mathbf{x}_i}^T \cdot \mathbf{x}_j\right) + {c_2}^2\left({\mathbf{z}_i}^T \cdot \mathbf{z}_j\right) = \beta_1 \ k_1\left(\mathbf{x}_i, \mathbf{x}_j\right) + \beta_2 \ k_2\left(\mathbf{z}_i, \mathbf{z}_j\right) \tag{3-13}$$

where $\beta_1 = {c_1}^2$, $\beta_2 = {c_2}^2$, k_1 and k_2 are both linear kernels.

In my work, I applied the following expression to combine two feature sets:

$$\beta \ k_1\left(\mathbf{x}_i, \mathbf{x}_j\right) + (1-\beta) \ k_2\left(\mathbf{z}_i, \mathbf{z}_j\right) \tag{3-14}$$

with a constraint: $0 \le \beta \le 1$. At the training stage, the kernel fusion is actually equivalent to the weighted feature combination with ${c_1}^2 = \beta$ and ${c_2}^2 = 1-\beta$, which is actually a kind of feature fusion.

At the classification stage, we apply the following decision function:

$$\mathrm{y} = \mathrm{sgn}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i}\alpha_\mathrm{i}\left(\beta k_1(\mathbf{x}_i, \mathbf{x}) + (1-\beta)k_2(\mathbf{z}_i, \mathbf{z})\right) + \mathrm{b}\right) \tag{3-15}$$

At first sight, Eq. (3-15) seems to use a score fusion to combine the contributions from two feature sets. Using Eq. (3-13), Eq. (3-15) can be rewritten as

$$\mathrm{y} = \mathrm{sgn}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i}\alpha_\mathrm{i}\left({c_1}^2({\mathbf{x}_i}^{\mathrm{T}} \cdot \mathbf{x}) + {c_2}^2({\mathbf{z}_\mathrm{i}}^{\mathrm{T}} \cdot \mathbf{z})\right) + \mathrm{b}\right) \tag{3-16}$$

We can further obtain

$$\mathrm{y} = \mathrm{sgn}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i}\alpha_\mathrm{i}([c_1{\mathbf{x}_i}^{\mathrm{T}}, c_2{\mathbf{z}_i}^{\mathrm{T}}] \cdot [c_1\mathbf{x}, c_2\mathbf{z}]) + \mathrm{b}\right) \tag{3-17}$$

where $[c_1\mathbf{x}_i{}^{\mathrm{T}}, c_2\mathbf{z}_i{}^{\mathrm{T}}]$ denotes the weighted combination of the support feature vector and $[c_1\mathbf{x}, c_2\mathbf{z}]$ is the weighted combination of the test feature vector. It means that the SVM classification with multiple linear kernels is equivalent to classifying the weighted combination of the test feature vector directly. It clearly shows that multi-kernel classification with linear kernels is actually a feature fusion method with weighted combination of different features.

For a non-linear kernel function, it involves a mapping from the original feature space to a higher-dimensional feature space:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \ \phi(\mathbf{x}_j) \tag{3-18}$$

For two non-linear kernels, we can get:

$$\beta_1 k_1\ (\mathbf{x}_i, \mathbf{x}_j) + \beta_2 k_2(\mathbf{z}_i, \mathbf{z}_j) = c_1^2\phi_1(\mathbf{x}_i)^T\ \phi_1(\mathbf{x}_j) + \ c_2^2\phi_2(\mathbf{z}_i)^T\ \phi_2(\mathbf{z}_j) \tag{3-19}$$

And, the right side of Eq. (3-19) can be written as for non-linear kernels,

$$\begin{aligned} &c_1^2\phi_1(\mathbf{x}_i)^T\ \phi_1(\mathbf{x}_j) + \ c_2^2\phi_2(\mathbf{z}_i)^T\ \phi_2(\mathbf{z}_j) \\ &= [c_1\phi_1(\mathbf{x}_i)^T, c_2\phi_2(\mathbf{z}_i)^T] \cdot [c_1\phi_1(\mathbf{x}_j), c_2\phi_2(\mathbf{z}_j)] \end{aligned} \tag{3-20}$$

It shows that multiple non-linear kernel fusion is also a type of feature fusion in the mapped high-dimensional feature space. For the SVM classification with multiple non-linear kernels, with Eq. (3-17) and Eq. (3-20), we can obtain

$$\mathrm{y} = \mathrm{sgn}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i}\alpha_\mathrm{i}([c_1\phi_1(\mathbf{x}_\mathrm{i})^T, c_2\phi_2(\mathbf{z}_\mathrm{i})^T] \cdot [c_1\phi_1(\mathbf{x}), c_2\phi_2(\mathbf{z})]) + \mathrm{b}\right) \tag{3-21}$$

where $[c_1\phi_1(\mathbf{x}_\mathrm{i})^T, c_2\phi_2(\mathbf{z}_\mathrm{i})^T]$ and $[c_1\phi_1(\mathbf{x}), c_2\phi_2(\mathbf{z})]$ denote the weighted combination of support feature vector and test feature vector in the mapped

high-dimensional feature space, respectively. It is reasonable to conclude that SVM classification with multiple non-linear kernels is also equivalent to classifying the weighted combination of the test feature vector in the mapped high-dimensional feature space.

3.3 Experiments and Discussion

3.3.1 Data Sets

In order to evaluate the proposed methods, I conduct the experiments on three public data sets: the Extended Cohn-Kanade (CK+) data set (Lucey et al., 2010), the GEMEP-FERA 2011 data set (Valstar et al., 2011) and the Acted Facial Expression in Wild (AFEW) 4.0 (Abhinav Dhall et al., 2012) data set. I first give a brief description to these three data sets.

The Extended Cohn-Kanade (CK+) data set contains 593 image sequences from 123 subjects. The face images in the database are collected under lab-controlled environment. The image sequences vary in duration from 10 to 60 frames. In total, 327 of 593 image sequences have emotion labels and each is categorized into one of the following seven emotion classes: anger (An), contempt (Co), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa) and surprise (Su). Each image sequence changes from the onset (the neutral frame) to the peak (the expressive frame). In addition, the

X–Y coordinates of 68 facial landmark points were given for each image in the database. The landmark points of key frames of each video sequence were manually labelled, while the remaining frames were automatically aligned using the AAM fitting algorithm (Matthews and Baker, 2004).

The GEMEP-FERA 2011 data set contains 289 video sequences from 10 actors who were trained by a professional director. It is divided into a training set of 155 sequences and a test set of 134 sequences. Each sequence is categorized into the following five emotions: anger (An), happiness (Ha), relief (Re), fear (Fe), and sadness (Sa). Only the training set provides emotion labels. This database is more challenging than the CK+ database, since there are head movements and gesture variations in image sequences.

The Acted Facial Expression in Wild (AFEW) 4.0 data set includes video clips collected from different movies which are believed to be close to real world conditions. The database splits into a training set, a validation set and a test set. There are 578 video clips in the training set. The validation and test sets have 383 video clips and 407 video clips, respectively. Each video clip belongs to one of the seven categories: anger (An), disgust (Di), fear (Fe), happiness (Ha), neutral (Ne), sadness (Sa), and surprise (Su). This database provides original video clips and aligned face sequences. They applied the model proposed in (Zhu and Ramanan, 2012) to extract

the faces from video clips and align the faces. Different from the CK+ and GEMEP-FERA 2011 data sets, facial expressions in AFEW 4.0 are more natural and spontaneous. The variations in illumination, pose and background in image sequences increase the complexity of facial expression analysis.

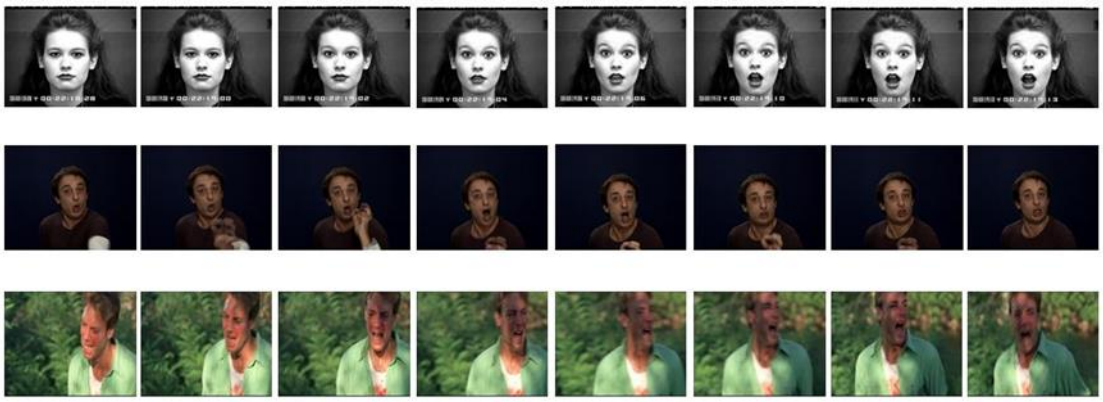


Figure 3-7. The selected image sequences from the three databases. From top to bottom: CK+, GEMEP-FERA2011 and AFEW 4.0.

Figure 3-7 shows the selected image sequences from the three databases. The first row is the face images from the CK+ database, which are frontal-view and lab-controlled faces. The middle row shows the images from the GEMP-FERA 2011 database. There exist head movements and gesture variations. The bottom row is an image sequence from the AFEW 4.0 database. It can be seen that the background is complex and there exist both illumination changes and pose variations.

3.3.2 Feature Extraction

In our experiments, three types of features were extracted, namely HOG-TOP,

geometric warp feature and acoustic feature.

In extracting HOG-TOP from image sequences, each face image was first cropped and resized to 128×128. The resized face image was then partitioned into 8×8 blocks with a size of 16×16. The bin number was set to 9 with an angle range of $0° - 180°$. In each block, an HOG-TOP was obtained with a dimension of 3×9=27. The HOG-TOP of the 8×8 blocks are further concatenated into a long feature vector with a dimension of 3×9×8×8=1728.

Facial landmarks were used to compute the geometric warp features. I computed the warp transform of facial landmarks between the neutral face and an expressive face. Each face contains 68 facial landmarks. These facial landmarks divide the face into many non-overlap sub regions by Delaunay triangulation. In this work, 109 pair of triangles (the smallest number of triangles available in face images) was acquired. Each pair of triangles between the neutral face and an expressive face can define a unique transform and each affine transform is determined by six parameters. The warp transform coefficients are finally concatenated as a feature vector of 6×109=654 elements to represent the geometric warp feature.

The acoustic features with a length of 1582 used in this work are provided by the database (Abhinav Dhall et al., 2012; Dhall et al., 2014). The acoustic features were extracted by applying the open-source Emotion Affect Recognition (openEAR) toolkit

(Eyben et al., 2009) backend OpenSMILE (Florian Eyben et al., 2010).

3.3.3 Experimental Results

3.3.3.1 A Comparison of HOG-TOP and LBP-TOP

I first compare the performance of HOG-TOP proposed in this work with LBP-TOP proposed in (Zhao and Pietikainen, 2007). When computing the LBP-TOP features, the general settings adopted in most reported works are used. The resized face image is partitioned into 4×4 blocks. The LBP-TOP is coded with a uniform pattern. The LBP-TOP histogram of each block is a feature vector of 3×59=177 elements. The length of the feature vector consists of 4×4 blocks is 3×59×4×4=2832.

There are 327 image sequences with emotion labels belonging to 118 subjects in the CK+ database. I followed the protocol proposed in (Lucey et al., 2010) and took the leave-one-subject-out cross validation strategy. Each time the samples from one subject were used for testing and the remaining samples from all other subjects were used for training. In order for each subject to be evaluated once, I carried out 118 validations. The classification accuracy obtained on the CK+ database by using two types of features is shown Table 3-2.

I also compare the performance of the two features in the GEMEP-FERA 2011 database. Since only the emotion labels of the training set are publicly available. Only

the training set is used to do the evaluation. There are seven subjects in the training set. I adopted the leave-one-subject-out strategy and carried out seven cross validations. Table 3-3 shows the performance obtained by applying the two feature descriptors.

As for the AFEW 4.0 database, I utilize the training set to train an SVM classifier and test the classifier on the validation set. The database provided a baseline method (Dhall et al., 2014) which employed LBP-TOP to represent the dynamic textures of the video sequence and trained an SVM with non-linear RBF kernel for emotion classification. The accuracy obtained on the AFEW 4.0 database by applying two types of features is shown in Table 3-4. The overall accuracy is used to evaluate the performance. The overall accuracy is defined as

$$O_{acc} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} m_{nk}}{\sum_{n=1}^{N}\sum_{k=1}^{K} M_{nk}} \tag{3-22}$$

where K is the number of classes, N is the number of cross validation folds, m_{nk} is the number of correctly predicted samples of the k-th class in the n-th fold, and M_{nk} denotes the total samples of the k-th class in the n-th fold. The classification rate of each individual facial expression (k-th class) over n validation folds is given by $\frac{\sum_{i=1}^{N} m_{ik}}{\sum_{i=1}^{N} M_{ik}}$.

Table 3-2. The classification accuracy of LBP-TOP and HOG-TOP on the CK+ database (%).

	Overall	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
LBP-TOP	**89.3**	75.6	88.9	93.2	80.0	98.5	78.6	92.8
HOG-TOP	**89.6**	88.9	66.7	94.9	76.0	95.6	67.9	97.6

Table 3-3. The classification accuracy of LBP-TOP and HOG-TOP on the GEMEP-FERA 2011 database (%).

	Overall	Anger	Fear	Joy	Relief	Sadness
LBP-TOP	**53.6**	56.2	26.7	58.1	51.6	74.2
HOG-TOP	**54.2**	43.7	36.7	61.3	54.8	74.2

Table 3-4. The classification accuracy of LBP-TOP and HOG-TOP on the AFEW 4.0 database (%).

	Overall	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
LBP-TOP	**30.6**	19.0	50.0	25.0	15.2	57.1	16.4	21.7
HOG-TOP	**35.8**	58.7	73.4	22.5	4.3	60.3	4.9	2.2

Experimental results show that the overall classification accuracy obtained by using HOG-TOP on the CK+ database and the GEMEP-FERA 2011 database is 89.6% and 54.2%, respectively. It is competitive with the result of 89.3% and 53.6% obtained by applying LBP-TOP on the two databases. While the overall classification rate of HOG-TOP on the AFEW 4.0 database is 35.8%, which is better than 30.6% obtained by using LBP-TOP, meaning that HOG-TOP is more robust in capturing the

subtle facial appearance changes in the wild. In addition, HOG-TOP with a length of 1728 is more compact than LBP-TOP with a length of 2832.

I also compare the computational speeds of the two features under the 64-bit Win 7 operating system with a Core i7 CPU. The two features are computed with Matlab 8.2. The computation time depends on the block size and sequence duration. With the same block size (16×16) and sequence duration (11 frames), the computation time of HOG-TOP and LBP-TOP is 0.027s and 0.042s, respectively, showing the computational efficiency of HOG-TOP.

3.3.3.2 Facial Expression Recognition Under Lab-controlled Environment

A model which combines HOG-TOP and geometric warp was developed to handle the problem of facial expression recognition under lab-controlled environment. The following feature sets are evaluated in this experiment: geometric warp feature, dynamic appearance feature (HOG-TOP), hybrid feature I and hybrid feature II. Hybrid feature I denotes the feature vector of concatenating HOG-TOP and geometric warp feature directly and hybrid feature II is the optimal combination of the HOG-TOP and geometric warp feature.

I first compare our proposed geometric warp feature with the other geometric features on the CK+ data set. All the methods take the leave-one-subject-out cross

validation. Experimental results are shown in Table 3-5. The method in (Lucey et al., 2010) applied a set of facial landmarks to characterize the face shape directly. Chew et al. (Chew et al., 2011) applied a constrained local model to extract similarity normalized shape as geometric features. The shifts of the facial landmarks between the neutral face and the expressive face were computed to represent the geometric feature in (Chen et al., 2015). Table 3-5 shows that our proposed geometric warp feature achieves a superior performance compared with the other geometric features, meaning that the proposed geometric warp feature is more effective to capture facial configuration changes.

Table 3-5. The results of the different geometric features on the CK+ database (%). (GWF is our proposed geometric warp feature).

	GWF	Lucey et al.	Chew et al.	Chen et al.
Anger	86.7	35.0	70.1	62.2
Contempt	94.4	25.0	52.4	72.2
Disgust	96.6	68.4	92.5	86.4
Fear	36.0	21.7	72.1	56.0
Happiness	98.5	98.4	94.2	91.3
Sadness	75.0	4.0	45.9	39.3
Surprise	96.4	100.0	93.6	95.2
Overall	89.0	66.7	82.3	79.2

I further evaluate hybrid feature I and hybrid feature II with the leave-one-subject-out cross validation on the CK+ database and compare the performance with that obtained by applying geometric feature and HOG-TOP alone.

Table 3-6 shows the classification accuracy obtained by using the two different feature sets and two different combination schemes. Figure 3-8 shows the confusion matrices of using two different feature sets and two different combination schemes. Experimental results illustrate that the emotions “disgust”, “happiness” and “surprise” have higher classification rates than the other emotions, indicating that these three emotions are easier to be distinguished than the others.

	An	Co	Di	Fe	Ha	Sa	Su
An	0.89	0.02	0.02	0.02	0.00	0.04	0.00
Co	0.17	0.67	0.00	0.05	0.05	0.00	0.06
Di	0.01	0.00	0.95	0.02	0.00	0.00	0.02
Fe	0.00	0.00	0.04	0.76	0.08	0.12	0.00
Ha	0.00	0.00	0.00	0.01	0.96	0.00	0.03
Sa	0.21	0.00	0.00	0.07	0.00	0.68	0.04
Su	0.00	0.01	0.00	0.00	0.00	0.01	0.98

(a)

	An	Co	Di	Fe	Ha	Sa	Su
An	0.87	0.00	0.06	0.00	0.00	0.07	0.00
Co	0.00	0.94	0.00	0.00	0.00	0.06	0.00
Di	0.01	0.00	0.97	0.02	0.00	0.00	0.00
Fe	0.00	0.00	0.04	0.36	0.28	0.04	0.28
Ha	0.00	0.00	0.00	0.01	0.99	0.00	0.00
Sa	0.14	0.00	0.07	0.04	0.00	0.75	0.00
Su	0.00	0.02	0.00	0.00	0.00	0.02	0.96

(b)

	An	Co	Di	Fe	Ha	Sa	Su
An	0.96	0.00	0.04	0.00	0.00	0.00	0.00
Co	0.00	0.94	0.00	0.00	0.00	0.00	0.06
Di	0.03	0.00	0.95	0.00	0.02	0.00	0.00
Fe	0.00	0.00	0.04	0.52	0.24	0.04	0.16
Ha	0.00	0.00	0.00	0.01	0.99	0.00	0.00
Sa	0.14	0.00	0.03	0.04	0.00	0.79	0.00
Su	0.00	0.02	0.00	0.00	0.00	0.02	0.96

(c)

	An	Co	Di	Fe	Ha	Sa	Su
An	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Co	0.00	0.94	0.00	0.00	0.00	0.00	0.06
Di	0.01	0.00	0.97	0.00	0.00	0.02	0.00
Fe	0.00	0.00	0.04	0.84	0.12	0.00	0.00
Ha	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Sa	0.21	0.00	0.00	0.00	0.00	0.79	0.00
Su	0.00	0.01	0.00	0.00	0.00	0.00	0.99

(d)

Figure 3-8. The confusion matrices obtained by using two feature sets and two combination schemes on the CK+ database: (a) HOG-TOP, (b) geometric feature, (c) hybrid feature I and (d) hybrid feature II. (An: Anger, Co: Contempt, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness, and Su: Surprise).

