

**AFFECT ANALYSIS IN VIDEO**

**XIAOHONG XIANG**

**NATIONAL UNIVERSITY OF SINGAPORE**

**2014**

**AFFECT ANALYSIS IN VIDEO**

**XIAOHONG XIANG**

*(B.Eng., Harbin Institute of Technology, China)*

**A THESIS SUBMITTED**

**FOR THE DEGREE OF DOCTOR OF  
PHILOSOPHY**

**DEPARTMENT OF COMPUTER SCIENCE  
SCHOOL OF COMPUTING**

**NATIONAL UNIVERSITY OF SINGAPORE**

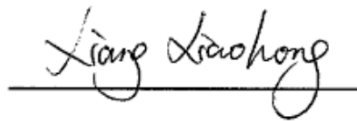
**2014**

## DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, reading "Xiang Xiaohong", is written over a horizontal line.

**Xiaohong Xiang**

10 Jul, 2014

## ACKNOWLEDGMENTS

First of all, my sincerest gratitude goes to my supervisor, Professor Mohan S Kankanhalli, who guided and encouraged me patiently and professionally throughout my doctoral study. Prof. Mohan has not only taught me all aspects of research but most importantly, independent thinking. He has always encouraged me to realize any idea, inspired and aided me when I was in trouble. It has been a very pleasant experience working with him which I have really enjoyed.

Also, I am grateful to have so many great labmates. Yangyang Xiang helped me so much when I joined this lab. Xiangyu Wang has always been available for discussions and I learned much from him. Karthik Yadati, Skanda Muralidhar, Yogesh Singh Rawat and Prabhu Natarajan supported me a lot in my paper writing, as well as with my spoken English.

Last, I would like to thank my husband and my families for being so encouraging and supportive. Without their unconditional love, support, and encouragement, I would not be able to finish my PhD study.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background And Motivation . . . . .	1
1.2	Overview . . . . .	2
1.3	Contributions . . . . .	4
<b>2</b>	<b>Literature Survey</b>	<b>5</b>
2.1	Emotional Models . . . . .	5
2.2	Facial Expression Analysis . . . . .	7
2.3	Multimodal Human’s Emotion Analysis . . . . .	15
2.4	Affective Content In Videos . . . . .	22
2.5	Summary . . . . .	25
<b>3</b>	<b>Sparsity-based Affect Representation And Modeling</b>	<b>26</b>
3.1	Introduction . . . . .	26
3.2	Related Work . . . . .	29
3.3	Methodology . . . . .	29
3.3.1	Overview of Sparse Representation . . . . .	30
3.3.2	Representation And Modeling . . . . .	32
3.3.3	Sample Matrix . . . . .	36
3.4	Experiments . . . . .	40
3.4.1	Over-complete Database . . . . .	42
3.4.2	Affective Classification Results . . . . .	45
3.4.3	Intensity Curve . . . . .	49
3.5	Summary . . . . .	55

<b>4</b>	<b>Affect-based Adaptive Presentation of Home videos</b>	<b>58</b>
4.1	Introduction . . . . .	59
4.2	Related Work . . . . .	62
4.2.1	Adaptive Presentation . . . . .	62
4.2.2	The Emotion Model . . . . .	63
4.2.3	Affective Video Analysis . . . . .	63
4.3	Methodology . . . . .	63
4.3.1	Affective Features Extraction . . . . .	64
4.3.2	Affective Labeling . . . . .	65
4.3.3	Presentation Construction . . . . .	67
4.4	Experimental Results . . . . .	75
4.4.1	Affective Classification Results . . . . .	75
4.4.2	Experimental Results For Presentation . . . . .	77
4.5	Summary . . . . .	79
<b>5</b>	<b>A Multimodal Approach For Online Estimation of Subtle Facial Ex- pression</b>	<b>80</b>
5.1	Introduction . . . . .	80
5.2	Related Work . . . . .	83
5.2.1	Facial Expression Recognition . . . . .	83
5.2.2	Multimodal Human’s Emotion Analysis . . . . .	83
5.3	Methodology . . . . .	84
5.3.1	Modeling The Changes of Human’s Emotion . . . . .	84
5.3.2	Subtle Expression Analysis . . . . .	85
5.4	Experimental Results . . . . .	89
5.4.1	Modeling Human’s emotion changes . . . . .	89
5.4.2	Sparse Representation In Analyzing Facial Expression . . . . .	90
5.4.3	Experimental Results For Subtle Facial Expression Analysis . . . . .	91
5.5	Conclusions . . . . .	93
<b>6</b>	<b>Social Photo Sharing</b>	<b>94</b>
6.1	Introduction . . . . .	94

6.2	Related Work . . . . .	97
6.3	Methodology . . . . .	98
6.3.1	Pre-Processing of The Photo Album . . . . .	99
6.3.2	Assessment Factor Features . . . . .	100
6.3.3	Social Groups . . . . .	103
6.3.4	Classifier Design . . . . .	103
6.4	Experiments . . . . .	104
6.5	Summary . . . . .	108
<b>7</b>	<b>Conclusions</b>	<b>109</b>
7.1	Summary . . . . .	109
7.2	Future Work . . . . .	111
7.2.1	Subtle Facial Expression Analysis . . . . .	111
7.2.2	Multimodal Emotion Analysis . . . . .	111
7.2.3	Utilizing Eye Gaze Data . . . . .	113

# Abstract

Affective computing is currently an active area of research, which is attracting an increasing amount of attention. With the diffusion of *affective computing* in many application areas, *affective video content analysis* is being extensively employed to help computers discern the affect contained in videos. However, the relationship between the syntactic content of the video, which is captured by low level features, and the expected emotion elicited in humans remains unclear, while not much work has been done on the evaluation of the intensity of discrete emotions.

In this thesis, we first propose a computational framework to build the representation and model from the affective video content to the categorical emotional states, while developing a computational measure for the intensity of *categorical* emotional states. Specifically, a sparse vector representation is proposed in this computational framework. The *intensity* of emotion can be represented by the values computed from the sparse vector. Then, the modeling of affective content video addresses the problem of obtaining the representative sparse vectors based on the low-level features extracted from video. The results demonstrate that the proposed approach manages to represent and model the affective video content based on the *categorical emotional states* model, and the obtained intensity time curve of the main emotion is in concurrence with the video content. The second aim of this thesis is to examine the importance of the *affect* in the area of multimedia systems, by utilizing the sparse representation modeling in applications. We therefore develop some useful applications towards this aim.

First, we propose an approach that employs affective analysis to automatically create video presentations from home videos. Our novel method adaptively creates presentations for family, acquaintances and outsiders based on three properties: emotional tone, local main character and global main character. Experimental results show that our



method is very effective for video sharing and the users are satisfied with the videos generated by our method.

Besides the adaptive presentation of home videos, this thesis also exploits the affective analysis (facial expression cue), eye gaze data and previous emotional states to develop an online multimodal approach for estimating the subtle facial expression. It is found that the performance of recognizing “surprise” and “neutral” emotions is improved with the help of eye pupil information; namely, this result demonstrates that the fusion of facial expression, pupillary size and previous emotional state is a promising strategy for detecting subtle expressions.

Furthermore, this thesis also utilizes the affective analysis to propose a novel approach to share home photos based on the aesthetic, affective and social features. This approach allows one to generate a suitable subset of photos from the personal photo collection for sharing with different social kinship groups. It can also be used to check whether an individual photo is appropriate for sharing with a particular kinship group. Our experiments demonstrate the utility of the proposed approach.

Thus, our work is the first to evaluate the intensity of emotions considering the categorical emotional states; the first work to fuse the facial expression, pupil size and previous emotional state to classify the subtle facial expressions; and the first work to propose the concept of adaptive sharing of photos as well. Based on the developed affective modeling approach, in future, more interesting and useful applications can be developed.

# List of Figures

2.1	Illustration of the 3-D emotion space from [DL99] . . . . .	7
2.2	Illustration of the 2-D emotion space from [DL99] . . . . .	7
2.3	Illustration of Circumplex Model [TYA11]. . . . .	7
2.4	Overview of face images analysis system in [LKCL00] . . . . .	9
2.5	Feature-based automatic facial action analysis system in [TKC01] . . . . .	11
2.6	The facial feature extraction and facial expression analysis system in [IRT <sup>+</sup> 05]	13
2.7	A Bayesian temporal manifold model of dynamic facial expressions in [SGM06]	14
2.8	The system framework for mono-modal and bi-modal emotion recogniton in [GP05] . . . . .	18
2.9	Diagram of the proposed methodology of [CMK <sup>+</sup> 06] . . . . .	20
3.1	An example for the “ideal case” of the relationship between the entry values of $x$ and each column of sample matrix $A$ based on the sparse representation: $y = Ax$ . . . . .	33
3.2	An example for the “practical case” of the relationship between the entry values of $\tilde{x}$ and each column of sample matrix $A$ by solving $y = Ax$ using the COSAMP [NT09]. . . . .	34
3.3	The classification rate curve of each emotion when increasing the training samples up to 10% of database. . . . .	43
3.4	The classification rate curve of each emotion when increasing the training samples up to 20% of database. . . . .	43
3.5	The classification rate curve of each emotion when increasing the training samples up to 30% of database. . . . .	44
3.6	The classification rate curve of each emotion when increasing the training samples up to 40% of database. . . . .	44
3.7	The classification rate curve of each emotion when increasing the training samples up to 50% of database. . . . .	45
3.8	The classification rate curve of each emotion when increasing the training samples up to 60% of database. . . . .	45
3.9	The classification rate curve of each emotion when increasing the training samples up to 70% of database. . . . .	46
3.10	The classification rate curve of each emotion when increasing the training samples up to 80% of database. . . . .	46
3.11	The classification rate curve of each emotion when increasing the training samples up to 90% of database. . . . .	47
3.12	Intensity time curve obtained for an excerpt from the film “E.T.”. . . . .	52
3.13	Intensity time curve obtained for an excerpt from the film “There is Some- thing about Mary (2)”. . . . .	53

3.14	Intensity time curve obtained for an excerpt from the film “Schindlers list (2)” . . . . .	53
3.15	Intensity time curve obtained for an excerpt from a news “Weather Forecast” . . . . .	54
3.16	Intensity time curve obtained for an excerpt from the film “Life is beautiful (La vita bella)(2)” . . . . .	55
3.17	Intensity time curve obtained for an excerpt from the film “Seven (2)” . . . . .	55
3.18	Intensity time curve obtained for an excerpt from the film “Trainspotting (1)” . . . . .	56
4.1	The overall framework of our proposed method . . . . .	64
4.2	Example of original videos. . . . .	78
5.1	The overall framework of our proposed approach . . . . .	82
5.2	A Markov Chain with 3 states (labeled $S_1, S_2, S_3$ ). . . . .	85
5.3	ne intuitive example for sparse representation of facial expression in the ideal situation. . . . .	86
5.4	The setup of experiments. . . . .	91
5.5	An screen shot of our developed system to identify the emotion of human. . . . .	93
6.1	The Framework of the proposed approach. . . . .	96
6.2	The overview of pre-processing when a photo album or collection is provided. . . . .	99
6.3	The algorithm for assessing which social groups the input photo is suitable for sharing. . . . .	104
6.4	Image examples for sharing with different social groups. . . . .	106
6.5	The classification results of second classifier design - $SVM_1$ . . . . .	106
6.6	The classification results of second classifier design - $SVM_2$ . . . . .	107

# List of Tables

2.1	Summarization of facial expression recognition algorithms. . . . .	16
2.2	Summarization of multimodal user’s emotion analysis . . . . .	21
2.3	Summarization of the related work of affective content in videos. . . . .	24
3.1	The number of shots and scenes for each emotion. . . . .	40
3.2	Recognition Results based on different shot-level features and fusion level. The bold decision-level ratio is the “optimal” ratio in our experiments. . .	42
3.3	Classification results based on different scene-level features and fusion level. . . . .	48
3.4	Labels describing the content of the test video clips. . . . .	51
4.1	Confusion matrices of classification based on feature-level fusion and decision- level fusion respectively. . . . .	76
4.2	Details of original videos and the corresponding three presentations. . . .	77
4.3	Results of user study . . . . .	78
5.1	The transition probability matrices for group and one person respectively.	89
5.2	Person-independent confusion matrix for classifying facial expressions us- ing sparse representation. . . . .	90
5.3	The experimental results for the proposed subtle facial expression recog- nition method. . . . .	92
6.1	The results of SVM classifier of person independent. . . . .	105
6.2	The results of SVM classifier of person dependent. . . . .	105

# List Of Symbols

Symbols	Meanings
$m$	The number of emotional states: $\in \mathbf{N}$
$k$	The number of features extracted: $\in \mathbf{N}$
$n_j$	The cardinality of the set of the representative feature vectors of the $j$ th emotional state: $\in \mathbf{N}$
$\alpha_{j,i}$	$i$ th representative feature vector of the $j$ th emotional state: $\in \mathfrak{R}^k$
$\beta_{j,i}$	Linear coefficient corresponding to $\alpha_{j,i}$ : $\in \mathfrak{R}$
$A_j$	Sub-sample matrix, i.e. sample matrix of the $j$ th emotional state: $\in \mathfrak{R}^{k \times n_j}$ $= [\alpha_{j1}, \dots, \alpha_{j,n_j}]$
$\beta_j$	Linear coefficient vector of $A_j$ : $\in \mathfrak{R}_j^n$
$A$	Sample matrix: $\in \mathfrak{R}^{k \times n}$ $= [A_1, \dots, A_m]$
$n$	Sum of the representative feature vectors of all emotional states: $\in \mathbf{N}$ $= \sum_{j=1}^m n_j$
$y$	Test feature vector: $\in \mathfrak{R}^k$
$q$	“Emotional property” of $y$ : $\in [1, \dots, m]$
$x$	Sparse solution: $\in \mathfrak{R}^n$
$\tilde{x}$	Approximation of $x$ : $\in \mathfrak{R}^n$
$\Psi$	Downsampling matrix in compressive sampling: $\in \mathfrak{R}^{k \times n}$
$f$	An arbitrary sparse or compressive signal: $\in \mathfrak{R}^n$
$\tilde{f}$	Approximation of $f$ : $\in \mathfrak{R}^n$
$s$	Sparsity factor: $\in [1, \dots, n_q]$
$\Upsilon_j$	Intensity of the $j$ th emotional state within $y$ : $\in [0, \dots, 1]$
$\Phi_j(x)$	Return a new vector consisting of the entries within $x$ which correspond to $A_j$ : $\in \mathfrak{R}^{n_j}$
$\tilde{y}_j$	Approximation of $y$ for the $j$ th emotional state: $\in \mathfrak{R}^k$ $= A_j \Phi_j(\tilde{x})$

*continued on next page*

continued from previous page

Symbols	Meanings
$\varphi_j$	Difference between $y$ and $\tilde{y}_j$ : $\in \mathfrak{R}^+$ $= \ y - A_j \Phi_j(\tilde{x})\ _2$
$\varphi$	Residual vector: $\in \mathfrak{R}^m$ $= [\varphi_1, \dots, \varphi_m]^T$
$\delta_s$	$s$ -restricted isometry constant [CW08];
$\theta_{s,t}$	The $s, t$ -restricted orthogonality constants [Can06, CT07];
$c/c_1/c_2/\gamma$	Constant
$\nu$	Visual feature vector: $\in \mathfrak{R}^k$
$v$	Audio feature vector: $\in \mathfrak{R}^k$
$\varphi_\nu$	Residual vector corresponding to visual feature vector $\nu$ : $\in \mathfrak{R}^m$
$\varphi_v$	Residual vector corresponding to audio feature vector $v$ : $\in \mathfrak{R}^m$
$w_1/w_2$	Weight parameter: $\in [0, \dots, 1]$ $w_1 + w_2 = 1$
$p$	Then number of key frames found in video clip: $\in \mathbf{N}$
$\hat{y}_i$	Visual feature vector extracted from $i$ th key frame
$y_a$	Audio feature vector extracted from audio component of video clip
$A_v$	Sample matrix only constructed by visual features
$A_a$	Sample matrix only constructed by audio features
$\hat{x}_i$	Approximation solution for $\hat{y}_i = A_v x$
$\hat{x}_a$	Approximation solution for $\hat{y}_a = A_a x$
$\hat{\varphi}_i$	Residual vector corresponding to $\hat{x}_i$ : $\in \mathfrak{R}^m$
$\hat{\varphi}_a$	Residual vector corresponding to $\hat{x}_a$ : $\in \mathfrak{R}^m$
$t_j$	term $j$ in Eq.(4.3) Tag/label $j$ in Eq.(4.4)
$n_{i,j}$	The number of occurrences of term $t_j$ in document $d_i$
$ D_c $	The total number of documents in the corpus in Eq.(4.3)
	The total number of shots in a video when computing LMC
	The total number of shots in a video collection when computing GMC and ET
$ d : t_j \in d $	the number of documents where the term $t_j$ appears
	the number of shots assigned with label $t_j$
$v_i$	$i$ th video in a video collection
$w_j^e$	The tf.idf weight of emotional label $t_j$
$w_{i,j}^l$	Local character weight (tf.idf weight) for person label $t_j$ in $v_i$
$w_j^g$	Global character weight (tf.idf weight) for person label $t_j$ in video collection

continued on next page

*continued from previous page*

<b>Symbols</b>	<b>Meanings</b>
$\varepsilon^l$	Threshold of Local character weight
$\varepsilon^g$	Threshold of global character weight
$s_i$	$i$ th shot
$D_{ij}$	Diversity between shots $s_i$ and $s_j$
$w_T$	Overall weight by fusing ET, LMC and GMC
$V_s$	Video collection
$S_j$	$j$ th emotional state in the Markov Chain figure
$I_j$	Temporary facial expression image for $j$ th emotion
$SC_j$	Sparse confidence for $j$ th emotional state
$\Lambda$	Pupil size detected by eye tracker
$\mu$	Mean of pupil size in the neutral emotional state
$\sigma$	Standard variation of pupil size in the neutral emotional state
$w_i^c$	Importance weight for $i$ th category of association in social network: $i \in [1, \dots, 6]$
$w^c$	$= [w_1^c, \dots, w_6^c]$
$w_{(p_i, p_j)}$	Degree weight of association of the linked people $p_i$ and $p_j$
$\tilde{w}_i$	Importance weight of $i$ th divided region: $i \in [1, \dots, 9]$
$\tilde{w}$	List of region weight $= [\tilde{w}_1, \dots, \tilde{w}_9]$
$w_1/w_2/w_3$	Weight parameter: $\in [0, \dots, 1]$

# List Of Abbreviations

Abbreviation	Meanings
V	Valence component in dimensional emotional space
A	Arousal component in dimensional emotional space
C	Control component in dimensional emotional space
ROI	Region Of Interest
HCI	Human-Computer Interaction
HCII	Human-Computer Intelligent Interaction
HMM	Hidden Markov Model
AU	Action Unit
ERBPS	Eye Region Biometric Processing System
FAP	Facial Animation Parameter
LBP	Local Binary Pattern
GB	GentleBoost
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
SVM	Support Vector Machine
FP	Feature Point
SVR	Support Vector Machine for Regression
CoSaMP	Compressive Sampling Matching Pursuit
NN	Nearest Neighbor
HOG	Histogram of Oriented Gradients
GMM	Gaussian Mixture Model
ACHMM	Auto-Regressive Coupled HMM
MFCC	Mel Frequency Cepstral Coefficients
UUP	Uniform Uncertainty Principle
RIP	Restricted Isometry Property
SVM-SC	Support Vector Machine on Sparse Coding

*continued on next page*



---

*continued from previous page*

<b>Abbreviation</b>	<b>Meanings</b>
SVD	Singular Value Decomposition
PCA	Principal Component Analysis
ET	Emotional Tone
LMC	Local Main Character
GMC	Global Main Character
SC	Sparse Confidence
JAFFE	Japanese Female Facial Expression

# Chapter 1

## Introduction

### 1.1 Background And Motivation

In recent times, with the advancement of technology, a variety of consumer electronic devices, such as digital cameras and computers, have become more and more popular in our daily life. It is much easier for individuals to produce and obtain multimedia material like videos and images. Concomitantly, the development of the multimedia analysis techniques, such as attention analysis and semantic analysis, has enabled a variety of multimedia applications, such as video and image retrieval, personalized television, and multilanguage learning. However, video data is becoming increasingly voluminous and redundant because of the steadily increasing capacity and content variety of videos. It is thus more difficult to effectively organize and manage videos in order to find the desired clips or video content.

Visual attention analysis and semantic analysis are two important traditional multimedia analysis techniques. Visual attention is a multidisciplinary endeavor which relates to multiple fields such as cognitive psychology, computer vision and multimedia. A great deal of research has been done on analyzing static attention and identifying Region of Interest (ROI) in still images [MZ03, IKN98, ZS06, YLSL07]. Visual attention has been used in many fields such as video summarization and video browsing. As the pivot of multimedia search engines, semantic video analysis aims to provide the semantic abstraction built on the original video data that is closer or even equal to the high-level understanding of human perceptual system. Both techniques help people to better un-

derstand and manage the multimedia material.

In addition to the above two major multimedia analysis techniques, affective computing is currently one of the active research topics, attracting increasingly intensive attention. This tremendous interest is driven by a wide spectrum of promising applications in many areas such as virtual reality, smart surveillance, perceptual interface, etc. As Picard [Pic00] chronicles in her paper, computing is not only a “number crunching” discipline, but also an interaction mechanism between humans and machines and sometimes even between humans. Trying to imbue computers with the human-like capabilities of observation, interpretation and generation of affective features [TT05], affective computing spans a multidisciplinary knowledge background such as psychology, cognition, physiology and computer science. It is very important for achieving harmonious human-computer interaction, by increasing the quality of human-computer communication and improving the intelligence of our computer system.

With the arrival of *affective computing*, affective video content analysis has come into being. Affective video content analysis makes use of both the psychological theories and content processing to detect the high level affect contained in the video. Compared to the traditional multimedia analysis techniques, this technique is better aligned with human’s perceptual mechanisms and the applications based on it thus tend to be more friendly, usable and natural. Till now, few works have been done on music and movie affective content analysis and the applications based on these technologies seem promising. For example, a video retrieval system with the help of *affective computing* may not only identify the scenes having your favorite actors, but it also can help people skip the *boring* scenes and fast forward to the most *exciting* or *interesting* scenes.

## 1.2 Overview

In the area of Human-Computer Interaction (HCI), *affective computing* is employed to help computers understand the humans’ affective state, and promotes the communication of machine and human beings. In order for a computer to correctly identify the affective state of people, two fundamental issues must be addressed. One of the two issues is how to represent the affective content, that is, to map the affective features to the

psychological model. In order to represent the affective content, Hanjalic and Xu [HX05] built a mapping from few low-level features to a 2D (arousal and valence) emotion space. However, this model is not complete because they only exploit four low-level features (out of at least 27 low-level features [ZTH<sup>+</sup>10]). Then, before representing the affective content of video, the second significant issue arises: what features are indeed related to the affect within a video. As one of the latest efforts on validating the relevant affective features, Zhang et al. [ZTH<sup>+</sup>10] have selected 13 arousal features and 9 valence features as described in their experimental results. The neglect of “control” component (it only reflects the distinction between two emotions which have similar valence and Arousal) of affect and user study for ground truth however make this result less objective. In addition, the subjectivity of humans also complicates this problem owing to the fact that different people could have different feelings for the same thing, which is quite common.

On the other hand, there are two main psychological models to represent the emotion: *dimensional emotion space* model and *categorical emotional states* model. The former one represents the emotion in a 3-dimensional space which are respectively “Valence” (V), “Arousal” (A) and “Control” (C). However, a 2-dimensional space which is represented by “Valence” (V) and “Arousal” (A) is more often used in research. The later one usually make use of some simple words such as “happy” and “sad” to describe the emotions. Many works have been done on classifying the emotions based on the low-level features. A variety of classifiers have been developed to solve the problem, for instance, Hidden Markov Model (HMM) [Kan03] and Bayesian Networks [TYA11]. However, all these works share one principal drawback: they fail to propose any computational approach to describe the *intensity* of emotion, instead of ill-defined adjectives, like “little”, and “very”.

Due to the importance of the two above mentioned issues and the lack of computational methods for describing the discrete emotion, in this thesis, our first work is to build a fundamental model which fills the gap of mapping the low-level features to the discrete emotional classes. We represent and model the affect with the sparsity-based framework considering the *categorical emotional states* psychological model. In parallel, we propose a computational and concise method to evaluate the *intensity* of each emotion. Second, we develop useful applications based on this fundamental theory: adaptive

presentation of home videos, a multimodal approach for online estimation of subtle facial expression, and social photo sharing.

### 1.3 Contributions

The main contributions of this thesis are as follows:

- An intuitive approach is proposed to map from low-level features and the “categorical emotional states” psychological model. This work fills *the gap* in the computational measurement of intensity of discrete emotional states.
- Our second work is the first work that proposed an affect-based approach for *adaptively* generating video presentation of *home videos* for *different* interested social group.
- Our third work is also the first work that introduced the *eye gaze* information into online estimation of *subtle* facial expression.
- Our last work is also the first work to utilize the *affect* factor of photos for the selection process, going beyond only facial expressions.

The remaining part of the thesis is organized as follow: Chapter 2 will provide a comprehensive literature survey on affect analysis of video. Chapter 3 will present a computational framework based on sparsity representation to represent and model the affect considering the *categorical emotional states* model. Chapter 4 will show an application about the adaptive representation of home videos based on affect. Chapter 5 will detail the multimodal approach for online estimation of subtle facial expression. Chapter 6 will elaborate on how to generate a suitable subset of photos from the personal photo collection for sharing with different social kinship groups, and how to check whether an individual photo is appropriate for sharing with a particular kinship group. Chapter 7 will draw the conclusion, followed by the future work.

## Chapter 2

# Literature Survey

Emotion is a complex psycho-physiological experience of an individual's state of mind as interacting with biochemical (internal) and environmental (external) influences. In humans, emotion fundamentally involves “physiological arousal, expressive behaviors, and conscious experience” [Mye04]. One question often asked is: How can these emotions be formally represented? In addition, another area within affective computing is the design of computational devices proposed to exhibit either innate emotional capabilities or the capability of convincingly simulating emotions. Thus, how to recognize these emotions is another issue. In the 2000s, research in computer science, engineering, psychology and neuroscience has been aimed at developing techniques that model emotions and recognize human affect. In the remaining part of this chapter, we first discuss the current main contemporary psychological models, followed by an analysis of the approaches proposed to recognize the emotions.

### 2.1 Emotional Models

In general, researchers have proposed two approaches to represent the psychological models of emotion. One of the two important and widely used psychological models is the “dimensional emotion space” model. As studied by Russell and Mehrabian [RM77], affect can be represented by three basic underlying dimensions as below:

- Valence - type of emotion;
- Arousal - intensity of emotion;

- Control - dominance;

For example, “anger” and “fear” have “negative” valence, while “joy” has “positive” valence. Arousal reflects the extent of reaction to stimuli from low to high. Additionally, control reflects the distinction between two emotions which have similar valence and arousal. Specifically, it reflects the emotional control on events and environments, ranging from feeling of overall lack of control to the opposite extreme of feeling in control. Therefore, the 3-Dimensional space consisting of V, A, and C as shown in Fig. 2.1 can represent the entire scope of human emotions as a set of points in this space. Furthermore, based on the fact that the “control” dimension has quite small and limited effect in characterizing various emotions [GCL89], the 2-Dimensional emotion space shown in Fig. 2.2 has often been used to model the smooth passage from one state to another in an infinite set of values [SYHH09, HX05]. Although the 2-D emotional model can represent rich affective states as pairs of (V, A), it is not easy for most people to vocalize their emotional experiences by describing the “Valence” and “Arousal”.

Instead, laypersons usually use simple words like “happy” to express their emotional experience. Consequently, an alternative model consisting of a set of discrete and distinct words has been proposed. This significant psychological model is named *categorical emotional states* model [TYA11]. In this area, the study of Ekman’s work [Ekm92] is one of the important basis for some of the recent research on emotions. He introduced six basic emotions: “happiness”, “anger”, “sadness”, “fear”, “disgust” and “surprise”, and any other emotions can be composed by a combination of these six basic emotions. What’s more, Ekman [Ekm93] proposed that an emotion should be considered to be a “family” since he and his colleague Friesen [EF78] showed that each emotion has not only one expression, but it has several related but visually dissimilar expressions. Likewise, Plutchik and Conte [RH97] developed the “Circumplex Model of Emotion” as shown in Fig. 2.3 which states that there are eight basic emotions: “anger”, “fear”, “sadness”, “disgust”, “surprise”, “anticipation”, “trust”, and “joy”. Surely, compared to the previous dimensional psychological model, this model also has an obvious drawback: how to computationally describe the *intensity* of emotion, instead of ill-defined adjectives like “little”, and “very”.

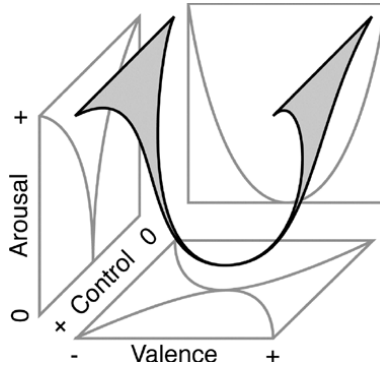


Figure 2.1: Illustration of the 3-D emotion space from [DL99]

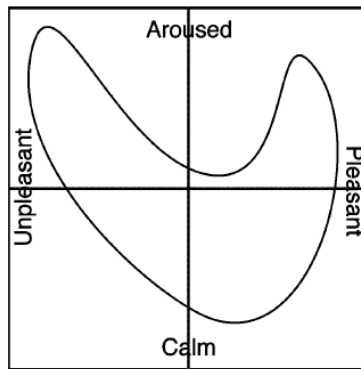


Figure 2.2: Illustration of the 2-D emotion space from [DL99]

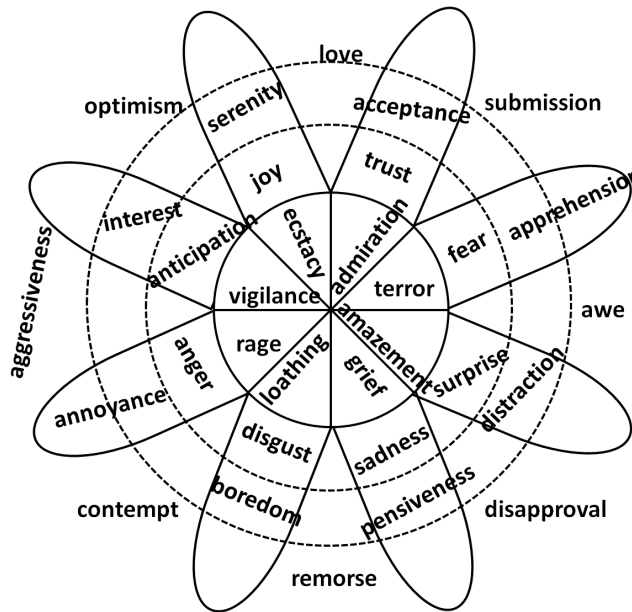


Figure 2.3: Illustration of Circumplex Model [TYA11].

## 2.2 Facial Expression Analysis

Human-computer *intelligent* interaction (HCII) is an emerging field aimed at providing natural ways for humans to use computers as aids. It is argued that for the



computer to be able to interact with humans, it needs to have the communication skills of humans. One of these skills is the ability to understand the emotional state of people. The most expressive way humans display emotions is through facial expressions. Therefore, extracting and validating emotional cues through the analysis of users' facial expressions is of high importance for improving the level of interaction in man machine communication systems.

Eisert and Girod [EG97] exploited Triangular B-Splines to construct a generic 3D head model and a table to describe the translation and rotation of the control points for the facial expressions according to the scheme proposed in [Sik97]. Their model reduced the computational complexity on estimating the facial movement and simplified the modeling of facial expressions. However, because a table built by them in advance was used to model the local movements, it could only deal with a small number of control points. With more points, the estimation of facial expression and movements can be more accurate. However, when they did the training and testing, they assumed that a person will display a neutral expression at the beginning of any video sequence. Although this assumption made the testing easy, it is in conflict with the reality in which any expression is possible in the video.

Black et al. [BY97] used the planar mode to recover qualitative information about the motion of the head, and used different parametric models to model the image motion of the facial features within local regions in space and time. Specifically, it used an affine model for eyes, and other affine models augmented with an additional curvature parameter for eye-brows and mouth during smiling. However, the system still imposed some limitations on the image sequences, such as transmission rate and larger image resolution. Meanwhile, their experimental design had some special challenges, for example, determining what expression was “actually” being displayed was difficult, because “different” expressions might appear quite similar leading to variation in human recognition of expressions. All of these limited the real time implementation of this method.

Cohn et al. [CZLK98] developed and implemented an optical flow based approach to detect the facial expression. Specifically, the first step was image alignment by which they mapped the face image to a standard face model based on three facial feature points. And next, they marked the key feature points in the first digitized frame manually. Thirdly,

they used a hierarchical optical flow method to automatically track feature points and get the displacements. Finally, they used different discriminant function analysis in each facial region. This system was sensitive to subtle motions in facial displays with high accuracy, and it already can deal with limited out-of-plane face. However, it needed manual marking of the feature points in the first frame, which is tedious.

Cohen et al. [CGH00] used a multilevel HMM (Hidden Markov Models) architecture to automatically do the segmentation and recognition of the facial expressions from live video input taking advantage of the temporal cues. The novelty of their architecture was that both segmentation and recognition of facial expressions were done automatically using a multilevel HMM architecture while increasing the discrimination power between the different affective classes. However, a database of only five people was used to demonstrate their system, which is too small.

Lien et al. [LKCL00] developed and implemented the first version of a face image analysis system (showed in Fig. 2.4) to detect, track and classify subtle changes in facial expression with convergent methods which utilized multiple types of feature information. It can automatically code input face image sequences into Facial Action Coding System (developed by Ekman in 1978 [EF78]) action units which were the smallest visibly discriminant changes in facial expression. However, it also needed some pre-processing to manually mark features, though marking of features in the initial frame was partially implemented. In addition, only small set of prototypic expressions can be recognized.

