7-13-2012

Affect-based Modeling and its Application in Multimedia Analysis Problems

Abhishek Bhattacharya
Florida International University, abhat002@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

AFFECT-BASED MODELING AND ITS APPLICATION IN MULTIMEDIA ANALYSIS PROBLEMS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Abhishek Bhattacharya

2012

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Abhishek Bhattacharya, and entitled Affect-based Modeling and its Application in Multimedia Analysis Problems, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Peter J. Clarke

Deng Pan

Chen Liu

Christine Lisetti

Zhenyu Yang, Major Professor

Date of Defense: July 13, 2012

The dissertation of Abhishek Bhattacharya is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2012

DEDICATION

To my wife, Ipsita.

ACKNOWLEDGMENTS

I would like to thank the Department of Computing and Information Science for all the support rendered to complete this dissertation successfully. I would like to thank my major academic advisor Dr. Zhenyu Yang for his constant support throughout the graduate program.

I also take this opportunity to thank my other dissertation committee members - Dr. Peter Clarke, Dr. Deng Pan, Dr. Christine Lisetti, and Dr. Chen Liu for being very supportive and reviewing my dissertation at various stages. Many thanks to all the department faculty members for teaching the various courseworks which helped me build a solid foundation. I would also like to thank my PhD candidacy committee member Dr. Shu-Ching Chen for his time.

I am indebted to my wife, Ipsita for her love, encouragement, endless support, and everything throughout my entire journey.

ABSTRACT

AFFECT-BASED MODELING AND ITS APPLICATION IN MULTIMEDIA ANALYSIS PROBLEMS

by

Abhishek Bhattacharya

Florida International University, 2012

Miami, Florida

Dr. Zhenyu Yang, Major Professor

The multimedia domain is undergoing a rapid development phase with transition in audio, image, and video systems such as VoIP, Telepresence, Live/On-Demand Internet Streaming, SecondLife, and many more. In such a situation, the analysis of multimedia systems, like retrieval, quality evaluation, enhancement, summarization, and re-targeting applications, from various context is becoming critical. Current methods for solving the above-mentioned analysis problems do not consider the existence of humans and their affective characteristics in the design methodology. This contradicts the fact that most of the digital media is consumed only by the human end-users. We believe incorporating human feedback during the design and adaptation stage is key to the building process of multimedia systems. In this regard, we observe that affect is an important indicator of human perception and experience. This can be exploited in various ways for designing effective systems that will adapt more closely to the human response.

We advocate an affect-based modeling approach for solving multimedia analysis problems by exploring new directions. In this dissertation, we select two representative multimedia analysis problems, e.g. Quality-of-Experience (QoE) evaluation and Image Enhancement in order to derive solutions based on affect-based modeling techniques. We formulate specific hypothesis for them by correlating system

parameters to user's affective response, and investigate their roles under varying conditions for each respective scenario. We conducted extensive user studies based on human-to-human interaction through an audio conferencing system. We also conducted user studies based on affective enhancement of images and evaluated the effectiveness of our proposed approaches. Moving forward, multimedia systems will become more media-rich, interactive, and sophisticated and therefore effective solutions for quality, retrieval, and enhancement will be more challenging. Our work thus represents an important step towards the application of affect-based modeling techniques for the future generation of multimedia systems.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Multimedia applications and systems are now experiencing a tremendous growth rate with their gradual penetration in our everyday lives, from digital cameras, professional high-end Single Lens Reflex (SLR), camcorders, mobile video terminals, Digital Video Disc (DVD) players, down-loadable games, radio stations on web, and a plethora of other different services. We use some form of multimedia data in the daily fabric of our digital life with the ever-increasing diffusion of computing in human society. With this advancement, there are increasing challenges being faced and a paradigm shift is required from our previous methods related to design, interaction, analysis, communication, and organization of multimedia data.

Any multimedia content analysis problem can be related to perception with the following three components: (1) data related to environment, (2) medium that transmits the data to the perceiver, and (3) the perceiver. Multimedia data such as visual (images/videos), aural (speech/music) and other types of sensory data are captured for a specific event that unfolds over time. We can observe that each sensor captures only one type of physical attribute from its perspective, such as a camera/camcorder records visual signal, a microphone records aural sensory data, and likewise. The multimedia stream is then formed by combining the correlated and complimentary information from individual streams to provide a more holistic information and experience in comparison to using any one medium. We believe that the context or perspective is also important along with the content in understanding and analyzing the user experience represented by the data. Present-day multimedia systems and applications are armed with a lot of contextual information sources, such as motion sensors in mobile phones, GPS location information, social networks, and many more. Each of these sources can be exploited to extract mean-

ingful information, which can help the design methodologies. Our guiding principle revolves around the human end-user which is supposed to be the most important component in a multimedia system. We consider human feedback signals to be a critical source of contextual information, which can be effectively exploited. We incorporate explicit/implicit human feedback signals in our proposed framework since they are known to relate closely to perception [AKJ09, SGL11], which is one of our essential design objective, as discussed next.

This principle of applying human feedback signals is known as human-centric computing (HCC) approaches, which are recently gaining importance in the multimedia research community with a paradigm shift from system/network-centered to user-centered design and evaluation methodologies [JSGP06]. HCC-based methodologies are designed to address the semantic gap between the raw data representation and the higher-level concepts, and are supposed to be more close and responsive to user perception. HCC approach to multimedia systems essentially considers the understanding and interpretation of multimedia signals by humans in the feature, cognitive, affective levels, and how humans interact naturally. Inevitably, this means an interdisciplinary approach covering neuroscience, psychology, cognitive science, and others, becomes essential for incoporating the knowledge in those domains to develop computational frameworks that integrate different media. As an illustration, if we view computing in a large space with a human in the center, then we can identify some applications that are closer to the human and some that are farther away. For example, packet routing is very important in communications, but is more distant from a human, than for instance, human computer interaction (HCI). The goal is thus to drive the computers–physically, conceptually, and functionally–closer to humans. The three main areas where the HCC approach for designing multimedia systems can benefit are identified as media production, analysis, and

interaction [JSGP06]. In this dissertation, we investigate the areas of interaction and analysis, and derive HCC approach based solutions in the context of affect or emotion of the user. We sense that affect or emotion will be an important candidate for user-centric design since it is closely related to human perception. Thereby, we adopt an affect-based modeling paradigm for solving the multimedia analysis problems.

Affect is an extensively studied area in psychology [Dam94, Rus03, Dar05], which is starting to get recognition in computer science with the advent of Affective Computing [Pic97]. It deals with the automated analysis of human affective behavior, which has attracted attention from researchers, due to its multidisciplinary nature spanning psychology, computer science, linguistics, neuroscience, and various other branches. Emotion is closely related to decision-making and thus plays a significant role in the action/perception of human beings as shown in research by psychologists and neuroscientists [Dam94]. The initial research studies in affect started with the problem of emotion detection and its role in possible application scenarios. Extensive studies have been performed for automated emotion recognition in the context of human-computer interaction by exploring multiple input modalities such as facial expression, speech, body gestures, physiological signals, posture, and neuroimaging for extracting implicit feedback using signal processing, linguistic analysis, text processing and various other techniques [ZPRH09, CD10]. Based on the success of emotion detection, many affect-aware applications were proposed, which found acceptance due to their human-centric attributes, that can adapt to the changing emotion of the user. Some possible application areas are found in user interface, gaming, mental health, learning technologies, customer services, intelligent automobile and entertainment industry [ZPRH09, VPBP08, KP05, Pel05]. The success of affect in emotion detection fueled interest to explore its role in multimedia applications.

The main challenge faced by the multimedia analysis problems emanate from the "semantic gap," which is the semantic difference between an user's representation or expectation from the system and the internal representation of multimedia data item in a collection. The usage of affect-based feedback channel to bridge the semantic gap found reasonable success in various applications, such as retrieval [MH10], summarisation [JJVS09] and search [AJG08].

Based on the above direction, we explore two interesting paths in this dissertation and derive affect-based models to produce solutions which will be closer to user experience. First, we examine the problem of QoE evaluation in audio communication or Voice-over-IP (VoIP) system, where a human-human interaction is mediated through a communication channel. Traditionally, the evaluation frameworks for voice communication systems were designed from the service provider perspective through objective assessment methodologies with the help of QoS-based attributes [CHHL06, MTK02, WCHL09, SW07, GMSL10]. QoS-based approaches for communication systems generally involve tracking certain network properties, such as delay, bandwidth and loss, followed by predicting the result using analytical models [MTK02, GMSL10]. More recently, QoE [Ebr09, Moe08, WAR$^+$09] concept has been developed to address the failure of capturing user's experience from QoS-based evaluation frameworks. The sematic gap between QoS and QoE is still ambiguous. A deterministic/non-deterministic fuzzy mapping between them to reveal user's perception of quality against fluctuating network/system/cognitive/behavioral and various other artifacts will be useful in deriving QoE. A prospective approach to model QoE-centric evaluation is a multi-dimensional construct of user perceptions and behaviors, where each dimension has a subjective or objective influence on the user experience. We hypothesize that QoE or perception of quality by the user is an implicit decision-making phenomenon related to human cognitive state, that is derived

from the sensory channels and is highly correlated to user's affective behavior. According to the best of our knowledge, the study of affective behavior in the context of remote human-human interaction integrated through a network channel and its dynamics with QoS (delay, loss-rate, and bandwidth) fluctuations is a relatively new and unexplored area. We investigate various affective cues from users interacting in a VoIP system which are extracted in an implicit manner and their association with the evaluation of QoE. We utilize three different types of information sources in our framework as follows: (i) Acoustic: pitch-related features, formant frequencies, timing features, and likewise; (ii) Lexical: modeled by the concept of salience, which measures the amount of mutual information between a specific word and a quality level; (iii) Discourse: modeled by repetition, which carry relevant cues for user's affective response since it was found to be an important indicator for trouble in human communication [BFH$^+$03] However, we build our ground-truth data from user feedback response and then perform training, classification and pattern recognition in a supervised learning approach to derive the relationship the between user's response and the content i.e., voice data features. We present detailed experimental methodology with real user studies and discuss outcome of the results.

Next, we analyze the problem of image enhancement where the objective is to enhance the image in a way, such that it increases the affective appeal of the photograph. Some earlier work in psychology literature explored a strong relationship between color and emotions [OLWW04a, OLWW04b, OLWW04c, VM94]. According to the best of our knowledge, the investigation of a computational framework for affective image enhancement is an unexplored area. Our goal in this topic is to learn the underlying adjustment rules associated with the emotional properties in images from a set of training examples. Given a pair of images before and after adjustment, we would like to discover the underlying mathematical relationships

optimally connecting the color-contrast properties between them. Some of the key issues that arise in order to achieve our goal are: (1) How do we choose a set of image parameters which will be used for enhancing the emotional appeal of images? (2) How do we capture the notion of enhancement of images in a mathematical fashion to understand the implicit relationship between an original image and its adjusted version? (3) Given an arbitrary image, how do we derive the enhancement operations that can be applied to any unseen/new image to enhance its emotional appeal? To solve these problems, as a first step we collected ground-truth data and constructed a training database consisting of example images and their different enhanced versions. Then we implemented a simple and effective web-based user-interface for this purpose and sent invitations for participation to a group of subjects. We asked a set of human participants to mark their preferences in the user-interface. To answer (1) in the above discussion, we selected a set of contrast and color properties, which were found to have a high influence on emotional impact of images [JDF+11]. We employed machine learning techniques to understand the implicit relationship between the original image and its adjusted version to address (2) and derived statistical solutions to develop an enhancement function in a mathematical form to address (3). We conducted user studies to evaluate the effectiveness of our approach.

CHAPTER 2

RESEARCH STATEMENT

In this chapter, we discuss the problem scenario with a formal definition in Section 2.1 followed by some of the challenges of designing a QoE assessment and image enhancement framework in Section 2.2. In Section 2.3, we discuss the possible impact of the solutions in a broader perspective and conclude this chapter with a discussion on dissertation contributions in Section 2.4.

2.1 Problem Scenario

Let us first illustrate the quality evaluation problem in a general context of a multimedia communication system with human-to-human interaction mediated by a network channel (such as Internet) for a VoIP system. The generalized block diagram of a No-Reference (NR) quality assessment system for multimedia signal is shown in Figure 2.1. The figure is based on NR quality assessment model, but can be easily modified to depict Full-reference (FR) and Reduced-Reference (RR) cases (refer to Section 3.2.2 for a detailed discussion). NR quality model is more desirable since the other options of FR, RR do not support many practical real-time applications such as VoIP and streaming video through the Internet. The input signal is treated as the reference which passes through the various stages of coder, tranmission channel and decoder, where each of these stages may distort the reference to a varying extent. The signal is then finally passed on to the measurement system for quality prediction. The formal definition of the quality assessment problem can be depicted as follows:

Given a multimedia system 'S', with 'C' as the set of different sensory channels of information flow such as C=$\{C_a, C_v, C_t, ..\}$ and $|C| = c$, where C_a is audio, C_v is

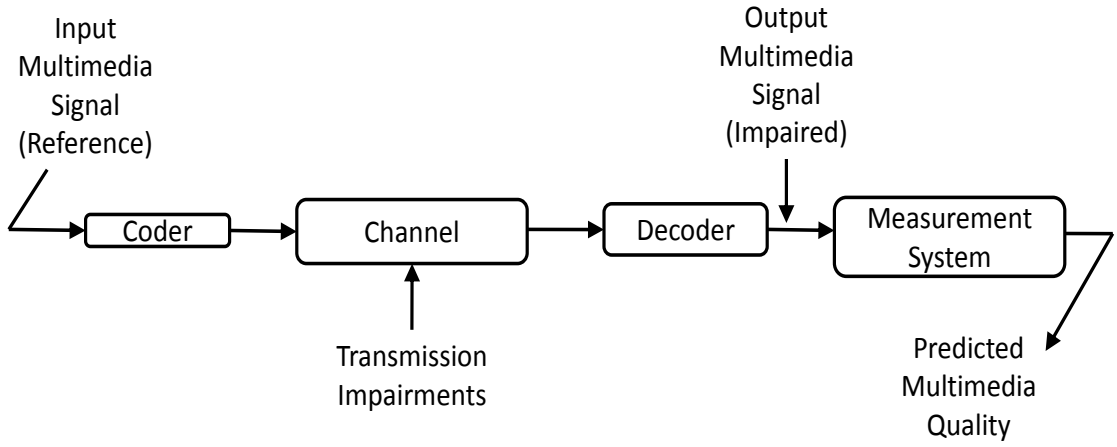

Figure 2.1: Block diagram of a NR multimedia quality assessment system.

video and C_t is tactile. A set of source codec Y=$\{Y_a, Y_v, Y_t, ...\}$, where Y_a is the audio source codec that can be G.729, iLBC, MP3, characterised by a sample rate of y_a^r kHz, bit-rate of y_a^b Kbit/s, number of bits/sample as y_a^nbits, and latency of y_a^l ms. Y_v is the video source codec that can be H.264/AVC, MPEG4, or WMV, and characterised by a bit-rate of y_v^b Kbit/s. A mediating network channel N=$\{N_w, N_l, N_m, ...\}$, where N_w is wired network channel with IP through Internet, LAN, or WAN, N_l is the wireless network channel, N_m is the mobile network channel, or other communication channels. The set of input (reference) multimedia sample is I=$\{s_1, s_2, .., s_n\}$ with each element s_x=$\{a_1, v_1, t_1, ..\}$, where a_1 is the 1st audio sample, v_1 is the 1st video sample, and likewise. The set of output (distorted) multimedia sample is O=$\{s'_1, s'_2, ...s'_n\}$ with each element s'_x=$\{a'_1, v'_1, t'_1, ..\}$, where O is a certain distortion function of I formed while travelling through the different components as shown in Figure 2.1. The total number of input/output multimedia samples is 'n' ($|I| = |O| = n$) and the set of subjective quality assessment of O by humans is Q=$\{q_1, q_2, .., q_n\}$. The final problem is to design a quality assessment model 'M',

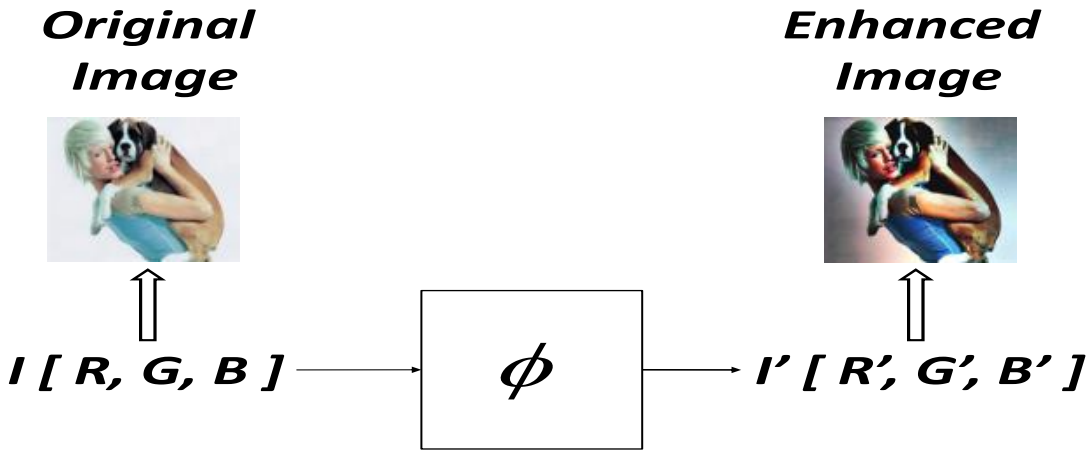

Figure 2.2: Block diagram of a basic image enhancement framework. The input or original image is I represented by RGB matrix and the output enhanced image is I' represented by $R'G'B'$ matrix. The transfer function is ϕ.

with the predicted quality set as Q'=$\{q'_1, q'_2, .., q'_n\}$, and the following optimization objective is to be satisfied:

$$Minimize \quad \sum_{i=1}^{n} \|q_i - q'_i\|$$

Next, let us examine an image enhancement problem scenario as follows:

Given an input image I represented by a RGB matrix, find the transfer function ϕ, that transforms I to the enhanced image I' represented by $R'G'B'$ matrix and formulated as follows:

$$\phi\Big(I\langle RGB\rangle \rightarrow I'\langle R'G'B'\rangle\Big) \quad s.t. \;\; e(I) > e(I')$$

where $e(I)$ is the emotional quotient of image I and $e(I')$ is the emotional quotient of image I'. The definition of emotional quotient (e) is derived from the computational model of affect as described in a later section. The block diagram of a basic image enhancement system is shown in Figure 2.2.

2.2 Challenges

There are a number of challenges which are required to be addressed for developing an effective quality assessment methodology, and the following discussion enumerates them. Let us start with the challenges generated from a QoE evaluation perspective. First, the general trend of QoE assessment in real-time interactive multimedia systems (such as VoIP, tele-conferencing, online games, and collaborative systems) is to follow the naive approach of modeling it with QoS parameters, since the interactivity is known to have a correlation with QoS metrics such as delay, delay jitters, network loss [CHW+06]. Another approach is to use application specific metrics such as conversation patterns, talk burst length in VoIP [CHHL06], playing times, player scores, ranking in online games [CHW+06], collaboration efficiency in collaborative systems [SKR+08]. Though the above approaches provide relative success, but these techniques do not have a high correlation with QoE [WAR+09]. Thus, it is challenging to explore the inter-relationship between QoS–QoE and the mutual influence between the different sensory modalities (i.e. audio or video is more important) with its implication on overall QoE.

Second, the multimedia systems are invoking the importance of QoE in various applications which are becoming increasingly rich with media content, following the advent of 3D technology. 3D movies and televisions have already entered the commercial market where the stereoscopic vision have been found to accompany visual strain and discomfort during extended viewing, which have a negative effect on quality perception [MB11a]. The challenging problem of depth perception in Human Visual System (HVS) will help to gain more insights for QoE evaluation, with the help of extensive psychophysical studies. Considerable challenges are also associated

with the development of QoE assessment algorithms for mobile devices such as smartphones, tablets, and other handheld devices, which are gaining popularity.

Third, media aesthetics is a measure of the perceived beauty of a stimulus and have been found to have a high correlation with human quality perception [MB11a]. Thus, it can be noted that the content of the stimulus will play an important role in the subjective quality ratings and adds a bias to the QoE assessment. Aesthetics and personal preferences of the stimuli involve a complex influence on the subjective user experience and are challenging to extract or filter out in order to estimate the exact QoE. [MB11a] suggested a naive approach to explicitly allow subjective rating from users on disjoint scales of content and quality which will help to measure the primary variable being assessed. The overlap of QoE evaulation and content assessment is an important research challenge.

Next, let us discuss about the challenges that are invoked from an image enhancement problem perspective: First, what are the set of image parameters to choose for defining the emotional appeal of images in a data-driven methodology? This is an open problem since the subject of human emotion is highly subjective and there are no exact answers. There are no defined rules that can explain the procedure to enhance any arbitrary image since each photographer uses his/her own expertise for adjustment. In spite of this, there are certain patterns that stand out with respect to emotional appeal in photography. For example, in nature photography, it can be observed that most professionals share common techniques or rules of thumb in their choices of colors, tonality, lighting, focus, vantage point and composition. Moreover, in scenery photography, they also use high-saturation and high-contrast colors rather than capturing the true colors of the real world. The purer the primary colors–red (sunset, flowers), green(trees, grass), and blue (sky)–the more striking is the scenery to the viewers.

Second, how do we capture the notion of enhancement in a mathematical fashion to understand the implicit relationship between an original image and its adjusted version? This is an important step if we need to derive an image enhancement function in a data-driven fashion.

Third, how to derive the enhancement operations that can be applied to the unseen/new image to enhance its emotional appeal? Assuming, that we have found a perfect enhancement function, the next question is, how to apply the function to an unseen or new image. Applying the same function to every image is a naive option already present in auto-enhance tools, and thus we need some advanced techniques to address this issue.

2.3 Impact

The impact of a quality assessment task can be justified by its role in the various aspects of system development. We enumerate them as follows:

- They can be primarily used for monitoring and managing the system quality in various control stages of the distributed multimedia systems such as VoIP, Internet video streaming systems, virtual environments and online gaming systems. In the acquisition stage, the system can use the quality metric to monitor and adjust itself automatically by varying bit-rate, in order to obtain the best quality audio and video data. In the distribution stage, a network video server (in server-client architecture) or supplier peers (in peer-to-peer architecture) can examine the quality of the digital video transmitted on the network and automatically adjust the streaming control according to the requirement [HCH09].

- They can be embedded into an audio and video processing system to optimize the algorithms and the parameter settings of the encoding and decoding phases. For example, in a audio-visual communication system, a quality metric can help in the optimal design of the pre-filtering and bit assignment algorithms at the audio/video encoder. It can also help in the optimal reconstruction, error concealment and postfiltering algorithms at the audio/video decoder [HHCW10], [TYHT08].
- They can be employed to benchmark audio and video processing systems/ algorithms. If multiple video processing systems are available for a specific task, then the quality assessment can help in determining which one of them provides the best perceptual quality.

The impact of an affective image enhancement application can be explained as follows:

- Computational frameworks for affective enhancement of images can open up a new space of opportunity, where images from low-end devices such as mobile phones can be enhanced to match the emotional appeal of professional cameras as a low-cost alternative.
- This technology can be incorporated in the processing pipeline of a digital camera as a preprocessing step, which can give innovative recommendations for views with higher emotional appeal, thereby helping an amateur to acquire the skills of a professional photographer.

2.4 Contributions

The dissertation contributions are listed as follows:

1. We provide a new affect-based methodology for QoE evaluation in voice communication. Different from previous approaches, we propose to assess quality directly from the user's affective response. Therefore, our method has the advantage of deriving subjective QoE measures in an implicit and non-intrusive manner.

2. The experimental results indicate very promising prospect of our approach of affect-based QoE evaluation. We evaluated the effectiveness of this approach by using classification techniques based on Support Vector Machines (SVM) and k-Nearest Neighbor (kNN) to discriminate different quality perceptions.

3. The accumulated evidence supports our initial hypothesis of exploiting affective response as a predictor of subjective experience of quality due to its correlation with human cognitive perception. Our best experimental performance achieved a prediction accuracy of 67.9%.

4. Regarding boarder impacts, as the communication systems become more media-rich and interactive (e.g., spatial audio, 3D/immersive space), measuring QoE via indirect methods will become more challenging [Ebr09]. Therefore, our work represents an important step towards the understanding of QoE for the future generation of communication systems.

5. As per our knowledge, affective enhancement of images is one of the first attempt to derive a computational framework for enhancing the emotional impact of images.

6. We collect ground-truth data for affective enhancement of images from human participants which is a valuable resource for understanding the underlying relationship between them.

7. We employ a data-driven systematic framework to learn models from training data, and derive generalized enhancement functions for arbitrary/unseen images using machine learning and statistical techniques.

8. We derive an objective methodology for evaluating the emotional enhancement of images using a color mood space model and used it to test our approach.

CHAPTER 3

LITERATURE REVIEW

We start this capter by presenting a literature review on Affective Computing in Section 3.1. Next, we present a detailed literature review of the various quality assessment methodologies from the viewpoint of audio in Section 3.2.1, video in Section 3.2.2, and audio-visual (multimodal) applications in Section 3.2.3. This is followed by Section 3.3, which reviews related work on various image enhacement techniques covered in the literature. Finally, we present arguments concerning the various advantages and disadvantages of the different techniques in Section 3.4 to end this chapter.

3.1 Affective Computing

We start this section with an introduction to affective computing in Section 3.1.1, followed by a discussion on emotional theories in Section 3.1.2. Then, we look into the various information channels of emotion in Section 3.1.3, and end this section with a description of the various affect-based applications present in the literature in Section 3.1.4.

3.1.1 Introduction

The importance of emotions in human cognition and perception are established through various evidences from psychological and neurological studies [Pic01, PLC06, SGL11]. Emotions play an essential role in social interactions [Sch03, RBFD03, SGP$^+$05], facilitate rational decision making and perception [Dam94]. Affective Computing involves the process of recognizing, expressing, modeling, communicating and responding to emotions [Pic03]. Affective Computing was incorporated in

the field of Human-Computer Interaction (HCI), which can adapt more closely by recognizing and responding appropriately to human emotions [CDCT+01]. Lately, affect-based applications are spreading to other domains such as multimedia information retrieval (MIR) [AKJ09], video summarization [JJVS09], understanding video/image semantics [HX05, KSK+10], image retrieval [MH10], visualization [ZHJ+10], video classification [HDL+10] and video browsing [ZTH+09].

3.1.2 Theory of Emotion

Emotion theory is at the intersection of multiple disciplines and mainly based on cognitive psychology, physiology, philosophy, economics, engineering and computer science. Specifically, psychology has the longest history of emotion research and has developed many framework for the theory of emotions. There are predominantly two approaches in the classical theory of emotions: First, the discrete emotion theory classifies the emotion in a number of categories usually between seven to ten core emotions and each is denoted by a descriptive word. Different theories propose various categories but the most well known core emotions are happy, surprise, sad, anger, disgust, contempt and fear which are universally displayed and recognised [Dar05, Ekm92, Ekm99a]. The argument in favor of this theory is that, these specific core emotions are biologically determined emotional responses whose expression and recognition are fundamentally the same irrespective of ethnic or cultural differences. Second, the dimensional theory of emotion argue that emotion should be conceptualized as points in a continuous (typically two or three) dimensional space [Rus03, MR74, Bar06, WT85]. The model gained support from other research, which revealed that people tend to perceive all kind of external stimuli in terms of valence and activation [Sch02]. Russell [RM77] proposed the use of independent bipolar dimensions of pleasure-displeasure, arousal and dominance-submissiveness in

support of the dimensional theory of emotion. Valence represents the pleasantness-unpleasantness in a positive-negative biploar axis and arousal indicates the intensity of the emotion in terms of activated-deactivated in a positive-negative bipolar axis of the other dimension.

3.1.3 Information Channels of Emotion

Human affect is manifested through various sensory channels which are rich information sources such as autonomic nervous system, facial, vocal, body gestures and electrodermal response.

Physiological signal processing for emotion detection involve the popular features such as galvanic skin response, electrocardiogram (ECG), electroencephalogram (EEG), skin temperature, heat flux and blood volume pulse. Emotional arousal are found to be detected by scanning brain activity, pulse rate, blood pressure or skin conductance [CD10]. The procedures for collecting physiological measures vary between a simple sensor on a finger (for monitoring pulse rate and skin conductance) to more invasive sensors such as ECG and EEG.