Table 3-6 also shows that hybrid feature I (91.4%) and hybrid feature II (95.7%) outperform the geometric warp feature (89.0%) and HOG-TOP (89.6%) applied individually. It can be concluded that different features (hybrid feature I) can provide complementary information and multiple feature fusion can further enhance the discriminative ability of the combined features (hybrid feature II).

Table 3-6. The classification accuracy obtained by using two feature sets and two combination schemes on the CK+ database (%).

Feature set	Overall	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
HOG-TOP	**89.6**	88.9	66.7	94.9	76.0	95.6	67.9	97.6
Geometric feature	**89.0**	86.7	94.4	96.6	36.0	98.5	75.0	96.4
Hybrid Feature I	**91.4**	95.6	94.4	94.9	52.0	98.5	78.6	96.4
Hybrid Feature II	**95.7**	100.0	94.4	96.6	84.0	100.0	78.6	98.8

I also compare our method with the other methods. All the methods compared follow the baseline method (Lucey et al., 2010) and take the leave-one-subject-out cross validation. The method in (Lucey et al., 2010) and (Chew et al., 2011) combined geometric feature and appearance feature and trained an SVM to perform the classification. In (X. Huang et al., 2011), a weighted component-based feature descriptor to extract dynamic appearance feature was utilized and multiple kernel learning was applied to train the SVM for recognition. A sparse temporal representation classifier was proposed for facial expression recognition in (Chew et al.,

2012). The method in (X. Huang et al., 2012) applied spatiotemporal local monogenic binary pattern (STLMBP) feature to handle the problem of facial expression recognition.

As can be seen from Table 3-7, the HOG-TOP (89.6%) and geometric feature (89.3%) proposed in our method can achieve a competitive performance compared with SPTS+CAPP (Lucey et al., 2010) (88.4%), CLM (Chew et al., 2011) (82.4%) and STLMBP (X. Huang et al., 2012) (88.4%). It demonstrates the effectiveness of our proposed features. The hybrid feature II as the optimal combination of HOG-TOP and geometric feature achieves a superior performance compared with the other methods tested, showing the effectiveness of the multiple feature method.

Table 3-7. Performance comparison with other methods on CK+ database.

Method	Accuracy (%)
HOG-TOP	89.6
Geometric feature	89.0
Hybrid Feature I	91.4
Hybrid Feature II	**95.7**
SPTS+CAPP (Lucey et al., 2010)	88.4
CLM (Chew et al., 2011)	82.4
STLMBP (X. Huang et al., 2012)	88.4
STR (Chew et al., 2012)	94.9
CFD (X. Huang et al., 2011)	93.2

3.3.3.3 Facial Expression Recognition in the Wild

HOG-TOP and acoustic feature are fused to tackle the problem of facial expression recognition in the wild. I first evaluate our method on the validation set. Four feature sets are explored: HOG-TOP only, acoustic feature only, hybrid feature I and hybrid feature II. Hybrid feature I concatenates the HOG-TOP and acoustic feature directly. Hybrid feature II is the optimal combination of the HOG-TOP and acoustic feature.

Table 3-8 shows the classification accuracy obtained by applying two different feature sets and two combination schemes. The corresponding confusion matrices are shown in Figure 3-9. The classification rates shown in Table 3-8 are much lower than the results shown in Table 3-6. Different from facial expressions under lab-controlled environment in which the actors or subjects can pose distinguished facial expressions, the facial expressions in the wild are more subtle. The factors including head movements, pose variations etc. also increase classification difficulties. And sometimes, several facial expressions in the wild may appear together, which makes a facial expression to be confused with other expressions.

It can be seen that the classification rate of emotion “surprise” is the lowest. The confusion matrices shown in Figure 3-9 show that the emotion “surprise” is mostly misclassified as emotions “anger”, “happiness” and “neutral”. The emotions “anger”

and “neutral” have higher recognition accuracies than the other emotions. Hybrid feature I and hybrid feature II outperform the HOT-TOP and acoustic feature used individually, indicating that two feature sets are complementary with each other. Hybrid feature II achieves a superior performance compared with hybrid feature I, demonstrating that the effectiveness of the multiple features in dealing with the facial expression recognition problem in the wild.

Table 3-8. The classification accuracy obtained by using two different feature sets and two combination schemes on the validation set of the AFEW 4.0 database (%).

Feature set	Overall	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
HOG-TOP	**35.8**	58.7	73.4	22.5	4.3	60.3	4.9	2.1
Acoustic feature	**32.9**	57.1	64.1	15.0	26.1	34.9	14.7	0.0
Hybrid Feature I	**37.6**	65.1	75.0	12.5	8.70	57.1	13.1	4.4
Hybrid Feature II	**40.2**	69.8	76.6	17.5	15.2	63.5	9.8	2.1

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.73	0.05	0.06	0.02	0.14	0.00	0.00
Di	0.23	0.23	0.08	0.13	0.33	0.03	0.00
Fe	0.52	0.11	0.04	0.11	0.17	0.04	0.00
Ha	0.05	0.06	0.08	0.60	0.19	0.02	0.00
Ne	0.16	0.03	0.05	0.10	0.59	0.08	0.00
Sa	0.31	0.13	0.00	0.08	0.43	0.05	0.00
Su	0.26	0.02	0.07	0.09	0.50	0.04	0.02

(a)

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.64	0.03	0.05	0.13	0.11	0.02	0.03
Di	0.18	0.15	0.00	0.18	0.35	0.10	0.05
Fe	0.33	0.02	0.26	0.13	0.15	0.02	0.09
Ha	0.24	0.03	0.11	0.35	0.27	0.00	0.00
Ne	0.08	0.10	0.06	0.14	0.57	0.03	0.02
Sa	0.02	0.15	0.11	0.23	0.30	0.15	0.05
Su	0.17	0.09	0.13	0.17	0.35	0.09	0.00

(b)

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.75	0.06	0.02	0.02	0.11	0.03	0.02
Di	0.23	0.13	0.00	0.23	0.38	0.05	0.00
Fe	0.41	0.11	0.09	0.15	0.20	0.04	0.00
Ha	0.11	0.03	0.06	0.57	0.19	0.02	0.02
Ne	0.08	0.14	0.00	0.06	0.65	0.06	0.00
Sa	0.21	0.23	0.03	0.07	0.33	0.13	0.00
Su	0.15	0.11	0.07	0.15	0.48	0.00	0.04

(c)

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.77	0.06	0.02	0.02	0.11	0.03	0.00
Di	0.28	0.18	0.00	0.15	0.33	0.08	0.00
Fe	0.39	0.09	0.15	0.15	0.20	0.02	0.00
Ha	0.08	0.05	0.06	0.63	0.16	0.02	0.00
Ne	0.10	0.08	0.00	0.08	0.70	0.05	0.00
Sa	0.21	0.18	0.00	0.07	0.44	0.10	0.00
Su	0.22	0.04	0.07	0.15	0.50	0.00	0.02

(d)

Figure 3-9. The confusion matrices obtained by using two feature sets and two combination schemes on the validation set of AFEW 4.0 database. (a) HOG-TOP, (b) acoustic feature, (c) hybrid feature I and (d) hybrid feature II. (An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Ne: Neutral, Sa: Sadness, and Su: Surprise).

Hybrid feature II which achieves the best performance on the validation set was further applied to evaluate the test set. The overall recognition accuracy on the test set is 45.2%. Table 3-9 shows the results compared with the other methods. The baseline method in (Dhall et al., 2014) combined LBP-TOP and the acoustic feature. Lip activity was incorporated with voice in (Ringeval et al., 2014) to tackle the emotion recognition problem. The method in (X. Huang et al., 2014) used dynamic textures only and the method in (Meudt and Schwenker, 2014) applied the voice only. The method in (Sun et al., 2014) employed audiovisual feature for emotion recognition. Table 3-9 shows that our method (45.2%) improves significantly compared with the baseline method (Dhall et al., 2014) and the method in (Ringeval et al., 2014), with an improvement of about 11% and 10%, respectively. Our method is also better than (X.

Huang et al., 2014) (41.5%) and EAC (Meudt and Schwenker, 2014) (40.1%). Compared with the method in (Kaya and Salah, 2014) (44.2%), our performance is still competitive. Moreover, our proposed method (Chen et al., 2014) participated in the second emotion recognition in the wild challenge (EmotiW 2014) (Dhall et al., 2014) and achieved the second runner-up award.

Table 3-9. Performance comparison with other methods on the test set of AFEW 4.0 database.

Method	Accuracy (%)
HOG-TOP+Voice (My method)	**45.2**
LBP-TOP + Voice (Dhall et al., 2014)	33.7
Lip activity + Voice (Ringeval et al., 2014)	35.3
STLMBP (Huang 2014 et al. 2014)	41.5
EAC (Meudt and Schwenker, 2014)	40.1
LBP+ELM (Kaya and Salah, 2014)	44.2
SIFT + BoW + Voice (Sun et al., 2014)	47.2
DCNN + SIFT + Voice (M. Liu et al., 2014)	50.2

Sun et al. (Sun et al., 2014) applied multimodal features (HOG, LBP, SIFT and audio features) and multiple kernel learning to handle this problem and won the first runner-up award with a classification accuracy of 47.17%. Liu et al. (M. Liu et al., 2014) utilized another database (Celebrity Faces in the Wild (CFW), (X. Zhang et al., 2012) to train a deep convolutional neural network (DCNN) and combined the learned features with other hand-designed features (HOG, SIFT and audio features) to tackle

this problem. With more training data, they acquired a classification accuracy of 50.37% and won the champion of this challenge.

3.3.3.4 Decision fusion versus Feature fusion

The multiple feature fusion method applied in this work is a kind of feature fusion method. Another technique, namely decision-level fusion, is also widely used in computer vision community to deal with multiple sets of features. In a preliminary study, the effectiveness of the two techniques for facial expression recognition in video was explored. A preliminary experiment was conducted on the AFEW 4.0 database. As mentioned above, the one-vs-one technique was employed to tackle the multiclass-SVM problem, and max-win voting strategy was used to conduct the classification. For decision-level fusion, I first applied the HOG-TOP and acoustic feature separately and then saved the predict results, i.e. the number of votes for each class of the two features, respectively. After that, the votes obtained by each individual feature were added and based on these combined votes, the max-win voting strategy was carried out again to make the final decision. The overall classification rate is computed as the performance of decision-level fusion method.

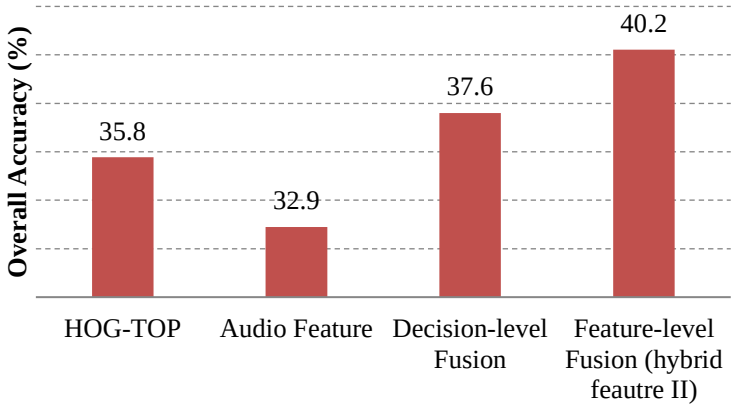


Figure 3-10. A comparison of different methods on the validation set of the AFEW 4.0 database.

Figure 3-10 shows the experimental results of the different methods tested. The overall accuracies of HOG-TOP, acoustic feature and feature-level fusion shown in Figure 3-10 are the same as those shown in Table 3-8. Experimental results show that feature fusion outperforms the decision fusion method, although they both utilize the same multiple sets of features and improve the recognition performance over individual features.

3.3.3.5 The Effect of Block Size on HOG-TOP

I further explore the representation ability of HOG-TOP with different block sizes, from 8×8 to 32×32. Table 3-10 shows the parameters used for extracting HOG-TOP. The blocks with a small size (12 and 16) are not overlapped and large blocks (24×24 and 32×32) are half overlapped. HOG-TOP is employed to evaluate on the CK+ database and the AFEW 4.0 database. Experimental results are shown in

Tables 3-11 and 3-12.

Table 3-10. The parameters used for extracting HOG-TOP.

Image size	Block size	Overlap	Blocks
96×96	12×12	No	8×8
96×96	24×24	Half	7×7
128×128	16×16	No	8×8
128×128	32×32	Half	7×7

It can be seen that the HOG-TOP with various block sizes achieve the similar overall accuracy. It can be concluded that HOG-TOP is robust to scales. For facial expressions under lab-controlled environment (Table 3-11), HOG-TOP with a small size (12) is more effective to recognize the facial expressions “fear” and “contempt” which have subtle facial muscle activities. A small block size is more robust to capture local subtle appearance changes than a large block size. For facial expressions in the wild (Table 3-12), HOG-TOP with a large size (24 and 32) achieves a superior performance for the facial expression “surprise”, indicating that HOG-TOP with a large block size is more robust to distinguish “surprise” expression from others in the wild. Table 3-12 also shows that HOG-TOP with various block sizes outperforms the LBP-TOP (30.6%) for facial expression recognition in the wild.

Table 3-11. The performance of HOG-TOP with various block sizes on the CK+ database (%).

	12×12	**24×24**	**16×16**	**32×32**
Angry	84.4	80.0	88.9	84.4
Contempt	72.2	66.7	66.7	66.7
Disgust	93.2	93.2	94.9	96.6
Fear	88.0	80.0	76.0	72.0
Happy	95.7	95.7	95.7	97.1
Sad	64.3	64.3	67.9	60.7
Surprise	96.4	96.4	97.6	97.6
Overall	89.3	87.8	89.6	88.7

Table 3-12. The performance of HOG-TOP with various block sizes on the AFEW 4.0 database (%).

	12×12	**24×24**	**16×16**	**32×32**
Neutral	54.0	38.1	58.7	46.0
Angry	71.9	71.9	73.4	73.4
Disgust	20.0	15.0	22.5	20.0
Fear	6.50	15.2	4.30	13.0
Happy	57.1	63.5	60.3	60.3
Sad	4.90	9.80	4.90	9.80
Surprise	2.20	13.0	2.20	8.70
Overall	34.2	35.2	35.8	36.0

3.3.4 Discussion

The experimental results reported above demonstrate that the proposed framework can efficiently handle the problem of facial expression recognition in video. Facial expressions under lab-controlled environment are different from those in the wild which are more natural and spontaneous. Two approaches are proposed to tackle the two different facial expression recognition problems in this work. The

HOG-TOP feature outperforms the geometric feature and audio feature in the two approaches, indicating that facial appearance plays an important role for both facial expression recognition problems. Compared with LBP-TOP, HOG-TOP is more compact and effective to characterize facial appearance changes. Facial configuration changes also provide useful clues for facial expression analysis. The facial landmarks can be located exactly on a face image under lab-controlled, representing the facial configuration changes caused by facial muscle movements. A new effective geometric feature based on warp transform of facial landmarks is proposed and the new geometric warp feature is robust to capture facial configuration changes. On the other hand, it is very challenging to locate facial landmarks on face images in the wild. However, the voice also plays an important role on affect recognition. Instead of using geometric feature, acoustic feature is employed for facial expression recognition in the wild. Experimental results show that different features can make different contributions to facial expression recognition and the multiple feature fusion can enhance the discriminative ability of the multiple features. It can be seen that for facial expression recognition in the wild, although our method outperforms the baseline method, the performance is in general not as good as that in facial expression recognition under lab-controlled. Facial expression recognition in the wild is much more challenging and it will be one of our future research focuses.

3.4 Summary

Video based facial expression recognition is a challenging and long standing problem. In this chapter, I discuss the potentials of audiovisual modalities and propose an effective framework with multiple-feature fusion to handle this problem. Both the visual modality (face images) and audio modality (speech) were utilized in this study. A new feature descriptor called histogram of oriented gradients from three orthogonal planes (HOG-TOP) is proposed to extract dynamic textures from video sequences to characterize facial appearance changes. Experiments conducted on three public databases (CK+, GEMEP-FERA 2011, AFEW4.0) have shown that HOG-TOP performs as well as a widely used feature LBP-TOP in representing dynamic textures from video sequences. Moreover, HOG-TOP is more effective to capture subtle facial appearance changes and robust in dealing with facial expression recognition in the wild. In addition, HOG-TOP is more compact and computationally more efficient. In order to capture facial configuration changes, an effective geometric feature deriving from the warp transform of the facial landmarks is also introduced. Realizing that voice is another powerful way for human beings to transmit message, I also explored the role of voice and employed the acoustic feature for affect recognition in video. The multiple-feature fusion was applied to deal with facial expression recognition under lab-controlled environment and in the wild. Experiments conducted on two

facial expression datasets, CK+ and AFEW 4.0 demonstrate that our approach can achieve a promising performance for facial expression recognition in video.