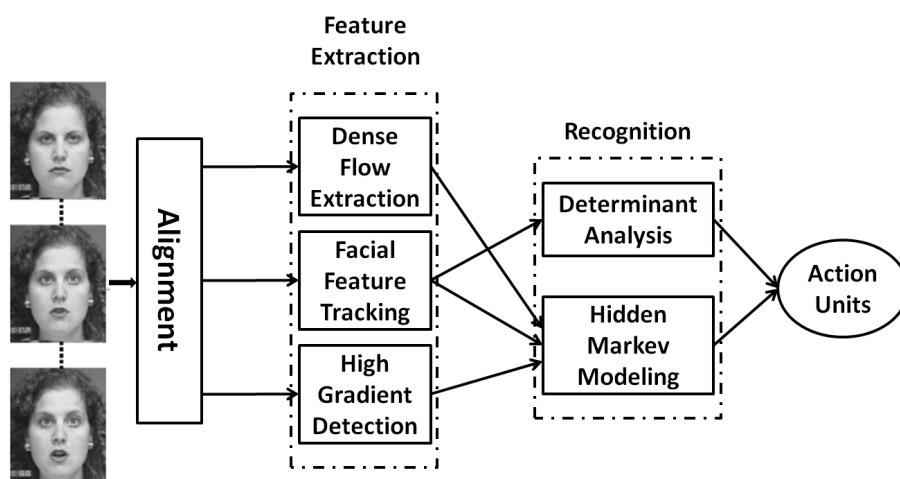


Figure 2.4: Overview of face images analysis system in [LKCL00]

Pantic et al. [PTR01] combined several distinct extraction techniques into a hybrid, knowledge-based approach to extract mouth features from facial images. Firstly, the Region of Interest (ROI) in the input facial image was determined by the color-based segmentation technique. After getting the right region, a function to get the position of ROI was applied. Secondly, the Curve Fitting of the Mouth and Mouth template Matching was used to localize the mouth contour in the input ROI. And thereafter the mouth movement can be classified. Finally, the four salient mouth feature points: top of the upper lip, bottom of the lower lip, left and right mouth corners were extracted respectively. However, this study can only deal with limited out-of-plane head rotations and sequences starting with an expressionless mouth appearance.

Tian et al. [TKC01] developed an Automatic Face Analysis system to analyze facial expressions based on both permanent facial features (eyebrows, eyes, and mouth) and transient facial features (brows, cheek, and furrow) in a nearly frontal-view face image sequence. In their work, they used the method described in [RBK96] to automatically extract the region of face and approximate the location of individual face features. Next, it needed manual adjustment of the contour of the face features and components in the initial frame. Then, multistate models of facial components were used to detect and track both transient and permanent features. On the other hand, for transient features, a Canny edge detector was used to quantify the amount of and orientation of furrows. Finally, they designed three-layers neural network with one hidden layer to recognize action units (AUs) by a standard back-propagation method. The system is shown in Fig. 2.5. However, there were some drawbacks which limit the real-time use of this system. First, it still needed manual adjustment for the contour of features. Second, it did not consider the large head motion. Last, in their experiment, the used image sequences began with a neutral face.

Cohen et al. [CSG<sup>+</sup>03] proposed two approaches to classify the facial expressions from the static and dynamic orientations respectively. They designed the face tracking subsystem based on the Piecewise Bezier Volume Deformation tracker. More specifically, they computed the 3D motions by the 2D image motions which are modeled as the projections of the true 3D motions and measured using template matching between frames. The authors proposed the Tree-Augmented-Naive Bayesian with Gaussian distribution

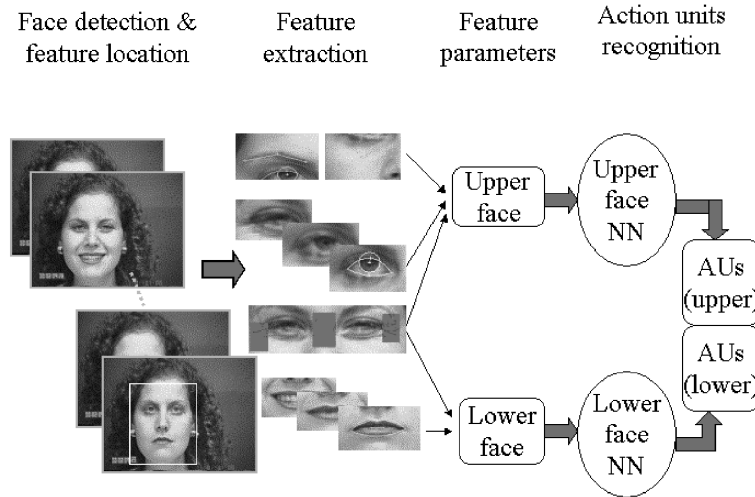


Figure 2.5: Feature-based automatic facial action analysis system in [TKC01]

which was considered as a static classifier. On the other hand, the authors also proposed multilevel Hidden Markov models (HMMs) as a dynamic approach to classify the facial expressions. However, although they mentioned that the system could be changed to adapt to the situation where a person can go from one expression to another without passing through a neutral expression, it was not done.

Heishman et al. [HDW04] identified eye region biometrics, that is, the fatigue and engagement for the interest region, within a particular HCI (Human Computer Interaction) context (e.g., video security system monitoring). In their work, they used five subjects to identify those biometrics that produced meaningful and measurable responses within the prescribed HCI scenarios. They designed four experimental sessions: Fatigued/Disengaged, Fatigued/Engaged, Non-Fatigued/Disengaged, and Non-Fatigued/Engage), and used the Eye Region Biometric Processing System (ERBPS) written by themselves to process the extracted video frames. The significant biometrics were found, manually analyzed, and used as input into the Fatigued/Engaged Matrix. However, in their work, the video needed manual processing and analysis using the ERBPS. Thus, if the test data set was very large, it was not practical. Moreover, it did not utilize the potential biometrics.

Cunningham et al. [CKBW04] discussed the necessary and sufficient facial motions for nine conversational expressions (agreement, disagreement, disgust, thinking, happy, sadness, surprise, clueless, and confusion). They first used the Max Planck Insti-

tute [KWB04] to record the facial expressions of six different people. After that, the sequences would be post-processed so that the selected regions associated with the expressions (mouth, eyes, eye-brows) was replaced with a static snapshot. They utilized a custom, image-based, stereo motion-tracking algorithm to recover the 3D location of the tracking target, and acquired a 3D model of the participant's head with a Cyberware 3D laser range scanner to determine the relative location of target to the individual's head. In their experiment, they set up the relationship by manual interactive initialization on the first frame of each recorded sequence. For the selected regions and the frozen regions, the final model was rendered with an alpha value of 0 and 1 respectively using the texture maps which refer to the texture map of the final 3D shape model and image pixels in the video footage. Although this method can detect many conversational expressions, their experimental conditions were stringent, which is not proper for the use of real system. In addition, the 3D model of each individual was required to be built in advance, which is tedious.

Ioannou et al. [IRT<sup>+</sup>05] developed an expression recognition system which could be robust to facial expression variations among different users, and evaluated facial expressions through the robust analysis of appropriate facial features. Finally, a neurofuzzy system was created, which was based on rules defined through analysis of Facial animation parameter (FAP) variations both in the discrete emotional space, as well as in the 2D continuous activation-evaluation one. This neurofuzzy system was allowed for further learning and adaptation to specific users' facial expression characteristics, measured through FAP estimation in real life application of the system, using the analysis of clustering of the obtained FAP values (the FAPs were defined by the ISO MPEG-4 standard). However, this system did not work well in terms of the real-time performance. An overview of the facial analysis and feature extraction system is given in Fig. 2.6.

Shan et al. [SGM06] proposed a Bayesian approach to modelling dynamic facial expression temporal transitions for a more robust and accurate recognition of facial expression given a manifold constructed from image sequences. Fig. 2.7 shows the flow chart of the proposed approach. They first derived a generalized expression manifold for multiple subjects, where Local Binary Pattern (LBP) features were computed for a selective but also dense facial appearance representation. Supervised Locality Preserving

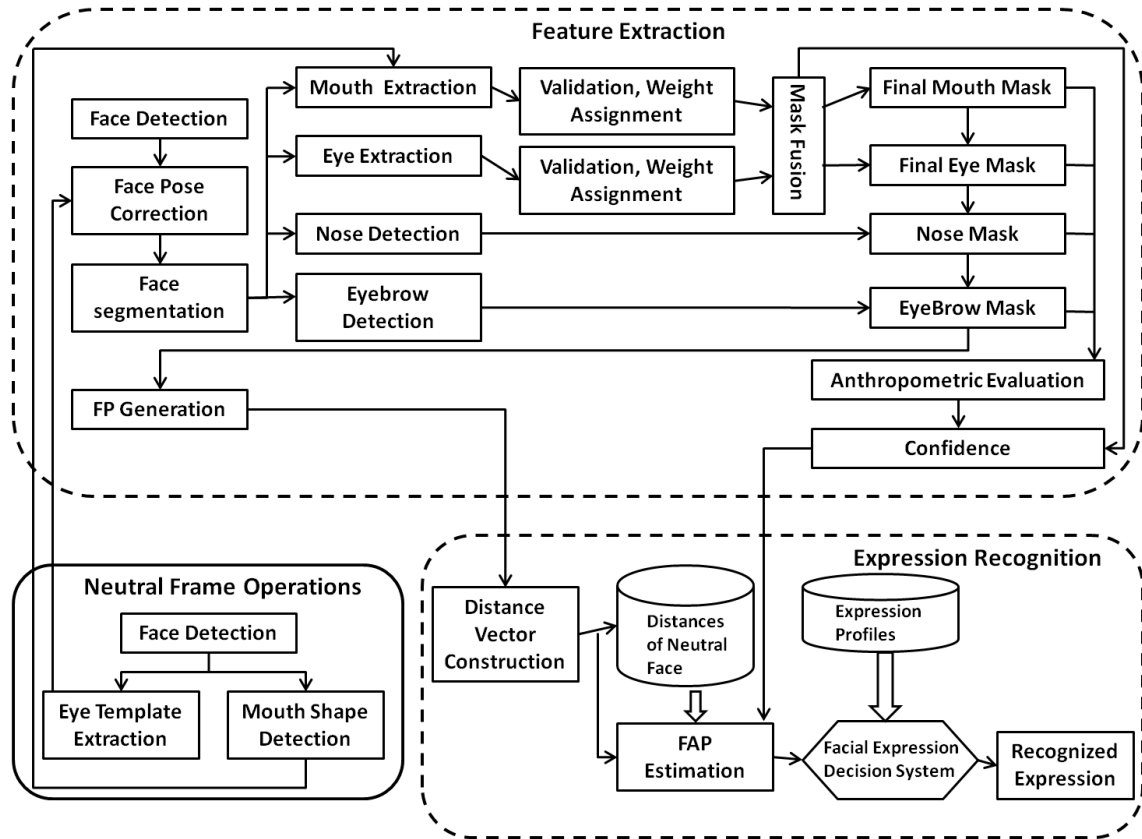


Figure 2.6: The facial feature extraction and facial expression analysis system in [IRT<sup>+</sup>05]

Projections was used to derive a generalised expression manifold from the gallery image sequence. Then, they formulated a Bayesian temporal model of the manifold to represent facial expression dynamics. For recognition, probed image sequences were first embedded in the low dimensional subspace and then matched against the Bayesian temporal manifold model. However, it required manual marking of features and preprocessing of the images.

Yeasin et al. [YBS06] first used a biologically-motivated face detector to detect and segment faces from the rest of the image. Second, the computed optical flow between consecutive frames of the sequence was projected to a lower dimensional space using the PCA (Principal Component Analysis). Third, the projected motion patterns were fed to a bank of linear classifiers to assign class labels from the set of universal expressions to each image of the sequence. The output of linear classifiers over a sequence of images was coalesced together to form a temporal signature. Fourth, the generated temporal signature was used to learn the underlying model of six universal facial expressions. Dis-

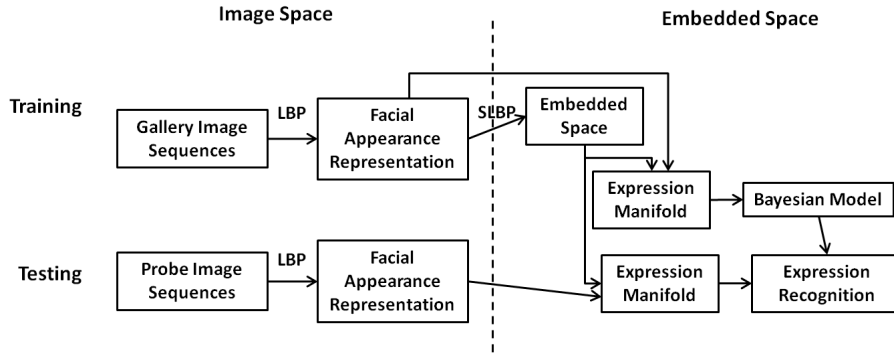


Figure 2.7: A Bayesian temporal manifold model of dynamic facial expressions in [SGM06]

crete HMMs were used in learning the models for facial expressions. Finally, recognized facial expressions was mapped to compute levels of interest based on 3-D affect spaces. However, the experimental database was generated by themselves, which is limited for the implementation of approach in real world.

Ying et al. [YWH10] proposed a new approach for facial expression recognition based on fusion of sparse representation. Specifically, the sparse representation were employed in both raw gray images and LBP of these images. The final recognition results were obtained by fusing this two sparse representation. However, the used test data set was too small.

Pai and Chang [PC11] presented a novel facial expression recognition scheme based on extension theory [Wan05]. Feature invariant approaches were employed to detect and segment the facial region, while the positions of lips were extracted as the features of face. Finally, the classification of facial expressions was performed by evaluating the correlation functions. However, only few emotions were classified and few facial features were considered.

Sandbach et al. [SZPR12] proposed a method that exploited 3D motion-based features between frames of 3D facial geometry sequences for dynamic facial expression recognition. GentleBoost (GB) classifier and HMM were used to recognize the onset/off-set temporal segments and model the full expression dynamics respectively. However, GB classifier can not capture the variability in the motion.

Almaev et al. [AV13] developed the novel dynamic appearance descriptor named Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) for automatic

facial expressions recognition in real-time. Combining the spatial and dynamic texture analysis with Gabor filtering, their LGBP-TOP method is relatively robust to face registration errors caused by rotational alignment errors. However, few action units were tested for this proposed approach.

Suja et al. [STD14] implemented two separate systems to recognize the facial expression from the face images. Considering neural network and K-nearest neighbor as classifiers, they used the Dual-tree complex wavelet transform and Gabor Wavelet Transform method respectively for the extraction of feature vectors from cropped face and whole face. However, both the training dataset and test dataset were small.

Totally, facial expression recognition contains three main components: face detection, feature extraction and expression classification. From our above survey, we make a table shown in Table 2.1 to summarize the existing facial expression recognition algorithms.

## 2.3 Multimodal Human’s Emotion Analysis

As introduced in [Pic00], “affective computing” should be thought of as an interfacing means between humans and machines and sometimes even between humans themselves. To achieve this, application design must take into account the ability of humans to provide multimodal input to computers, thus “moving away from the monolithic window-mouse-pointer interface paradigm and utilizing more intuitive concepts, closer to human perceptual mechanisms. A large part of this naturalistic interaction concept is expressivity, both in terms of interpreting the reaction of the user to a particular event or taking into account their emotional state and adapting presentation to it, since it alleviates the learning curve for conventional interfaces and makes less technology-savvy users feel more comfortable” [CMK<sup>+</sup>06]. As shown in the previous section, most facial expression analysis systems focus on facial expressions to estimate emotion-related activities. Furthermore, the introduction and correlation of multiple channels may increase robustness, as well as improve interpretation disambiguation in real-life situations. Multimodal emotion recognition is therefore currently gaining ground.

Zeng et al. [ZTL<sup>+</sup>04] presented their effort towards audio-visual HCI-related affect recognition. In their work, a tracking algorithm called Piecewise Bezier Volume Defor-



Reference	Model		EM	Methods Used	Comments
	2D	3D			
Eisert & Girod [EG97]	✓		-	Triangular B-spline/Optical Flow	No rigid head rotations
Black et al. [BY97]	✓		6	Affine Model	1)Limited rigid head motion; 2)Prior knowledge of face location
Cohn et al. [CZLK98]	✓		-	Optical Flow	1)No rigid head motion; 2)Limited out-of-plane motions; 3)Manually mark feature points.
Cohen et al. [CGH00]	✓		6	HMMs/ML Classifier	A small database
Lien et al. [LKCL00]	✓		-	Wavelet motion/HMMs	1)Manually mark feature points; 2)limited head rotations
Pantic et al. [PTR01]	✓		-	Color-based technique/Curve Fitting of Mouth/Template Matching	1)Limited out-of-plane motion; 2)Require no expression in first frame
Tian et al. [TKC01]	✓		-	Canny Edge Detector/Neural Network	1)Manually adjust feature points; 2)Limited out-of-plane motion; 3)Neutral face in first frame
Cohen et al. [CSG <sup>+</sup> 03]	✓		6	Naive Bayesian Classifier	Neutral state between two emotions
Cunningham et al. [CKBW04]		✓	9	Cyberware 3D Laser	1)Manual initialization; 2)Small dataset
Ioannou et al. [IRT <sup>+</sup> 05]	✓		-	Neurofuzzy Model	Limited real-time performance
Shan et al. [SGM06]	✓		6	Bayesian Approach	The features are manually marked
Yeasin et al. [YBS06]	✓		6	Optical Flow/PCA/HMMs	Low classification rates on fear and disgust emotion
Ying et al. [YWH10]	✓		7	Sparse Representation	The test data is too small
Pai and Chang [PC11]	✓		3	Extension Theory	1) few emotions are classified; 2) few facial features are used.
Sandbach et al. [SZPR12]		✓	6	GentleBoost classifier/HMM	GB Classifier can not capture the variability in the motion.
Almaev et al. [AV13]	✓		-	Local Gabor Binary Patterns/SVM	The performance only tests few Action Units.
Suja et al. [STD14]	✓		6	Neural Network/K-Nearest Neighbor	The training data set and test data set are both small.

Table 2.1: Summarization of facial expression recognition algorithms. Notes: “2D” = “Planar Model”; “3D” = “3-Dimensional head model”; “EM” = “Emotions”

mation tracking was applied to extract facial features in their experiment. An optical flow method was applied to track these AU movements as facial features. The movements of facial features are related to both affective states and content of speech. Therefore, based on the assumption that the influence of speech on face features is temporary, and the influence of affect is relatively more persistent, a smoothing method was applied to reduce the influence of speech on facial expression to some extent. They used three kinds of prosody features for affect recognition: logarithm of energy, syllable rate, and two pitch candidates and corresponding scores, and applied *Sparse Network of Window* to build two affect classifiers individually based on face-only and prosody-only features. Finally, they applied a voting method to combine the classification outputs from face and prosody modalities. Compared with the four previous reports of bimodal affect recognition, those which contributed to this field include the following points. Firstly, more affective states are analyzed, especially including four HCI-related affective states (confusion, interest, boredom, and frustration) besides the basic emotions. Secondly, more subjects are tested which improve the generality of their algorithm. Thirdly, they consider the fact that a facial expression is influenced by both an affective state and speech content, and apply a smoothing method to reduce the influence of speech on facial expression to some extent. However, their tracking results are very sensitive to the initial frame, because the face tracker they used required that the expression of the initial frame is neutral with closed mouth. Also, only person-dependent experiments were done.

Gunes and Piccardi [GP05] presented an approach to automatic visual emotion recognition from two modalities: expressive face and body gesture. In their work, face and body movements were captured simultaneously using two separate cameras. For each face and body image sequence, single “expressive” or “apex” frames were selected manually for analysis and recognition of emotions using Weka, a tool for automatic classification [WFT<sup>+</sup>99]. Individual classifiers were trained from individual modalities for uni-modal emotion recognition. They fused facial expression and affective body gesture information at the feature and at the decision-level. Finally, they further extended the affect analysis into a whole image sequence by a multi-frame post integration approach which chose the emotion with the maximum amount of recognized frames as the “as-

signed emotion” or final decision for a whole video over the single frame recognition results. In their experiment, they created their own bi-modal database by capturing

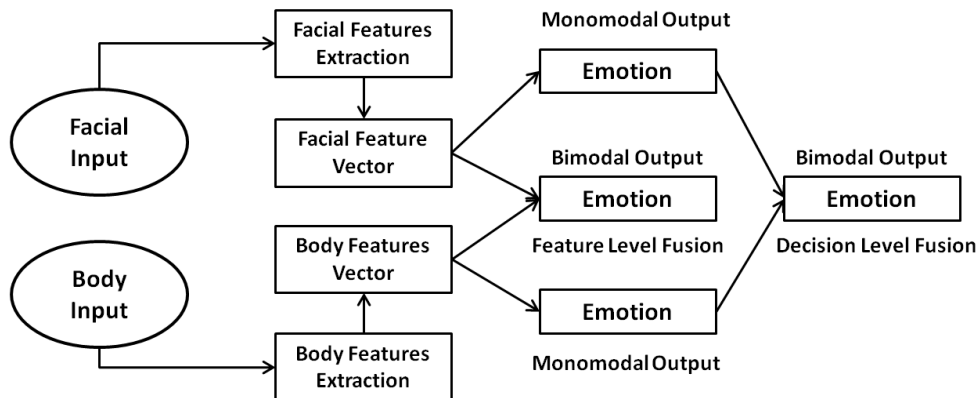


Figure 2.8: The system framework for mono-modal and bi-modal emotion recognition in [GP05]

face and body simultaneously from 23 people using two cameras (as shown in Fig. 2.8), since they were not able to find a publicly available database with bi-modal expressive face and body gesture. Based on the survey asking the participants to evaluate their own performance, a number of recorded sequences were treated as outliers and not included in their work by which the experiment results were more accurate. However, it was an extra task to manually select the neutral frame and a set of previous frames for feature extraction and tracking. In their experiment, the training and test datasets were person-dependent with just four subjects, which influences the generality of the system. And few hand gestures and postures were considered.

Jaimes et al. [JNL<sup>+</sup>05] examined the affective content of meeting videos. First they asked five subjects to manually label three meeting videos using continuous response measurement (continuous-scale labeling in real-time) for arousal (excited/calm) and valence (pleasure/displeasant) (the two dimensions of the human affect space). Then they automatically extracted audio-visual features to characterize the affective content of the videos. Finally, they compared the results of manual labeling and low-level automatic audiovisual feature extraction.

However, in the visual analysis step, when they applied the Visual Trigger Templates framework to detect large posture changes, which may indicate interest level changes. The used templates were manually constructed. So it limits the size of dataset of testing

video. In this study, the techniques they used were simple, and they only considered low-level features of audio-visual features. Therefore, there may be scope to improve their work.

Caridakis et al. [CMK<sup>+</sup>06] described a multi-cue, dynamic approach in naturalistic video sequences using the Valence-Arousal space as the representation of emotion. Specifically, the framework of the recognition of facial expressions is described in Fig. 2.9. They first located the face to estimate the approximate facial feature locations from the head position and rotation, and the head was segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose and mouth. In this stage, because the naturalistic video can have some frames without face, they applied the nonparametric discriminant analysis with a Support Vector Machine (SVM) to classify face and non-face areas. They chose the MPEG-4 Facial Animation Parameters (FAPs) to measure the deformation of these feature points and identify the expressions. Then, they fused intermediate feature masks of every isolated area to generate the final mask. Finally, 19 feature points(FPs) were extracted from the final mask, and the FAPs were obtained compared to FPs from the neutral frame. They also exclusively analyzed the vocal expressions based on prosody and related to pitch and rhythm. They extracted several types of features: pitch interval based features, nucleus length features and distances between nuclei by analyzing each tune with a method employing prosodic representation based on perception called “Phonogram”. The fusion of visual and acoustic features was performed on a frame basis, meaning that the values of the segment-based prosodic features were repeated for every frame of the tune considering preserving the maximum of the available information. The final recognition was performed via a “Simple Recurrent Network” which lends itself well to modeling dynamic events in both user’s facial expressions and speech.

However, in their facial expression recognition, the system required the neutral frame for the subject because of the use of MPEG-4 FAPs. Here, the neutral frame means that the subject’s expression is neutral in that frame. So they needed to manually select the neutral frame from video sequences to input into the system, which increases the extra work.

Caridakis et al. [CCK<sup>+</sup>07] presented a multimodal approach for the recognition of

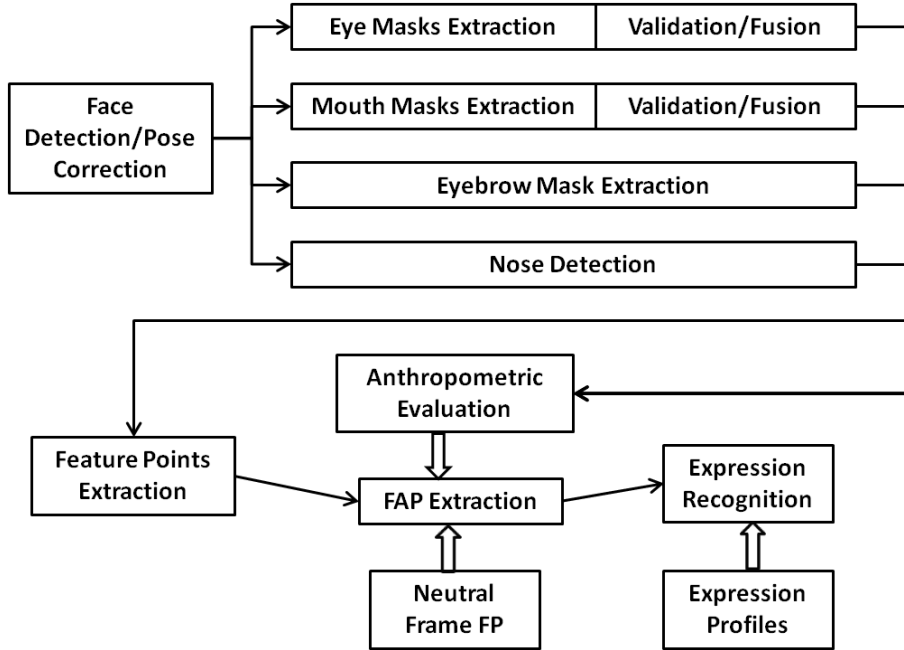


Figure 2.9: Diagram of the proposed methodology of [CMK<sup>+</sup>06]

eight emotions that integrated information from facial expressions, body movement and gestures and speech. A Bayesian Classifier was trained and tested for each modality. Finally, both feature-level fusion and decision-level fusion were exploited on these multimodal data. However, few samples were tested for each emotion.

Nicolaou et al. [NGP11a] proposed a method for continuous prediction of spontaneous affect from multiple cues and modalities in Valence-Arousal Space. Using bidirectional Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) technique, facial expression, shoulder gesture and audio cues were fused for dimensional and continuous prediction of emotions in valence and arousal space. However, the set of subjects was small.

Prado et al. [PSLD12] made use of the dynamic Bayesian network to classify the emotion from facial expression and vocal expression. Based on the recognized emotion result from face and voice, a Bayesian Mixture Model Fusion method was employed to do the final decision of emotional state. However, the experimental data are collected from the pre-defined good environment, which differs the real environment. Thus, maybe it is not suitable for the implementation in the real environment.

Chen et al. [CTLM13] proposed a novel framework to model the temporal dynamics information of face expression and body gesture. Employing the Histogram of Oriented

Gradients (HOG) on the Motion History Image and Image-HOG features, this framework made use of SVM classifier to classify the emotion of subject into six basic emotions. However, the classification rates on sadness and surprise were relatively low. Also, both training set and test set were small.

The emotion of people can be reflected by many channels: facial expression, body language, physiological signal, etc. On the whole, combining different signals can improve the final accuracy of emotion recognition. Table 2.2 summarizes the above reviewed papers.

Reference	Sources			Fusion		Methods	Comments
	FE	SP	BG	DL	FL		
Zeng et al. [ZTL <sup>+</sup> 04]	✓	✓		✓		Optical Flow	Requires initial frame
Gunes & Piccardi [GP05]	✓		✓	✓	✓	Bayesian Network	1)Manual selection of neutral frame; 2)Small subject dataset size
Jaimes et al. [JNL <sup>+</sup> 05]		✓	✓	✓		Continuous Response Measurement	1)Limited dataset; 2)Small user set
Caridakis et al. [CMK <sup>+</sup> 06]	✓	✓			✓	SVM/Recurrent Network	1)Requires neutral frame; 2)Manual selection of neutral frame
Caridakis et al. [CCK <sup>+</sup> 07]	✓	✓	✓	✓	✓	Bayesian Classifier	Test samples for each emotion are too few.
Nicolaou et al. [NGP11a]	✓	✓	✓	✓	✓	Neural Network/SVR	Only four subjects are tested.
Prado et al. [PSLD12]	✓	✓		✓		Dynamic Bayesian network	Experimental data is collected in good environment.
Chen et al. [CTLM13]	✓		✓		✓	SVM	1)The classification rates on sadness and surprise are relatively low; 2) the training set and test set are both small.

“FE” = “Facial Expression”; “SP” = “Speech”; “BG” = “Body Gestures”; “DL” = “Decision Level”; “FL” = “Feature Level”;

Table 2.2: Summarization of multimodal user’s emotion analysis

## 2.4 Affective Content In Videos

Based on psychological theories and models mentioned above, the emerging area of affective video content analysis hopes to enable computers to recognize the emotions and affect contained in videos. As a relatively new multimedia analysis technique, it faces some challenges. One critical issue is to understand the mapping from the low-level features and the affect in terms of different psychological models, that is, how to recognize the emotions by the affective features.

One significant work based on “dimensional emotion space” model was reported by Hanjalic and Xu [HX05]. In their computational framework, the affective content of a given video clip was defined as the intensity and type of feeling or emotion. They extracted the features from both audio and visual signals to model Arousal and Valence components considering the 2-dimensional emotional psychological model. Besides, they combined the obtained Arousal and Valence time curves into the affect curve that can serve to determine the prevailing mood per segment of a video.

Specifically, they proposed a function integrating these three components: motion, rhythm and sound energy, and evaluated it on a number of representative test sequences. Therefore, they modeled the arousal time curve in general as a function of three components. Valence is modeled similarly. Considering that their models need to be psychologically justifiable because arousal and valence are psychological categories, the components selected to model the Arousal and Valence should satisfy: comparability, compatibility, and smoothness. Thus, they chose the motion component, rhythm component, and sound energy component as the low-level features. On the other hand, for simplicity, they just chose the Pitch-Average component as the low-level feature. Thus, specifically, the complete arousal model is defined as follow: %begin{equation}

However, they don’t describe how to select the general user, which is important because the curves elicited from him/her will be different based on the differences among users. Also, only three low-level features to model the Arousal and one for Valence is perhaps not enough as [SVE<sup>+</sup>12] mentioned that there at least 1841 low-level features. More low-level features should be taken into account. Additionally, when they integrated the feature function, the weight  $w_i$  used in formula is not validated in any study, and

at the end of this paper, the authors mentioned that the relations known so far are rather vague and therefore difficult to map onto reliable models for arousal or valence components which perhaps have the possibility for further improvement of the obtained representation in searching for more concrete relations between the affect dimensions (arousal and valence) and low-level features.

Based on the work of Hanjalic and Xu, Sun et al. [SYHH09] proposed an improved 2-dimensional Valence-Arousal emotional space to represent and recognize the affective video content. The V-A emotional space was divided into a set of typical fuzzy emotional subspaces representing the certain discrete affective states. Subsequently, a Gaussian Mixture Model (GMM) was employed to determine the maximum membership principle and the threshold principle which represent and recognize the affective video content. Consequently, Soleymani et al. [SKCP09] introduced a Bayesian classification framework for affective video tagging that allows taking contextual information into account. In their method, informative features extracted from three information streams -video (visual), sound (auditory), and subtitle (textual)- were linearly combined to compute the arousal at the shot level using a relevance vector machine. Consequently, the Bayesian classification based on the shots arousal and content-based features allowed tagging these scenes into three affective classes: positive excited, negative excited and calm. Zhang et al. [ZTH<sup>+</sup>10] built a Support Vector Regression (SVR) model for Arousal and Valence respectively to map the features to the affective states. In their paper, they extracted quite rich audio-visual features, and employed SVR model with RBF kernel to select the most effective features. All the above mentioned papers have done the classification on valence and arousal dimension respectively, assuming the valence and arousal are independent. However, various psychological findings indicate that these affective dimensions, such as valence, arousal, and control, are correlated; therefore, the affective video content analysis based on the “dimensional emotion space” model has a new trend lately to consider the correlations between these dimensions. In order to model inter-dimensional correlations, Nicolaou et. al [NGP11b] proposed a novel, multi-layer hybrid framework utilizing a graphical model named Auto-Regressive Coupled HMM (ACHM-M) for emotion classification by geometric features based on symmetric spatio-temporal characteristics of facial expressions.



Reference	Model		Features			EMs	Methods
	DM	CM	V	A	T		
Kang [Kan03]		✓	✓			3	HMMs
Hanjalic & Xu [HX05]	✓		✓	✓		-	Linear weight Function
Sun & Yu [SY07]		✓	✓	✓		4	HMMs
Sun et al. [SYHH09]	✓		✓	✓		-	GMM
Soleymani et al. [SKCP09]	✓		✓	✓	✓	-	Bayesian Classifier
Zhang et al. [ZTH <sup>+</sup> 10]	✓		✓	✓		-	SVR
Teixeira et al. [TYA11]		✓	✓	✓		6	HMMs / Bayesian Network
Nicolaou et. al [NGP11b]	✓				✓		ACHMM
Xu et al. [XWH <sup>+</sup> 12]		✓	✓	✓	✓	5	HMMs
Cui et al. [CLT <sup>+</sup> 13]	✓		✓	✓			SVR
Acar et al. [AHA14]	✓		✓	✓			SVMs

Note: “DM” = “Dimensional Model”; “CM” = “Categorical Model”; “V” = “Visual”; “A” = “Audio”; “T” = “Texture”; “EMs” = “Emotions”. This table is organized by their publication year.

Table 2.3: Summarization of the related work of affective content in videos.