Facial Expressions are a rich source of information for affect as previous research indicates that facial cues (smiles, chuckles, smirks, frowns and likewise) are important indicators of social interactions [RBFD03], which help to determine the focus of attention [PR00]. Ekman developed a formulation known as Facial Action Coding System (FACS), which essentially encodes the facial feature points with respect to their respective emotion categories. It used an unobtrusive technique by measuring muscular movements in terms of distance between feature points on the face (also known as action units) [RE93, Ekm99b].

Speech Analysis for emotion is a well studied area where the objective is to derive certain features, which can effectively capture the various human emotions. Though

there are many proposals in the literature which address this problem in different context, but the most common feature set is found to be pitch, intensity, speech rate, pitch contour and phonetic features [ZPRH09, CDCT+01, LN05, BFH+03, Sch03]. Some of the systems that analyse speech for detecting emotion are described in [SCGH05, SS01, SBCL04]. These systems achieve an accuracy rate of 72%-85% when detecting one or more basic emotions from noise-free audio-visual input.

Body Gesture Recognition tends to associate certain body movements to specific emotions [BC98, Mei05, Wal98]. A fused methodology for bi-modal emotion recognition by considering facial expressions and body gesture is reported in [GP07]. They provided a list of expressive gestures and their correlated emotion categories, whereby it is observed that hand gesture is the most common and natural interactive media. The tracking aparatus can be glove-based (intrusive) or vision-based (nonintrusive).

Eye Tracking can be utilized to record the gaze direction and eye fixation, which are found to be strong indicators of visual attention [SPS+05, KSK+10]. Even though eye-tracking related features are not directly associated to the affective state of the user, they remain good indicators of attention that can help to determine potential sources of emotional stimuli and identify their importance to the observer. The systems required for eye-tracking are categorised into wearable, non-wearable, infrared-based and appearence-based.

User-feedback procedure is explicit in nature and is based on the assumption that human participants are willing to recognise and report their emotions. There are two major methods: (1) the discrete emotion approach and (2) the dimensional approach. The discrete emotion approach relies on the categorisation that is reflected in the classification of the semantics for emotions in natural language. This approach has the following disadvantages: (i) the possibility that one or several feedback

alternatives may bias the user to choose them, (ii) the user may wish to refer to a category that is not provided in the list, (iii) some of the emotion labels may be unfamiliar to the user. The dimensional approach is based on the pleasure-arousal-dominance model of emotion as described earlier where an user can simply report his/her emotional experience by indicating coordinates in a three-diemnsional space with a scale of 0-9 for each dimension in the general case. Due to high correlation between pleasure and dominance dimensions, in practice only two dimensions are applied, thus forming a two-dimensional valence-arousal space.

3.1.4 Affect-based Applications

Affective Computing based applications found a natural place in Human-Computer Interaction (HCI) scenarios, where the system is made to adapt more closely in response to human emotion. Affective interfaces which have the ability to detect subtleties of user's emotional behavior and to initiate interactions based on this information was proposed in [LN02]. The multimodal system of [DGH$^+$02] applied a model of embodied cognition that can be seen as a detailed mapping between the user's affective states and the type of interface adaptations. The tool proposed in [MP06], is capable of learning and analyzing the user's context-dependent behavioral pattern from multi-sensory data and adapting the interface accordingly. The Automated Learning Companion in [KBP07] combines information from cameras, a sensing chair and mouse, wireless skin sensor and task state to detect frustration in order to predict when the user needs help. Affect-based systems are also found to be effective in applications for customer services, call centers, intelligent automobile systems, game and entertainment industry. Lately, there is a lot of interest in applying affective techniques for multimedia applications with the aim of bridging the well-know "semantic-gap" present between the lower level data and

the higher level semantic concepts. Affective feedback as implicit source of evidence was proposed in [AKJ09], to evaluate the topical relevance of information items in multimedia search systems. The authors also explored similar techniques for web search systems [AJG08], and the effetiveness of personalized affective models for predicting topical relevance [AAJ10]. Affective video summarisation was proposed in [JJVS09], where the user's facial expressions were analysed to infer personalised affective scenes from a video that can be tailored to individual preference. Affective modeling of videos with the help of content analysis was explored in [HX05], and a computational framework was developed by considering the two-dimensional theory of valence-arousal. Affect-based image retrieval was demonstrated in [MH10], by extracting various low-level features such as color, texture, composition and content to represent the emotional content of an image and developing classification techniques by leveraging the standard affective image database known as International Affective Picture System (IAPS). An affective recommender system was illustrated in [TBK10], based on user's emotional response with the underlying assumption that affective parameters are more closely related to the user's experience than generic metadata (e.g. genre) and are thus more suitable for separating the relevant items from the non-relevant. Affect-based presentation of home videos was exhibited in [XK11], for automatically creating video presentations by considering emotional tone, other content related features and social networks for deriving local and global main characters based on face detection.

3.2 Media Quality Assessment

We classify media quality based on the three most popular forms i.e., audio, visual and audio-visual categories, which are discussed in the following sections.

3.2.1 Audio Quality Assessment

We classify available quality evaluation methods of voice communcation in three groups: (*a*) *user feedback*, (*b*) *QoS-based estimation*, and (*c*) *media quality analysis*.

User Feedback

This group of methods obtain explicit input from the user for quality measurement. For example, in the popular format of *Mean Opinion Score* (MOS) [itu96], users are asked to complete a questionnaire based on a 1-to-5 Likert scale. It is a simple method that provides subjective measure of user's perception. The main disadvantage is its intrusiveness [AJG08]. The laboratory settings are often not transparent to the participants, which destroy the eco-psychological validity of a naturalistic study. Thus, the user feedback may not easily elicit spontaneous expressive behavior. Another disadvantage is the issue of scaling quality with numbers [KMK99, WS98]. To alleviate these problems, an interesting idea of OneClick [CTX09] was proposed where the user only needs to click the mouse whenever he/she feels dissatisfied with the quality. Compared to traditional user feedback, it is less intrusive: the user task is reduced from a multiple-choice decision to a dichotomous one and the test can be performed "during" user interaction instead of "after". However, OneClick still requires direct user attention which poses cognitive overhead. For more interactive systems like gaming or 3D immersion, this technique implies that the user may have to pause his/her ongoing activity from time to time.

QoS-based Estimation

A good amount of research has been done in this area [CHHL06, MTK02, WCHL09, SW07, GMSL10], where QoE is modeled with the underlying QoS parameters. QoS-

based methods are implicit and non-intrusive which makes them an appealing choice. However, these methods are essentially objective approximation of QoE due to lack of user engagement. Accordingly, they cannot cover all QoE dimensions that may effect user perception and experience [CTX09], and the discrepancy among the user population tend to be ignored [CWCL09]. For example, users have different sensitivity to delays under varying conversational dynamics (e.g., various talk/silent spurt duration) [SW07]. It is hard to accommodate such features in QoS-based estimation. The correlation between QoS and QoE also becomes more intractable for advanced communication systems, which greatly complicate the modeling process [WAR+09]. Moreover, there are non-trivial technical details regarding its deployment in the field such as messaging overhead, traffic detection [BS06, BMM+07] and the buffer masking effect on the interaction between QoE and delay/loss [WCHL09]. A crowdsourceable evaluation framework was proposed in [CWCL09], where the experimental process of QoE evaluation is outsourced to the Internet crowd thereby gaining a rich set of rated samples at a lower economic cost with a diverse set of participants alongwith their subjective wisdom.

Media Quality Analysis

Media quality analysis methods assess the quality by measuring the distortion of the signal based on certain analysis models like signal-to-noise ratio (SNR). More sophisticated ones attempt to incorporate human auditory perception such as Enhanced Bark Spectral Distortion (EBSD) and Perceptual Speech Quality Measure (PSQM) [FC06]. Similar to QoS-based estimation, these methods do not require explicit user input. The state-of-the-art standard in this category is ITU-T P.862 [pes01], also known as the Perceptual Evaluation of Speech Quality (PESQ). The drawback of PESQ is that it is double-ended: the algorithm requires both the

original and the degraded signals to compute the quality difference [RBK+06]. Further, it fails to consider factors such as various listening levels, sidetone/talk echo, and conversational delay/interaction [CTX09, CHHL06].

We argue that a suitable QoE evaluation method in voice communication should capture the subjective measures from the user in a non-intrusive manner. It is clear that none of the existing methods comprise of both characteristics. The contribution of this article is the proposition of a new affect-based approach which opens such possibility.

3.2.2 Video Quality Estimation

The user feedback and QoS-based methodologies are similar to the above audio techniques which can be used likewise, since they are not dependent on the content. With regards to media content, video quality assessment can be extrapolated from image quality techniques on a frame-by-frame basis and applying some operator such as Minkowski summation of the frame-level scores. Objective algorithms based on media quality analysis are generally categorized as follows: (a) full-reference (FR), (b) reduced-reference (RR), and (c) no-reference (NR).

FR algorithms are double-ended since it requires both the original/reference and distorted signals for comparing them and predict a quality score which signifies the subjective assessment of the distorted stimulus. The ITU-T J.247 [itu08] recommedation for perceptual video quality transmitted over error-prone channels with coding impairments and transmission errors take 4 spatial and 2 temporal distortion indicators. The most widely used objective video quality metric is Peak Signal-to-Noise Ratio (PSNR) which does not have a high correlation with the perceptual quality since it does not take the human visual system (HVS) into consideration. Perception-based quality assessment models exploit the HVS using pyscho-visual

processing of the human vision and perception by integrating certain typical factors such as contrast and orientation sensitivity, frequency selection, spatial-temporal pattern masking and color perception [YRH+10a]. These factors are derived from several psychovisual experiments to study the human perception with varying spatial and temporal distorted visual stimulus. Some of the techniques in psychovisual based quality assessment are MPQM [vdBLV96], PDM [Win99], DVQ [WHM01], and wavelet-based metric [MHS06]. Psychovisual methods generally derive good performance since they mimic the HVS and the human perception but the drawback of these techniques are its computational complexity. They are thereby unable to process in real-time which is a critical requirement in many applications. To counter this problem, simplified metrics (based primarily on image content and distortion analysis due to coding and transmission errors while psychophysical effects are considered with a lesser extent) are developed which are mathematical-based algorithms. Some of the methods in this category are [BBBK07], [AJD+02], but the most popular one is Video Quality Metric or VQM [PW04], where the video sequence is divided into spatio-temporal blocks and seven parameters are extracted for comparison which are from spatial gradient activity, chrominance information, contrast information and absolute temporal information. Some of the most popular FR image quality assessment algorithms are based on simplified metrics such as SSIM (Structural Similarity Index Measure) [WLB04] and VIF (Visual Information Fidelity) [SB06]. These techniques can be extrapolated to the video domain with a frame-by-frame computation. Other interesting ideas are as follows: (a) motion information is recently found to be an important component and MOVIE [SB10] index performed extremely well in terms of correlation with human perception; (b) just noticeable distortion (JND) take advantage of the fact that HVS cannot perceive any changes between adjacent pixels below a threshold and several approaches based

on sub-band (DCT or wavelet) domain or pixel domain are proposed [MZLN10]; and (c) visual attention analysis [Koh03] based methods are found to be important attributes of the perceptual system currently ignored in most existing metrics and basically provides differentiated importance to various salient regions based on human attention [YPHG09].

RR algorithms operate without the use of pristine reference and instead, use additional information extracted from the original signal along with the distorted one. The extracted features can be PSNR [LKJ+02], attention based [YPHG09], wavelet transform [LK03] and SSIM [WLB04] which require lower bandwidth for transmission rather than using the entire reference signal for aiding the quality assessment task.

NR algorithms are completely blind i.e., only uses the distorted signal for assessing quality, which makes it difficult and challenging. Most work on NR so far relies on prior knowledge of the distortion process, such as degradation from compression that creates characteristic artifacts such as blocking, blurring, or ringing, to develop an algorithm [YGEMD07], [WSM01]. Typically coding and transmission parameters such as number of lost packets, type of packet lost and likewise are used as mapping artifacts for quality assessment. A truly NR algorithm that predicts the quality with a high probability is still to materialize and remains an interesting research problem. Human attentional mechanisms and motion analysis are attractive research directions to pursue [SB11, MB11b, YRH+10b].

3.2.3 Audio-Visual Quality Estimation

Audio-Visual quality assessment can be referred to Quality-of-Experience (QoE) due to its multiple modalities and is the end-goal of most communication systems, which are derived by the end user of the system. All the present techniques are

unfortunately related to assessment in one single modality and either focus on audio or video quality alone, without taking into account the possible cross-modal effects. ITU-R P.910 [itu99] describes a subjective assessment method that can be used to evaluate one-way overall video quality for multimedia applications such as video-conferencing in interactive scenarios.

Synchronization between the individual media components is a major indicator of the quality of multimedia signals. It has been found that the perception of asynchrony is not symmetric and humans are more tolerant of video being ahead of audio in the playback timeline rather than the opposite situation (audio stream leads the video stream) [YRH+10b]. It has also been proved by subjective tests that there is a strong mutual influence between audio and video on the experienced overall quality [YRH+10b]. To solve the multimodal quality assessment problem, a deeper understanding of both the human perceptual process to audio-visual stimuli as well as, at what stage the audio and visual processes are fused to form a single overall quality experience is required. The state-of-art multimodal quality assessment techniques follow a naive approach of fusion by a weighted linear combination of the different sensory modalities. The different weights are associated to the importance of audio or video channels which are found to be task-dependent i.e., video quality dominates audio quality in a number of situations such as high motion video, while audio quality dominates overall quality for specific stimuli such as "talking head" videos [SB11].

3.3 Image Enhancement

The image enhancement problem has been studied in the literature from various contexts, and thus can be classified across various dimensions: point-based process-

ing in Section 3.3.1, spatial processing in Section 3.3.2, object-based processing in Section 3.3.3 and data-driven processing in Section 3.3.4

3.3.1 Point-based Processing

Point-processing algorithms enhance each pixel separately without considering spatial interactions and dependencies with the neighboring pixels. Typical digital image enhancement techniques fall in this category such as histogram stretching/shrinking, gamma and power law transformation [GW02, Pra01, Jai84]. Log transformation compresses the dynamic range of an image after Discrete Fourier Transfrom (DFT) is applied to the image. Adaptive histogram equalization flattens the histogram throughout the entire range of the pixels, which is motivated by the information theory principle stating that a uniform probability density function contains the largest amount of information. All the above techniques are involved in shaping the contrast in some way by applying the function to the monochromatic channel for gray-level images. These methods have also been adapted for color images by transformation to other color spaces where the chromatic components are more uncorrelated from the achromatic components [LS84, KR95, PK96, LMM01].

3.3.2 Spatial Processing

Spatial enhancement techniques apply some form of filter or mask on the image and thereby each pixel and its spatial neighborhood has a dependency relationship. Some of the operations of this category are convolution filter in the spatial and Fourier frequency domains, Wiener filter, homomorphic filters, high-pass and low-pass filters [GW02, Pra01, Jai84]. Other variants of this category are the non-linear filters (transform function is nonlinear), order-statistic filters (intensity values based

on their rank within a spatial processing window), morphology filters (utilizing the maximum or minimum value of the order statisitic in the local window) [KA97, Ser82, APM00].

Another form of nonlinear enhancement algorithms is the transform-mapping framework, where the image is first processed with a transformation operator that separates the original image into two components: one is semantically meaningful and the other is noise. A nonlinear mapping then removes the noise and the inverse transformation operator is applied to generate the enhanced image. An example of this category is image denoising filters with wavelet transformation. Edge-adaptive filters focus on the preservation of edges in the image frame. The edges can be easily detected as corresponding to siginificant differences of pixel intensities between adjoining pixels. Maintaining the edges is critical for the spatial integrity of the objects in the scene. A typical edge preserving filter utilizes an adaptive gaussian kernel as its transformation operator.

3.3.3 Object-based Processing

The success of the above techniques paved the way to conceive more sophisticated controls with complex characteristics. Shape and orientation based adjustments for the various objects within the image were proposed by many researchers which led to the development of interesting application areas. Carroll et.al. [CAA10] presented a system for artistic perspective manipulation to produce a variety of effects such as changing the perspective composition of a scene not realizable with a camera using shape-preserving image warps. Zhou et.al. [ZFL$^+$10] proposed an easy-to-use enhancement technique for realistic reshaping of human bodies in a single image producing visually pleasing result with a variety of poses and shapes by a novel body-aware image warping technique. A data-driven enhancement system for facial

attractiveness in frontal portraits is proposed in [LCODL08] which learns a set of distances from a variety of facial feature locations using a ground-truth training dataset.

3.3.4 Data-driven Processing

Data-driven processing is another line of research followed by enhancing images using data features such as color and tonal properties of images, to achieve various objectives. One general trend in this direction is to transfer the image characteristics of a source image to a target image by color harmonization [SJMP10], tonal and textural features [BPD06]. Other proposals in this category involve learning of characteristic image properties from example images and then using this knowledge to enhance arbitrary/unseen images. Examples are: color harmonization [COSG+06], relighting effects by neutralizing the light colors using spatially varying while-balance [HMP+08], personal photo enhancement [JMAK10], color theme [WYW+10], color and tone style [WYX11], and collaborative personalization [KKL10], [CKK11]. Dale et.al. [DJS+09] described an image restoration method using various operations such as white balance correction, exposure correction and contrast enhancement by leveraging a large database of images gathered from the web. An effective decolorization algorithm is proposed in [AAB11], that preserves the appearence of the original color image by blending the luminance and the chrominance information in order to conserve the initial color disparity while enhancing the chromatic contrast. An efficient method for recovering reliable local sets of dense correspondences between two images with some shared content is described in [HSGL11], with applications in automatic example-based photograph enhancement such as adjusting the tonal characteristics of a source image to match a reference, transferring a known mask to a new image, and kernel estimation for

image deblurring.

Quality enhancement of images is one of the heavily studied areas with many proposals in the literature. The general objective of all these techniques are essentially to improve the quality of pictures by manipulating color, contrast, tonal, textural features of the degraded images [RJW02], [MI07], [MM08]. Computational models for media aesthetics [JDF+11] and emotion [MH10] are slowly gaining importance in bridging the semantic gap [DJLW06] and to create new applications [ZTH+09] due to its close correlation with human visual perception. Though there are many image-editing software present in the market such as Adobe Photoshop and the open-source GIMP, these applications require manual control for enhancing images which is not practical in large-scale environments. Though they are provided with automated controls such as the Retinex filter [RJW02], but these techniques do not consider affect or emotion while adjusting the image, which is our focus in this paper. We follow a data-driven approach in our framework since it is non-intrusive (no human involvement after collection of training data) in nature and more practically feasible due to its computational characteristics.

3.4 Drawback of Previous Approaches

In this section, let us summarize the drawbacks of all the different approaches that are discussed till now as follows:

- *FR/RR* algorithms are the most popular and the most studied ones based on the different spatio-temporal metrics such as MSE, PSNR, SSIM, MOVIE on a frame-by-frame basis. But the most primary drawback of this approach is its unusability in many real-time multimedia applications such as VoIP, Internet

video streaming and similar applications. where the reference signal is not available at the client end.

- *Subjective/MOS-based* techniques are intrusive since they require user's explicit input during the evaluation process, which makes them expensive and slow with high resource overhead and limited efficiency. Moreover, subject's feeling can be exposed to various effects such as personal biases and limited human memory effect. The main weakness of MOS-based evaluation is its varied and vague interpretation of levels i.e., how should a user differentiate between levels 1/2 or 2/3 (e.g. if a user feels the quality is average then whether he/she should rate it as 2, 3 or 4 [CTX09]) and the cognitive distance between Bad (1) and Poor (2) is usually different from that between Good (4) and Excellent (5) [WS98].
- *QoS-based* methodologies estimate quality from underlying transport/network level properties such as delay, loss-rate, and bandwidth by continuous probing and monitoring. Some of the drawbacks are as follows: (a) deterministic modeling from QoS to QoE is not able to cover various factors related to cognitive and behavioral perceptions [WAR+09], (b) message overhead for continuous probing, (c) detecting multimedia traffic is nontrivial [BS06], [BMM+07] and (d) modeling delay/loss-rate effect on QoE is delicate due to buffer masking [WCHL09].
- *Media-based* quality assessment mechanisms are heavily studied and found to perform well in estimating perceptual media quality. But most of them are FR-based algorithms which have the obvious drawbacks as discussed above. Moreover these methods fail to consider certain factors from real-time interactive applications such as variable listening levels, sidetone/talk echo effects, conversational delay/interaction [CHHL06]. Multimedia signals generated from

multiple sensory modalities such as audio+video and their mutual influence which can have a significant impact on the final multimodal quality or QoE. These are also not addressed in media-based techniques since they only model media-content related metrics extracted through various audio, video, and image processing algorithms.

CHAPTER 4

QUALITY OF EXPERIENCE EVALUATION

Our choice of affect-based approach for QoE evaluation is linked to the fact that human emotion have a high degree of semblance to subjective user perception which plays a decisive role in the cognitive decision process of assessing quality. We propose and examine the following research hypothesis:

- H: *The user perception of multimedia communication quality is correlated to his/her affective feedback response (i.e., audio, visual, speech, physiological), which will vary across different conditions that cause quality distortion (such as network channel properties, source coding artifacts).*

We start this chapter with an introduction of the affective framework in Section 4.1 followed by a detailed discussion in Section 4.2. Then we illustrate the experimental design in Section 4.3 and we end this chapter with a demonstration of the results in Section 4.4.

4.1 Introduction

The voice communication industry is undergoing a rapid phase change with technologies such as cellular, mobile and Internet telephony replacing the conventional telephone networks. Service providers are faced with offering high communication quality under more heterogeneous and dynamic networking conditions. Effective evaluation of system performance is becoming critical, which will serve as an important instrument for quality monitoring and management. Traditional evaluation methods are very system-oriented where Quality of Service (QoS) metrics have been the de facto standards for voice communication technologies.

Recently, there is a paradigm shift towards user-oriented methodologies with the introduction of human-centric computing in the systems area [JSGP06], and the concept of *Quality of Experience* (QoE) is gradually gaining popularity [Jai04, TYHT08, Ebr09]. Since QoE metrics are closely related to human perception, they could potentially serve as more valuable quality indicator from the user perspective.

There are no universal definition of Quality of Experience (QoE) due to its varied interpretation in different context. Some consider the notion of QoE to be the totality of QoS mechanisms provided to ensure smooth transmission of audio and video over IP networks. Others consider QoE to be the perception of elements of the network and performance relative to the expectation of users/subscribers. By combining the different interpretations, we can deduce QoE to be a multi-dimensional construct of user perceptions at the end-point application layer and can be conceptualised as the experience derived from the total QoS provided by the different layers of the communication stack.

The main challenge of how to evaluate QoE remains largely unsolved due to its complexity. Over the years, different meanings have been attached to the term [qoe98, BC07, WAR+09]. Theoretical studies from various disciplines characterize QoE as a multi-dimensional construct which involves both subjective and objective factors intertwined in the user interaction such as perception, emotion, behavior, need, context, system and networking [FB04, Csi90, VTR92, HS82, AMR02, DBW89, VMDD03]. In practice, system developers have applied QoE assessment techniques, ranging from user feedback [CTX09, itu96], QoS-based estimation [CHHL06, SW07] to media quality analysis [pes01, itu05]. Despite the value demonstrated, each approach has limitations and weaknesses as elucidated in related work (Chapter 3).

In this paper, we investigate the usability of *affective computing* in evaluating QoE of voice communication. Affective computing deals with the analysis of human emotional variables naturally revealed during the user-system interaction. In the process, emotions have been shown to have strong association with user experience regarding interest, satisfaction, motivation and performance [Nah04, Nah98, TWZ+08, Kra02]. Using signal processing, linguistic analysis, and psychophysiological techniques, automated emotion recognition is feasible by aggregating affective cues from multi-modal user input such as facial expression, speech, body gesture, and neuroimaging [ZPRH09, KP05, Pel05]. Leveraging on these findings, affect-aware systems are emerging that dynamically adapt according to the change of user emotions in applications of user interface, health care, education, customer service, intelligent automobile, entertainment, information retrieval and social signal processing [CD10, AJG08, AKJ09, VPBP08, VPB09, CP07].

Guided by the above evidences, we hypothesize that QoE or the user perception of quality in voice communication is correlated to his/her *affective behavior* (e.g., pitch, voice, timing and articulation), which will vary across networking conditions. To the best of our knowledge, the analysis of affective behavior and its role in the QoE evaluation of voice communication is an unexplored area. As an initial step, we focus on experimental user studies to record changes of user affect state and examine at what level QoE is reflected from these changes.

The general scheme of the experiment is given in Figure 4.1, which shows the network setting, the voice communication, the audio signal input, and the affect analysis framework. Our study is based on the context of human-human voice

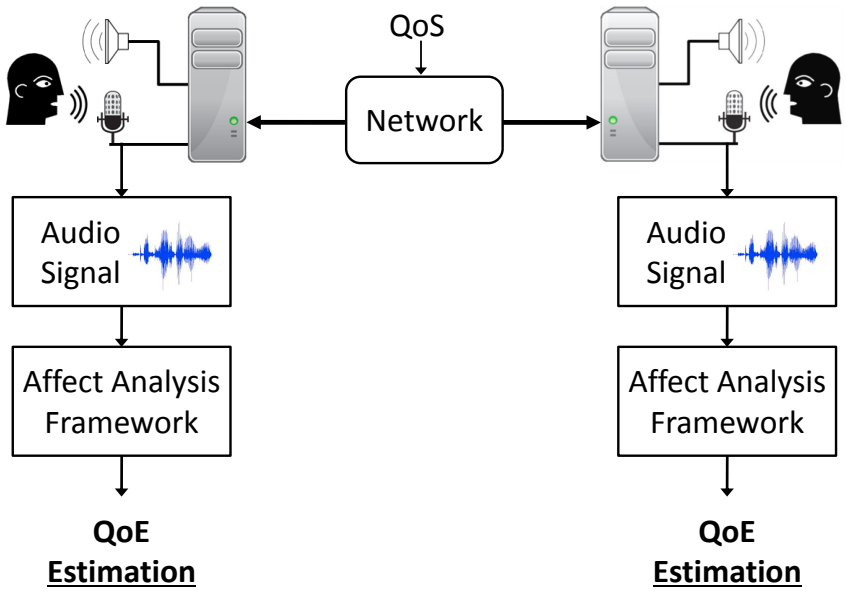

Figure 4.1: The overview of affect-based approach for QoE evaluation in voice communication.

communication. In a test session, two participants are engaged in natural conversation through a VoIP connection mediated with varying QoS parameters (e.g., delay, loss rate, and bandwidth). During the conversation, the communication quality is tagged by user feedback ("*Good*", "*Average*" and "*Bad*"). The conversation is saved and further processed by the affect analysis framework offline. The front end of the framework performs feature extraction to derive samples. The set of features are drawn from three major categories of affective signals involving *acoustic*, *lexical* and *discourse* features as suggested by previous research in emotion detection [CDCT+01, ZPRH09, BFH+03, LN05]. The affective behavior is analyzed individually for each category as well as combined. A subset of samples are used to train classification modules. Then, the rest samples are used for performance testing. Finally, the accuracy analysis is provided by comparing the output of the affect analysis framework (i.e., QoE estimation) with the user tagged feedback.

Our contributions in this paper are the followings. Most of all, *we provide a new affect-based methodology of QoE evaluation in voice communication.* Different from previous approaches, we propose to assess quality directly from the user affective responses. Therefore, our method has the advantage of deriving subjective QoE measures in an implicit and non-intrusive manner. The experimental results indicate very promising prospect of this approach. Regarding boarder impacts, as the communication systems become more media-rich and interactive (e.g., spatial audio, 3D/immersive space), measuring QoE via indirect methods will become more challenging [Ebr09]. Therefore, our work represents an important step towards the understanding of QoE for the future generation of communication systems.

4.2 Affect Analysis Framework

The basic approach of affect-based framework is to apply *multi-modal analysis* for QoE evaluation. The front end of the framework performs feature extraction from the audio signals of user conversation. In our study, three major categories of affective signals are extracted including *acoustic* features, *lexical* features and *discourse* features. The importance and usage of acoustic features for automatic emotion detection has been well-studied [CDCT$^+$01]. However, acoustic-based methods are effective for posed expression in staged scenarios but degrades in natural human interaction with spontaneous expression [ZPRH09]. The common wisdom is to incorporate acoustic features along with lexical and discourse features for performance enhancement [ZPRH09, BFH$^+$03]. Since we target normal human-human voice communication, we include all these features in the framework. After feature extraction, the affective response is analyzed individually by the classification module of each category. In the last stage, the results are then combined to generate the final output of the framework (i.e., QoE estimation).