Chapter 4 A New Framework with Multiple Tasks for Detecting and Locating Pain Events in Video

4.1 Introduction

Pain analysis in video has two problems to be solved: pain event detection and pain event locating. The first problem is to detect whether there exists any pain event in a video sequence; the second problem is to locate the pain events in a video sequence if there is any. In this chapter, a new framework with multiple tasks is proposed to jointly tackle the two problems. Considering that information with various time scales (frame, segment and sequence) can make different contributions, I propose to combine three different tasks, that is frame-level, segment-level and sequence-level detection, to effectively detect and locate pain events in video. Figure 4-1 shows the pipeline of our proposed framework. In this framework, HOG of fiducial points (P-HOG) is used to characterize spatial features from video frames and an SVM classifier is trained for frame-level detection. In order to further exploit spatial-temporal information among contiguous frames, segment-level detection is proposed to assist the frame-level detection. HOG from three orthogonal planes (HOG-TOP) are applied to model dynamic textures of video segments. An SVM

classifier is trained as a segment-level paint event detector. I further apply the max pooling strategy to obtain global P-HOG and HOG-TOP to represent the whole video sequence and employ multiple kernel fusion to optimally combine the two types of global features. An SVM with multiple kernels is trained to perform the sequence-level (pain event) detection. At last, an effective probabilistic fusion method is proposed to integrate the detection results of the three different tasks (frame-level, segment-level and sequence-level detection) to locate pain events in video. By integrating three different tasks, the proposed method provides a more robust and precise detection of pain events in video than the other previously reported techniques which usually focus on one of these tasks only.

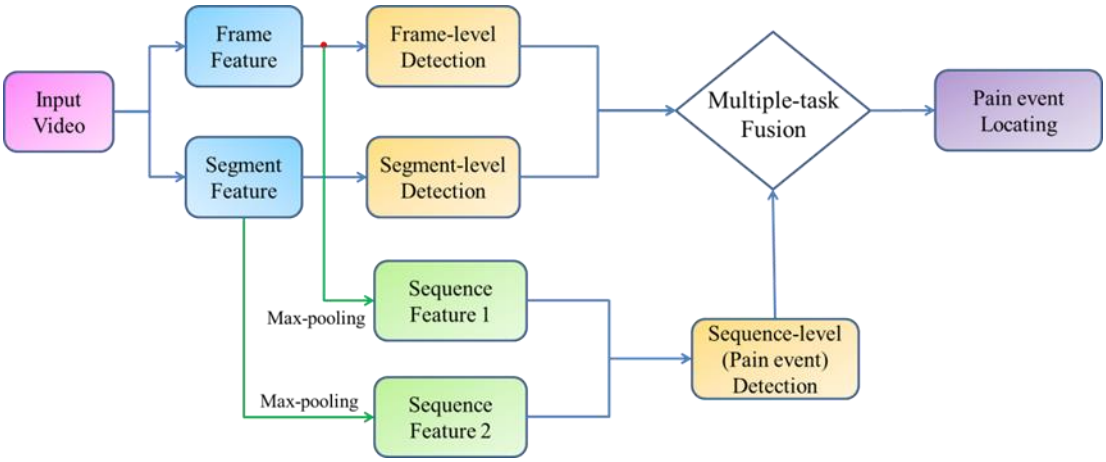


Figure 4-1. The pipeline of our proposed framework for joint pain event detection and locating in video.

4.2 Methodology

4.2.1 Frame-level Detection

Frame-level detection tries to detect the pain presence/absence of each frame, which is a binary classification problem. For frame-level detection, an SVM with spatial features extracted from video frames was trained for this task. For facial expression analysis, there are two major types of features considered: appearance and geometric features. Geometric features often use facial fiducial points to describe the face shape while appearance features mainly characterize the textures of faces. In this study, both geometric information (facial fiducial points) and appearance information (HOG) were utilized. I extracted HOG from the neighborhoods of facial fiducial points (P-HOG) to characterize the spatial feature of each video frame and trained an SVM to perform the classification.

Facial expressions are caused by facial muscle movements, especially the muscles around the mouth, nose and eyes. Features can be extracted from the interesting regions directly. A face image was first tracked with active appearance models (AAMs) (Matthews and Baker, 2004), and some facial landmarks were labeled on the face, such as the blue points shown in Figure 4-2. The facial landmarks around the face outline are ignored and only the landmarks around the mouth, nose

and eyes are considered. We can define a neighborhood (a local patch) centered in each landmark and extract the appearance features from each local patch. Here HOG is employed to characterize appearance features. The HOG features extracted from each local patch are concatenated to form a global feature vector to represent the whole spatial appearance feature of each frame.

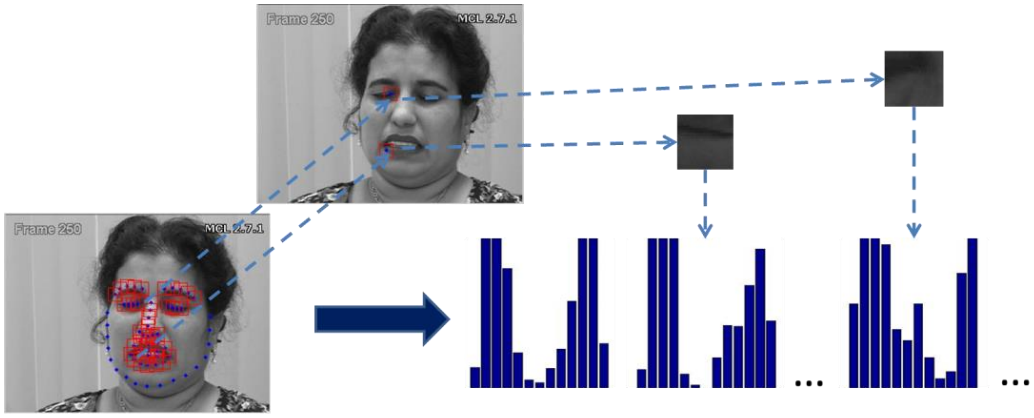


Figure 4-2. Extracting HOG features from the neighborhoods (red rectangles) around the fiducial points (P-HOG). The fiducial points around the face outline are ignored.

In our experiments, each video frame was tracked with 66 facial landmarks. The 49 fiducial points which cover the brows, eyes, nose and mouth were utilized, as shown in Figure 4-2. A 16×16 local patch centered on each fiducial point was cropped and 49 local patches from each frame were obtained. HOG was further used to encode each local patch. The vector length of HOG extracted from a local patch with the default setting is 36 (Dalal and Triggs, 2005). The global feature including the HOGs from 49 local patches is a vector with a length of 36×49 = 1764.

4.2.2 Segment-level Detection

Some previous studies have pointed out that facial expression is a dynamic and contiguous process (Koelstra et al., 2010; Scherer and Ekman, 1982). It means that in a video sequence, pain frames and no-pain frames are clustered by themselves. Figure 4-3 shows the ground truth frame labels in a video sequence. It can be seen that there are three pain events in this video sequence, although the duration of each pain event is different. The pain frames of each pain event are contiguous and the no-pain frames between pain events are also contiguous. It inspires us to consider segment-level detection. Each long video sequence can be partitioned into many non-overlap segments with each segment containing a set of contiguous frames. We can locate the clustered pain/no-pain frames by classifying each segment.

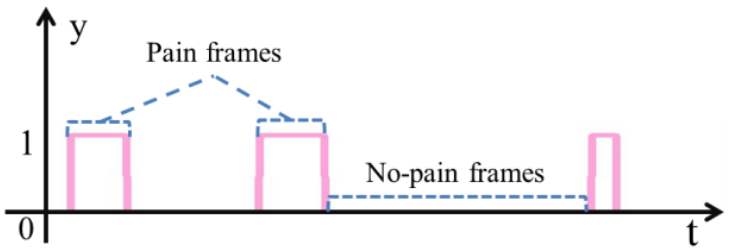


Figure 4-3. The ground truth frame labels in a video sequence. We can find the pain/no-pain frames are generally contiguous and clustered.

In this work, HOG from Three Orthogonal Planes (HOG-TOP) is applied to extract the dynamic textures from segments, as I have described in Section 3.2.1 of

Chapter 3.

A block-based method is also applied in this study. The video segment is divided into a number of block volumes and HOG-TOP features are extracted from each block volume. The HOG-TOP features of all the block volumes are concatenated together to represent the whole video segment. In experiments, the face image is first cropped from the original image and resized to 64×64. The face image is partitioned into 8×8 blocks with each block with size of 8×8. The number of bins is set to 9 with an angle range of $0° - 180°$. Each block can generate a HOG-TOP with a dimension of 3×9=27. All the HOG-TOP features of the 8×8 blocks are concatenated into a long feature vector with a dimension of 3×9×8×8 = 1728.

After feature extraction, an SVM was trained for segment-level detection. In our experiment, a segment contains at least one pain frame is labeled as a positive instance (pain segment) and a segment which contains only no-pain frames is labeled as a negative instance (no-pain segment).

4.2.3 Sequence-level Detection

In order to detect whether there exist pain events in a video sequence, a sequence-level detection method which is based on multiple-feature fusion is proposed. At first, the max-pooling strategy is adopted to transform the frame-level feature (P-HOG) and segment-level feature (HOG-TOP) to a global P-HOG and a

global HOG-TOP, respectively. After that, multiple kernel fusion is applied to find an optimal combination of the two kinds of features. Finally, an SVM with multiple kernels is trained to perform the classification.

Given a long video sequence with N frames segmented into M segments, after feature extraction, we can obtain N P-HOG features and M HOG-TOP features. Here, the max-pooling strategy is used to get the global P-HOG and global HOG-TOP. Suppose that we have a set of features $\mathbf{S} = \{\mathbf{x}_i \in \mathbb{R}^D | i = 1,2, \dots N\}$, and $\mathbf{x}_i$ is a D-dimension feature vector, i.e. $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots x_{iD})$. Here I denote the final feature vector after the max pooling as $\mathbf{F}$. Then the elements in $\mathbf{F}$ should satisfy

$$F_j = \max_{i=1,2,\dots N} x_{ij} \qquad j = 1,2, \dots D \tag{4-1}$$

where F_j is the j-th element of $\mathbf{F}$ and x_{ij} is the j-th element of $\mathbf{x}_i$. Eq. (4-1) shows that each element in $\mathbf{F}$ is the maximum value of all the corresponding elements in the feature set $\mathbf{S}$, where $\mathbf{S}$ is a $N \times D$ matrix, with N being the number of feature vectors and D the dimensionality of each feature vector. Then $\mathbf{F}$ contains the maximum value of each column in $\mathbf{S}$. Max pooling transforms a set of feature vectors to a global feature vector, which is used to represent the whole video sequence.

After obtaining the global features, the next thing is to train a classifier to perform the detection. SVM is a widely used classification model. However, a

traditional SVM with a single kernel is not efficient to handle the training problem of multiple features. Recently, multiple kernel fusion has attracted a growing attention. Previous works have showed that multiple kernels can enhance the discriminative power of the SVMs (Chen et al., 2014; Karan Sikka et al., 2013).

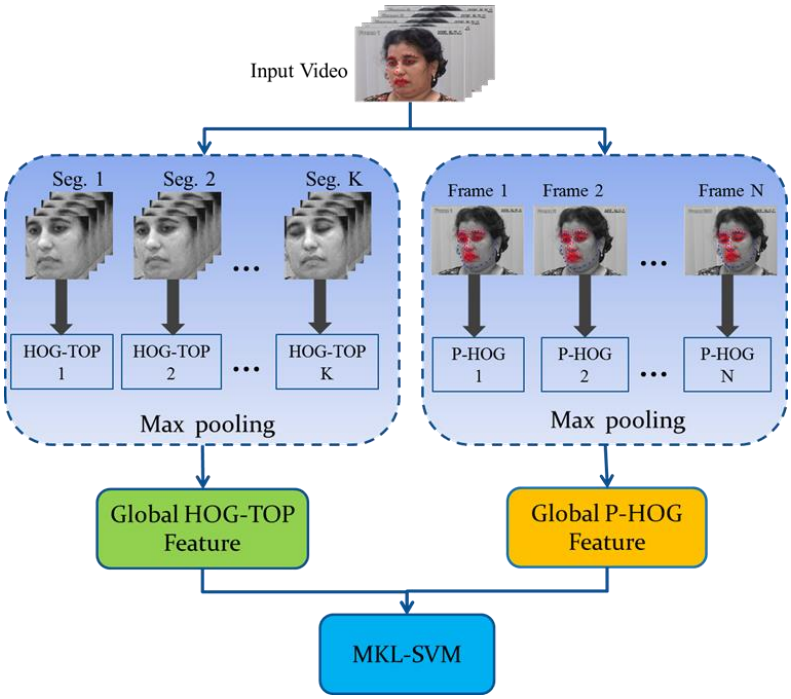


Figure 4-4. The flowchart of sequence-level detection with a multiple kernel SVM.

In this study, there are two types of features: global P-HOG and global HOG-TOP. Multiple kernel fusion is applied to find an optimal fusion of the two types of features, as I have described in Section 3.2.4 of Chapter 3. The flow chart of the framework is shown in Figure 4-4. I adopt the formulation proposed in (Rakotomamonjy et al., 2008) in which the kernel is actually a convex combination of several basis kernels. Define the global P-HOG as $\mathbf{x}$ and global HOG-TOP as $\mathbf{z}$,

then we have

$$K_{ij} = \beta_1 k_1(\mathbf{x}_i, \mathbf{x}_j) + \beta_2 k_2(\mathbf{z}_i, \mathbf{z}_j) \quad (4\text{-}2)$$

with $\beta_1, \beta_2 > 0,\ \beta_1 + \beta_2 = 1$.

The basis kernels can be linear kernels, radial basis function (RBF) kernels and polynomial kernels, etc. In this study, I use a linear kernel for each type of features and adopt the grid search with LIBSVM (Chang and Lin, 2011) to learn the kernel weights β_1, β_2 and coefficients α.

Given a test sample which contains global P-HOG $\mathbf{x}$ and global HOG-TOP $\mathbf{z}$, the label y can be predicted by

$$y = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \left(\beta_1 k_1(\mathbf{x}_i, \mathbf{x}) + \beta_2 k_2(\mathbf{z}_i, \mathbf{z})\right) + b\right) \quad (4\text{-}3)$$

4.2.4 Probabilistic Fusion of Three Tasks

The proposed framework aims to effectively detect and locate pain events in video. I have illustrated how to apply a sequence-level detection method which incorporated multiple features (global P-HOG and HOG-TOP) and multiple kernel fusion for pain event detection. For pain event locating, a probabilistic fusion method is proposed to integrate three different tasks (frame-level, segment-level and sequence-level detection) to achieve this goal. Figure 4-5 shows the semantic diagram of combining three tasks for pain event locating. Frame-level detection can detect the pain presence/absence of each individual frame. Segment-level detection can obtain

the clustered pain/no-pain frames in a video sequence. And sequence-level detection can eliminate some false positives caused by frame and segment detection in a true negative video sequence. I will show that the combined multiple-task outperforms any individual task, and each task plays a different but vital role for pain event locating.

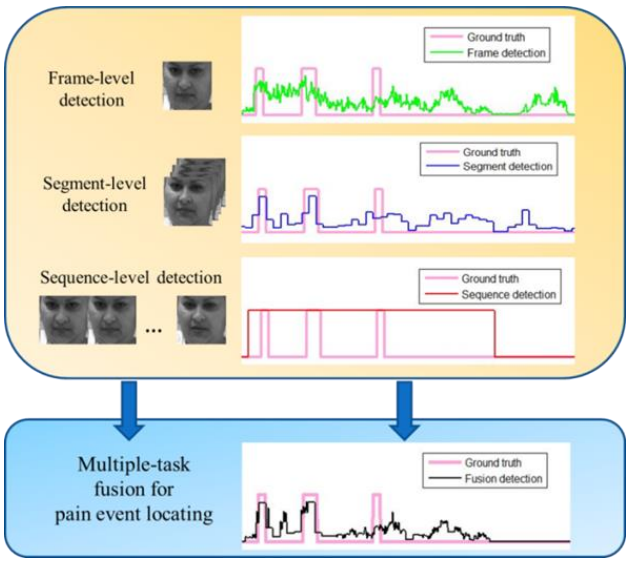


Figure 4-5. Multiple-task (frame, segment and sequence detection) fusion for pain event locating.

The classifiers trained for the three tasks in our work are all SVMs. The output of an SVM is a decision value obtained by a linear function, $\delta = \mathbf{wx} + \mathrm{b}$. The sigmoid function is applied to transform the decision value δ to a probability $\mathrm{p} = 1/(1 + e^{-\delta})$. Note that for the frame-level detection, each frame has a probability; while for the segment-level detection; the frames contained in the same segment share the same probability.

I denote the probability of a frame which is predicted as a pain frame by p_f and

it can be acquired from the frame-level detection. p_s is the probability of a segment which is predicted as a positive segment, meaning that the segment contains at least one pain frame. p_s can be obtained from the segment-level detection. I define the following rule to fuse the frame probability and segment probability:

$$p = \begin{cases} p_f & |p_f - \tau| > |p_s - \tau| \\ p_s & |p_f - \tau| < |p_s - \tau| \end{cases} \quad (4\text{-}4)$$

where τ is a threshold and it is set to 0.5 in our experiment. $|p_f - \tau|$ is the distance between the frame probability and the threshold, while $|p_s - \tau|$ is the distance between the segment probability and the threshold. The distance shows the confidence of a prediction. The greater the distance is, the higher confidence of a prediction is. However, combining the frame-level and segment-level detection might still not be able to achieve a reliable prediction. For instance, considering a true negative video sequence, meaning there are no pain frames in it. If we integrate the frame-level detection and the segment-level detection only, we would most likely see some false positives. In order to eliminate these false positives, I bring in the sequence-level detection.

Suppose that the prediction of sequence-level detection is $y = 1$ when the sequence is predicted as a positive instance and $y = 0$ when the sequence is predicted as a negative instance. The fused frame probability defined in Eq. (4-4) is multiplied by the sequence prediction y. Then, the final frame probability which

combines the three detections is defined as

$$p = \begin{cases} p_f \cdot y & |p_f - \tau| > |p_s - \tau| \\ p_s \cdot y & |p_f - \tau| < |p_s - \tau| \end{cases} \quad (4\text{-}5)$$

Compared with Eq. (4-4), Eq. (4-5) brings in the sequence-level prediction which can wipe out the false positives caused by the frame-level and segment-level detection in a true negative video sequence.

4.3 Experiments and Discussion

4.3.1 Data Sets

In order to evaluate the method, I conducted the experiments on the UNBC-McMaster shoulder pain dataset (Lucey, Cohn, Prkachin, et al., 2011). This dataset includes 200 sequences from 25 subjects. Each subject was suffered from some kind of shoulder pain and was requested to make some passive or active movements. Active tests were performed with the patient in a standing position. Passive tests were performed with the help of a physiotherapist. More details about the dataset can be found in (Lucey, Cohn, Prkachin, et al., 2011).

