

**SOCIAL INTERACTION ASSISTANT**  
**An Assistive and Rehabilitative Technology to Enrich Social Interactions for**  
**Individuals who are Blind and Visually Impaired**

**PhD Proposal**  
Sreekar Krishna

**Committee:**

Dr. Sethuraman (Panch) Panchanathan, Chair

Dr. Gang Qian

Dr. Baoxin Li

Dr. Michelle (Lani) Shiota

Dr. John Arthur Black

Department of Electrical Engineering  
School of Electrical, Computer and Energy Engineering  
Ira A. Fulton School of Engineering  
Arizona State University  
December 2009

## Contents

<b>Chapter 1</b>	<b>1</b>
1.1 Need for Social Interactions:	1
1.1.1 Psychological Support:	1
1.1.2 Social Intelligence:	2
1.1.3 Summary:	3
1.2 Non-verbal Cues – Essential component of Social Interactions:	4
1.2.1 Encoding of Non-verbal Cues:	4
1.2.2 Decoding of Non-verbal Cues:	5
1.2.3 Summary:	6
1.3 Components of Non-verbal Communication:	6
1.3.1 The Communication Environment:	6
1.3.2 The Physical Characteristics of the communicators:	7
1.3.3 Behavior of the Communicator:	8
1.3.3.1 <i>Gesture</i> :	8
1.3.3.2 <i>Posture</i> :	8
1.3.3.3 <i>Touch</i> :	8
1.3.3.4 <i>Face</i> :	9
1.3.3.5 <i>Eye</i> :	10
1.3.4 Summary:	11
1.4 Visual Impairment - a hindrance to Social Interaction:	11
1.4.1 Inability to learn social skills due to the lack of visual feedback:	12
1.4.2 Development of stereotypic body mannerisms, especially body rocking, as they don't get a reinforcement visual feedback on their mannerisms:	12
1.4.2.1 <i>Intervention</i> :	13
1.4.2.2 <i>Self Monitoring</i> :	13
1.4.3 Case study on a student how is blind	14
1.4.4 Summary:	14
1.5 Design of assistive technology towards social interactions:	14
1.5.1 Observations:	16
1.5.1.1 <i>Observation 1</i> :	16
1.5.1.2 <i>Observation 2</i> :	17
1.5.1.3 <i>Observation 3</i> :	17

1.5.1.4 Observation 4:	17
1.5.1.5 Observation 5:	17
1.5.2 Summary:	17
1.6 Sensing Non-verbal Cues:	17
1.6.1 Exocentric sensing:	18
1.6.2 Facial Expression Research:	19
1.6.3 Summary:	21
1.6.4 Egocentric sensing	21
1.6.4.1 Summary:	22
1.7 Delivering Non-verbal Cues:	22
1.7.1 Haptics:	22
1.7.2 Haptic interfaces for delivering interpersonal information:	23
1.7.3 The Vibrotactile Belt:	28
1.7.4 Vibrotactile Glove:	29
1.7.5 Summary:	32
1.8 Research Questions:	32
1. What non-verbal cues are important from the perspective of an individual who is blind or visually impaired?	32
2. What assistive technology framework can be developed towards addressing the important social needs of individuals who are blind and visually impaired?	32
3. How effectively can the non-verbal assistive and rehabilitative cues be identified from state-of-the-art sensors used in developing the above social interaction assistant framework?	32
a. How effectively can social interaction cues be identified from an exocentric perspective using camera as the primary input sensors?	32
b. How effectively can social interaction cues be identified from a egocentric perspective using body motion sensors?	32
4. How effectively can data be delivered back to the users of the social interaction assistant by using haptic processing technologies?	32
1.9 References:	32
<b>Chapter 2</b>	45
2. 1 Requirements for a Social Interaction Assistant	46
2.2 Online Survey	46
2.3 Results:	47
2.3.1 Mean Score Table:	47
2.3.2 Histogram of Responses:	48
2.3.3 Box Plot Analysis:	48

2.3.4 Response Ratio: .....	49
2.3.5 Rank Average and F-score: .....	50
2.3.6 Average Response per Group: .....	51
2.4 Analysis of the survey responses .....	51
2.4.1 Histogram of the responses: .....	51
2.4.2 Box Plot Analysis: .....	52
2.4.3 Response Ration - Questionnaire Bias: .....	52
2.4.4 Rank Average Response: .....	52
2.4.5 Average Response per Group: .....	52
2.5 Summary: .....	53
2.6 Alternative Sensing Platforms for a Social Interaction Assistant .....	53
2.6.1 Concept Social Interaction Assistant Prototypes: .....	53
2.6.1.1 Concept 1: .....	53
2.6.1.2 Concept 2: .....	54
2.6.1.3 Concept 3: .....	54
2.6.2 Social Interaction Assistant Prototype: .....	54
2.6.2.1 System Architecture .....	55
2.6.2.2 Prototype System: .....	55
2.6.3 The Haptic Belt: .....	56
2.6.3.1 Hardware .....	57
2.6.3.2 Software .....	59
2.7 Summary: .....	60
<b>Chapter 3</b> .....	61
3.1 The Hardware: .....	62
3.2 Extracting Body Rock Information from Motion Sensor Data .....	63
3.2.1 Features: .....	63
3.2.1.1 Group 1 – Popular features used by the motion analysis research community [2] [1]: .....	64
3.2.1.2 Group 2 – Authors insights into body rocking data: .....	64
3.2.2 Learning Algorithm: .....	66
3.2.2.1 Classic AdaBoost Learning Framework: .....	66
3.2.2.1 Modest AdaBoost .....	68
3.3 Data Collection .....	68
3.3.1 Controlled Data Collection: .....	68

3.3.1.1 Routine A: .....	69
3.3.1.2 Routine B: .....	69
3.3.1.3 Routine C: .....	69
3.3.2 Uncontrolled Data Collection: .....	69
3.4 Experiments .....	69
3.5 Results .....	70
3.6 Discussion of Results .....	74
3.6.1 Packet Length, and Detection Efficiency .....	75
3.6.2 Generalization Capabilities .....	75
3.7 Summary: .....	76
3.8 References: .....	77
<b>Chapter 4</b> .....	78
4.1 Proxemics .....	78
4.2 Conceptual Framework .....	79
4.3 Accurate Face Detection through the Wearable Camera .....	80
4.3.1 Face Validation Framework: .....	82
4.3.1.1 Module 1: Human Skin Tone Detector with Dynamic Background Modeler .....	82
4.3.1.2 Module 2: Evidence-Aggregating Human Face Silhouette Random Field Modeler .....	85
4.3.1.3 Combining Evidence .....	89
4.3.1.4 Coarse Pose estimation .....	90
4.3.2 Experiments .....	90
4.3.3 Results .....	91
Discussion of Results .....	92
4.4 Delivering Proxemics Information .....	92
4.4.1 Delivering Direction Data - <i>Localization of Vibrotactile Cues</i> .....	93
4.4.1.1 Subjects: .....	93
4.4.1.2 Apparatus: .....	93
4.4.1.3 Procedure: .....	93
4.4.1.4 Results: .....	93
4.4.1.5 Discussion: .....	94
4.4.2 Delivering Distance Data .....	94
4.4.2.1 Tactile Rhythm Design .....	94
4.4.2.2 Experiment .....	95
Summary: .....	98

References: .....	98
<b>Chapter 5</b> .....	100
5.1 Sensing Facial and Head Mannerisms and Expressions: .....	101
5.1.1 FaceAPI: .....	101
5.1.2 Facial Image Features under Investigation: .....	102
5.1.2.1 Image features for expression recognition: .....	103
5.2 Importance of Facial Features: .....	105
5.3 Delivering Facial Mannerisms and Expressions: .....	106
5.3.1 Design Considerations: .....	106
5.3.2 Construction of the Vibrotactile Glove: .....	106
5.3.3 Mapping for facial expressions: .....	107
5.3.4 Experiments: .....	108
5.3.5 Preliminary Results: .....	109
5.3.5.1 Confusion Matrix: .....	109
5.3.5.2 Response Time: .....	111
5.3.6 Proposed work: .....	112
5.4 References: .....	112

# Chapter 1

---

## Introduction & Background Work

---

### 1.1 Need for Social Interactions:

People participate in social interactions every day with friends, family, co-workers and strangers. A strong set of social skills is critical in life—for example, they help us make new friends or make good first impressions at job interviews. Sociologists believe that social interactions are the underpinnings of over modern society and are essential for social development and acceptance of an individual within our society [1]. Social interactions refer to all forms of interpersonal communication between the participants that they deem necessary to initiate and maintain interactions. This could be bilateral (between two individuals) or group interactions (between multiple people). Irrespective, all the participants are engaged in continuous exchange of social information through their behaviors, mannerisms, gaze, posture, proxemics and kinesics [2]. Years of research in human communication, behavioral psychology, and social psychology has culminated in our understanding of social interactions to have two important functions in our everyday lives.

#### 1.1.1 Psychological Support:

Recent studies by Segrin et al. [3] have shown that poor social skills are antecedents to psychosocial problems including depression, loneliness, social anxiety, etc. The authors conducted a battery of tests on college students to determine the effect of stress on the students when they live at away from home. Figure 1 shows Depression and Loneliness plotted against stress levels of undergraduate students. Depression was measured using the Beck Depression Inventory [4], while Loneliness was measured on the UCLA Loneliness Scale (version 3) [5] as an index into the students experience of loneliness. For both of these tests, the participating students were categorized into high, medium and low social skilled groups based on the Social Skills Inventory [6] (a battery of tests administered to determine the socialization ability of an individual).