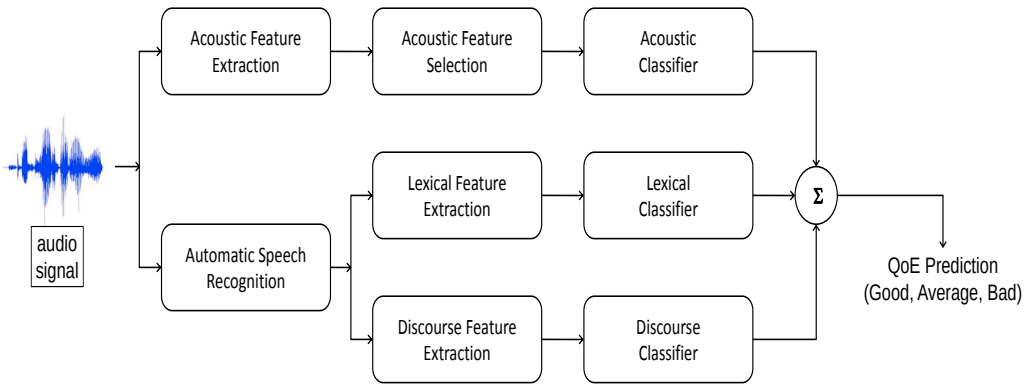

Figure 4.2: The block diagram of the audio analysis framework.

We follow the general methodologies and technical justifications as presented in [LN05] for the design of the framework, but we independently complete the implementation and integration work. We briefly introduce each module of the framework in the following sections as it is not our major focus of this research. An overview of the framework is presented in Figure 4.2.

4.2.1 Acoustic Features

We consider 22 different acoustic attributes related to segmental and suprasegmental information of speech signals. These attributes are derived from turn-level statistics and transformations in the domains of *fundamental frequency*, *energy*, *duration*, and *formants*.

- *Fundamental Frequency* ($F0$): the lowest frequency of the signal wave. We use statistical functionals of mean, median, standard deviation, maximum, minimum, range (maximum-miniumn) and linear regression coefficient.

- *Energy* (E_r): computed as the Root Mean Square (RMS) of the signal, i.e., $E_r = \sqrt{\frac{\left(\sum_{n=0}^{N} x_n^2\right)}{N}}$ for a PCM frame of size N. Similarly, we use statistical functionals of mean, median, standard deviation, maximum, minumum, range, and linear regression coefficient.

- *Duration*: computed by comparing various voiced and non-voiced regions in a temporal-domain analysis of the speech waveform. The individual attributes are speech-rate, duration of the longest voiced speech and ratio of voiced and unvoiced region.

- *Formants*: basically the resonance of human vocal tract. Formant location and bandwidth are used to identify phonetic property of human speech. We use the first and second formant frequencies ($F1$, $F2$), their corresponding bandwidths ($BW1$, $BW2$) and the mean.

The module of acoustic feature extraction was implemented based on the tool of `openSMILE` [EWS10], an open-source software that can extract many low-level acoustic features and statistic functionals. The 22 acoustic attributes form the base feature set (denoted as $Base$). Generally, not all features are equally significant for affective analysis. Following the work of [LN05], we perform feature selection with an expectation to improve prediction accuracy by dropping off the least significant attributes. We employ the *leave-one-out* method for feature selection and the *nearest neighborhood* rule for estimating accuracy. We generate two subsets from the base feature set, one has 10 best features (denoted as $f10$) and the other has 15 best features (denoted as $f15$). Lastly, we obtain another feature set by performing Principal Component Analysis (PCA) on the base feature set, which is denoted as PCA.

4.2.2 Lexical Features

Lexical features here refer to language related information regarding the fact that people tend to choose specific words for various expressions, e.g., "*can't*", "*damn*", "*bad*", and "*great*" [Csi90]. In our study, we adopt the notion of *mutual information* [GLM+90] to establish the correlation between words and different QoE levels.

Denote the vocabulary $V = \{v_1, v_2, ..., v_n\}$ and the set of quality levels $Q = \{q_1, q_2, ..., q_m\}$. As mentioned earlier, we choose $m = 3$ (i.e., "*Good*", "*Average*", and "*Bad*"). The mutual information is given as:

$$I(v, q) = \log \frac{\Pr\{q|v\}}{\Pr\{q\}} \tag{4.1}$$

for $v \in V$ and $q \in Q$. Intuitively, if a word v is correlated with a quality level q, then $\Pr\{q|v\} > \Pr\{q\}$ and $I(v, q)$ is positive. If there is no correlation, then $\Pr\{q|v\} = \Pr\{q\}$ and $I(v, q)$ is zero. If v makes q less likely, $I(v, q)$ is negative.

Given an utterance, the *action* for quality level q_k can be calculated as:

$$a_k = \sum_{i=1}^{n} O_i I(q_k, v_i) + \log(q_k) \tag{4.2}$$

where O_i is the likelihood that word v_i is recognized in the utterance by the speech recognition module. For simplicity, we use a one-layer network where the order and sequence of words are not considered. Finally, we use $\{a_0, a_1, a_2, a_0 - a_1, a_1 - a_2, a_2 - a_0\}$ as the feature set for the lexical information. We took leverage from the Automatic Speech Recognition (ASR) system from the `HTK Toolkit` [WOVY94] of Cambridge University for predicting text output of voice signals, which is based on the Hidden Markov Models (HMM). The models of HMM are accessed and trained on Wall Street Journal (WSJ) corpora [Ver06] and the generated tied-state cross-word triphones are utilized for later recognition purposes. We develop a 3-gram language model provided in a 1,200 million English Gigaword corpus [GC03]

indexed with the Linguistic Data Consortium catalogue and also coupled with a 125,000-word CMU pronounciation dictionary [Rud07].

4.2.3 Discourse Features

The value of discourse information for emotion recognition has been noted in the literature, particularly combined with acoustic information to improve the accuracy [ADK+02, AGA+01]. In our study, we take a simplified approach by modeling only *repetition*, which is found to be the most important indicator of trouble in communication [BFH+03]. We choose from 1-word to 5-word repetitions and formulate the dimensions of discourse feature set as: number of 1-word repetition, number of 2-word repetitions, and likewise. Moreover, we construct another repetition metric as the following:

$$R = \sum_{i=1}^{5} r_i * w_i \tag{4.3}$$

where r_i is the number of i-word repetitions and w_i represents a proportional weightage which assigns higher weight to the repetition of longer sequence of words (e.g., 5-word repetition). The discourse feature extraction relies on the same automatic speech recognition module as described in the lexical features (Section 4.2.2).

4.2.4 Classifiers

The classifiers form the core component of the framework, which is used to provide QoE prediction based on an input set of features extracted from previous feature extraction modules (shown in Figure 4.2). The original data set contain samples tagged with user quality feedback, which is divided into training set and testing set. For the initial training phase, samples of the training set are utilized to build a classifier. During the testing phase, samples of the testing set are fed into the

classifier (with user feedback removed) which will output QoE predictions. Then, the output is used to compare with the user feedback. Overall, 75% of the data samples are used for training and the rest 25% for testing.

Two basic types of classifiers are used here, namely *Support Vector Machines* (SVM) and *k-Nearest Neighbors* (kNN). SVM involves the construction of a set of hyperplanes by maximizing the separating distance between the nearest training data points among all classes. In contrast, kNN works by computing the k nearest neighbors to the input sample based on a distance metric (by default, Euclidean) and using a majority vote among the neighbors to determine the class label of the sample. Our rationale for choosing SVM and kNN classifiers are mainly due to their effectiveness and performance benefit. Though we test various classification methods, our framework does not depend on any one particular technique.

We make no assumption on the dependency between these features and the user experience of quality. Instead, we take an unequivocal classification approach, using the ground truth from the training data. For SVM classifier, we trained our models with the radial basis function (RBF) kernel. To identify the optimal values of C (cost) and γ (kernel parameter), we apply *cross-validation* with iteratively refined grid-search method [AKJ09]. The notations of the four different SVM derivatives used in our study are given as follows.

- *SVM*: uses fixed training and testing sets.
- *SVM–5CV*: divides the data into five randomly chosen segments of equal size and run five times with each run comprising of four segments for training and one segment for testing (5-fold cross-validation).
- *SVM–5WC* and *SVM–10WC*: employs a 2-layer hierarchical SVM model that trains 5 and 10 bottom-layer weak classifiers (*5WC* and *10WC* respectively)

on different subsets of the training set and the output of each is used to train a top-layer meta-classifier.

For kNN classifier, we trained two derivatives. Their notations are given as follows.

- *kNN*: similar to that of *SVM* with $k = 10$.
- *kNN–5CV*: similar to that of *SVM–5CV* with an iterative number of nearest neighbors from $k = 1$ to 15.

Since we incorporate multiple affective sources (i.e., acoustic, lexical, discourse) in our framework, we need aggregation scheme to produce a single output. One simple method is to employ *feature-level* fusion which merges all features from different sources into a large feature vector as the input to a single classifier. In contrast, *decision-level* fusion takes results from multiple classifiers to compute a single value. We take the latter approach and calcuate the average of results from each classifier as the final output. As simple as it is, such method achieves pretty good performance [KHDM98, TvBDK00]. For software, we use the tools of `libSVM` [CL01] for SVM classifiers and `Weka` [HFH+09] for kNN classifiers.

4.3 Experimental Design

We designed the user study experiment for the examination of the following research hypothesis:

H: *The user perception of voice communication quality is correlated to his/her affective response, which will vary across networking conditions.*

The main purpose of the experiment was to collect two types of data: audio signal of user conversation and the user quality feedback. For that purpose, we engaged two participants in natural conversation through a VoIP connection. During the

test, we tuned the QoS parameters of delay, loss rate and bandwidth to simulate various networking conditions. We expect that the affect state of user as reflected from his/her voice will change accordingly, as well as his/her subjective perception of the communication quality. Thus, we recorded the voice conversation and asked the user to rate the quality. The collected data were used for training and testing by the affect analysis framework in the next stage (Section 5.2).

4.3.1 Networking

We installed two desktop computers with Intel i686 Core 2 Quad CPU 32-bit (2.39GHz) and 2GB RAM running Linux kernel (Ver. 2.6.31.5) as each end of a VoIP connection. The two computers were placed in two separate and quiet rooms of our department building (one in the second floor and the other in the third floor). During the experiment, no other person was allowed inside the room to avoid any psychological influence on the participants.

We configured a third computer (with similar hardware settings and the same operating system) to be used as a layer-2 bridge. The bridge computer had two network interface cards to connect with the two VoIP end computers. We installed the Linux `brctl` utility on the bridge computer to form a LAN: (VoIP end computer 1) $\leftrightarrow$ (bridge computer) $\leftrightarrow$ (VoIP end computer 2). All physical connections were based on dedicated wired lines with no interference from external network traffic. To simulate various networking conditions, we instrumented the traffic flow between the two VoIP end computers by applying `dummynet` [Riz97] on the bridge computer. The dummynet software allows us to tune the network condition with different delay, loss rate and bandwidth settings.

4.3.2 Voice Channel

Each VoIP end computer was equipped with a Logitech headset with speaker/microphone for voice capturing, playback, and recording. We used the library of `PortAudio` [Ben01] to record audio signals from the microphone into wav files that can be further processed by the feature extraction modules. For simplicity, each end only records its local signal for QoE evaluation. The interplay between local and remote signals would be an interesting topic for future research. We deployed `PJSIP` [PIIM05] as the VoIP software with G.711 codec. It is an open-source, comprehensive, and highly portable system with a small memory footprint. Although PJSIP contains lots of features, it is console-based and the basic command set is very simple. After a few minutes' instruction, all participants were capable of operating it. To facilitate the experiment, we automated the communication and session management process for call initiation, connection, and recording, so that two end users only need to press a few keys to launch a test session.

4.3.3 Sample Collection

We organized each test session based on a maximum 15-minute run. The limit here was chosen basically due to our observation that it was not natural to engage two strangers in a phone conversation for over-stretched duration. Thus, to improve the efficiency of every test we divided a run into multiple 20-second intervals. Timescale of each interval is an important issue for QoE evaluation which is justified in Section 4.3.4. Each interval corresponded to a fixed setting of QoS parameters. We adopted the approach of OneClick [CTX09] in the sense that at the end of each interval a console prompt was shown to ask for user input of quality. Different from OneClick, we employed a trichotomous or 3-point scale decision of quality

levels: "*Good*", "*Average*", and "*Bad*." Although this experimental methodology still bears some intrusiveness, it is unavoidable in the initial study since we need user feedback for training and testing the affect analysis framework. The complete establishment of the framework will eventually eliminate the need of user feedback.

The interval was saved as one sample after it was tagged with QoS parameters and user feedback. Then, a different set of QoS parameters was configured into the network and the next interval was started (shown in Figure 4.3). All the steps were automatically synchronized. Besides pressing a key every 20 seconds, there was no extra distraction to the end user. Overall, one test run provided 90 samples (i.e., 45 samples from each end).

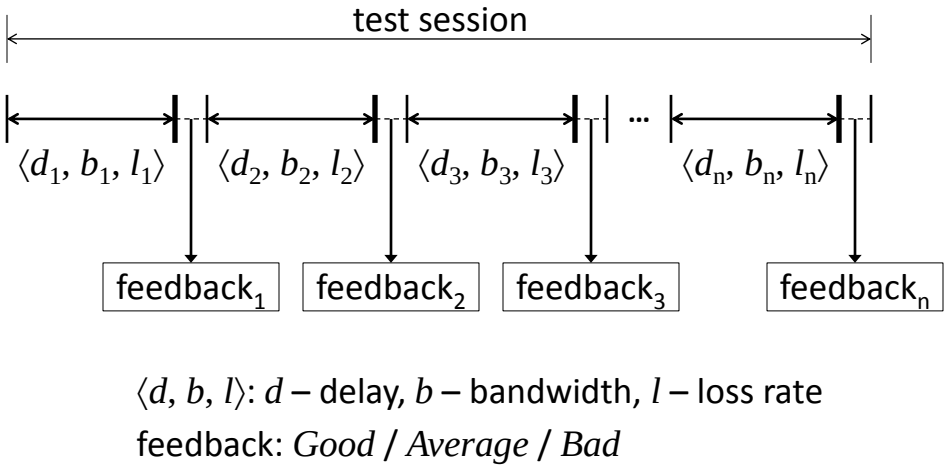

Figure 4.3: The organization and sample collection procedure within one test session.

4.3.4 Timescale

It has been recognized that user perception of speech quality varies under different temporal scale [RBK+06]. For example, in short-term test ($\approx$ 30 seconds) the beginning portion (at least $>$ 8 seconds) of the speech sample carries the greatest weight in the overall rating of MOS (*primary effect*) [GMH97], while in long-term test ($\geq$ 60 seconds) the last portion carries the greatest weight (*recency effect*) [Ros98]. Regarding its psychological merit of human memory mechanism, these findings give us the guideline for deciding the timescale of test interval. Because user feedback is requested at the end of each interval, we should choose short duration for each interval to leverage on primary effect (but not too short). On the other hand, since the QoS parameters are fixed in one interval, there is no need to apply long duration due to the recency effect. Considering previous research results and the possible interference to the user, we chose 20 seconds as the fixed timescale for each test interval.[1]

4.3.5 QoS Classes

We employed delay (d), bandwidth (b) and loss rate (l) as the basic QoS parameters. For notation, each networking configuration is denoted as a tuple of $\langle d, b, l \rangle$. Because the space of possible configurations is quite large, we applied a selection procedure to pick the most "*distinctive*" value set. In the pilot experiment, we performed the empirical study to pick up a meaningful range for each parameter. First, we set the best configuration as $\langle$50 ms, 100 Kbps, 0.06$\rangle$. Then the worst configuration of each parameter was determined incrementally while keeping the other two as the

[1]The ITU-T recommendation for the subjective assessment of sound quality also suggests the clip duration from 15 to 20 seconds, and the minimum number of participants in an experiment to be between 10 and 20 [itu03].

best until the quality became intolerable. In such a way, the worst delay was set as 1200 ms, worst bandwidth 52 Kbps, and worst loss rate 0.3. Next, we picked 5 values from each parameter range evenly and aligned them together from best to worst and generated 5 QoS classes.

- C_1: $\langle d = 50$ ms, $b = 100$ Kbps, $l = 0.06\rangle$
- C_2: $\langle d = 300$ ms, $b = 88$ Kbps, $l = 0.12\rangle$
- C_3: $\langle d = 600$ ms, $b = 76$ Kbps, $l = 0.18\rangle$
- C_4: $\langle d = 900$ ms, $b = 64$ Kbps, $l = 0.24\rangle$
- C_5: $\langle d = 1200$ ms, $b = 52$ Kbps, $l = 0.3\rangle$

For further validation, we asked a few graduate students to run MOS-based test to rate the quality for each configuration. The test was done in a random and blind way (i.e., the students were not aware of the configuration). The average rating results confirmed clearly the quality difference between each configuration.

4.3.6 Participants

We recruited 15 participants from two mid-level undergraduate classes, namely, Java Programming and Computer Networks.[1] We found that selecting students from same classes provided better conversation control, for example, the familiarity and the common interest (details in Section 4.3.8). All students were proficient with English language (6 native, 7 advanced and 2 intermediate speakers). Among them, 11 were male and 4 were female, between 18-38 years of age ($M = 25.54$, $SD = 3.38$). They had a mean of 3.67 years of using VoIP service and all claimed to have been using at least one VoIP service in the past (with the most popular being Skype followed by GTalk).

For better control of the test (more in Section 4.3.8) and regarding the difficulty of accommodating each participant's schedule, in each test session a graduate student was assigned to play one VoIP end with another undergraduate student at the other end (i.e., the real participant). So there were totally 15 test sessions. For statistic validity, the samples from the graduate student were discarded. Participation in the experimental study was considered as a benefit by the students and no monetary compensation was involved, since the conversation topics were chosen from course materials as discussed later.

4.3.7 Questionnaire

Before the experiment, each participant was introduced with the experimental details following Institutional Review Board (IRB) required procedures. Then he/she was asked to complete an Entry Questionnaire at the beginning of the study for collecting background/demographic information, and their previous experience with VoIP if any. Exit Questionnaires were also provided at the end of each session to elicit subjective experience of the user for the entire audio conversation. The questionnaire gathered information on the subjective quality perception and the evaluation criterion, as well as the participant's view of the importance of affective feedback regarding the usability and ethical outcome (for IRB purposes). Some of the questions are listed below.

Q: *What is your overall perception of the system?*
Q: *Did the system perform as per your expectation?*
Q: *Were you able to interact with the system without any problems?*
Q: *Did you feel any external or environmental influence during the experiments?*
Q: *Did you expect any more help from the system?*

4.3.8 Conversation

Since conversational materials are highly correlated to the quality metrics [WA04], a careful design of dialog exchange is important for the experiment. We wanted to invoke natural conversation from the participants. Initially, we thought about selecting a few topics. However, the cultural and emotional influence of the topic on each user would be hard to predict and may interfere with his/her experience. The ITU recommendation also suggests that the content should be "neither so interesting nor so disagreeable or boring" [itu03]. To make the communication more controllable, we decided to give quiz-like conversation formed using course-related materials with low-level difficulty. The graduate student (actually the TA of the class) presented some questions and discussed with the student. This kind of conversation can be easily engaged and the topic is neutral. Moreover, to avoid over-burdening the participants, we embedded some brief comments on general topics or short riddles from time to time to provide cognitive relief.

It is observed that in some cases the conversation itself might have invoked emotional responses that are not related to communication quality perception. In the experiment, we attempted to reduce this kind of effect by preferring neutral conversational content as described. However, in the real scenario, people do engage with heated conversational dialog among themselves. In our preliminary experiment, we tried to filter out "emotional noise" from the perspective of conversational content by considering lexical and discourse features such as salient words and repetitions. Note that, such drawback is inherent in any type of subjective assessment including MOS. This challenging issue deserves further study which is the focus of our future work.

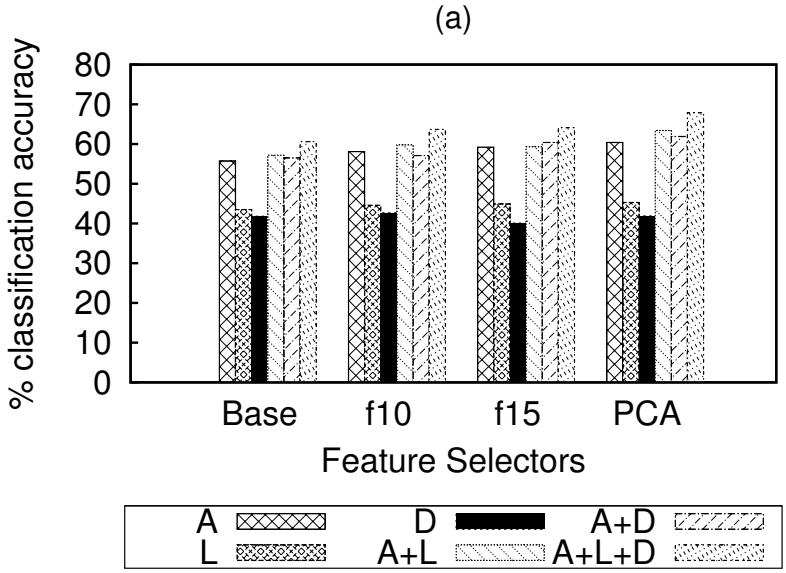

Figure 4.4: Classification accuracy for different combinations of affective sources (A=Acoustic, L=Lexical, D=Discourse), feature selection schemes (*Base*, $f10$, $f15$, PCA) and SVM classification technique.

4.4 Results

In this section we present the experimental results of our study based on test sessions collected from 15 subjects. One finding of interest is the performance of QoE prediction where we compared the output from the affect-based framework with the user feedback (both on a 3-point scale of quality levels). We also highlight other results and implications regarding the interaction between QoS setting, user feedback and quality prediction.

4.4.1 Performance of Estimation

We conducted a comprehensive study of estimation accuracy along three dimensions as introduced in Section 5.2: (1) different classifiers of SVM and kNN, (2) different combinations of affective sources (acoustic: A, lexical: L and discouse: D), and (3) different feature selection schemes (Base, $f10$, $f15$, and PCA), which are

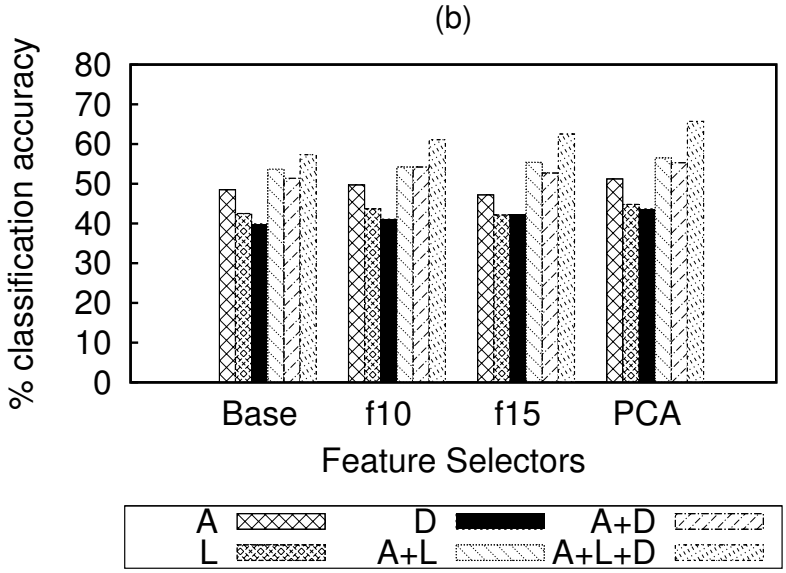

Figure 4.5: Classification accuracy for different combinations of affective sources (A=Acoustic, L=Lexical, D=Discourse), feature selection schemes (*Base*, $f10$, $f15$, PCA) and *SVM–5WC* classification technique.

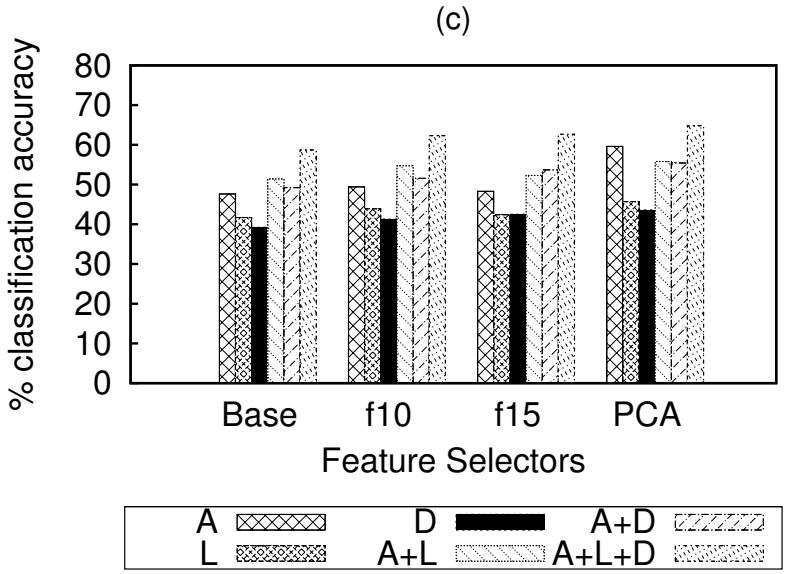

Figure 4.6: Classification accuracy for different combinations of affective sources (A=Acoustic, L=Lexical, D=Discourse), feature selection schemes (*Base*, $f10$, $f15$, PCA) and *SVM–10WC* classification technique.

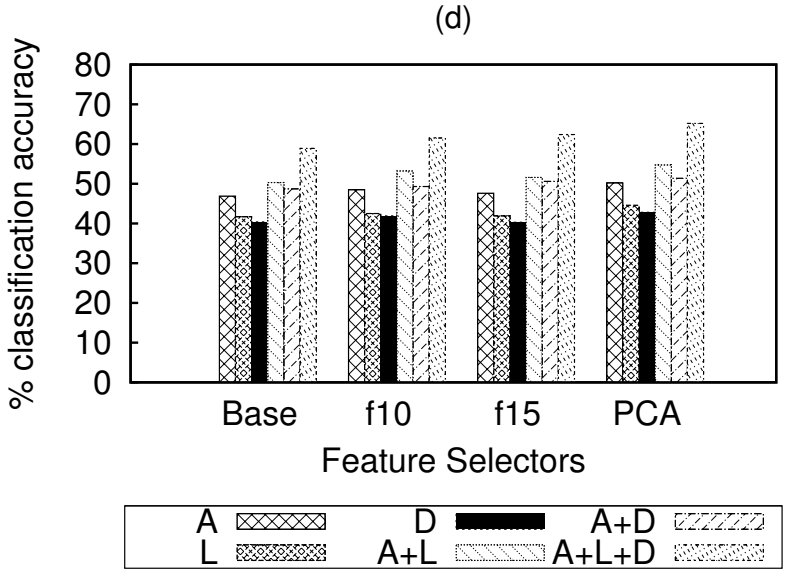

Figure 4.7: Classification accuracy for different combinations of affective sources (A=Acoustic, L=Lexical, D=Discourse), feature selection schemes (*Base*, $f10$, $f15$, PCA) and $SVM–5CV$ classification technique.

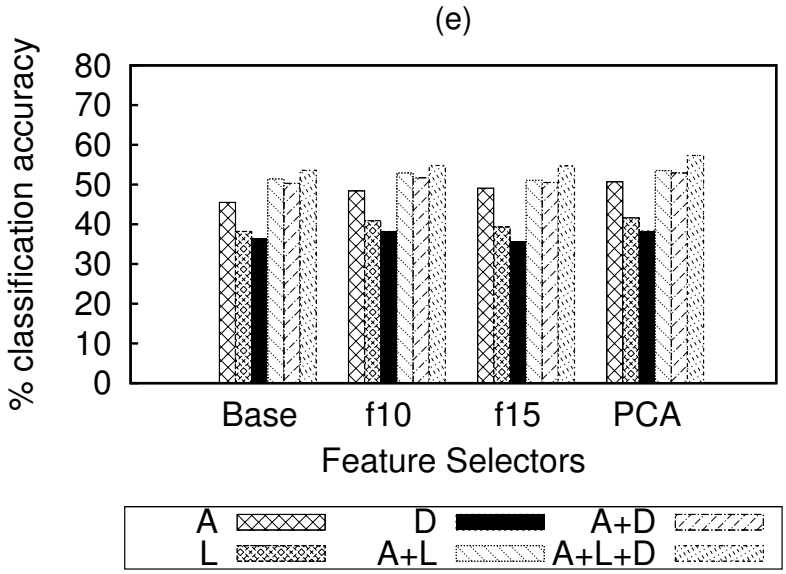

Figure 4.8: Classification accuracy for different combinations of affective sources (A=Acoustic, L=Lexical, D=Discourse), feature selection schemes (*Base*, $f10$, $f15$, PCA) and kNN classification technique.