The facial expressions recorded in the dataset are spontaneous with head movements. Each video frame provides 66 facial landmarks tracked by using the active appearance model. The dataset provides two kinds of labels: frame-level label and sequence-level label. The frame-level label is called Prkachin and Solomon pain

intensity (PSPI) metric which is a sum of the intensities of certain facial action units from Facial Action Coding System (FACS) (Essa and Pentland, 1997). The PSPI (with a range from 0 to 16) can describe the pain intensity of each frame. My work focuses on detecting pain presence/absence on each frame. A frame with PSPI greater than 1 is considered as a pain frame; otherwise the frame is a no-pain frame. This database also provides the sequence-level label called Observer Pain Intensity (OPI) rating that categorizes each sequence into one of the six intensities from 0 (no-pain) to 5 (strong pain). Following the protocol proposed in (Lucey et al., 2008; K. Sikka et al., 2014), all the video sequences are classified into "pain" and "no pain" in our study. When $\mathrm{OPI} \geq 3$, the sequence is a positive instance (pain) and when $\mathrm{OPI} = 0$, the sequence is a negative instance (no-pain). The video sequences with a pain intensity of 1 or 2 were removed. I consider both of frame-level and sequence-level labels and select 139 video sequences with 50 positive instances and 89 negative instances for the experiments. Table 4-1 shows the distributions of positive video sequences and negative video sequences. A positive video sequence has its OPI of greater than or equal to 3 and there at least exist some pain frames in the video sequence. On the other hand, a negative sequence contains no-pain frames with an OPI equal to 0.

Table 4-1. The description of positive and negative sequences.

Positive sequence	(OPI=3,4,5) Containing some pain frames	50 sequences
Negative sequence	(OPI=0) Containing no-pain frames only	89 sequences

4.3.2 Pain Event Detection

I first evaluate my method for pain event detection. As mentioned above, two types of features are employed: P-HOG and HOG-TOP. HOG features of each frame were extracted from local patches around the facial landmarks provided by the dataset. In order to compute HOG-TOP, the facial landmarks were used to crop the face region from each frame and the face region was resized to 64×64. After that, the face sequences were separated into a number of non-overlapping fixed length segments and HOG-TOP feature descriptor were used to extract the dynamic textures from each segment. The max pooling was further used to acquire the global features from the feature set.

In this experiment, I took the leave-one-subject-out cross validation strategy. The video instances from one subject were used for testing and the video instances from the other subjects were used for training. In each cross validation, there was no overlapping between the subjects in the training and test data. Since there are 25 subjects, I carried out 25 cross validations. I followed the strategy employed in

(Lucey et al., 2008; K. Sikka et al., 2014) and used the overall classification rate for performance evaluation.

At first, the global P-HOG and HOG-TOP were applied individually. An SVM with a linear kernel was trained for the classification. The overall classification accuracy acquired by using P-HOG is 87.1%. For HOG-TOP, different segment lengths were set to explore the performance of HOG-TOP under different scales. Table 4-2 shows the overall classification rates obtained by applying HOG-TOP with different scales. S1-10 is the scale setting 1 with a segment length of 10 (10 frames in each segment). The similar interpretation can be given to S2-15 and S3-20. Since P-HOG is the frame-level feature, the segment length does not affect P-HOG. The performance of P-HOG is the same in all of the three settings. Experimental results show that global P-HOG can achieve a higher accuracy than global HOG-TOP. Since P-HOG extracts features from specific regions, it is more effective to characterize subtle facial appearance changes.

Table 4-2. Accuracy obtained by using individual and combined feature sets (%).

Name-Segment length	HOG-TOP	P-HOG	Hybrid Feature
S1-10	82.0	87.1	91.4
S2-15	79.9	87.1	91.4
S3-20	83.4	87.1	90.6

Multiple kernel fusion was further applied to combine the P-HOG and

HOG-TOP features optimally and trained an SVM with multiple kernels to perform the classification. I compare the results with P-HOG and HOG-TOP features applied individually. Table 4-2 demonstrates that the hybrid feature outperforms the individual features, with an improvement of around 5%, meaning that P-HOG and HOG-TOP can make different contributions to the classification and multiple kernel fusion can further enhance the discriminative power of an SVM.

Table 4-3. A comparison of our method with other methods for pain event detection in video (MS-Multiple Segments).

Methods	**Accuracy (%)**	**Subjects-samples**
Lucey et al.	81.0	20-142
Ashraf et al.	68.3	20-142
BoW + Max + SVM	81.5	23-147
MS-MIL	83.7	23-147
Our method-S1-10	**91.4**	25-139
Our method-S2-15	**91.4**	25-139
Our method-S3-20	**90.6**	25-139

I further compare our method with the other methods. The results are shown in Table 4-3. The best performance in our method is S2-15 with a classification rate of 91.4%. The improvement is significant compared with the methods reported in (Ashraf et al., 2009) (68.3%) and (Lucey et al., 2008) (81.0%). The methods reported in (Ashraf et al., 2009; Lucey et al., 2008) used frame-based detection. Experimental results shows that a sequence based detection method can achieve a better

performance than a frame based detection method. In addition, the algorithms reported in (Wang et al., 2012) employed BoW to encode each frame and adopted the max pooling strategy to obtain a global feature from all the frames of a video sequence, and an SVM was trained for classification. The performance (81.5%) is comparable with HOG-TOP applied individually in our method, but is not as good as when the hybrid feature is used in our method. It shows that multiple features can make different contributions and achieve better performance. Note that Multiple Instance Learning (MIL) was applied in (K. Sikka et al., 2014). Our best performance (91.4%) is better than that reported in (K. Sikka et al., 2014) (83.7%), with an improvement of about 8%. Although the classification accuracy in our method is affected by the segment length, as shown in Table 4-3, I have shown that even the lowest accuracy (90.6%) achieved in our method is still better than the other methods.

4.3.3 Locating Pain Events

Locating pain events focus on predicting the pain presence/absence at the frame level. Like pain event detection, I also took the leave-one-subject-out cross validation strategy. In our experiment, two evaluation metrics used in (K. Sikka et al., 2014) were utilized: classification accuracy and maximum F1-score. F1-socre is defined as $\mathrm{F1} = 2R \cdot P/(R + P)$, known to give a trade-off between recall ($\mathrm{R} = \frac{TP}{TP+FN}$) and precision ($\mathrm{P} = \frac{TP}{TP+FP}$), where TP is the true positive. In this experiment, it means

that a pain predicted was actually labeled in the test face image. FP is the false positive, meaning that a pain predicted does not exist in the test face image. FN is the false negative, meaning when a pain appears in the test face image but it is missed in the prediction. The dataset provides a PSPI with a range of 0~16 to indicate the pain intensity for each frame. My study attempts to detect pain/no-pain frames. I first transformed the PSPI to a binary label. In my experiment, a frame is labeled as a pain frame when PSPI >= 2; otherwise, it is a no-pain frame.

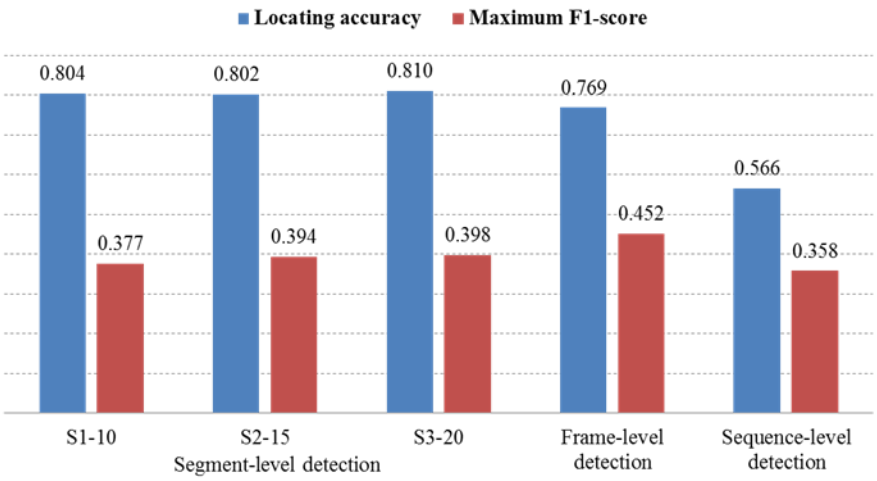


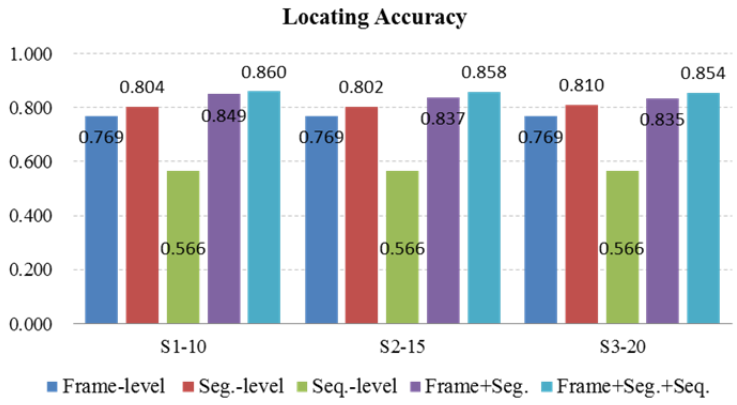
Figure 4-6. Performance of pain event locating obtained by three different tasks.

For frame-level detection, P-HOG was extracted from each frame and an SVM was trained to perform the binary classification. The locating accuracy is 76.9% and the maximum F1-socre is 0.452. For the segment-level detection, HOG-TOP was applied to extract the dynamic textures from each segment and an SVM was trained to perform the classification. I set different segment lengths to explore the performance

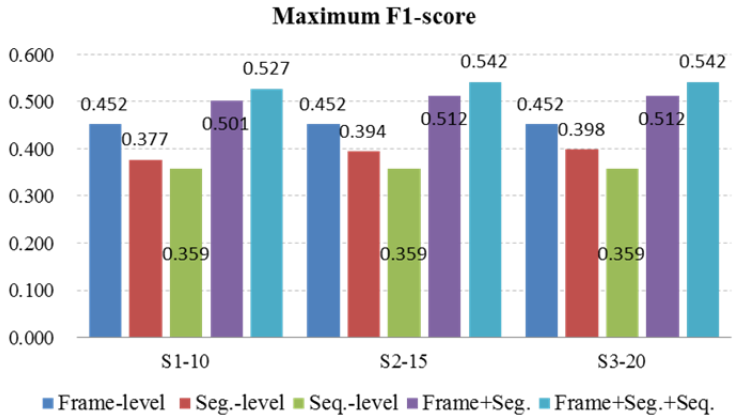
of segment-level detection at different scales. Figure 4-6 shows the locating accuracy and the maximum F1-score of segment-level detection with different scales. Experimental results show that there is no much difference on the performance under different scales. Compared with the frame-level detection, the segment-level detection achieves a higher locating accuracy while obtaining a lower maximum F1-score.

Figure 4-6 also shows the performance of pain event locating based on the sequence-level detection. In this case, all the frames contained in a video sequence share the same prediction result. The locating accuracy and the maximum F1-score are 56.9% and 0.358, respectively. Since there are a small amount of pain frames in a positive video sequence, when all the frames share the same prediction output, there are too many false positives which make the sequence-level detection is not as good as the frame-level detection.

From the experimental results of three individual detections, I realize that information with various time scales (frame-level, segment-level and sequence-level) can make different contributions. Each piece of information is complementary with one another, which inspires us to combine three detection methods to enhance the paint event locating performance.



(a)

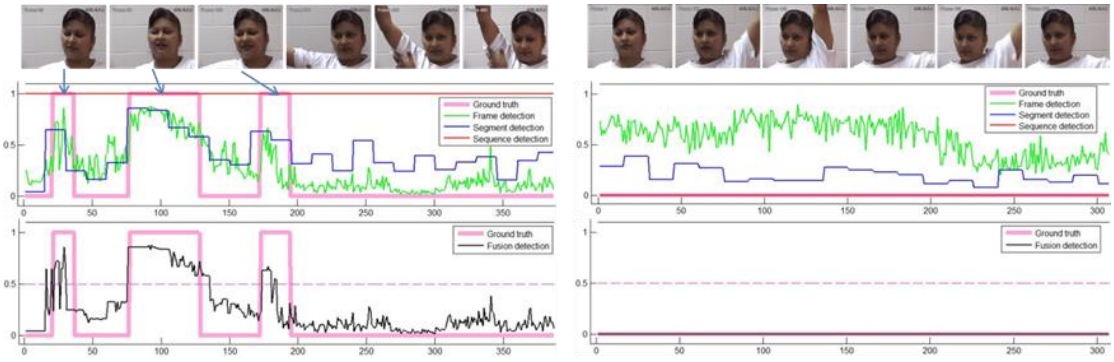


(b)

Figure 4-7. Performance of three different detection methods and the combined detection methods. (a) Locating accuracy; (b) Maximum F1-score.

Three detection methods (frame-level, segment-level and sequence-level detection) are combined by applying Eq. (4-5). Experimental results are shown in Figure 4-7. The segment length is only meaningful to segment-level detection. It is obvious to find that the combined method outperforms individual detection methods. It also can be seen that a combination of three detection methods performs better than a combination of two detection methods (frame-level and segment-level detection),

especially for the maximum F1-score, with an improvement of about 3%. Bringing in the sequence-level detection can eliminate some false positives in a negative video sequence. Figure 4-8 shows the pain event locating results in a positive video sequence and a negative video sequence, respectively. We can see that three different tasks can complement with one another and integrating the three tasks can achieve a better performance. Figure 4-8 (a) shows that some false positives generated by the frame-level detection method can be wiped out when I bring in the segment-level detection method. Figure 4-8 (b) also illustrates that the sequence-level detection method can eliminate some false positives generated by the frame-level detection method on a negative video sequence.



(a) (b)

Figure 4-8. Locating pain events in a positive video sequence (a) and a negative video sequence (b). (Top) The first three frames are frames 20, 100 and 180 which are pain frames and the other no-pain frames. (Middle) The ground truth and the prediction results made by three individual tasks. (Bottom) The ground truth and the fusion results by integrating three different tasks.

I also compare our method with some other methods. There are little previous work trying to tackle both pain detection and locating tasks except for the study reported in (K. Sikka et al., 2014) in which Sikka et al. proposed a method called MS-MIL to jointly detect and locate pain events in video. They also designed an MS-SVM$_{max}$ method for the comparison purpose. I compare my method with MS-MIL and MS-SVM$_{max}$ in Table 4-4. The detection accuracy measures the performance of video based pain event detection as discussed in Section 4.3.2. Locating accuracy and maximum F1-score demonstrate the performance of pain event locating. We can find that the locating accuracy in our method is much higher than MS-SVM$_{max}$ and MS-MIL reported in (K. Sikka et al., 2014), with an improvement of about 13% and 10%, respectively. The maximum F1-score is also higher than MS-SVM$_{max}$ and slightly better than MS-MIL. Although the number of samples in our experiments is slightly less than that in (K. Sikka et al., 2014), it is reasonable to conclude that our method compares favorably with MS-SVM$_{max}$ and MS-MIL for pain event detection. For pain event locating, our method can achieve a comparable maximum F1-score as MS-MIL while obtaining a much higher locating accuracy.

Table 4-4. A comparison of our method with some other methods for joint pain event detection and locating in video.

Method	Location Acc.	Max-F1	Detection Acc.	Subjects-Samples
MS-SVM$_{max}$	72.64%	0.471	77.17%	23-147
MS-MIL	76.08%	0.523	83.70%	23-147
Our method-S1	85.96%	0.527	91.37%	25-139
Our method-S2	**86.08%**	**0.542**	**91.37%**	25-139
Our method-S3	85.37%	0.542	90.65%	25-139

4.3.4 The Effect of Segment Length

A challenging problem for this work is to set the length of a segment when computing HOG-TOP. In my experiments, I explored the effect of the segment length for pain event detection and locating. Since there is no overlap between the segments, segment length is the only parameter I need to consider. I set different segment lengths with a range from 10 to 30 and evaluated the performance of pain event detection and locating. Figure 4-9 shows the detection accuracy obtained by three different feature sets under different segment scales. P-HOG is a frame feature which is not affected by segment length. The performance of P-HOG is therefore the same under six different segment lengths. On the other hand, HOG-TOP is a kind of dynamic texture feature to characterize spatial-temporal information of an image sequence. The segment length affects the detection performance of HOG-TOP significantly. It can be seen that the highest accuracy achieved is 83.4% when the

segment length is set to 20 (Seg.-20). The performance falls to 76.9% when the segment length is set to 30 (Seg.-30). It suggests that a too long segment for HOG-TOP will reduce discrimination power of the feature and result in a low classification rate. However, when multiple kernel fusion is used to combine P-HOG and HOG-TOP, it can be seen that the performance achieved by different hybrid features is comparable. The gap between the highest accuracy (91.4%) and the lowest accuracy (90.6%) is small.

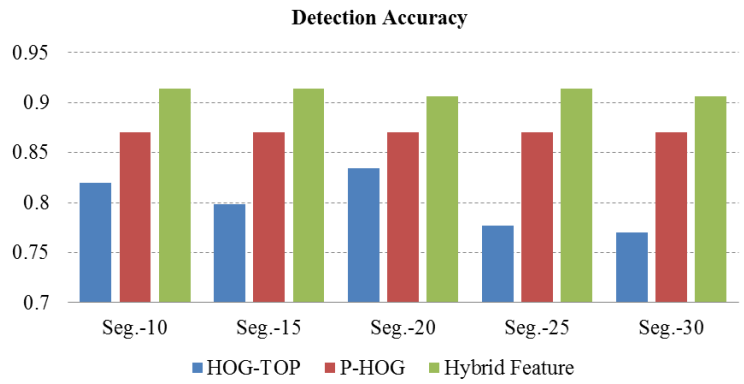
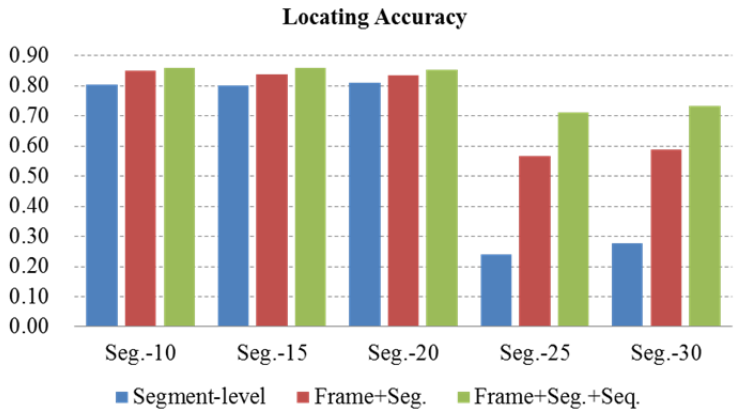
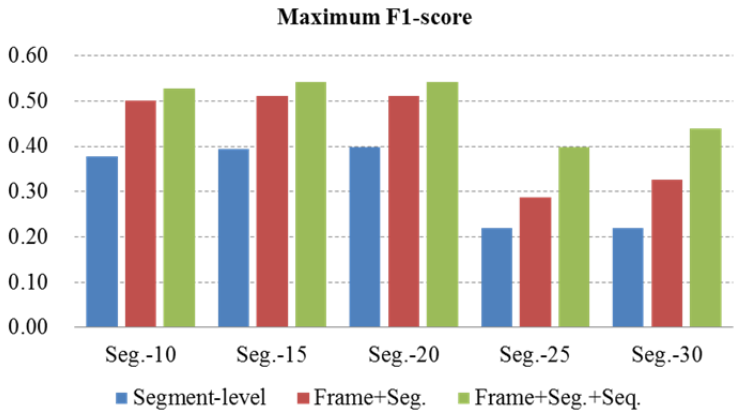


Figure 4-9. The detection accuracy obtained by three different feature sets under different segment scales.