In addition to the works on the “dimensional emotion space” model, many works have put efforts on the “categorical emotional states” model, since this model is easier to articulate the emotional experience. Kang [Kan03] first performed an empirical study on the relation between the emotional events and low-level features (color, motion and shot cut rate) in terms of manual labeling of the training data, and then constructed Hidden Markov Model (HMM) to detect affective events contained in video. In their paper, three types of emotions: joy, sadness and fear were classified. However, only three emotional events were detected. Compared to their work, an affective video content representation and recognition framework presented by Sun and Yu [SY07] could detect one more emotional event. In their work, affective video content units of different granularities were first located by excitement intensity curves, and then the selected affective content units were used to construct video affective tree. Based on the excitement intensity curve, the affective intensity of each unit at various levels of video affective tree can also be quantified into several levels from weak to strong. Many middle level audio and visual affective features, which represent emotional characteristics, were designed and extracted to construct observation vectors. Based on these observation vector sequences, HMM-based affective video content recognizers were trained and tested to recognize the basic emotional events of audience (joy, anger, sadness and fear). What’s more, Teixeira

et al. [TYA11] created the affective model from real data acquired through a series of user experiments to reflect the affective state of a viewer after they watched the video clip. Bayesian network topology and Hidden Markov Models were employed to recognize six emotional events in their work. Recently, additional information is taken advantage of to recognize more emotions. Xu et al. [XWH<sup>+</sup>12] proposed a three-level affective content analysis framework using the textual information except the audio and visual features. Specifically, they obtained a mid-level representation like dialog, audio emotional events and textual concepts from machine learning on low-level features, and then infer high-level affective content with the assistance of these mid-level representations. Cui et al. [CLT<sup>+</sup>13] has employed support vector regression (SVR) to model arousal and valence of affect based on the audio-visual features. Acar et al. [AHA14] utilizes multi-class support vector machines (SVMs) for the affective classification of music video clips in the Valence-Arousal space. However, the classification accuracy is not very high (below 60%).

In this part, we can find that a variety of methods are proposed to represent and detect the affective content within videos. The reviewed papers are summarized in Table 2.3.

## 2.5 Summary

Affective video content analysis has been proposed to help people to better understand the semantics of video, and help make applications more friendly, and natural. In this chapter, we presented a survey on the psychological emotional model and affective video content analysis, respectively.

## Chapter 3

# Sparsity-based Affect Representation And Modeling

### 3.1 Introduction

Affective computing [Pic00] is currently an active research area, due to the increased users' expectation of natural interaction with computers. Affective video content analysis is an important sub-area that makes use of both the psychological theories and computational methods to recognize the high level affective content present in videos. It is better aligned with humans' perceptual mechanisms, which enables more friendly and usable applications.

We first present some background information on “what is the affect of a video clip” and “how to represent the affect with the psychological models”. In modern psychology, the affective domain represents one of the three divisions: the cognitive, the conative, and the affective [McK76]. Affect is the experience of feeling or emotion. In this chapter, we define the affect (emotion) of a video clip as the type of emotion that is expected to arise in the viewers when they are watching that video clip. The expected emotion refers to the one that is either intended to be felt by the viewers (by the video creator), or felt by the most viewers who are watching the video clip.

“Dimensional emotion space” and “categorical emotional states” are the two most widely used psychological models. The dimensional emotion space model considers the emotion space as a 3-dimensional space of valence, arousal and control. In the categorical

emotional states model, emotional experiences are represented by a set of discrete and distinct words such as “happy”, “sad” and “angry”. We choose this model because it is very natural for us to relate to these categorical states and hence it is very intuitive. How to model and represent these emotional categories in video is a challenge. Moreover, this model has an obvious drawback – it is not clear how one can compute the “intensity” of emotion, instead of using ill-defined adjectives like “little”, and “very”. We present a solution to this problem as well in this chapter. We take a sparse representation based approach [Can06] in this chapter.

How to best map the low-level video content features (such as color and motion) into the discrete emotional states, and explicitly determine the extent of each emotional state is the most significant objective of this chapter. As stated in [Zet12], colors or particular color groups can influence our emotions, and the intelligent use of colors can produce a variety of specific overall emotional effects. Specifically, warm colors are perceived to possess high energy and excite us, but cold colors are of low energy and calm us down. The deviation around the main “hue” is what determines the warmth or coldness of a color. This was extensively studied by Rudolf Arnheim, a well-known perception psychologist and art theorist [Zet12]. He found that cold colors of less saturation can dampen the mood of people, whereas highly saturated warm colors can excite them. Therefore, low-level features like the color content are related to the emotions conveyed by the visual component of video. However, the relationship between the low-level features of videos and the expected emotions elicited in humans is still not well understood. How does the combination of low-level features contribute to affect is still an open problem in affective computing, and most existing approaches have not yielded good results. It is also very difficult to determine if the number of features and the construction of features are sufficient to recognize the affect within video. For sparse representation, as long as the number of features employed is large enough, even randomly chosen features are sufficient to recover the sparse representation (i.e. recover the important information related to affect in our problem) [Can06]. Sparse representation offers a new perspective on feature selection – it shows that the number of features is much more important than the details of how they are constructed [YWMS07]. Therefore, we use many features resulting in a high-dimensional space from which we extract

the right sparse representation. Interestingly, [WMM<sup>+</sup>10] has argued that “the sparse representation to uncover semantic information derives in part from a simple but important property of the data: although the images (or their features) are naturally very high dimensional, in many applications images belonging to the same class exhibit *degenerate structure*”. Given the fact that humans agree on a small set of adjectives for emotional experiences across languages, cultures and ages does point to the existence of some basic degenerate structure. Thus, sparse representation can be taken advantage of to capture/recover the basic characteristic of each emotion. Our findings show that the sparsity based approach is indeed effectively able to represent the categorical emotion model. It corroborates the utility of the sparse representation for extracting semantic information [WMM<sup>+</sup>10]. It must be noted that many of these features have been used separately in the past, motivated by psychological considerations.

In this chapter, we propose a computational framework to link the affective features with the emotional states considering the psychological model of “categorical emotional states”. We also try to address the lack of an intensity measure in this categorical psychological model. We develop a sparse vector representation in this computational framework, with a method to compute the “intensity” of the emotions. In addition, we show how to obtain the representative sparse vectors from the low-level features extracted from video. The approach is flexible - features extracted from any modality (audio, visual, dialog and even subtitles) can be used in this representation framework. The key contributions of this chapter are:

- A simple, fast and intuitive method is proposed to map the low-level features to the “categorical” emotional states.
- A computational measure is proposed to capture the “intensity” of “discrete” emotional states.

This chapter is organized as follows. Section 3.2 reviews the related work to serve as a preamble. Section 3.3 elaborates the sparse representation and modeling of affective content within videos, and discusses the construction of sample matrix. Section 3.4 describes the relevant experimental results. Finally, conclusions are drawn in Section 3.5.

## 3.2 Related Work

As discussed in subsection 2.4, [HX05, SYHH09, SKCP09, ZTH<sup>+</sup>10, NGP11b, CLT<sup>+</sup>13, AHA14] took advantage of different techniques to analyze the arousal and valence factors of emotion in the 2D *dimensional emotion space* psychological model. In addition, many efforts have been put on the *categorical emotional space* psychological model. [Kan03, SY07, TYA11, XWH<sup>+</sup>12] developed different classifiers, such as HMMs, GMM, and Bayesian network, to identify the discrete emotional states.

## 3.3 Methodology

Emotions are an integral part of story-telling via videos. For example, movie directors are always concerned whether a shot evokes what they expect the audience should feel at that moment [BTA90], which indicates the central role of emotion for videos.

As observed by Sun and Yu [SY07], Zhang et al. [ZTH<sup>+</sup>10], Kang [Kan03] and Hanjalic and Xu [HX05], the affect within a video clip can often be captured by the low-level features, such as color features, MFCCs, and sound energy. For example, Sun and Yu [SY07] have validated that motion, shot length and sound energy are robust enough to characterize the “happy” content in a video. In order to represent the affective content of a video clip, these features therefore can be extracted to form a feature vector whose “*emotional property*” is defined as the corresponding emotional state induced in viewers when they are watching that video clip. If multiple emotions are elicited on viewers, then the “emotional property” refers to the dominant emotion among them.

[Phi99] stated that “When it comes to expressing emotions, members of widely different cultures have much in common,..., Such findings imply that beneath all the cultural complexity of mankind, there is a core of basic emotional expression that is understood all over the world.” This is also evidenced by the fact that most languages have distinct words for the common emotions. Therefore it is reasonable to assume that there is a universally shared core information hidden behind each emotional expression. Movie directors therefore use a broad but common set of audio-visual recipes for evoking different kinds of emotions. It must be recognized that it is broad enough to allow for tremendous creativity in conveying the same emotion. Yet there is often a common discernible

thread. Thus, each emotion has core features or the combination of features that differ from other emotions, e.g. sound energy to distinguish “happy” and “sad” states. Then, we can further assume that the vectors of core features extracted from different affective videos are distinctive and separated, while the feature vectors from videos with the same affect are more similar. This is like a “template pattern” in the template matching problem. However, template matching methods do not perform well, for example: [JJVS09] has a problem that because there are several times of project (3D to 2D, then 2D to 3D to find the motions), the obtained features is not so reliable. Based on this hypothesis, we can define that the representative feature vector of each emotion to be the feature vector consisting of the core features, and which is extracted from a video clip with only that emotion. Then, any feature vector representing one kind of emotion can only be represented as a linear combination of the corresponding representative feature vectors. However, as the emotions elicited by a video clip are mixed in general, which has been verified by [Pet09] stating “...They can have many emotions at the same time. Emotions are mixed in...emotions are possibly composed of many more components...”, we have to note that the feature vector generally contains more than one emotion. We call this typical situation as the “practical case” while in the “ideal case”, the feature vector contains only one emotion.

In the remaining part of this section, we first present a general overview of sparse representation in the subsection 3.3.1, followed by the details on how to exploit sparsity to represent and model the affective video content in the subsection 3.3.2. Finally, in the subsection 3.3.3, we discuss the conditions that the sample matrix needs to satisfy in order to obtain a better representation and we discuss methods for constructing the “optimal” sample matrix.

### 3.3.1 Overview of Sparse Representation

The algorithmic problem of computing sparse linear representations has seen a recent surge of interest in the statistical signal processing area [Can06]. In all of what follows, we will adopt a somewhat abstract and general point of view about the sparse representation. And then we will relate it to our problem in the next sub-section.

The representation and compression of the signal is the original objective of studying

the sparse representation. The basic idea of compressive sampling is that for certain types of signals, just a small number of nonadaptive samples carries sufficient information to approximate the signal well. Research in this field can be categorized into two classes: sampling and reconstruction. The way to construct a signal approximation given a vector of noisy samples is the major algorithmic challenge in this research field. Thus an arbitrary discrete signal  $f$  is sparse and compressive if the signal value sequence of  $f$  is concentrated on a small set and a small number of samples are enough to approximate the signal well. We say that the signal  $f$  is  $s$ -sparse when  $\|f\|_0 \leq s$ , where  $\|\cdot\|_0$  denotes the  $\ell^0$ -norm, i.e., it counts the number of non-zero entries in a vector. In the sensing mechanism, the partial information about  $f$  is obtained by linear functionals recording the values.

$$y_i = \langle f, \psi_i \rangle, \quad i = 1, \dots, k \quad (3.1)$$

Eq.(3.1) can also be written as Eq.(3.2) where  $\Psi$  is the  $k \times n$  matrix with  $\psi_i$  as rows or equivalently.

$$y = \Psi f \quad (3.2)$$

The above equation can also be explained as the action of a downsampling matrix  $\Psi$  on the target signal  $f$  in the theory of compressive sampling. Given  $y$ , we can solve Eq.(3.2) to find the original signal  $f$ . However,  $\Psi$  is a real  $k \times n$  matrix and  $k < n$ , the number of solutions of Eq.(3.2) is infinite. The objective of sparse representation is to find a sparse signal  $\tilde{f}$  to represent the original signal  $f$  by choosing the solution of the optimization problem under the  $\ell^0$ -norm (or  $\ell^1$ -norm, or  $\ell^2$ -norm).

$$\ell^i : \tilde{f} = \arg \min_f \|f\|_i, \quad \text{Subject to } y = \Psi f, i = 0, 1, 2 \quad (3.3)$$

Since the signal  $f$  is compressive, and even sparse, the approximation  $\tilde{f}$  is close enough to represent  $f$  accurately based on the above discussion. Next, we will discuss how to employ the “sparse representation” into the analysis of affective content within videos.



### 3.3.2 Representation And Modeling

Let us first consider the case where the test video clip can elicit only one emotion from the viewers. Then given the test feature vector  $y \in \mathfrak{R}^k$ , which is a  $k$ -dimensional feature vector that represents the affective content of a test video clip. We assume there are  $m$  basic emotional states (emotional categories). Let  $\alpha_{j,i} \in \mathfrak{R}^k$  for  $i = 1, \dots, n_j$  be the representative feature vectors of the  $j$ th emotional state, where  $n_j$  is the cardinality of this set of representative feature vectors. We put the representative feature vectors of the  $j$ th emotional state into a  $k \times n_j$  matrix  $A_j$ , which is called the “sub-sample matrix” and  $A_j = [\alpha_{j1}, \dots, \alpha_{jn_j}]$ . Accordingly, we form a  $k \times n$  matrix  $A = [A_1, \dots, A_m]$ , which is called the “sample matrix”. Then, we have  $n = \sum_{j=1}^m n_j$ . If we write  $y$  as a linear combination of the representative feature vectors, then for each emotional state we can have:

$$y = \alpha_{j1}\beta_{j1} + \alpha_{j2}\beta_{j2} + \dots + \alpha_{jn_j}\beta_{jn_j} = A_j\beta_j \quad (3.4)$$

where  $\beta_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jn_j}]^T \in \mathfrak{R}^{n_j}$ , for  $j = 1, \dots, m$ , and  $\beta_j$  is the linear coefficient vector.

Furthermore, we have assumed that any feature vector representing an emotional state can be represented only as a linear combination of the corresponding representative feature vectors. As a consequence, if  $y$  belongs to the  $q$ th emotional state, where  $q \in \{1, \dots, m\}$ , then, the linear coefficient vector  $\beta_j$  for  $j = 1, \dots, m$  is equal to the zero vector  $\mathbf{0}$  except  $\beta_q$  as shown in Eq.(3.5). Moreover, based on Eq.(3.5), we can write  $y$  as a linear combination of the representative feature vectors of all the emotional states using Eq.(3.6) as follows.

$$\beta_j \begin{cases} = \mathbf{0} & j \neq q \\ \neq \mathbf{0} & j = q \end{cases} \quad (3.5)$$

$$y = A_1\beta_1 + \dots + A_q\beta_q + \dots + A_m\beta_m = Ax \quad (3.6)$$

where,  $A = [A_1, \dots, A_m]$

$$x = [\beta_1, \dots, \beta_{q-1}, \beta_q, \beta_{q+1}, \dots, \beta_m]^T$$

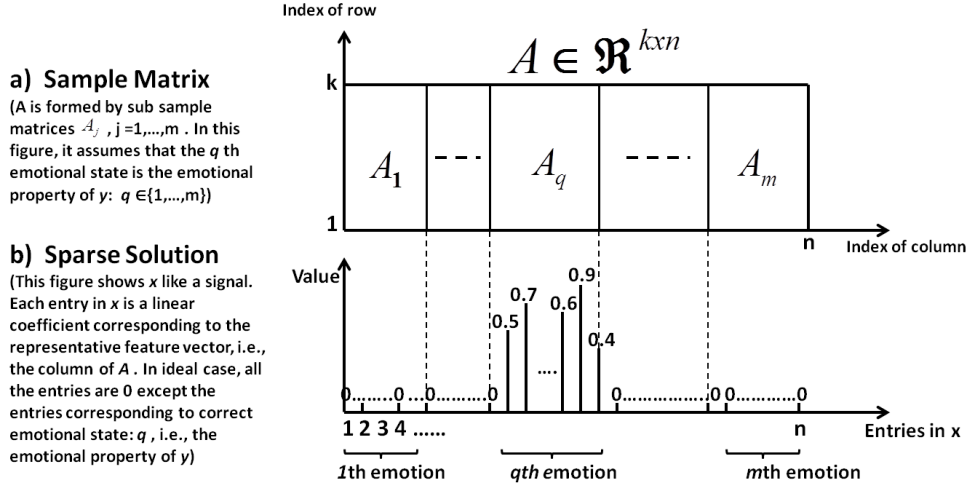


Figure 3.1: An example for the “ideal case” of the relationship between the entry values of  $x$  and each column of sample matrix  $A$  based on the sparse representation:  $y = Ax$ .

where,  $A$  is the complete “sample matrix”, and  $x \in \mathbb{R}^n (n = \sum_{j=1}^m n_j)$  which is called the “sparse solution” is a sparse vector within which the entries should be zero, except for the ones associated with the  $q$ th emotional state. Specifically, let us treat  $x$  as a signal. When there are a large number of representative feature vectors for each emotion ( $n_j$  for  $j = 1 \dots m$  is large), the signal  $x$  is very long because of the large  $n$ . However, the signal  $x$  only has  $s$  non-zero coefficients ( $s = \|\beta_q\|_0 \leq n_q \ll n$ ), while Eq.(3.5) is satisfied. Therefore,  $x$  is sparse and compressive because its non-zero coefficients are concentrated on a small set. Referring to the previous discussion about sparse representation,  $x$  can be compared to  $f$  with  $A$  being the downsampling matrix  $\Psi$ ,  $y$  being the partial information signal of  $f$ . We therefore aim to find  $x$  by solving Eq.(3.6). Intuitively, all the non-zero coefficients of the solution of Eq.(3.6) should only correspond to the columns of  $A_q$ , which happens in the “ideal case”, that is, when Eq.(3.5) is satisfied. Accordingly, the location of the non-zero entries of  $x$  predicts the emotional property of the test feature vector  $y$  – the emotional state whose corresponding entries in  $x$  are non-zero. Fig. 3.1 shows an example to visually depict the relation of the entry values of  $x$  to each sub-sample matrix in  $A$  in the ideal case.

However, Eq.(3.6) is under-determined when  $k \ll n$ , which is more typical in practice, that is, the number of solutions of Eq.(3.6) is infinite. When we solve this equation by considering the  $\ell^1$ -minimization problem using COSAMP [NT09], we can obtain the

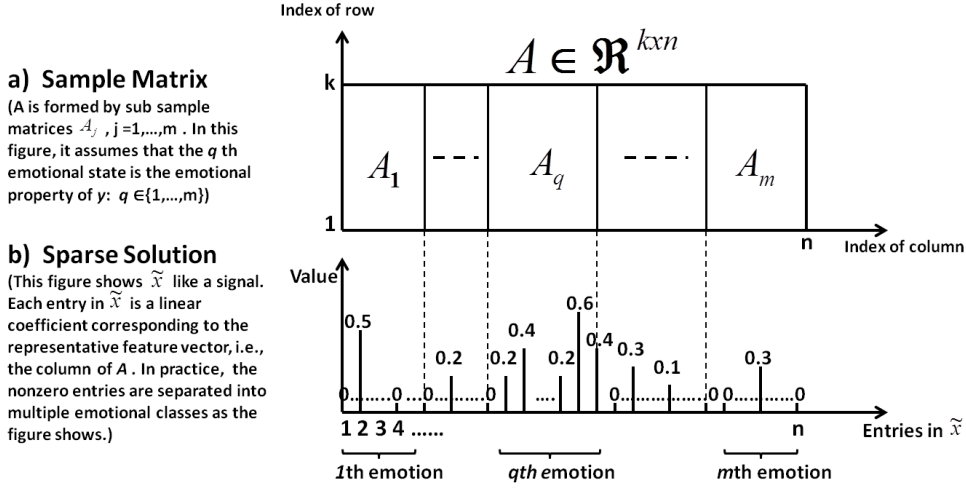


Figure 3.2: An example for the “practical case” of the relationship between the entry values of  $\tilde{x}$  and each column of sample matrix  $A$  by solving  $y = Ax$  using the COSAMP [NT09].

approximation  $\tilde{x}$  of  $x$ .

$$\ell^1 : \tilde{x} = \arg \min_x \|x\|_1, \text{ Subject to } y = Ax \quad (3.7)$$

It is possible that  $\tilde{x}$  cannot guarantee that its non-zero coefficients would only be associated with the columns of  $A_q$ , that is, the coefficients of  $\tilde{x}$  do not satisfy Eq.(3.5). Other than the under-determined nature of the linear equation, two other reasons can also explain this phenomenon. First, the noise, such as environmental noise and camera motion, possibly impacts the values of the extracted features. Second, it could be an instance of the “practical” case. The test video clip represented by  $y$  simultaneously contains several emotions, such as happiness mixed with sadness in case of nostalgia. Therefore, the non-zero entries of  $x$  are separated into several emotional categories. Fig. 3.2 shows the visualization of the practical case of the solution  $\tilde{x}$ .

We need to discuss about how to determine the emotional property of  $y$  and the intensity of each emotion within  $y$  when the coefficients of its corresponding  $\tilde{x}$  do not satisfy Eq.(3.5). We focus on the dominant emotion. We would like to guess that the columns from the “dominant” (or say “main” or “correct”) emotional category in  $A$  should have the most contribution into  $\tilde{x}$ ; therefore, we not only capture how well the entries in  $\tilde{x}$  are associated with each “sub-sample matrix”, but also discuss the intensity

of each emotion within  $y$ .

For the first point, we introduce the function denoted by  $\Phi_j(x)$  that returns a new vector which is composed of all the coefficients of  $x$  corresponding to  $A_j$ , and  $\Phi_j(x) \in \mathfrak{R}^{n_j}$ . Then, let  $\tilde{y}_j = A_j \Phi_j(\tilde{x})$ , for  $j = 1, \dots, m$ , represents the approximation of  $y$  for the  $j$ th emotional state in terms of  $\tilde{x}$ . The difference  $\varphi_j$  between  $y$  and its approximation  $\tilde{y}_j$  can be computed by Eq.(3.8).

$$\varphi_j = \|y - \tilde{y}_j\|_2 = \|y - A_j \Phi_j(\tilde{x})\|_2 \quad (3.8)$$

The value  $\varphi_j$  can be interpreted as how close or how well the coefficients within  $\tilde{x}$  are associated with the sub-sample matrix  $A_j$  for  $j = 1, \dots, m$ . Namely, the smaller  $\varphi_j$  is, the closer  $y$  and  $\tilde{y}_j$  are, which means the representative feature vectors of the  $j$ th emotional state match better with  $y$ . We therefore determine the ‘‘correct’’ emotional property of  $y$  is the  $q$ th emotional state using Eq.(3.9).

$$q = \arg \min_j \varphi_j = \arg \min_j \|y - A_j \Phi_j(\tilde{x})\|_2 \quad (3.9)$$

This classification rule is corroborated by the experimental results in Section 3.4.

From another angle, the difference  $\varphi_j$  represents the degree of match between  $y$  and the columns of  $A_j$ . In other words, the smaller the difference  $\varphi_j$  is, the more significant is the  $j$ th emotional state’s (representative feature vectors) contribution to  $y$ . Obviously, the intensity of an emotional state is directly proportional to the importance of this emotional state within  $y$ . It means the intensity of the  $j$ th emotional state is inversely proportional to the difference  $\varphi_j$ . In order to computationally describe the intensity of the  $j$ th emotional state, let  $\boldsymbol{\varphi} = [\varphi_1, \dots, \varphi_m]^T \in \mathfrak{R}^m$  be the residual vector of  $y$ . Then, we can compute the intensity  $\Upsilon_j$  of the  $j$ th emotional state within  $y$  using Eq.(3.10).

$$\Upsilon_j = 1 - \frac{\varphi_j}{|\boldsymbol{\varphi}_{\max}|} \quad (3.10)$$

Here, the  $\Upsilon_j$  is scaled to the range between 0 (weakest) and 1 (strongest).

### 3.3.3 Sample Matrix

In previous sub-section, we have detailed the sparse representation and modeling of the affective content within videos, as well as the recognition of emotion and the computation of its intensity. However, we do not discuss how to construct the sample matrix  $A$ , which is the main focus of this section. Specifically, the question is: how to find the representative feature vectors of the  $j$ th emotional state for  $j = 1, \dots, m$ ? However, since we intend to solve Eq.(3.6) instead of Eq.(3.4), we only need to consider the overall sample matrix  $A$  instead of the sub-sample matrix  $A_j$  for  $j = 1, \dots, m$ . Yet, the construction possibilities and choices of  $A$  are infinite. Accordingly, the solutions  $\tilde{x}$  based on different  $A$  are distinct. Naturally, we want the best solution  $\tilde{x}$ . Also, the sample matrix  $A$  which enables Eq.(3.7) to generate the best solution is called the “optimal” sample matrix. In the remaining part of this subsection, 3.3.3.1 will discuss the properties that the “optimal”  $A$  should satisfy mathematically; and 3.3.3.2 will describe two methods for constructing  $A$ .

#### 3.3.3.1 Property of Sample Matrix

Our final aim is that the solution  $\tilde{x}$  to the equation  $y = Ax$  should be exact, i.e.,  $\tilde{x} = x$  in terms of the “optimal” sample matrix  $A$ . Therefore, in our model, we would like to answer the question: how well can  $\tilde{x}$  recover (or approximate) the signal  $x$  (if we treat the sparse solution  $x$  as a signal) that is sparse or compressive in terms of varied  $A$ ? In other words, we have to specify the conditions that the sample matrix  $A$  should satisfy such that  $\tilde{x}$  will best recover  $x$ . So, we have

**Theorem 3.3.1 ([CT05])** *Assume that  $x$  is  $s$ -sparse and suppose that  $\delta_{2s} + \delta_{3s} < 1$  or, better  $\delta_{2s} + \theta_{s,2s} < 1$ . Then the solution  $\tilde{x}$  to Eq.(3.7) is exact, i.e.,  $\tilde{x} = x$ .*

Candès and Tao [CT06] proposed the above strong condition of sample matrix, which is called uniform uncertainty principle (UUP) and defined in [CT05]. The UUP essentially states that the  $k$  by  $n$  sample matrix  $A$  obeys a “restricted isometry property (RIP) [CT05]” which makes the exact solution possible. In other words, the geometry of sparse signals should be preserved under the action of the sample matrix. To quantify this idea, let  $A_T$ ,  $T \subset \{1, \dots, n\}$  be the  $k \times |T|$  submatrix obtained by extracting the

columns of  $A$  corresponding to the indices within  $T$ . Then, they defined the  $s$ -restricted isometry constant of a matrix  $A$  as the least number  $\delta_s$  for which

$$(1 - \delta_s)\|x\|_2^2 \leq \|A_T x\| \leq (1 + \delta_s)\|x\|_2^2 \text{ whenever } \|x\|_0 \leq s \quad (3.11)$$

$\|\cdot\|_2$  denotes the  $\ell^2$ -norm. For all subsets  $T$  with  $|T| \leq s$  and coefficient sequences  $(x_j)_{j \in T}$ , Eq.(3.11) holds. The above equation implies that each collection of  $s$  columns from  $A$  approximately behaves like an orthonormal system. In short, if the UUP holds at the sparse level  $s$ , it can be proven that exact recovery is possible via the  $\ell^1$ -norm minimization.

Therefore, we can say: if the columns of sample matrix obeys RIP, then we can obtain the best recovery of  $x$ , i.e. the best solution  $\tilde{x}$  of Eq.(3.6). However, when  $s$  is very large, it is computationally difficult and time-consuming to check whether a matrix provably satisfies the UUP or not. We know nothing about the value of  $s$ .  $A$  is a  $k \times n$  matrix, which means a total of  $n$  representative feature vectors are selected from all the emotional states. Then, a natural question now arises: how many representative feature vectors are necessary to acquire  $s$ -sparse signal? That is, what is the possible relationship between  $s$  and  $n$ , even when counting the  $k$  (the number of features extracted)? Needell and Tropp [NT09] argued that the minimum number of basic samples  $n \geq 2s$  on account of the following simple argument: two different  $s$ -sparse signals cannot be mapped to the same one. Therefore, each collection of  $2s$  columns from the sample matrix must be non-singular. We need to formalize this intuition.

We would like the conditions in Theorem 3.3.1 to hold for large values of  $s$ , ideally of the order of  $k$ ! How to construct one sample matrix  $A$  so that any collection of  $s$  columns from  $A$  is almost orthogonal? And what are the possible values of  $s$ ? Although it is computationally difficult to check if a sample matrix provably satisfies the UUP for large  $s$ , we know that trivial randomized constructions will do so with overwhelming probability [NT09]. The high-dimensional sphere is mostly empty, it is possible to pack many vectors while maintaining approximate orthogonality, which is due to the ‘‘blessing of high-dimensionality’’ [Don00][Can06]. Therefore, the conditions  $\delta_{2s} + \delta_{3s} < 1$  and  $\delta_{2s} + \theta_{s,2s} < 1$  hold for  $s = O(k/\log(n/k))$  with probability  $1 - O(\exp(-\gamma^n))$  for some  $\gamma > 0$ ,

when  $n$  vectors (i.e.,  $n$  feature vectors) are obtained on the unit sphere of  $\mathbb{R}^k$  (i.e.,  $k$  features extracted) independently and uniformly at random, see [Can06]. Consequently, we can say that: if  $n, k, s$  have the following relationship:  $s = O(k/\log(n/k))$ , then we can obtain the exact recovery of  $x$ , i.e., the conditions in Theorem 3.3.1 hold for  $A$ .

In summary, if one sample matrix  $A$  of dimension  $k \times n$  satisfies one of the following conditions.

- $A$  obeys the UUP.
- $s = O(k/\log(n/k))$

Then, the best recovery of  $x$  occurs, i.e.,  $\tilde{x} = x$  [Can06]. Our original aim is thus achieved.

### 3.3.3.2 Construction of the *optimal* Sample Matrix

In our modeling, “optimal” refers to the situation when the sample matrix can help Eq.(3.7) obtain the best solution. In other words, the solution  $\tilde{x}$  better represents the original  $x$  when the “optimal” sample matrix is used. It is easy to see that a sample matrix  $A$  satisfying one of the above mentioned conditions is “optimal”. Therefore, when we have a sample matrix, checking whether it satisfies one of the two conditions is a possible way to verify whether the sample matrix is *optimal* or not.

As discussed earlier, the individual representative feature vector was not assumed to have any requirement or any particular semantic meaning. They are typically chosen from standard bases such as Fourier, Gaussian, Binary or even generated from random matrices [CT06][Can06][WYG<sup>+</sup>09]. Two methods therefore can be made use of to construct an optimal sample matrix. They are:

1. Form  $A$  by taking all the training samples which is an over-complete database.
2. Form  $A$  by sampling  $n$  columns vectors uniformly and randomly on the unit sphere of  $\mathbb{R}^k$ .

For Method (1), the reason for  $A$  being optimal is based on the following argument. “The sparse representation is naturally discriminative: among all subsets of base vectors, it selects the subset which most compactly expresses the input signal and rejects all other

possible but less compact representations” [WYG<sup>+</sup>09]. If there are sufficient training feature vectors for each emotional state, it has a large probability to represent the test feature vector as a linear combination of just these training feature vectors with the same emotional property. Therefore, this representation is naturally sparse, as it only employs a very small subset of the overall training feature vectors. Wright et. al. [WYG<sup>+</sup>09] argue that in many interesting problems, the sparsest linear representation can occur in terms of this over-complete database, and the recovery of  $x$  is efficient via  $\ell^1$ -minimization. Therefore, the sample matrix constructed by Method (1) can yield the best approximation of  $x$ , as well as be “optimal”.

For Method (2), there is a claim [BDDW08][CW08][WMM<sup>+</sup>10] using fairly standard results in probability theory that, with overwhelming probability, the matrix  $A$  of dimension  $k \times n$  generated by Method (2) obeys the “Restricted Isometry Property (RIP)” (i.e., the condition of Theorem (3.3.1) provided that

$$k \geq c_1 \cdot s \log(n/s) \tag{3.12}$$

where  $c_1$  is some constant depending on each instance [BDDW08][CW08]. It means that Theorem 3.3.1 holds for this sample matrix  $A$ , that is, the columns of sample matrix  $A$  are orthogonal, and the exact recovery of  $x$  occurs. Therefore,  $A$  is optimal. The inequality can be written as:

$$n \leq s \cdot 10^{c_2 \cdot k/s} \tag{3.13}$$

where  $c_2 = 1/c_1$ . The probability of constructing a sample matrix which does not obey the “Restricted Isometry Property” is exponentially small in  $k$ ; see also [CW08]. Additionally, the equation  $s = O(k/\log(n/k))$  can also be expressed as Eq.(3.14):

$$n = k \cdot 10^{c \cdot k/s} \tag{3.14}$$

where  $c$  is a constant. Consequently, as long as the relationship of  $n$ ,  $k$  and  $s$  satisfies this equation when the sparsity  $s$  is very large, the sample matrix  $A$  is also optimal based on the discussion in the previous section.



Emotion	Happy	Anger	Sadness	Fear	Disgust	Neutral	Tenderness
$N_{shots}$	171	285	327	317	276	96	258
$N_{scenes}$	132	201	227	229	180	58	187

Table 3.1: The number of shots and scenes for each emotion.

In summary, Method (1) and (2) can both be utilized to construct an optimal sample matrix. On the other hand, there is a small probability that the matrix may not be optimal.

### 3.4 Experiments

Ideally, it would be best to use video clips with pure basic emotions to construct sample matrix. However, it is very hard to obtain video clips with only one single emotion since emotions are very subjective. There is no standard public database of such affective videos. As discussed in the previous section, the sample matrix does not have any requirement on the feature selection but there is a relationship between the sparsity factor ( $s$ ), the number of features ( $k$ ), and the number of training samples ( $n$ ). As long as the sample matrix contains sufficient (over-complete) samples, it will work well.

To the best of our knowledge, Schaefer et al. [SNSP10] provide the only publicly available affective videos data-set. Therefore, our testing database is composed of the database provided by [SNSP10] and 7 neutral video clips of our own. A total of 73 video clips are used in this work. [SNSP10] have provided a database of brief video clips intended to elicit emotional states in experimental psychology experiments. In order to generate this database, they invited fifty film experts to recall specific video scenes that elicited fear, anger, sadness, disgust, amusement, tenderness, as well as emotionally neutral scenes. For each emotion, the 10 most frequently mentioned scenes were selected and cut into video clips. Therefore, these video clips can be used as a database for our research. However, they do not provide neutral video clips. Although the neutral clips are missing, Schaefer et al. provided the details of these neutral clips which are validated by fifty film experts. Thus, we download the neutral clips from YouTube based on these descriptions. Furthermore, we also invited ten people to rate them in order to

ensure that these neutral clips are indeed reliable. The neutral clips are used in our experiments when ten people all vote it with a “neutral” label. We take advantage of the shot segmentation technique [CKP03] to segment each clip into several shots labeled with the corresponding emotion. Besides, we also segment these video clips into scenes using the method [ZMM95], and extract the visual-audio features for each scene based on the shot level. Finally, a total of 1730 shots and 1214 scenes are obtained, and the number of the shots for each emotion in our experiments is listed in Table 3.1. Here, please note that the shots of length less than 1-second are discarded in our experiments in order to ensure the reliability of the extracted features.