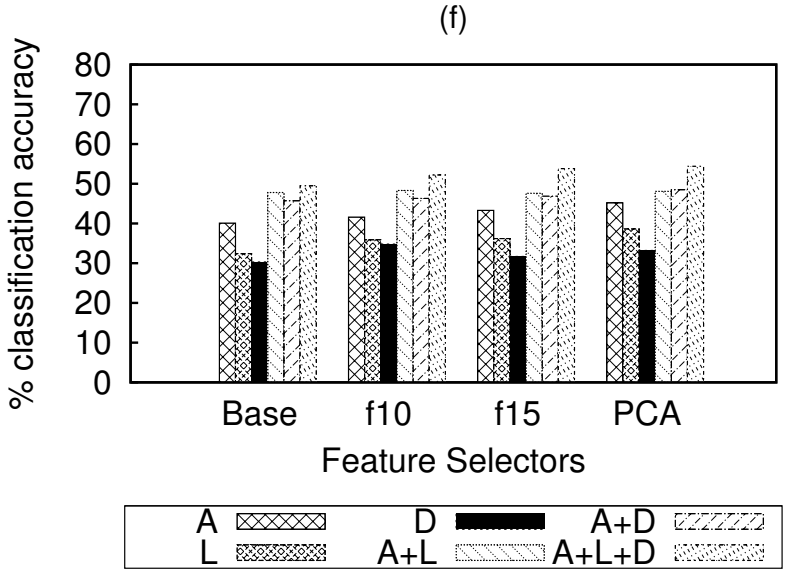

Figure 4.9: Classification accuracy for different combinations of affective sources (A=Acoustic, L=Lexical, D=Discourse), feature selection schemes (*Base*, $f10$, $f15$, PCA) and kNN–$5CV$ classification technique.

shown in Figures 4.4, 4.5, 4.6, 4.7, 4.8, 4.9. The overall findings are summarized as follows: (a) combining other affective sources with acoustic features consistently improves the performance where the best results appear from the aggregation of all sources (A+L+D), (b) considering single source, QoE prediction based on acoustic source is more accurate than the other two whereas the performance of lexical source is slightly better than that of discourse source, and (c) the performance impact of different feature selection schemes is less noticeable in kNN classifiers than SVM classifiers.

Table 4.1 shows the cross-view of the results based on all affective sources (A+L+D) with different feature selection schemes and classifiers. In all cases, the PCA feature selection scheme provides the highest performance. The $f10$ and $f15$ feature selection schemes provide comparable accuracy which is consistently higher than the base feature set. Comparing different classifiers, it is observed that SVM with 5-

fold cross-validation (*SVM–5CV*) gives the highest accuracy. As a short summary, the best performance is achieved with SVM 5-fold cross-validation and PCA-based feature selection, which gives an accuracy of 67.9%.

Table 4.1: Classification accuracy with different feature sets versus varying classifiers from the combination of all affective sources (A+L+D).

Classifier Model	Base	$f10$	$f15$	PCA
SVM	57.3	61.1	62.5	65.7
SVM–5WC	58.7	62.3	62.7	64.8
SVM–10WC	58.9	61.5	62.4	65.2
SVM–5CV	60.6	63.7	64.1	67.9
kNN	49.7	52.8	53.1	54.3
kNN–5CV	53.5	54.2	54.9	57.4

4.4.2 QoS Distribution

We next examine the distribution of QoS classes among various user quality ratings. Recall that, each conversation interval is initialized by applying a particular QoS class (i.e., $C_1, C_2, ...C_5$, Section 4.3.5) to set the networking condition and a quality rating is recorded from the user feedback at the end of the interval. Figure 4.10 shows the distribution of the five different QoS classes with respect to the user feedback for "*Good*", "*Average*", and "*Bad*" quality ratings. As seen, the highest portion of "*Good*" quality rating is associated with QoS class C_1 (42.0%), and the overall trend is decreasing from C_1 to C_5. However, quite interestingly the results also show that C_3 contributes more for "*Good*" quality rating (30.0%) in comparison with C_2 (20.0%). In a similar way, C_4 takes the highest distribution for "*Bad*" quality rating (34.0%) which is more than C_5 (30.67%). For the "*Average*" case, the quality rating is more evenly distributed across all QoS classes. The finding of our experimental study does imply that the correspondence between different quality ratings and network QoS classes does not always bear a close coherent relationship.

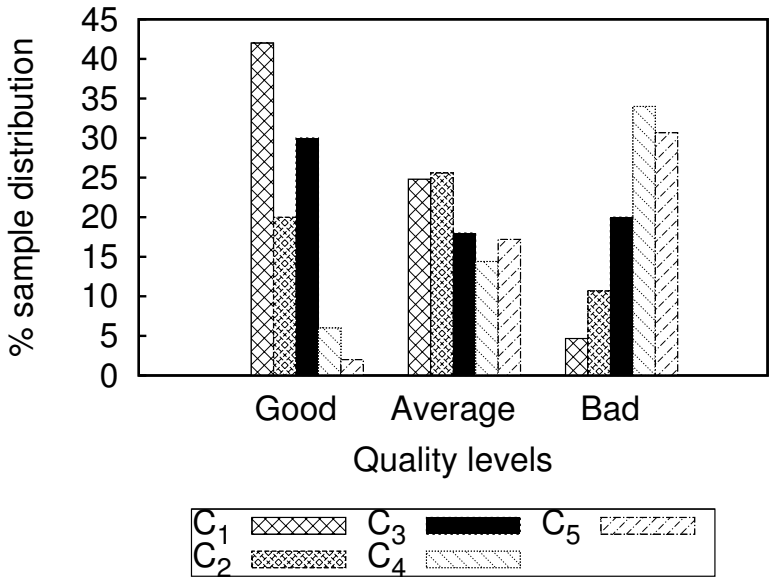

Figure 4.10: Distribution of QoS classes with respect to quality ratings of total samples.

This seems to validate our previous observation that solely QoS-based estimation is not sufficient for QoE prediction as discussed in related work.

4.4.3 Correlation in QoS Classes

From another angle, the results of correlation between the user quality feedback and the quality prediction are given for each QoS class in Table 4.2 using different calculation methods (Pearson, Kendall and Spearman). It is observed that different network QoS classes have varied correlation impact: QoS classes of C_1 and C_5 demonstrate the highest correlation, C_2 and C_4 stay as the intermediate cases, while C_3 has the lowest correlation. The implication we can draw here is that QoE prediction tends to be more accurate if the underlying networking condition is either very good or very bad, and becomes less accurate for mid-range networking conditions. The finding also suggests that we should be more careful to apply QoS-based

estimation. The usage of such methods need to take into account of different QoS contexts (e.g., extreme vs. average conditions).

Table 4.2: Correlation tests of the quality prediction from the testing samples.

Selection	Pearson	Spearman	Kendall
C_1	0.8068	0.7856	0.784
C_2	0.5424	0.5338	0.507
C_3	0.1026	0.0871	0.0752
C_4	0.5572	0.4766	0.4743
C_5	0.7321	0.7319	0.7325
Overall	0.5697	0.5351	0.5608

4.4.4 Implication of Quality Ratings

Finally, we examine the interaction between user quality ratings and the predictions made by the affect analysis framework. We refer to the results of *SVM–5CV* only since it is the classifier of the best performance as observed in Section 4.4.1. The testing samples are plotted in Figure 4.11 which shows the values of user quality rating as well as the prediction. For better visualization, the testing samples are ordered with respect to the user quality ratings from "*Good*", "*Average*", to "*Bad*."

Table 4.3 summarizes the number of user ratings in each quality level and the corresponding prediction accuracy. Some of the observations from the table are as follows: (*a*) though the number of "*Good*" quality ratings are low but it still has a high degree of prediction accuracy (80%), (*b*) the number of "*Bad*" quality ratings are much more in number with a comparable degree of prediction accuracy (77.78%), and (*c*) the "*Average*" quality ratings have a relatively low degree of accuracy (58.67%) compared to the other two but it has the most number of cases. The study shows that quality rating of "*Average*" seems to be more ambiguous than the ratings of "*Good*" and "*Bad*." The reason may be that people tend to rate

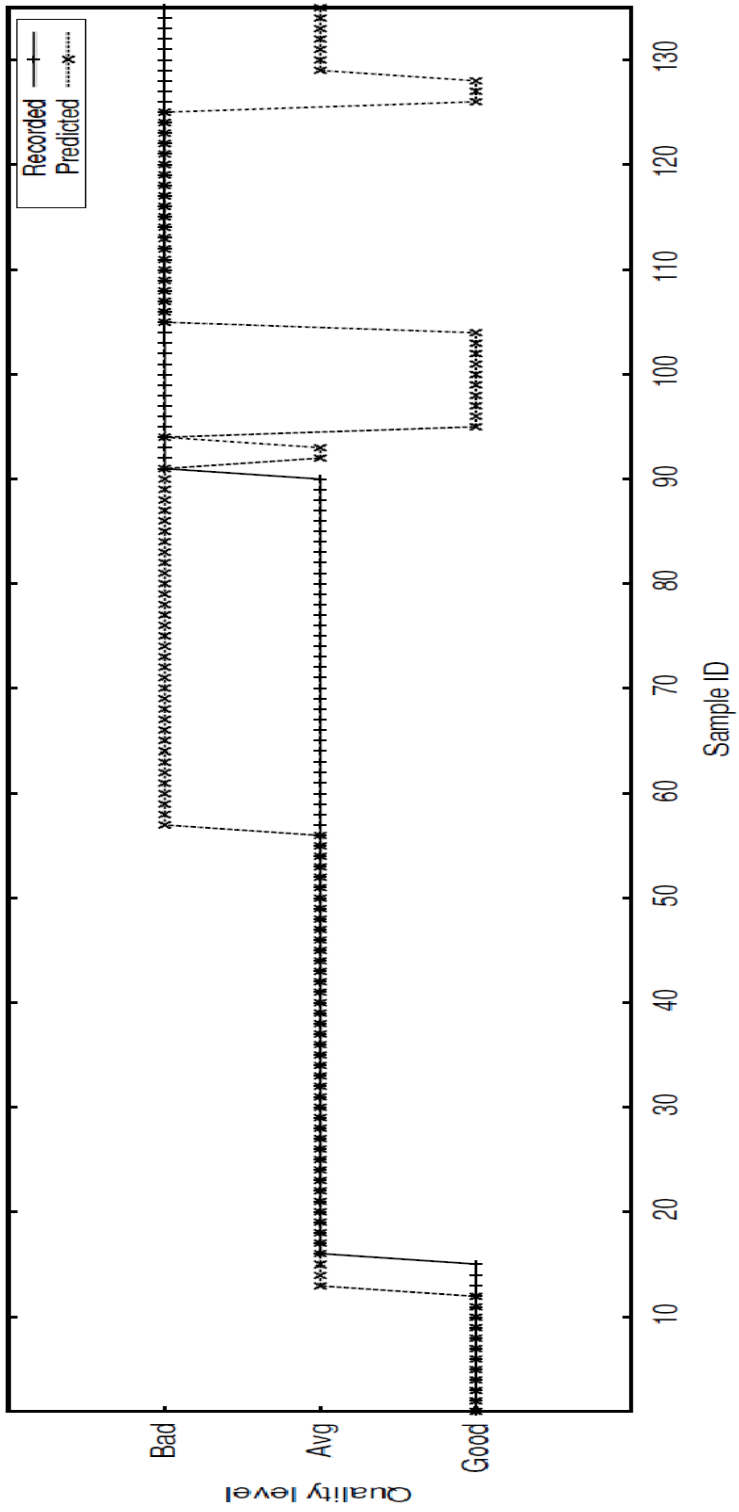

Figure 4.11: Comparison of the distribution of different quality levels for the testing samples with respect to the user feedback and predicted values for *SVM–5CV*.

"*Average*" more frequently across different network conditions (also shown in other studies as well).

Table 4.3: Correlation tests of the quality prediction from the testing samples.

User Quality Rating	number of cases	accuracy (%)
"*Good*"	15	80.0
"*Average*"	75	58.7
"*Bad*"	45	77.8

4.5 Summary

In this chapter, we presented an affect-based framework for QoE evaluation in voice communication systems with human-to-human interaction mediated by a network channel. We also performed experimental studies based on human user for the purpose of evaluation. The purpose of the study is to examine how user affective behavior changes with the communication quality as mediated through different network QoS conditions, and how such changes can be detected and used to predict QoE from the user perspective. We employed supervised machine learning techniques for classification based on SVM, kNN, and the various variants to derive predictions by building training models and testing on a different set. The accumulated evidence supports our initial hypothesis of exploiting affective responses as the predictor of QoE due to its correlation with human cognitive perception.

CHAPTER 5

AFFECTIVE IMAGE ENHANCEMENT

5.1 Introduction

The increasing availability of digital cameras from mobile phones, hand-held devices, to high-end SLR and professional cameras are driving a revolution in digital photography world. Digital cameras are getting increasingly familiar with decreasing cost and expanded functionalities. With this current trend of digital camera revolution, it is increasingly important and desirable to generate high-impact photography that can have a lasting influence on the viewer. This still remains a challenge for the typical user, since they are not familiar with the sophisticated photographic techniques generally employed by the professionals. These specialized techniques to produce impactful photography are termed as correction, adjustment, or enhancement, each of which may have distinctive characteristics in various context, but the general objective is to improve the image.

In a camera processing pipeline, the enhancement functionalities can be applied at either real-time or post-capture. Real-time processing is performed at the time of capturing images (manual or automatic controls) with a wide variety of hardware and software techniques involving signal processing and filter functions. This type of functionalities (such as exposure control, image stabilization, and likewise) are available with advanced devices such as high-end professional cameras, and generally not present in mobile phone cameras. The other type of processing is post-capture, which is our focus in this chapter. Post-processing enhancement techniques are generally supported by automatic or manual editing controls in software systems such as Adobe Photoshop, GIMP, and Microsoft Picture Manager. Many post-processing techniques that improve the quality of images have been proposed where

different correction techniques are used to improve the various properties such as contrast, brightness, color, and tonality. [RJW02], [MI07], [MM08].

Recently, there is an increasing interest of computational/data-driven methodologies for more sophisticated controls such as aesthetics and emotions [JDF+11]. These techniques are attractive due to its computational nature, which allows to automate the enhancement processing function through software control. Psychological studies confirm the strong influence of image and its properties such as color, contrast, textures on the human perception and emotion [VM94], [TMY78], [PLC06]. Affect-based techniques are thus used to explore a variety of multimedia analysis applications such as retrieval [MH10] [LP11], presentation [XK11], summarisation [JJVS09], browsing [ZTH+09], deriving affective film semantics [HX05], visualization [ZHJ+10]. In this chapter, we would like to address the following problem: How do we modify the characteristics (such as color and contrast) of an image to enhance its emotional impact?

The problem is challenging due to the "semantic gap" that presents between the image RGB channel data and the higher level emotional concepts such as fear, happy, or anger. Moreover, emotion is inherently subjective in nature which varies with different contexts and there are no explicit rules guiding them, which makes it an exciting problem to explore. We know, that an artist use different colors and their variations to invoke emotion in paintings. Similarly, a professional photographer carefully tune the temperature and tint of existing colors in a photograph to convey specific emotional feeling. Professionals usually rely on experience and intuition to choose appropriate color compositions, which makes this process tedious and labor-intensive. Thus, it is desirable if such emotion styles can be mathematically formulated in an equation format, which can then be applied to new images for invoking emotional appeal.

Our goal in this chapter is to learn underlying adjustment rules associated with the emotional properties of images from a set of training examples. Given a pair of image before and after adjustment, we would like to discover the underlying mathematical relationships optimally connecting the color-contrast properties between them. Some of the key issue that arises in order to achieve our goal are: (1) How to choose the set of image parameters which will be used for enhancing the emotional appeal of images? (2) How do we capture the notion of enhancement of images in a mathematical fashion to understand the implicit relationship between an original image and its adjusted version? (3) Given an arbitrary image, how to derive the enhancement operations that can be applied to a new/unseen image to enhance its emotional appeal? To seek a reasonable solution to these problems, as a first step we collected ground-truth data and constructed a training database consisting of example images and their different enhanced versions, and asked a set of human participants to mark their preferences. We implemented a simple and effective web-based user-interface for this purpose and sent invitations for participation to a group of subjects. A web-based interface allows to reach out to a larger group of audience in a scalable fashion in comparison to a local desktop interface which is tied with the restriction of lab premises and time. To answer (1) in the above discussion, we selected a set of contrast and color properties which is found to have a high influence on emotional impact of images [JDF+11]. We employ supervised machine learning techniques to understand the relationship between an original image and its adjusted version to address (2), and derive statistical solutions to develop an enhancement function in a mathematical form to address (3). We conduct user studies to evaluate the effectiveness of our approach.

The contributions of this chapter are listed as follows: (a) As per our knowledge, this is one of the first attempts to derive a computational framework for enhancing

the emotional impact of images. (b) We collect ground-truth data for affective enhancement from human participants which is a valuable resource for understanding the underlying relationship between them. (c) We employ a data-driven systematic framework to learn models from training data and derive generalized enhancement functions for arbitrary/unseen images using machine learning and statistical techniques. (d) We derive an objective metric for evaluating the emotional enhancement of images using a color mood space model and used it to test our approach.

5.2 Image Enhancement Framework

The high-level schematic framework of our image enhancement system is shown in Figure 5.1. The first step involves a training phase which essentially builds a database comprising of a set of images and their corresponding metadata, which are rating values. We utilize the International Affective Picture System (IAPS) [LBC08] database which is a standarized benchmark of emotion/affect ratings in images, widely acknowledged by the researchers in psychology-emotion domain and distributed for the purpose of research. The input database images I^{in} are fed into our enhancement processing framework to derive the corresponding output images I^{out}, which are then presented to human particpants for rating metadata $\mathcal{R}$ through a web-based interface. The enhancement processing framework is a set of color-tonal image operations denoted by a vector of enhancement features ϕ. The training database inserts each participant observation $\mathcal{R}_i^u$ and their corresponding enhancement vector ϕ_i^u, for each user u and image I_i in a table. Once the training database is prepared with sufficient samples, we analyze it to learn a statistical model that encodes the relationship between $\mathcal{R}$ and ϕ. Specifically, we cluster the set of original images I^{in} in a number of different groups and then estimate the enhancement vectors of each group using a regression based learning approach followed by an

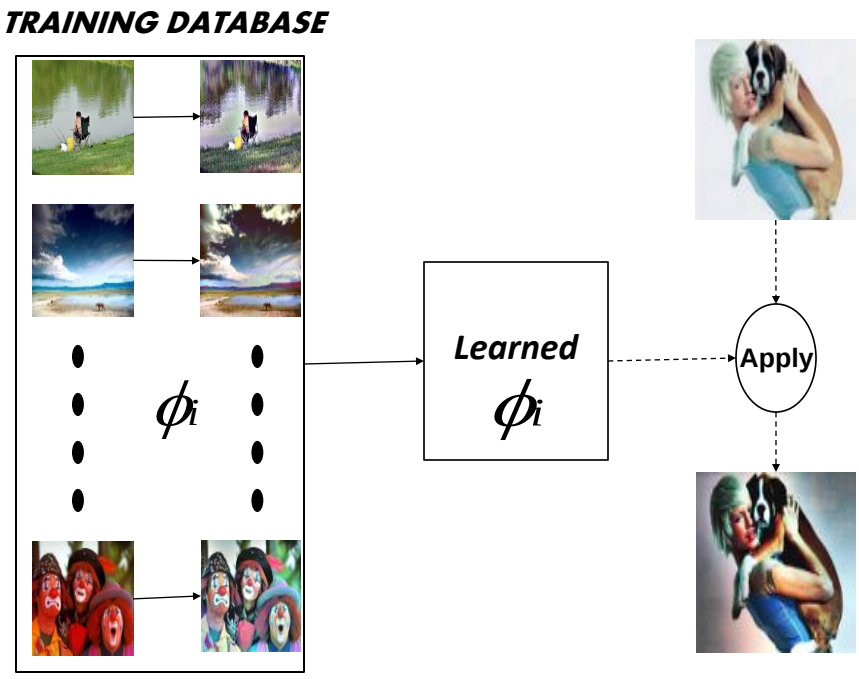

Figure 5.1: Block diagram of the enhancement framework. ϕ is the enhancement vector which is a set of control features based on color and tonal properties.

optimization solver. Once such a statistical model is learned, then we can enhance a new or unobserved image by mapping the image to one of the trained clusters and applying the enhancement operator for that cluster.

5.2.1 Image Database

As already mentioned, we utilize an affective image database known as International Affective Picture System (IAPS) as a ground-truth for building our models. The IAPS currently includes around 1200 images depicting a variety of human experience: joyful, sad, fearful, angry, threatening, attractive, and many more with a virtual world of pictures. The stimuli are standarized on the basis of ratings of pleasure and arousal in a 9-point scale for each dimension. Each image is rated by approximately 100 participants covering a diverse group based on gender, cross-

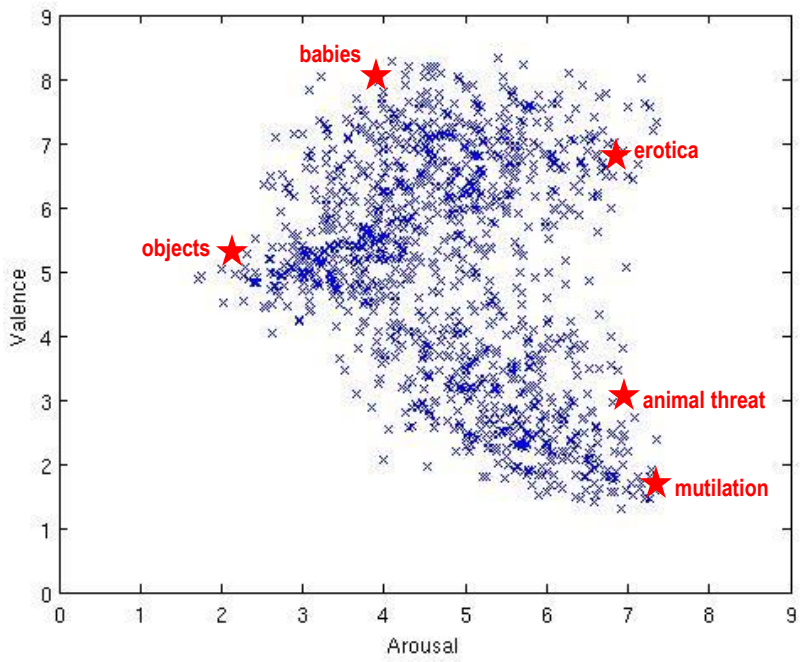

Figure 5.2: Each image in the IAPS is placed in a 2-dimensional affective space on the basis of its mean valence and arousal rating in a 9-point scale with markers at some distinctive image categories to get an insight into the relationship between emotional sematics and the 2d space.

culture, age, and many more. All the images are in color and have a resolution of 1024x768. Each trial is associated with a 5 sec preparatory cue followed by a 6 sec presentation of the to-be-rated image, and then a 15 sec for final ratings of pleasure, arousal. The affective space that is defined by the mean ratings of pleasure and arousal are illustrated in Figure 5.2, by plotting in the 2-dimensional space. Some of the key empirical facts that can be derived are: (1) as pictures are rated as more pleasant or more unpleasant, arousal ratings increase as well, and (2) pictures that are rated as neutral tend to be rated low in arousal. We manually selected a representative subset from the original set, since it requires a lot of human resource to collect sufficient amount of training data for the total set of 1000 images.

5.2.2 User Interface

We implemented a web-based user interface for collection of training data related to the affective enhancement of images, specifically the V_i^u, A_i^u columns in the database table (discussed later in Section 5.4). Our user interface is kept simple and intuitive so that an ordinary users can easily enter their ratings as shown in Figure 5.3. Engaging professional photographers for manual image adjustment to build training data appears to be an attractive option, but it is more expensive with high resource consumption for even collecting a small number of samples. Thus, our target set of participants are ordinary users from university students or any other field, without any prior knowledge of professional photography. The possible options for designing a rating procedure to capture emotional enhancement are manual or decision. In manual-based procedure, the participant is provided with various image edit controls so that the subject can perform the enhancement based on his/her judgement and submit. This process is geared for professional photographers (not easy for ordinary users to enhance emotional impact with edit controls) and local desktop setting which does not fit our design choices. The decision-based procedure generally involves in a binary yes/no or good/bad type of opinion which is very simple for the ordinary users but carry less information.

We identified a simple procedure with higher meaningful information than the decision-based, using a 2D coordinate grid which is located in the right side of the user interface as shown in Figure 5.3. The 2D coordinate system is calibrated with valence-arousal dimension in a 9-point scale which is similar to the ratings in IAPS dataset. The user just needs to mark a point in the 2D affective space with the help of a mouse which will indicate its rating for emotional impact with respect to the image displayed on the left side of the interface. We illustrate the relationship

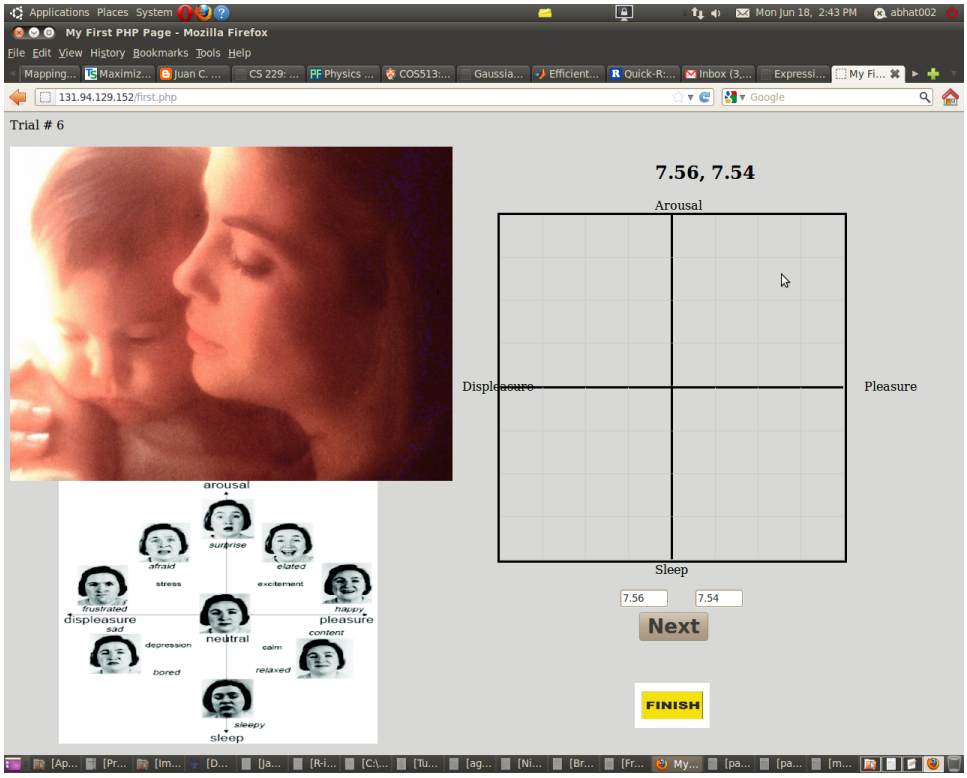

Figure 5.3: The web-based user interface for collecting data of affective enhancement for the images displayed on the left of the panel. This can be found at http://131.94.129.152/start.php

between image and the corresponding 2D affective spatial coordinates using examples from the IAPS dataset in the web interface before the start of the final rating procedure. This will provide sufficient insight to the participant regarding the spatial position in the affective space. During the rating procedure we allow multiple clicks but the coordinate valence-arousal values of the final click are only submitted. As soon as the partcipant clicks on a position in the 2D space, the system will display the value of the clicked coordinate positions on the top of the grid as shown in Figure 5.3 which allows for better judgement by the subject. A reference image is always kept in the bottom-left portion of the user interface as shown in Figure 5.3 which provides the position of some higher-level emotional concepts such as “fear”, “anger”, “sad”, and likewise in the 2D affective space. The images on the

top-left portion of the interface are various enhanced versions of the base training set from the IAPS dataset, and they are generated dynamically and displayed after each submit click by the participant. The various relevant information such as image metadata, enhancement parameters, and valence-arousal emotional ratings are stored in the training database for data analysis. The web-based interface can be found at http://131.94.129.152/start.php and we welcome any reader to submit their emotional ratings which will enable us for more robust data collection and distribution to the community.