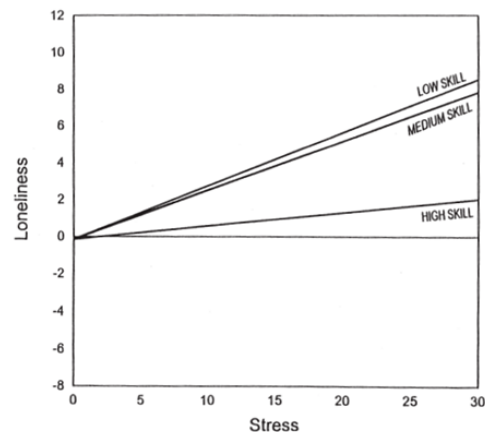
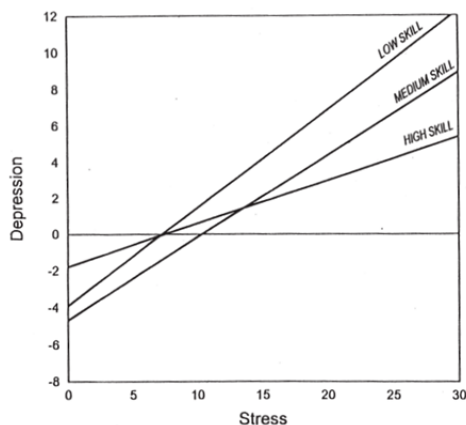


Figure 1: Depression and Loneliness of students plotted against stress levels in high, medium and low social skilled undergraduate students. (Please see text for the scales used for the measurement.)

One can immediately identify a positive correlation between stress and an increased experience of psychosocial problems in all the students, but the ones that rank higher on social skills show higher resistance to stress and in turn higher resistance to mental breakdown. Students assessed with mild or lesser social skills were highly vulnerable to social issues as the stress increased.

Similar results were found in [7] where the authors conclude that people with high competence in communication are known to display immense capability towards adapting their social behavior based on others in their surroundings. Such competence has been acknowledged to reinforce social skills thereby creating a reinforcement feedback that allows these individuals to be successful in their social endeavors [8] and in turn successful in their life. In a tangential study, though Magnusson [9] was not looking for social interaction needs in people, found that social interaction is an important dimension in the cognitive organization of human behaviors. When college students were assessed individually, and as a group, to determine how they classified everyday activities into different situations, Magnusson discovered 5 dimensions (Principle Dimensions). These included two dimensions based on whether the students perceived a situation as being positive (*positivity*) or negative (*negativity*) influence on their behavior, two dimensions based on whether the situations were *active* or *passive*, and finally, the fifth dimension was based on *social interaction* with others. His study emphasizes how social interactions are perceived by individuals as an important scale for judgments on their activity of daily living.

### 1.1.2 Social Intelligence:

Studies in Cognitive Psychology support the hypothesis that social interactions play a vital role in the overall development of intelligence in humans, especially, in the development of Social Intelligence (or Interpersonal Intelligence [10]) and Emotional Intelligence [11]. Social and Emotional intelligence are vital components in an individual understanding the importance of other people and things in their surroundings. Without active social interactions, a large part of the learning component is lost.

*Social Intelligence (SI)* can be defined as the competence in initiating and maintaining group interactions and behaviors. First defined by Edward Thorndike, Social Intelligence is “the ability to understand and manage men and women, boys and girls, to act wisely in human relations” [12]. Karl Albrecht [13] argues that Social Intelligence is the basis for five important aspects for an individual to mingle into his/her society, including, 1) Situational awareness, 2) Sense of Presence, 3) Authenticity (or Individuality), 4) Clarity (of action), and 5) Empathy.

*Emotional Intelligence (EI)* describes the ability, capacity, and skill to identify, assess and manage the emotions of one’s self, others and of groups of individuals. Many models have been proposed in the past to explain EI, such as Ability based models [14], Mixed models [15] and Trait based models [16] and all these models point towards the fact that reduction in social interactions can reduce the overall understanding of an individual of their place in the society. Recently, EI metric scales have been used to diagnose autism spectrum disorders, including autism and Asperger syndrome, semantic pragmatic disorder or SPD, schizophrenia, and Attention-deficit hyperactivity disorder (ADHD). These measurements have



shown a direct correlation of one's ability to increase their overall emotional involvement within the society by increasing their social interactions.

While most SI and EI models have provided theoretical understandings of the importance of social interactions, primate researcher, Humphrey [17], has demonstrated the real-world effect of social interactions to cultural transmission of knowledge and the development of intelligence. His studies with rhesus monkey have emphasized the positive influence of social interactions on the development of general intelligence. For example, Helen (a rhesus monkey) had her visual cortex surgically removed and studies were conducted on her recovery of spatial vision. Over four years, isolated within the laboratory, Helen hardly recovered any of her spatial knowledge. However, when she was taken out of the laboratory into the real world and allowed to interact with objects and other monkeys, she regained three dimensional spatial vision within a few weeks. Humphrey argues that the interactions with other monkeys were key to Helen's learning of interactions (both with objects and other monkeys).

From a neuro-physiological perspective, advanced functional brain imaging is enabling researchers to study the workings of human brain under various functional conditions and they are confirming the role of social intelligence to an important aspect of human learning. Brothers [18] has worked extensively on the neuro-physiological patterns in primate brains that are associated with social behavior. Her work has established the presence of dedicated brain regions involved only in *social cognition* (Social cognition is the processing of information that culminates in the accurate perception of dispositions and intentions of other people). She has proposed a network of neural regions that comprise the social brain and she argues that a malfunction of the any component of the social brain results in reduced social cognition. Her work has been recently bolstered by [19], where the authors study autistics and controls under functional Magnetic Resonance Imaging (fMRI). The subjects watched another person's eye expressions, and guessed what that person was thinking or feeling. The fMRI images confirmed Brothers observations of STG and amygdala activations during social cognition, and showed that people with autism display a cognitive disability in the amygdala which prevents them from making appropriate mental inferences of other people's emotions or facial expressions. Authors conclude that a social brain does exist, and that teaching children and adults social skills could offer a means of increasing activations in the social brain. This conclusion is supported by the behavioral research in autism that employs social interaction training and language skill training in children to ameliorate the social deficits characteristic of autism spectrum disorders (ASD).

### 1.1.3 Summary:

*In summary, social interactions are vital to the workings of our society. Humans learn through their social interactions and these interactions form the basis of our psychological and mental stability. Any disruption to the social interactions of an individual will definitely affect their ability to assimilate into the society. Years of research and observation has shown that sensory, perceptual and cognitive disabilities are a leading cause of such disruption in social interactions.* Once there is a disability set in, the loop of social learning through social interactions is permanently damaged thereby causing a chain of related problems.

## 1.2 Non-verbal Cues – Essential component of Social Interactions:

Social interactions and social skills primarily correspond to the two main channels of communication [20]

- *Verbal communication*: Explicit communication through the use of words in the form of speech or transcript.
- *Non-verbal communication*: Implicit communication cues that use prosody, body kinesis, facial movements and spatial location to communicate information that may be unique or overlapping with verbal information.

From a communication point of view, nearly 64% of all information communication happens through non-verbal cues [21]. Out of this large chunk, 48% of the communication, is through visual encoding of face and body kinesis and posture, while the rest is encoded in the prosody (intonation, pitch, pace and loudness of voice). Inability or difficulty to access any part of this non-verbal cues, seriously affects the overall understanding of the social scene and reduces the involvement of an individual in the social interactions.

### 1.2.1 Encoding of Non-verbal Cues:

From the perspective of encoding information into non-verbal cues, speech, voice, face and body form the primary channels of communication in any social interaction. Speech forms the primary channel for verbal communication, while prosody (intonation, pace and loudness of one's voice), face, and body (posture, gesture and mannerisms) form the medium for nonverbal communication. Unlike speech, which is mostly under the conscious control of the user, the non-verbal communication channels are engaged from a subconscious level. Though people can increase their control on these channels through training, innately, individuals demonstrate certain inability to control their non-verbal cues. This inability to control non-verbal channels is referred to as the leakiness [22] and humans (evolutionarily) have learnt to pick up these leaked signals during social interactions. For example, people can read very subtle body mannerisms very easily to determine the mental state of their interaction partner. Eye Gaze is a classic example of such subtle cues where interaction partners can detect interest, focus, involvement and role play, to name a few. On this leakiness scale, it has been found that the voice is the leakiest of all channels, implying that emotions of individuals are revealed first in their voice before any of the other channels are engaged. The voice is followed by body, face and finally the verbal channel, speech. The leakiness is plotted on the abscissa of Figure 2 with the ordinate showing the amount of information encoded in the other three non-verbal communication channels. It can be seen that the face communicates the most amount of non-verbal cues, while the prosody (voice) is the first channel to leak emotional information.

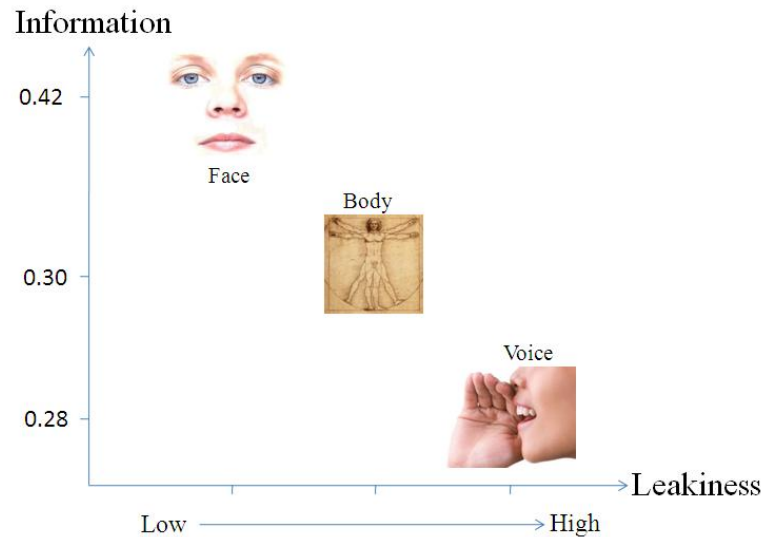


Figure 2: Plot of communicative information encoded in the three important non-verbal channels of encoding. Speech forms the verbal channel. Face, body and voice form the non-verbal communication channels.