I further analyze the effects of segment length for pain event locating. Figure 4-10 (a) and Figure 4-10 (b) show the locating accuracy and maximum F1-score acquired by the segment-level detection and the combined detection method, respectively. Experimental results demonstrate that when the segment length is set to a large value (e.g. greater than or equal to 25), the locating accuracy and maximum F1-score obtained by the segment-level detection drop quickly.



(a)



(b)

Figure 4-10. The performance of segment-level detection and the combined detection method with different segment lengths. (a) Locating accuracy; (b) Maximum F1-score.

Moreover, when I integrate the three different tasks (frame-level, segment-level and sequence-level detection), since the frame and sequence detection does not rely on the segment length, the performance of the combined detection is only affected by the segment-level detection. It can be seen that with a large segment length, the

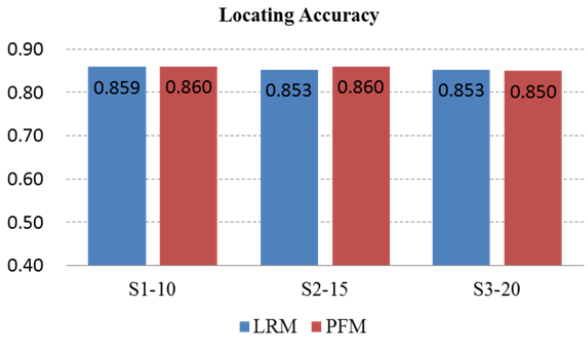
combined detection method cannot perform as well as that with a smaller segment length, suggesting that a too long segment tends to bring in more false positives and also weaken the performance of the combined detection method. In experiments, I determined the segment length by a try-and-error method. I found that a segment length between 10 and 20 frames can achieve a promising performance for both pain detection and locating.

4.3.5 Multi-Task Fusion

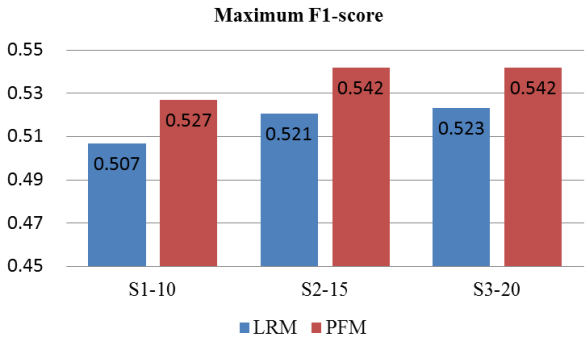
This work proposes a probabilistic fusion method to combine the three tasks. Another widely used technique is linear regression. Suppose that there are three probabilities p_1 (frame level), p_2 (segment level), p_3 (sequence level) from the three tasks. The integrated probability is $w_1p_1 + w_2p_2 + w_3p_3$, where the weights w_i can be trained by the linear regression method. I conducted a comparison study between the linear regression model and our proposed probabilistic fusion model. The comparison results of the two methods are shown in Figure 4-11.

Figure 4-11 shows that with different segment lengths (10, 15 and 20 shown in the figures), the two methods obtain compatible pain event locating accuracies, while the maximum F1-score obtained by our proposed probabilistic fusion model is higher than that by the linear regression model. F1 score is a trade-off between recall accuracy and precision. In this work, the number of negative samples is much greater than that of

positive samples. This is an imbalanced training problem and therefore F1 score is a better measure on the performance of a model. It means that our probabilistic fusion model (PFM) is more robust to combine the three tasks.



(a)



(b)

Figure 4-11. The performance of the linear regression model (LRM) and our probabilistic fusion model (PFM). (a) Locating accuracies; (b) Maximum F1-scores.

4.4 Discussion

Experimental results reported in Section 4.3.2 show the effectiveness of my proposed method in pain event detection. P-HOG can effectively characterize the

spatial appearance of facial expressions. HOG-TOP can characterize the dynamic appearance changes caused by facial activities. In fact, these two types of features capture information from different time scales. Pain, as a kind of facial expression, actually is a dynamic process. These two kinds of information can be used to build a more useful model to characterize this dynamic process. Moreover, in order to take advantages of both types of features, multiple kernel fusion is used to integrate the two features optimally. Multiple features with an optimal combination can achieve a better performance than individual feature.

Experimental results reported in Section 4.3.3 show that a multiple-task fusion based on three different tasks is more robust to deal with pain event locating than any individual task. In the dataset I used, a negative video sequence contains no-pain frames. Even in a positive video sequence, the number of pain frames accounts for a small part of the whole sequence. It means that there are much more negative instances than positive instances in these videos. This is an unbalance training problem. The frame-level and segment-level detection tend to make a large number of false positives. But when I combine the two detection methods, they can complement each other and achieve a better performance. It suggests that the pieces of information from different time scales are complementary and they can be used to reduce false positives. I also observe that the sequence-level detection can help in eliminating

some false positives made by the frame-level and segment-level detection in negative video sequences. The maximum F1-score can be improved slightly when I bring in the sequence-level detection, as shown in Figure 4-7.

4.5 Summary

In this chapter, a novel framework is presented for joint pain event detection and locating in video. I propose to combine three detection tasks, frame-level, segment-level and sequence-level detection, to handle (1) pain event detection which determines whether there exist pain events in a video (a pain/no-pain classification problem) and (2) pain event locating (predicting pain presence/absence in each frame).

For pain event detection, a multiple-feature fusion method which combines spatial feature and spatial temporal feature is presented. Both the static attributes of video frames and dynamic attributes of sequences were explored in this study. P-HOG is applied to extract spatial features from video frames to represent static attributes. HOG-TOP is used to characterize the dynamic textures of video segments. Max pooling strategy is employed to form a global P-HOG and global HOG-TOP to characterize the whole video sequence. Multiple kernel fusion is used to find an optimal combination of these two types of global features. Finally, an trained SVM with multiple kernels is utilized to detect whether there exist any pain events in video.

For pain event locating, a multiple-task fusion method is proposed. I first address three sequential tasks namely frame-level detection, segment-level detection and sequence-level detection. For the frame-level detection, an SVM with P-HOG features was trained to predict pain presence/absence on each frame. Noting that pain frames or no-pain frames are contiguous in a video sequence, I also propose to detect the pain segments of contiguous frames. An SVM was trained with HOG-TOP features extracted from video segments to perform the segment-level detection. At last, the two detection methods were coupled with the sequence-level detection to locate pain events in video. This method utilizes information from different time scales and achieves a promising performance on the public UNBC-McMaster Shoulder Pain dataset.

Chapter 5 Smile Detection in the Wild with Deep Convolutional Neural Network

5.1 Introduction

In this chapter, an effective approach for smile detection in the wild with deep learning is presented. A favorable superiority of deep learning lies in that it not only can perform the classification effectively, but also can learn some high level abstract representations from raw inputs because of the hierarchical multiple layers of a deep learning model. The activations of the hidden layers can be extracted as the learned abstract representations. These representations can be used to train a traditional classifier such as SVM or AdaBoost.

In order to take advantages of deep learning, I apply deep convolutional network, a widely used deep learning model, to handle this problem. This work addressed the following tasks: 1) developing a deep convolutional neural network namely Smile-CNN to perform the smile detection in the wild; 2) feature learning with Smile-CNN; 3) training SVM and AdaBoost with these learned features to evaluate their discriminative power; 4) investigating into the effects of the image background and pose variations to the problem of smile detection in the wild.

5.2 Methodology

5.2.1 Multilayer Perceptron

A feedforward network, also known as MultiLayer Perceptron (MLP), has become a widely used neural network model. An MLP consists of multiple layers of nodes and each layer is fully connected with the next layer. Figure 5-1 shows an MLP with one hidden layer.

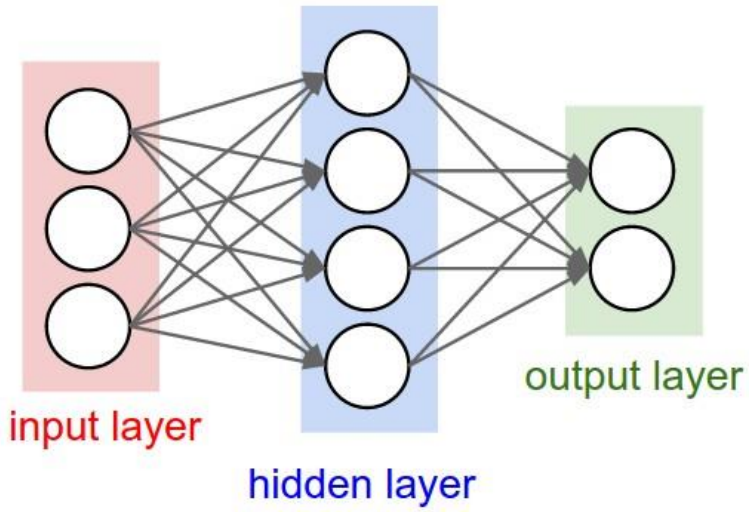


Figure 5-1. An MLP with one hidden layer.

Each node in hidden layer and output layer is normally a neuron with a non-linear activation function. The outputs of each layer can be defined as:

$$\mathbf{x}^l = f(\boldsymbol{u}^l), \text{ with } \boldsymbol{u}^l = \mathbf{w}^l\mathbf{x}^{l-1} + \boldsymbol{b}^l \tag{5-1}$$

where $\mathbf{x}^l$ denotes the output of the l layer, $\mathbf{x}^{l-1}$ indicates the output of the $l-1$ layer, $\mathbf{x}^0$ is the input, $\mathbf{w}^l$ is the connection weights, $\boldsymbol{b}^l$ is the bias, f is the

activation function. The sigmoid function $(f(\mathrm{x}) = 1/(1+e^{-x}))$ is commonly adopted.

MLP can be trained by applying a supervised learning technique called the error back propagation algorithm. Given a training set with N samples, each training sample is categorized into one of K classes and the ground truth label $\mathbf{t}$ is a K-length vector with 0 and 1. For each training sample i, we can obtain the following error:

$$e_i = \frac{1}{2}\sum_{j=1}^{K}(y_j^i - t_j^i)^2 = \frac{1}{2}\left\|\mathbf{y}^i - \mathbf{t}^i\right\|_2^2 \tag{5-2}$$

where e_i is the error, y_j^i is the j-th predicted result, and t_j^i is the j-th ground truth of the i-th sample. We can further define the accumulated error of N training samples:

$$\mathrm{E} = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{1}{2N}\sum_{i=1}^{N}\left\|\mathbf{y}^i - \mathbf{t}^i\right\|_2^2 \tag{5-3}$$

With error back propagation, we have,

$$\boldsymbol{\delta}^l = (\mathbf{w}^{l+1})^T\boldsymbol{\delta}^{l+1} \circ f'(\boldsymbol{u}^l) \tag{5-4}$$

where δ is the "errors" back propagated from output layer, the operator "$\circ$" denotes the element-wise multiplication. For the output layer, the δ takes a slight different form:

$$\boldsymbol{\delta}^L = (\mathbf{y} - \mathbf{t}) \circ f'(\boldsymbol{u}^L) \tag{5-5}$$

where $\mathbf{y}$ is the predicted output vector, $\mathbf{t}$ is the target output vector. $f(\cdot)$ is the activation function. We can easily get the derivative of the sigmoid function as:

$f'(x) = f(x)\big(1 - f(x)\big)$. We can get the update rules for the trainable weights $\mathbf{w}$ and bias $\mathbf{b}$.

$$\frac{\partial E}{\partial \mathbf{w}^l} = \boldsymbol{x}^{l-1}(\boldsymbol{\delta}^l)^T,\ \Delta\mathbf{w}^l = -\eta\frac{\partial E}{\partial \mathbf{w}^l}\ ,\ \mathbf{w}^l = \mathbf{w}^l + \Delta\mathbf{w}^l \tag{5-6}$$

$$\frac{\partial E}{\partial \boldsymbol{b}^l} = \frac{\partial E}{\partial \boldsymbol{u}^l}\frac{\partial \boldsymbol{u}^l}{\partial \boldsymbol{b}^l} = \boldsymbol{\delta}^l,\ \Delta\boldsymbol{b}^l = -\eta\frac{\partial E}{\partial \boldsymbol{b}^l}\ ,\ \boldsymbol{b}^l = \boldsymbol{b}^l + \Delta\boldsymbol{b}^l \tag{5-7}$$

5.2.2 Convolutional Neural Network

MLP takes full connection strategy and treats all the inputs equally. There are two drawbacks for MLP: 1) ignoring local property of images, which is that nearby pixels are more strongly correlated than more distant pixels; 2) full connection tends to produce a quantity of weight parameters which are easy to result in over fitting. In order to address the two issues, local connection and weight sharing are taken into account. Convolutional neural networks (CNNs) are such network models which satisfy these two conditions. CNNs have been successfully applied by LeCun (Yann LeCun et al., 1989) for handwritten digit recognition and has attracted growing attention with the rapid development of deep learning and GPU computing.

There are three principle mechanisms in CNN: 1) local receptive field, 2) weight sharing and 3) subsampling or pooling. Figure 5-2 illustrates part of a CNN (Bishop, 2007).

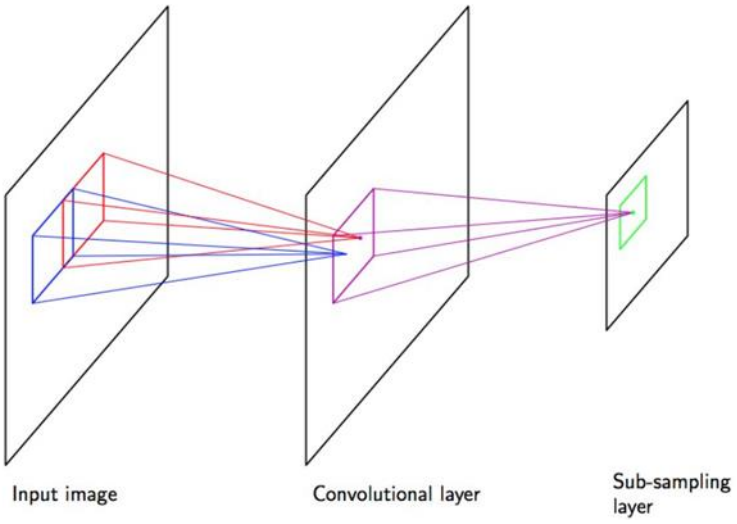


Figure 5-2. The Diagram illustrating part of a convolutional neural network, showing a layer of convolutional units followed by a layer of subsampling units (Bishop, 2007).

In the convolutional layer, the units are organized into planes, each of which is called a feature map. As we can see, the nodes in convolutional layer do not fully connect all the nodes of the input layer but only connect the nodes of a local input patch. This is the core idea of local receptive field. Each node in convolutional layer only covers a sub region of the input layer, which is more effective to exploit local spatial correlation and extract robust local features. Another fascinating property of CNN is weight sharing. Unlike MLP in which the nodes in hidden layers have different connection weights, while in CNN, the nodes belonging to the same feature map in convolutional layer share the same connected weights. There are two advantages for weight sharing. It not only reduces the number of free parameters but also promotes the reuse of local features. Since local features that are useful in one

region of the image are likely to be useful in other regions of the image, shared weights are able to extract similar features from different regions. The third important concept of CNN is subsampling or pooling. Pooling layer often follows convolutional layer and performs down sampling to reduce the feature map size. It partitions the input image into a set of non-overlapping rectangles and gets the outputs from each sub region. Average pooling was widely used before, and recently, max pooling has become the first choice in most practical applications.

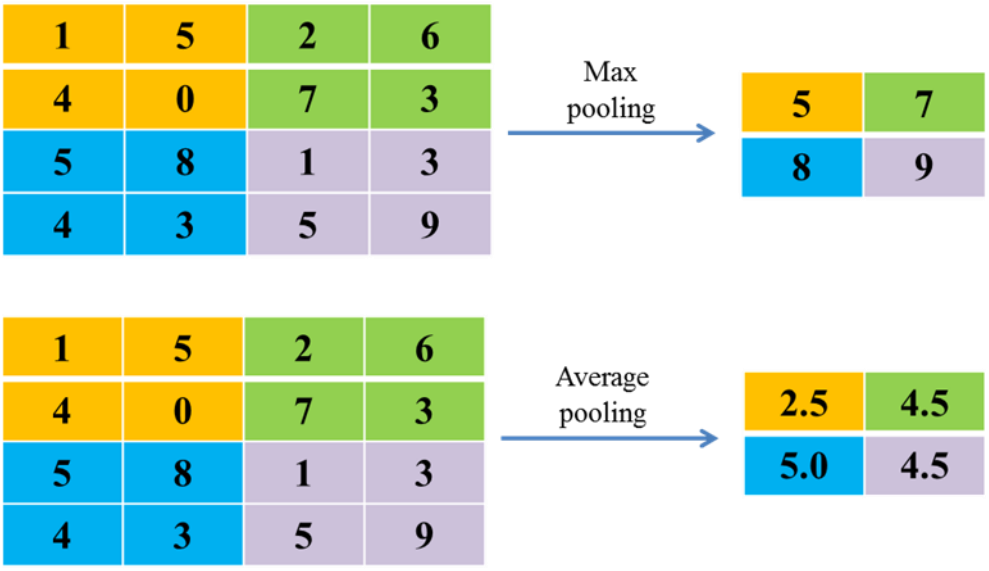


Figure 5-3. The max pooling and average pooling from 2×2 sub region with a stride of 2.