We forwarded voluntary requests for participation through emails containing the web-link. The participant group is entirely from the university students of various level (undergrad to graduate) and no one is an expert of photograhy as per our knowledge. We don't have any control over the user session length and the user can stop anytime, but still we requested each user before the session started to submit at least 50 each. We collected training data from 26 participants in this procedure. The average age of the participants is around 18-40 years. The size of the training set image stimulus is set to be 166 which was selected entirely from the IAPS database. The image selection and the adjustment function is completely dynamically controlled by the webserver. The total number of training samples collected in the database is 1073 which is roughly around 40 per user in average.

5.3 Enhancement Channel

For designing the enhancement channel, we first need to identify the vital image parameters which are known to influence human affect. There are already ample evidences in the literature where it is found that images can influence human emotions to a considerable extent [VM94]. Moreover, recently there are many pro-

posals to indicate that different image properties such as color, contrast, saturation, birghtness, and similar attributes [BB03, SL10, WH08, WY05, WnYlSm06, WZW05, YvGR$^+$08] contribute to invoke human emotions, and moreover these properties can be exploited to address retrieval [MH10], summarisation [JJVS09] and presentation [XK11].

After carefully studying the literature, we have found evidence of various features that are related to emotion (such as color, texture, composition, content, and likewise) but the two most widely used parameters are contrast and color [MH10]. Following our observation, we design the enhancement channel based on the following parameters: power curve and S-curve shaping for contrast adjustment, color tint and temperature for color adaptation. Although the framework can be extended with more effective parameters, we limit the number of parameters primarily to limit the complexity since the computational space will increase exponentially with the number of parameters. Figure 5.4 shows the entire block of components that are incorporated in our enhancement channel. The initial preprocessing step involves the linearization of the nonlinear RGB space with gamma curves applied to the input image I_{in}. This is followed by another preprocessing step of auto-enhancement of the images, which is intended for normalizing the difference in image qualities and bring them within comparable limits. Then the channel applies two contrast adjustment functions i.e., power curve and S-curve shaping, followed by the color operators i.e., temperature and tint adaptation. These are the core operations and the enhancement vector is derived from these transformations. Finally, a postprocessing step of inverse linearization is performed to revert back to the nonlinear space to produce the final enhanced image I_{out}.

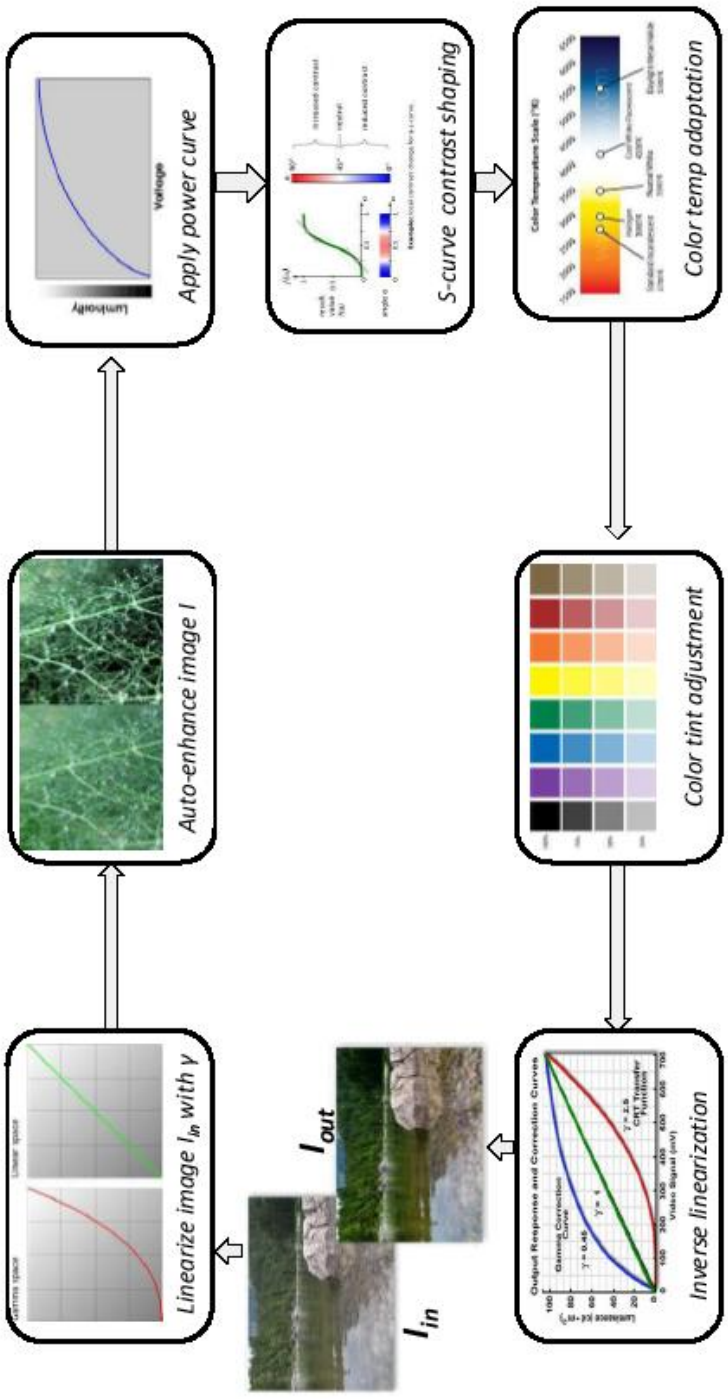

Figure 5.4: Block diagram of the enhancement framework. ϕ is the enhancement vector which is a set of control features based on color and tonal properties.

5.3.1 Linearization

Linearization is generally followed to minimize the variation between device hardware since different cameras may have been calibrated with different intrinsic properties. We use the simple power curve for the process of linearization as follows:

$$c = c^{\gamma} \tag{5.1}$$

where c is the normalized values of the R,G,B channel and $\gamma = 2.2$ [KKL10]. The inverse linearization process or reverting back to the nonlinear space which is the last component block of the enhancement channel in Figure 5.4 is the same operation but with $\gamma = \frac{1}{2.2} = 0.45$.

5.3.2 Auto-Correction

We initiate an auto-correction step to factor out the differences of qualities among the various images in the IAPS database. This will allow to correct the degraded images if present in the database to a more acceptable level upon which our enhancement operator will function. We follow a generic methodology consisting of white balance and contrast stretch which are the two most simple operations for correcting an image.

White balance is the process of removing unrealistic color casts from images which are mostly generated from different acquisition condition with varying illuminations. The human visual system is generally able to render the perceived colors of objects almost independent of illuminations by a phenomenon known as Color Constancy, which is mimicked in the digital cameras by a process of Automatic White Balancing to suppress unrealistic colors. From a computational perspective, white balance is a two-step procedure: the illuminant is estimated, and the image colors are then corrected on the basis of the estimate. For estimating the illuminant, we follow the

Gray World algorithm which assumes the average surface color in a scene is gray and the shift from gray of the measured averages on the three channels corresponds to the color of the illuminant [CKK11]. We find the average values of the RGB color components and use their average to find an overall gray value for the image. Each color component is then scaled by a factor ratio of the grayvalue to the approriate average of each color component.

Contrast stretch is the process of evenly distributing the pixel intensities within a defined span which is essentially used to stretch the compact region of the pixel values in the image histogram. We fix the span at 0.5% on the darker side and 1% on the brighter side followed by linear stretching of the original values to the new range with a shift and scale operation applied to all the color bands [CKK11].

5.3.3 Contrast Shaping

We apply two transformations for contrast adjustment i.e., power curve and S-curve as shown in Figure 5.4 as follows:

Power curve is similar to the gamma curve that was employed in the linearization process but the value of γ is not fixed at 2.2 as before, but rather is adjusted at different levels which will allow to learn the affective preference.

$$i_{out} = i_{in}^{\gamma} \tag{5.2}$$

where i_{in} and i_{out} are the normalized input and output intensities of each pixel.

S-curve is the most common tool for adjusting contrast and is found in almost all commercial products such as Adobe Photoshop. The contrast of the tonal curve can be increased or decreased to varying degrees by applying different shaping pa-

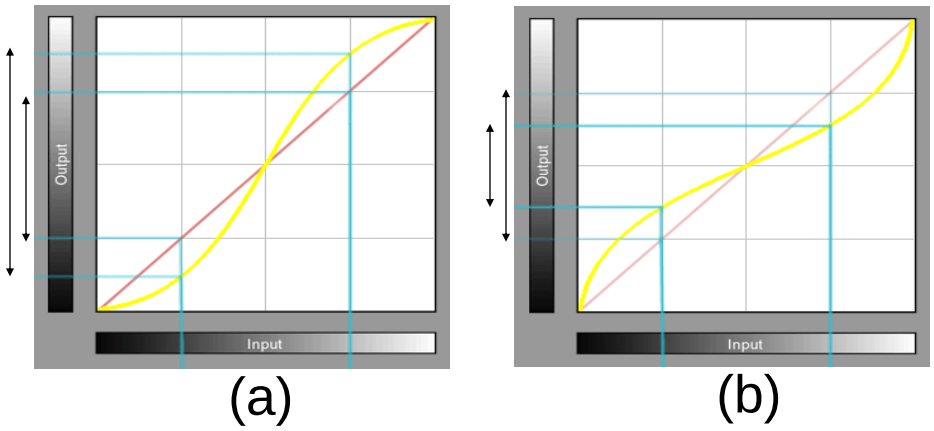

Figure 5.5: (a) shows increase of contrast and (b) shows decrease of contrast by varying the shaping parameters of S-curve.

rameters for the S-curve as shown in Figure 5.5. The formula for S-curve shaping is as follows:

$$i_{out} = \begin{cases} q - q(1 - \frac{i_{in}}{q})^{\alpha} & \text{if } i_{in} \leq q \\ q + (1 - q)(\frac{i_{in} - q}{1 - q})^{\alpha} & \text{otherwise} \end{cases} \tag{5.3}$$

where i_{in} and i_{out} are the normalized input and output intensities of each pixel; q and α are the adjustment parameters of the S-curve which can be described as follows: Refering to Figure 5.5, q will be the mid-point of the curve where the curve and the diagonal intersects, and $\frac{\alpha}{2}$ is the maximum deviation of the curve from the straight diagonal line.

5.3.4 Color Temperature

Color temperature have been found to have a high influence on emotion whereby the "warmthness" or "coolness" of an image can indicate the nature of affect generated by the picture [VM94]. Many professional photographers adjust color temperature in pictures using commercial software such as Adobe Photoshop to generate desired effects of emotion. Warmer temperatures give the images a sense of warmth and

coziness, while cooler temperatures can make images seem cold and harsh [VM94]. The color temperature of an image can be determined by comparings its chromaticity values with that of an ideal black-body radiator. The scale of color temperature varies from 1900 Kelvin to almost around 10,000 Kelvin and the standard is calibrated at 6,500 Kelvin which sginifies daylight settings. Color temperatures over 5,000 Kelvin are known as cool colors (blueish white) while lower color temperatures around 2,500 3,000 Kelvin are known as warm colors (yellowish white to red) as illustrated in Figure 5.6. The computation algorithm estimates the correlated color temperature, T_c by way of interpolarions from lookup tables and charts using the popular Robertson's method [ROB68]. The first step involves in converting the color space from RGB to CIE XYZ, and then the chromaticity values (u, v) are obtained using the CIE 1931 coordinate system as follows:

$$u = \frac{4x}{-2x + 12y + 3} \qquad v = \frac{6y}{-2x + 12y + 3} \tag{5.4}$$

Then a search is initiated through the set of standard isotherms from the lookup table to find the two tightest adjacent lines between which the test chromaticity i.e., (u, v) lies. Then, the correlated color temperature or T_c can be calculated as follows:

$$\frac{1}{T_c} = \frac{1}{T_i} + \frac{d_i}{d_i - d_{i+1}} \left(\frac{1}{T_{i+1}} - \frac{1}{T_i} \right) \tag{5.5}$$

The distance between the test point (u_T, v_T) and the i^{th} isotherm (u_i, v_i) is given by:

$$d_i = \frac{(v_T - v_i) - m_i(u_T - u_i)}{\sqrt{1 + m_i^2}} \tag{5.6}$$

where m_i is the slope of the i^{th} isotherm computed by $m_i = -\frac{1}{l_i}$ and l_i is the slope of the locus at (u_i, v_i).

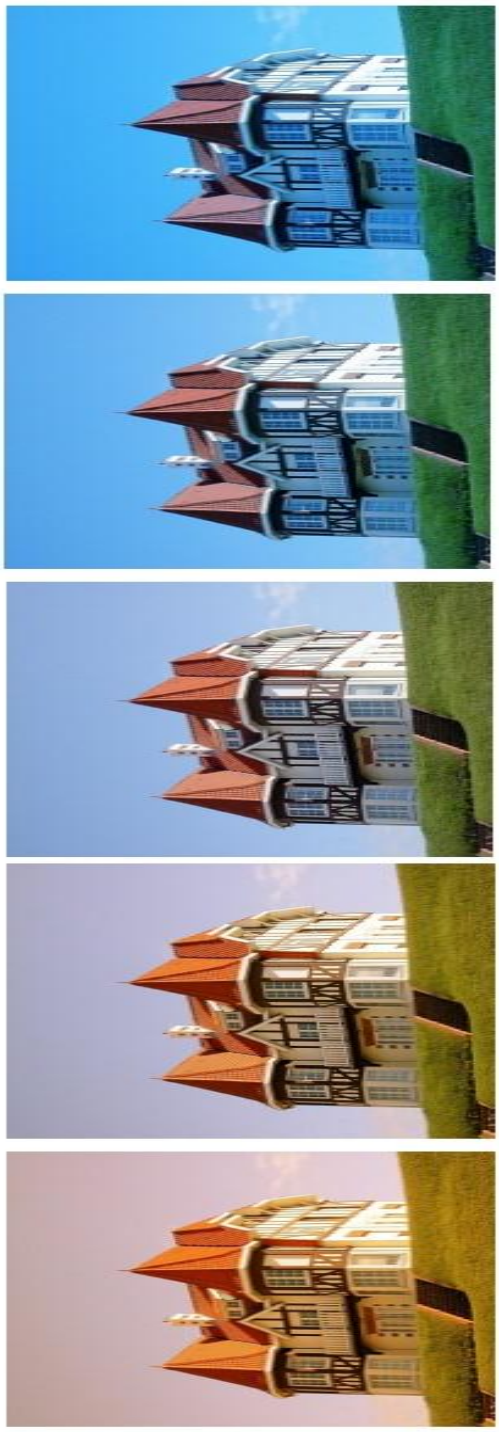

Figure 5.6: Block of image showing the variation of the different color temperatures on a single image. The color temperature generally varies in a blue-yellow axis. The middle image is the original one and moving towards the left increases the color temperature (yellow end) and towards the right decreases the color temperature (blue end).

5.3.5 Color Tint

Color tint can be viewed as orthogonal to color temperature whereby the control changes along the green-magenta axis which are at two complementary positions in the color wheel as illustrated in Figure 5.7. To adjust color tint, we utilize the hue channel by converting to the HSV color space. Now, we need to desgin an adjsutment function which can change the tint of the hue channel by varying degrees. The strategy involves in shifting the hue of each pixel towards the desired tint i.e., along the green-magenta axis. Instead of a simple linear shift of the hues which may produce unwanted color artifacts, we apply a technique adapted from [COSG+06] using Gaussian function whereby the overall tone of the image is not greatly affected. Let the hue of pixel p is denoted by h_p and the tint hue of the sector associated with green-magenta axis to be h_t, then we can derive the value of shifted hue h'_p as follows:

$$h'_p = h_t + \frac{w}{2}\Big(1 - G_\sigma(\|h_p - h_t\|)\Big) \tag{5.7}$$

where w is the arc-width of the green-magenta sector and G_σ is the normalized Gaussian function (such that $G_\sigma(x) \in (0, 1]$) with mean 0 and standard deviation σ; the hue distance $\|.\|$ is the arc-length distance on the hue wheel measured in radians. The width of the Gaussian σ is a user-defined parameter that may vary between zero and w, where larger values of σ will create hue concentrations near the sector center and smaller values will lead to concentrations near the sector boundaries. In our implementation, we use the mid-point $\sigma = w/2$. Thus, the adjustment parameter for color tint or T is the value of w which can be varied to achieved different effects of tint. Thus the final enhancement vector can be found to be $\phi =< \gamma, q, \alpha, K, T >$ where γ is from power curve, k, α from S-curve, K from color temperature, and T from color tint.

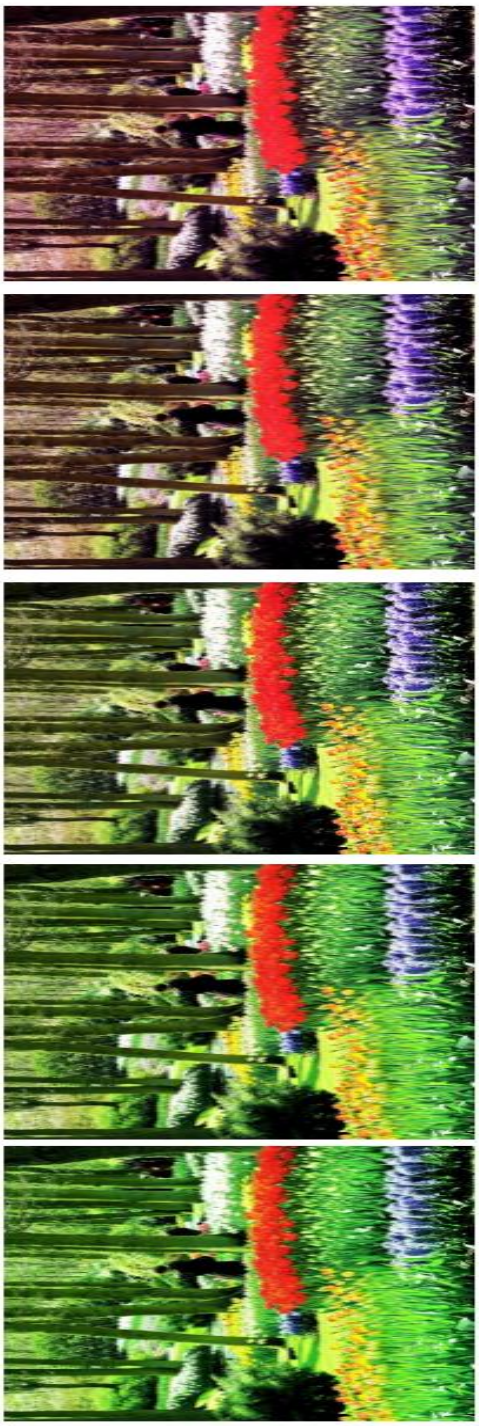

Figure 5.7: Block of image showing the variation of the different color tints on a single image. The color tint generally varies in a green-magenta axis. The middle image is the original one and moving towards the left increases the color tint (green end) and towards the right decreases the color tint (magenta end)

5.4 Learning Framework

In this section we will describe our learning framework which is basically involved in deducing a best set of enhancement parameters (with respect to its match with the user's preference) which will be later used for applying to unseen/test images. Till now, we have observed the enhancement feature vector covering various adjustments of image, which is utilized to gather training data through the web-based interface. The input to the learning framework will be a table of training data where each row is denoted as follows: $< I_i, \phi_i^u, V_i^{gt}, A_i^{gt}, V_i^u, A_i^u >$ for user u and training/seen image I_i as described before, and the expected output of the best set of enhancement vector is $\phi' = \langle \gamma', k', \alpha', K', T' \rangle$ for test/unseen image I_t.

In order to solve this objective, we need to resolve the following problems: (1) How to find an enhancement vector ϕ for an unseen/test image? (given the fact, that the training samples which are labeled with seen/training images I_i and the new/test images I_t are distinct subsets.), (2) How to define the degree of enhancement in each training sample or specifically, how to define the relationship strength between IAPS ground-truth (V_i^{gt}, A_i^{gt}) and user-rated values of (V_i^u, A_i^u) for each image I_i since as per our design interface each sample of the training database will have varying degrees of enhancement? (3) How to learn the relationship between the enhancement vector E and the valence-arousal based emotional values (V, A)? (4) How to compute the best set of enhancement vector ϕ' from the training data samples? We take a modular approach to address each of the above concerns as follows: k-means clustering for (1), a mapping function for (2), regression-based learning for (3), and an optimization solver for (4) as shown in Figure 5.8.

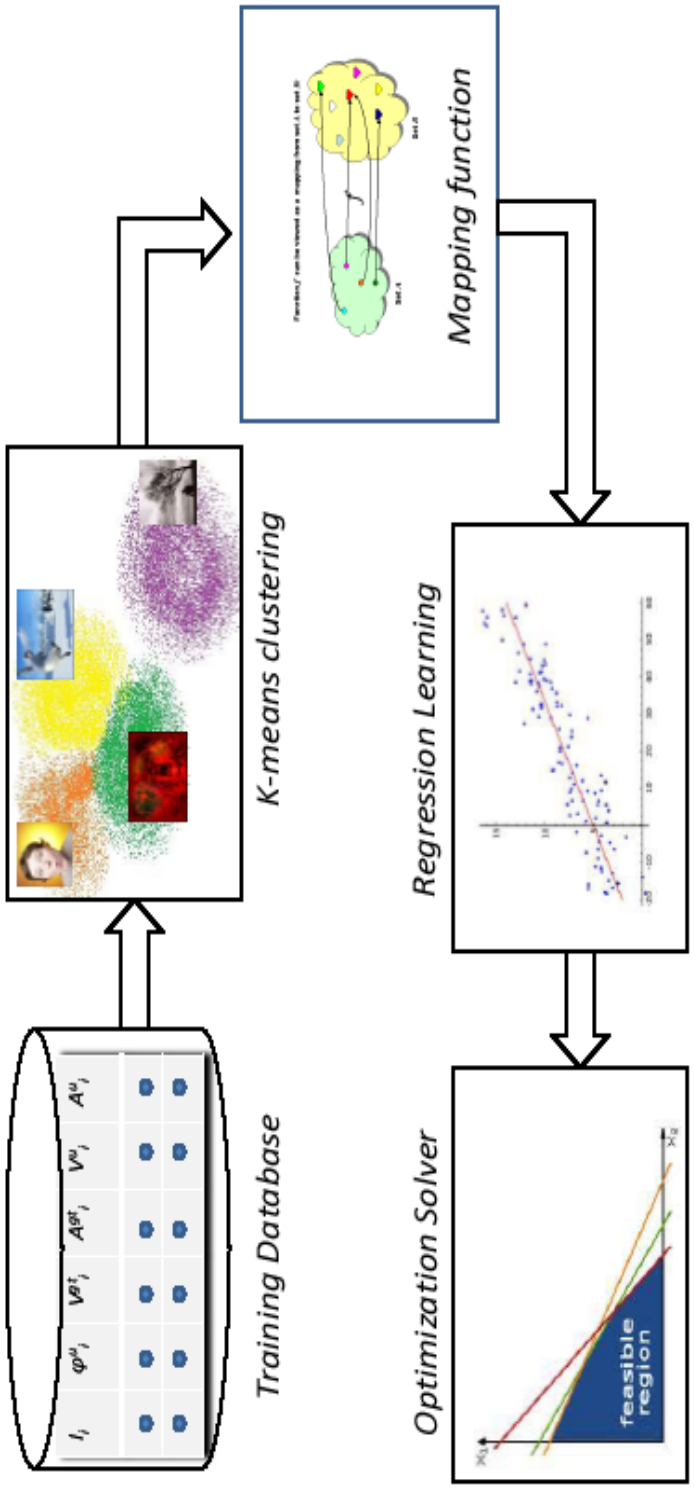

Figure 5.8: Block diagram of the learning framework where the objective is to derive a best set of enhancement vector ϕ' from a training database of samples.

5.4.1 Clustering

This is the first component in the learning framework where the samples from the training database are fed to a k-means clustering algorithm for processing. Note, that the clustering is performed on the ground-truth values from the IAPS dataset (V_i^{gt}, A_i^{gt}). The rationale behind the adoption of clustering is as follows: IAPS images are tagged with user-rated valence-arousal ratings in a 2-dimensional scale, but we wish to divide them into separate classes (in line with discrete theory where the emotions are classified as happy, anger, sad, fear, disgust and surprise) and then derive an enhancement vector for each of those classes. This observation follows from previous research where it has been shown that each discrete classes of emotion possess distinctive color/contrast/tonal characteristics of image which is exploited for the affect-based retrieval of images [LP11]. Thus, our hypothesis is to enhance the emotional characteristics of image from each class with a distinctive function, so that the emotional semantics (in the form of color/contrast properties) can be embedded in the images effectively. Another way, is to explore the enhancement in the valence-arousal dimensions, whereby we can learn enhancement functions for valence and arousal separately. In such a case, we only need to omit this component of clustering and the rest of the framework can be used to attain the objective. Enhancing an image in terms of valence and arousal is an interesting direction which has not being explored before, and we intend to explore this area in future.

The k-means clustering is a popular unsupervised learning algorithm where the data points are grouped into a user-defined (K) number of clusters. Given a set of training examples from the database $\{x^{(1)}, x^{(2)}, .., x^{(m)}\}$ where each $x^{(j)}$ is a valence-arousal tuple $\langle V_{i,(j)}^{gt}, A_{i,(j)}^{gt} \rangle$ and $j \rightarrow 1$, m is the number of samples in the database, the algorithm starts by initializing cluster centroids $\mu_1, \mu_2, .., \mu_K$ randomly. Then at each

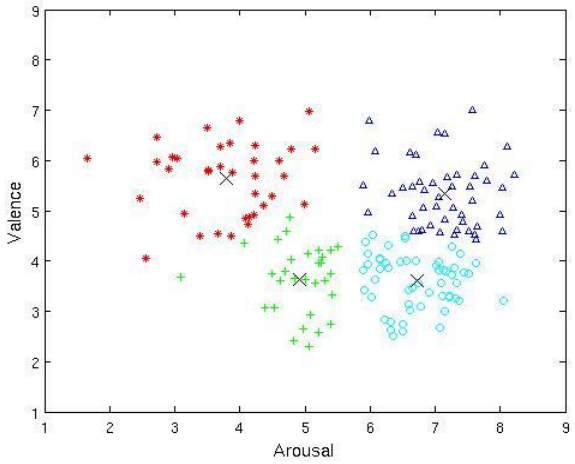

Figure 5.9: The clustered images from the IAPS database in the affective space and the respective centroids marked with a X.

iteration step till convergence, it assigns training example $x^{(j)}$ to the closest cluster centroid μ_k by minimizing the distance between them, followed by moving each cluster centroid μ_k to the mean of the member points assigned to it as follows [CH67]:

$$\forall i: \quad c^{(k)} := arg\ min_k\ \|x^{(i)} - \mu_k\|^2 \tag{5.8}$$

$$\forall k: \quad \mu_k := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = k\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = k\}} \tag{5.9}$$

where $c^{(j)}$ is the cluster membership of sample j, $1\{.\}$ is a membership function between sample i and cluster j evaluating 1 if present in cluster or zero otherwise, and $\|.\|$ is a sum of squared distance function between training sample $x^{(i)}$ and the assigned cluster centroid μ_k. After k-means clustering, each sample in the training database will be: $\langle I_i, \phi_i^u, V_i^{gt}, A_i^{gt}, V_i^u, A_i^u, c_i^k \rangle$ where the new field is c_i^k is the cluster membership of image I_i in cluster k. In our framework, we assume K=4, which can intuitively correspond to major affect categories (i.e., happy, sad, anger, and fear [MH10]). Figure 5.9 shows the plot corresponding to the deduced clusters on a subset of 167 images from the IAPS dataset.