### 1.2.2 Decoding of Non-verbal Cues:

From the perspective of decoding non-verbal communication cues, the human input channels can be analyzed under,

- a) **The auditory channel** (includes conscious, verbal speech and unconscious, nonverbal voice),
- b) **The visual channel** (includes nonverbal face and body mannerisms and gestures, which are distributed fuzzily between the conscious and unconscious mediums),
- c) **The combined Audiovisual channel** (includes simultaneous verbal and nonverbal communication mediums), and
- d) **The touch** (includes the nonverbal conscious haptic sensory perceptions).

Figure 3 shows the encoding and decoding components of non-verbal communication (scales for the verbal, non-verbal, audio and visual components are arranged based on the numbers). It can be seen that most part of interpersonal communication is encoded into the non-verbal channels with the visual media conveying the most amount of information.

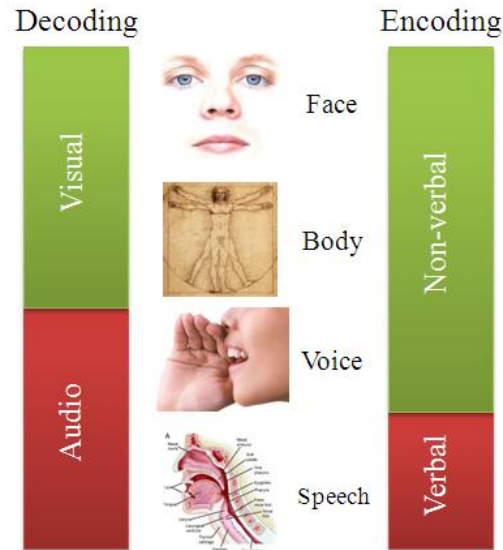


Figure 3: Shows the encoding and decoding aspects of interpersonal communication. From an encoding perspective, humans use verbal and non-verbal cues to communicate while from a decoding perspective, face and body encoded data is received visually and verbal speech and non-verbal prosody are received through audio.

### 1.2.3 Summary:

Non-verbal communication cues form the basis for human interpersonal interaction with vital information encoded into two important channels of vision and hearing. *A large component of these cues are visual in nature thereby requiring individuals to decode non-verbal visual cues before getting involved meaningfully in any social interactions.*

## 1.3 Components of Non-verbal Communication:

Non-verbal communications are inherently complex in nature. In order to understand the nature of these cues, psychologists have been studying these cues under three subdivisions based on what affects individual's non-verbal cueing [2]. These subdivisions include,

- a) The communication environment,
- b) The physical characteristics of the communicators, and
- c) The behaviors of the communicators.

### 1.3.1 The Communication Environment:

The communication environment or surroundings where the interactions are taking place make a huge difference of how humans respond or react [23] [24]. For example, lengthy periods of extreme heat [25] are known to increase discomfort, irritability, reduced work output and unfavorable evaluations of other. Along with the interaction partners, the environment either reinforces or depreciates the emotional experience of an individual. For example, wide open spaces and natural environments are known to be conducive for psychological stability [26]. Though the environmental factors just perceptual, they impose a lot of control on how humans

react towards them. Some of the important environmental factors that affect interpersonal communication and non-verbal cueing are shown in the table below. These are some of the well identified factors towards which psychologists and sociologists are working towards.

<b>The Communication Environment</b>	
Familiarity of the environment	[27][28]
Colors in the environment	[29][30]
Other people in the environment	See next two subsections.
Architectural Designs	[31]
Objects in the environment	[32]
Sounds	[33][34]
Lighting	[35]
Temperature	[25]

### 1.3.2 The Physical Characteristics of the communicators:

The physical appearance of a person is very important aspect of non-verbal cueing. People draw impressions of their communication partner as soon as they see them. The human body acts like means for communicating important sociological parameters like status, interest, dominance etc. Researchers have found cultural and global preferences in overall body image and any deviations from the norm affects interactions between people. For example, facial babyishness [36] has been found affect judgment of facial attractiveness, honesty, warmth and sincerity. Any deviation from the babyishness has been correlated to immediate reduction in the judgment of these traits. A similar such example is the clothing that people wear. It has been found that first impressions are positive if the interviewer and interviewee are clothed similarly [37]. The table below shows the important aspects of a person's physical appearance that affects the interpersonal interaction. Various psychological studies have been conducted towards understanding the model of human perception of character. Very little is known on the reasons for some of the human norms, but it is an active area of research that is being explored rigorously, especially, in the context of group behaviors and personal mannerisms with work environments [38].

<b>Physical Characteristics that affect interpersonal communication</b>	
The human facial attractiveness	[36][39][40]
body shape	[41][42]
height of a person	[43]
self image	[44]
body color	[45]
body smell	[46][47][48]
body hair	[49]
clothing	[37][50]
personality	[51][52]
body decoration or artifacts	[53]

### 1.3.3 Behavior of the Communicator:

The last of the three units of non-verbal communication is the behavior of the communicators. While the term behavior is used loosely in defining this unit, this encompasses both static posture and dynamic movements demonstrated by communicators. Of the three units of non-verbal communication, the behavior forms the most important aspect. Most part of the emotional information encoded by humans is delivered through the behavior of individuals during social interactions. Gestures, Posture, Touch and Voice form the basic subdivisions in behavioral non-verbal cueing. While the entire human body is important for the communication of these cues, the face and eyes play a major role.

#### 1.3.3.1 Gesture:

Gestures are dynamic movement of face and limbs displayed during interpersonal communication. Together, they convey a lot of information that is sometimes redundant (with speech) while other times deliver emotional information about the enactor. Most often gestures are classified based on their occurrence with speech. Accordingly, there are

- a) Speech-independent gestures, or emblems (like shrug, thumbs up, victory sign etc), that are mostly visual in nature and convey the user's response to the situation [54][55].
- b) Speech-related gestures, or illustrators (pointing to a thing, drawing a shape while describing etc) [56].
- c) Punctuation gestures, that emphasize, organize and accent important segments of a communication, like pounding the hand, raising a fist in the air etc.

#### 1.3.3.2 Posture:

Posture refers to the temporary limb and body positions assumed by individuals during interpersonal interactions. Posture is a very effective medium for communicating some of the important non-verbal cues like leadership, dominance [57], submissiveness and social hierarchy [58]. For example, people who show a tendency of dominance tend to extend their limbs out while sitting thereby displaying an overall larger body size. Similarly, submissiveness seems to be correlated to reducing the overall body size by keeps the limbs together.

Both gestures and postures are influenced heavily by the cultural background of the individual and also varied with the geographical location [59]. Though the cultural influence if true with other non-verbal and verbal cues, the perceived difference is the highest in gestures and posture displayed by individuals.

#### 1.3.3.3 Touch:

Social touch has been a very important aspect of non-verbal communication in humans. Developmental biologists believe that the first set of sensory responses in a human fetus is touch [60]. From a social context this sensory channel is very well used in conveying important interpersonal cues such as interest, intimacy, warmth, confidence, leadership and sympathy [61]. Touch is a powerful means of unconscious interaction and it is believed that people who are very good in their social skills rely upon touch a lot [62].

Historically, the sense of touch (Haptics Communication [63]) has been studied by psychologists in the perspective of understanding the human sensory system, but recently, haptics has grown out into the technology front providing human machine interfaces that augment or replace visual and auditory interfaces [64].

#### **1.3.3.4 Face:**

The face is the primary channel for non-verbal communication. Humans are efficient in conveying and receiving plethora of information through subtle movements of their face and head. This focus on the face develops from a very young age and it has been shown that by 2 months, infants are adept in understanding facial gestures and mannerisms [65]. The human face has very fine muscular control allowing it to perform complex patterns that are common to humans, while at the same time being vastly individual [66]. The facial appearance of an individual is due to their genetic makeup, transient moods that stimulate the facial muscles and due to chronically held expressions that seem to set in and become permanent. Human visual system has developed the ability to read these subtleties on people's faces and interpret all the three aspects of the face - genetic makeup (person's identity through face recognition), transient mood (facial expression and emotion recognition), and permanent expression on the face (default neutral face of individuals). While the aspects of permanent facial appearance are important in the recognition of the individual, from a non-verbal communication perspective, the primary function of the face is directed towards communicating emotions and expressions.

The understanding of the human facial expression space was immensely increased by the work of Ekman, Friesen [67] and Izard [68] in the late 1970s. They independently measured precise facial movement patterns and correlated these individual movements with facial expressions on the human face. While Izard developed these patterns on infants, the Facial Action Coding System (FACS) developed by Ekman and Friesen has become the de facto standard for measuring facial expressions and emotions. FACS allow expression and emotion researchers to encode facial movements into accurate contraction and relaxation of facial muscles. Based on these facial actions, Ekman and Friesen discovered the global occurrence of seven basic judged emotions. As psychologists have started to master the FACS system of analyzing facial actions, human computer interaction specialists have started to use the same FACS encodings for building better interfaces that can determine human affect and respond accordingly.