For our experiments, we extract the features from the visual component as well as the audio component. Color is a very important cue of the emotional events [BTA90][Kan03] in the visual component. To compute the color feature, each frame in the RGB color space is transformed into the HSV color space, and then the pixel values are quantized into 11 culture colors such as red, green, blue, yellow, brown, purple, orange, gray, pink, black and white [Gol99]. For the key frame of each shot, we compute the histograms of 11 culture colors, as well as the mean and covariance of color histogram, saturation and value separately. A total of 17 features are extracted and composed into a visual feature vector for each key frame. Finally, we average the visual feature vectors of key frames to get one final visual feature vector denoted by  $\boldsymbol{\nu}$  for each shot.

It is well known that the associated audio can contain significant affective information [XWH<sup>+</sup>12]. Therefore, the 19 audio features, such as pitch average, silence ratio, short time energy, zero-crossing rate, brightness, rolloff and 13 MFCCs, are computed to construct the audio feature vector, which is denoted by  $\boldsymbol{v}$ . Therefore, two kinds of feature vectors (visual feature vector:  $\boldsymbol{\nu} \in \mathbb{R}^{17}$  and audio feature vector:  $\boldsymbol{v} \in \mathbb{R}^{19}$  in our experiments) are constructed and normalized for each shot. Here, we have to point out that the features that we use are not arbitrary but they have been informed by past research [BTA90][Kan03] and they all capture some psychophysiological aspects of affect. And effectively the sparsity approach provides a principled approach to reduce the dimensionality of feature-based representation, which is validated in our experiments.

Method		EM						
		HA	AN	SA	FE	DI	NE	TE
Audio	NN	0.14	0.2	0.17	0.25	0.19	0.06	0.22
	SVM-SC	0.28	0.66	0.38	0.59	0.36	0.62	0.38
	Our Method	0.27	0.43	0.78	0.38	0.54	0.58	0.51
Visual	NN	0.87	0.77	0.75	0.79	0.64	0.8	0.72
	SVM-SC	0.91	0.84	0.83	0.87	0.81	0.98	0.84
	HMMs [Kan03]	0.87	-	0.31	0.55	-	-	-
	Our Method	0.88	0.87	0.88	0.93	0.77	0.95	0.91
Feature Fusion	NN	0.87	0.78	0.7	0.79	0.62	0.83	0.71
	SVM-SC	0.91	0.89	0.83	0.95	0.79	0.95	0.84
	Our Method	0.89	0.93	0.88	0.94	0.77	0.98	0.91
Different weight ratio for decision-level fusion using our method								
$w_1/w_2= 1/9$		0.29	0.56	0.87	0.55	0.63	0.68	0.62
$w_1/w_2= 2/8$		0.33	0.70	0.92	0.74	0.66	0.77	0.77
$w_1/w_2= 3/7$		0.43	0.78	0.92	0.83	0.71	0.81	0.85
$w_1/w_2= 4/6$		0.60	0.82	0.94	0.87	0.74	0.84	0.87
$w_1/w_2= 5/5$		0.69	0.84	0.94	0.90	0.77	0.88	0.88
$w_1/w_2= 6/4$		0.81	0.85	0.94	0.94	0.80	0.88	0.90
$w_1/w_2= 7/3$		0.87	0.87	0.92	0.94	0.79	0.95	0.91
$w_1/w_2= \mathbf{8/2}$		0.88	0.88	0.92	0.95	0.79	0.95	0.91
$w_1/w_2= 9/1$		0.88	0.87	0.91	0.94	0.79	0.95	0.90

Table 3.2: Recognition Results based on different shot-level features and fusion level. The bold decision-level ratio is the “optimal” ratio in our experiments.

### 3.4.1 Over-complete Database

We mentioned in previous section that the sample matrix  $A$  formed by all the training samples of an over-complete database is the optimal sample matrix. Therefore, before we demonstrate the sparse representation of affective video content, we should verify whether the database we use in our experiments is over-complete or not since we want to make use of the Method (1) to form the sample matrix  $A$ . We use a heuristic method for this purpose. If the classification rate tends to be stable with the increase of training samples, then we establish that this set of training samples is an over-complete data set.

We divide this database into two components: the test samples and the training samples. Also, we use different ratios of these two components. The ratios (the number of the test samples : the number of the training samples) that we test are 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, and 9:1. This experiment is summarized in Algorithm 1. In this experiment, the feature vector we use is the combination of visual and audio features:  $[\nu, \mathbf{v}]^T$ . Finally, we show the results in Fig. [3.3- 3.11].

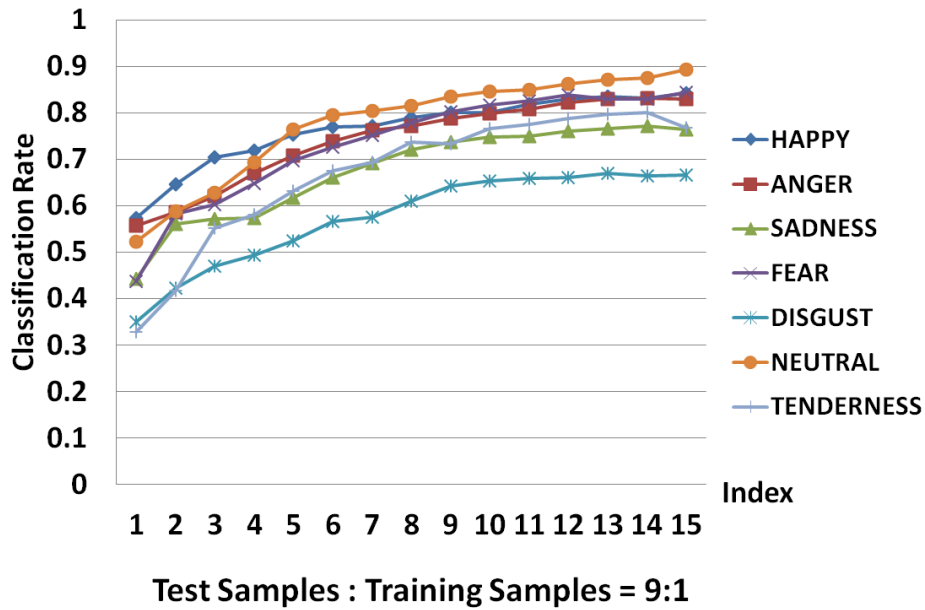


Figure 3.3: The classification rate curve of each emotion when increasing the training samples up to 10% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

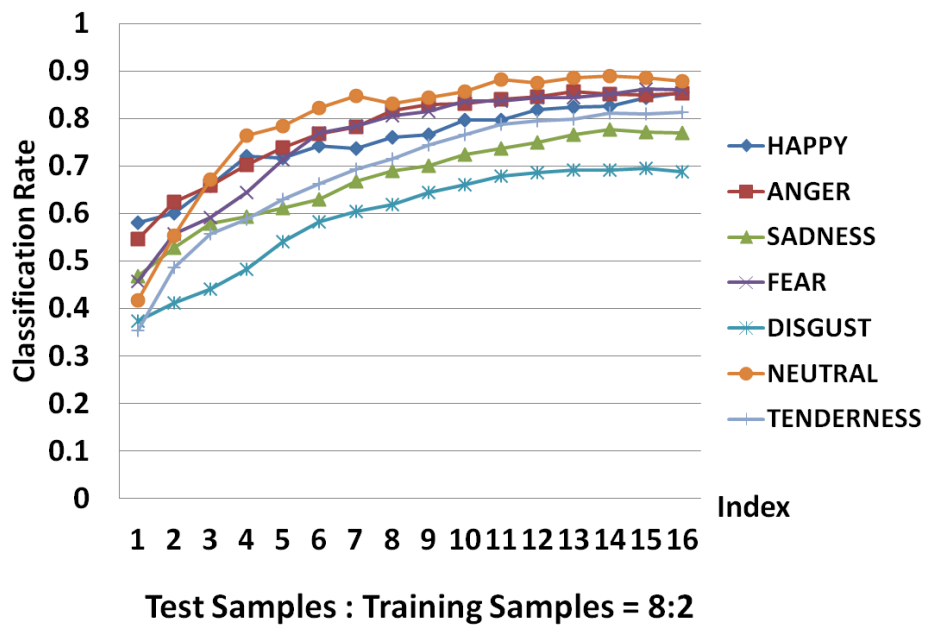


Figure 3.4: The classification rate curve of each emotion when increasing the training samples up to 20% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

From these figures, we can see that when the ratio is 9:1, the classification rate curves for all of the emotions always keeps increasing. However, when the ratio is changed to 6:4, the curves tend to stabilize, which means the current set of training samples we use

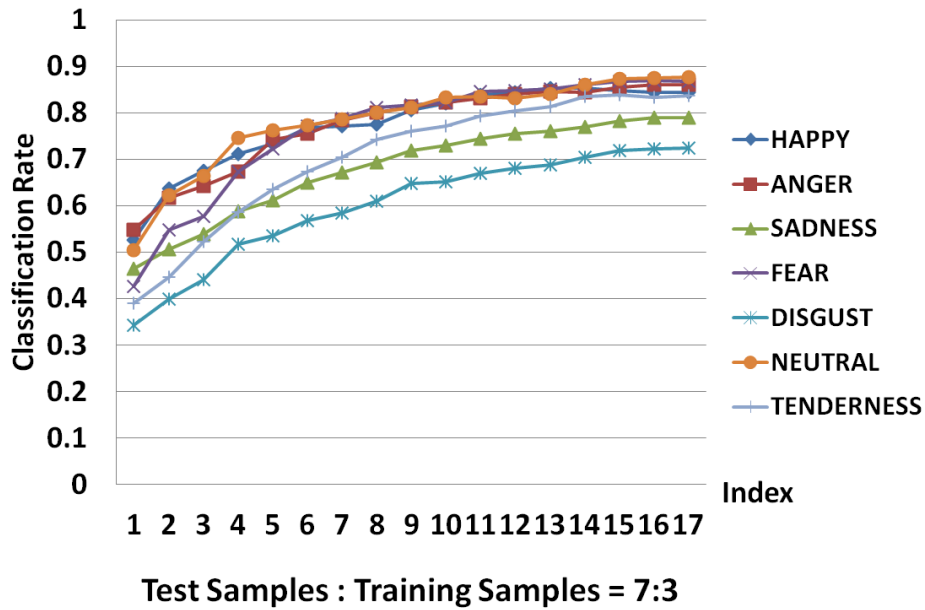


Figure 3.5: The classification rate curve of each emotion when increasing the training samples up to 30% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

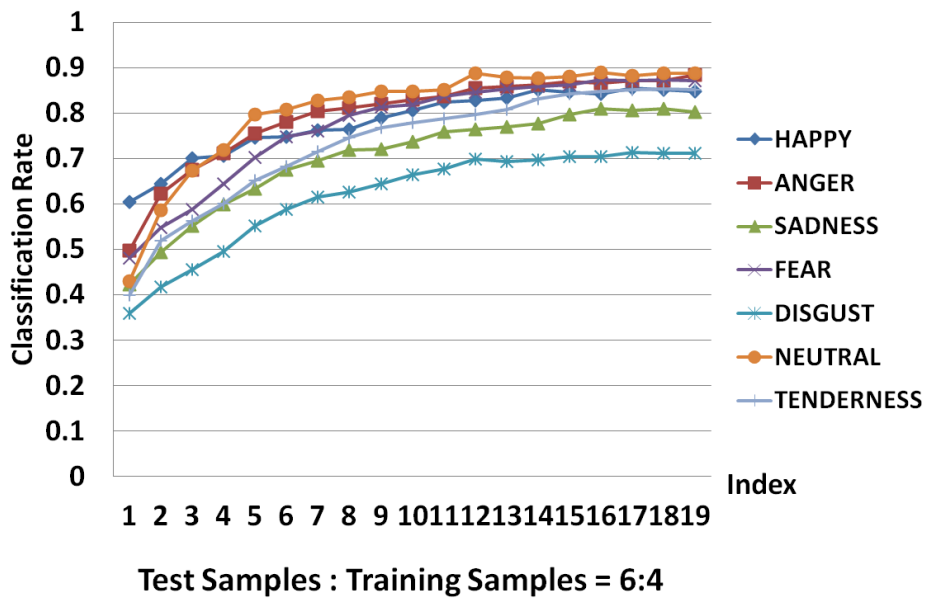


Figure 3.6: The classification rate curve of each emotion when increasing the training samples up to 40% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

is an over-complete data set. When the ratio reaches to 2:8, the curves saturate at the “index” value of 13 (the variable “Index” in Algorithm 1).

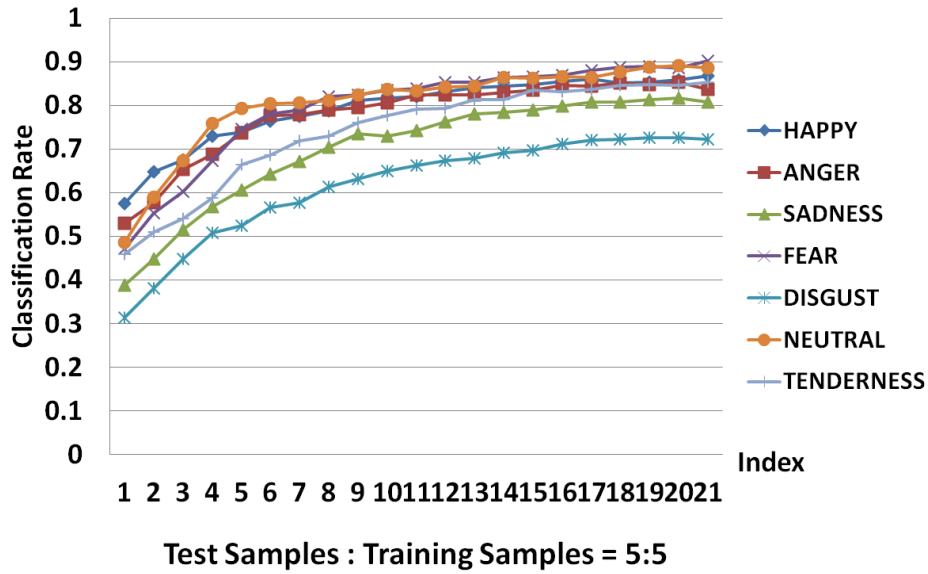


Figure 3.7: The classification rate curve of each emotion when increasing the training samples up to 50% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

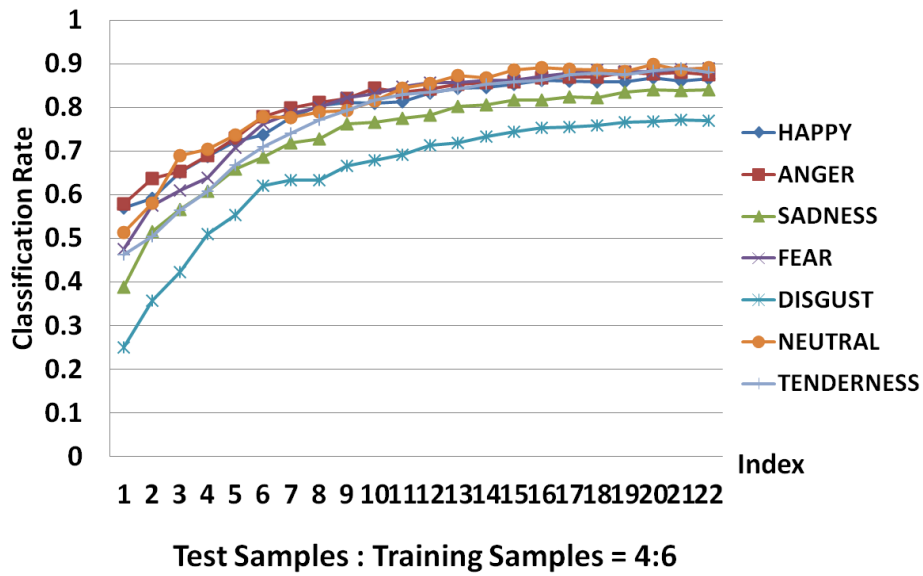


Figure 3.8: The classification rate curve of each emotion when increasing the training samples up to 60% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

### 3.4.2 Affective Classification Results

We use the same database for this experiment. The sample matrix constructed by Method (1) is optimal since the database we use is over-complete when the proportion of the test samples to the training samples is 6:4. Therefore, we randomly select forty

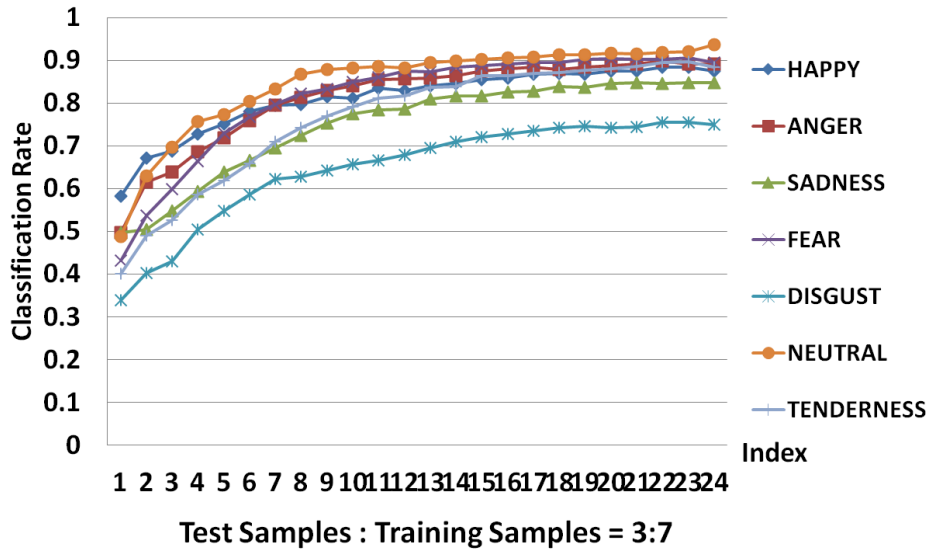


Figure 3.9: The classification rate curve of each emotion when increasing the training samples up to 70% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

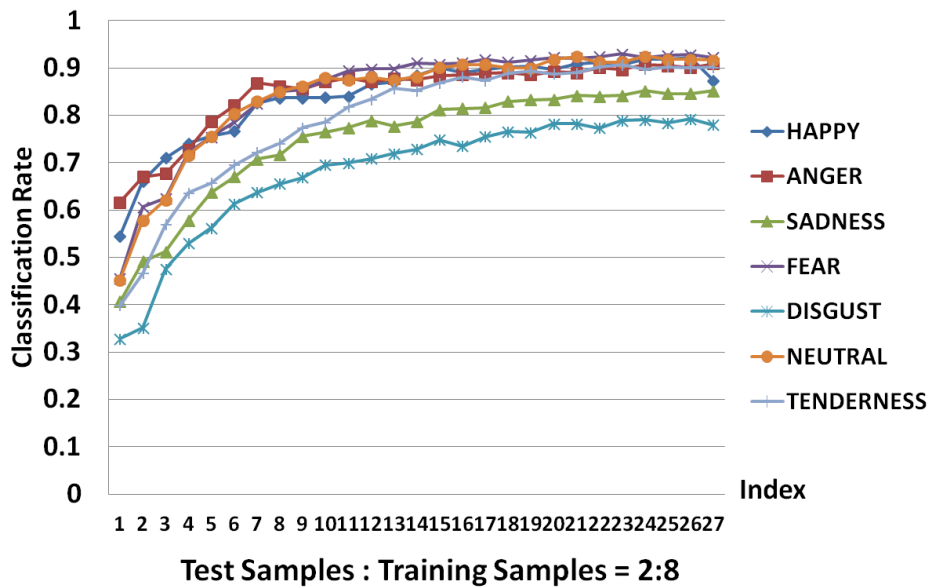


Figure 3.10: The classification rate curve of each emotion when increasing the training samples up to 80% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

percent of samples in the database to form the sample matrix  $A$ . We do the classification not only based on visual and audio features respectively, but we also do it based on different fusion levels: feature level fusion and decision level fusion, in order to test the performance of fusing visual and audio components. In this experiment, two different types of units of video - shot and scene - are considered as the classification units. In

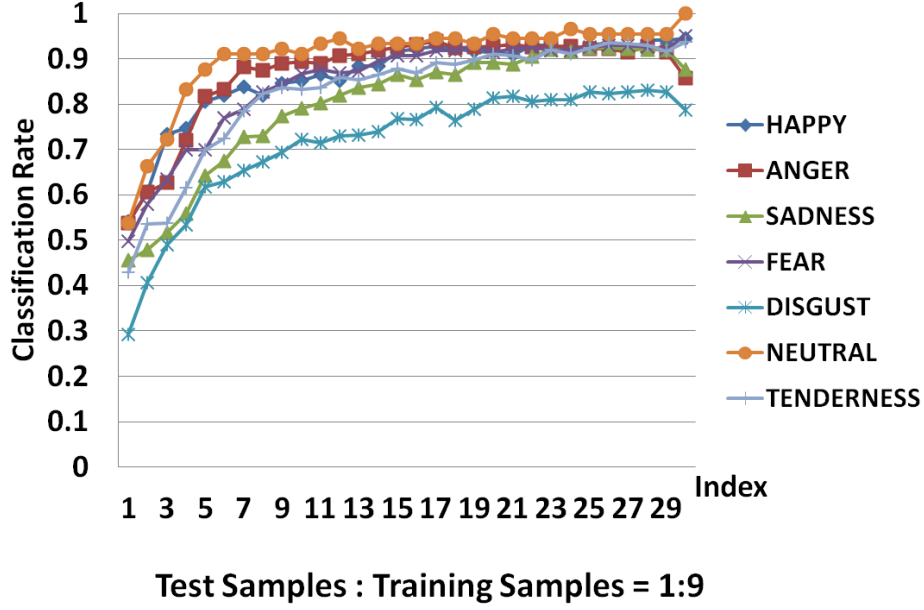


Figure 3.11: The classification rate curve of each emotion when increasing the training samples up to 90% of database. Note: because the increasing parameter  $P$  we use for different dividing ratios is the same, the “index”-axis in each graph is distinct.

parallel, we also implement the Nearest Neighbor (NN) method [CH67] and the Support Vector Machine on sparse coding (SVM-SC) [YYGH09] method to compare with our approach.

Let  $\varphi_\nu$  denote the residual vector corresponding to visual feature vector  $\nu$ , and  $\varphi_v$  denote the residual vector corresponding to audio feature vector  $v$ . Then, for feature level fusion, the feature vector we use is  $[\nu, v]^T$ . For decision level fusion, we compute the residual using Eq.(3.15)

$$\varphi = w_1 \times \varphi_\nu + w_2 \times \varphi_v \quad (3.15)$$

We adjust the proportion of  $w_1$  to  $w_2$  to try to find out the “optimal” ratio which gives us the best classification rate. The test proportions of  $w_1$  to  $w_2$  are: 1/9, 2/8; 3/7; 4/6; 5/5; 6/4; 7/3; 8/2; 9/1. In addition, we know nothing about the sparsity factor  $s$ . The emotional category is also unknown. As a result, we can further establish that  $s$  satisfies inequality  $s \leq n_q \leq \min_j n_j$  as  $q \in [1, \dots, m]$ , and we test the sparsity factor  $s \in [1, \dots, \min_j n_j]$  for  $j \in [1, \dots, m]$  in this experiment.

Table 3.2 reveals that first, the classification rate based on only audio features is



relatively low, and is the lowest in all experiments, because when the shot is very short, the features extracted from audio component are not very informative compared to the visual features. Second, the performance based on visual features is much better than that based on only audio features. Third, both the classification rates based on the feature-level fusion and decision-level fusion are higher compared to those based on visual or audio features only. It means the fusion of visual and audio components can improve the performance. Also, as the  $w_1/w_2$  ratio increases, the performance of decision-level fusion improves until the ratio becomes 8/2. We find that the relative weights of visual to audio in the 8:2 proportion is “optimal” for decision-level fusion. Fourth, in terms of the “optimal” ratio, the classification rates on “sadness”, “fear” and “disgust” of decision-level fusion are slightly higher than that of feature-level fusion. Yet, the classification rates for “happy”, “anger” and “neutral” are marginally smaller. Finally, the classification rates of “disgust” obtained for all the feature combinations are relatively low. This may be because this emotion requires a much more higher level of semantics processing than the others. In addition, our method outperforms the Nearest Neighbor (NN) method and Kang’s approach [Kan03]. However, the classification results of our method are better only on some emotions like “sadness” when compared to the SVM-SC method. We chose Kang’s work for comparison because it is the only other existing work that uses the “categorical emotional state” model. The rest of the works use the valence-arousal model. This helps in doing a direct comparison. Moreover, their approach also considers shot-level emotions.

Method	EM								Avg.time(s)
	Source	HA	AN	SA	FE	DI	NE	TE	
NN	Audio	0.12	0.23	0.21	0.25	0.14	0.04	0.22	0.09
	Visual	0.88	0.81	0.78	0.73	0.67	0.73	0.71	
	Fusion	0.9	0.81	0.75	0.75	0.65	0.74	0.72	
Our Method	Audio	0.24	0.30	0.64	0.29	0.30	0.47	0.54	0.25
	Visual	0.88	0.84	0.78	0.86	0.73	0.85	0.84	
	Fusion	0.89	0.88	0.78	0.87	0.74	0.84	0.83	

Table 3.3: Classification results based on different scene-level features and fusion level.

Table 3.3 shows the comparison between the Nearest Neighbor (NN) method and our approach based on the scene level. This table also shows the average computation time of each algorithm. As we can see that the classification rates of our method are higher

than those of the NN algorithm. In addition, the computation time of our method is 0.25s which means the speed of our method is reasonably fast (though NN is faster).

Furthermore, as we discussed in the previous section, the classification of affect within video depends on the construction of sample matrix ( $A$ ). Therefore, our obtained sample matrix, which is composed of the 10 most frequently mentioned film scenes would work well on the video clips in which the content also is one of the 10 scenes.

### 3.4.3 Intensity Curve

This experiment aims to evaluate the intensity of emotions computed by our approach using Eq.(3.10). Measuring emotional intensity is much harder than the recognition of the emotion for the “categorical emotional states” psychological model. Since multiple emotions co-exist simultaneously in videos, it is difficult to tease out each emotion’s intensity value. We assume that if the intensity curve of the main emotion within a video can be computed accurately, then it should work for the other emotions as well. Therefore, we focus on computing the intensity curve of the main emotion. We have chosen test video clips such that their main emotions are very obvious and clear. In addition, we would like the obtained intensity curves to be accurate. We make use of the sample matrix comprising of all the training samples, so that the obtained intensity curves are close to the ground truth. Each emotional category presents an example of intensity time curve as shown in Fig.[3.12 - 3.18] when it is the main emotion. In order to evaluate the correlation between the obtained intensity curve of the main emotion and the actual content of video clip, each figure is divided into several segments corresponding to the partitions of its video content which are outlined in Table 3.4.

Segment	Content Description
Figure 3.12 - Excerpt (4'35'') from the movie “E.T.”	
1	Elliotts brother talks to a man with a white coat
2	Scientists operates the machines to do analysis, and shows results
3	A man with a white coat comes to join this lab
4	ET weakly lies, and going to die
5	The man talks to Elliott, and Elliott tells people ET needs go home
<i>continued on next page</i>	

*continued from previous page*

Segment	Content Description
6	Machine shows results of the heart rates of Elliott and ET
7	Elliott sings the song to E.T. because ET turns more and more weak
8	Scientists find Elliott and ET are separated from the heart rate
9	Elliott holds out ETs hand, and ET dies

Figure 3.13 - Excerpt (2'26") from the movie "There is Something about Mary (2)"

1	Ted makes sperm at the rest room, pressing background music and man's fast breath
2	After a short break, Ted looks for the sperm, quite quiet but several knocks on door
3	Ted goes to open the door for Mary
4	Ted and Mary are talking
5	Mary takes sperm from Ted's ear mistaking it for hair gel
6	Mary and Ted talk at the cash register. Mary has her hair shaped into a peak

Figure 3.14 - Excerpt (1'55") from the movie "Schindlers list (2)"

1	A soldier picks the prisoners chosen to work in the factory
2	The commander stands on the balcony and leisurely looks at the square
3	Soldiers drive the prisoners to work
4	The commander picks rifle up and look for the prisoners who don't work
5	The commander freely shoots two prisoners one after another
6	The commander stops shooting and uses his rifle to stretch his arms behind him

Figure 3.15 - Excerpt (1'20") from a news "Weather Forecast"

1	Compere starts to forecast the weather, the map of earth turns up. Audio is the stable voice of compere
2	Cities' name is shown on the map. Compere focuses on those city. Audio is the stable voice of compere
3	The whole scene turns from bright to shadow, and then recover back. Audio is the stable voice
4	Compere continually focuses on those city. Audio is the stable voice
5	Compere starts to focus on the specific city. Audio is the stable voice

Figure 3.16 - Excerpt (1'45") from the movie "Life is beautiful (La vita bella) (2)"

1	The father, surrounded by other prisoners, brings his son hidden in a wheelbarrow. Audio is soft music
2	The father and the son find the broadcast studio, and try to talk to the mother using the louder speaker
3	The mother interrupts her work, stands up and smiles to listen to her husband and son

*continued on next page*

<i>continued from previous page</i>	
Segment	Content Description
Figure 3.17 - Excerpt (1'43") from the movie "Seven (2)"	
1	Several soldiers come into a dilapidated house and find something lying on the bed covered by the quilt
2	One soldier opens the quilt, and find a extremely skinny man having been savagely tortured
3	Soldiers can not bear this scene and the smell, and tight background music arise
4	The boss also come into this house, music is off
5	A "SLOTH" is found on the wall, and tight background music arise again
6	The boss finds some picture in a box besides the body, and audio is tight music
7	Unexpectedly, the dead man wakes up
Figure 3.18 - Excerpt (1'37") from the movie "Trainspotting (1)"	
1	A women screams at night, along with the actor's soliloquies
2	"Sick Boy" tries to calm her down, in the meantime, the others wake up
3	They eventually find out that the womans newborn baby is dead, audio is women's scream
4	Peoples keep silent, audio are women's scream and actor's soliloquies
5	Sick boy asks Mark to say something, Mark then loudly says he will make a "fix", audio is women's scream

Table 3.4: Labels describing the content of the test video clips.

Fig. 3.12 presents the obtained intensity curve of one "sad" video clip from the movie "E.T.". It is divided into a total of nine segments corresponding to the video content detailed in Table 3.4. The values at the beginning of the "sadness" intensity curve in Fig. 3.12 are lower than that of some other emotions. It is inevitable that not all the values of intensity curve representing the main emotions are the largest because the best classification rate of "sad" emotion is not quite 1. Since these curves of the second and third segments in this figure are not that accurate, we start from the fourth segment to analyze the correlation between the curve and the video content. The intensity values of the fourth, fifth and seventh segment are relatively higher in the entire curve. This reveals that relatively stronger sad scenes occur during these intervals, which are as expected and in line with the video content. There is a slight fall followed by slow rise around the 3500th frame in the middle of the fifth part. This is because of the change of shot. The obvious drop of the intensity in the sixth and eighth segment is also expected

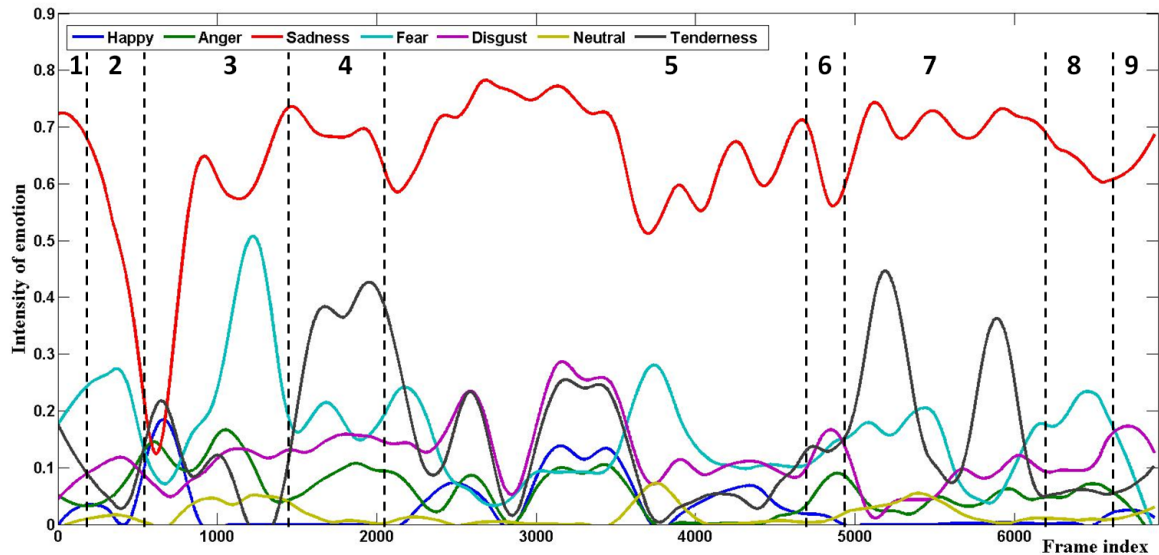


Figure 3.12: Intensity time curve obtained for an excerpt from the film “E.T.”. This video clip is divided into 9 parts based on their content as the broken lines show. The main emotion contained in this excerpt is *sadness*, so the intensity of “sadness” emotion is almost higher than intensities of all the other emotions, and in line with the video content briefly described in Table 3.4.

as the corresponding video content is rather uninteresting. At the tail of this curve, i.e. the ninth segment, the intensity value gradually increases, which means the atmosphere of sadness is coming back and is gradually enhanced in the video clip. Thus the obtained intensity curve tallies with the expectation based on viewing the content.

Fig. 3.13 shows the intensity time curve obtained for the selected test video clip from the movie “There is Something about Mary (2)” whose main emotion is “happy”. Overall, the intensity values of “happy” emotion are almost always higher than that of other emotions except the intensity values at around 1000th frame to 1250th, 2250 – 2400th and after 3200th which are lower than “sadness”. Besides, this test video clip is divided into six segments based on scene descriptions as shown in Table 3.4. The most interesting and funny scene of this video clip is at the fifth segment. Correspondingly, Fig. 3.13 shows that the intensity curve of “happy” emotion for the fifth segment is not only larger than that of other emotions, but also has the highest sharp peak which exactly corresponds to the most funny part within the video content.