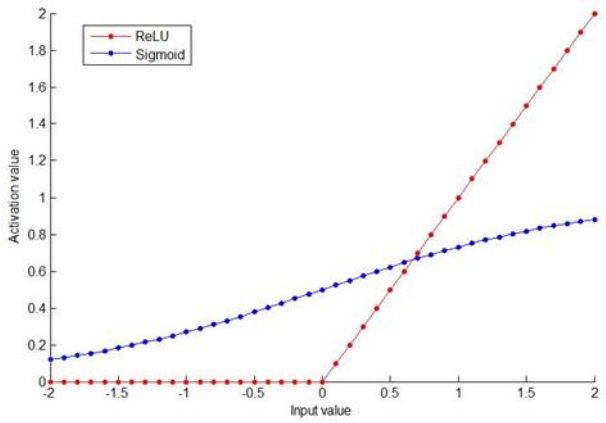
Figure 5-3 illustrates the rules of max pooling and average pooling, respectively. Max pooling outputs the maximum value of each sub region while average pooling outputs the mean value. Pooling layer plays an important role for feature extraction. The objective of pooling is to transform the joint feature representation into a new,

more usable one that preserves important information and discards irrelevant information. It can progressively transform the low level features to high level features.

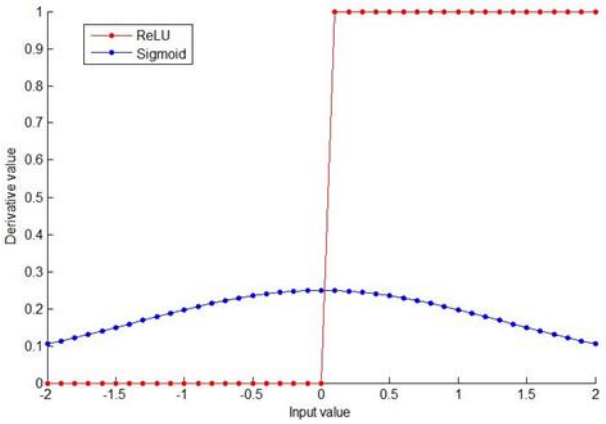
In a practical architecture, there may be several pairs of convolutional and pooling layers. After several pairs of convolutional and pooling layers, the final layer would typically be a fully connected layer to represent the output.

In a CNN, activation functions are generally applied on the units of convolution layer. In the past, sigmoid function was widely used as activation function in hidden layers. Recently, the so-called Rectified Linear Unit (ReLU) has been proposed in (Nair and Hinton, 2010), which has been successfully applied in many deep learning models. The ReLU function is defined as $f(x) = \max(0, x)$, where x is the input to a neuron. ReLU has a number of advantages over sigmoid function. Figure 5-4 shows a comparison of the two functions with different input values. Figure 5-4 (a) shows the activation values and (b) shows the derivative values of the two functions. We can find that ReLU is easier to compute, the output of ReLU is either 0 or the input. In addition, ReLU can suppress negative input, which seems more biologically plausible (Glorot et al., 2011). ReLU is able to produce sparse representation, since the negative inputs become 0 after applying ReLU. Moreover, Figure 5-4 (b) illustrates that the derivative of ReLU is constant. It avoids the gradient vanishing problem during error

back propagation. This is a very useful feature for training deep networks.



(a)



(b)

Figure 5-4. A comparison of the two functions with different input values. (a) the activation values of the two functions; (b) the derivative values of the two functions.

5.2.3 Smile CNN

This work developed a CNN model for smile detection in the wild. Choosing an appropriate CNN architecture relies heavily on experiences. Based on the input size (64×64) and database size (4000 samples), I found that three convolutional layers are

sufficient to handle our classification problem. Once the number of convolutional layers is determined, the suitable filter size is needed to map the input (64×64) to the output (1×2). In order to control the number of free parameters, it also needs to set an appropriate number of filters in each layer. The architecture of Smile-CNN built in this study is a balanced choice between the representation learning ability and computational complexity. The architecture of Smile-CNN model is shown in Figure 5-5. This model maps an input 64×64 image into 2 output nodes with one node indicating "smiling" and the other "non-smiling". Between the input and the output, there are three convolutional layers with each followed by a max-pooling layer. The first convolution layer (C1) filters the 64×64 input image with 16 learnable kernels of size 9×9. The second convolution layer (C3) filters the 16 28×28 output feature maps of P2 layer with 16×8 kernels of size 5×5. The third convolution layer (C5) filters the 8 12×12 output feature maps of P4 layer with 8×16 kernels of size 5×5.

The role of these layers is to extract the features hierarchically. The hidden nodes of each layer are referred to as feature maps or output maps. The convolutional layers extract the features from the input by applying a number of learnable filters or kernels sliding across the input image. The convolution operation can be expressed as

$$\mathbf{x}_j^l = f(\sum_{i\in\Omega_j} \mathbf{x}_i^{l-1} * k_{ij}^l + b_j^l) \tag{5-8}$$

where $\mathbf{x}_i^{l-1}$ and $\mathbf{x}_j^l$ are the i-th input feature map of layer $(l-1)$ and j-th output

feature map of layer l, respectively. Ω_j represents a set of input feature maps. k_{ij}^l is the convolutional kernel which connects the i-th and j-th feature maps. b_j^l is the bias of the j-th output feature map. $f(\cdot)$ is the activation function which performs the non-linear transformation. I apply the ReLU as the activation function.

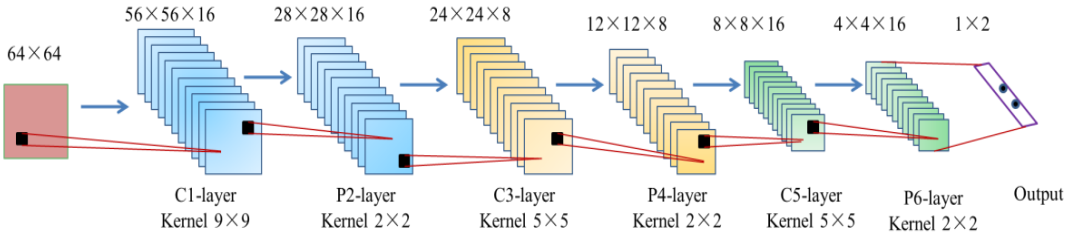


Figure 5-5. The Smile-CNN applied in our study. The input is a gray image with size 64×64. The C1, C3 and C5 are the convolutional layers while P2, P4 and P6 are the max-pooling layers.

Each convolutional layer is followed by a pooling layer, which is used to reduce the spatial size of representation and control overfitting. The pooling layer takes small square blocks ($s \times s$) from the convolutional layer and subsamples it to produce a single output from each block. The most common pooling form is average pooling or max pooling. Here I employ the max pooling strategy, which can be formulated as

$$y_{j,k}^i = \max_{0 \le m,n \le s} \{x_{j \cdot s+m, k \cdot s+n}^i\} z \tag{5-9}$$

where i indicates the feature map of the previous convolutional layer. This expression simply takes an $s \times s$ region and outputs a single value, i.e. $y_{j,k}^i$, which is the maximum value in that region. This operation reduces an $N \times N$ input map to an

$\frac{N}{s} \times \frac{N}{s}$ output map. In this study, I set the block size s to 2.

Following several convolution and max pooling layers, the final output is a 1-by-2 vector with one node indicating "smiling" and the other "non-smiling". Each neuron of the output layer fully connects to the nodes from the previous hidden layer. Let $\mathbf{x}$ denotes the output of the last hidden layer nodes, $\mathbf{w}$ is the connected weights between the last hidden layer and output layer. The output is defined as $\mathbf{f} = \mathbf{w}^T\mathbf{x}+\mathbf{b}$. The output is fed to a 2-way softmax which produces a distribution over the two class labels. We then have

$$p_k = \frac{\exp(f_k)}{\sum_{j=1}^{2} \exp(f_j)} \tag{5-10}$$

where p_k indicates the probability of the k-th class and $\sum_{k=1}^{2} p_k = 1$. In general, the deep convolutional network model is trained by minimizing the cross-entropy loss:

$$L = -\sum_{k=1}^{2} y_k \log(p_k) \tag{5-11}$$

Finally, the predicted class would be $\tilde{k} = arg \max_i p_i = arg \max_i f_i$. I trained the Smile-CNN using stochastic gradient descent with a batch size of 50 samples. I followed the method in (Krizhevsky et al., 2012) and adopted a momentum of 0.9 and a weight decay of 0.0005 to update the weights. The update rule for weight w is defined as

$$v_{i+1} = 0.9 \cdot v_i - 0.0005 \cdot \alpha \cdot w_i - \alpha \frac{\partial L}{\partial w} \tag{5-12}$$

$$w_{i+1} = w_i + v_{i+1} \tag{5-13}$$

where i is the iteration index and α is the learning rate.

Note that a deep convolutional network contains a large number of hidden nodes and free parameters. A serious problem is the overfitting during training, especially when the dataset is not large enough. In order to lessen the over-fitting problem, the dropout strategy which has been shown as an efficient method is adapted to handle this problem (Srivastava et al., 2014). Dropout is a simple but effective way to prevent the overfitting during training. The term "dropout" refers to dropping out some units in the neural network during the training process and the choice of which unit to drop is random. In general, we can set a fixed probability p for each unit to be reserved or dropped out. In my experiments, the probability p was set to 0.5. It is not necessary to drop out the units from all the hidden layers and therefore I just randomly dropped out the units of C5-layer (the last convolutional layer) with a probability of 0.5.

In this work, the MatConvNet toolbox (Vedaldi and Lenc, 2015) was used to construct and train the Smile-CNN. MatConvNet toolbox is a MATLAB toolbox implementing Convolutional Neural Networks (CNNs) for computer vision applications.

As mentioned above, an attractive advantage of deep learning is that it not only

performs the classification, but also learns some abstract representations from the raw data. These learned abstract representations generally lie in a low dimensional space and have good intra-class similarity and inter-class diversity. Some research works, e.g. those reported in (F. J. Huang and LeCun, 2006; Lee et al., 2009), have demonstrated that SVM trained from the features learned by CNN usually can improve the classification performance.

In my work, the outputs of the last hidden layer, i.e., P6-layer in Figure 5-5 are extracted as the learned features. The input size is 64-by-64 and P6 layer contains 16 feature maps with size 4×4. The 16 feature maps are reorganized as a 256×1 column vector.

5.2.4 Classification

In this study, I further applied the SVM to explore the discriminative power of the learned features. SVM is a popular machine learning method for classification. It has been widely used in various pattern recognition tasks. It is believed that SVM can achieve a near optimum separation among classes. Given a training set with labeled samples: $(\mathbf{x}_n, y_n), \mathrm{n} = 1,2, \ldots \mathrm{N}$, $\mathbf{x}_n \in \mathbb{R}^D$, $y_n \in \{-1, +1\}$, SVM tries to find the maximum margin between data points of different classes by solving the following constrained optimization:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.} \qquad y_i(\boldsymbol{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \tag{5-14}$$

$$\xi_i \geq 0, i = 1, \ldots N$$

where ξ_i are slack variables and b is the bias. It needs to obtain the weights $\boldsymbol{w}$ and bias b. Given a unseen input $\boldsymbol{x}$, the predicted label y is computed as

$$\text{y} = \text{sgn}(\boldsymbol{w}^T\mathbf{x} + b) \tag{5-15}$$

In addition, another widely used machine learning algorithm, i.e. Adaptive Boosting, short for AdaBoost (Freund and Schapire, 1997) was also applied in this work to evaluate the learned features. Boosting is an approach based on the idea of combining many relative weak learners to create a strong classifier and achieve better performance. The core concept lies in that instead of pursuing a learning algorithm that is accurate over the entire instance space, it can focus on finding weak learning algorithms and combining them optimally.

Given a training set $D = (\boldsymbol{x}_1, y_1) \ldots (\boldsymbol{x}_m, y_m)$, where $\boldsymbol{x}_i \in \text{X}$, $y_i \in \text{Y} = \{-1, +1\}$, AdaBoost tries to maintain a distribution or a set of weights over the training set. The weak learning algorithm is first used to find a weak hypothesis $h_t: X \to (-1, +1)$. The final output hypothesis is weighted majority vote of the weak hypotheses:

$$H(x) = \mathrm{sgn}(\sum_{t=1}^{T} \alpha_t h_t) \quad (5\text{-}16)$$

where T is the number of the weak hypotheses and α_t is the weight assigned to h_t.

5.3 Experiments and Discussions

5.3.1 Database

I conducted the experiments on the GENKI4K (Whitehill et al., 2009) database to evaluate the method. The database consists of 4000 images taken in the real world. Figure 5-6 shows examples of face images in the database. It can be seen that these images are diverse in illumination, pose and background. The pose range (yaw, pitch, and roll parameters of the head) of most images was within approximately $\pm 20°$ of frontal. In addition, the images span a wide range of imaging conditions like age, gender, facial hair, and glasses (Whitehill et al., 2009). All the images were manually labeled. There are 2162 “smiling” faces and 1838 “non-smiling” faces in total.

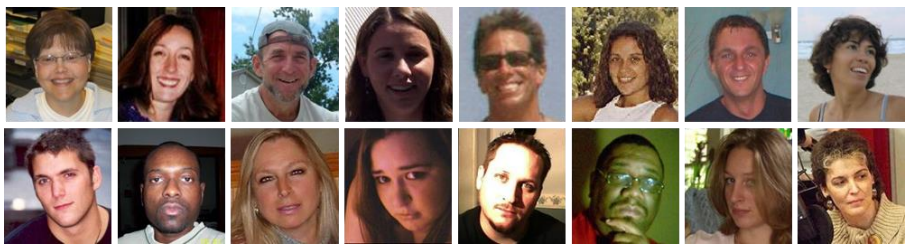


Figure 5-6. Examples of face images from GENKI4K. Top: smiling face images; Bottom: non-smiling face images.

I followed the experiment setting reported in (Whitehill et al., 2009). All the face

images were first converted to gray images and the face detector (Z. Liu et al., 2015) was applied to detect the faces. The faces were normalized to reach a canonical face of 64×64 pixels. Figure 5-7 shows some of the normalized faces.

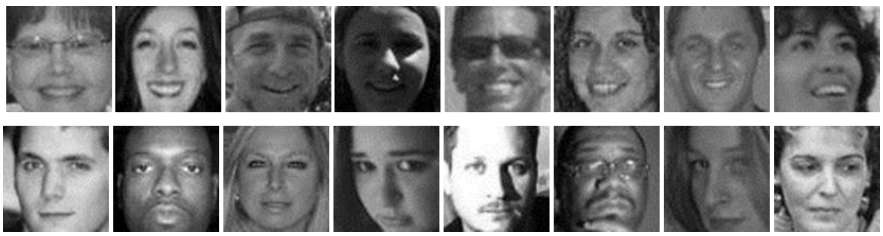


Figure 5-7. Examples of the normalized faces.

The dataset is randomly divided into 4 subsets with each subset having 1000 samples. The numbers of "smiling" and "non-smiling" face images in each subset are shown in Table 5-1.

Table 5-1. The distribution of "smiling" and "non-smiling" face images in each subset.

Subset	1	2	3	4
"smiling" faces	540	541	540	541
"non-smiling" faces	460	459	460	459

5.3.2 Smile Detection

5.3.2.1 MLP versus Smile-CNN

I first evaluated MLP and Smile-CNN on the database and made a comparison of

the two networks, respectively. The input for the Smile-CNN is a 64×64 image, while for MLP is a 1×4096 vector. The Smile-CNN has three pairs of convolutional and pooling layers. I also set three hidden layers in MLP. Table 5-2 illustrates the overview of the two networks. Note that pooling layers of Smile-CNN do not have trainable parameters. Both MLP and Smile-CNN apply ReLU as activation functions and the cross-entropy loss is computed as loss function.

Table 5-2. The overview of Smile-CNN and MLP applied in our work.

Smile-CNN			MLP		
Input	64×64	No. of parameters	Input	1×4096	No. of parameters
Conv. 1	16@9×9	1312	Hidden 1	750	3072750
Pool 2	16@2×2	0			
Conv. 3	8@5×5	3208	Hidden 2	150	112650
Pool 4	8@2×2	0			
Conv. 5	16@5×5	3216	Hidden 3	25	3775
Pool 6	16@2×2	0			
Output	1×2	514	Output	1×2	52

Table 5-2 illustrates that the number of trainable parameters in Smile-CNN is much less than that in MLP. Take the first hidden layer as an example. The first convolutional layer includes 16 9×9 filters, the number of trainable parameters of the first convolutional layer is 16×9×9+16=1312. However, the first hidden layer with 750 nodes in MLP have 4096×750+750=3072750 trainable parameters, which is extremely high. It means that MLP is not efficient to deal with large scale images.

MLP tends to produce a large quantity of free parameters.

Table 5-3. The accuracy obtained by the MLPs with different numbers of hidden layers and Smile-CNN (%) (MLP-1: an MLP with one hidden layer; MLP-2: an MLP with two hidden layers; MLP-3: an MLP with three hidden layers)

	MLP-1	**MLP-2**	**MLP-3**	**CNN**
Fold 1	86.9	85.2	85.4	91.6
Fold 2	86.7	86.1	87.6	91.1
Fold 3	85.8	88.4	87.1	92.5
Fold 4	84.7	85.8	87.9	93.0
Avg.	86.0	86.4	87.0	92.1

I further compare the performance acquired by the two neural networks. Table 5-3 shows the accuracy obtained by the MLPs with different numbers of hidden layers and Smile-CNN. Experimental results show that Smile-CNN achieves higher accuracy than an MLP at each fold. The average accuracy acquired by Smile-CNN is 92.1%, with an improvement of about 5%, compared with an MLP with the best average accuracy of 87.0%. Since the two networks use the same database, activation function and loss objective function, it means that architecture plays an important role for improving recognition ability of Smile-CNN. Compared with an MLP, Smile-CNN includes much more hidden nodes while including much less trainable weights. With a local receptive field, these hidden units in convolutional layer can fully exploit correlations of adjacent pixels. Weight sharing promotes the reuse of features. Deep architectures can potentially lead to progressively more abstract features at higher layers of representations.