##### **1.3.3.4.1 Facial Action Coding System (FACS):**

FACS defines all possible facial feature movements into Action Units (AU) which represent movement of facial features (like lips, eye brow, chin etc). The AUs are the net effect of facial muscle contraction and relaxation, though they are not directly related to the muscles. Table below shows the different AUs that form the basis of FACS based facial coding with the appropriate number and the associated facial feature movement.

- |                       |                          |
|-----------------------|--------------------------|
| • 1 Inner Brow Raiser | • 9 Nose Wrinkler        |
| • 2 Outer Brow Raiser | • 10 Upper Lip Raiser    |
| • 4 Brow Lowerer      | • 11 Nasolabial Deepener |
| • 5 Upper Lid Raiser  | • 12 Lip Corner Puller   |
| • 6 Cheek Raiser      | • 13 Cheek Puffer        |
| • 7 Lid Tightener     | • 14 Dimpler             |



- 15 Lip Corner Depressor
- 16 Lower Lip Depressor
- 17 Chin Raiser
- 18 Lip Puckerer
- 19 Tongue Out
- 20 Lip stretcher
- 21 Neck Tightener
- 22 Lip Funneler
- 23 Lip Tightener
- 24 Lip Pressor
- 25 Lips part
- 26 Jaw Drop
- 27 Mouth Stretch
- 28 Lip Suck
- 29 Jaw Thrust
- 30 Jaw Sideways
- 31 Jaw Clencher
- 32 Lip Bite
- 33 Cheek Blow
- 34 Cheek Puff
- 35 Cheek Suck
- 36 Tongue Bulge
- 37 Lip Wipe
- 38 Nostril Dilator
- 39 Nostril Compressor
- 41 Lid Droop
- 42 Slit
- 43 Eyes Closed
- 44 Squint
- 45 Blink
- 46 Wink

#### **1.3.3.5 Eye:**

Like the human face, eyes are very important for the control of non-verbal communication. This involvement of human eyes comes from the functions that gaze and mutual gaze play in everyday human interpersonal communication [69]. People use their gaze to convey subtle information that enables smooth verbal interaction which eventually leads to information exchange [70]. From a research perspective, the function of gaze has been classified into four important functional categories [71][2]. These include

##### **1.3.3.5.1 Regulating the flow of communication.**

One of the most important functions of gaze is the regulation of verbal communication in bilateral and group communications. People use gaze to shift focus, bring the attention of a group of people to one thing, turn taking in group conversations [72] and eliciting response from communication partners [73].

##### **1.3.3.5.2 Monitoring feedback.**

Gaze provides a means for individuals to get feedback during conversations and communications. Feedback is a very important tool while people converse. Humans study the eyes of the listener to cognitively inject or eliminate more verbal information into the conversation [74].

##### **1.3.3.5.3 Reflective of cognitive activity.**

Both listeners and speakers tend not to gaze at others when they are processing complex ideas or tasks. Studies have shown that people can answer better when they close their eyes and are allowed to process their thoughts [75]. Thus, cognitive processing is displayed very elegantly by monitoring eye gaze patterns.

##### **1.3.3.5.4 Expressing emotions.**

Along with the facial muscular movements, the eyes play a vital role in the expression of emotions. In fact, in human computer interaction research, it has been found that relying on the eyes and the eyelids alone can provide more accurate delivery of affect information when compared to the entire face [76]. Verbal communication tends to move the lips and mouth



quickly and randomly that can make image and video processing of expressions very tough. Some of the more recent *spontaneous expression* recognition research is focusing on the eyes for this very reason.

#### 1.3.4 Summary:

In summary, non-verbal cue communication specifically depends on the environment, physical appearance of the communicators and the behaviors of the communicators. While all three components play important roles in determining the net effect on the interpersonal communication, the most important aspect of the non-verbal communications are focused on the face, eye and body mannerisms of the communicators. Further, most part of these communications are visual in nature and require the communicator's visual attention to determine the subtle cues. Unfortunately, this implies that the visual senses of all the communicators need to be fully engaged in this process and people who are blind or visually impaired face problem during such scenarios.

### 1.4 Visual Impairment - a hindrance to Social Interaction:

As explained above, most part of the non-verbal encoding happens through visual media. While some parts of these cues are delivered along with speech, most part of the nonverbal communication is inaccessible to someone with visual impairment or blindness. This disconnect from the visual stimulations deprive the individuals of vital communicative cues that enrich the experience of social interactions. People who are blind cannot independently access this visual information, putting them at a disadvantage in daily social encounters. For example, during a group conversation it is common for a question to be directed to an individual without using his or her name—instead, the gaze of the questioner indicates to whom the question is directed. In such situations, people who are blind find it difficult to know when to speak because they cannot determine the direction of the questioner's gaze. Consequently, individuals who are blind might be slow to respond or talk out of turn, possibly interrupting the conversation. As another example, consider that people who are blind cannot use visual cues to determine when their conversation partners change positions (e.g., pacing the floor or moving to a more comfortable chair). In this scenario, an individual who is blind might inadvertently create a socially awkward situation by speaking in the wrong direction.

To compound these problems, sighted individuals are often unaware of their non-verbal cues and often do not (or cannot) make appropriate adjustments when communicating with people who are blind. Also, people who are blind often do not feel comfortable asking others to interpret non-verbal information during social encounters because they do not want to burden friends and family. The combination of all these factors can lead people who are blind to become socially isolated [3], which is a major concern given the importance of social interaction. While people who are blind and visually impaired face a difficulty in social interactions, research in rehabilitation training for these populations recommends that the social involvement for these individuals have to substantially increase in order to enable their acceptance of the society.

National Center for Health Statistics reported in 2007 that the estimated number of visually impaired and blind people totals up to 21.2 million in the United States alone<sup>1</sup>. Global numbers are daunting. In 2002 more than 161 million people were visually impaired, of whom 124 million people had low vision and 37 million were blind<sup>2</sup>. WHO reports that more than 82% of the populations who are blind or visually impaired are of age 50 or older. With the life expectancy going up in most developing countries, the percentage of general population entering into some sort of visual impairment is going to increase in the coming years.

Recently, Jindal-Snape [77][78][79] carried out extensive research in understanding social skill development in the blind and visually impaired. She has studied individual children (who are blind) from India where the socio-economic conditions do not provide for trained professionals to work with children with disabilities. Her seminal work in understanding social needs of children who are blind have revealed two important aspects of visual impairment that restricts seamless social interactions. These include.

#### **1.4.1 Inability to learn social skills due to the lack of visual feedback:**

Jindal-Snape observed that significant others in the environment often fail to give feedback, and even when they do, it is not meaningful or understandable to an individual who is visually impaired—for example, nodding one's head in reply to a question or gesturing. Lack of meaningful feedback could make it difficult for visually impaired persons to comprehend a conversation [78] [80]and, at times, may stop conversing. Similar studies carried out by Celeste [81] indicated that social intervention by parents and teachers are very important in the formative years of a child with visual impairment. Developing on the work by [82], which emphasizes that short-term feedbacks are never effective, Celeste insists that professionals must identify strategies related to social skills that work, provide consistent support and follow children longitudinally to ensure effective development of social skill set.

People who are sighted do not necessarily have the training to work with individuals who are blind or visually impaired. Thus, unconsciously they tend to neglect people who are blind. For example, sighted people use gaze as a primary means of keeping attention with people they communicate with. While conversing with a person who is blind or visually impaired, sighted individuals expect the same gaze feedback. The lack of such a feedback distracts the sighted individuals to turn their attention to or assume disinterest from the visually impaired individual. Research indicates that blind individuals with the ability to accommodate social requirements of their sighted counterparts have exhibited immense personal and professional growth.

#### **1.4.2 Development of stereotypic body mannerisms, especially body rocking, as they don't get a reinforcement visual feedback on their mannerisms:**

Due to the lack of visual feedback, people who are blind and visually impaired do not have access to learn mannerisms from their social counterparts. Especially, people who are impaired at a very young age find it very difficult to learn appropriate social actions and mannerisms. A

---

<sup>1</sup> J.R. Pleis and M. Lethbridge-Çejku, *Summary health statistics for U.S. adults: National Health Interview Survey, 2006*, National Center for Health Statistics, Vital Health Stat 10 (235), 2007.

<sup>2</sup> World Health Organization: *Magnitude and causes of visual impairment*, Fact Sheet N°282 November 2004.

stereotypic body mannerism is one such scenario where positive reinforcement through visual stimulation would have prevented the individual from developing acute non-social conditions.

For over three decades, researchers in behavioral psychology have been publishing case studies on individuals who exhibit stereotypic body rocking. Most of these studies have targeted at reducing or controlling stereotypic body rocking. The methodologies used by these researchers, though varying in nature, can be broadly classified into two important categories.

#### **1.4.2.1 Intervention:**

Intervention relates to any form of feedback provided to an individual at the moment of exhibiting stereotype behaviors. Researchers have attempted to reduce body rocking by providing audio and/or tactual intervention whenever an individual started to rock. They have tried aversive punishment as well as less restrictive positive feedback in such situations. Felps and Devlin [83] issued an annoying tone in the ears of the subject while [84] used a recording of stone scratching on blackboard as the feedback tone whenever the individual started rocking. Both reported that the subjects responded well to the intervention. In contrast, [85], [86] and [87] have used verbal praise, physical guidance, verbal reprimands, and brief time-outs as intervention tools. Most of these researches have shown that intervention has worked in reducing and controlling body rocking without the use of aversive techniques. Aversive or not, these techniques validate a claim that it is possible to control or reduce body rocking (or any other stereotypic body mannerism) through feedback.