An example for the “anger” emotion is presented in Fig. 3.14. This test video clip is an excerpt from the movie “Schindler’s list”. It is found that the main emotion of this “anger” clip can be properly recognized by our method as its intensity curve is above all

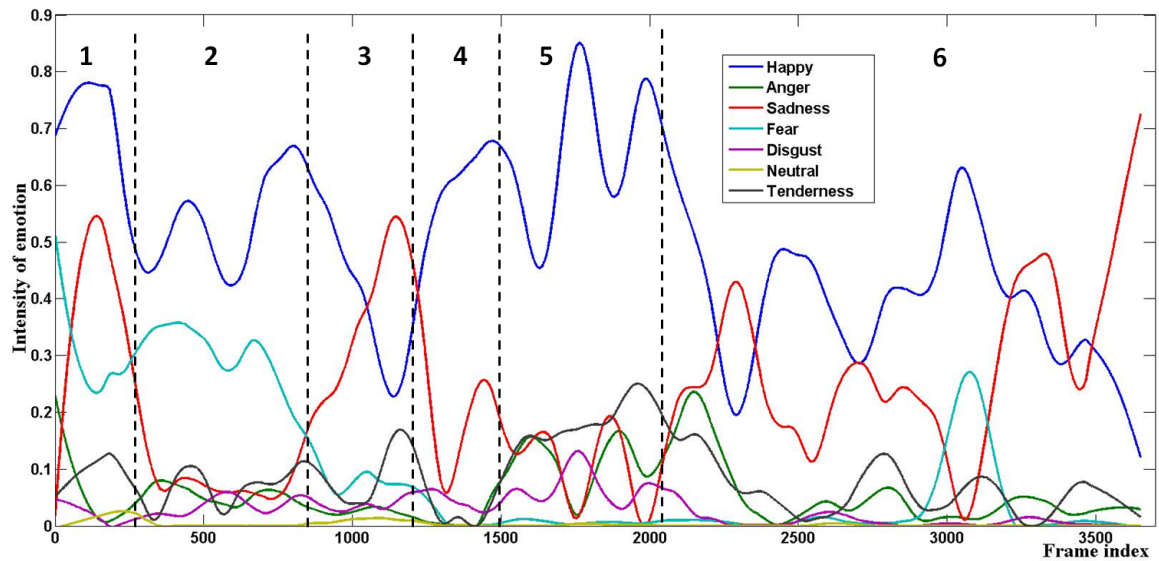


Figure 3.13: Intensity time curve obtained for an excerpt from the film “There is Something about Mary (2)”. This video clip is divided into 6 parts based on their content as the broken lines show. The main emotion contained in this excerpt is *happy*, so the intensity of “happy” emotion is almost higher than intensities of all the other emotions, and largely consistent with the video content briefly described in Table 3.4.

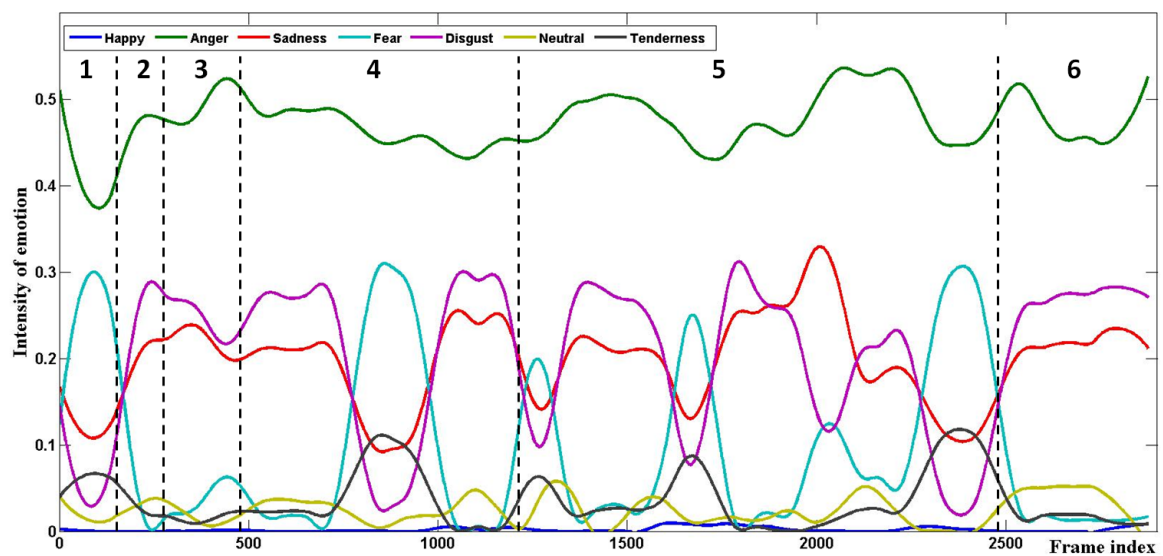


Figure 3.14: Intensity time curve obtained for an excerpt from the film “Schindlers list (2)”. This video clip is divided into 6 parts based on their content as the broken lines show. The main emotion contained in this excerpt is *anger*, so the intensity curve of “anger” emotion is completely separated from the intensity curves of all the other emotions, and in line with the video content briefly described in Table 3.4.

the other emotions’. Based on the scene segments and descriptions shown in Table 3.4, the segment of the strongest “anger” emotion is the fifth one. Correspondingly, the fifth segment of Fig. 3.14 displays two large peaks which separately indicate the action of

shooting from the commander, which dovetails with the scene description.

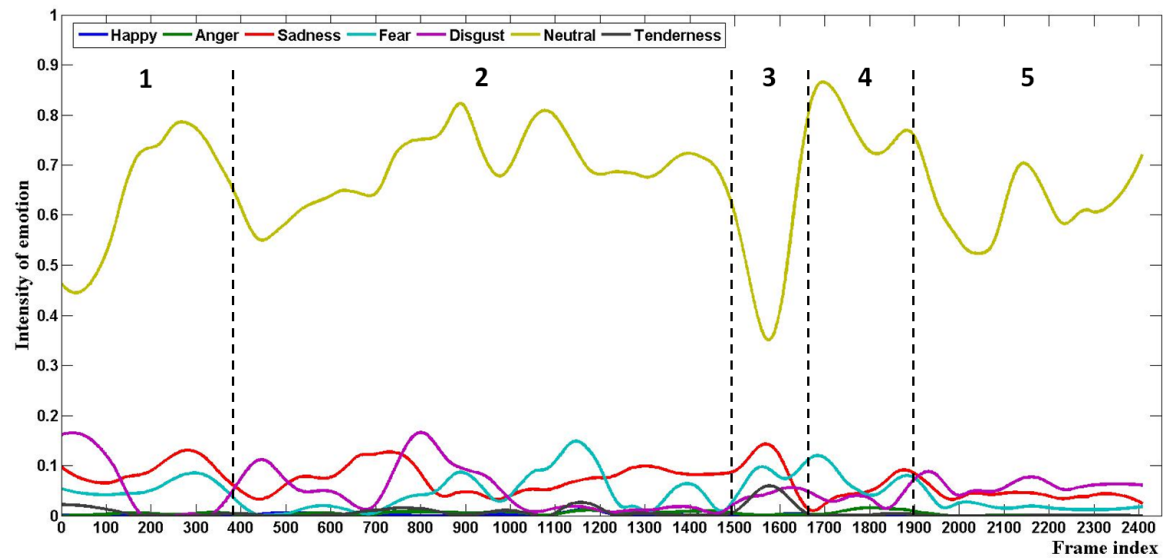


Figure 3.15: Intensity time curve obtained for an excerpt from a news “Weather Forecast”. This video clip is divided into 5 parts based on their content as the broken lines show. The main emotion of this video clip is *neutral* whose intensity values are much higher than all the other emotions’, and keep up with the changes of video content.

Fig. 3.15 exhibits a relatively stable curve without strong changes except at the boundary of each segment. This is expected as this neutral video clip is rather stationary and only contains one speech without too many changes. As for the drastic changes on the boundary of each segment, this can be because of the switch of shots using some shot editing technique. The other figures also mostly correspond to the expectation based on the video content description. Therefore, we managed to obtain intensity time curves which prove to be largely in line with what is expected. To the best of our knowledge, our work is the first work to evaluate the intensities of 7 discrete emotions. Almost all the existing affective curves are based on the “dimensional emotional space” psychological model, that is, the arousal curve and valence curves. They are completely different and they are not comparable.

Most people are used to express their emotional states with words (happy, sad, angry), and describe the degree of their emotion using the sentences like “Oh, I am so happy today!” or “I am very angry now!” Therefore, compared to the valence arousal curves, our intensity curves are easier to relate to and therefore can be more easily used in applications.

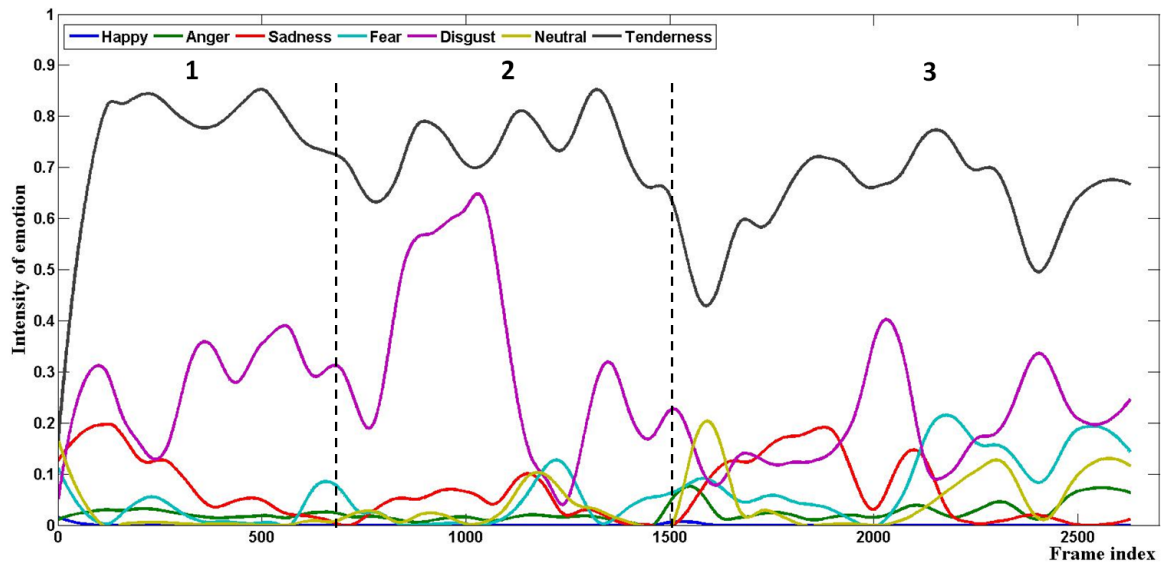


Figure 3.16: Intensity time curve obtained for an excerpt from the film “Life is beautiful (La vita bella)(2)”. This video clip is divided into 3 parts based on their content as the broken lines show. The main emotion of this video clip is *tenderness*.

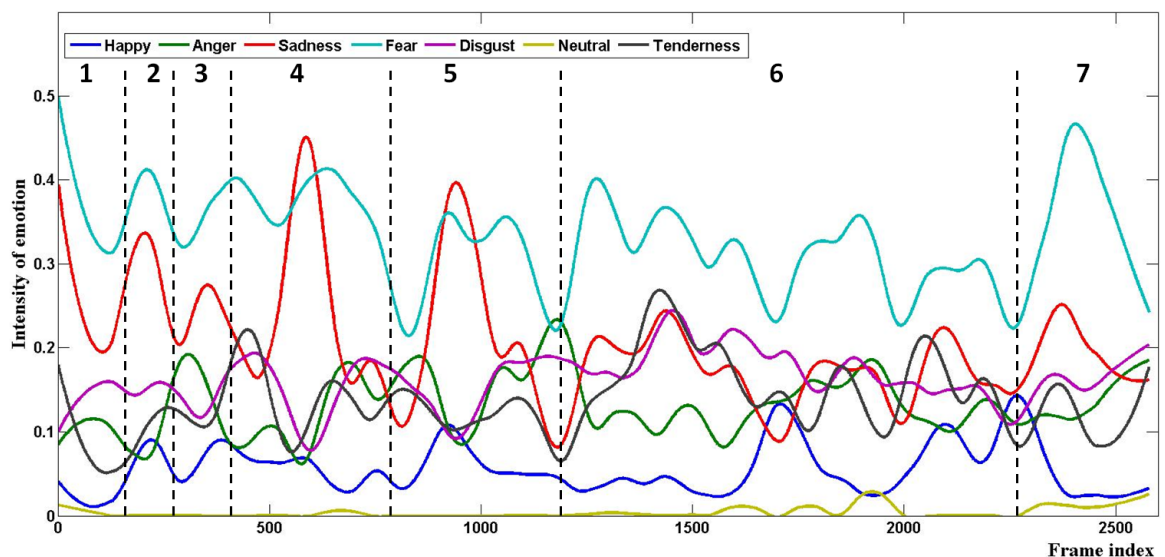


Figure 3.17: Intensity time curve obtained for an excerpt from the film “Seven (2)”. This video clip is divided into 7 parts based on their content as the broken lines show. The main emotion of this video clip is *fear*.

### 3.5 Summary

In this chapter, we first present a brief overview of affect or emotion in the “affective computing” area, and outlined two important psychological models representing the affect. Then, defining the links between the “categorical emotional states” and low-level features, we elaborated the proposed computational framework with sparse representa-



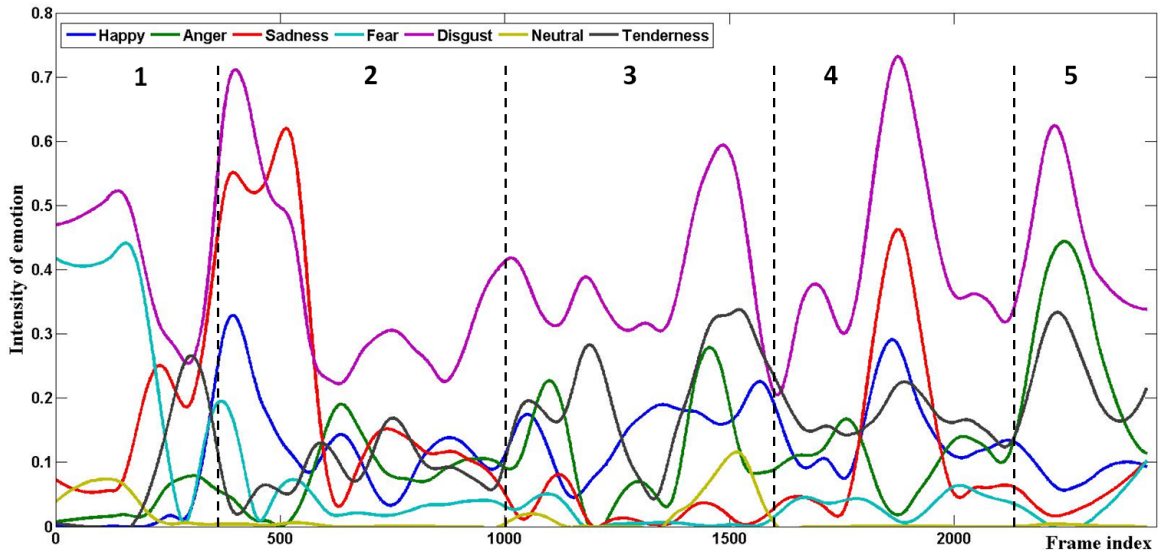


Figure 3.18: Intensity time curve obtained for an excerpt from the film “Trainspotting (1)”. This video clip is divided into 5 parts based on their content as the broken lines show. The main emotion of this video clip is *disgust*.

tion to represent and model the affective video content in Section 3.3, as well as the computation of intensities of discrete emotions. Subsequently, we discussed the conditions the “optimal” sample matrix should satisfy leading to the best classification result, while proposing two methods to construct the “optimal” sample matrix to gain the best classification rate. We tested our model on a number of video shots. The results demonstrated that our approach can well represent and model the affective video content based on the “categorical emotional states” model and the methods to construct the sample matrix performed efficiently. Also, we managed to obtain intensity time curves that represent the degree of contributing to the overall affect within a video for each emotion along the video. The chief curves were largely in line with the video content. Therefore, the proposed model does not only perform the classification of affective video content, but also provides a reliable approach for obtaining the intensity of discrete emotion. We can claim that the proposed approach is more in line with the way that people articulate their emotion experience. In the next chapter, we will present an application which utilizing the proposed model.

**Require:**

- Dividing ratio:  $R$ ;
- All the samples (feature vectors) in database:  $D$ ;
- The increasing parameter:  $P$ ;
- The number of emotional categories:  $m$ ;

**Ensure:**

- True/False
- 1: Extract the features from  $D$ , and normalized;
- 2: **for**  $\forall i, i \in [1, \dots, m]$  **do**
- 3:    $D_i = \{d | d \in D \text{ and } d \in i\text{th emotional categories}\}$ ;
- 4:   Randomly construct test sample set  $S_i$  such that  $S_i = \{s | s \in D_i\}$  and  $|S_i| = R * |D_i|$ ;
- 5: **end for**
- 6: The set of test samples:  $S = S_1 \cup \dots \cup S_m$ ;
- 7: The set of training samples:  $T = \{t | t \in (D_i - S)\}$ ;
- 8:  $A = \emptyset$ ;
- 9:  $Index = 1$ ;
- 10: **while**  $T \neq \emptyset$  **do**
- 11:    $T_{sub} = \{t | t \in T\}$  and  $|T_{sub}| = P$ ;
- 12:    $T = T - T_{sub}$ ;
- 13:    $A = A \cup T_{sub}$ ;
- 14:    $M(Index, \cdot) =$  the classification rates of emotions: Classify  $S$  based on  $A$
- 15:    $Index = Index + 1$ ;
- 16: **end while**
- 17: **if** the values in each column of  $M$  gradually increase **then**
- 18:   **return** True;
- 19: **else**
- 20:   **return** False;
- 21: **end if**

**Algorithm 1:** The algorithm for verifying if the database is over-complete. Returned “True” means the database is over-complete, and the input dividing ratio  $R$  is the best to choose the training samples; otherwise, it is not over-complete.

## Chapter 4

# Affect-based Adaptive Presentation of Home videos

In the previous chapter, we proposed a computational framework to bridge the representation and modeling from the affective video content to the *categorical emotional states* psychological model. We also developed a computational measure for the intensity of categorical emotional state. Specifically, a sparse vector representation  $x$  was obtained in this computational framework composed of a linear equation  $y = Ax$ . ( $y$ : feature vector extracted from the video;  $A$ : sample matrix consisting of the representative feature vectors of each emotional class.) This sparse vector  $x$  can be used to decide the emotional state of video clip by Eq.(3.9), as well as the *intensity* of emotion by Eq.(3.10). Since we can recognize the affect within videos with the sparse representation, the video can be assigned a label representing the emotion. We therefore can take advantage of these affective labels to adaptively create distinct representations of home-video for different people classes. In this chapter, utilizing affective labels, we also consider the face labels which are categorized into two classes. Consequently, based on different rules, three kinds of representations are constructed for families, acquaintances and outsiders respectively.

## 4.1 Introduction

The advancement of technology has brought forth into daily life, a plethora of consumer electronic devices such as digital cameras, digital video recorders and computers. Holidays, weddings, children's birthdays - those moments in life that we would like to hold on forever can now be preserved in the form of videos and images with the aid of such devices. Sharing these rich media with family, friends and even outsiders via Internet has become a significant means of communication between people [Bab07] due to the popularity of social community websites like Facebook and YouTube.

However, several issues are encountered when people want to share their home videos. First, significant redundancy within these video and image collections makes it boring and time-consuming to browse through. Second, we observe distinct individual preferences for such video and image content. This in turn makes it necessary to select and display different content to suit the diversity of preferences. Taking a birthday party video as an example, family members would probably be interested in media containing the host, rather than other unfamiliar guests. However, outsiders perhaps only care whether this media clip can interest them or not, if it is shared via Internet. Furthermore, it is impossible to keep all the media in online albums because of the restrictions on upload size and the limited capacity. As a result, it is probably a tedious job to choose and keep the desired videos or images. In the light of such prevailing problems, there is an urgent need for effective media synthesis method for creation of videos and images that are brief, rich and appropriate for individual preferences.

YouTube is one of the most well-known participatory platform providing a new generation of video sharing service in the contemporary online environment. Videos shared on it usually are relatively short [BG09], which means short video clips are generally more popular than longer videos. Therefore, people might need to cut long videos into shorter and easily digestible versions. Nowadays, there have been a variety of video editing software available, like Boilsoft video cutter and VisiCrop. However, as far as we know, all of them require the user to manually choose the needed part by providing beginning and ending times for video segments. This can be tedious and very time-consuming. Therefore, automatically analyzing the video and selecting the part needed

for the user is preferred. We find that the automatic selection of desired segments of the videos is a preferable strategy.

Among the various multimedia analysis techniques, visual attention analysis [Bun90] and semantic video analysis [RHC99] are two well-known approaches. “Everyone knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. ....” [Jam90]. Visual attention analysis is a multidisciplinary endeavor involving multiple fields such as cognitive psychology, computer vision and multimedia, etc. Semantic video analysis is to provide the semantic abstraction built on the original video data that is closer to and attempts to reach to the high-level understanding of human perceptual system. Both approaches aim to help understand the video content better. A relatively new and important multimedia analysis technique, termed as affective video analysis [HX05], has been introduced to improve the understanding between media and human beings’ feeling. Differing from the traditional techniques, affective video content analysis takes advantage of both psychological knowledge and computational methods to enable computers to recognize the emotions and affective states present in the video. Consequently, applications based on this method are more consistent with human abstract thinking and cognitive processes, which results in being more friendly, usable and natural. Combination of the affective video analysis and media synthesis not only helps towards the automatic affect-based video analysis, classification and labeling, but can also help identify all the attractive, strong-emotional and interesting moments. Affective video analysis has significant potential in choosing and generating satisfactory, interesting and natural media segments. Exploiting the affective cues contained in home videos and presenting them adaptively are the distinguishing features of this research.

In this chapter, we propose a method to adaptively present home videos using affective video analysis. We achieve this with the following steps: First, we extract the affective states within home video content by affective video analysis. In order to get the corresponding affective label for each video clip, the shot-based audio and visual features are extracted separately, and input to the affective analysis model. Thereafter, the affective analysis model generates a corresponding affective label for the analyzed

video shot. Simultaneously, the face recognition model also identifies the persons contained in each shot according to the face database provided in advance, and labels the video as well. Consequently, two kinds of labels (affective label and face label) are presented along with each video for the post processing. Second, three properties of video: emotional tone, local main character and global main character are found in each video and collection of videos by means of a modified tf.idf weight. Moreover, we assign a “diversity” value for each pair of shots. Finally, considering the diversity factor, we use these properties to connect different video clips and create presentations. Three example categories adopted from social groups identified in the Social Sciences considered in this chapter are: family, acquaintance and outsider respectively. Based on the experimental results, we show that our method can generate good presentations for each category. In this chapter, our main contribution is:

- Three concepts: emotional tone, local main character and global main character are proposed for video presentation. We demonstrate these to be important reference factors in generating adaptive presentations.

A preliminary version of this work was reported earlier [XK11]. Compared to that work, there are several improvements presented here. First, we use more audio features such as MFCCs in affective labeling. This improves the quality of labeling. Second, we compare the classification results of two fusion levels: feature-level and decision-level. This facilitates looking for a better way of fusion of low-level features to emotions. Third, we classify more emotions in our experiments. A finer granularity of emotions thus helps in better choice of segments. Fourth, we introduce a new factor - “diversity” to improve our algorithms for generating family and acquaintance presentation. Fifth, the types of home videos in our experiments have slight changes. We use a new test video which is a raw video that has not been edited.

This chapter is organized as follows. Section 4.2 reviews the related work to serve as a preamble, and Section 4.3 describe the proposed methodology. Section 4.4 presents the experimental results of the proposed methodology, and finally, conclusions are drawn in Section 4.5.

## 4.2 Related Work

Within this section, we will review and discuss the related work on the adaptive presentation of home videos and video content affective analysis.

### 4.2.1 Adaptive Presentation

CeWe Color [Col07] has provided CeWe Photobook software program to manually create photobooks as customer wishes with a variety of templates of backgrounds, layouts and designs. The concept of adaptive presentation has been innovatively exploited. Based on different objective, individual characteristics, and profiles, the presentation can be created adaptively to satisfy different demands. Since CeWe Photobook has got the support of the majority of customers and has become a popular gift among family and friends, we have reasons to believe that creating elegant, attractive and adaptive media presentations is promising.

Demerdash et al. [DBKL05] have proposed a framework for generating adaptive multimedia presentations by dynamic selection of files from a large data repository. They have made use of concepts from the fields of discourse analysis and rhetorical structure. They also exploited the technical(syntactic), semantic and relational textual annotation as well as context-sensitive rules and patterns of selection to create the presentation. Rabbath [RSB10] et al. have established that the representation of photos in appealing physical photo books has been highly appreciated by many people, and sharing the photos via social networking sites is becoming a popular communication means in the world . Accordingly, it is reasonable to develop an approach to collect and bundle photos from the same events but uploaded by different persons into one story. They have taken advantage of content analysis of text and images to automatically and semi-automatically detect media elements and select photos of a specific story. Our method is the generalization of the Rabbath's work, using videos instead of photos.

Adaptive media presentation is relatively new, thus, as far as we know not much literature is available in this field.

### 4.2.2 The Emotion Model

As discussed in subsection 2.1, there are two most important psychological models of emotion: a) continuous emotion space [Rus80, Bra94] in which the emotion is represented by a set of point; b) discrete emotion states [Ekm92, Hev36] which uses the discrete word to describe the emotional states.

In this chapter, we choose the second emotion model to represent the affect within videos for the following consideration: first, because we collect affective ground truth from user study, a discrete emotion model can enable participants to express their emotional experience better. Second, it is easier to label video shots with a single emotion compared to continuous dimensional emotion model. We identify the affect of each shot of video, and label them with happy, sad, fear, angry or neutral tags. If several adjacent shots are identified to have the same affect, then, we label them all with a single affective tag.

### 4.2.3 Affective Video Analysis

As discussed in subsection 2.4, [Kan03, HX05, SYHH09] have put efforts to understand the relationship between the low level features and affective events. Additionally, most studies have focused on the classification of affective events within videos/images. Currently, all the proposed approaches can be summarized into four categories: Gaussian Mixture Model [SYHH09], Hidden Markov Models [Kan03, SY07, TYA11], Bayesian Model [SKCP09] and Support Vector Regression Model (SVR) [ZTH<sup>+</sup>10, CLT<sup>+</sup>13, AHA14]. They all totally differ from our proposed approach: sparsity-based affective analysis, which is one of our main contributions.

## 4.3 Methodology

In this section, we will discuss the details of the proposed method. Fig. 4.1 shows the overall framework of our method. Shot is the basic unit of affective analysis and face recognition. So, a Singular Value Decomposition (SVD)–based approach [CKP03] is first implemented to segment the video. After that, shot-based audio and visual feature vectors are extracted respectively as input to the affective analysis model. Simultaneous-



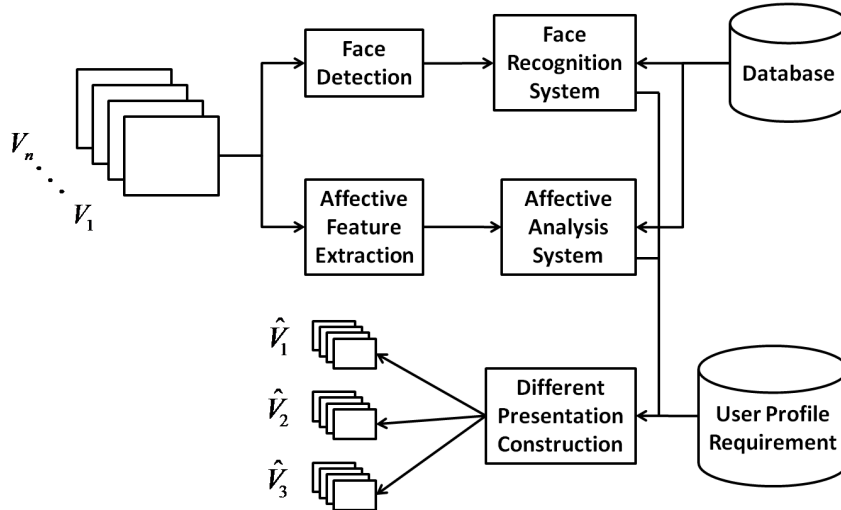


Figure 4.1: The overall framework of our proposed method

ly, faces are also detected as input to the face recognition model [TP91]. Then, affective and face labels are assigned to each shot, and treated as the terms in “document”: shot and video. A modified tf.idf weighed scheme is used to find the three *properties* of video: emotional tone, local main characters and global main characters. Based on them, adaptive presentations can be created for different target viewers. As our main contributions, the technique of affective analysis has been elaborated in Chapter 3, and the algorithms for the construction of presentation will be detailed in the following.

### 4.3.1 Affective Features Extraction

The relationship between emotional events and low level features has been studied by many researchers [BTA90, Kan03]. For example, in the “sad” event, the colors are usually “dark”, and “low saturated”, but “happy” event is along with “bright” colors in general. So the color cue is taken into account to represent the emotional events. To compute the color feature, each frame of RGB color space is transformed into HSV color space, and then the pixel values are quantized into 11 culture colors [Gol99]. For key frames of each shot, we computed the histograms of 11 culture colors, as well as the mean and covariance of color histogram, saturation and value separately. A total of 17 features are extracted for each key frame, and composed into a visual feature vector. Totally,  $p$  number of visual feature vectors are extracted for shot with  $p$  key frames.

As we know, movies can express rich and strong emotions because of the meticulous

editing of film editors. In contrast, usual home videos without any editing not only has no strong emotions, but it also has much noise. However, the associated audio can contain significant affective information. Thus, the 19 audio features, such as pitch average, silence ratio, short time energy, zero-crossing rate, brightness, rolloff and 13 MFCCs, were computed to construct the audio feature vector. So far, two kinds of feature vectors ( $p$  visual feature vectors and 1 audio feature vector) were constructed and normalized.

### 4.3.2 Affective Labeling

The approaches reported in the past papers can be classified into several categories: Gaussian Mixture Model, Hidden Markov Models, Bayesian Model and Support Vector Regression Model. Here, a totally different model called *sparse representation of affect* which has been detailed in Chapter 3 to do the classification and labeling of affective video. Therefore, we briefly review this approach in this section.

Since we do not well understand that how does the combination of low-level features contribute to affect, we extract the right sparse representation from a high-dimensional feature space. Sparse representation essentially involves solving of a linear system of equations. Considering a linear equation  $y = Ax$ , where  $A$  is a real  $k$  by  $n$  matrix and  $k < n$ , the solution  $x$  is infinite. Sparse representation aims to solve it by choosing the optimization problem  $\ell^0$ -norm solution, or  $\ell^1$ -norm solution, or  $\ell^2$ -norm solution as Eq.(3.3) and Eq.(3.7) point out.

#### 4.3.2.1 Sparse Representation

As discussed in Section 3.3.2, the feature vectors extracted from videos with different affect are distinctive. Given two video clips with same affective state, the extracted feature vectors should be very similar. A set of representative feature vectors for each affective event can be found as the basic feature vectors of that affective event subspace, and then each feature vector coming from that affective event can be represented as a linear combination of the corresponding basic feature vectors. Therefore, we take advantage of Eq.(3.4) to represent the test feature vector  $y$ .

Unfortunately, it is difficult to learn if those feature vectors can represent that e-

motional subspace. However, we already provided two methods to construct the reliable sample matrix  $A$  in Section 3.3.3.2. Given feature vector  $y$  extracted from the video with  $q$ -th emotion can only be represented by those basic vectors from the same affect, as Eq.(3.6) shows in Section 3.3.2. Then, sparse representation vector  $x$  can be obtained by solving the linear equation  $y = Ax$ .

#### 4.3.2.2 Sparse Solution

In this chapter, we exploit the discriminative nature of sparse representation to perform classification of affect content within videos. In sparse representation area, the basic idea of compressive sampling is: for certain types of signals, just a small number of nonadaptive samples carries sufficient information to approximate the signal well [Can06], which is discussed in Section 3.3.1. We have a complete discussion and algorithm about converting the affective analysis problem into the reconstruction of a sparse signal in Section 3.3.2. The solution  $\tilde{x}$  therefore can be obtained by solving  $\ell^1$ -minimization problem as shown in Section 3.3.2.

#### 4.3.2.3 Fusion Of Two Components

Once we obtain the solution  $\tilde{x}$ , we test how well or how close the coefficients within  $\tilde{x}$  is associated with each emotion. The function denoted by  $\Phi_j(x)$  is introduced to compute a  $\varphi_j$  for  $j$ th emotional class ( $j = 1, \dots, m$ ) as Eq.(3.8) states. Then the emotion of  $y$  can be found by  $q = \arg \min_j \varphi_j$ .

Given a new test shot of home video,  $p$  visual feature vectors  $\hat{y}_i$  for  $i = 1, \dots, p$  are extracted from  $p$  key frames, and one audio feature vector  $y_a$  is extracted from the shot.  $A_v$  and  $A_a$  respectively denote the sample matrix of visual component and audio component. Since we do not precisely know how to fuse visual and audio components to obtain better classification results, two kinds of fusion levels: feature-level and decision-level have been used to do a comparison. For feature-level fusion of visual and audio components, we combine the visual feature vectors  $\hat{y}_i$  for  $i = 1, \dots, p$  with the audio feature vectors  $y_a$  to get the overall feature vector  $\bar{y}$  using the following equation. Then

the final classification decision of emotions is based on this overall feature vector  $\bar{y}$ .

$$\bar{y} = \left[ \frac{\sum_{i=1}^p \hat{y}_i}{p}, y_a \right]^T \quad (4.1)$$

Besides, we also consider the decision-level fusion of visual and audio components. In this case, a sparse solution  $\hat{x}_i$  will be obtained for  $\hat{y}_i$  for  $i = 1, \dots, p$  based on the visual sample matrix  $A_v$ , and  $\hat{x}_a$  is computed for  $y_a$  based on  $A_a$ . Consequently,  $\hat{\varphi}_i$  would also be computed corresponding to  $\hat{x}_i$  for  $i = 1, \dots, p$ , and  $\hat{\varphi}_a$  would also be computed corresponding to  $\hat{x}_a$ . Finally, an overall residual vector  $\hat{\varphi}$  is computed by fusing these residual vectors  $\hat{\varphi}_i$  for  $i = 1, \dots, p$  and  $\hat{\varphi}_a$ . The final classification decision is made based on  $\hat{\varphi}$  according to Eq.(3.8) and Eq.(3.9).

$$\hat{\varphi} = w_1 \hat{\varphi}_a + w_2 \frac{\sum_{i=1}^p \hat{\varphi}_i}{p} \quad (4.2)$$

### 4.3.3 Presentation Construction

Before we create presentations of a collection of videos, we also have to recognize the faces in the videos to create the person labels of shot. In order to create the basic face database, the users in advance need to upload as many photos as possible, of individuals who they usually care, to train our face recognition model. Though it is tedious, fortunately, users just need to do it once. Once the basic face database is established, the PCA-based approach [TP91] can be applied to identify the human faces within shot and label each shot with the recognized face.