5.4.2 Mapping Function

Now, we try to derive a metric that can quantify the degree of enhancement (D) of each sample from the training database. Formally, the amount of change from $VA^{gt}_{(j)} = \langle V^{gt}_{(j)}, A^{gt}_{(j)} \rangle$ to $VA^{u}_{(j)} = \langle V^{u}_{(j)}, A^{u}_{(j)} \rangle$ for sample j of ground-truth rated V-A values in IAPS database with enhancement rated V-A values in training database. One naive way is to use a metric of $D_{(j)} = \|VA^{u}_{(j)} - VA^{gt}_{(j)}\|$ where $\|.\|$ is an Euclidean distance function, but this is not a good choice since it does not consider the direction between two data points. In reference to the ground-truth point $VA^{gt}_{(j)}$ as the origin in the affective space, $VA^{u}_{(j)}$ can be placed in four quadrants i.e., first quadrant $[V^+, A^+]$ when $V^{u}_{(j)} > V^{gt}_{(j)}$ && $A^{u}_{(j)} > A^{gt}_{(j)}$, second quadrant $[V^+, A^-]$ when $V^{u}_{(j)} > V^{gt}_{(j)}$ && $A^{u}_{(j)} < A^{gt}_{(j)}$, and likewise follows. Here, we would like to take another hypothesis as follows: the degree of enhancement D should also consider the quadrant relationship between $VA^{gt}_{(j)}$ to $VA^{u}_{(j)}$ unlike the naive approach discussed before. The rationale behind this hypothesis is intuitive: enhancement to $[V^+, A^+]$ should have higher relevance than $[V^+, A^-]/[V^-, A^+]$, which should be higher than $[V^-, A^-]$, since both the valence and arousal values are increased in the first case, either valence/arousal is increased and the other one decreased in the second case, and both the valence and arousal are decreased in the last case. So, the enhancement degree D should consider three levels of relevance based on the directional aspect of the comparison, apart from the positional aspect defined by $\|.\|$ as discussed above.

Now, we can easily derive D using simple scaling and shifting techniques as follows: Assuming, we want the value of D to be distributed in [0,1], a, b, c are the starting point for each quadrant mapping such that $0 \leq a < b < c < 1$, the diagonal end-point 2d coordinate of each quadrant opposite to origin be Z^q (where Z^I for

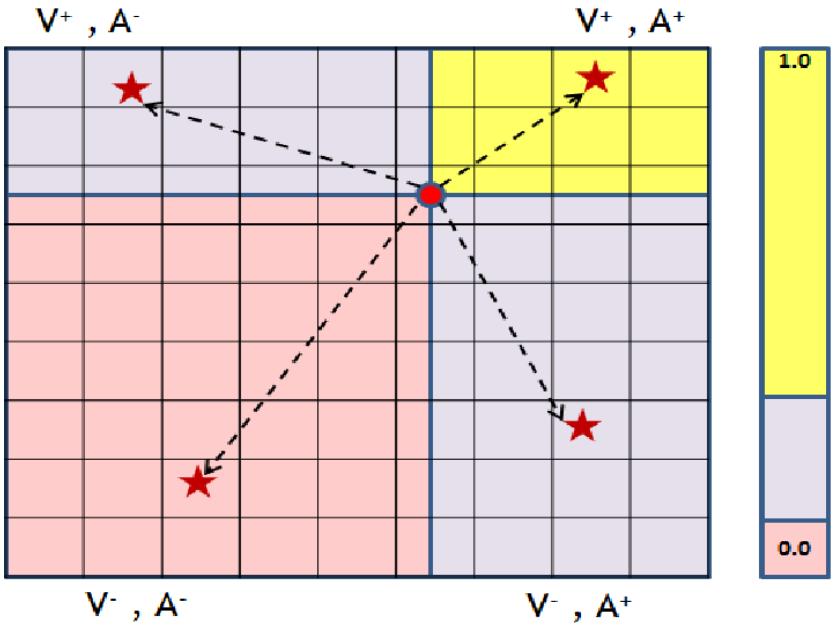

Figure 5.10: The pictorial depiction of the mapping function using scaling and shifting techniques from the 2D affective space on the left to the 0.0-1.0 linear scale on the right to capture the degree of affective enhancement.

$[V^+, A^+]$ is (9.0,9.0), Z^{II} is (0.0,9.0), and likewise since we already know that the IAPS affective 2d space is calibrated in a 9-point scale):

$$R_{(j)} = \frac{\|VA^u_{(j)} - VA^{gt}_{(j)}\|}{\|Z^q_{(j)} - VA^{gt}_{(j)}\|} \tag{5.10}$$

$$D_{(j)} = s + \frac{R_{(j)}}{e - s} \tag{5.11}$$

where [s=a;e=b] is the segment interval for enhancement quadrant $[V^-, A^-]$, [s=b;e=c] for $[V^+, A^-]/[V^-, A^+]$, and [s=c;e=1.0] for $[V^+, A^+]$ for each training sample j as shown in Figure 5.10. In our experimental study, we heuristically assume a=0, b=1.0, c=3.0 which classifies varying degree of relevance for different enhancement quadrants though various other strategies can also be supported in our framework. Thus, each training sample can be denoted as $\langle I_i, \phi^u_i, D, c^k_i \rangle$ (some fields are dropped since they are not required for further processing) where we quantified the degree of enhancement by D. Design choices different from our hypothesis can also be easily

integrated in the framework by little modifications of the mapping function. One such choice of importance could be based on voting, where the number of samples in each quadarant is calculated for each image, and then the importance is computed in proportion to the majority share. If there are more number of training samples for image I_i with enhancement to $[V^+, A^-]$ (majority of user's prefer enhancement of I_i by increasing valence and decreasing arousal), then highest importance is given for this quadrant mapping, unlike our hypothesis where we assume a fixed bias.

5.4.3 Regression Learning

The goal of this component is to learn a relationship function between the enhancement vector $\phi^{(k)} =< \gamma^{(k)}, q^{(k)}, \alpha^{(k)}, K^{(k)}, T^{(k)} >$ and $D^{(k)}$ for each cluster k. First, we divide the training samples based on its cluster membership $c^k_{(j)}$, and then we learn a regression function for each cluster. The rationale for choosing a regression learning approach is due to its suitability for deriving a function of input variables to correlate with the output variable. Regression is a supervised learning algorithm which intially starts with a representation of function/hypothesis [Fis22]:

$$h(\phi) = \sum_{i=0}^{n} \theta^i \phi^i = \theta^T \phi \tag{5.12}$$

where θ_i are the regression coefficents parameterizing the space of linear function mapping from ϕ, n=5 in our case where $\phi = \langle \gamma, q, \alpha, K, T \rangle$ and $x_0 = 1$ which makes θ_0 the intercept term. We compute the coefficient vector θ^T using a gradient descent approach to solve the least mean squares optimization problem of minimizing the error term as per the following update rule:

$$\forall i \rightarrow 1 \; to \; 5 : \quad \theta_i := \theta_i + \alpha \sum_{j=1}^{m} (D_{(j)} - h(\phi_{(j)})) \phi^i_{(j)} \tag{5.13}$$

where α is the learning rate, m is the number of training samples, and the above process is repeated till convergence. After termination, a set of five regression coef-

ficients and an intercept term θ^T are derived, and the process is repeated for each cluster $c^{(k)}$ to obtain a set-of-set of regression parameters $\{\theta^T\}^{(k)}$.

5.4.4 Optimization Solver

Now, we proceed to the last component of the learning framework which involves to solve an optimization problem of deriving a set of enhancement vector ϕ^k for each cluster k by maximization of $D^{(k)}$. This problem involves in maximizing the function learned in Eq: 5.12 so that we can derive a best set of enhancement vector $\boldsymbol{\phi}'$ (our initial objective when we started with the design of the learning framework as discussed before) which will statistically provide the maximum degree of enhancement (or highest value of $D^{(k)}$). We formulate the problem as a methodology for linear programming (LP) as follows [BTN00]:

$$\textbf{maximize}: \quad \boldsymbol{\theta}^T\boldsymbol{\phi} + \theta_0 \tag{5.14}$$

$$\textbf{subject to}: \quad 0 \preceq \boldsymbol{\phi} \preceq h \tag{5.15}$$

$$\boldsymbol{\phi} \succeq \boldsymbol{\phi}_{(j),max} \tag{5.16}$$

where $\boldsymbol{\phi}$ is the optimization vector, $\boldsymbol{\theta}$ is a vector of known coefficients, h is the upper bound of $\boldsymbol{\phi}$ variables which is 1 in our case since we have them normalized, and $'\preceq'$, $'\succeq'$ denotes elementwise inequality. $\boldsymbol{\phi}_{(j),max}$ is one set of $\boldsymbol{\phi}_{(j)}$ values for a training sample j with a maximum value of $D_{(j)}$ selected from the training database since we want to derive a best set of vector $\boldsymbol{\phi}'$ which is explicitly more than the current maximum $\boldsymbol{\phi}_{(j),max}$ from the database. We use the CVX [GB08] optimization solver which exploits the widely known interior-points method to solve the linear problem by moving through the interior of the feasible region unlike traversing the edges between vertices on a polyhedral set typically performed in the simplex algorithm. We repeat the process to derive the best set of enhancement vector $\boldsymbol{\phi}'_{(k)}$ for each

cluster k. Thus, we are now ready with a separate function of enhancement variables $\langle \gamma'_{(i)}, k'_{(i)}, \alpha'_{(i)}, K'_{(i)}, T'_{(i)} \rangle$ for each cluster i which can be applied to any unseen/test images for affective enhancement.

5.4.5 Applying Enhancement

We are interested in predicting enhancement preference for test image I_{test} for arbitrary/unseen source. The key observation lies in the fact that, once we know the cluster membership of the new image, then just select the corresponding enhancement vector $\langle \gamma'_c, k'_c, \alpha'_c, K'_c, T'_c \rangle$ to find the individual parameters and then adjust the contrast and color based on the values accordingly. The adjustment procedure is described in Sec: 5.3 with fixed values of the derived best set of enhancement vector. Thus, the problem really reduces down to establishing the cluster membership of the new image I_{test}. Images with V-A ratings in a 9-point scale from IAPS database or some other sources can be easily integrated in our framework. We find the closest cluster centroid μ_k and assign the image to that cluster, thereby applying the enhancement operator from the corresponding cluster.

The scenario is a little bit complicated for unrated new images where we need to somehow identify the cluster membership of the image. One solution is to employ a classifier such as SVM to learn a model of image features and its associated class of cluster membership from the training database samples (traning sample are grouped into various clusters in Section 5.4.1). A single multi-class classifier or mutiple binary-class classifiers can be designed for this purpose. For multiple binary-class classifers, an aggregation technique is required to derive a single prediction of cluster membership. The selection of image features is a critical issue and is still an open problem, but recent research provides considerable success with high classifier accuracy for affective retrieval of images [LP11]. Some of the possible

image features are: saturation/brightness contrast, hue statistics, colorfulness, Itten contrasts, wavelet features, level of detail, dynamics, human faces and skins [MH10]. The classifier predicts the cluster membership and the rest of the procedure follows as discussed before. We can also employ regression to learn the valence and arousal ratings with image features as input variables from training database samples. The learned model predict the V-A values for unseen/new images which can then be mapped to a cluster based on the distance to the centroids.

5.5 Experimental Evaluation

Evaluating the quality of affective enhancement is difficult, since there is no universal agreement on the perceptual metric. Its nature is also more subjective and so developing objective evaluation models is challenging. In this section we start with subjective user studies in Section 5.5.1, followed by objective evaluation methodologies in Section 5.5.2, and finally we illustrate some images in the form of examples in Section 5.5.3.

5.5.1 User Study

We have carried out a user study to evaluate the effectiveness of our enhancement framework with respect to a set of testing images. A simple web-based interface is built for this purpose as shown in Figure 5.11. The interface facilitates pairwise comparison between the original and the adjusted images. The user feedback mechanism is provided with three options: left (indicates the original invokes more emotion than the adjusted i.e., negative case for our method), right (adjusted image is more emotional than the original i.e., success of our method), and neutral (within comparable limits i.e., enhancement does not produce significant gain). The users

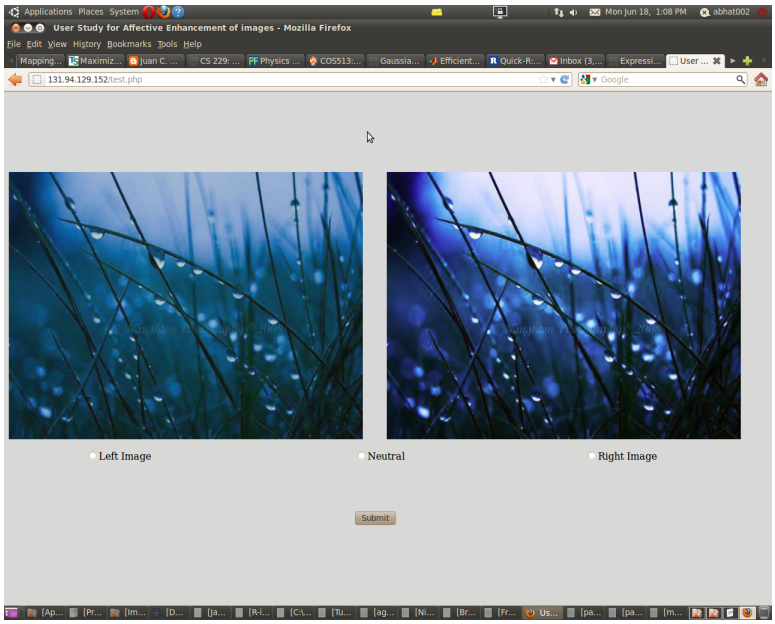

Figure 5.11: The web-based interface for user study which allows the subject to indicate its preference of affective enhancement between the original (left) and the adjusted (right) images.

rating are stored in a database for later processing.

A total number of 18 subjects participated in this test study. The participant group is entirely from the university students of various level (undergrad to graduate) and no one is an expert of photograhy as per our knowledge. The average age of the participants is around 18-40 years. We sent the web link for the user study to a mixed participant group with a few users who already provided feedback for building training ground-truth data and some new users. This will provide an evaluation which can test the generalization of our approach by allowing some new participants in the user study, thereby avoiding any preference bias with a fixed set of user group (those already provided feedback for training ground-truth data collection). The image stimulus for the testing procedure is selected from the IAPS database but distinct from the training set. The size of the testing set of images is set to be 50.

We initially selected a random set of 100 and then assigned them to the respective clusters by its distance to the closest centroid, and then trimmed them evenly for equal representation of each cluster i.e., around 12 images for each cluster. It was completely a voluntary participation without any renumeration offered. The total number of testing samples collected is 427 which is roughly around 23 per user in average. We explicitly mentioned the participants to judge the results purely from an emotional aspect and avoid any bias due to quality, since our main target is to enhance the affective features rather than the general perceptual quality of the image. We provided some introductory notes in the web-interface before the study starts, about some examples and insights of judging the emotional aspect of an image similar to the procedure done in Section 5.2.2.

For the purpose of comparison, we are not aware of any related affective enhancement tool, and thus we compare with a generic baseline i.e., GIMP auto-enhance function. GIMP's auto-enhance tool is geared for enhancing the perceptual quality of images and applies the same enhancement function to each images without considering the emotional aspect. This is unlike our approach of applying different functions for different emotional categories. Each enhanced image stimuli (i.e., image on the right side in Figure 5.11 alternately rotated between our approach and GIMP auto-enhance in the web-interface during user feedback collection. The results are plotted in Figure 5.12. The plot depicts the various percentage ratios with respect to positive, negative, and neutral ratings as collected from user feedback for GIMP auto-enhance tool and our affective enhancement mechansim. The positive label indicate the users preference for the enhanced version of the pairwise comparison (right button option in Figure 5.11), negative label indicate users preferred the original version (left button option in Figure 5.11), and neutral indicates no strong

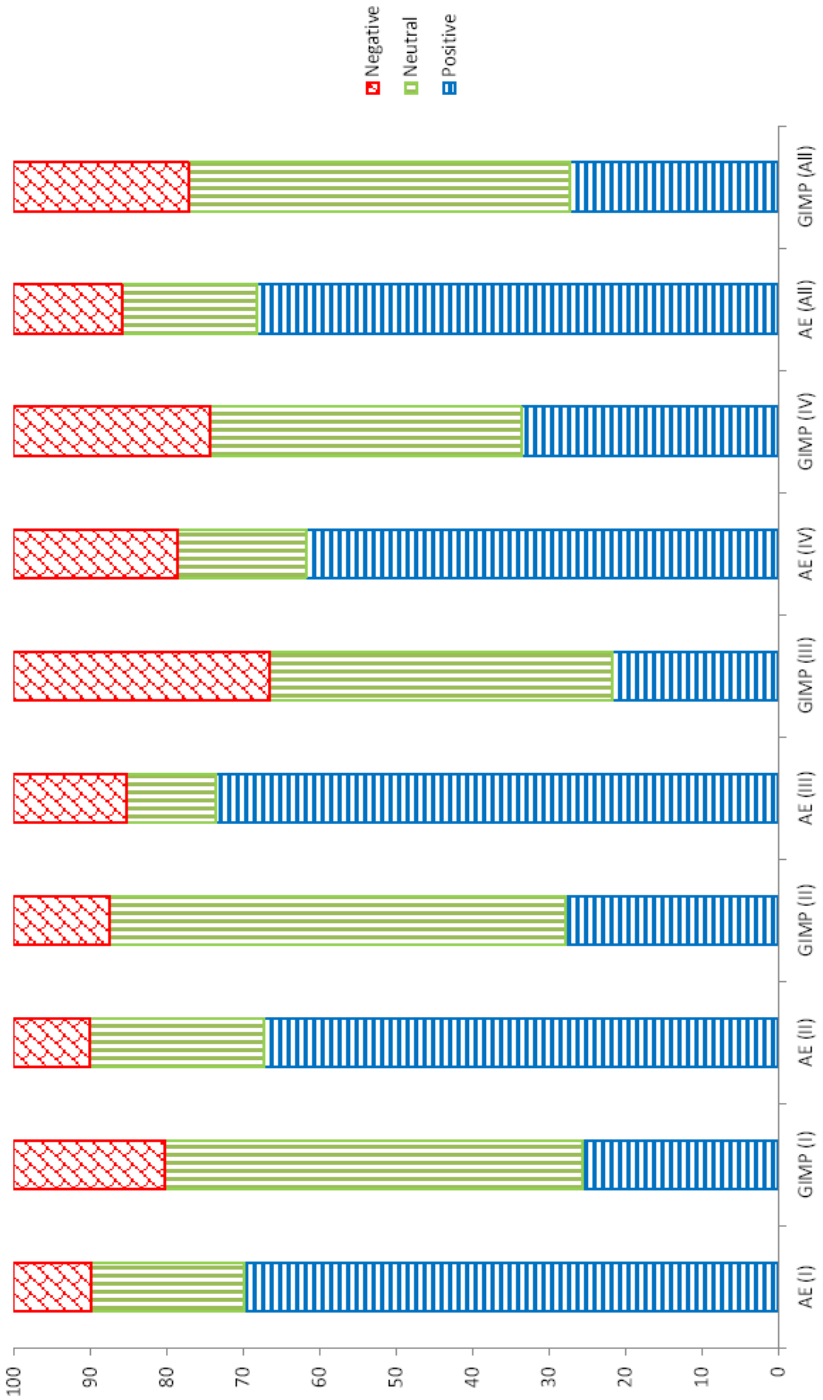

Figure 5.12: The X-axis labels are associated with the two comparative methodologies i.e., AE for our Affective Enhancement and GIMP auto-enhance tool. The symbol within the parentheses indicates the cluster membership and the 'All' label considers the combined results. The Y-axis labels are the percentage ratio collected from user feedback for three different cases i.e., Positive, Neutral, and Negative.

preference or rejection for the enhancement (middle button option in Figure 5.11). The plot also analyses the variation of performance for different clusters (considers the case for the naumber of clusters $k = 4$.) and the final plot on the right end corresponds to the total combined results.

Some of the observations from the plot are as follows: (1) The average case results for our affective enhancement are 68.15%-17.73%-14.13% for positive-neutral-negative respectively which is considerably higher than GIMP auto-enhance with 27.23%-49.9%-22.88% where it can be justified by the fact that, GIMP auto-enhance is not designed for emotional adjustments and only considers quality adaptation. (2) Comparing the results cluster-wise, there is not much significant variation for our method with Cluster: 3 having a slight gain over the others with no apparent reason. This shows that our method performs relatively satisfactorily across the different clusters. GIMP have some variation among the cluster results, but generally it can be observed that a high bias is for the neutral case wherein the user didnot gave any strong preference or rejection for the emotional enhancement.

5.5.2 Objective Evaluation

Deriving an objective metric for affective enhancement of images is a hard problem due to the inherent subjectivity of human emotion. In this section, we design an objective metric to quantify the degree of emotional enhancement in images strictly form image properties and without any ground-truth data. We exploit the color mood space conversion introduced in [OLWW04a], [OLWW04b], [OLWW04c] to achieve our objective. The color mood space gives specific formulae to calculate mood scales from color appearence attributes such as luminance, hue, and chroma which makes it an exciting candidate for objective evaluation of affective enhancement. The mood space is a three coordinates axes of the space called as activity,

weight, and heat. The methodology provides empirical formulations of the transformation from the CIELAB color space to the proposed color mood space. Given a color $\mathbf{c} = \langle L^*, a^*, b^* \rangle$, its corresponding point $\mathbf{e} = \langle a, w, h \rangle$, the color mood space is a nonlinear function of $\mathbf{c}$ defined by the following equations [WYW+10]:

$$a = -2.1 + 0.06 \left[(L^* - 50)^2 + (a^* - 3)^2 + \left(\frac{b^* - 17}{1.4} \right)^2 \right]^{\frac{1}{2}} \tag{5.17}$$

$$w = -1.8 + 0.04(100 - L^*) + 0.45 * cos(h - 100^o) \tag{5.18}$$

$$h = -0.5 + 0.02(C^*)^{1.07} * cos(h - 50^o) \tag{5.19}$$

where L^* is CIELAB lightness, C^* is CIELAB chorma, h is CIELAB hue angle, and a^*, B^* are CIELAB coordinates. On generalization of the additivity relationship [OLWW04c], we compute the color mood of an image by averaging the color of every pixel. Given, $\mathbf{e}$ computed from original image I_i, and $\mathbf{e}'$ computed from enhanced image I_i' using the above transformation function, we quantify the degree of enhancement as follows:

$$D_{(i)} = \|\mathbf{e}' - \mathbf{e}\| \tag{5.20}$$

where $\|\cdot\|$ is the Euclidean distance and at the end we normalize the values to (0,1). Figure 5.13 plots the distribution of affective enhancement using Eq: 5.20 on the same testing sample set employed in Sec: 5.5.1. We cannot directly compare the objective metric with the subjective results from Sec: 5.5.1 due to the difference in scales. To find the degree of correlation between the subjective ratings and the objective values, we analysed the test samples and found the following. By classifying the objective scale as: 0-0.1 (negative), 0.1-0.3 (neutral), and 0.3-1.0 (positive), the % ratios computed are: 13.7, 19.2, and 67.1 respectively, which correlates well with the subjective % ratios (9.15, 12.75, and 78.1 respectively) from the "AE-All" column inf Figure 5.12. The Pearson's correlation index between the subjective and

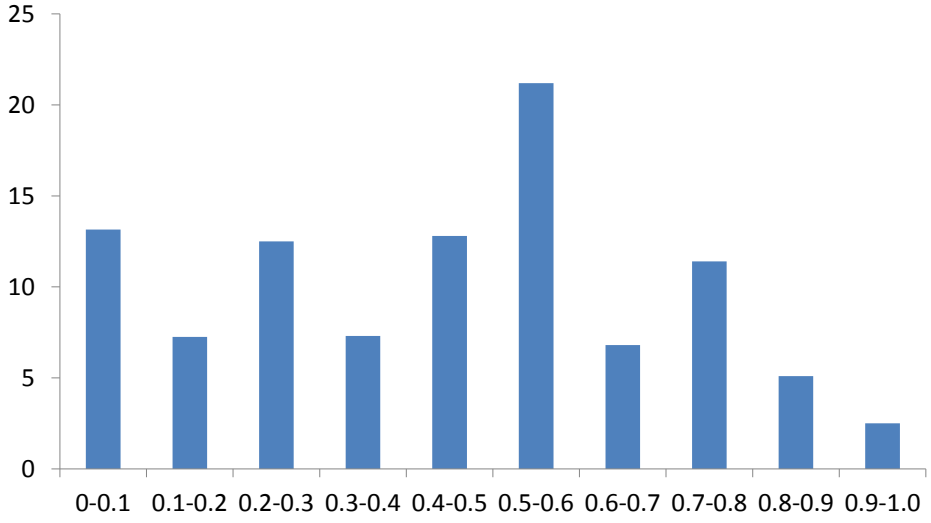

Figure 5.13: Histogram plot depicting the distribution of enhancement degree using the objective function as discussed in Sec: 5.5.2. The X-axis shows the specified intervals objective metric values computed from Equation 5.20 and the Y-axis defines the % ratio of the number of testing image samples within the respective histogram bins.

objective results for the test samples is found to be 0.8125, which indicates that the objective metric of Eq: 5.20 is a good indicator of affective enhancement in images.

Next, we compare the correlation between the computed objective metric to the method used in Sec: 5.4.2 for calculating the enhancement between the IAPS dataset ground-truth and the user collected response from our web-interface for samples in the training database. The CDF plot is shown in Figure 5.14. It can be observed from the plot that both the techniques have a reasonable correlation between them based on the computed metric of affective enhancement. We can also observe that, based on the analysis discussed, we can say that an user-rating with higher valence and higher arousal $[V^+, A^+]$ (as described in Section 5.4.2 from the user feedback for training database) translates to a high value of $D_{(i)}$, valence/arousal higher and the other lower $[V^+, A^-]/[V^-, A^+]$ corresponds to moderate values of $D_{(i)}$, and lower valence-arousal $[V^-, A^-]$ corresponds to very low values of $D_{(i)}$. So, our design in

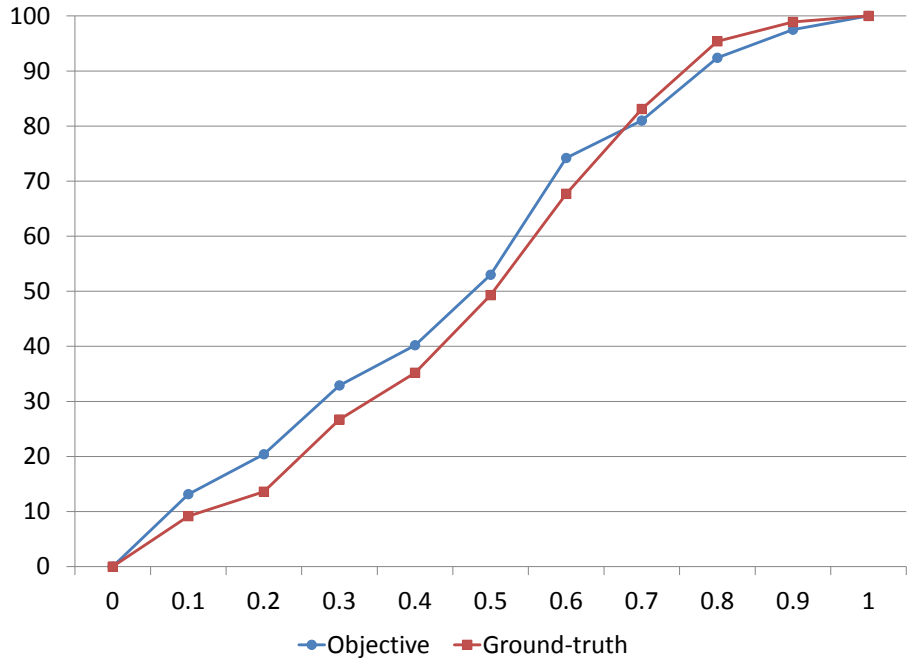

Figure 5.14: CDF plot of the ditribution of enhancement degree between the Ground-truth (methodology proposed in Sec: 5.4.2 for calculating the enhancement between the IAPS dataset ground-truth and the user collected response) and the objective technique described in 5.5.2.