#### **1.4.2.2 Self Monitoring:**

In contrast to intervention, self-monitoring does not stop at intervening into the activities of the individual. It attempts to teach these individuals subtle cognitive skills to replace the current mannerism with more socially acceptable behavior, exercise, or medications. McAdam and O'Cleirigh [88] identifies that self monitoring is a very effective way of reducing the body rock behavior. They introduce the case of a congenitally blind individual who is trained (with constant monitoring and positive feedback) to count the number of body rocks he goes through. Researchers noticed that the individual slowly waned off body rocking as he came to recognize and count his body's oscillatory movements. The research concludes that a well designed self monitoring program could benefit in reducing stereotypic body rocking. Shabani, Wilder and Flood [89] presents the case of a 12 year old child who was diagnosed with attention deficit hyperactivity disorder (ADHD) having an excessive body rocking and hand flapping stereotypy. The authors introduce an elaborate and positively rewarding self monitoring scheme that allows the child to improve on his behavior effectively. A follow-up with the child's teacher indicated that the social outlook of the child had improved over the course of rehabilitation and the case further reiterates ability to rehabilitate individuals with stereotypic behavior. Estevis and Koenig [90] introduces a cognitive approach to reducing body rocking on an 8 year old congenitally blind child through self monitoring. Teachers or family members would tap on the shoulders of the child when he started rocking, while the child was taught to recite his own monitoring script. The authors conclude that rocking can be significantly reduced through notification to the individual combined with self monitoring.

Supporting such case studies of behavioral mannerisms, psychologists have been studying intervention and feedback as an integral component of social development. Feedback can be defined as the provision of evaluative information to an individual with the aim of either maintaining present behavior or improving future behavior [91]. According to [92], feedback is critical to social development because after an individual receives information about his or her performance, he or she can make the necessary modifications to improve social skills. Most social skills develop during early years and in order for children to evaluate themselves accurately and to modify social skills, it is essential that children to be given feedback [77][79], since without clear feedback, the children are unable to identify how their social behavior differs from others or is perceived by others in the environment [93]. Based on these studies there is enough evidence that feedback that offers intervention, possibly followed by a well planned self-monitoring program could benefit in reducing or controlling body rocking behavior.

### 1.4.3 Case study on a student how is blind

Technology specialist Shinohara [94][95][96], observed the everyday activities of a college student who was blind named Sara. Shinohara categorized Sara's daily needs into functional categories and has arrived with 5 important aspects in Sara's life where she needs assistance. These include (in order of importance) increased *socialization*, increased *independence* in doing things, increased *control* over things she does, *feedback* from objects around her, and increased *efficiency* in her activities. As seen from the list, socialization was a very important aspect of this college student's requirement. Shinohara concludes that design ideas for technology that supports socialization capabilities for people with visual impairment is of absolute necessity.

### 1.4.4 Summary:

*In summary, individuals who are blind and visually impaired find it difficult in engage in social interactions and any technology that could be developed towards enhancing their access to social cues on an everyday basis might provide opportunities for both social learning and social rehabilitation. Currently, there exists no technology that is focused on developing social interaction assistance and rehabilitation.*

## 1.5 Design of assistive technology towards social interactions:

Historically, the development of assistive devices has tended to be characterized by a technology-centric approach, which begins by asking "What can we do?" This approach is often inspired by a newly emerging technology, and it tends to produce one-size-fits-all technological solutions to the obvious problems that people with disabilities might have already largely solved for themselves. One example of this type of technology-centric approach is a recent research project at Utah State University's *Robotic Guide* [97], which is a robot that employs multiple sensors to provide navigational assistance to users who are blind within a shopping environment. The user interacts with the robot through speech, a wearable keyboard, and audio icons. Although the multimodality approach offers significant advantages, feedback from the participants who are blind and who used the robotic guide indicated that the robot problems that reveal a very technology centric approach to the problem. The robotic guide moved at an average of 0.5 miles per hour which was too restrictive for any person. Additionally, the navigation system for the robot was based on SONAR, which caused jerky movements, and

sometimes provided unreliable results, due to specular reflections and cross talk. The feedback from the focus group indicated that a major portion of decision-making was unnecessarily being off-loaded to the robot thereby restricting their freedom, which was viewed as an undesirable feature. This solution approached the problem from a navigational view point rather than as an accessibility issue. This is an important limitation because people who are blind can navigate independently through an environment using traditional methods, but they cannot read the printed signs, shelf tags, or package labels, nor can they determine the size, color, or pattern in a fabric of clothing in a retail shop. Focusing on the right problem is very important, especially while building assistive technologies.

Another problem with the technology-centered approach is that it often focuses only on the disabilities of the user, without taking into full account the user's abilities. For example, people who are blind are often able to perceive the presence of large objects in the environment around them. Ambient sound sources in the environment provide a form of audio illumination and the resulting sounds bouncing off of objects (or sounds shadowed by objects) allow a person who is blind to detect the presence of those objects. Sometimes in attempting to overcome a disability, developers of assistive devices unintentionally interfere with the user's abilities. For example, assistive devices that require the user to wear headphones or earphones [98] deprive the user of sounds that are vital to the perception of the environment.

In an attempt to develop an assistive technology for delivering facial expression information to individuals who are blind, [99] [100] developed a *haptic chair* for presenting facial expression information. It was equipped with vibrotactile actuators on the back of the chair in a three arm star configuration. The vibrations on the chair are related to the facial expression pattern of the interaction partner. For this experiment, the authors focus only on the mouth of the participant and deliver sad, happy and surprise expressions to the user. Experiments conducted by the researchers showed that people were able to distinguish between three basic emotions. However, this solution had the obvious limitation that the user needed to be sitting in the chair to use the system. The practical applicability of an assistive technology lies in its ubiquity in an everyday environment. Devices should be mobile and/or wearable for them to be useful in different professional and personal settings.

People with disabilities are not always able to perceive or interpret implicit social feedback as a guide to improving their social interaction. However, they might be able to use explicit feedback provided by a technological device. Rana and Picard [101] developed a device called Self Cam, which provides explicit feedback to people with Autism Spectrum Disorder (ASD). The system employs a wearable, self-directed camera that is supported on the users own shoulder to capture the user's facial expressions. The system attempts to categorize the facial expressions of the user during social interactions to evaluate the social interaction performance of the ASD user. Unfortunately, the technology does not take into account the social implication of assistive technologies. Since it is being developed to address social interaction problems, it is important to take into account the social artifacts of technology. A device that has unnatural extensions could become more of a social distraction for both the participants and users than as an aid.

Current trends in pervasive and wearable computing allow miniature sensors to be placed on an individual discretely and inconspicuously. Vinciarelli et. al. [102] have described the use of

discrete technologies for understanding social interactions within groups, specifically targeting professional environments where individuals take decisions as a group. They analyze the use of bodily mannerisms and prosody to extract nonverbal cues that allow group dynamics analysis. They rely on simple sensors in the form of wearable tags [103] which detect face to face interaction events along with prosody analysis to determine turn taking, emotion of the speaker, distance to an individual etc. Pentland describes these signals captured during group interactions as [104] *honest signals*. Some of his recent works [105] in the area of social monitoring hopes to capture these signals and provide feedback to individuals about their social presence within a group. The use of social feedback is illustrated elegantly in their work but their findings relied on sensors carried by all individuals involved in the study. Having everyone in a group wear sensors has proved to be a viable and productive approach for studying group dynamics. However, this approach is not viable as a strategy for developing an assistive technology for people who are blind, as it is not realistic to assume that everyone who interacts with that individual will wear sensors. Thus, it is important to develop technologies that are both egocentric and exocentric in nature, thereby allowing the monitoring of self and others in their environment.

In two independent experiments [106] and [83], researchers developed a social feedback device that provides intervention when a person with visual impairment starts to rock their body displaying a stereotypy. [106] designed a device that consisted of a metal box with a mercury level switch that detects any bending actions. The feedback was provided with a tone generator that was also located inside the metal box. The entire box was mounted on a strap that the user wears around his/her head. The authors tested it on a congenitally blind individual who had severe case of body rocking and they conclude that the use of any assistive technology is useful only temporarily while the device is in use. They state that the body rocking behavior returned to baseline levels as soon as the device was removed. Since the time of this experiment, behavioral psychology studies have explored short term feedback for rehabilitation [78], and these studies support the above observation that short term feedback is often detrimental to rehabilitation and subject's case invariably worsens. Unfortunately, due to the prohibitively large design of the device developed by these researchers, it was impossible to have the individual wear the device over long durations.

In [83] researchers used a 'Drive Alert' (driver alerting system that monitors head droop) to detect body rocking and provide feedback to a congenitally blind 21 year old student. The research concludes that they were able to control body rocking effectively, but the device could not differentiate between body rocks from any other functional body movements. This device, primarily built to sense drooping in drivers provides no opportunity to differentiate between a body rock and a functional droop. Use of such devices could only be negative on the user as a large number of false alarms would only discourage an individual from using any assistive technology.

### **1.5.1 Observations:**

#### **1.5.1.1 Observation 1:**

Assistive technology designed towards social assistance should be portable and wearable so that the users can use them at various social circumstances without any restriction to their everyday life.



#### 1.5.1.2 Observation 2:

Assistive technology designed towards social assistance should allow seamless and discrete embodiment of sensors or actuators making sure the device does not become a social distraction.

#### 1.5.1.3 Observation 3:

Assistive technology designed towards social assistance should incorporate mechanisms embodied on the user to determine both self and other's social mannerism.

#### 1.5.1.4 Observation 4:

Assistive technology designed towards social assistance and behavioral rehabilitation should be used over long durations in such a way that the feedback is slowly tapered off over a significantly longer duration of time.

#### 1.5.1.5 Observation 5:

Assistive technology designed towards social assistance and behavioral rehabilitation should be effective in discriminating social stereotypic mannerisms from other functional movements to keep the motivation of device use high.