Through the affective model and face recognition model of our method, each shot has been assigned with several person labels and one emotional label. In order to obtain satisfying home video presentations by exploiting the labels of affect state and person, various rules need to be developed on the basis of the human perception of different purposes respectively. Next, we will introduce three significant concepts for constructing our presentation. These three proposed concepts are one of our contributions in this work.

#### 4.3.3.1 Properties Of Videos

It is quite frequently possible that the original video is noisy, of low quality or has long boring segments. However we find some “inherent” properties contained in it. We list them as follows:

- Emotional Tone
- Global Main Character
- Local Main Character

Specifically, in this chapter, the main feeling of audience is considered as the emotional tone of that video, when he/she watches the video. The main characters of a single video are referred to as local main characters, to whom the story is closely related. Main characters found in the entire collection of videos of the same theme are referred to as global. Later, we will explain why the three properties are chosen, describe how to define these three concepts in videos, and how to utilize them to select shots and construct the final presentation.

According to Murch [OM02], emotion is the most important factor when it comes to film editing: whether the cut reflects what the editor believes the audience should be feeling at that moment or not [BTA90]. Therefore, emotional continuity is proposed as the most important goal of scene editing in film editing area to maintain the emotional flow of a movie or show. It is better to keep the overall tone continuity of the film, because that is what the audiences mostly experience. On the other hand, the audience attention span for our own home videos is often quite short, and the shots with strong emotion catch audience’s attention mostly, since the interesting parts of a home video are intermixed with longer, less interesting regions. What’s more, such video is often of poor quality resulting from abrupt camera movement or too short or too long views, making it uninteresting to watch while waiting for the next interesting segment [GBC<sup>+</sup>00]. Therefore, the final presentation must preserve the emotion of videos as much as possible and greatly attract audience’s attention and interest.

People overwhelmingly attend to humans in images and videos with the faces dominating visual attention [CXF<sup>+</sup>03]. In other words, the faces within images/frames catch

most attention of audiences. Based on this fact, the person labeling can be taken into account as an important cue for the construction of presentation. On the other hand, since we have made an assumption that the collection of videos is related to the same theme, the main people in these videos are relatively constant. Thus, the idea of local main character and global main character make sense for videos.

The term frequency-inverse document frequency (tf.idf) weight [SM83] is a statistical measure used to evaluate how important a term is to a document in a collection or corpus. The weight of *tf.idf* is calculated by Eq.(4.3):

$$(tf.idf)_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}} \times \log \frac{|D_c|}{|d : t_j \in d|} \quad (4.3)$$

where  $n_{i,j}$  is the number of occurrences of the considered term  $t_j$  in document  $d_i$ ,  $|D_c|$  is the total number of documents in the corpus, and  $|d : t_j \in d|$  is the number of documents where the terms  $t_j$  appears.

It is a popular method often used in text mining and information retrieval. According to the definition of tf.idf, the importance of a term increases proportional to the number of times the term appears in that document but decreases with an increase in the frequency of the word in the collection. However, it conflicts with our definitions of emotional tone, and main characters. The emotion mostly felt by audience in video is the emotional tone of that video, which means that corresponding emotion term will appear with high frequency in all shots of that video. In this case, shot is regarded as “document”, one video is treated as the “corpus”. Similarly, local main characters refer to the persons who appear with relatively high frequency in video, and global main characters are the set of person labels which are assigned in whole collection of videos more often. Therefore, the tf.idf weight of emotion term and person label can not really reflect the importance of the term, because it violates the intuition that the final weight should increase proportionally to the frequency of the label in the corresponding “corpus”. So, a new tf.idf formulation which can really reflect the relationship between importance of label and its frequency in the corresponding “corpus” is introduced as Eq.(4.4).

$$w_j = \frac{n_{i,j}}{\sum_k n_{i,k}} \times (\log |D_c| - \log \frac{|D_c|}{|d : t_j \in d|}) \quad (4.4)$$

An emotional term like “happy” or “neutral” has been assigned to each shot. In this case, we treat the shots as the documents, and video as corpus in modified tf.idf scheme, that is,  $|D_c|$  is the total number of shots of video  $v_i$ , and  $|d : t_j \in d|$  is the number of shots labeled with  $t_j$  (refers to  $j$ th emotional class) in video  $v_i$ . According to the formulation (4.4), each emotional term in video has a tf.idf weight. We treat emotional term with highest weight as the emotional tone of this video. We denote the tf.idf weight of emotion label  $t_j$  by  $w_j^e$ , so the emotional tone  $ET$  of the video is:

$$ET = \arg \max_j (w_j^e) \quad (4.5)$$

Here, we point out that the emotional tone can tolerate or hide certain errors of affective labeling. Even if there are few errors in affective labeling, as long as the rate of errors is less than a threshold, the found emotional tone is also consistent with that of the video. We make use of the weights and emotional tone to select appropriate shots from each video for the construction of presentation.

Similar to the weight of emotional term, a weight  $w_{i,j}^l$  called local character weight is computed for each person label in video  $v_i$ . Differing from local character weight computation, global character weight  $w_j^g$  is based on the whole collection of videos. Specifically,  $|D_c|$  is sum of shots of whole collection of videos, and  $|d : t_j \in d|$  is sum of shots labeled with person label  $t_j$  in all videos.

In order to find local and global main characters of videos, we introduce two adaptive thresholds  $\varepsilon^l$  and  $\varepsilon^g$  as the judgement of local and global main characters respectively. When the local weight is larger than  $\varepsilon^l$ , the corresponding person label is one of main characters of that video. Similarly, comparing global character weight with global thresholds  $\varepsilon^g$ , the global main characters can be found. The formulations to find local main character and global main character are:

$$LMC_i = \{j | w_{i,j}^l \geq \varepsilon^l\} \quad (4.6)$$

$$GMC = \{j | w_j^g \geq \varepsilon^g\} \quad (4.7)$$

where  $LMC_i$  is the set of local main characters of video  $v_i$ , and  $GMC$  represents the set of global main character of collection of videos. The thresholds can vary for different videos, which is more reasonable. In our implementation, we choose the threshold as the half of maximum:

$$\varepsilon_i^l = 0.5 * \max_i(w_{i,j}^l) \quad (4.8)$$

$$\varepsilon^g = 0.5 * \max_{i,j}(w_{i,j}^g) \quad (4.9)$$

Here,  $\varepsilon_i^l$  is the threshold of  $v_i$ .

#### 4.3.3.2 Diversity Factor

Besides the above three concepts, we introduce another factor named *diversity*. The algorithms proposed in our previous work have one problem: all the shots could be selected from the same video in some cases, and the information of other videos is lost. Therefore, to avoid the occurrence of this situation and to consider as much information as possible, we assign a “diversity” value denoted by  $D_{ij}$  for each pair of shots  $s_i$  and  $s_j$ . According to the fact the content of one shot might be similar to the adjacent shots within the same video, and dissimilar to the far shots, *diversity* is inversely proportional to the similarity of the content of two shots. Therefore, we define a new function  $D$  as the following:

$$D_{ij} = \begin{cases} 0 & \forall i, j : s_i, s_j \in v_p \\ 1 & \forall i, j : s_i \in v_p, s_j \in v_q, \text{ and } v_p \neq v_q \end{cases} \quad (4.10)$$

Specifically, the diversity between two shots from the same video is 0; otherwise, it is 1. Finally, we can establish a diversity matrix  $D$  whose elements are 1 or 0 and  $i$  and  $j$  in (4.10) represent the row and columns respectively. We always try to construct a presentation which has larger diversity based on the above three properties.

#### 4.3.3.3 Algorithms For Presentation

We now have already found the emotional tone, and local and global main characters of total collection of videos. Each shot has three kinds of weight: emotional tone weight  $w_i^e$ , local main character weight  $w_j^l$  and global main character weight  $w_j^g$ . Eq.(4.11) is



used to compute three fusion weights for each shot with different parameters  $w_1$ ,  $w_2$  and  $w_3$ .

$$w_T = w_1 * w_j^e + w_2 * \sum w_j^l + w_3 * w_j^g \quad (4.11)$$

For different purposes of presentation, we focus on different weights to construct our characteristic presentations. Final presentation must satisfy that the order of the selected shots should be consistent with the original videos, and any other styles of film are not taken into account.

In the social sciences perspective [KK85], a society can be viewed as a large group consisting of many smaller social groups. The members of a true social group should share some characteristics which is called *social interaction* in social sciences, such as representations, interests, values, ethnic or social background, and even kinship ties. There are two main types of social groups classified [KK85]. One is primary groups which are small groups with intimate, kinship-based relationships. The other one is secondary groups which are large groups including formal and institutional relationships. In contrast to the primary groups, they may last for years. Besides, the individuals which do not belong to the previous mentioned groups can be treated as the third group. In terms of this knowledge and social interaction, we identify *family* as primary group, since it is more global and has a confident and firm tie, which make this group stable in any case. *Acquaintance* is considered secondary group in humans' communication network, as it can change with events, environments, social identity, and even social status of individual. However, from the view of humans' communication network, it is relatively stable in one particular case. In addition, the individuals outside the previous two groups can be classified into one group named *outsider* in this case. Therefore, we derive the three types of social groups for our presentation purposes.

In addition, we have conducted a survey to learn the relative importance of three properties for the different groups. A total of 12 people (6 females and 6 males) were invited to participate this survey. These subjects were asked ten questions, such as "Would you definitely show the videos containing you with your family?", "Would you definitely share the videos containing only your friends with your family?", and "Would you like to share the videos containing only your family with your friends?".

**Require:**

- A set of video's shots:  $V_s$ ;
- The fusion weight:  $w_T$ ;
- Diversity matrix:  $D$ ;
- The required upper bound of length of presentation:  $L_u$ ;
- The required lower bound of length of presentation:  $L_L$ ;

**Ensure:**

- The set of shots selected for the construction of presentation:  $S$ ;
- 1: Sort  $w_T$ , consequently, shots in set  $V_s$  come into being a set  $\hat{V}_s$  keeping the same order of sorted weights;
- 2: **while** *TRUE* **do**
- 3:   **for**  $\forall i, s_i \in \hat{V}_s$  **do**
- 4:      $\hat{S} = \{s_i \mid D(S \cup s_i) \text{ is largest and } s_i \not\subseteq S\}$ ;
- 5:   **end for**
- 6:    $i = \arg \max w_T(S \cup s_i)$ , such that,  $s_i \in \hat{S}$ ;
- 7:    $S = S \cup s_i$ ;
- 8:   **if**  $L_L \leq \text{length}(S) \leq L_u$  **then**
- 9:     **return**  $S$ ;
- 10:   **end if**
- 11: **end while**

**Algorithm 2:** The algorithm of constructing presentation videos for family and acquaintance.

The results of this survey indicate that the subjects pay more attention to the global main characters than local main characters when they share the video with their family. Yet, when sharing with friends, they would like to share the video containing the friends, which implies that the local main characters are more important. On the other hand, they indicate that as long as the video does not have any privacy issue, they would like to share it on the internet. Therefore, the presentations for different groups will focus on different aspects of the video.

- *Presentation for family*

We focus on the global main characters (*GMC*) and emotional tone (*ET*) to generate a particular presentation for family members. Generally speaking, the main persons appearing in the videos relevant to the same theme are relatively constant, and the role of these videos relatively appears at most times. The persons who go through the whole videos perhaps are the main character of these videos, in other words, the main character of this theme perhaps is the owner of these videos. As the family members of the owner, they care more about the activities

of their families rather than the others. According to the definition of global main character, global main characters play the lead in this video collection, that is, *GMC* represents the role of this theme. Therefore, the presentation focusing on the *GMC* is meaningful, on the other hand, we should keep the emotional tone of this collection in the terms of emotional continuity. The emotional tone is the primary emotion of that video the viewer should feel. Thus, the final presentation should preserve this property at least. As a result, the fusion weight is computed for each shot with  $w_2 = 0$  and  $w_1, w_3 \neq 0$ .

- *Presentation for acquaintance*

We focus on the local main characters (*LMC*) and emotional tone (ET) to generate a presentation specially for acquaintance. Compared to the global main character, the local main characters only stand for a short focus or interest, that is, *LMC* represents the role of that specific video, which perhaps is just a scene in this theme. Relatively speaking, local main characters could be those persons who are related to the role of theme. In other words, in the home videos, the local main characters perhaps are the friends or acquaintance of owner based on the fact that the local main characters are recognized depending on the user's basic face database. Therefore, compared to the family members, acquaintances concern with the local main characters who perhaps are themselves more than global main characters. Meanwhile, the final presentation also should preserve the emotional tone as well because the emotional tone is the primary emotion of that video. As a result, the fusion weight is computed for each shot with  $w_3 = 0$  and  $w_1, w_2 \neq 0$ .

- *Presentation for outsider*

Compared to family members and acquaintance, outsiders probably don't care what is the content of presentation but care more whether this presentation can interest them. As we know, faces and emotions within video both are the important aspects which can help video attract the viewers' attention. In our work, emotion of each shot has already been found by the affective labeling. Also, the face cue can be represented by the local and global main characters. A weighted scheme is used to combine the main character information and emotion information. Then the

importance of shot increases proportionally to the value of weight  $w_T$ . Selecting shots with high weight  $w_t$  can not only preserve the emotional tone, but also keep the main characters of videos. As a result,  $w_1, w_2$  and  $w_3$ , all  $\neq 0$ .

Specifically, given the collection of videos  $V_s$ , based on different fusion weight, we use the algorithm 2 to select the shots for family and acquaintance presentation respectively. For outsider presentation, we select the shots with higher fusion weight  $w_T$  to construct the needed presentation. The presentation for outsiders is created by algorithm 3.

**Require:**

- A set of video’s shots  $V_s$ ;
- The fusion weight:  $w_T$ ;
- The required upper bound of length of presentation:  $L_u$ ;
- The required lower bound of length of presentation:  $L_L$ ;

**Ensure:**

- The set of shots selected for the construction of presentation:  $S$ ;
- 1: Sort  $w_T$ , consequently, shots in set  $V_s$  come into being a set  $\hat{V}_s$  keeping the same order of sorted weights;
- 2: **for**  $i = 1 : \|\hat{V}_s\|$  **do**
- 3:    $S = S \cup s_i$  such that  $s_i$  is the  $i$ -th shot in  $\hat{V}_s$ ;
- 4:   **if**  $L_L \leq \text{length}(S) \leq L_u$  **then**
- 5:     **return**  $S$ ;
- 6:   **end if**
- 7: **end for**

**Algorithm 3:** The algorithm for construction of presentation for outsider

## 4.4 Experimental Results

### 4.4.1 Affective Classification Results

We collect affective ground truth from user study. Six people (3 males and 3 females) are invited to accomplish this task. They are required to describe their affective states with discrete words: “happy”, “neutral”, “angry”, “fear”, “surprise”, “disgust” and “sad” labels. A video of about one hour is provided to those participants. To the best of our knowledge, we just find one public video database established by Schaefer et al. [SNSP10] in which videos are labeled with discrete emotional words such as “happy” and “sad”. However, this database of films is not appropriate in our experiments since we focus on home videos. Therefore, the experimental videos are downloaded from

YouTube and Tudou. We assign an emotional label to each shot, only if six participants all choose that label. Because the selection of sample matrix is still a challenging problem in compressive sampling, all sample matrices randomly select their training samples from the ground truth as their columns. Therefore, we totally generated 100 samples matrices and 3600 experiments have been done based on different sparse level parameter in sparse presentation. Of course, these sample matrices have overlap but not be the same. For each emotion, we find the largest classification rate, and they are organized as a confusion matrix as shown in Table 4.1. As stated in previous section, we try to find a better fusion of visual and audio components for our sparse representation method. Thus, two fusion methods: feature fusion and decision fusion both are tested in our experiments as Table 4.1 shows.

	Feature-level Fusion				Decision-level Fusion			
	Fear	Happy	Sad	Neutral	Fear	Happy	Sad	Neutral
Fear	<b>0.95</b>	0.16	0.46	0.81	<b>0.87</b>	0.817	0.54	0.76
Happy	0.80	<b>0.46</b>	0.54	0.76	0.80	<b>0.97</b>	0.54	0.90
Sad	0.73	0.29	<b>0.62</b>	0.62	0.60	0.49	<b>0.77</b>	0.67
Neutral	0.93	0	0.31	<b>0.90</b>	0.80	0.95	0.46	<b>0.95</b>

Note: The left part of this table is the confusion matrix of classification based on feature-level fusion of two sources; The right part is the confusion matrix of Classification based on the decision-level fusion of two sources.

Table 4.1: Confusion matrices of classification based on feature-level fusion and decision-level fusion respectively.

From Table 4.1, we can see that, compared with feature-level fusion of visual and audio components, the classification rate of decision-level fusion has obvious improvements for “happy” (0.46  $\rightarrow$  0.97), “sad” (0.62  $\rightarrow$  0.77), and “neutral” (0.90  $\rightarrow$  0.95) on the largest classification rate, but the performance on recognizing “fear” emotion is slightly down (0.95  $\rightarrow$  0.87) . Additionally, because the test samples of other emotions we get after user study are very rare, the results of other emotions are not listed in Table 4.1. We will try to collect more videos to improve the experiment on other emotions in the future.

Emotion is a subjective feeling which relies on audiences perceptions. Talking about the emotion inevitably leads to a discussion about subjectivity. In addition, finding a good sample matrix is a very tough problem. We consider the similarity of pairs of feature vectors. In order to get the “optimal” sample matrix, we choose the feature

vectors with lowest similarity between them for each emotion. Of course, it suffers from the limitations of the data set. Therefore, we will seek other better methods in our future work or increase the number of test videos. Due to its inherent nature, home videos with other emotions such as “fear” and “angry” are much fewer. Therefore, it is difficult to compare our experimental results with the others’ results in home video area.

#### 4.4.2 Experimental Results For Presentation

It is very difficult to evaluate the quality of a video presentation as results are hard to quantify computationally and the related factors to consider are highly complex. Since there is no absolute measurement of presentation quality available, we evaluate the quality of our presentation with a user study. For the experiment, we invited 15 people as our test subjects. A total of 9 video presentations of three different theme videos were generated by our method, and used as the test videos. The information about three original videos and the generated presentations is summarized in Table 4.2.

Video	Theme	$L_o$	$L_u$	$L_L$	$L_f$	$L_a$	$L_s$
A	Graduation Ceremony	3m 47s	60s	40s	40.47s	51.39s	40.53s
B	Birthday Party	3m 34s	60s	40s	49.17s	40.03s	42.63s
C	Wedding Party	6m 10s	60s	40s	40.95s	40.35s	40.52s

Note: Video A, B and C are the collections of several short videos;  $L_o$  refers to the total length of original video collection;  $L_u$  and  $L_L$  provided by user is the upper bound and lower bound of length of final presentation respectively;  $L_f$ ,  $L_a$  and  $L_s$  are the length of created presentation for family, acquaintance and outsider respectively.

Table 4.2: Details of original videos and the corresponding three presentations.

Before viewing the video presentations, the subjects were given the aim of these presentations and the relevant information: Video A is related to one specific graduation ceremony; Video B is related to a birthday party; Video C is related to a wedding party; We give an example of the original videos in Fig. 4.2.  $L_o$  is the total length of original video collection;  $L_u$  and  $L_L$  provided by user is the upper bound and lower bound of length of final presentation respectively;  $L_f$ ,  $L_a$  and  $L_s$  are the length of created presentation for family, acquaintance and outsider respectively. All relevant videos can be found in the URL: <http://mmas.comp.nus.edu.sg/NUSAAP/nusaap-JP.html>. After viewing each presentation, each subject was asked to compare them with the original



Figure 4.2: Example of original videos. Note: The region in the yellow circle represents the detected face.

one, and rate the presentation on a scale of 1 to 5, with 1 represents worst, 5 represents excellent. Table 4.3 shows the results of the evaluation.

Presentation	A	B	C	Average
Family	3.2	3.7	3.3	3.4
Acquaintance	3.3	3.4	3.7	3.5
Outsider	3.4	3.7	3.6	3.6

Table 4.3: Results of user study

From Table 4.3, we can see that on the whole, the presentations performed quite well with scores of over 3 in all. Looking at the scores of each individual summary, all the presentations constructed for video A - "Graduation Ceremony" - get the relative lowest scores. This can be explained: this original video is recorded by an amateur with a relatively inexpensive camera. As a result, low quality of original video such as shaking effect, sound noise and low resolution affects the shot detection, the emotion classification and even the identification of persons within video. All of these can affect the construction and impression of this presentation, thus influencing the score of family presentation of video A. Furthermore, we have noticed that all home videos suffer with the problem of noise [YK02]. Therefore, as one of our future works, removing noise and consideration of aesthetics should be introduced into our method in order to create more satisfactory presentations.

On the other hand, compared to the scores of presentations for family and acquaintance members in each video theme, the scores of presentation for outsider are almost the highest scores. Since an outsider presentation just cares if the content can mostly catch the viewer's attention, this presentation reflects the interesting parts in that collection.

Therefore, subjects can easily judge this presentation.

## 4.5 Summary

In this chapter, sparse presentation of affective analysis is used to classify the affective states in home video for the first time. By comparing different fusion of visual component and audio component, it looks like that the decision-level of fusion is a better fusion way. Additionally, based on the proposed three properties of videos and “diversity” factor, home video presentations for family, acquaintance and outsider are automatically created. The test results are good and satisfactory.



## Chapter 5

# A Multimodal Approach For Online Estimation of Subtle Facial Expression

In Chapter 3, we proposed a computational framework to bridge the representation and modeling from the affective video content to the *categorical emotional states* psychological model. Furthermore, we developed a useful system to adaptively present the home video based on our affective model in Chapter 4. Likewise, in this chapter we develop another system to recognize the subtle expression of humans by combining the facial expression with eye gaze information.

### 5.1 Introduction

Human Centered Computing is an emerging field which aims to provide natural ways for humans to use computers as aids. It is argued that for the computer to be able to interact with humans, it needs to have the communication skills of humans. One of these skills is the ability to understand the emotional state of people. In humans, emotion fundamentally involves “physiological arousal, expressive behaviors, and conscious experience” [Mye04]. Facial expression is the most expressive way humans display their emotions. Therefore, extracting and validating emotional cues through analysis of users’ facial expressions is of high importance for improving the level of interaction in man

machine communication systems.

A facial expression results from one or more motions or positions of the muscles of the face [EF78]. These movements, including both global motions like head rotation and local motions like eye or mouth movements, all convey the emotional state of the individual to observers. Ekman [EF78] found six classical categories, referred to as the universal emotions: happiness, sadness, surprise, fear, anger, and disgust. However, humans rarely display those emotions in a clear unambiguous manner [KWS98]. Subtle facial expressions are more often involuntary as expressions are closely related to emotion. It is nearly impossible to avoid expressions for certain emotions, even when it would be strongly desirable to do so. In other words, there always exist some subtle cues, no matter how unnoticed they are. The current existing approaches to measure facial expression are categorized into three characteristics [PK09]: the location of facial actions, the intensity of facial actions, and the dynamics of facial actions. However, it is difficult to acquire these three characteristics of subtle facial expression. Additionally, people do not always portray extreme facial expressions in normal situations. Therefore, there is an urgent need for effective computational methods for analyzing subtle facial expression which are brief, precise and appropriate for an online system.

The idea that the eyes are clues to emotions - “windows of the soul,” as the French poet Guillaume de Salluste wrote - is almost commonplace in literature and everyday language [Hes65]. A person’s eyes can reveal much about how person are feeling, or what person are thinking. For example, the blink rate can show how nervous or at ease a person may be, and even the stress levels the person is feeling [Tsu99]. Among these eye responses, pupillary response is a physiological response that varies the size of the pupil, either resulting in constriction or dilation (expansion), via activation or deactivation of the iris dilator muscle. There are a variety of causes which result in changes on pupillary size, such as an involuntary reflex reaction to exposure or in exposure to light, and interest in the subject of attention. Specifically, dilated pupils indicate greater affection or attraction, while constricted pupils send a colder signal [Hes65]. Thus, the pupillary dilation and constriction of eyes is a significant cue to convey some messages of emotion to observers. Moreover, the pupillary reaction is involuntary which can not be faked [PS03]. Therefore, taking advantage of eye pupil data into analyzing subtle

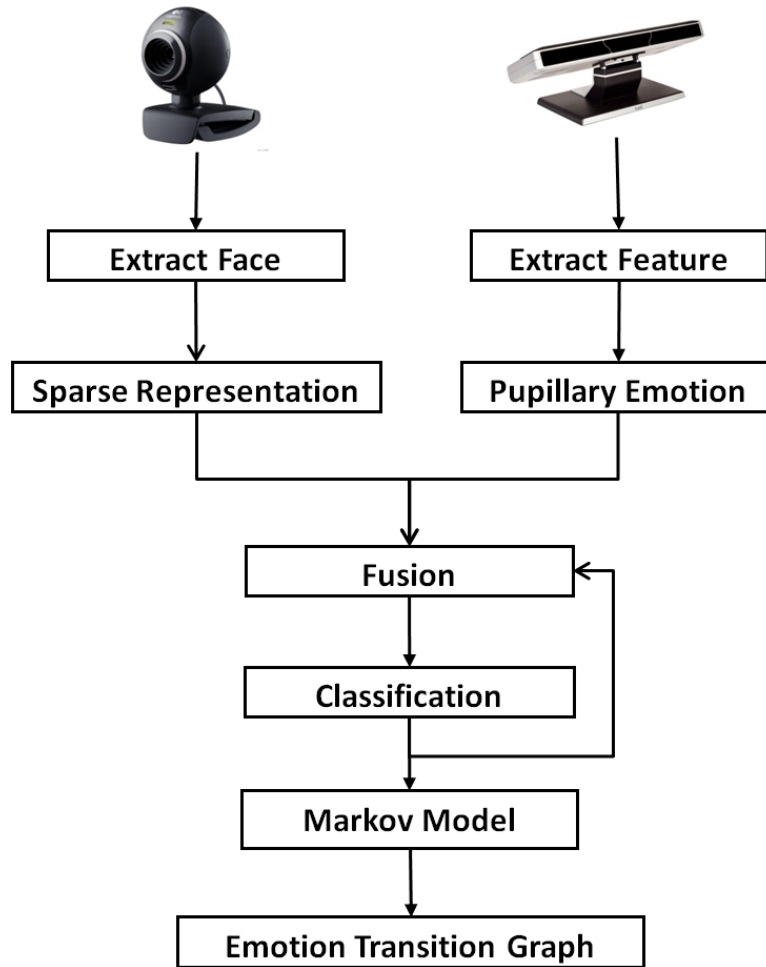


Figure 5.1: The overall framework of our proposed approach

expression is preferred. We find that the concurrent use of multiple modalities: facial expression and eye pupil data is a viable strategy [XK12].

In this chapter, we propose an approach for online estimation of subtle expression exploiting multiple modalities: facial expression, pupil size and previous emotional state. The whole framework of our approach is shown in Fig 5.1. In this chapter, our key contributions are:

- Our novel sparsity-based, multimodal method is proposed to analyze the subtle expressions of human. Compared to existing approaches of analyzing the subtle expression, our method is relatively simpler and faster.
- A novel concept to model person’s emotion changes is proposed. This provides us a predictive estimation about the probabilities of transition among individual’s emotions in the future research on analyzing individual’s emotion.

- To the best of our knowledge, our work is the first work to fuse facial expression, pupillary emotional response and previous emotion for analyzing the current emotional state of human.

This chapter is organized as follows. Section 5.2 reviews the related work to serve as a preamble, and Section 5.3 describes the proposed methodology. Section 5.4 presents the experimental results of the proposed methodology, and finally, conclusions are drawn in Section 5.5.

## 5.2 Related Work

### 5.2.1 Facial Expression Recognition

As discussed in subsection 2.2, many researchers [CZLK98, CSG<sup>+</sup>03, SWB<sup>+</sup>06, PK09, PC11, SZPR12, YWH10] have focused on facial expression recognition, and proposed a variety of methods to recognize human’s facial expression, such as: optical flow method [CZLK98], Bayesian classifier [SGM06], and Hidden Markov Models [CGH00]. As far as we know, few researchers have focused on detecting subtle facial expressions, and they all are offline implementations.

### 5.2.2 Multimodal Human’s Emotion Analysis

As discussed in subsection 2.3, many works [ZTL<sup>+</sup>04, GP05, JNL<sup>+</sup>05, CCK<sup>+</sup>07, NGP11a] have taken advantage of multiple channels to increase the robustness of system and improve interpretation disambiguation in real-life situations. The emotion of people can be reflected by many channels: facial expression, body language, physiological signal, etc. The eyes also create the obvious and immediate cues that lead to the formation of impressions [FM07]. To the best of our knowledge, no earlier work has taken into account the eye pupillary response information to help infer the human’s emotion state in facial expression analysis, especially used it as an aid to help analyze the subtle expressions.

## 5.3 Methodology

### 5.3.1 Modeling The Changes of Human's Emotion

Six basic emotion classes - happiness, sadness, surprise, fear, anger, and disgust - has been categorized by Ekman [EF78]. A person's emotional state can vary between these six basic emotions in time. We try to answer the question: is the probability of change from one emotional state to any other emotional state in a natural situation the same? Specifically, we also ask do people change from "happy" emotional state to the other extreme, "sad" emotional state suddenly? Is this a common case? And if it occurs often, what is the probability? Therefore, we propose a model to compute and represent these probabilities of transitions. In addition, we argue that the transition probabilities for different people are different. In order to solve this problem, we make the following assumptions:

- People can switch their emotions only among seven emotions: neutral, happiness, sadness, surprise, fear, anger, and disgust.
- There is only environmental (external) stimulation (ignoring biochemical (internal) stimulation), which means that if and only if there is an environmental stimulus, the switch of emotions occurs.
- People are initially in "neutral" state.
- The switch of emotions is not sudden, that is, the current emotion is related to the emotion state at previous instant time.

Markov chain is the simplest model which is a stochastic model that assumes the Markov property, namely that, given the present state, the future and past states are independent. It models the state of a system with a random variable that changes through time. Fig 5.2 gives an example of Markov Chain. Therefore, the Markov chain is very suitable for modeling our problem.

Specifically, the seven emotional states are referred as the state space. The switch of human's emotion has become the transition among emotional states in the state space along with time. Therefore, the transition probability of each pair is calculated by

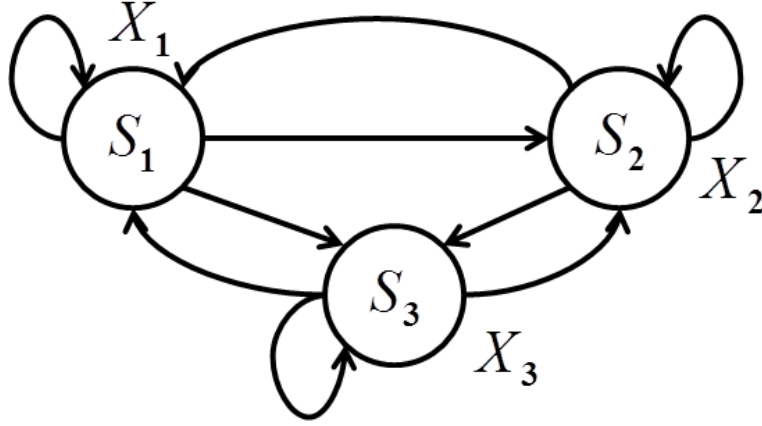


Figure 5.2: A Markov Chain with 3 states (labeled  $S_1, S_2, S_3$ ).

Eq.(5.1).

$$P(S_i \rightarrow S_j) = \frac{\|S_i \rightarrow S_j\|}{\|S_i\|}, i, j \in 0, 1, 2, 3, 4, 5, 6 \quad (5.1)$$

Here,  $S_i$  represents the  $i$ -th emotion. “ $S_i \rightarrow S_j$ ” represents the event that the occurrence of emotional state  $S_j$  is following emotional state  $S_i$  (i.e, the switch from emotion  $S_i$  to emotion  $S_j$ ); “ $\|S_i \rightarrow S_j\|$ ” means the times of occurrence of the event -  $S_i \rightarrow S_j$ ; “ $\|S_i\|$ ” is the times of occurrence of  $S_i$  in entire sequence.

Therefore, based on the collection of prior knowledge of emotion switches, we use Eq.(5.1) to compute the model graph. Finally, one directed graph for group and one special personal directed graph for each subject are created respectively.