Sec: 5.4.2 for classifying the enhancement degree based on its directional orientation of $VA^{u}_{(j)} = \langle V^{u}_{(j)}, A^{u}_{(j)} \rangle$ with respect to the original image IAPS ratings $VA^{gt}_{(j)} = \langle V^{gt}_{(j)}, A^{gt}_{(j)} \rangle$ is a fair choice to make.

5.5.3 Image Examples

In this section, we observe several image examples with their corresponding enhanced versions as produced by our affective enhancement function for the four different clusters. Clusters: I, II, III, IV are shown in Figure 5.15, 5.16, 5.17, 5.18 respectively. It can be observed that the same enhancement function is applied for all images in the same cluster, thereby the enhancement color and contrast ratios are different across different clusters but remains the same within the cluster. The top row of

two image pairs are rated as "positive"by a majority of users, bottom-left pair as "neutral", and bottom-right pair as "negative".

Now, let us plot some images from the IAPS database with very high/ low arousal/ valence values and observe their output from affective enhancement. In Figure 5.19, the image have a very high valence and high arousal with V=8.1; A=6.28 and so gets mapped to Cluster:I from the clustering algorithm. The left column of all the four rows contain the original image. The right column plots the enhanced version i.e., correctly mapped cluster which is Cluster:I in this case is in the first row; the other rows compose of applying the enhancement function of the other clusters to the same image for the purpose of qualitative analysis. We received six responses from different participants in the subjective study of Section 5.5.1 and five of them gave a positive response in favor of the enhancement of the first row and one participant remained neutral. Figure 5.20 is a low-valence/ low-arousal image with V=2.87; A=3.85 from the IAPS database. The organization is similar as above where the first row is the correctly mapped cluster i.e., Cluster:IV in this case, and the other rows denote the enhancements from non-matching clusters. We received three responses from different participants from the test database with no one giving a positive response, two negative response and one neutral. Figure 5.21 depicts a high-valence/ low-arousal image with V=7.82; A=1.34 and mapped to Cluster:II. We received five responses from the testing database and all of them are positive in this case corresponding to the first row right image. Finally, Figure 5.22 depicts a low-valence/ high-arousal image with V=1.62; A=7.15 and mapped to Cluster:III. We received four user response from the testing database, with two positive instances, one negative, and one neutral.

5.5.4 Summary

In this chapter, we proposed an affect-based image enhancement framework which is targetted to enhance the affective image charactersitics related to color and contrast of an image.

In Section 5.1, we started with a detailed introduction about the current image enhancement schemes in general and how it differs from affective image enhancement and also provided a broad overview of the framework. We discussed the various associated challenges and our proposed solutions for them, followed by a list of our contributions in this chapter.

In Section 5.2, we presented a detailed design of the framework, starting with the construction of the training database in Section 5.2.1, followed by the web-based user interface in Section 5.2.2.

We focussed on the enhancement channel in Section 5.3 which comprises of the various components: Linearization in Section 5.3.1, Auto-correction in Section 5.3.2, Contrast Shaping in Section 5.3.3, Color Temperature in Section 5.3.4, and Color Tint in Section 5.3.5. All these steps after combination, generates an enhancement vector which was used for the adjustment purpose.

The main objective of the supervised learning framework in Section 5.4 is to learn the best enhancement vector which can generate the highest degree of enhancement as detailed through its various components: Clustering in Section 5.4.1, Mapping function in Section 5.4.2, Regression Learning in Section 5.4.3, and Optimization

Solver in Section 5.4.4. We also discussed the process of applying the learned enhancement function to new and unseen images in Section 5.4.5.

We illustrated our experimental results in Section 5.5 using subjective user studies in Section 5.5.1 which gave an average result of 68.15% agreement with user feedback, followed by objective evaluation in Section 5.5.2 using the color mood space which computed a Pearson's correlation index of 0.8125 between the subjective and objective results, and finally we portrayed some direct examples of images before and after enhancement from IAPS dataset using our enhancement function for the different categories in Section 5.5.3.

Figure 5.15: For each pair, the left image is the original one and the right image is enhanced one. The results are for Cluster: I. The top 2 rows are rated by users as "positive", 3rd. row as "neutral", and last row as "negative".

Figure 5.16: For each pair, the left image is the original one and the right image is enhanced one. The results are for Cluster: II. The top 2 rows are rated by users as "positive", 3rd. row as "neutral", and last row as "negative".

Figure 5.17: For each pair, the left image is the original one and the right image is enhanced one. The results are for Cluster: III. The top 2 rows are rated by users as "positive", 3rd. row as "neutral", and last row as "negative".

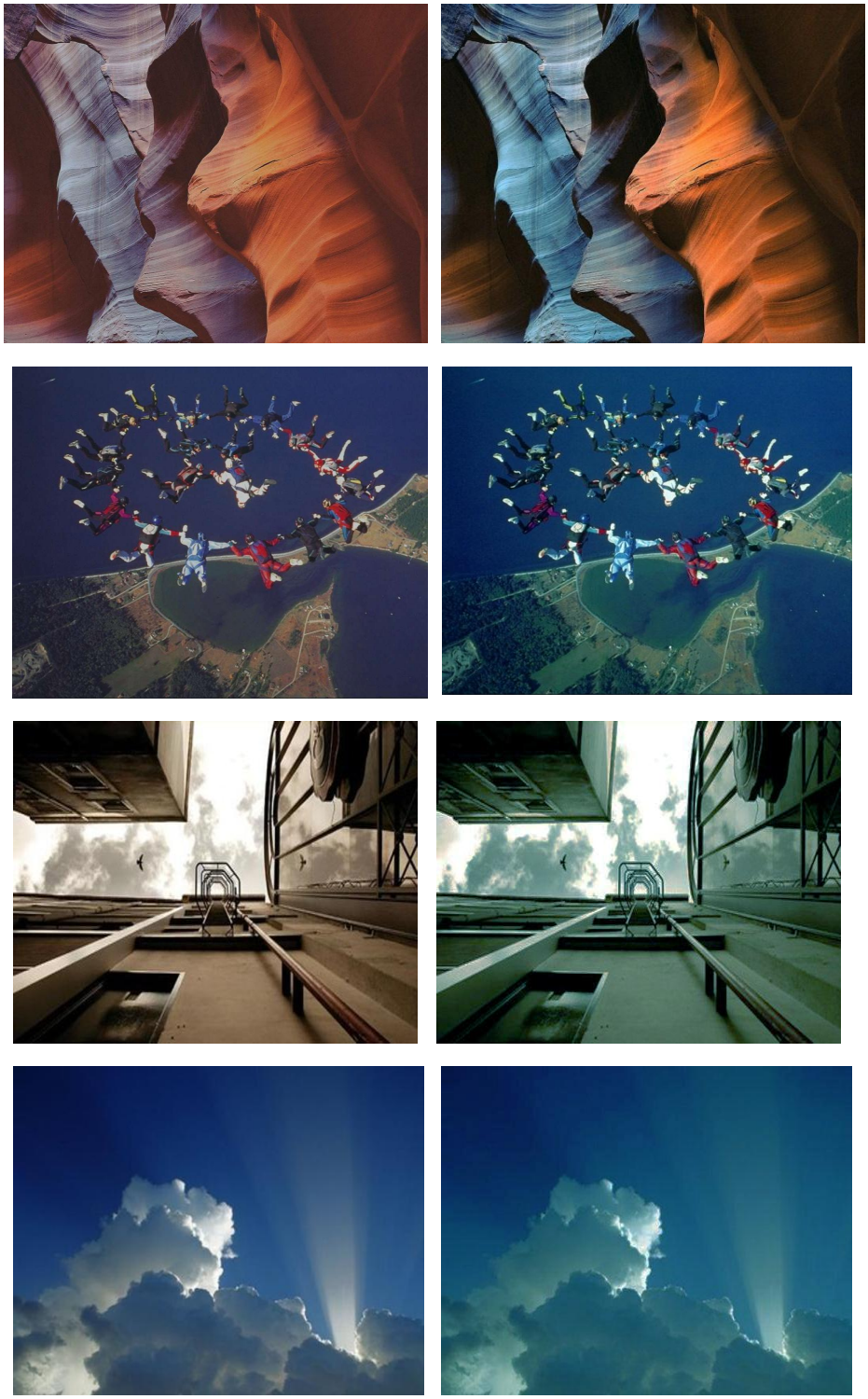

Figure 5.18: For each pair, the left image is the original one and the right image is enhanced one. The results are for Cluster: IV. The top 2 rows are rated by users as "positive", 3rd. row as "neutral", and last row as "negative".

Figure 5.19: High-valence/ High-arousal image from the testing set. The left column in all the rows is the original image. The right column first row is the correct enhancement version with matching cluster i.e., Cluster:I in this case. Other rows contain non-matching cluster enhancement images for comparison purpose.

Figure 5.20: Low-valence/ Low-arousal image from the testing set. The left column in all the rows is the original image. The right column first row is the correct enhancement version with matching cluster i.e., Cluster:IV in this case. Other rows contain non-matching cluster enhancement images for comparison purpose.

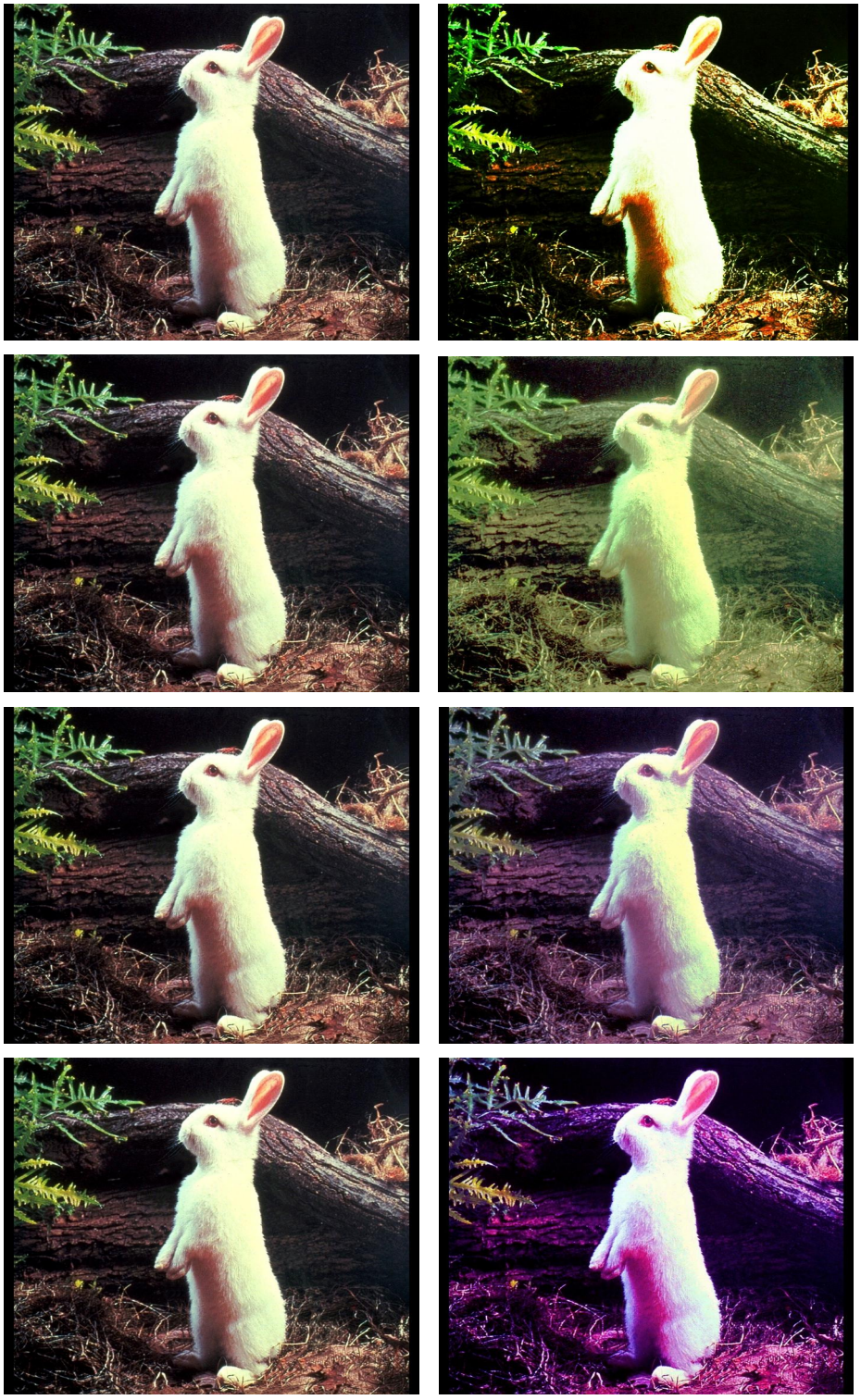

Figure 5.21: High-valence/ Low-arousal image from the testing set. The left column in all the rows is the original image. The right column first row is the correct enhancement version with matching cluster i.e., Cluster:II in this case. Other rows contain non-matching cluster enhancement images for comparison purpose.

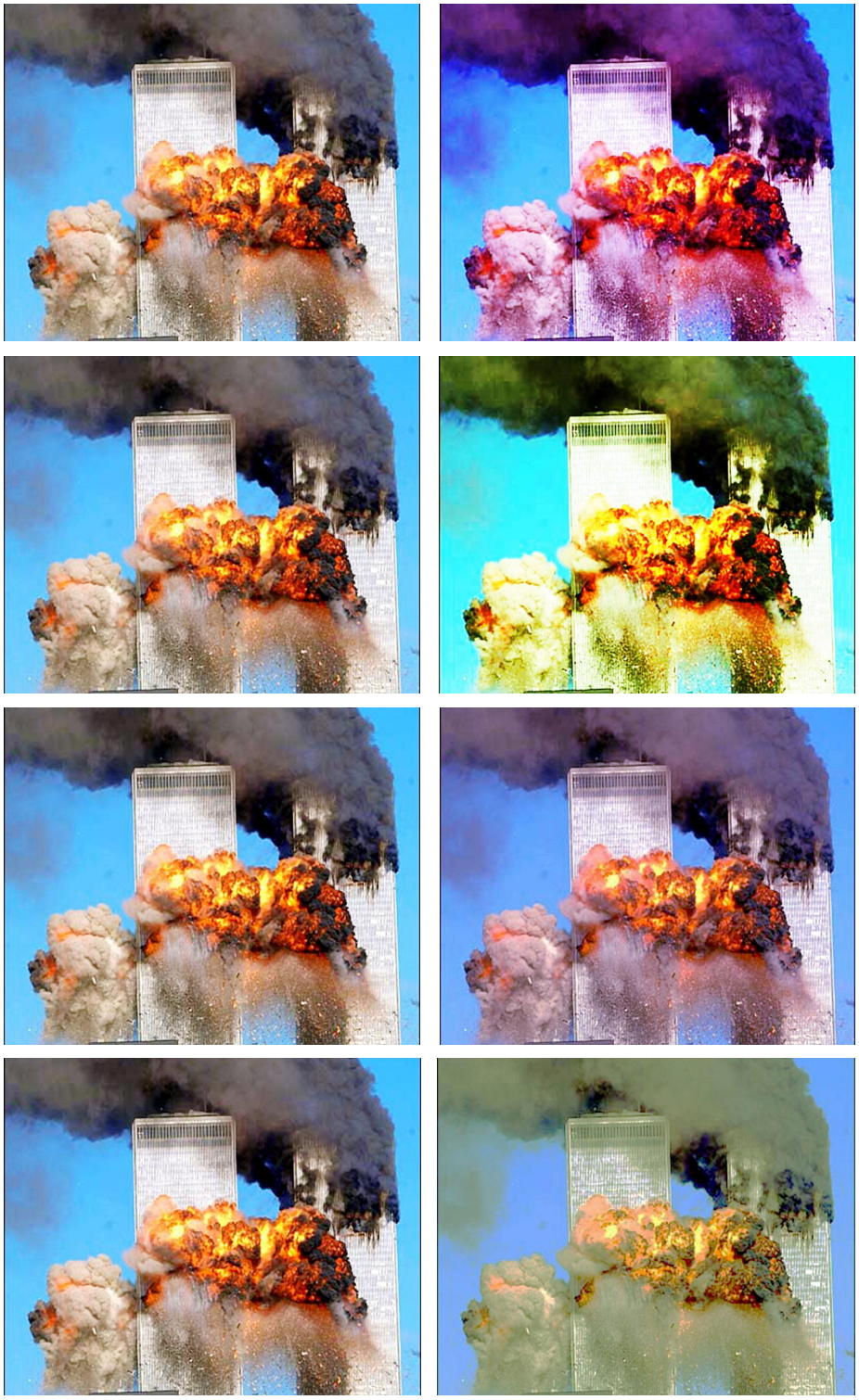

Figure 5.22: Low-valence/ High-arousal image from the testing set. The left column in all the rows is the original image. The right column first row is the correct enhancement version with matching cluster i.e., Cluster:III in this case. Other rows contain non-matching cluster enhancement images for comparison purpose.

CHAPTER 6

CONCLUSION

In this chapter, we start with our thesis objectives in Section 6.1, followed by achievements in Section 6.2, succeeded by conclusions in Section 6.3 and finally wrapping the section with future work in Section 6.4.

6.1 Objectives

This dissertation explored the affect-based computational modeling in multimedia analysis problems, specifically in two applications i.e., QoE evaluation and image enhancement. It is a new and emerging field with lots of possibilities and many unexplored research questions. The objective was to aid the QoE evaluation and image enhancement problems from a user-centric approach by investigating the affective characterisitcs of these processes. The two problems are associated with the two different human sensory channels i.e., aural in QoE evaluation and visual in image enhancement. Moreover, the two problems also differ in their interactiveness i.e., QoE evaluation is associated with a human-human voice interaction scenario and image enhancement is a completely non-interactive scenario. For both the problems, we pursued our objectives as follows:

1. Collected ground-truth data from human users in the form of feedback response in different affective dimensional spaces.
2. Extracted meaningful and effective features from the data in a systematic fashion.
3. Correlated and derived the underlying relationship between the feature data and the user feedback in a computational framework.

4. Predicted different variables in the form of mathematical functions which help in generalizing them to unseen and new samples.
5. Evaluated the results from the framework with different subjective and objective methodologies.

6.2 Achievements

We collected ground-truth data in the form of feedback response as illustrated in Section 4.3.6 for QoE evaluation and Section 5.2.2 for image enhancement. In Section 4.3.6, we collected affective feedback data in the form of voice signals while the participants were engaged in a human-to-human interaction scenario. We also assembled explicit quality ratings after periodic intervals from the user in a discrete dimensional space involving three options: good/bad/average. We conducted realistic human-to-human voice communication experiments (Section 4.3.1) mediated through a network channel, and systematically controlled the channel by imposing various QoS constraints as discussed in Section 4.3.5. In Section 5.2.2, we built affective ground-truth data from the images, by asking the human users to explicitly rate the pictures in a two-dimensional valence-arousal affective space with a 0-9 scale in both the axes. We leveraged the standard IAPS database to construct our ground-truth dataset by replicating the original image ratings from the IAPS results and merged our collected ratings for the adjusted images of their corresponding original images. We employed a web-based user interface for user participation and invited users to visit a link for providing their respective ratings which is similar to the notion of crowdsourcing.

We extracted meaningful affect-related data features from the voice data in Section 4.2 and image data in Section 5.3. Specifically, for voice data, we computed

acoustic features related to fundamental frequencies, energy, duration, and formants in Section 4.2.1; lexical features for capturing the salient information after processing through an ASR in Section 4.2.2; and discourse features for repititions in Section 4.2.3. For image data, we employed contrast and color related features such as power-curve and S-curve for contrast shaping in Section 5.3.3, color temperature in Section 5.3.4 and color tint in Section 5.3.5.

We employed supervised machine learning tehniques for deriving the underlying relationship between the feature data and the user affective ratings in a mathematical form for voice data in Section 4.2.4 and image data in Section 5.4. For voice data, we exploited supervised classification techniques based on SVM and kNN to discriminate different quality perceptions and compare them with user feedback in Section 4.2.4. We experimented different variants such as SVM–5CV, SVM–5WC, SVM–10WC, and kNN–5CV, and compare the performances of each of them with various features. For image data, we utilized clustering in Section 5.4.1, mapping function to encode the degree of enhancement in Section 5.4.2, regression learning in Section 5.4.3, and statistical optimization solver in Section 5.4.4 to achieve the process of learning an optimal enhancement function from the data.

We analysed experimental results for QoE evaluation in Section 4.4 using subjective methodologies by dividing the user collected data into training and testing segemtns using a ratio of 75%:25%. We compared the prediction accuracies by comparing the classifier outputs with the user provided ratings. For image enhancement, we examined subjective methodologies in Section 5.5.1 by user studies through a web-based interface and collected their agreement or non-agreement for affective enhancement. In Section 5.5.2, we applied an objective methodology by exploiting a color mood

space for quantifying the image enhancement degree and it was found to have a reasonable correlation to the subjective counterpart.

6.3 Conclusion

In Chapter 4, we have presented a user study experiment to evaluate an affect-based approach of QoE in voice communication, which represents a new and unexplored area. The purpose of the study is to examine how user's affective behavior changes with the communication quality as mediated through different network QoS conditions and how such changes can be detected and used to predict QoE from the user's perspective. We evaluated the effectiveness of this approach by using classification techniques based on SVM and kNN, to discriminate different quality perceptions and compare with user feedback values. The accumulated evidence supports our initial hypothesis of affective response to be a predictor of QoE due to its correlation with human cognitive perception. Our best performance achieved a prediction accuracy of 67.9%. The experimental results also support our design assumption of incorporating other information sources, such as lexical and discourse features, where it was found that the best results appeared from the aggregation of all sources (A+L+D) consistently as mentioned in Section 4.4.1. Although we refrain from claiming that our methodology can cover the entire spectrum of QoE evaluation factors, our study provided contributory illustrations for affective information to be considered as a relevant indicator. Since our work represents the first attempt in this area, a cross comparison with related existing approaches will be difficult. For example, our system is not directly comparable to QoS-based estimation methods because the latter provides objective measures. In objective methodology, the QoE estimation will be the same for one single QoS setting, which ignores the variation of the user group. We are aware that our present study still has certain limitations such as the poten-

tial emotional influence. It seems likely that in some cases the conversation itself might invoke emotional responses that are not related to the communication quality perception.

In Chapter 5, we presented an affect-based image enhancement methodology which can adjust an input image based on the emotional characteristics of the underlying relationship. As per our knowledge, this was one of the first attempts to derive a computational framework for enhancing the emotional impact of images. The experimental results support our objective, where it was shown that the average results for our affective enhancement are 68.15%-17.73%-14.13% for positive-neutral-negative respectively, with a reasonable advantage over its GIMP counterpart. The objective metric for calculating the affective enhancement is found to have a reasonably high correlation with their subjective counterpart by a Pearson's correlation index of 0.8125. Our experimental results also support the design for calculation of affective enhancement as discussed in Section 5.4.2, with a 3-layered preference model which generated a reasonably high correlation between the subjective and objective results. Though emotion in images is a highly subjective issue with various degrees of dimension ranging from gender, culture, age, and many more, but we achieved to provide a generalized framework by leveraging the standard IAPS database to build our models. More research work is required to understand the various facets of human perception/emotion and its association with affective image enhancement.

6.4 Future Work

We plan to continue our current research along the following possible directions in the future:

- To study influence of other affective cues (e.g., laughter and sigh) on the subjective quality of experience or QoE in multimedia communication.
- To integrate other discourse related attributes (e.g., rephrase, reject and ask over) in the present QoE evaluation framework for further evaluation.
- To improve the implementation for real-time processing of the various voice analysis components in QoE evaluation methodology.
- To investigate an integration of both subjective and objective QoE evaluation methodologies.
- Our testing method applies QoS settings to the communication channel in a random fashion, which can be improved by following real Internet traces to simulate a more realistic testing environment.
- To filter out the "emotional noise" component which is usually present as a bias in normal human-to-human communication, generated due to the communication content. We are presently working on machine learning strategies to cancel the content bias by analysing the emotional context of the transcribed text from the voice signals.
- We also plan to integrate other affective sources of information related to facial expression and physiological response, which will definitely help to improve the results.

- We can also extend the QoE evaluation framework to other media forms i.e., video communication, 3D audio/video systems, smart interactive spaces such as telepresence, and virtual environments.
- Affective enhancement of images can be transformed to enhance the valence and arousal of images separately with different enhancement functions unlike our clustered approach.
- Enhancement of images in a finer granularity level by segmenting the image to different semantic sections and enhance each section differently by preserving the various characteristics of the object in the image.
- Extending the affective enhancement to the video domain by incorporating motion related features inherent in the temporal domain of videos.

BIBLIOGRAPHY

[AAB11] C. O. Ancuti, C. Ancuti, and P. Bekaert. Enhancing by saliency-guided decolorization. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 257–264, Washington, DC, USA, 2011. IEEE Computer Society.

[AAJ10] Ioannis Arapakis, Konstantinos Athanasakos, and Joemon M. Jose. A comparison of general vs personalised affective models for the prediction of topical relevance. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 371–378, New York, NY, USA, 2010. ACM.

[ADK+02] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *in Proc. ICSLP 2002*, pages 2037–2040, 2002.

[AGA+01] Sudha Arunachalam, Dylan Gould, Elaine Andersen, Dani Byrd, and Shrikanth Narayanan. Politeness and frustration language in child-machine interactions. In *IN PROC. EUROSPEECH*, pages 2675–2678, 2001.

[AJD+02] Hekstra A.P., Beerends J.G., Ledermann D., de Caluwe F.E., Kohler S., Koenen R.H., Rihs S., Ehrsam M., and Schlauss D. Pvqm - a perceptual video quality measure. *Signal Processing: Image Communication*, 17(10):781–798, 2002.

[AJG08] Ioannis Arapakis, Joemon M. Jose, and Philip D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, New York, NY, USA, 2008.

[AKJ09] Ioannis Arapakis, Ioannis Konstas, and Joemon M. Jose. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *ACM international conference on Multimedia*, MM '09, New York, NY, USA, 2009.

[AMR02] G.D. Abowd, E.D. Mynatt, and T. Rodden. The human experience [of ubiquitous computing]. *Pervasive Computing, IEEE*, 1(1):48 – 57, jan-mar 2002.

[APM00] G. R. Arce, J. L. Paredes, and J. Mullen. *Nonlinear Filtering for Image Analysis and Enhancement*. Handbook of Image and Video Processing, 2000.

[Bar06] L. F. Barrett. Emotions as natural kinds? *Perspectives on Psychological Science*, pages 28 – 58, 2006.

[BB03] Nadia Bianchi-Berthouze. K-dime: An affective image filtering system. *IEEE MultiMedia*, 10(3):103–106, July 2003.

[BBBK07] M. Barkowsky, J. Bialkowski, R. Bitto, and A. Kaup. Temporal registration using 3d phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality. In *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pages 195 –198, oct. 2007.

[BC98] R. T. Boone and J. G. Cunningham. Children's decoding of emotion in expressive body movement: The development of cue attunement. *Developmental Psychology*, pages 1007–1016, 1998.

[BC07] Russell Beauregard and Philip Corriveau. User experience quality: a conceptual framework for goal setting and measurement. In *International conference on Digital human modeling*, ICDHM'07, Berlin, Heidelberg, 2007. Springer-Verlag.

[Ben01] Ross Bencina. PortAudio: Portable Cross-Platform Audio I/O. http://www.portaudio.com/, 2001.

[BFH+03] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Commun.*, 40, 2003.

[BMM+07] Dario Bonfiglio, Marco Mellia, Michela Meo, Dario Rossi, and Paolo Tofanelli. Revealing skype traffic: when randomness plays with you. In *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '07, New York, NY, USA, 2007.

[BPD06] Soonmin Bae, Sylvain Paris, and Frédo Durand. Two-scale tone management for photographic look. In *ACM SIGGRAPH 2006 Papers*, SIGGRAPH '06, pages 637–645, New York, NY, USA, 2006. ACM.

[BS06] S. A. Baset and H. G. Schulzrinne. An analysis of the skype peer-to-peer internet telephony protocol. In *IEEE International Conference on Computer Communications*, 2006.

[BTN00] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88:411–424, 2000. 10.1007/PL00011380.

[CAA10] Robert Carroll, Aseem Agarwala, and Maneesh Agrawala. Image warps for artistic perspective manipulation. *ACM Trans. Graph.*, 29(4):127:1–127:9, July 2010.

[CD10] R.A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18 –37, jan. 2010.

[CDCT+01] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1), 2001.