### 1.5.2 Summary:

In summary, any assistive technology developed towards social interaction assistance and rehabilitation will have to consider some of the important repercussions of social training, social actions and social impact of technology.

## 1.6 Sensing Non-verbal Cues:

As described in Section 1.2.1, most important aspects of the non-verbal communication cues are visual in nature. After speech, face delivers the most important cues for everyday interpersonal communication [2]. Further, people who are blind or visually impaired are very good at processing some part of the non-verbal cues through auditory signals. For example, they can sense large abrupt movements made by their interaction partners caused due to their cloths, furniture and other objects in the environment. It is the finer details of motion pertaining to the facial expression, hand gestures and eye gaze that becomes a problem in everyday interactions. Thus, introducing sensing technologies that can augment their abilities should be capable of providing access to the visual nature of some of the important non-verbal cues.

In the past two decades, machine vision technologies have advanced tremendously. This includes both the engineering aspects of developing ever smaller cameras and also the computing aspects of developing pattern recognition and machine learning tools that enable real-time analysis of images and videos. This advancement in image and video processing has resulted in advanced algorithms that are capable of sensing some of the important non-verbal cues that were identified in Sections 1.3.1 through 1.3.3. Though these techniques were not developed with social interaction assistance as being the focus, it is possible to adapt some of these techniques towards developing assistive technologies. In the table below non-verbal cues are presented along the rows and the columns present some of the popular computer vision algorithms. Each cell in the table presents appropriate research work that represents potential algorithm for specific non-verbal cue extraction. This is represented here as exocentric sensing as it allows a user to observe the field of view in front of them and understand the non-verbal cues.

### 1.6.1 Exocentric sensing:

	Scene Change Detection	Background Modeling	Face & Object Detection	Environment Analysis	Person Recognition	Clothing Recognition	Body Part Segmentation	Facial Feature Segmentation	Gender Race Recognition	Facial Motion Analysis	Body Motion Analysis	Eye Detection	Eye Tracking
<b>Interaction Environment</b>													
Proxemics		[107]	[108] [109]				[110] [111]						
Objects in the scene	[112]	[113] [114]	[108]										
Natural vs manmade environment	[112]			[115]									
<b>Physical Characteristics of the Communicator</b>													
Race & Body Color							[116] [117] [111]		[118]				
Body Shape					[119] [120]	[121] [122]	[123] [110] [116] [117] [111] [124]				[118] [120]		
Body Decoration					[125]								
Facial Hair								[126]					
Eye Glasses								[127]				[128]	
Clothing						[121] [122]							
Hair							[123] [129]						
Age					[130]								
Gender					[119]				[118]		[131]		
Identity					[132] [121] [133] [134] [135]					[136]			
<b>Behavior of the Communicator</b>													
Description of facial features								[137] [136]					
Body Mannerisms							[138] [139] [140]		[118]		[141] [142] [143] [144]		
Eye Gestures												[128]	[145] [146]
Gaze										[147]		[148] [149]	[147] [150]
Expressions & Emotions					[135]			[127] [137]		[151] [126] [127] [152] [153]	[154] [146] [144]	[155]	[145]
Personality					[119]		[139]		[118]		[143]		
Posture					[119]		[111]	[137]			[142] [126]		



The table above represents a comprehensive list of various technologies that exist in the computer vision and pattern recognition community that can offer solutions for providing non-verbal cues to people who are blind and visually impaired. While many interesting research questions exist under each of these sections, we focus on the two boxes that are shown in here as shaded cells. The first cell under “Face & Object Detection” along “Proxemics” and the second cell under “Facial Motion Dynamics” along “Expressions and Emotions” refer to the important non-verbal cues that can be extracted from the face of an interaction partner. The later chapters in this document will introduce these computer vision techniques that will be used for improving face detection and facial analysis.

### 1.6.2 Facial Expression Research:

Facial expression research has been popular among computer vision scientists for over a decade. Encompassed within the broader research issue of affect recognition, expression recognition has been growing alongside with other bodily sensor signal processing. A history of affect recognition from audio and visual sensors along with state-of-the-art algorithms can be found in [156]. The table below shows the state-of-the-art in vision based facial expression recognition. The table is not a comprehensive representation of the algorithms that have been developed, but it provides a brief glimpse into the various feature selection algorithms, learning methodologies and classification paradigms that are being used. While the state-of-the-art in expression recognition research is very high, from the perspective of developing real-time systems that can provide assistive technology support is still not at a pragmatic stage. This can be attributed to some important concepts that are not being addressed by the computing community. These include.

1. Most research is focused on developing classification algorithms for facial expressions focused on seven basic expressions as described by Ekman and Friesen [67], namely Happy, Sad, Surprise, Angry, Disgust, Fear and Neutral.
2. The research works primarily with standard posed expression databases that are not very well representative of the spontaneous expressions that occur during everyday interpersonal interactions.
3. While facial expression recognition has been analyzed from a classification perspective, in assistive technology solutions, it is more useful to develop a regression framework as it will allow access to subtle changes in facial mannerisms of the interaction partner.
4. Most important, these algorithms have not been developed from the perspective of delivering the information back to a human (except certain assistive aides like [101]). They have only focused on using these classifications in determining the affect of the user from a human computer interaction perspective. While this will allow enhanced interactions in human-machine interfaces, the technology cannot be used effectively as an interface in human-human interaction. The focus of the social interaction assistant is to enhance the interactions between humans.

Reference	Features	Classifier	Performance					
			Exp	Per	Class	Sub	Samp	Acc (%)
[157]	AAM	SVM	S	I	2	21	?	81
[158]	Gabor	SVM + HMM	S	I	3 AUs	17	V	98
[159] [160]	Gabor	AdaBoost SVM	S P	I	17 AUs	119+12	I	93+90.5
[161]	12 motion units	Tree DBN HMM	P	D I	6	5 + 53	V	66.5+73.2
[162]	Shape Models, Gabor	LDC	S	I	3 AUs	21	I	76
[163]	24 facial points	DBN	P	D	6	30	V	77
[164]	Intensity	NN	P	?	7	?	I	68
[165]	Shape fea, Optic flow	C4.5 Bayes Net	P	?	8	4	I	100
[166]	FAPs	Neurofuzzy network	S	I	3	?	I	78
[167]	Shape fea	DBN	S	?	2	8	V	95.3
[168]	Facial and head gesture	GP SVM HMM NN	S	?	2	8	V	86
[169]	Pixel diff of mouth	GP SVM HMM NN	S	I	2	24	V	79
[170]	Intensity of face	Decomposable model	P	I	6	8+16	V	61
[171]	Gabor	AdaBoost SVM	S	I	2	26	V	72
[172]	AAM	SVM	S P	I	AUs	100	?	95
[173]	Facial profile	Rule-based	P	I	27 AUs	19	V	86.3
[174]	Frontal & profile facial points	Rule and case based	P	I	9	8	I	83
[175]	12 motion units	kNN	S	I	4	53+28	V	93+95
[176]	Gabor	AdaBoost DBN	P	I	14 AUs	100+10	I	93+93
[177]	Motion history	SNoW kNN	P	I	15 AUs	19+100	V	61+68
[178]	8 facial points	Gentle Boost SVM	S P	I	2	27+32+65	V	90
[177]	20 facial points	Gentle SVM	S P	I	2	52	V	94
[179]	Shape fea & Intensity	NN	S	?	7	14	I	84
[180]	3D surface	LDA	P	I	6	60	I	83
[181]	Geometric ratio	GMM	P	I	4	47	I	75
[182]	Harr	AdaBoost	P	I	11 AUs	?	I	92
[159]	Intensity	kNN HMM	S P		6	97+21	V	90.7 + 82
[183]	Texture with LPP	SVDD	S	D	2	2	I	87

exp: Spontaneous/Posed expression,  
 per: person Dependent/Independent,  
 class: the number of classes,  
 sub: the number of subjects.  
 samp: sample size (the number of utterances),  
 acc: accuracy,  
 ?: missing entry.

### 1.6.3 Summary:

In summary, developing algorithms that can process the facial features of interaction partners for developing social interaction assistive technologies will require that they be able to

1. Deliver subtle facial movements to the users of the technology.
2. Allow the users to cognitively process spontaneous expressions.
3. Provide real-time tracking of facial movements in an unobtrusive manner.
4. Offer a mechanism to deliver these high bandwidth facial data back to the user in an effective manner.