## 5.3.2 Subtle Expression Analysis

### 5.3.2.1 Sparse Representation

The main idea of *sparse representation* is that the recognition of facial expression is converted to the reconstruction of a sparse signal. To quote Patterns [Phi99], “When it comes to expressing emotions, members of widely different cultures have much in common,..., Such findings imply that beneath all the cultural complexity of mankind, there is a core of basic emotional expression that is understood all over the world.” For a photograph of facial expression, the interpretations of different cultures people are in accord. We assume that there exists a common pattern for each facial expression to display expression separately. However, which and how the low dimensional features of

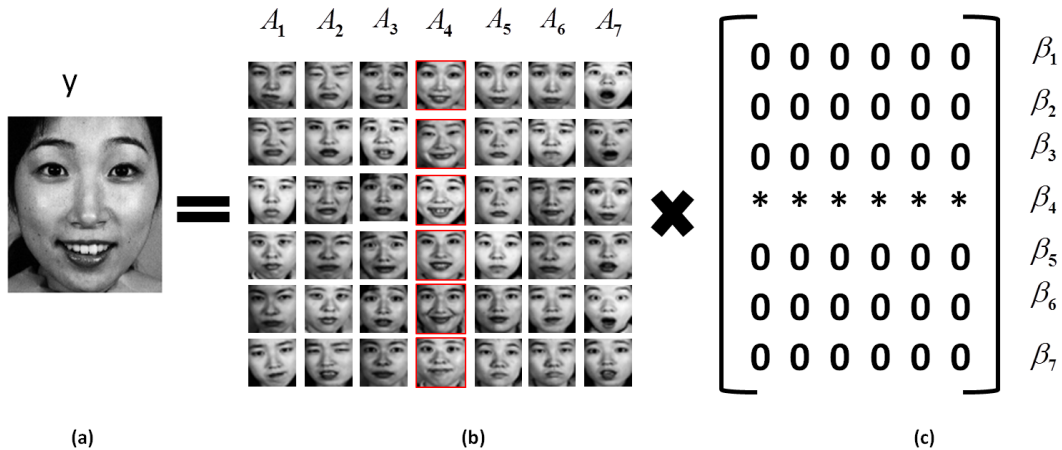


Figure 5.3: One intuitive example for sparse representation of facial expression in the ideal situation. Fig 3(a): test “happy” facial expression image; Fig 3 (b): one example of sample matrix  $A$ . Each column is corresponding to one emotion basis. From the top to bottom, the emotions of rows are anger, disgust, fear, happy, neutral, sad and surprise respectively. Fig 3 (c): sparse solution. In order to view conveniently, the sparse solution which should be a column vector is organized as a matrix, and each entry in it corresponds to the sample in Fig 3 (b). (“\*” represents the nonzero value, and “x” represents the product of matrices); All these facial expression images are from JAFFE database [LAKG98].

a facial expression images are the most relevant for recognition? It is difficult to give a specific emotional pattern in terms of age, skin color, etc. Fortunately, sparse representation can overcome this problem as long as there are sufficient and over-complete samples. Given a set of representative facial expression images for each emotion class can be found as the basic expression patterns of that emotion class, each facial expression image coming from that emotion class can be represented as a linear combination of the basic expression patterns. Therefore, the equations in Section 3.3.2 can be employed in this case but the  $y$  refers to the test facial expression, the sample matrix  $A$  is formed by the presentative facial expression pattern images of each emotion. We provide an intuitive example about this concept as Fig 5.3.

Considering the global sparse representation, we classify test facial expression image based on how well the sparse entries associated with the patterns of each facial expression. Furthermore, for each facial expression, we can construct a temporary image to represent it based on the sparse solution. Similarly, the function  $\Phi_j(x)$  defined in Section 3.3.2 returns a new vector whose nonzero entries correspond to the coefficients belonging to  $j$ -th class in the obtained solution  $\tilde{x}$ . The temporary image for  $j$ -th facial expression

denoted by  $I_j$  is constructed using Eq.(5.2).

$$I_j = A * \Phi_j(\tilde{x}) \quad (5.2)$$

In addition, the difference between original feature image  $y$  and temporary image  $I_j$  is named *Sparse Confidence* denoted by  $SC_j$ . Finally, we use the following Eq.(5.3) to obtain it.

$$SC_j = \|y - I_j\|_2 \quad (5.3)$$

### 5.3.2.2 Eyes' Pupillary Response

As mentioned in previous section, combining different informative cues to analyze humans subtle emotion is preferred, since humans rarely display extreme facial expressions on their faces. “eye tracking and pupil size variation can provide useful cues to discriminate emotional states” [PS03]. So taking advantage of eye tracking and pupil size into analyzing humans emotion is a possible way. On the other hand, Hess [Hes65] argue that “dilation and constriction of pupils reflect not only changes in light intensity but also ongoing mental activity” and extreme dilation to interesting or pleasing stimuli and extreme constriction to unpleasant or distasteful material.

When people receive stimuli, the pupil size will change: dilation or constriction along with the changes of stimuli. One key point or fact is that people can not control their pupillary response. Thus, the pupil size can provide a more reliable cue for analyzing human emotional state even when people try to hide or control their expression such as facial expression. On the other hand, there is considerable variation in the maximum pupil size in any human age group [ASS00]. Therefore, it is not reasonable to use a single and fixed value to represent the pupil size of all people. The definition of dilation or constriction of pupil of each person also should be different.

$$\left\{ \begin{array}{ll} \textit{Neutral} & |\Lambda - \mu| \leq 2\sigma \\ \textit{Dilation} & \Lambda - \mu > 2\sigma \\ \textit{Constriction} & \mu - \Lambda > 2\sigma \end{array} \right. \quad (5.4)$$



**Require:**

“Sparse Confidence” values - SCs;

Previous emotional state -  $E_0$ ;

The current pupil emotion -  $E_p$ ;

**Ensure:**

$E_c$

1: **if**  $E_p == \text{NEUTRAL}$  **then**

2:      $E_c = \text{“neutral”}$

3: **else**

4:     **if**  $E_p == \text{DILATION}$  **then**

5:          $E = \text{the set of positive stimuli};$

6:     **end if**

7:     **if**  $E_p == \text{CONSTRICTION}$  **then**

8:          $E = \text{the set of negative stimuli};$

9:     **end if**

10:    **if**  $E_0 \in E$  **then**

11:        $E_c = E_0$

12:    **end if**

13: **end if**

14:  $E_c = \arg \min(\text{SCs}) \text{ in } E.$

**Algorithm 4:** The Overall Algorithm For Recognition Of Subtle Expression

We use Eq.(5.4) to quantitatively define the dilation and constriction of pupil size for individuals.  $\mu$  and  $\sigma$  represent the mean and standard variation of individual pupil size in neutral stimuli respectively, and  $\Lambda$  represent the instant pupil size of a person.

### 5.3.2.3 Recognition of Emotion

Based on the *Sparse Confidence* values (SCs) computed by Eq.(5.3), we take into account the pupillary cue and the previous emotional state using the decision-level fusion. The overall algorithm is summarized in Algorithm (4). Finally, the result  $E_c$  of classification of emotion is fed into the Markov Chain to update the personal directed graph of transition among seven emotions. Upon obtaining sufficient individual data of emotion changes, the personal directed graph is closer to the approximate of *real* transition directed graph.

## 5.4 Experimental Results

We design three experiments as for the following objectives: firstly, we aim to understand the transition probabilities between seven basic emotions for human, because the emotional state of human at previous instant time is considered in our algorithm. secondly, we try to test the performance of sparse representation used on facial expression analysis; finally, compared to only using facial expression information, the performance would improve with the aid of eye gaze information and previous emotional state information.

Emotion	Neutral	Happy	Surprise	Anger	Disgust	Fear	Sad
Neutral	75.58	0.30	0.53	9.73	1.29	0.53	12.03
Happy	0.13	91.95	1.49	1.34	0.89	3.09	1.11
Surprise	0.39	1.19	89.16	0.28	0.34	1.53	7.10
Anger	8.79	2.00	0.48	71.14	6.45	0.95	10.18
Disgust	1.69	1.71	0.73	8.09	82.34	0.94	4.49
Fear	1.67	13.44	6.57	2.99	2.33	64.53	8.47
Sad	4.06	0.67	4.43	3.71	1.27	1.13	84.72

Note: Each element  $p_{i,j}$  represents the transition probability from  $i$ th emotion to  $j$ th emotion. Here, all values are 100 \* probability.

Table 5.1: The transition probability matrices for group and one person respectively.

### 5.4.1 Modeling Human’s emotion changes

In order to compute the group graph of average transition probabilities between seven basic emotions: happy, sad, fear, anger, surprise, disgust and neutral, we invite 10 people as our subjects (five males and five females). We download about 55 videos from Youtube and Tudou. We downloaded videos as we were unable to find any standard dataset of videos for emotion recognition. We assign an emotion label to a video, only if three participants all choose that emotion label. Finally, 27 videos have been selected and each emotion class has three videos. We concatenate these videos into one long video of 23 mins and 14 seconds. The ten participants were seated in front of a common camera of a resolution of 640 x 480 and required to show their facial expressions in a natural way. An emotion application which is developed by Sebe group [SLCH02] is used to record their emotions along with time. The classification obtained from the emotion application is manually verified to eliminate the classification error by the application.

Emotion	Happy	Surprise	Sad	Anger	Disgust	Fear	Neutral
Happy	<b>83.87</b>	57.14	58.07	60.00	65.52	62.50	56.67
Surprise	41.94	<b>89.29</b>	32.26	33.33	34.48	40.63	36.67
Sad	61.29	53.57	<b>83.87</b>	66.67	65.52	62.50	53.33
Anger	67.74	32.14	38.71	<b>90.00</b>	44.83	56.25	16.67
Disgust	12.90	25.00	9.68	20.00	<b>89.66</b>	15.63	36.67
Fear	25.81	32.14	32.26	43.33	48.78	<b>93.75</b>	26.67
Neutral	64.52	28.57	51.61	50.00	37.93	56.25	<b>86.67</b>

Table 5.2: Person-independent confusion matrix for classifying facial expressions using sparse representation.

The collection of data is fed into our Markov Model to compute a group directed graph model representing the average transition probabilities among seven basic emotions as shown in Table 5.1 and one personal directed graph model for each subject by the method we propose.

#### 5.4.2 Sparse Representation In Analyzing Facial Expression

We use the Japanese Female Facial Expression (JAFFE) Database [LAKG98] to do the experiments of sparse representation of facial expressions. Because the choice of sample matrix is still a challenge in compressive sampling area [NT09], random generation is one way to choose the basic expression pattern images from database for us. Therefore, 250 sample matrices are generated, and 16000 experiments have been done based on different sparse level parameter in sparse representation. Of course, the columns of these sample matrices have overlap. For each emotion, we find the largest classification rate, and they are organized in a confusion matrix as shown in Table 5.2. From Table 5.2, we can see that based on different sample matrices, all of the best classification rates for each emotion are over 85% in person-independent experiment. Moreover, the classification rates of “anger” and “fear” for person-independent are even over 90%. Specially, the results on the recognition of emotions: fear, surprise, disgust, anger has the best performance in person-independent experiments of facial expression recognition with JAFFE database. These results convince us that the sparse representation of facial expression would have higher performance on larger database for some other optimal sample matrices.



Figure 5.4: The setup of experiments.

### 5.4.3 Experimental Results For Subtle Facial Expression Analysis

For our experiment to determine subtle expression, we use the videos as stated in Section 5.4.1. These videos have been classified under different emotional tags: happy, sad, surprise, anger, fear, disgust and neutral. We try to use those videos to stimulate the subject's emotion. Here, we have to point out that it must be the first time for the subject to watch these videos, so that their responses are direct, reliable and authentic, required for testing our approach. The camera we use in the experiment is a common camera of 640 x 480 solution. The eye-tracker we use is the product of SMI group. This SMI eye tracker provides a robust and real time analysis for gaze position, eye movements and pupil dynamics, record stimulus screen content and automatically analyze user-defined areas of interest within the subject's field of view. The eye-tracker uses 250 Hz to capture the diameters of pupils. Fig. 5.4 shows an example of that one subject is seated in front of eye tracker, web camera, and a computer which is playing a video stimuli to evoke the emotion of subject.

We design two experiments for different goals. In the first experiment, the subject

Emotion	Happy	Surprise	Neutral	$EP_2$
$N_{detected}$	50	51	103	95
$N_{correct}$	13	50	95	87
Rate	26.00	98.04	92.00	91.58

Note:  $N_{detected}$  represents the total number of captured subtle facial expression images.  $N_{correct}$  represents the number of subtle expression images which are recognized correctly.  $EP_2$  represents the second experiment that tests the ability to tell the “real” emotion.

Table 5.3: The experimental results for the proposed subtle facial expression recognition method.

is required to be seated in front of the camera and eye-tracker to watch the videos from each of the emotional tags and display their natural emotions. When the subject finishes watching one video, she is also required to state her emotional state when she is watching that video. If and only if the emotion declared by the subject is in accord with the video’s emotion, then the corresponding experimental results can be used. This is because different people display different emotions for the same video. Thus, this step is to make sure the experimental results are reliable. After that, the experimental results corresponding to exaggerated facial expressions are eliminated since we aim to test the classification rate of subtle facial expression. In the second experiment, the subject is required to display wrong facial expression, that is, the emotional expressions on face are not the “real” emotions she is feeling. Finally, the experimental results of two experiments are shown in Table 5.3.

From Table 5.3, we can see that the ability to tell the *real* emotion of humans reaches 0.9158 by our approach. On the other hand, all of the recognitions of “surprise”, and “neutral” subtle facial expressions are over 90%. Compared to the results in Table 5.2, the performances of recognizing “surprise”, and “neutral” emotions get improvements with the help of eye pupil information. The detected images of other emotions after we eliminate the expressive images are very rare, so the results of these emotions are not listed in this table.

Finally, we build a complete system with friendly interface in terms of the proposed method as Fig. 5.5 shows.

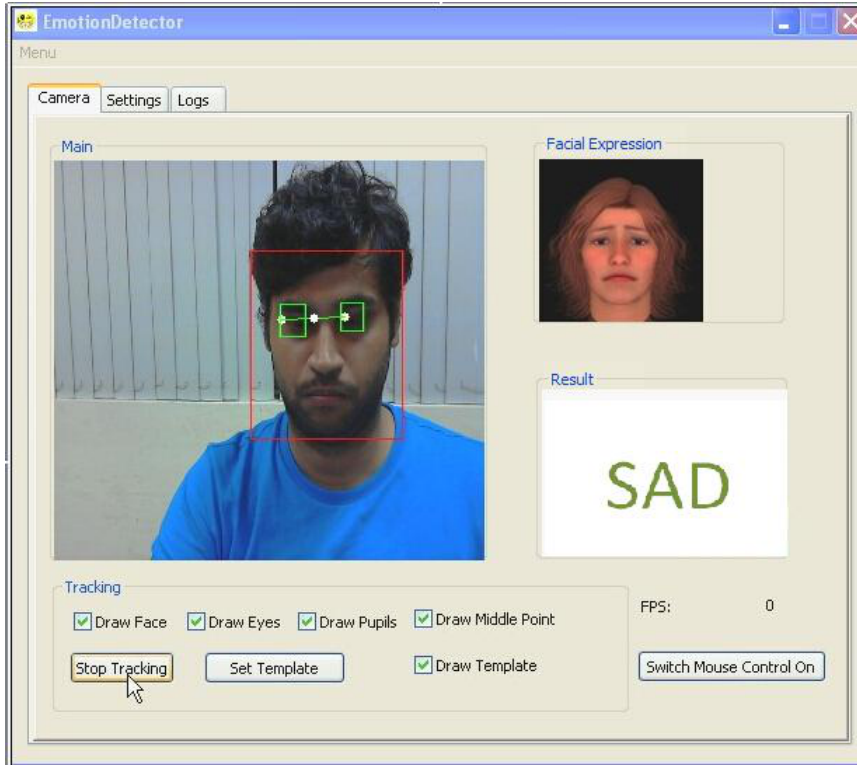


Figure 5.5: An instant of our developed system to identify the emotion of human. The left window is showing the image captured by camera, and the red window is showing the detected face. Green window represents the position fixed by eye tracker. The right side of this photo is showing the current recognized emotion.

## 5.5 Conclusions

In this chapter, we propose a multimodal approach for online estimation of subtle facial expression. A novel sparsity-based facial expression analysis is proposed to recognize the facial expressions. To the best of our knowledge, our work is the first work to fuse the facial expression, pupil size and previous emotional state to classify the subtle facial expressions. Additionally, we propose a novel concept to model the transition probabilities among the seven emotions. The establishment of the model can provide a predictive estimation about switch of emotions, which is good for the future research of individuals' emotion changes. The experimental results show that: first, the sparse representation has a good classification rate on facial expression. Second, the fusion of facial expression, pupil size and previous emotional state is a promising strategy for analyzing subtle expressions.

## Chapter 6

# Social Photo Sharing

In Chapter 3, a computational framework to bridge the representation and modeling from the affective video content to the *categorical emotional states* psychological model has been built. Furthermore, we developed two different applications to take advantage of the proposed affective representation and modeling respectively in Chapter 4 and Chapter 5. In this chapter, considering the affective information of images, we also develop another application to generate a suitable subset of photos from the personal photo collection for sharing with different social kinship groups. It can also be used to determine whether an individual photo is appropriate for sharing with a particular kinship group.

### 6.1 Introduction

The current proliferation of consumer cameras has enabled people to easily record and collect a huge amount of multimedia content like personal digital photos. The digital camera has become a very convenient means for home users to preserve their meaningful moments or experiences such as weddings, holiday trips and birthdays. The bundling of mobile phones with built-in cameras has also contributed towards the enlargement of personal digital photo collections. While home users build large photo collections, they usually pick a few selected photos to print and decorate their homes and offices, and make them beautiful, comforting, and warm. Moreover, with the popularity of DIY (do it yourself) gifts like DIY mugs, T-shirts, and posters with personal photos are becoming

popular gifts, especially among young people. Personal digital photos are thus becoming an important part of our lives. Additionally, many platforms to enhance the sharing experience with people at a distance have sprung up with the advancement of social networks, such as FaceBook and Flickr. Sharing personal photos through participatory media have become a significant means of interaction between people [Bab07].

However, there are some issues encountered when people try to select from their personal photo collections. First, the photo collection is massive that grows with time. Manually organizing and managing the photo collection is a tedious and time-consuming task. Second, duplicate photos are often present in the photo collection, as people are used to taking multiple photos to guarantee that at least one of them is of good enough quality for preservation and sharing. As a result, it is very inefficient and time-consuming to manually browse through collections or select photos from them. Third, most photos are taken by people with amateur skills using amateur devices. Therefore, the low quality and artifacts-ridden photos increase the work needed for filtering. Furthermore, distinct individual preferences have been observed for such video and image content [XK11]. This in turn makes it necessary to select and show different photos to cater for the diversity of preferences. Taking a photo collection of birthday party as an example, people would rather print a photo with family members than with unfamiliar guests. However, an amusing photo could catch the attention of outsiders as compared to a family photo, when it is shared over a public network. Furthermore, it is not a good idea to keep all the photos in online albums. This is because that it wastes other people's time to go through the entire album. In view of these prevailing problems, there is an urgent need for an effective solution for selecting a small set of personal photos that are engaging yet appropriate for the target viewer preferences.

Not surprisingly, many research efforts have been expended on developing techniques for organizing and managing the personal photos, such as the "albuming" tools like MyPhotos [SZZL02], SmartAlbum [TCMK02]. For instance, a query-based "albuming" tool allows user to select the photos with the annotations including the query keywords. To the best of our knowledge, none of them help users *automatically* select the appropriate photos when they want to share with different social kinship groups, without requiring extra manual input from them. The selection of photos for sharing is not com-



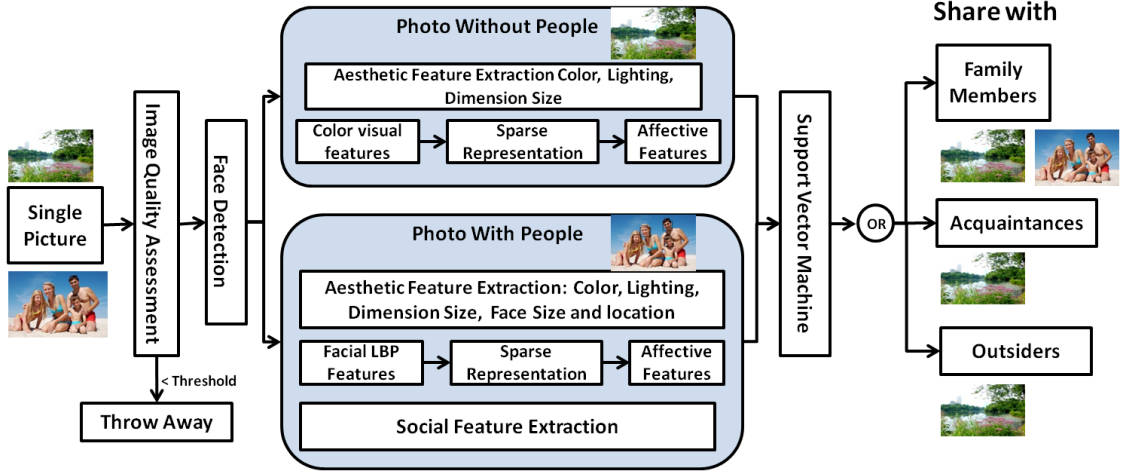


Figure 6.1: The Framework of the proposed approach. The input single photo would be fed into different processing components (with detected face or without detected face) depending on the result of face detection.

pletely equal to choose the photos which are more memorable. Recent studies shows that it is not an inexplicable phenomenon that variation in memorability of images is consistent across subjects, and indicating that some images are intrinsically more memorable, independent of a subjects' contexts and biases [IPO11]. However, the photo sharing is not independent of a subjects' contexts and biases based on the considerations of the privacy of subjects, which is also validated by our experiments.

In this chapter, we first consider three social groups: family members, acquaintances, and outsiders, and then consider two cases of the provided personal photos: a photo collection (album) and a single photo. For the photo album case, we first make use of content similarity to cluster photos. Then, one single photo possessing the best image quality would be selected as the representative photo for each photo cluster. Consequently, the single photo will be processed by our proposed approach whose framework is presented in Fig. 6.1. Aesthetic features, affective features, and social features are extracted from the photo separately. Finally, the support vector machine (SVM) classifier and the maximum likelihood method are employed for classification. The main contributions of our work are as follows:

- We propose the concept of adaptive sharing of photos with different social kinship groups.
- We utilize the *affect* factor of photos for the selection process, going beyond only

facial expressions.

- We have developed an approach that works on a photo collection and for an individual photo.

This chapter is organized as follows. Section 6.2 reviews the related work to serve as a preamble, and Section 6.3 describes the proposed approach. Section 6.4 shows the experimental results, and finally, the conclusions are provided in Section 6.5.

## 6.2 Related Work

Many efforts have been put on the development of techniques to help home users organize, manage, browse their huge photo collections including indexing and retrieval. Mulhem et al. [MLLK03] proposed a query-based methodology exploiting the learning-based meaningful visual vocabularies to identify and describe the concepts and relations in the photo contents to semantically index and retrieve photos. Neil et al. [OLC<sup>+</sup>06] presented a MediAssist demonstration system to manage personal photo collections based on contextual information and content-based analysis. It is also a query-based tool to retrieve photos in terms of semiautomatic annotation techniques. MiAlbum developed by Liu et al. [WSZ00] was a system with semi-automatic image annotation methods to tag, browse, retrieve, group and export (send or print) photos. Besides, MyPhotos [SZZL02], SmartAlbum [TCMK02] Photoware both had a friendly visualization interface with thumbnails to allow people to access their collection as well as to organize and manage photos based on the folder metaphor. Sentic Album [CH12] is a novel online personal photo management system which intelligently organizes, annotates, and retrieves photos based on the content concept with contextual information. Chu et al. [CL08] selected the representative photo by modeling the mutual relation of near-duplicate photo pairs. Guldogan et al. [GKG13] could generate the personalized image(s) subset from a give album based on the “interest set” of that particular user.

However, to the best of our knowledge, none of them consider the concept of adaptive selection of photos for different social kinship groups. Moreover, they do not use the affect of photos other than using the facial expressions.

## 6.3 Methodology

In order to automatically decide whether a photo is appropriate for sharing with a particular social group, or to generate a suitable subset of photos from the photo collection to share with a particular social group, we consider three factors:

- The Aesthetic Factor
- The Social Factor
- The Affect Factor

Aesthetic visual quality assessment of photos has been studied in detail since it is known to influence the photo selection process of most people. As we know, humans are more interested in the things that are more visually appealing than others. Therefore the aesthetic quality of photo is critical for assessing the utility of a photo.

People overwhelmingly attend to humans in images and videos with the faces dominating visual attention [CXF<sup>+</sup>03]. In other words, faces within the photos are often the main areas of attention. The face information implicit in a personal photo, such as the relative importance of faces, the personal relationship between the faces, and even the relationship between a face and the owner of the photo, plays a vital role in the consideration for sharing that photo. For most people, photos with unknown people is less important than a photo capturing themselves. We refer the above-mentioned information as the *social features* of a photo. Thus, when faces are detected in a photo, it is necessary to consider the social features hidden in that photo.

Finally, we argue that people may like photos which make them think or make them emotional. The photos could make them happy either explicitly because of the content of the photo or implicitly because of the associations or memories linked to that photo. In either case, this phenomenon reflects the fact that the affect contained in the photo is also an important factor in the photo selection process of most users. Here, the affect of the photo mainly refers to the emotion evoked in the viewer when he or she views this photo. Additionally, as stated in [CL09], the facial expression is a very important attribute of a photo with people. Therefore, when faces are detected, we also need to recognize the facial expressions contained.

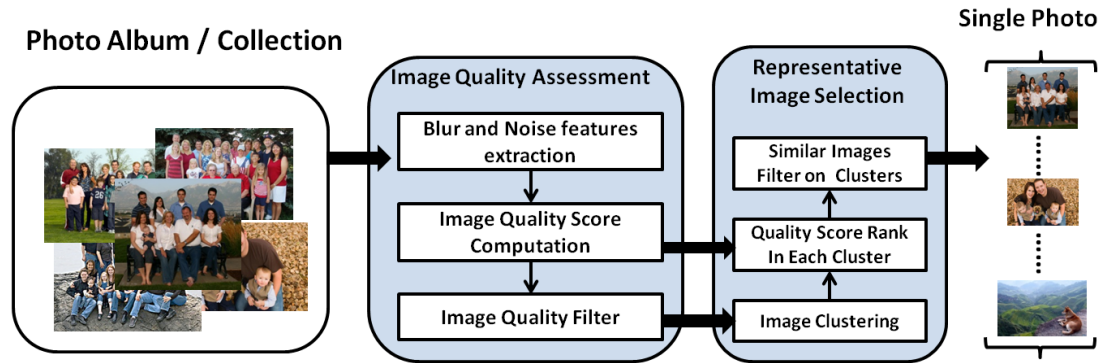


Figure 6.2: The overview of pre-processing when a photo album or collection is provided.

So far, we have provided the motivation behind considering the aesthetic, social, and affect features of photos. In the following subsection, we will detail the proposed approach.

### 6.3.1 Pre-Processing of The Photo Album

Clustering of the photos is a necessary basic step for photo organization. One photo cluster can contain any number of photos, and a photo could belong to a few different clusters in many “albuming” tools. This is different from our approach which only concerns with the content similarity of photos, and each photo therefore belongs only to one photo cluster.

With the proliferation of digital cameras, a large number of personal photos are being rapidly accumulated. Duplicate photos, photos with similar content and photos of low quality are unavoidable. It is not very useful to retain all such photos. Extracting the blur and noise features of each photo, we first take advantage of the no-reference image quality assessment method [CJJ09] to compute a score for each photo. The photos of low image quality would be filtered out based on the threshold score (0.91 in our experiments). Then, the remaining photos are categorized into a few clusters based on the content similarity between photos within a cluster. The content similarity is computed by block-based least square error combined with time stamp. The photo with the highest image quality score in each photo cluster would be selected as the representative photo to feed into the assessment component. This whole framework is shown in Fig. 6.2 when users provide a photo collections instead of a single photo.

Next, based on the results of face detection procedure, the photos would be divided into people or non-people category. Then the aesthetic feature, affective feature and social feature extraction are applied to the two categories and is detailed in the next subsection.

### 6.3.2 Assessment Factor Features

#### 6.3.2.1 The Aesthetic Factor

In the art and photography field, aesthetics refers to the appreciation of beauty and principles of the nature. Since it is a fact that humans are more interested in visually appealing things, the photos with high aesthetic quality would attract more interest and are more willingly shared by people. Therefore, it is meaningful to consider its aesthetic visual quality when we recommend a photo to share.

The aesthetic features of personal photo is highly related to the following factors: the environmental conditions, the quality of the digital camera, and the photographic technique. [LC09][CL09] have empirically studied the perception of aesthetics in photography, and demonstrated that color, lighting, size and location of main subject were the important factors when judging the beauty of a photo. Their strategic use would highly affect the aesthetic visual quality, as well as the impression of viewer. In the case of personal photos, we consider the most significant objects to be the faces. Therefore, the color and lighting features [LC09] are extracted across the entire photo when it has no face detected. On the other hand, we focus on the color and lighting features of its background region (the region outside of faces) while faces are detected in that photo. Additionally, the size and its relative location of faces are also computed as the aesthetic feature. Specifically, each image is divided uniformly into 9 regions which index from left to right, and from top to bottom. Each region will be assigned a weight  $\tilde{w}_i \in [1, \dots, 9]$  representing the importance in terms of the aesthetic knowledge. In order to learn  $\tilde{w}$  for each region, we compute the normalized histogram of face occurrences of the 9 regions from all the samples with faces to represent the  $\tilde{w}$ . In terms of our database,  $\tilde{w}_1, \dots, \tilde{w}_9$  are 0.03, 0.23, 0.03, 0.12, 0.35, 0.14, 0.02, 0.06, 0.02. The size of face is computed by  $size(face)/size(image)$ .

### 6.3.2.2 The Social Factor

As discussed earlier, faces in personal photos are considerably important elements drawing more attention. It is imperative to discover the potential information among the faces in the photo, such as the importance of faces, the relationship between the faces and the relationship between the face and user.

In the chapter we assume that the social network of the user would be built in advance to indicate the interpersonal relationships. This could be built either via explicit social network platforms or implicitly (such as by using a conversational social network). This social network not only includes the daily interpersonal relations, such as family members, intimate kinship-based relationships as well as formal institutional relationships but it can also contain special relationships depending on the preference of the user, such as his/her most favorite celebrities. Intuitively, when someone deliberately takes a photograph with a movie-star or the President, he/she probably would like to share this photo. Although this situation usually occurs rarely, we could take this into account for the sake of completeness.

Thus the initial social network is built from the connection lists of the user, such as phone contact list, email, Facebook, Weibo, and Twitter. Data mining technique can be taken advantage of to extract the required information. For example, we can extract many information, such as the name, phone number, email address, person photo, and group category from the contact information stored in user's phone. We can also find the related information from all the channels, the user takes advantage of, to keep contact with others. As long as the person has at least some contact with the user in the user-defined recency period, he/she would be one node of the social network. Based on the extracted words indicating the relationship, the link in the social network would be assigned an attribute: one of the six categories (family, relative, familiar friend, acquaintance, stranger, and another special category). The special category would be composed of the celebrities – famous persons based on the preference of user. Meanwhile, based on the frequency of their communication, the link would also be assigned a corresponding link weight representing the degree of association between each other. Therefore, each link in the social network has a label of (category, weight). An initial complete graph is established from our approach to represent the social network of user. Each catego-

ry has an importance weight denoted by  $w_i^c$  for  $i \in [1, \dots, 6]$  respectively. The weight representing the degree of association of the linked people  $p_i$  and  $p_j$  is denoted  $w_{(p_i, p_j)}$  and range from 1 (weakest) to 10 (strongest). The weight  $< 0$  represents the fact that the linked people is a complete unknown person, or a disliked celebrity. The larger the value, the stronger is the association of the two linked persons.

While faces are detected, the method by [TP91] is employed to identify the faces in the existing social network. The corresponding social features in terms of the current social network have: the number of the faces, the association value of each pair of faces, between-face distances in the photo, the association value between the user and each face, and the importance of faces. The association value is computed by  $w_i^c * w_{(p_i, p_j)}$ , and the importance of faces is computed by  $\tilde{w} * w^c * size(face)/size(image)$ .

### 6.3.2.3 The Affective Factor

This is one of our contributions that affect within photos other than the facial expressions is taken into account for photo sharing. Depending on whether the photo has faces or not, its corresponding affect is either defined as the emotion evoked in most viewers when it has no faces, or the expressions on the detected faces within it. Thus we define the affect as the emotion evoked in most viewers if there is no face detected in the photo. We further define the affect as the facial expression if face is detected. Also, the affect within a photo is represented by a set of seven basic discrete emotional states: neutral, happy, sad, surprise, disgust, anger and fear. In our work, we take advantage of sparse representation approach to recognize the affect of a photo.

The main idea of utilizing sparse representation into affective analysis is that any feature vector representing one emotional state can be represented by a linear combination of a comprehensive set of representative feature vectors representing that corresponding emotional state. More details about the motivation of exploiting sparse representation theory into affective analysis on photo and facial expression are described in Chapter 4 and Chapter 5 respectively.

**Photos without people.** [Kan03] has studied the affect of image can be captured by visual color features. Thus, the extracted features in test feature vector  $y$  and corresponding sample matrix  $A$  are composed of the visual color features.

**Photos with people.** In order to recognize the facial expression using sparsity based approach, we extract the faces from photo, and then exploit face alignment. Linear binary pattern (LBP ) features are extracted for each detected face, and used to form  $y$  and corresponding  $A$ .

Therefore, given the test feature vector  $y \in \mathbb{R}^k$ , which is a  $k$ -dimensional feature vector represents the affective content of a test photo (either facial expression features or the color visual features) and the corresponding sample matrix  $A$ , we have  $y = Ax$  in terms of sparse representation theory. The approximation of  $x$  which is a sparse vector is obtained by solving  $y = Ax$ . Finally, the approximation of  $x$  is the feature we need for affect factor.

### 6.3.3 Social Groups

This is another contribution of our work: introducing the adaptive photo sharing concept. Different people may be interested in different photo content. Ideally, a personalized assessment of a photo should be done for each person individually. However, this is not yet practical. Therefore, as stated in Section 4.3.3.3, the people the user would like to share with are also categorized into three different interest groups for our final sharing assessment.

- Family Members: consists of the people with intimate kinship-based relationships.
- Acquaintances: contains the people with formal and institution relationships.
- Outsiders: consists of the individuals who do not belong to the previous two groups.

### 6.3.4 Classifier Design

The final classification decision of a photo has slight differences because of the used features. Aesthetic, affective and social features are extracted for a photo with face(s) detected, while only aesthetic and affective features are extracted for a photo without detected faces.

Since it is normal to share a photo with multiple social groups, there is a total of seven combinations possible based on the pre-defined social kinship groups: family member, acquaintance, and outsider. They are: 1) family members; 2) acquaintances;



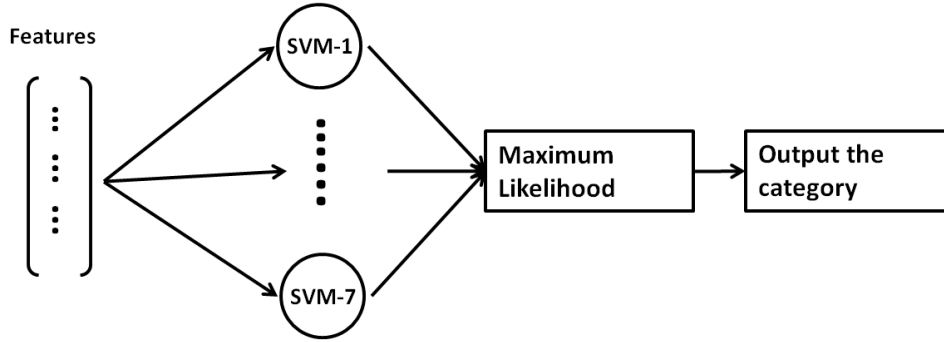


Figure 6.3: The algorithm for assessing which social groups the input photo is suitable for sharing.