5.3.2.2 Learned Features versus Hand-crafted Features

As I have discussed above, the convolutional layers can extract the features hierarchically. I investigated into the deep convolutional network and tried to explore the features learned by this model. Figure 5-8 shows some feature maps of each layer in the Smile-CNN model. There are 16 feature maps with size 56×56 in C1-layer and P2-layer consists of 16 feature maps with size 28×28. C3-layer includes 8 feature maps with size 24×24 and P4-layer contains 8 features maps with size 12×12. C5-layer contains 16 feature maps with size 8×8 and P6-layer consists of 16 feature maps with size 4×4.

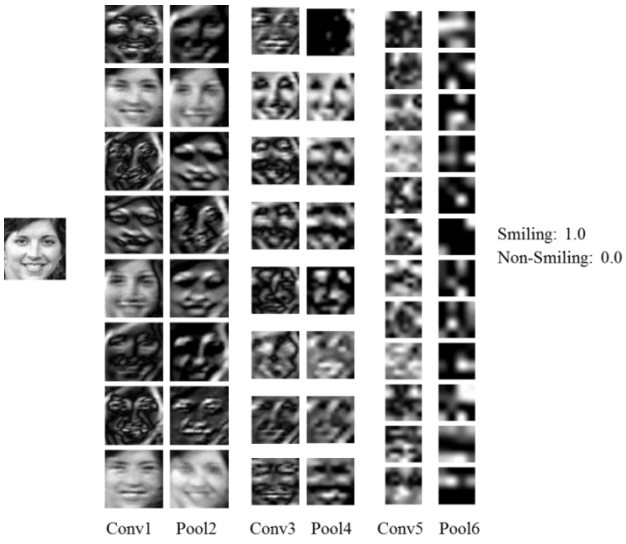


Figure 5-8. The feature maps of each layer in the smile-CNN model.

I can see that the deeper the layer, the representation becomes more abstract and sparse. In C1-layer, it is easy to identify these “smile” faces. In C2-layer, although the feature maps become ambiguous, it can still recognize the outline of the faces. In

C5-layer and P6-layer, the feature maps are difficult to recognize. Since the deep learning model extract the features layer by layer and the deeper layer can capture higher level representations. The representations in P6-layer are more compact with a better discriminative ability. The activation outputs of the last hidden layer (P6-layer) were extracted as the learned features, as shown in Figure 5-9. These learned features were used to train an SVM classifier and an AdaBoost classifier. The average accuracy yielded by the SVM and AdaBoost classifiers is 92.4% and 91.8%, respectively. Note that when I trained SVM or AdaBoost with the raw data, the average accuracy achieved by SVM and AdaBoost is only 84.0% and 81.0%, which is much lower than the performance achieved with the learned features extracted from Smile-CNN. It demonstrates that Smile-CNN can learn some discriminative representations from the raw data.

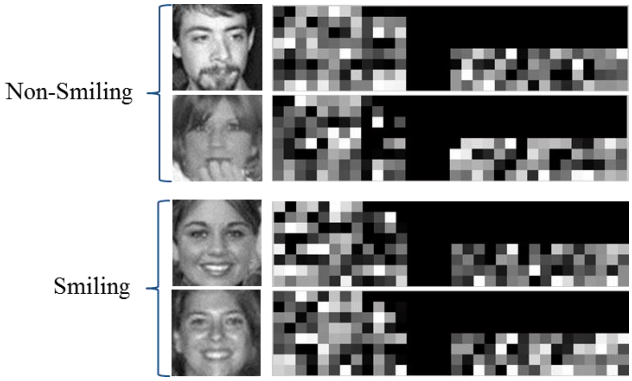


Figure 5-9. The learned features extracted from the Smile-CNN. Left column: the original input images; Right column: the activation outputs of the last hidden layer (P6-layer). There are 16 4-by-4 feature maps and I reshape them as 8×32 for the convenience of illustration.

I also compare our results with the other state of the art methods. The results are shown in Table 5-4. Note that all the other methods compared apply the hand-crafted features, such as HOG in (M. Liu et al., 2012) and (An et al., 2015)), LBP in (Shan, 2012) and (An et al., 2015)) and pixel comparison in (Shan, 2012). Our method performs better than that acquired by LBP and pixel comparison. Compared with HOG, our performance is still competitive. Smile-CNN can not only do well in the smile recognition but also effectively learn some powerful representations from raw pixels. Consider that the hand-crafted features have a higher dimension, it is reasonable to conclude that the learned features are more compact and efficient to represent facial expressions.

Table 5-4. A comparison of our method with the other methods on the GENKI4K database.

Method	**Feature**	**Dimension**	**Classifier**	**Accuracy (%)**
(An et al., 2015)	LBP	500	ELM	85.2
	HOG	500	ELM	88.2
(Shan, 2012)	LBP	944	SVM	87.1±0.76
	Pixel Comparison	500	AdaBoost	89.7±0.45
(M. Liu et al., 2012)	HOG (labeled)	1200	SVM	91.8±0.97
	HOG (labeled + unlabeled)	1200	SVM	92.3±0.81
Our method	Raw pixels	64×64	SVM	84.0±0.91
	Raw pixels	64×64	AdaBoost	81.0±0.76
	Learned Features	256	SVM	**92.4±0.59**
	Learned Features	256	AdaBoost	91.8±0.95

5.3.2.3 Effect of Nuisance Factors

I further investigate the impacts of the alignment and background for this problem. Three different types of face images are set for a comparison. Figure 5-10 illustrates the three different types of face images used in our work. The original face images are shown in the top row. The second row demonstrates the cropped aligned face images, which are widely used in most previous works. The third row shows the cropped face images without alignment, in which we can see that there are pose variations. The face images in the bottom row are without preprocessing and these images vary in poses, backgrounds and face scales. All the three types of face images have been resized to 64×64 and take the same 4-fold cross validation for the convenience of a comparison.

Experimental results are shown in Table 5-5. It can be seen that the average accuracies of Type Ⅰ (92.1%) and Type Ⅱ (90.6%) are quite close, meaning that Smile-CNN is robust to deal with pose variations and face alignment is not very important for this problem. However, the average accuracy of Type Ⅲ has dropped to 78.1%, with a large decline of about 12%, illustrating that the background and scale variations can seriously weaken the recognition ability of the Smile-CNN. It means that cropping the face and eliminating the background play a vital role for this problem.

Figure 5-10. Examples of three different types of face images used in our experiments together with the original face images. The top row: the original faces; the second row: the cropped aligned faces (Type Ⅰ); the third row: the cropped faces without alignment (Type Ⅱ); the bottom row: the face images without preprocessing (Type Ⅲ). All of the three types of face images have been resized to 64×64.

Table 5-5. The accuracy acquired by Smile-CNN on three types of face images (%).

	Type Ⅰ	**Type Ⅱ**	**Type Ⅲ**
Fold 1	91.6	89.7	78.8
Fold 2	91.1	90.8	77.5
Fold 3	92.5	90.2	78.5
Fold 4	93.0	91.7	77.6
Avg.	92.1	90.6	78.1

5.3.3 Discussion

The experimental results reported in this chapter illustrate that a deep convolutional network can effectively perform smile detection in the wild by hierarchical representation learning. From the comparison results of MLP and Smile-CNN, it can be concluded that architecture is the key to enhance the

recognition ability of Smile-CNN. Compared with MLP, Smile-CNN consists of much more hidden units while including much less trainable weights. Each hidden unit only covers a local patch or sub region of the input feature map, which is helpful to fully exploit correlations of adjacent pixels. Pooling operation is able to transform the joint feature representation into a new, more usable one that preserves important information and discards irrelevant information.

An attractive advantage of the deep learning model is to learn the representations from raw inputs. In general, with multiple layers, a deep convolutional network can learn very complex mapping functions which can transform the representation at low level (starting with the raw input) into a representation at a higher, more abstract level (Y. LeCun et al., 2015). The high level representation is generally more compact (from 64×64 to 256). Table 5-4 illustrates that when classifying raw pixels directly, the average accuracy is about 84%. However, when classifying the features learned by Smile-CNN from raw pixels, the performance increases to about 92%, with a significant improvement of about 8%. Consider that raw inputs are of size of 64×64 and the learned features with a dimension of 256, meaning that the deep convolutional network can remove redundant information and keep the distinct useful information from raw inputs.

In addition, from Table 5-5, I also find that face alignment is not very important,

meaning that the deep convolutional network is robust to deal with small pose variations. On the other hand, the background and scale variations tend to increase the complexity of this problem and undermine the classifying power of the Smile-CNN. A reasonable explanation lies in that different from the object recognition, such as face recognition, facial expression is more abstract and thus the features used to represent a facial expression are more subtle. The background would likely to cover or pollute subtle information and make facial expression recognition become more difficult.

5.4 Summary

This chapter introduces an effective approach based on deep learning to address the problem of smile detection in the wild. Different from some previous research works which performed feature extraction and classification separately, deep learning can effectively combine the two stages into a single trainable model. In this study, a deep convolutional network called Smile-CNN was developed to perform the smile detection. Although a deep learning model is generally developed for dealing with "big data", I found that it could also effectively handle "small data" in this task with an appropriate model architecture and parameter setting. During the training process, drop out technique was adopted to lessen the overfitting problem. In addition, considering that a deep convolutional network can extract features hierarchically and higher level representations are more abstract and compact, I obtained the outputs

from the last hidden layer as the learned features which were used to train the SVM and AdaBoost classifiers for the purpose of comparison. Experimental results demonstrate an impressive discriminative power of these features. Smile-CNN achieves a promising performance on the public GENKI4K database compared with other methods. A broad investigation on the impacts of pose variations and backgrounds is also presented in this chapter. It has been observed that the Smile-CNN is robust to deal with pose variations while it is less effective in handling a complex background.

Chapter 6 Conclusion and Future Work

6.1 Conclusion of the Thesis

In this thesis, I present my research findings on novel hand-crafted features, multi-task approach, and deep learning for facial affect recognition. The thesis can be concluded as follows.

In Chapter 3, an efficient approach for facial expression recognition in video with novel features and multiple feature fusion is proposed. Facial expressions are caused by facial muscle activities. These muscle activities generate facial appearance and configuration changes. A new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) is proposed to extract dynamic textures from video sequences to characterize facial appearance changes. Compared with LBP-TOP, a widely used feature descriptor for modeling dynamic textures, our proposed HOG-TOP has several advantages. One is the feature dimensionality. The size of LBP-TOP coded using a uniform pattern is 59×3, while the size of HOG-TOP quantized into 9 bins is 9×3, which is much more compact than that of LBP-TOP. In addition, HOG-TOP is more effective to characterize facial appearance changes than LBP-TOP. Moreover, HOG-TOP is more computationally efficient than LBP-TOP. I also propose a new effective geometric feature derived from the warp transformation

of facial landmarks to capture facial configuration changes. The proposed geometric feature is more robust to capture facial non-rigid configuration changes than the other geometric features. Moreover, the role of audio modalities on recognition is also explored. Audio features can provide complementary information for visual features. Multiple feature fusion developed to tackle the video-based facial expression recognition problems under lab-controlled environment and in the wild are also presented in the chapter.

In Chapter 4, I present my research findings on pain analysis in video. There are two problems to be solved: pain event detection and pain event locating. The first problem is to predict whether there exists any pain event in a video sequence and the second problem is to locate the pain events in a video sequence. I address the role of facial information at various time scales (frame-level, segment-level and sequence-level) and propose a new framework for pain event detection and locating in video. A multiple-feature fusion method for pain event detection and a multiple-task fusion method for locating pain events are introduced. Both spatial and spatial-temporal features were utilized in this study. In our framework, HOG of fiducial points (P-HOG) is used to characterize spatial features from video frames and a trained SVM classifier is used for frame-level detection. In order to further address spatial-temporal information among contiguous frames, segment-level detection is

proposed to assist the frame-level detection. HOG from three orthogonal planes (HOG-TOP) is applied to model dynamic textures of video segments. A trained SVM classifier is used to perform segment-level detection. I further applied the max pooling strategy to obtain global P-HOG and HOG-TOP to represent the whole video sequence and employed multiple kernel fusion to optimally combine the two types of global features. An SVM with multiple kernels was trained to perform the sequence-level (pain event) detection. At last, an effective probabilistic fusion method was utilized to integrate the detection results from the three different tasks (frame-level, segment-level and sequence-level detection) to locate pain events in video. By integrating three different tasks, the proposed method provides a more robust and precise detection of pain events in video than the other previously reported techniques which usually focus on one of these tasks only.

In Chapter 5, the study for smile detection in the wild with deep convolutional neural networks (CNNs) is presented. Different from conventional facial expression recognition techniques which extracted hand-crafted features from face images and trained a classifier to perform smile recognition in a two-step approach, deep learning can effectively combine feature learning and classification into a single model. A deep convolutional neural network called Smile-CNN was developed to perform feature learning and smile detection simultaneously. By comparing Smile-CNN with MLP, I

found that architecture is the key for the improved recognition power of Smile-CNN. Compared with MLP, Smile-CNN consists of much more hidden units while including much less trainable parameters. Each hidden unit only covers a local patch or sub-region of the input feature map, which is helpful to fully exploit correlations of adjacent pixels. Pooling operation is able to transform the joint feature representation into a new, more usable one that preserves important information and discards irrelevant information. I further investigated into the discriminative power of the learned features, which were taken from the neuron activations of the last hidden layer of the Smile-CNN model. By using the learned features to train an SVM or AdaBoost classifier, I show that the learned features have impressive discriminative ability. Experimental results demonstrate that the proposed approach can achieve a promising performance in smile detection. I also present in this chapter an investigation on the impacts of pose variations and backgrounds. It is observed that the Smile-CNN model is robust in dealing with pose variations while it is still challenging to handle a complex background.

6.2 Future Research Directions

More work can be pursued along the line of our research, which is discussed below.

In Chapter 3, I have demonstrated that our approach can achieve a promising

performance on facial expression recognition under lab-controlled environment. However, for affect recognition in the wild, the performance is far from satisfactory. How to effectively address the problem of affect recognition in the wild is one of our future research directions. The key to solve the problem is to explore the representation capability of multiple modalities. Multimodality can enrich the representation space and improve emotion inference. These modalities include faces, voice, body gestures, actions and physiological information (brain signals). The color and depth information can also be explored. I believe that multiple modalities can provide complementary cues and make different contributions to affect recognition. Nevertheless, how to mine useful representations from different modalities and how to integrate these different representations optimally need to be carefully studied.

In Chapter 4, I present a new framework with multiple tasks for detecting and locating pain events in video. Although our framework achieves a promising performance compared with other methods, careful engineering and considerable domain expertise are required to design effective features. Deep learning has attracted growing attention recently and many deep learning models have been successfully applied for most computer vision applications including image classification, object recognition and face detection etc. Pain analysis with deep learning methods is another potential research direction. Convolutional neural networks have illustrated its

superiority to deal with 2-D images. We can explore a CNN solution to tackle the problem of pain detection and locating. In order to fully explore the spatial and spatial-temporal information, we may design two different CNNs to deal with two types of inputs: single frame and continuous frames (frame segment). A decision-level fusion method can also be developed to combine the prediction results of the two CNNs optimally.

In this thesis, our research focuses on facial affect recognition, i.e. inferring an emotion behind a displayed facial expression. However, facial affect recognition only covers a small subset of emotions. At times facial muscle movements do not reveal a kind of specific emotion, but convey some intentions. It is necessary to analyze and define these subtle facial muscle actions. Facial action units provide other important and valuable cues to explain facial expressions and understand the emotions of human beings. Facial action unit detection is another fundamental research direction in affective computing. The goal of facial action unit detection is to measure and describe facial muscle activities appeared on faces. Detecting multiple action units simultaneously remains a very challenging task, which is another research direction which we may focus on.

References

Abhinav Dhall, Roland Goecke, Simon Lucey and Gedeon, T. (2012). "A semi-automatic method for collecting richly labelled large facial expression databases from movies". *IEEE Multimedia*.

An, L., Yang, S. and Bhanu, B. (2015). "Efficient smile detection by Extreme Learning Machine". *Neurocomputing, 149*, pp. 354-363.

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M. and Solomon, P. E. (2009). "The Painful Face - Pain Expression Recognition Using Active Appearance Models". *Image and Vision Computing, 27*(12), pp. 1788-1796.

Bengio, Y. (2009). "Learning Deep Architectures for AI". *Foundations and Trends in Machine Learning, 2*(1), pp. 1-127.

Bengio, Y., Courville, A. and Vincent, P. (2013). "Representation Learning_ A Review and New Perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), pp. 1798-1828.

Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H. (2007). "Greedy Layer-Wise Training of Deep Networks". *Advances in Neural Information Processing Systems*, pp. 153-160.

Bishop, C. M. (2007). "Pattern Recognition and Machine Learning": *Springer.*

Chang, C.-C. and Lin, C.-J. (2011). "LIBSVM: a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3).

Chen, J., Chen, Z., Chi, Z. and Fu, H. (2014). "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning". *ACM International Conference on Multimodal Interaction 2014*, pp. 508-513.

Chen, J., Chen, Z., Chi, Z. and Fu, H. (2015). "Dynamic texture and geometry features for facial expression recognition in video". *2015 IEEE International Conference on Image Processing (ICIP),* pp. 4967-4971.

Chew, S. W., Lucey, P., Lucey, S., Saragih, J., Cohn, J. F. and Sridharan, S. (2011). "Person-independent facial expression detection using constrained local models". *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* pp. 915-920.

Chew, S. W., Rana, R., Lucey, P., Lucey, S. and Sridharan, S. (2012). "Sparse Temporal Representations for Facial Expression Recognition" *Advances in Image and Video Technology* (pp. 311-322).

Cohn, J. F. and Ekman, P. (2005). "Measuring facial action" *The new handbook of methods in nonverbal behavior research*: New York: Oxford University Press.

Corneanu, C. A., Simon, M. O., Cohn, J. F. and Guerrero, S. E. (2016). "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression

Recognition: History, Trends, and Affect-Related Applications". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(8), pp. 1548-1568.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G. (2001). "Emotion recognition in human-computer interaction". *Signal Processing Magazine, IEEE, 18*(1), pp. 32-80.

Dahmane, M. and Meunier, J. (2011). "Emotion recognition using dynamic gridbased hog features". *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* pp. 884-888.

Dalal, N. and Triggs, B. (2005). "Histograms of Oriented Gradients for Human Detection". *IEEE Conference on Computer Vision and Pattern Recognition, 2005.* pp. 886-893.