### 1.6.4 Egocentric sensing

From the discussions above, egocentric sensing mostly pertains to the behavior patterns of an individual who is blind or visually impaired. Specifically, we are monitoring their body movements and detecting stereotypic mannerisms. Recently, human activity detection and recognition using motion sensors have taken a front seat in technology and behavioral research. This is due to the availability of micro mechanized electronic systems (MEMS) that have started to implement complex mechanical systems at a micro scale on integrated circuit chips. These offers advantages like reliability, cheaper cost of production, smaller form factor and above all extremely precise measurement with least or no maintenance. One such sensor is the accelerometer that is capable of measuring the effect of gravity on three perpendicular axes. When mounted on any moving object, the opposing motion (opposing gravity) of the entity allows these sensors to measure the speed and direction of motion. Integrating the magnitude and orientation information over time it is possible to accurately measure the exact motion pattern of the moving entity. These accelerometers have been used by researchers to track motion activity in almost every joint of the human body [184]. Researchers have used single, double or triple orthogonal axis accelerometers to detect various activities of humans

In [185], the researchers provide a nice discussion on some of the ambulatory movements that can be extracted from accelerometers. Five bi-axial accelerometers are used in [184], along with a decision tree classifier to detect and recognize 20 different activities of daily life. They report a recognition rate of over 85%. In [184], the authors evaluated different meta classifiers for recognizing seven lower body motion patterns from a single biaxial accelerometer data and reported the best performance for boosted Support Vector Machines (SVM) [186] with a subject independent accuracy of 64%. Since each dimension of the accelerometer data is similar to audio waveform, popular Hidden Markov Models [187] can be used to learn motion patterns. Reference [188] used HMM to learn the accelerometer data for specific tasks performed by participants and reports a recognition rate of over 90%. In [189], researchers have used two accelerometers placed on the arms of Kung-Fu practitioner and report a recognition accuracy of 3 Kung-Fu arm movements at 96.6%. Research work [190] demonstrates the use of accelerometer data to not only recognize activity, but also localize people within a building. Though the technique is rudimentary, the authors report a high accuracy in recognition of activities while localization still remains a research topic. [191] have demonstrated the use of accelerometers in not only monitoring movements, but also static posture of the human body. They report a recognition rate of 95% using four sensors placed on the chest, thigh, forearm and wrist of participants. Extending this work, [192] have demonstrated an assistive technology solution that uses low cost accelerometers on stroke patients and monitor their posture and walking patterns. Using this information, a feedback is provided to the patient to self-correct

their posture and walking pattern. While these motion sensors are capable of extraction very subtle motion patterns, they have not been exploited in detecting stereotypic mannerisms.

#### **1.6.4.1 Summary:**

*In summary, with the current technology in motion sensors and the ability to provide accurate measurement of motion patterns, stereotypic body mannerisms can be modeled, provided enough training data is acquired for the particular body mannerism that is of interest.*

## **1.7 Delivering Non-verbal Cues:**

The human visual system is a very high bandwidth channel through which immense amount of data is acquired and processed. Providing an auxiliary channel that can handle the immensity of this data is impossible. It is only possible that an alternate modality of information could be made available through assistive technologies which can be used by the individuals who are blind when they deem necessary to access certain non-verbal cues from their interaction partners. Historically, this alternate channel for information delivery has been auditory signals. Various types of information have been encoded into audio signals and delivered in the form of varying frequencies, amplitudes and pulses. But this can only introduce a higher cognitive load on the individuals as they are already processing most of their environmental data in the form of auditory signals. It is imperative that they should not be overload with more information on this channel. To this end, haptics (sense of touch; a mostly unexplored area of human interface design) is introduced as an alternate modality for information delivery.

### **1.7.1 Haptics:**

Recent developments in the area of haptics have resulted in innumerable number of interfaces and interface design principles. Researchers have explored various dimensions of touch (associated with various mechanoreceptors and thermoreceptors on the human skin) including vibratory stimulation (Pacinian corpuscle [193] & Meissner's corpuscles[194]), pressure and texture stimulation (Merkel cell [195]), temperature differential (thermoreceptor [196]), and proprioception (Ruffini Ending [197]). Given these dimensions of haptic actuation and the very large surface area of the human skin, it is possible to develop various technologies that can deliver data in various modalities that can work independently or coactively with the auditory system.


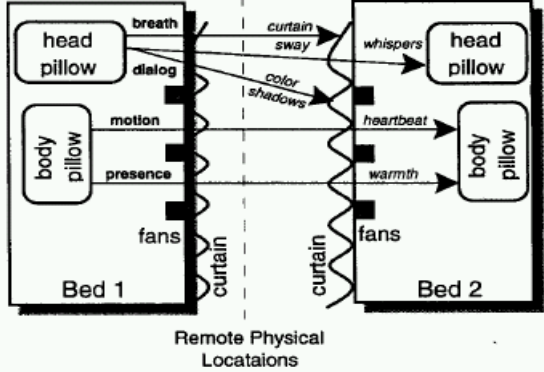
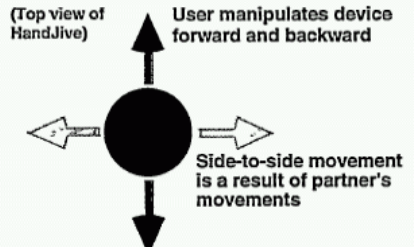
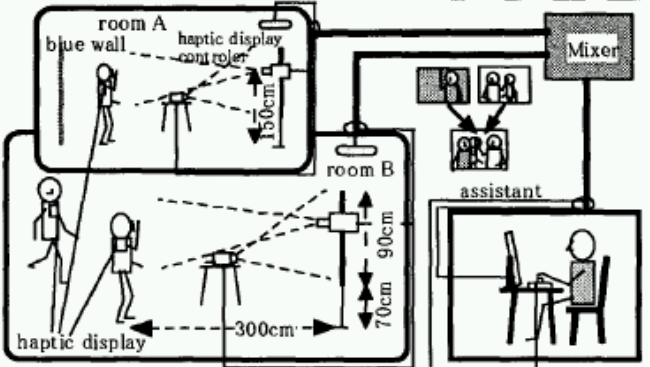
While the entire human body is covered with skin, the brain mapping of the various neural receptors is not consistent across regions. Based on the site on the human skin, the density of the receptors and their mapping to the brain varies. The image of a human exaggerated based on the mapping of the receptors is referred to as the somatosensory homunculus [198]. As shown in the Figure 4, the homunculus has very large hands, lips and genitals. These areas are very sensitive to touch and have very high resolution when compared to other parts of the body. This offers a mapping of where haptic based delivery of information could be places depending on the data bandwidth.



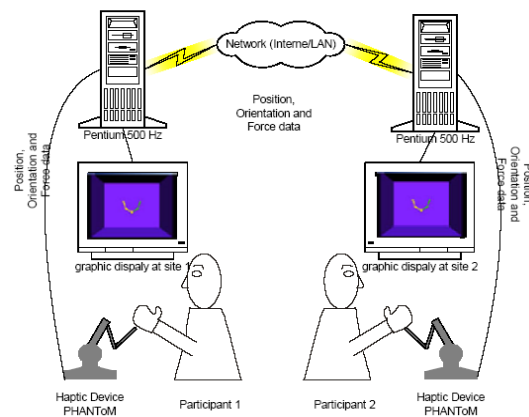


Figure 4: The somatosensory homunculus with exaggerated body parts based on the mapping of haptic receptors in the brain.



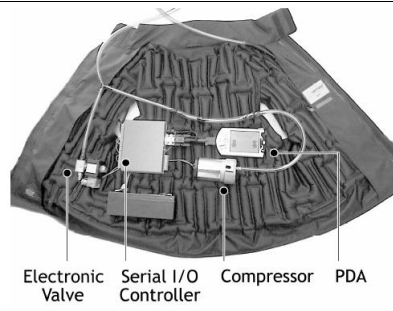
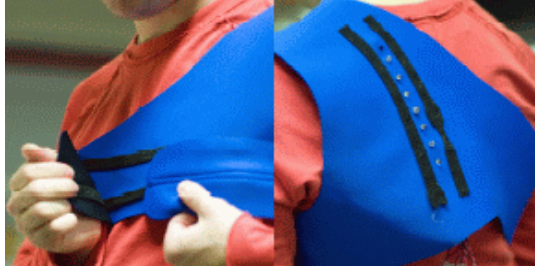
### **1.7.2 Haptic interfaces for delivering interpersonal information:**

Haptics technology has been heavily embraced by the human computer interface community in the past two decades primarily due to the penetration of computing resources into everyday lives of individuals. This has resulted in the emergence of a new area of interface design that focuses primarily on delivering interpersonal information across distances thereby allowing remote interpersonal interactions, both professional and personal. Similar work in the area of telerobotics for surgery, teleoperation of unmanned vehicles, gaming etc have exploded research towards developing haptic interfaces. Below we present a comprehensive assessment of various interfaces that have been developed towards interpersonal communication.


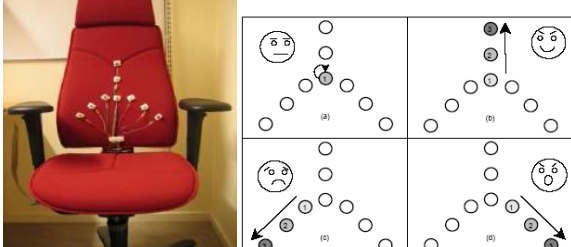
Name	Device	Ref	Actuations	Major Organ	Non-verbal Cue	Application	User Experience
inTouch		[199]	Vibration Pressure Texture	Hand	Touch	-Interpersonal communication through remote touch	No Testing
The Bed	 Remote Physical Locations	[200]	Pressure Texture Temperature	Body	Touch	-Remote interpersonal intimate communication	Self report - System found to produce feelings of intimacy.
HandJive		[201]	Pressure Proprioception	Hand	Handshake Touch	-Interpersonal communication through new cueing language	
HyperMirror		[202]	Pressure	Shoulders	interpersonal distance, relative position, crossing paths	-Remote crossing of paths. -Initiating interaction in strangers across distance. Tap to initiate conversation.	Eye contact was made across distant users. Tap signal aroused attention. Crossing paths initiated conversations.