[CH67] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21 –27, january 1967.

[CHHL06] Kuan-Ta Chen, Chun-Ying Huang, Polly Huang, and Chin-Laung Lei. Quantifying skype user satisfaction. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '06, pages 399–410, New York, NY, USA, 2006. ACM.

[CHW+06] K.-T. Chen, P. Huang, G.-S. Wang, C.-Y. Huang, and C.-L. Lei. On the sensitivity of online game playing time to network qos. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, pages 1 –12, april 2006.

[CKK11] J.C. Caicedo, A. Kapoor, and Sing Bing Kang. Collaborative personalization of image enhancement. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 249 –256, june 2011.

[CL01] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[COSG+06] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. *ACM Trans. Graph.*, 25(3):624–630, July 2006.

[CP07] Jared R. Curhan and Alex Pentland. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802 – 811, 2007.

[Csi90] Mihaly Csikszentmihalyi. *Flow: The psychology of Optimal Experience.* Harper and Row, 1990.

[CTX09] Kuan-Ta Chen, Cheng-Chun Tu, and Wei-Cheng Xiao. OneClick: A framework for measuring network quality of experience. In *INFOCOM 2009, IEEE*, 2009.

[CWCL09] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowdsourceable qoe evaluation framework for multimedia content. In *ACM international conference on Multimedia*, MM '09, New York, NY, USA, 2009.

[Dam94] A. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Publishing, 1994.

[Dar05] C. Darwin. *The Expression of the Emotions in Man and Animals*. Kessinger Publishing, 2005.

[DBW89] Fred D. Davis, Richard P. Bagozzi, and Paul R. Warshaw. User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8):pp. 982–1003, 1989.

[DGH+02] Z. Duric, W.D. Gray, R. Heishman, Fayin Li, A. Rosenfeld, M.J. Schoelles, C. Schunn, and H. Wechsler. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE*, 90(7):1272 – 1289, jul 2002.

[DJLW06] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European conference on Computer Vision - Volume Part III*, ECCV'06, pages 288–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[DJS+09] K. Dale, M.K. Johnson, K. Sunkavalli, W. Matusik, and H. Pfister. Image restoration using online photo collections. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2217 –2224, 29 2009-oct. 2 2009.

[Ebr09] Touradj Ebrahimi. Quality of multimedia experience: past, present and future. In *ACM international conference on Multimedia*, MM '09, New York, NY, USA, 2009.

[Ekm92] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, pages 169 – 200, 1992.

[Ekm99a] P. Ekman. Basic emotions. *The Handbook of Cognition and Emotion*, pages 45 – 60, 1999.

[Ekm99b] P. Ekman. Facial expressions. *The Handbook of Cognition and Emotion*, pages 301 – 320, 1999.

[EWS10] Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: the munich versatile and fast open-source audio feature extractor. In *ACM international conference on Multimedia*, MM '10, New York, NY, USA, 2010.

[FB04] Jodi Forlizzi and Katja Battarbee. Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, DIS '04, pages 261–268, New York, NY, USA, 2004. ACM.

[FC06] T.H. Falk and Wai-Yip Chan. Single-ended speech quality measurement using machine learning methods. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6), 2006.

[Fis22] R. A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):pp. 597–612, 1922.

[GB08] M. Grant and S. Boyd. Matlab software for disciplined convex programming. *http://stanford.edu/ boyd/cvx/*, 2008.

[GC03] David Graff and Christopher Cieri. English Gigaword Linguistic Data Consortium, Philadelphia. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp ? catalogId=LDC2003T05, 2003.

[GLM+90] A.L. Gorin, S.E. Levinson, L.G. Miller, A.N. Gertner, A. Ljolje, and E.R. Goldman. On adaptive acquisition of language. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, volume 1, April 1990.

[GMH97] Phil Gray, Rob Massara, and Mike Hollier. An experimental investigation of the accumulation of perceived error in time-varying speech distortions. In *Audio Engineering Society Convention 103*, 1997.

[GMSL10] Maxim Graubner, Parag S. Mogre, Ralf Steinmetz, and Thorsten Lorenzen. A new qoe model and evaluation method for broadcast audio contribution over ip. In *International workshop on Network and operating systems support for digital audio and video*, NOSSDAV '10, New York, NY, USA, 2010.

[GP07] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, pages 1334–1345, 2007.

[GW02] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley, 2002.

[HCH09] Te-Yuan Huang, Kuan-Ta Chen, and P. Huang. Tuning skype's redundancy control algorithm for user satisfaction. In *INFOCOM 2009, IEEE*, pages 1179 –1187, april 2009.

[HDL+10] Xintao Hu, Fan Deng, Kaiming Li, Tuo Zhang, Hanbo Chen, Xi Jiang, Jinglei Lv, Dajiang Zhu, Carlos Faraco, Degang Zhang, Arsham Mesbah, Junwei Han, Xiansheng Hua, Li Xie, Stephen Miller, Lei Guo, and Tianming Liu. Bridging low-level features and high-level semantics via fmri brain imaging for video classification. In *Proceedings of the international conference on Multimedia*, MM '10, pages 451–460, New York, NY, USA, 2010. ACM.

[HFH+09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11, November 2009.

[HHCW10] Te-Yuan Huang, P. Huang, Kuan-Ta Chen, and Po-Jung Wang. Could skype be more satisfying? a qoe-centric study of the fec mechanism in an internet-scale voip system. *Network, IEEE*, 24(2):42 –48, march-april 2010.

[HMP+08] Eugene Hsu, Tom Mertens, Sylvain Paris, Shai Avidan, and Frédo Durand. Light mixture estimation for spatially varying white balance. *ACM Trans. Graph.*, 27(3):70:1–70:7, August 2008.

[HS82] Donald Hobbs and Blank Stuart. *Sociology and the human experience*. Wiley, 1982.

[HSGL11] Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph.*, 30(4):70:1–70:10, July 2011.

[HX05] A. Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143 – 154, feb. 2005.

[itu96] ITU-T Recommendation P.800:Methods for subjective determination of transmission quality, 1996.

[itu99] ITU-T Recommendation P.910:Subjective video quality assessment methods for multimedia applications, 1999.

[itu03] Itu-t rec. bs.1284-1: General methods for the subjective assessment of sound quality, 2003.

[itu05] ITU-T Recommendation G.107. The E-model, a computational model for use in transmission planning, 2005.

[itu08] Objective perceptual multimedia video quality measurement in the presence of full reference, 2008.

[Jai84] A. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1984.

[Jai04] R. Jain. Quality of experience. *Multimedia, IEEE*, 11(1), jan.-march 2004.

[JDF+11] D. Joshi, R. Datta, E. Fedorovskaya, Quang-Tuan Luong, J.Z. Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94 –115, sept. 2011.

[JJVS09] Hideo Joho, Joemon M. Jose, Roberto Valenti, and Nicu Sebe. Exploiting facial expressions for affective video summarisation. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 31:1–31:8, New York, NY, USA, 2009. ACM.

[JMAK10] Neel Joshi, Wojciech Matusik, Edward H. Adelson, and David J. Kriegman. Personal photo enhancement using example images. *ACM Trans. Graph.*, 29(2):12:1–12:15, April 2010.

[JSGP06] Alejandro Jaimes, Nicu Sebe, and Daniel Gatica-Perez. Human-centered computing: a multimedia perspective. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 855–864, New York, NY, USA, 2006. ACM.

[KA97] P. Kuosmanen and J. T. Astola. *Fundamentals of Nonlinear Digital Filtering*. CRC press, 1997.

[KBP07] Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.*, 65(8):724–736, aug 2007.

[KHDM98] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), mar 1998.

[KKL10] Sing Bing Kang, A. Kapoor, and D. Lischinski. Personalization of image enhancement. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1799 –1806, june 2010.

[KMK99] Hendrik Knoche, Hermann G. De Meer, and David Kirsh. Utility curves: Mean opinion scores considered biased. In *International conference on Quality of Service*, IWQoS'99, 1999.

[Koh03] T. Kohonen. A computational model of visual attention. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 4, pages 3238 – 3243 vol.4, july 2003.

[KP05] Ashish Kapoor and Rosalind W. Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 677–682, New York, NY, USA, 2005. ACM.

[KR95] S.-S. Kuo and M. V. Ranganath. Real time image enhancement for both text and color photo images. In *Proceedings of the 1995 International Conference on Image Processing (Vol. 1)-Volume 1 - Volume 1*, ICIP '95, pages 159–, Washington, DC, USA, 1995. IEEE Computer Society.

[Kra02] Jacqueline Kracker. Research anxiety and students' perceptions of research: an experiment. part i: Effect of teaching kuhlthau's isp model. *J. Am. Soc. Inf. Sci. Technol.*, 53:282–294, February 2002.

[KSK+10] Harish Katti, Ramanathan Subramanian, Mohan Kankanhalli, Nicu Sebe, Tat-Seng Chua, and Kalpathi R. Ramakrishnan. Making computers look the way we look: exploiting visual attention for image understanding. In *Proceedings of the international conference on Multimedia*, MM '10, pages 667–670, New York, NY, USA, 2010. ACM.

[LBC08] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report, University of Florida, Gainesville, FL, 2008.

[LCODL08] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Data-driven enhancement of facial attractiveness. *ACM Trans. Graph.*, 27(3):38:1–38:9, August 2008.

[LK03] Chulhee Lee and Ohjae Kwon. Objective measurements of video quality using the wavelet transform. In *Opt. Eng. 42, 265 (2003)*, 2003.

[LKJ+02] C. Lee, O. Kwon, T. Jeong, S. Cho, and H. Kim. Weighted psnr for objective measurement of video quality. In *Proceeding (364) Visualization, Imaging, and Image Processing*, 2002.

[LMM01] L. Lucchese, S.K. Mitra, and J. Mukherjee. A new algorithm based on saturation and desaturation in the xy chromaticity diagram for enhancement and re-rendition of color images. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pages 1077 –1080 vol.2, oct 2001.

[LN02] Christine L. Lisetti and Fatma Nasoz. Maui: a multimodal affective user interface. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, pages 161–170, New York, NY, USA, 2002. ACM.

[LN05] Chul Min Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293 – 303, march 2005.

[LP11] Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *Multimedia, IEEE Transactions on*, 13(5):1031 –1039, oct. 2011.

[LS84] A. Lehar and R. Stevens. High-speed manipulation of the color chromaticity of digital images. *IEEE Comput. Graph. Appl.*, 4(2):34–39, February 1984.

[MB11a] Anush Moorthy and Alan Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51:675–696, 2011. 10.1007/s11042-010-0640-x.

[MB11b] Anush Moorthy and Alan Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51:675–696, 2011. 10.1007/s11042-010-0640-x.

[Mei05] M. Meijer. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, pages 247–268, 2005.

[MH10] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, MM '10, pages 83–92, New York, NY, USA, 2010. ACM.

[MHS06] M. Masry, S.S. Hemami, and Y. Sermadevi. A scalable wavelet-based video distortion metric and applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(2):260 – 273, feb. 2006.

[MI07] Aditi Majumder and Sandy Irani. Perception-based contrast enhancement of images. *ACM Trans. Appl. Percept.*, 4(3), November 2007.

[MM08] J. Mukherjee and S.K. Mitra. Enhancement of color images by scaling the dct coefficients. *Image Processing, IEEE Transactions on*, 17(10):1783 – 1794, oct. 2008.

[Moe08] Sabine A. Moebs. A learner, is a learner, is a user, is a customer: Qos-based experience-aware adaptation. In *ACM international conference on Multimedia*, MM '08, New York, NY, USA, 2008.

[MP06] Ludo Maat and Maja Pantic. Gaze-x: adaptive affective multimodal interface for single-user office scenarios. In *Proceedings of the 8th international*

conference on Multimodal interfaces, ICMI '06, pages 171–178, New York, NY, USA, 2006. ACM.

[MR74] A. Mehrabian and J. A. Russell. *An Approach to Environmental Psychology.* MIT Press, 1974.

[MTK02] A.P. Markopoulou, F.A. Tobagi, and M.J. Karam. Assessment of voip quality over internet backbones. In *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, 2002.

[MZLN10] Lin Ma, Fan Zhang, Songnan Li, and K.N. Ngan. Video quality assessment based on adaptive block-size transform just-noticeable difference model. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2501 –2504, sept. 2010.

[Nah98] Diane Nahl. Learning the internet and the structure of information behavior. *J. Am. Soc. Inf. Sci.*, 49:1017–1023, September 1998.

[Nah04] Diane Nahl. Measuring the affective information environment of web searchers. *Proceedings of the American Society for Information Science and Technology*, 41(1):191–197, 2004.

[OLWW04a] Li-Chen Ou, M. Ronnier Luo, Andrae Woodcock, and Angela Wright. A study of colour emotion and colour preference. part i: Colour emotions for single colours. *Color Research Applications*, 29(3):232–240, 2004.

[OLWW04b] Li-Chen Ou, M Ronnier Luo, Andre Woodcock, and Angela Wright. A study of colour emotion and colour preference. part ii: Colour emotions for two-colour combinations. *Color Research Application*, 29(4):292–298, 2004.

[OLWW04c] Li-Chen Ou, M. Ronnier Luo, Andre Woodcock, and Angela Wright. A study of colour emotion and colour preference. part iii: Colour preference modeling. *Color Research Applications*, 29(5):381–389, 2004.

[Pel05] Catherine Pelachaud. Multimodal expressive embodied conversational agents. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 683–689, New York, NY, USA, 2005. ACM.

[pes01] Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codes, 2001.

[Pic97] R. W. Picard. *Affective Computing*. MIT Press, 1997.

[Pic01] R. W. Picard. Building hal: Computers that sense, recognize, and respond to human emotion. *Society of Photo-Optical Instrumentation Engineers*, 4299:pp. 518–523, 2001.

[Pic03] Rosalind W. Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55 – 64, 2003.

[PIIM05] Benny Prijono, Perry Ismangil, Nanang Izzuddin, and Sauw Ming. PJSIP: Open source SIP stack and media stack for presence, instant messaging, and multimedia communication. http://www.pjsip.org/, 2005.

[PK96] I. Pitas and P. Kiniklis. Multichannel techniques in color image enhancement and modeling. *Trans. Img. Proc.*, 5(1):168–171, January 1996.

[PLC06] E.A. Phelps, S. Ling, and M. Carrasco. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science,17 (4)*, pages 292–299, 2006.

[PR00] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424 –1445, dec 2000.

[Pra01] W. Pratt. *Digital Image Processing*. John Wiley and Sons, 2001.

[PW04] M.H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Transactions on*, 50(3):312 – 322, sept. 2004.

[qoe98] Subjective audiovisual assessment methods for multimedia experience. Technical report, ITU-T Rec. P.911, 1998.

[RBFD03] J. A. Russell, J. A. Bachorowski, and J. M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 2003.

[RBK+06] A.W. Rix, J.G. Beerends, Doh-Suk Kim, P. Kroon, and O. Ghitza. Objective assessment of speech and audio quality - technology and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6), 2006.

[RE93] E. L. Rosenberg and P. Ekman. Facial expression and emotion. *Neuroscience Year:Supplement 3 to the Encyclopedia of Neuroscience*, pages 51–52, 1993.

[Riz97] Luigi Rizzo. Dummynet: a simple approach to the evaluation of network protocols. *SIGCOMM Comput. Commun. Rev.*, 27, 1997.

[RJW02] Zia-Ur Rahman, Daniel J. Jobson, and Glenn A. Woodell. Retinex processing for automatic image enhancement. *Human Vision and Electronic Imaging VII*, 4662(1):390–401, 2002.

[RM77] J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, pages 273–294, 1977.

[ROB68] A. R. ROBERTSON. Computation of correlated color temperature and distribution temperature. *J. Opt. Soc. Am.*, 58(11):1528–1535, Nov 1968.

[Ros98] J. H. Rosenbluth. Testing the quality of connections having time varing impairments. *ITU-T del. cont. COM12-D64*, 1998.

[Rud07] Alex Rudnicky. CMU Pronounciation Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2007.

[Rus03] J. A. Russell. Core affect and the psychological construction of emotion. *Annual Review of Psychology*, pages 145 – 172, 2003.

[SB06] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430 –444, feb. 2006.

[SB10] K. Seshadrinathan and A.C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 19(2):335 –350, feb. 2010.

[SB11] Kalpana Seshadrinathan and Alan Bovik. Automatic prediction of perceptual quality of multimedia signals: a survey. *Multimedia Tools and Applications*, 51:163–186, 2011. 10.1007/s11042-010-0625-9.

[SBCL04] Mingli Song, Jiajun Bu, Chun Chen, and Nan Li. Audio-visual based emotion recognition - a new approach. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–1020 – II–1025 Vol.2, june-2 july 2004.

[SCGH05] E. Sebe, N.and Bakker, I. Cohen, T. Gevers, and T. S. Huang. Bimodal emotion recognition. In *5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[Sch02] K. R. Scherer. Emotion, the psychological structure of emotions. *International Encyclopedia of the Social Behavioral Sciences*, 2002.

[Sch03] Klaus R. Scherer. Vocal communication of emotion: a review of research paradigms. *Speech Commun.*, 40(1-2):227–256, April 2003.

[Ser82] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982.

[SGL11] J. K. Stefanucci, K. T. Gagnon, and D. A. Lessard. Follow your heart: Emotion adaptively influences perception. *Social and Personality Psychology Compass*, pages pp. 296–308, 2011.

[SGP+05] David Sander, Didier Grandjean, Gilles Pourtois, Sophie Schwartz, Mohamed L. Seghier, Klaus R. Scherer, and Patrik Vuilleumier. Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *NeuroImage*, 28(4):848 – 858, 2005.

[SJMP10] Kalyan Sunkavalli, Micah K. Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. In *ACM SIGGRAPH 2010 papers*, SIGGRAPH '10, pages 125:1–125:10, New York, NY, USA, 2010. ACM.

[SKR+08] Renata M. Sheppard, Mahsa Kamali, Raoul Rivas, Morihiko Tamai, Zhenyu Yang, Wanmin Wu, and Klara Nahrstedt. Advancing interactive collaborative mediums through tele-immersive dance (ted): a symbiotic creativity and design environment for art and computer science. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 579–588, New York, NY, USA, 2008. ACM.

[SL10] Martin Solli and Reiner Lenz. Emotion related structures in large image databases. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 398–405, New York, NY, USA, 2010. ACM.

[SPS+05] J. Salojarvi, K. Puolamaki, J. Simola, L. Kovanen, I. Kojo, and S. Kaski. Inferring relevance from eye movements: Feature extraction. *Publications in computer and information science, report a82*, 2005.

[SS01] Emilio Schapira and Rajeev Sharma. Experimental evaluation of vision and speech based multimodal interfaces. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, PUI '01, pages 1–9, New York, NY, USA, 2001. ACM.

[SW07] Batu Sat and Benjamin W. Wah. Playout scheduling and loss-concealments in voip for optimizing conversational voice communication quality. In *ACM International conference on Multimedia*, MULTIMEDIA '07, New York, NY, USA, 2007.

[TBK10] Marko Tkalčič, Urban Burnik, and Andrej Košir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, 20(4):279–311, October 2010.

[TMY78] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460 –473, june 1978.

[TvBDK00] David M. J. Tax, Martijn van Breukelen, Robert P. W. Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9), 2000.

[TWZ+08] Carol Tenopir, Peiling Wang, Yan Zhang, Beverly Simmons, and Richard Pollard. Academic users' interactions with sciencedirect in search tasks: Affective and cognitive behaviors. *Inf. Process. Manage.*, 44:105–121, January 2008.

[TYHT08] Shuji Tasaka, Hikaru Yoshimi, Akifumi Hirashima, and Nuno Toshiro. The effectiveness of a qoe-based video output scheme for audio-video ip transmission. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 259–268, New York, NY, USA, 2008. ACM.

[vdBLV96] Christian J. van den Branden Lambrecht and Olivier Verscheure. Perceptual quality measure using a spatio-temporal model of the human visual system. In *Proceedings of SPIE*, SPIE' 96, pages 450 – 461, 1996.

[Ver06] Keith Vertanen. Baseline wsj acoustic models for htk and sphinx: Training recipes and recognition experiments, 2006.

[VM94] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General, Vol 123(4)*, pages 394–409, 1994.

[VMDD03] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):pp. 425–478, 2003.

[VPB09] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vision Comput.*, 27:1743–1759, November 2009.

[VPBP08] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *ACM international conference on Multimedia*, MM '08, New York, NY, USA, 2008.

[VTR92] Francisco Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, 1992.

[WA04] Allison Woodruff and Paul M. Aoki. Conversation analysis and the user experience. In *ACM SIGCHI Workshop on Exploring Experience Methods Across Disciplines*, 2004.

[Wal98] H. G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 1998.

[WAR+09] Wanmin Wu, Ahsan Arefin, Raoul Rivas, Klara Nahrstedt, Renata Sheppard, and Zhenyu Yang. Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 481–490, New York, NY, USA, 2009. ACM.

[WCHL09] Chen-Chi Wu, Kuan-Ta Chen, Chun-Ying Huang, and Chin-Laung Lei. An empirical evaluation of voip playout buffer dimensioning in skype, google talk, and msn messenger. In *International workshop on Network and operating systems support for digital audio and video*, NOSSDAV '09, New York, NY, USA, 2009.

[WH08] Weining Wang and Qianhua He. A survey on emotional semantic image retrieval. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 117 –120, oct. 2008.

[WHM01] Andrew B. Watson, James Hu, and John F. McGowan. Digital video quality metric based on human vision. *J. Electron. Imaging 10*, pages 20 – 29, 2001.

[Win99] Stefan Winkler. Perceptual distortion metric for digital color video. In *Proceedings of SPIE*, SPIE 3644, pages 175 – 184, 1999.

[WLB04] Zhou Wang, Ligang Lu, and Alan C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121 – 132, 2004.

[WnYlSm06] Wang Wei-ning, Yu Ying-lin, and Jiang Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, volume 4, pages 3534 –3539, oct. 2006.

[WOVY94] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young. Large vocabulary continuous speech recognition using htk. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume ii, pages II/125 –II/128 vol.2, 1994.

[WS98] Anne Watson and M. Angela Sasse. Measuring perceived quality of speech and video in multimedia conferencing applications. In *ACM international conference on Multimedia*, MULTIMEDIA '98, New York, NY, USA, 1998.

[WSM01] Stefan Winkler, Animesh Sharma, and David McNally. Perceptual video quality and blockiness metrics for multimedia streaming applications. In *Proceedings of the International Symposium on Wireless Personal Multimedia Communications*, pages 547–552, 2001.

[WT85] D. Watson and A. Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, pages 219 – 235, 1985.

[WY05] Wei-Ning Wang and Ying-Lin Yu. Image emotional semantic query based on color semantic description. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 7, pages 4571 – 4576 Vol. 7, aug. 2005.

[WYW+10] Baoyuan Wang, Yizhou Yu, Tien-Tsin Wong, Chun Chen, and Ying-Qing Xu. Data-driven image color theme enhancement. *ACM Trans. Graph.*, 29(6):146:1–146:10, December 2010.

[WYX11] Baoyuan Wang, Yizhou Yu, and Ying-Qing Xu. Example-based image color and tone style enhancement. *ACM Trans. Graph.*, 30(4):64:1–64:12, July 2011.

[WZW05] Qingfeng Wu, Changle Zhou, and Chaonan Wang. Content-based affective image classification and retrieval using support vector machines. In *Proceedings of the First international conference on Affective Computing and Intelligent Interaction*, ACII'05, pages 239–247, Berlin, Heidelberg, 2005. Springer-Verlag.

[XK11] Xiaohong Xiang and Mohan S. Kankanhalli. Affect-based adaptive presentation of home videos. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 553–562, New York, NY, USA, 2011. ACM.

[YGEMD07] Kai-Chieh Yang, C.C. Guest, K. El-Maleh, and P.K. Das. Perceptual temporal quality metric for compressed video. *Multimedia, IEEE Transactions on*, 9(7):1528 –1535, nov. 2007.

[YPHG09] Junyong You, Andrew Perkis, Miska M. Hannuksela, and Moncef Gabbouj. Perceptual quality assessment based on visual attention analysis. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 561–564, New York, NY, USA, 2009. ACM.

[YRH+10a] Junyong You, Ulrich Reiter, Miska M. Hannuksela, Moncef Gabbouj, and Andrew Perkis. Perceptual-based quality assessment for audio-visual services: A survey. *Image Commun.*, 25:482–501, August 2010.

[YRH+10b] Junyong You, Ulrich Reiter, Miska M. Hannuksela, Moncef Gabbouj, and Andrew Perkis. Perceptual-based quality assessment for audio-visual services: A survey. *Image Commun.*, 25:482–501, August 2010.

[YvGR+08] V. Yanulevskaya, J.C. van Gemert, K. Roth, A.K. Herbold, N. Sebe, and J.M. Geusebroek. Emotional valence categorization using holistic image features. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 101 –104, oct. 2008.

[ZFL+10] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. *ACM Trans. Graph.*, 29:126:1–126:10, July 2010.

[ZHJ+10] Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao, and Qi Tian. Affective visualization and retrieval for music video. *Multimedia, IEEE Transactions on*, 12(6):510 –522, oct. 2010.

[ZPRH09] Zhihong Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1), 2009.

[ZTH+09] Shiliang Zhang, Qi Tian, Qingming Huang, Wen Gao, and Shipeng Li. Utilizing affective analysis for efficient movie browsing. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1853 –1856, nov. 2009.

VITA

ABHISHEK BHATTACHARYA

December 23, 1975	Born, Kolkata, India
2006–present	Ph.D.Candidate, Computer Science Florida International University Miami, Florida
1996–2000	B.E., Electronics and Communication Engineering Bangalore University Bangalore, India

PUBLICATIONS

Abhishek Bhattacharya, Zhenyu Yang and Christine Lisetti, (2012). *Data-driven Affective Enhancement of Images.* (under submission).

Abhishek Bhattacharya, Deng Pan and Zhenyu Yang, (2012). *An Exploration of A New Design Approach for Incorporating DHT with P2P Applications Multimedia Systems.* (under review at Springer Multimedia Systems).

Abhishek Bhattacharya, Zhenyu Yang and Deng Pan, (2012). *Query Adaptation Techniques in Temporal-DHT for P2P Media Streaming Applications.* International Journal of Multimedia Data Engineering and Management (IJMDEM).

Abhishek Bhattacharya, Zhenyu Yang and Wanmin Wu, (2012). *Quality of Experience Evaluation of Voice Communication: An Affect-based Approach.* Springer Journal on Human-centric Computing and Information Sciences (HCIS).

Abhishek Bhattacharya, Zhenyu Yang and Deng Pan, (2011). *Using Affective features as intrinsic estimator for Quality Assessment in VoIP Systems.* IEEE Global Communications Conference (GLOBECOM), Houston, Texas, Dec 05-09.

Abhishek Bhattacharya, Zhenyu Yang and Deng Pan, (2011). *Popularity Awareness in Temporal-DHT for P2P-based Media Streaming Applications.* IEEE International Symposium on Multimedia (ISM), Dana Point, California, Dec 05 - 07.

Abhishek Bhattacharya, Wanmin Wu and Zhenyu Yang, (2011). *Quality of Experience Evaluation of Voice Communication Systems using Affect-based Approach.* ACM International Conference on Multimedia (ACM MM), Scottsdale, Arizona, Nov 28 - Dec 1.

Abhishek Bhattacharya, Zhenyu Yang and Shiyun Zhang, (2010). *Temporal-DHT and its Application in P2P-VoD Systems.* IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, Dec.13-15.

Shiyun Zhang, Abhishek Bhattacharya, Zhenyu Yang and Deng Pan, (2010). *MERIT: P2P Media Streaming with High Content Diversity and Low Delay.* The 7th. International ICST Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), Houston, Texas, November 17 - 19.

Abhishek Bhattacharya, Zhenyu Yang and Deng Pan, (2010). *Multi-source latency variation synchronization for Collaborative Applications.* IEEE Global Communications Conference (GLOBECOM), Miami, FL, Dec.

Abhishek Bhattacharya, Zhenyu Yang and Deng Pan, (2009). *COCONET: Co-Operative Cache Driven Overlay Network for P2P VoD Streaming.* The 6th. International ICST Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), Las Palmas de Gran Canaria, Spain, November 23 - 25.

Abhishek Bhattacharya and Zhenyu Yang, (2009). *DVBMN-l: Delay Variation Bounded Multicast Network with Multiple Paths.* The IEEE International Conference on Multimedia and Expo (ICME), New York, NY, USA, June 28 - July 3.