Devries, T., Biswaranjan, K. and Taylor, G. W. (2014). "Multi-Task Learning of Facial Landmarks and Expression". *2014 Canadian Conference on Computer and Robot Vision (CRV),* pp. 98-103.

Dhall, A., Asthana, A., Goecke, R. and Gedeon, T. (2011). "Emotion recognition using PHOG and LPQ features". *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, pp. 878-883.

Dhall, A., Goecke, R., Joshi, J., Sikka, K. and Gedeon, T. (2014). "Emotion

Recognition In The Wild Challenge 2014: Baseline, Data and Protocol". *ACM International Conference on Multimodal Interaction 2014*, pp. 461-466.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R. and Pal, C. (2015). "Recurrent neural networks for emotion recognition in video". *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 467-474.

Egede, J., Valstar, M. and Martinez, B. (2017). "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation". *arXiv preprint arXiv:1701.04540*.

Ekman, P., Friesen, W. V. and Hager, J. C. (2002). "Facial Action Coding System: The Manual on CD ROM. A Human Face".

Eslami, S. A. and Williams, C. (2012). "A generative model for Parts-based Object Segmentation". *Advances in Neural Information Processing Systems*, pp. 100-107.

Essa, I. A. and Pentland, A. P. (1997). "Coding, Analysis, Interpretation, and Recognition of Facial Expressions". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), pp. 757-763.

Eyben, F., Wollmer, M. and Schuller, B. (2009). "OpenEAR — Introducing the munich open-source emotion and affect recognition toolkit". *3rd International*

Conference on Affective Computing and Intelligent Interaction and Workshops. ACII 2009, pp. 1-6.

Feichtinger, H. G. and Strohmer, T. (1998). "Gabor analysis and algorithms: Theory and applications": *Springer*.

Florian Eyben, Martin Wöllmer and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor". *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459-1462.

Freund, Y. and Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences, 55*(1), pp. 119-139.

Fukushima, K. (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". *Biological Cybernetics, 36*(4), pp. 193-202.

Gönen, M. and Alpaydın, E. (2011). "Multiple Kernel Learning Algorithms". *The Journal of Machine Learning Research, 12*, pp. 2211-2268.

Glauner, P. O. (2015). "Deep Convolutional Neural Networks for Smile Recognition". *arXiv preprint arXiv:1508.06535*.

Glorot, X., Bordes, A. and Bengio, Y. (2011). "Deep Sparse Rectifier Neural Networks". *Proceedings of the Fourteenth International Conference on*

Articial Intelligence and Statistics, pp. 315-323.

Hammal, Z. and Cohn, J. F. (2012). "Automatic detection of pain intensity". *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pp. 47-52.

He, M., Horng, S.-J., Fan, P., Run, R.-S., Chen, R.-J., Lai, J.-L., . . . Sentosa, K. O. (2010). "Performance evaluation of score level fusion in multimodal biometric systems". *Pattern Recognition, 43*(5), pp. 1789-1800.

Hicklin, A., Ulery, B. and Watson, C. (2006). "A brief introduction to biometric fusion". *National Institute of Standards and Technology*.

Hinton, G. E., Osindero, S., Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation.* Retrieved 7, 18

Hinton, G. E. and Salakhutdinov, R. R. (2006). "Reducing the dimensionality of data with neural networks". *Science, 313*(5786), pp. 504-507.

Huang, F. J. and LeCun, Y. (2006). "Large-scale Learning with SVM and Convolutional for Generic Object Categorization". *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 284-291.

Huang, X., He, Q., Hong, X., Zhao, G. and Pietikainen, M. (2014). "Improved Spatiotemporal Local Monogenic Binary Pattern for Emotion Recognition in The Wild". *ACM International Conference on Multimodal Interaction 2014*,

pp. 514-520.

Huang, X., Zhao, G., Pietikäinen, M. and Zheng, W. (2011). "Expression Recognition in Videos Using a Weighted Component-Based Feature Descriptor". *Proceedings of the 17th Scandinavian Conference on Image Analysis*, pp. 569-578.

Huang, X., Zhao, G., Zheng, W. and Pietikainen, M. (2012). "Spatio temporal Local Monogenic Binary Patterns for Facial Expression Recognition". *Signal Processing Letters, IEEE, 19*(5), pp. 243-246.

Hubel, D. H. and Wiesel, T. N. (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". *The Journal of Physiology, 160*(1), pp. 106-154.

Ijjina, E. P. and Mohan, C. K. (2014). "Facial Expression Recognition Using Kinect Depth Sensor and Convolutional Neural Networks". *2014 13th International Conference on Machine Learning and Applications (ICMLA),* pp. 392-396.

Jain, V., Crowley, J. L. and Lux, A. (2014). "Local Binary Patterns Calculated over Gaussian Derivative Images". *2014 22nd International Conference on Pattern Recognition (ICPR),* pp. 3987-3992.

Kaltwang, S., Rudovic, O. and Pantic, M. (2012). "Continuous Pain Intensity Estimation from Facial Expressions" *Advances in Visual Computing* (pp.

368-377).

Kanade, T., Cohn, J. F. and Tian, Y. (2000). "Comprehensive database for facial expression analysis". *Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings*, pp. 46-53.

Kanou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., . . . Bengio, Y. (2013). "Combining modality specific deep neural networks for emotion recognition in video". *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 543-550.

Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M. and Cun, Y. L. (2010). "Learning Convolutional Feature Hierachies for Visual Recognition". *Advances in Neural Information Processing Systems*, pp. 1090-1098.

Kaya, H. and Salah, A. A. (2014). "Combining Modality-Specific Extreme Learning Machines for Emotion Recognition in the Wild". *ACM International Conference on Multimodal Interaction 2014*, pp. 487-493.

Kim, Y., Lee, H. and Provost, E. M. (2013). "Deep learning for robust feature generation in audiovisual emotion recognition". *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3687-3691.

Koelstra, S., Pantic, M. and Patras, I. (2010). "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(11), pp. 1940-1954.

Krizhevsky, A., Sutskever, I. and Hinton, G. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing Systems 25*, pp. 1106-1114.

Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E. and Jordan, M. I. (2004). "Learning the Kernel Matrix with Semi-Definite Programming". *The Journal of Machine Learning Research, 5*, pp. 27-72.

Lawrence, S., Giles, C. L., Tsoi, A. C. and Back, A. D. (1997). "Face recognition a convolutional neural-network approach". *IEEE Transactions on Neural Networks, 8*(1), pp. 98-113.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). "Deep learning". *Nature, 521*, pp. 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989). "Backpropagation Applied to Handwirtten Zip Code Recognition". *Neural computation, 1*(4), pp. 541-551.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). "Gradient Based Learning

Applied to Document Recognition". *Proceedings of the IEEE, 86*(11), pp. 2278-2324.

Lee, H., Grosse, R., Ranganath, R. and Ng, A. Y. (2009). "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations". *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609-616.

Li, S. Z. and Jain, A. K. (2011). "Handbook of face recognition": *springer*.

Li, Y., Wang, S., Zhao, Y. and Ji, Q. (2013). "Simultaneous Facial Feature Tracking and Facial Expression Recognition". *IEEE Transactions on Image Processing, 22*(7), pp. 2559-2573.

Liang, D., Yang, J., Zheng, Z. and Chang, Y. (2005). "A facial expression recognition system based on supervised locally linear embedding". *Pattern Recognition Letters, 26*(15), pp. 2374-2389.

Liu, M., Li, S., Shan, S. and Chen, X. (2012). "Enhancing Expression Recognition in the Wild with Unlabeled Reference Data". *Computer Vision–ACCV 2012*, pp. 577-588.

Liu, M., Li, S., Shan, S. and Chen, X. (2013). "AU-aware Deep Networks for facial expression recognition". *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),* pp. 1-6.

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z. and Chen, X. (2014). "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild". *ACM International Conference on Multimodal Interaction 2014*, pp. 494-501.

Liu, P., Han, S., Meng, Z. and Tong, Y. (2014). "Facial Expression Recognition via a Boosted Deep Belief Network". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805-1812.

Liu, W. and Wang, Z. (2006). "Facial Expression Recognition Based on Fusion of Multiple Gabor Features". *18th International Conference on Pattern Recognition, 2006*, pp. 536-539.

Liu, Z., Luo, P., Wang, X. and Tang, X. (2015). "Deep learning face attributes in the wild". *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730-3738.

Liu, Z., Wu, W., Tao, Q. and Yang, J. (2013). "Facial Expression Recognition Using a New Image Representation and Multiple Feature Fusion". *Intelligent Science and Intelligent Data Engineering*, pp. 441-449.

Long, F., Wu, T., Movellan, J. R., Bartlett, M. S. and Littlewort, G. (2012). "Learning Spatiotemporal Features by Using Independent Component Analysis with Application to Facial Expression Recognition". *Neurocomputing, 93*, pp.

126-132.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. (2010). "The Extended Cohn-Kanade Dataset (CK+)_ A complete dataset for action unit and emotion-specified expression". *2010 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 94-101.

Lucey, P., Cohn, J. F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J. and Prkachin, K. M. (2011). "Automatically Detecting Pain in Video Through Facial Action Units". *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 41*(3), pp. 664-674.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., Chew, S. and Matthews, I. (2012). "Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database". *Image and Vision Computing, 30*(3), pp. 197-205.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E. and Matthews, I. (2011). "Painful data_ The UNBC-McMaster shoulder pain expression archive database". *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* pp. 57-64.

Lucey, P., Howlett, J., Cohn, J., Lucey, S., Sridharan, S. and Ambadar, Z. (2008). "Improving pain recognition through better utilisation of temporal

information". *International Conference on Auditory-Visual Speech Processing-AVSP*, pp. 167-172.

Lynch, M. E., Craig, K. D. and Peng, P. W. (2011). "*Clinical pain management: a practical guide"*: John Wiley & Sons.

Lyons, M., Akamatsu, S., Kamachi, M. and Gyoba, J. (1998). "Coding Facial Expressions with Gabor Wavelets". *Third IEEE International Conference on Automatic Face and Gesture Recognition, Proceedings.*, pp. 200-205.

Lyons, M. J., Budynek, J. and Akamatsu, S. (1999). "Automatic Classification of Single Facial Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(12), pp. 1357-1362.

Mangai, U. G., Samanta, S., Das, S. and Chowdhury, P. R. (2010). "A survey of decision fusion and feature fusion strategies for pattern classification". *IETE Technical review, 27*(4), pp. 293-307.

Martinez, A. and Du, S. (2012). "A model of the perception of facial expressions of emotion by humans_ Research overview and perspectives". *The Journal of Machine Learning Research, 13*(1), pp. 1589-1608.

Matsugu, M., Mori, K., Mitari, Y. and Kaneda, Y. (2003). "Subject independent facial expression recognition with robust face detection using a convolutional neural network". *Neural Networks, 16*(5-6), pp. 555-559.

Matthews, I. and Baker, S. (2004). "Active appearance models revisited". *International Journal of Computer Vision, 60*(2), pp. 135-164.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. and Cohn, J. F. (2013). "DISFA_ A Spontaneous Facial Action Intensity Database ". *IEEE Transactions on Affective Computing, 4*(2), pp. 151-160.

Meudt, S. and Schwenker, F. (2014). "Enhanced Autocorrelation in Real World Emotion Recognition". *ACM International Conference on Multimodal Interaction 2014*, pp. 502-507.

Nair, V. and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines". *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807-814.

Nandakumar, K., Chen, Y., Dass, S. C. and Jain, A. (2008). "Likelihood ratio-based biometric score fusion". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(2), pp. 342-347.

Ng, H.-W., Nguyen, V. D., Vonikakis, V. and Winkler, S. (2015). "Deep learning for emotion recognition on small datasets using transfer learning". *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 443-449.

Ojala, T., Pietikainen, M. and Maenpaa, T. (2002). "Multiresolution gray-scale and

rotation invariant texture classification with local binary patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), pp. 971-987.

Orrite, C., Gañán, A. and Rogez, G. (2009). "HOG-Based Decision Tree for Facial Expression Classification" *Pattern Recognition and Image Analysis* (pp. 176-183).

Poultney, C., Chopra, S. and Cun, Y. L. (2007). "Efficient Learning of Sparse Representations with an Energy-Based Model". *Advances in Neural Information Processing Systems*, pp. 1137-1144.

Prkachin, K. M. (1992). "The consistency of facial expressions of pain: a comparison across modalities". *Pain, 51*(3), pp. 297-306.

Prkachin, K. M. and Solomon, P. E. (2008). "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain". *Pain, 139*(2), pp. 267-274.

Rakotomamonjy, A., Bach, F. R., Canu, S. and Grandvalet, Y. (2008). "SimpleMKL". *Journal of Machine Learning Research, 9*(11), pp. 2491-2521.

Ramirez Rivera, A., Castillo, R. and Chae, O. (2013). "Local Directional Number Pattern for Face Analysis_ Face and Expression Recognition". *IEEE Transactions on Image Processing, 22*(5), pp. 1740-1752.

Rifai, S., Bengio, Y., Courville, A., Vincent, P. and Mirza, M. (2012). "Disentangling Factors of Variation for Facial Expression Recognition". *Computer Vision–ECCV 2012*, pp. 808-822.

Ringeval, F., Amiriparian, S., Eyben, F., Scherer, K. and Schuller, B. (2014). "Emotion Recognition in the Wild Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion". *ACM International Conference on Multimodal Interaction 2014*, pp. 473-480.

Saeed, A., Al-Hamadi, A., Niese, R. and Elzobi, M. (2012). "Effective geometric features for human emotion recognition". *IEEE 11th International Conference on Signal Processing (ICSP), 2012*, pp. 623-627.

Salakhutdinov, R. and Hinton, G. E. (2009). "Deep Boltzmann Machines". *International Conference on Artificial Intelligence and Statistics*, pp. 448-455.

Scherer, K. and Ekman, P. (1982). "Handbook of Methods in Nonverbal Behavior Research": *Cambridge, UK: Cambridge Univ. Press.*

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A. and Narayanan, S. S. (2010). "The InterSpeech 2010 Paralinguistic Challenge". *InterSpeech*, pp. 2794-2797.

Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K. and Prevost, L. (2011). "Combining LGBP histograms with AAMm coefficients in the multi-kernel

svm framework to detect facial action units". *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pp. 860-865.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U. and Poggio, T. (2007). "A quantitative theory of immediate visual recognition". *Progress in Brain Research, 165*, pp. 33-56.

Shan, C. (2012). "Smile detection by boosting pixel differences". *IEEE Transactions on Image Processing, , 21*(1), pp. 431-436.

Shan, C., Gong, S. and McOwan, P. W. (2005). "Robust facial expression recognition using local binary patterns". *IEEE International Conference on Image Processing*, pp. 370-373.

Sikka, K., Dhall, A. and Bartlett, M. S. (2014). "Classification and Weakly Supervised Pain Localization using Multiple Segment Representation". *Image and Vision Computing, 32*(10), pp. 659-670.

Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G. and Bartlett, M. (2013). "Multiple kernel learning for emotion recognition in the wild". *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 517-524.

Sikka, K., Wu, T., Susskind, J. and Bartlett, M. (2012). "Exploring bag of words

architectures in the facial expression domain". *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 250-259.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). "Dropout_ A Simple Way to Prevent Neural Networks from Overfitting". *The Journal of Machine Learning Research, 15*(1), pp. 1929-1958.

Sun, B., Li, L., Zuo, T., Chen, Y., Zhou, G. and Wu, X. (2014). "Combining Multimodal Features with Hierarchical Classifier Fusion for Emotion Recognition in the Wild". *ACM International Conference on Multimodal Interaction 2014*, pp. 481-486.

Taheri, S., Turaga, P. and Chellappa, R. (2011). "Towards view-invariant expression analysis using analytic shape manifolds". *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* pp. 306-313.

Tang, Y. (2013). "Deep Learning using Linear Support Vector Machines". *arXiv preprint arXiv:1306.0239*.

Turan, C. and Lam, K.-M. (2014). "Region-based feature fusion for facial-expression recognition". *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5966-5970.

Turk, D. C. and Melzack, R. (2001). "The measurement of pain and the assessment of

people experiencing pain" *Handbook of Pain Assessment* (pp. 1-11): Guilford, New York, USA.

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

Vedaldi, A. and Lenc, K. (2015). "Matconvnet: Convolutional neural networks for matlab". *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 689-692.

Wang, X., Wang, L. and Qiao, Y. (2012). "A comparative study of encoding, pooling and normalization methods for action recognition". *Computer Vision–ACCV 2012*, pp. 572-585.

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M. and Movellan, J. (2009). "Toward practical smile detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(11), pp. 2106-2111.

Williams, A. C. d. C., Davies, H. T. O. and Chadury, Y. (2000). "Simple pain rating scales hide complex idiosyncratic meanings". *Pain, 85*(3), pp. 457-463.

Wong, J.-J. and Cho, S.-Y. (2009). "A face emotion tree structure representation with probabilistic recursive neural network modeling". *Neural Computing and*

Applications, 19(1), pp. 33-54.

Yu, Z. and Zhang, C. (2015). "Image based static facial expression recognition with multiple deep network learning". *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435-442.

Zavaschi, T. H., Britto, A. S., Oliveira, L. E. and Koerich, A. L. (2013). "Fusion of feature sets and classifiers for facial expression recognition". *Expert Systems with Applications, 40*(2), pp. 646-655.

Zeng, Z., Pantic, M., Roisman, G. I. and Huang, T. S. (2009). "A Survey of Affect Recognition Methods Audio, Visual, and Spontaneous Expressions". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), pp. 39-58.

Zhang, K., Huang, Y., Wu, H. and Wang, L. (2015). "Facial Smile Detection Based on Deep Learning Features". *Chinese Academy of Sciences Institute of Automation*.

Zhang, S., Li, L. and Zhao, Z. (2012). "audio-visual emotion recognition based on facial expression and affective speech". *Multimedia and Signal Processing*, pp. 46-52.

Zhang, X., Zhang, L., Wang, X.-J. and Shum, H.-Y. (2012). "Finding celebrities in billions of web images". *IEEE Transactions on Multimedia, 14*(4), pp. 995-1007.

Zhang, Y. and Ji, Q. (2005). "Active and dynamic information fusion for facial expression understanding from image sequences". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(5), pp. 699-714.

Zhao, G. and Pietikainen, M. (2007). "Dynamic texture recognition using local binary patterns with an application to facial expressions". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), pp. 915-928.

Zhu, X. and Ramanan, D. (2012). "Face detection, pose estimation and landmark localization in the wild". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*, pp. 2879-2886.