3) outsiders; 4) family members and acquaintances; 5) family members and outsiders; 6) acquaintances and outsiders; 7) family members, acquaintance and outsiders. A support vector machine (SVM) with kernel function [CL11] is trained for each possible combination using the combination features of two or three factors. The reason of choosing SVM is because that it can easily capture the complex relationships between the data without having to perform difficult transformations. The combination features of two or three factors are separately fed into these seven SVM classifiers, and the category with the maximum likelihood in the seven SVM classifiers is the final decision for the current photo. Fig. 6.3 illustrate the entire classification process.

In addition, we also train a individual SVM classifier for each social group (family members, acquaintances and outsiders). In terms of slight difference on the feature vector of photos with detected face and photos without detected face, each photo group has completely individual SVM classifiers of social groups. The final decision is made by the vote of the result of each classifier. Specifically, for example, if SVM for family members group output the positive result, then the input image will be shared with family member no matter what results we obtain from other SVM classifier. If and only if all three SVM classifiers (for family members, acquaintance and outsiders) obtain the negative results, then the input image won't be shared with anyone.

## 6.4 Experiments

To the best of our knowledge, there is no publicly available personal photo database for the purpose of studying personal photo management. Thus, we collected a total of

		Decision-level Fusion	Feature-level Fusion
Group of Photos Without People	$SVM_1$	36.16%	32.93%
	$SVM_2$	36.97%	32.2%
Group of Photos With People	$SVM_1$	60.34%	58.87%
	$SVM_2$	46.42%	66.67%

Note:  $SVM_1$  refers to the linear SVM classifier, and  $SVM_2$  refers to the non-linear SVM classifier.

Table 6.1: The results of SVM classifier of person independent.

		Decision-level Fusion	Feature-level Fusion
Photo Without Face Group	$SVM_1$	44.13%	41.85%
	$SVM_2$	46.97%	50.5%
Photo With Face Group	$SVM_1$	85.57%	69.57%
	$SVM_2$	72.23%	69.73%

Table 6.2: The results of SVM classifier of person dependent.

2744 personal photos from 7 subjects. About fifty percent of these photos contain people, including photos which may do not show face but show the body. In order to collect the ground truth, the subjects were asked to answer the question: who do you want to share this photo with? Additionally, we also requested the subjects to modify the initial social network obtained by data mining in order to make our experiments more reliable. Based on the results of face detection algorithm, this photo collection is divided into two classes: photos without people (i.e. no face detected) and photos with people (i.e. face detected). However, based on the face detection algorithm, only 265 personal photos have faces detected. The reasons causing low accuracy of face detection in our home photo collection may be that: these home photos used for experiments directly collect from subjects. We do not make any restrictions about the quality of photos and the content of photos, and we do not also make any selection or changes on these raw home photos when we obtain these home photos. Therefore, faces appearing in these photos can have any pose, any size, and even they may be not clear. Then, it is very normal that the accuracy of face detection reduces in this particular and real-life home photo dataset. The remaining photos are classified as photos without people.

In terms of the extracted features from the social, aesthetic, and affect factors, we use support vector machine (SVM) to train the classifier for each possible category.



Figure 6.4: Image examples for sharing with different social groups.

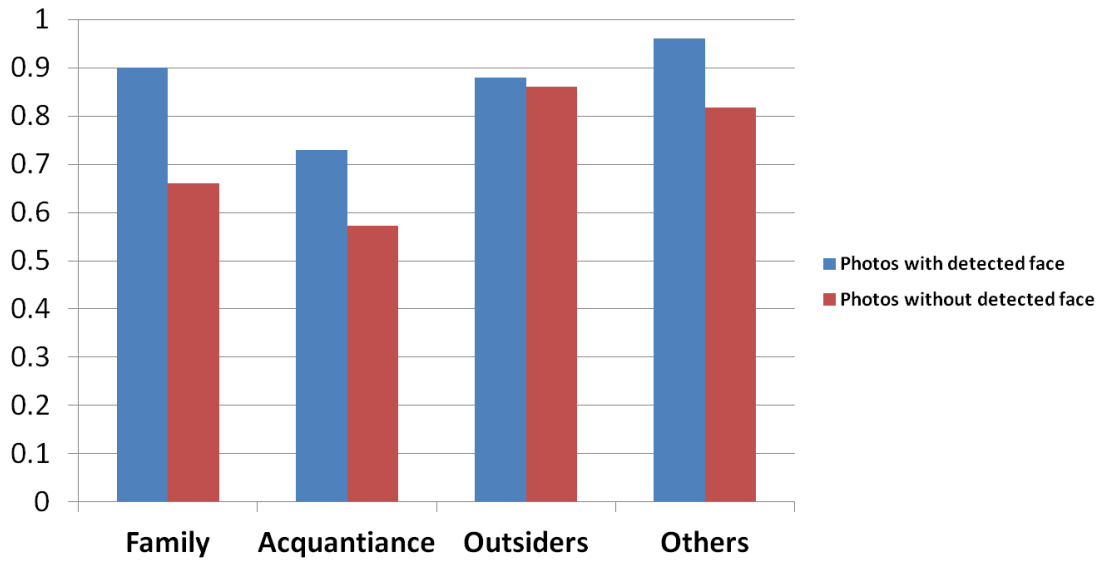


Figure 6.5: The classification results of second classifier design -  $SVM_1$ . Note: “Others” refers to a set of photos which won’t be shared with other people by user.

Meanwhile, leave-one-out cross validation method is used to test the obtained SVM classifier on group of photos with people, and K-fold cross-validation is used for group of photos without people. We design two main scenarios in order to find a better strategy of training SVM classifiers, and understand the selection process of ordinary users as much as possible. One is person dependent which means we train the classifiers based on the samples of each person. The objective is to adequately consider the personal characteristics. Correspondingly, the other one is person independent which considers all the samples regardless of who is the owner of the photo. Fig. 6.4 shows image examples for sharing with different social groups.

Table 6.1 shows the results based on the entire dataset when we consider different classifier and distinct fusion method. From this table, we can see that the classification

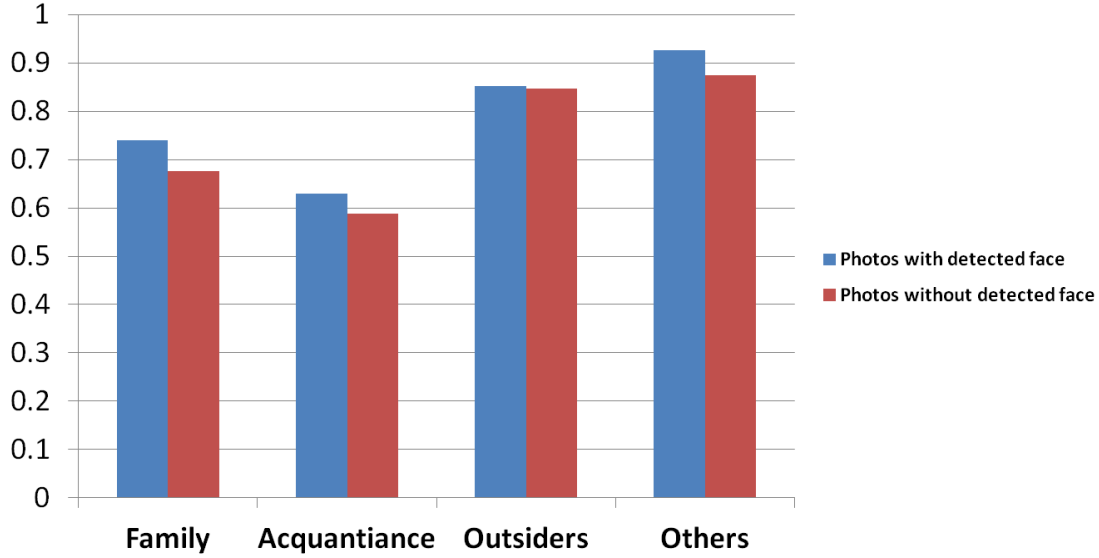


Figure 6.6: The classification results of second classifier design -  $SVM_2$ .

results of group (photos with people) are higher than group of photos without people. This phenomenon demonstrates that the social information hidden among the persons of the photo plays an important role in the selection process of common people.

Table 6.2 presents the person-dependent classification results. As we can see that the results are better than the results of Table 6.1. The performance of classifiers improves. It validates that it is necessary to independently consider the preference of each user enabling the results to better satisfy the user. In future, a self-learning method can be built to learn the preference of current user through each sharing experience. Meanwhile, Table 6.1 and Table 6.2 both show the social information hidden among the persons of the photo plays an important role in the selection process of common people. Compared to Table 6.1, Table 6.2 shows that the preference of user is also very important. Therefore, we claim that the social features are more important and favored by most audiences. Then, we can do more experiments to substantiate this point in the future.

Fig. 6.5 and Fig. 6.6 present the classification results of second classifier design which both are person-independent respectively using linear and non-linear SVM classifier. As we can see that the histograms for group of photos with detected face are higher than these for group of photos without detected face in the both figures. In other words, the results for group of photos with detected face is better than the results for another photo group. We have obtained one significant insight that the social factor of images does

appear to have the most significant role in the photo selection process of most people when they intend to share photos.

Finally, we have to point out one observation from our ground truth collection that no photo (among the 2744) is thought to be appropriate for sharing with family members and outsiders but not for acquaintances. Maybe this photo collection is not sufficient to validate this conclusion.

## 6.5 Summary

In this chapter, we first presented a brief overview of the current photo management methods and system, and outlined several issues we intended to address here. Then, given the motivation of considering aesthetic, social and affect factors of photo, we elaborate on the pre-processing, the extracted features, as well as the division of social kinship groups. Subsequently, our experiments demonstrate the utility of the proposed approach. This method not only automatically generates a suitable subset of photos from the personal photo collection for sharing with different social kinship groups, but it can also be used to check whether an individual photo is appropriate for sharing with a particular kinship group.

# Chapter 7

## Conclusions

### 7.1 Summary

This thesis examined the mapping from the affective content in videos to the categorical psychological models, that is, defining the links between the *categorical emotional states* and low-level features. The detailed modeling of affective content of videos has revealed that the proposed sparse representation can effectively represent and model the affective video content based on the *categorical emotional states* model. The results show the classification rate of the fusion of visual and audio components outperforms that using the audio features only, while slight improvement compared to visual features only. This is perhaps because when the test unit - “shot” - is very short, the features extracted from audio component are not very precise compared to visual features, which influences the final classification result. Also, the results demonstrate that the proposed methods to construct the sample matrix performed efficiently. In parallel, we manage to obtain intensity time curves that represent the degree of contribution of each emotion to the overall affect within a video. The dominant emotion curves are largely in line with the video content. Therefore, the proposed model not only performs the classification of affective video content, but also provides a reliable approach for obtaining the intensity of discrete emotion. We can claim that the proposed approach is more in line with the way that people articulate their emotion experience. To the best of our knowledge, our work is the first work that compute the intensity of emotions considering the categorical emotional states.

Besides, this thesis also presented the importance of the *affect* in the area of *affective computing*, and tested the application of the sparse representation modeling of affective content. A very useful framework has been successfully developed to construct an adaptive presentation of home videos for various social groups: family, acquaintance, and outsider in terms of the *affect* factor and *face* factor. The results of classifying emotions of shots of home videos show that the performances of classification of “happy” decrease, but the performance of “neutral” increases slightly. This may be because of the noise of the home videos which hindered the quality of the audio features of “happy” shot compared to “neutral” shot. By decision fusion of visual component and audio component, only affective events “happy” and “sad” are detected with high accuracy based on a good sample matrix. In addition, the results of user study for the adaptive presentation generated by the proposed algorithms demonstrate that our method is very effective in video sharing and the users are satisfied with the videos generated by our method. Emotion is a subjective feeling which relies on perceptions. Talking about the emotion inevitably leads to a discussion about subjectivity. Thus, it is difficult to compare our experimental results with the others’ results in home video area.

Besides the adaptive presentation of home videos, this thesis also exploited the affective analysis to develop a multimodal approach exploiting the facial expression, eye gaze data and previous emotional states have been successfully proposed for online estimating the subtle facial expression. It is found that the performances of recognizing “surprise” and “neutral” emotions are improved with the help of eye pupil information. Additionally, the results demonstrate that the fusion of facial expression, pupillary size and previous emotional state is a promising strategy for analyzing subtle expression. To the best of our knowledge, this work is also the first work to fuse the facial expression, pupil size and previous emotional state to classify the subtle facial expressions.

Finally, this thesis also utilizes the affective analysis technique to develop a novel approach based on the aesthetic, affective and social features for photo sharing. The results demonstrate the utility of the proposed approach to generate a suitable subset of photos from the personal photo collection for sharing with different social kinship groups. It can also be used to check whether an individual photo is appropriate for sharing with a particular kinship group.

## 7.2 Future Work

In this section, we will discuss the challenges for affective analysis in video and outline the issues that need to be addressed.

### 7.2.1 Subtle Facial Expression Analysis

“Affect” also implies affective display, such as facial expression, or gestural behavior that indicates the affect sometimes [Van07]. Facial expression is the most expressive way humans display their emotions. A facial expression results from one or more motions or positions of the muscles of the face [EF78]. These movements, including both global motions like head rotation and local motions like eye or mouth movements, all convey the emotional state of the individual to observers. However, humans rarely display those emotions in a clear unambiguous manner [KWS98]. Moreover, people do not always portray extreme facial expressions in normal situations. The current existing approaches to measure facial expression are categorized into three characteristics [PK09]: the location of facial actions, the intensity of facial actions, and the dynamics of facial actions. However, the most difficult thing for subtle facial expression is to acquire these three characteristics. So, compared to the methods to recognize the *extreme* facial expression, the fundamental issues for *subtle* facial expression analysis are:

- capture more and reliable visual affective information or features.
- real-time system implementation is still a challenge.
- Depth information using Kinect as well as audio information can be tried in future.

### 7.2.2 Multimodal Emotion Analysis

Intrinsically, the fusion of various modalities can increase the confidence of results of classification. For example, current facial expression analysis techniques are sensitive to the head orientation, luminance, and occlusion. While, the speech processing also is sensitive to auditory noise in current technique. But, the fusion of visual and audio clues is able to make use of the complementary information to improve the robustness and confidence of system, as well as interpretation disambiguation in real-life situations. Many



psychological studies have theoretically and empirically demonstrated the importance of the integration of information from multiple modalities (vocal and visual expression) to yield a coherent representation and inference of emotions [AR92]. As a result, an increasing amount of research effort is being put on this field.

The fusion just takes advantage of the diverse and complementary information, but does not solve problems raised in each source. Therefore, in order to gain a better performance, the issues left in each area unavoidably need to be solved. For example, the fundamental issue mentioned in subsection 7.2.1 for subtle facial expression is still unsolved.

Moreover, we notice that as the information sources are fused, a critical issue about data set appears. The most used dataset [GP05] is generated by asking the subjects to perform the corresponding emotional expressions in front of a camera, microphone, and/or even wearing some special devices. As reported in [GP05], there is no a publicly available database with bi-modal expressive face and body gesture. This situation leads to the algorithm lacking the generality and fair comparison, because of the controlled condition of generation training and test dataset.

Authentic affective expressions are difficult to collect because they are relatively rare, short lived, and filled with subtle context-based changes that make it difficult to elicit affective displays without influencing the results [ZPRH09]. Additionally, user study for ground truth of emotional expressions is very time-consuming and less reliable. Moreover, a large number of affective states are much more difficult (if possible at all) to elicit, like fear and stress. This state of affairs makes the analysis of spontaneous emotional expression a very difficult task. Until now, many databases of human emotional behavior do exist, such as Cohn-Kanade facial expression database [KTC00]. However, most of data contained in those databases currently lack labels. One of the reasons for this situation is that there is no standard metadata to identify the affective states in a video and the context in which this affect was shown.

In summary, two main issues also exist in this area. We list them as following:

- Reliable features for facial expression are still needed.
- An authentic and public database with multi-modal emotional events is needed.

### 7.2.3 Utilizing Eye Gaze Data

As one of the most important features of the human face, eyes and their movements is a useful cue in expressing the human's desires, focus, cognitive processes, and even emotional states. "The importance of eye movements to the individual's perception of and attention to the visual world is implicitly acknowledged as it is the method through which we gather the information necessary to negotiate our way through and identify the properties of the visual world" [HJ10]. The point or region of gaze usually represents the interesting part which mostly attracts the humans' attention. Therefore, it is useful to help people in understanding semantic information from video/image. For example, basic categorizations of eye movements include saccades and fixations. A fixation occurs when the gaze rests for some minimum amount of time on a small predefined area, usually within 2-5 degrees of central vision, usually for at least 80-100 ms [HJ10]. Saccades are fast, jump-like rotations of the eye between two fixated areas, bringing objects of interest into the central few degrees of the visual field. Smooth pursuit movements are a further categorization that describe the eye following a moving object [HJ10].

A number of efforts have been put on the mechanisms and dynamics of eye rotation. However, it is important to point out that the eye tracker does not provide the absolute gaze direction, but rather can only measure changes in gaze direction. Thus, some calibration procedure is required to help people to know precisely what a subject is looking at. Even those techniques that track features of the retina cannot provide exact gaze direction because there is no specific anatomical feature that marks the exact point where the visual axis meets the retina, if indeed there is such a single, stable point [HJ10]. An accurate and reliable calibration is essential for obtaining valid and repeatable eye movement data, and this can be a significant challenge for non-verbal subjects or those who have unstable gaze.

In addition, each existing method of eye tracking has their advantages and disadvantages, and the choice of an eye tracking system depends on considerations of cost and application. There is a trade-off between cost and sensitivity, with the most sensitive systems costing many tens of thousands of dollars and requiring considerable expertise to operate properly. Interpretation of the results still requires some level of expertise, however, because a misaligned or poorly calibrated system can produce wildly erroneous

data.

If we want to use the eye gaze tracking into affective analysis, some issues need to be solved. they are:

- An accurate and reliable calibration method is needed to obtain valid and repeatable eye movement data and better understand what the user is looking at.
- A simpler, low cost eye gaze tracking system is required.
- Characterization of the relation of eye gaze data (like pupillary dilation) with affect needs to be studied.

# Bibliography

- [AHA14] E. Acar, F. Hopfgartner, and S. Albayrak. Understanding affective content of music videos through learned representations. In *The International Conference on MultiMedia Modeling*, pages 303–314, 2014.
- [AR92] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.
- [ASS00] D. A. Atchison, G. Smith, and G. Smith. *Optics of the human eye*. Butterworth-Heinemann Oxford, 2000.
- [AV13] Timur R Almaev and Michel F Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 356–361, 2013.
- [Bab07] E.R. Babbie. *The practice of social research*. Wadsworth Pub Co, 2007.
- [BDDW08] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.
- [BG09] J.E. Burgess and J.B. Green. *YouTube: Online video and participatory culture*. Polity Press, 2009.
- [Bra94] M.M. Bradley. Emotional memory: A dimensional analysis. *Emotions: Essays on emotion theory*, pages 97–134, 1994.
- [BTA90] D. Bordwell, K. Thompson, and J. Ashton. *Film art: An introduction*. McGraw-Hill, 1990.
- [Bun90] C. Bundesen. A theory of visual attention. *Psychological review*, 97(4):523, 1990.

- [BY97] M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.
- [Can06] E.J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452, 2006.
- [CCK<sup>+</sup>07] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaïou, L. Malatesta, S. Asteriadis, and K. Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *Artificial intelligence and innovations: From theory to applications*, pages 375–388. 2007.
- [CGH00] I. Cohen, A. Garg, and T.S. Huang. Emotion recognition from facial expressions using multilevel hmm. In *Neural Information Processing Systems*, 2000.
- [CH67] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.
- [CH12] E. Cambria and A. Hussain. Sentic album: content-, concept-, and context-based online personal photo management system. *Cognitive Computation*, 4(4):477–496, 2012.
- [CJJ09] M Choi, J Jung, and J Jeon. No-reference image quality assessment using blur and noise. *International Journal of Computer Science and Engineering*, pages 76–80, 2009.
- [CKBW04] D.W. Cunningham, M. Kleiner, H.H. Bühlhoff, and C. Wallraven. The components of conversational facial expressions. In *Proceedings of ACM 1st Symposium on Applied perception in graphics and visualization*, pages 143–150, 2004.
- [CKP03] Z. Cernekova, C. Kotropoulos, and I. Pitas. Video shot segmentation using singular value decomposition. In *Proceedings of IEEE International Conference on Multimedia and Expo*, volume 2, pages 301–304, 2003.
- [CL08] W. T. Chu and C. H. Lin. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 829–832, 2008.
- [CL09] C. D. Cerosaletti and A. C. Loui. Measuring the perceived aesthetic quality of photographic images. In *IEEE International Workshop on Quality of Multimedia Experience*, pages 47–52, 2009.

- [CL11] C.C. Chang and C.J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- [CLT<sup>+</sup>13] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. S. Jin. Mutual information-based emotion recognition. In *The Era of Interactive Media*, pages 471–479. 2013.
- [CMK<sup>+</sup>06] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaïou, and K. Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the ACM international conference on Multimodal interfaces*, pages 146–154, 2006.
- [Col07] C.W. Color. *CeWe Photobook*. Photoworld, 2007.
- [CSG<sup>+</sup>03] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.
- [CT05] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- [CT06] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.
- [CT07] E. Candes and T. Tao. The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, 35(6):2313–2351, 2007.
- [CTL<sup>+</sup>M13] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, 31(2):175–185, 2013.
- [CW08] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.
- [CXF<sup>+</sup>03] L.Q. Chen, X. Xie, X. Fan, W.Y. Ma, H.J. Zhang, and H.Q. Zhou. A visual attention model for adapting images on small displays. *Multimedia systems*, 9(4):353–364, 2003.
- [CZLK98] JF Cohn, AJ Zlochower, JJ Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtledifferences in facial expression. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.

- [DBKL05] O.E. Demerdash, S. Bergler, L. Kosseim, and PK Langshaw. Generating Adaptive Multimedia Presentations Based on a Semiotic Framework. *Advances in Artificial Intelligence*, pages 417–421, 2005.
- [DL99] R. Dietz and A. Lang. Affective agents: Effects of agent affect on arousal, attention, liking and learning. In *International Cognitive Technology Conference*, volume 99, 1999.
- [Don00] D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.
- [EF78] P. Ekman and W. V. Friesen. Facial action coding system: A technique for the measurement of facial movement. palo alto, 1978.
- [EG97] P. Eisert and B. Girod. Facial expression analysis for model-based coding of video sequences. *Proceedings of Picture Coding Symposium*, pages 33–38, 1997.
- [Ekm92] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200, 1992.
- [Ekm93] P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384, 1993.
- [FM07] A Freitas-Magalhães. The psychology of emotions: The allure of human face. *University Fernando Pessoa Press, Oporto*, 2007.
- [GBC<sup>+</sup>00] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox. A semi-automatic approach to home video editing. In *Proceedings of ACM symposium on User interface software and technology*, pages 81–89, 2000.
- [GCL89] M.K. Greenwald, E.W. Cook, and P.J. Lang. Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of psychophysiology*, 1989.
- [GKG13] E. Guldogan, J. Kangas, and M. Gabbouj. Personalized representative image selection for shared photo albums. In *International Conference on Computer Applications Technology*, pages 1–4, 2013.
- [Gol99] E.B. Goldstein. *Sensation and perception*. Brooks/Cole Publishing, 1999.
- [GP05] H. Gunes and M. Piccardi. Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. *Affective Computing and Intelligent Interaction*, pages 102–111, 2005.

- [HDW04] R. Heishman, Z. Duric, and H. Wechsler. Using eye region biometrics to reveal affective and cognitive states. In *CVPR Workshop on Face Processing in Video*, 2004.
- [Hes65] E. H. Hess. Attitude and pupil size. *Scientific American*, 212:46–54, 1965.
- [Hev36] K. Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268, 1936.
- [HJ10] D.W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010.
- [HX05] A. Hanjalic and L.Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.
- [IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.
- [IPO11] P. Isola, A. Parikh, D. and Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *In Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.
- [IRT<sup>+</sup>05] S.V. Ioannou, A.T. Raouzaïou, V.A. Tzouvaras, T.P. Mailis, K.C. Karpouzis, and S.D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4):423–435, 2005.
- [Jam90] W. James. *The principles of psychology*. New York: Holt, 1890.
- [JJVS09] H. Joho, J. M. Jose, R. Valenti, and N. Sebe. Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 31, 2009.
- [JNL<sup>+</sup>05] A. Jaimes, T. Nagamine, J. Liu, K. Omura, and N. Sebe. Affective meeting video analysis. In *IEEE International Conference on Multimedia and Expo*, pages 1412–1415, 2005.
- [Kan03] H.B. Kang. Affective content detection using HMMs. In *Proceedings of the eleventh ACM international conference on Multimedia*, page 262, 2003.
- [KK85] A. Kuper and J. Kuper. *The social science encyclopedia*. Routledge/Thoemms Press, 1985.



- [KTC00] T. Kanade, Y. Tian, and J.F. Cohn. Comprehensive database for facial expression analysis. *Florida Geographer*, page 46, 2000.
- [KWB04] M. Kleiner, C. Wallraven, and H.H. Bülthoff. The MPI VideoLab-A system for high quality synchronous recording of video and audio from multiple viewpoints. *MPI-Technical Reports*, 123, 2004.
- [KWS98] S Kaiser, T Wehrle, and S Schmidt. Emotional episodes, facial expressions, and reported feelings in human-computer interactions. In *Proceedings of Xth Conference of the International Society for Research on Emotions*, 1998.
- [LAKG98] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.
- [LC09] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009.
- [LKCL00] J.J.J. Lien, T. Kanade, J.F. Cohn, and C.C. Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.
- [McK76] W.J. McKeachie. Psychology in america’s bicentennial year. *American Psychologist*, 31(12):819, 1976.
- [MLLK03] P. Mulhem, J. H. Lim, W. K. Leow, and M. Kankanhalli. Advances in digital home photo albums. *Multimedia Systems and Content-Based Image Retrieval*, pages 201–226, 2003.
- [Mye04] D. G. Myers. Theories of emotion. *Psychology: Seventh Edition, New York, NY: Worth Publishers*, 500, 2004.
- [MZ03] Y.F. Ma and H.J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, page 381, 2003.
- [NGP11a] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.
- [NGP11b] M. A. Nicolaou, H. Gunes, and M. Pantic. A multi-layer hybrid framework for dimensional emotion classification. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 933–936, 2011.

- [NT09] D. Needell and J.A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.
- [OLC<sup>+</sup>06] N. Ohare, H. Lee, S. Cooray, C. Gurrin, G. J. Jones, J. Malobabic, N.E. Oconnor, A.F. Smeaton, and B. Uscilowski. Mediassist: Using content-based analysis and context to manage personal photo collections. In *Image and video retrieval*, pages 529–532. 2006.
- [OM02] M. Ondaatje and W. Murch. *The conversations: Walter Murch and the art of editing film*. Knopf, 2002.
- [PC11] N. S. Pai and S. P. Chang. An embedded system for real-time facial expression recognition based on the extension theory. *Computers & Mathematics with Applications*, 61(8):2101–2106, 2011.
- [Pet09] M. Pettinelli. *The psychology of emotions, feelings and thoughts*. Retrieved from the Connexions Web site: <http://cnx.org/content/col10447/1.11/>, 2009.
- [Phi99] E Phillips. The classification of smile patterns. *Journal of the Canadian Dental Association*, 65:252–254, 1999.
- [Pic00] R.W. Picard. *Affective computing*. The MIT Press, 2000.
- [PK09] S. Park and D. Kim. Subtle facial expression recognition using motion magnification. *Pattern Recognition Letters*, 30(7):708–716, 2009.
- [PS03] T. Partala and V. Surakka. Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, 59(1):185–198, 2003.
- [PSLD12] J. A. Prado, C. Simplicio, N. F. Lori, and J. Dias. Visuo-auditory multimodal emotional structure to improve human-robot-interaction. *International Journal of Social Robotics*, 4(1):29–51, 2012.
- [PTR01] M. Pantic, M. Tomc, and L.J.M. Rothkrantz. A hybrid approach to mouth features detection. In *IEEE International Conference on System Man And Cybernetics*, volume 2, pages 1188–1193, 2001.
- [RBK96] HA Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 203–208, 1996.
- [RH97] Plutchik R and Conte H. *Circumplex models of personality and emotions*. American Psychological Association, 1997.

- [RHC99] Y. Rui, T.S. Huang, and S.F. Chang. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.
- [RM77] J.A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.
- [RSB10] M. Rabbath, P. Sandhaus, and S. Boll. Automatic creation of photo books from stories in social media. In *Proceedings of ACM SIGMM workshop on Social media*, pages 15–20, 2010.
- [Rus80] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.
- [SGM06] C. Shan, S. Gong, and P.W. McOwan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *Proc. BMVC*, volume 1, pages 297–306, 2006.
- [Sik97] T. Sikora. The mpeg-4 video standard verification model. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):19–31, 1997.
- [SKCP09] M. Soleymani, JJM Kierkels, G. Chanel, and T. Pun. A bayesian framework for video affective representation. In *Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009.
- [SLCH02] N. Sebe, M. S. Lew, A. Cohen, I. and Garg, and T. S. Huang. Emotion recognition using a cauchy naive bayes classifier. In *IEEE International Conference on Pattern Recognition*, volume 1, pages 17–20, 2002.
- [SM83] G. Salton and M.J. McGill. *Introduction to modern information retrieval*, volume 1. McGraw-Hill New York, 1983.
- [SNSP10] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172, 2010.
- [STD14] P. Suja, S. Tripathi, and J. Deepthy. Emotion recognition from facial expressions using frequency domain techniques. In *Advances in Signal Processing and Intelligent Recognition Systems*, pages 299–310. 2014.
- [SVE<sup>+</sup>12] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456, 2012.

- [SWB<sup>+</sup>06] M. Song, H. Wang, J. Bu, C. Chen, and Z. Liu. Subtle facial expression modeling with vector field decomposition. In *IEEE International Conference on Image Processing*, pages 2101–2104, 2006.
- [SY07] K. Sun and J. Yu. Video affective content representation and recognition using video affective tree and hidden Markov models. *Affective Computing and Intelligent Interaction*, pages 594–605, 2007.
- [SYHH09] K. Sun, J. Yu, Y. Huang, and X. Hu. An improved valence-arousal emotion space for video affective content representation and recognition. In *IEEE International Conference on Multimedia and Expo*, pages 566–569, 2009.
- [SZPR12] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 30(10):762–773, 2012.
- [SZZL02] Y. Sun, H. Zhang, L. Zhang, and M. Li. Myphotos: a system for home photo management and processing. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 81–82, 2002.
- [TCMK02] T. Tan, J. Chen, P. Mulhem, and M Kankanhalli. Smartalbum: a multi-modal photo annotation system. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 87–88, 2002.
- [TKC01] Y.I. Tian, T. Kanade, and JF Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.
- [TP91] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition*, pages 586–591, 1991.
- [Tsu99] K Tsubota. Blink of an eye. *Newsweek, Personal Essay*, 134:6, 1999.
- [TT05] J. Tao and T. Tan. Affective computing: A review. *Affective Computing and Intelligent Interaction*, pages 981–995, 2005.
- [TYA11] R.M.A. Teixeira, T. Yamasaki, and K. Aizawa. Determination of emotional content of video clips by low-level audiovisual features. *Multimedia Tools and Applications*, pages 1–29, 2011.
- [Van07] G.R. VandenBos. *APA Dictionary of Psychology*. American Psychological Association, 2007.
- [Wan05] M. H. Wang. Extension neural network-type 2 and its applications. *Neural Networks, IEEE Transactions on*, 16(6):1352–1361, 2005.

- [WFT<sup>+</sup>99] I.H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S.J. Cunningham. Weka: Practical machine learning tools and techniques with Java implementations. In *Proceedings of ICONIP/ANZIIS/ANNES'99 International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems*, volume 99, pages 192–196, 1999.
- [WMM<sup>+</sup>10] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.
- [WSZ00] L. Wenyin, Y. Sun, and H. Zhang. Mialbum-a system for home photo management using the semi-automatic image annotation approach. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 479–480, 2000.
- [WYG<sup>+</sup>09] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.
- [XK11] X. Xiang and M.S. Kankanhalli. Affect-based adaptive presentation of home videos. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 553–562, 2011.
- [XK12] X. Xiang and M. S. Kankanhalli. A multimodal approach for online estimation of subtle facial expression. In *Advances in Multimedia Information Processing*, pages 402–413. 2012.
- [XWH<sup>+</sup>12] M. Xu, J. Wang, X. He, J.S. Jin, S. Luo, and H. Lu. A three-level framework for affective content analysis and its case studies. *Multimedia Tools and Applications*, pages 1–23, 2012.
- [YBS06] M. Yeasin, B. Bulot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE transactions on multimedia*, 8(3), 2006.
- [YK02] W.Q. Yan and M.S. Kankanhalli. Detection and removal of lighting & shaking artifacts in home videos. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 107–116, 2002.
- [YLSL07] J. You, G. Liu, L. Sun, and H. Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):273–285, 2007.

- [YWH10] Z. L. Ying, Z. W. Wang, and M. W. Huang. Facial expression recognition based on fusion of sparse representation. *Advanced Intelligent Computing Theories and Applications*, pages 457–464, 2010.
- [YWMS07] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry. Feature selection in face recognition: A sparse representation perspective. *UC Berkeley Tech Report UCB/EECS-2007-99*, 2007.
- [YYGH09] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.
- [Zet12] H. Zettl. *Sight Sound and Motion*. Cengage Learning, 2012.
- [ZMM95] R. Zabih, J. Miller, and K. Mai. Feature-based algorithms for detecting and classifying scene breaks. Technical report, Cornell University, 1995.
- [ZPRH09] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
- [ZS06] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th annual ACM international conference on Multimedia*, page 824, 2006.
- [ZTH<sup>+</sup>10] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li. Utilizing affective analysis for efficient movie browsing. In *IEEE International Conference on Image Processing (ICIP)*, pages 1853–1856, 2010.
- [ZTL<sup>+</sup>04] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, and S. Levinson. Bimodal HCI-related affect recognition. In *Proceedings of the ACM international conference on Multimodal interfaces*, pages 137–143, 2004.