<b>VinroBod</b>			[203]	Pressure Temperature	Hands	Touch	-Convey remote interpersonal cues.	15 subjects found the device useful and intuitive
<b>What's Shaking</b>			[203]	Vibration Temperature	Hands	Proxemics	-Heat corresponds to the number of people. -Vibration corresponds to the amount of activity in the environment.	12 subjects found the glove intuitive and were able to identify activity around them.
<b>Tele Handshake</b>			[204]	Proprioception	Hnads	Touch	-Remote handshake between interaction partners	65% Satisfaction 55% Convincing 60% Intiutive



Com Touch			[205]	Vibration Pressure	Hands	Touch Emotions	-Bidirectional operation. -Remote participant squeezes one end and a recipient at the other end feels vibrations	-24 subjects. -Subjects came up with their own cueing. -In Desert Survival Task, 15 items were sorted based on importance and 5 were ranked. -83% of participants used atleast one gesture. -67% developed their own gestures																												
Haptic Instant Messenger	<table><tr><th>Icon</th><th>Emoticon</th><th>Meaning</th><th>Hapticon</th></tr><tr><td></td><td>: )</td><td>regular smile</td><td></td></tr><tr><td></td><td>: D</td><td>big smile</td><td></td></tr><tr><td></td><td>: (</td><td>sad face</td><td></td></tr><tr><td></td><td>; -)</td><td>wink</td><td></td></tr><tr><td></td><td>(k)</td><td>kiss</td><td></td></tr><tr><td></td><td>: \$</td><td>embarrassed</td><td></td></tr></table>	Icon	Emoticon	Meaning	Hapticon		: )	regular smile			: D	big smile			: (	sad face			; -)	wink			(k)	kiss			: \$	embarrassed			[206]	Audio Vibrations	Hands	Emotions	Based on user selections at a remote location, haptic and audio codes are transmitted to the receiver.	No user testing
Icon	Emoticon	Meaning	Hapticon																																	
	: )	regular smile																																		
	: D	big smile																																		
	: (	sad face																																		
	; -)	wink																																		
	(k)	kiss																																		
	: \$	embarrassed																																		
Hug over Distance	 		[207]	Pressure Proprioception	Upper Body	Touch Hug	At one end the user rubs tummy of a stuffed toy and based on the pressure applied, air bags are filled at the remote end to simulate hug.	- Air compressor at the receivers end makes a lot of noise - Six couple focus group found the concept weird.																												
TapTap			[208]	Pressure	Shoulders	Touch Tap	Solenoids and vibrators used on the shoulder to simulate tapping.	- 8 men and 8 women tested on the device found based on the tap, it reminded them of someone.																												



United Pulse	 <p>heart rate monitor/inside microcontroller and wireless connection to the mobile (Bluetooth, RadioFrequency)</p>	[209]	Vibrations	Finger	Intimacy	<p>-Vibrators on the ring stimulated to initiate communication between remote couple. -Simulated heart beats were delivered</p>	<p>-20 couples tested with the device. - 22 liked the idea. - 5 were irritated.</p>
Haptic Chair		[100]	Vibrations	Back	Emotions	<p>- Vibrations corresponding to emotions are delivered to the back of the user. - Has sensing of the emotions inbuilt through vision technologies</p>	<p>- 3 expressions tested. - 100% recognition on expressions. - 10% of participants complained of cognitive load.</p>

While most of the devices presented above were developed for the application of interpersonal communication, their focus was restricted to a specific non-verbal cue, especially intimacy. The Haptic Chair [100] is one research that comes close to developing an assistive technology for people who are blind. Unfortunately, the device is not portable and any assistive technology should not be restrictive to the users, especially in a professional setting. Further, the researchers have tried to acquire the emotion data from videos by analyzing the mouth region of the interaction partners. This works in scenarios where the participants are posing expressions and not in spontaneous emotion generation. Verbal movements of the mouth can render the system ineffective due to random movements of the mouth.

As explained above, the number of haptic devices possible is innumerable as the form factor and modality of delivery can vary significantly. Two important form factors are explored in detail below as

1. The vibrotactile belt, develops on the intuitive cueing that is possible in haptic technology for low bandwidth non-verbal cues, like number of people in the vicinity, proxemics, eye gaze of an interaction partner etc.
2. The vibrotactile glove, that is placed on the human hand, which has the highest haptic sensitivity thereby allowing us to deliver high bandwidth non-verbal cues like facial mannerisms.

### 1.7.3 The Vibrotactile Belt:

Vibrotactile cues are vibratory signals defined by signal frequency, intensity, rhythm, and duration [210] of the vibration in contact with the human body. Vibrotactile cues have found uses in a variety of application areas including human navigation [211] [212] [213], human spatial orientation [214][215], human postural control [216] and human communication [210]. The idea of using vibrotactile cues on a haptic belt for information delivery is not a new idea. However, the use of vibrotactile cues for non-verbal communication during social interactions is novel and provides an exciting opportunity to provide assistance with daily tasks to individuals who are blind. This section introduces several approaches for using vibrotactile belts to convey navigation and/or orientation information, which inspired the design of our haptic belt.

In an early haptic navigation system for individuals who are blind [211], Ertan *et al.*, proposed a tactile display (worn on the back) consisting of a 3x3 array of tactors that convey directional information through pulsing columns and rows. In [212], the authors proposed the ActiveBelt, a haptic belt to guide the user to a destination using eight tactors placed around the waist, a GPS unit and an orientation sensor. Another system for human navigation is a tactile vest proposed by Jones *et al.* [213], which utilizes a 3x3 array of tactors placed on the back to convey directional information.

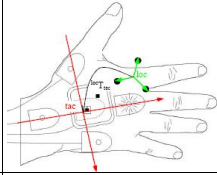
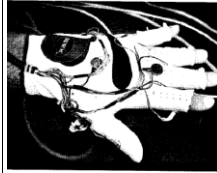
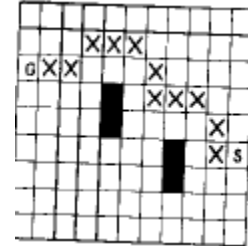
Another application of vibrotactile cues is the Tactical Situation Awareness System (TSAS) [214], which is a tactile suit designed to help reduce spatial disorientation that is sometimes experienced by pilots in flight due to a lack of visual cues. The TSAS uses vibrations to indicate critical information such as the direction of the gravity vector. Similarly, tactile displays have been developed to help astronauts compensate for spatial disorientations [215]. Finally, tactile display devices have been developed to assist people with damage to their vestibular system. For example, in [216], balance control is achieved using a haptic belt system composed of a tilt sensor and three rows of tactors used to indicate body tilt information.

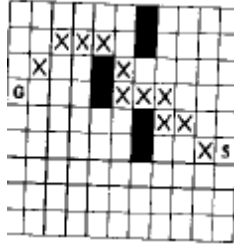
A variety of belt designs have been proposed in the literature. Most of these designs are motivated by a particular application. Furthermore, most of these implementations lack usability and performance studies as the central focus tends to be a proof of concept. In fact, usability and performance are rarely discussed. Implementations often have excessive cabling and bulky modules, while the required robustness and rigidity for real-world use are completely ignored. Table II presents a sample of seven haptic belts, chosen based on maturity and availability of information regarding design choices and implementation details. Table below provides a comparison of these belts, based on the functionality design requirements. A cross indicates that a feature is available, whereas a blank entry indicates that the feature is either not available or its availability is unknown.

<b>Function</b>	[217]	[218]	[212]	[219]	[220]	[221]	[222]
<b>Amplitude</b>	X			X	X		X
<b>Frequency</b>	X		X		X	X	X
<b>Timing</b>	X	X	X		X		X
<b>Location</b>	X	X	X	X	X	X	X
<b>Add/Del motors</b>	X			X			
<b>Adjust tactor position</b>	X	X		X		X	
<b>API</b>				X	X		
<b>Wearable</b>		X	X	X	X	X	X
<b>Wireless</b>		X	X	X	X	X	X

#### 1.7.4 Vibrotactile Glove:

As can be seen from the homunculus in Figure 4, the hand is represented in the brain by a very large somatosensory cortical area. This allows for the human hand to be a receptor for large bandwidth of information. Vibrotactile stimulations have been used in the past to take advantage of the human haptic sensory system and hands form a perfect medium for conveying large variations in such stimulations. The palm of the human hand is haptically more sensitive, when compared to the hairy side of the hand. This has prompted researchers to explore the palm as their medium for communicating haptic cues. Unfortunately, any hindrance to the palm of the hand renders the hand functionally useless. If an assistive technology has to use the hand to deliver haptic cues, it is important that the hairy side of the hands be used. The table below presents some of the important research work in that uses vibrotactile gloves for delivering information in various applications. Detailed description of the application of the glove, the structure of the glove and the accuracy of the system working are shown in the table also.

Ref	Application	No. of vibrators	Location of vibrators	Vibration Pattern	Encoding	Experiments	User Study
[223]	Convey color information to people who are blind.	3	- Distal phalanges of index, middle and ring fingers. (T) - Three phalanges of the index finger. (O)	- Continuous on all three vibrators. (S) - 0.5s time gap between vibrators. (D)	- Encode R, G and B channel to each of the 3 vibrators. - Amplitude of vibration proportional to the intensity of the color channel.	- Convey only colors individually (C). - Allow users to explore a down sampled color image using a mouse. (I)	- 5 participants who are blind. - 2 sighted participants.  - COS: 71% - CTS: 87% - ITS: 100% - IOD: 67% - IOS: 92% - COD: 87% - CTD: 90%
[224]	Vibrotactile cueing to improve target acquisition in virtual 2D environment using mouse as input.	4	- 2 on the lower part of the palm just above the wrist. - 2 on the back of the lower palm just above the wrist.	- 100ms vibratory cues to indicate direction of the target and on-target signals.	- Frequency of the vibration was proportional to the direction and distance from the target location. - Two vibrators were turned on to indicate arrival on a target.	- Expt 1 tested vibrators on the front and back of palm. - Expt 2 tested continuous distance cueing with suppressing or increasing frequency as the target is approached.	- The location of the tractors did not have an effect. Front and Back worked the same, - Suppressing the frequency as the user approaches the target worked better than enhancing.
[225]	Vibrotactile array for delivering distance to an obstacle from a wheelchair driven by a visually impaired person.	9	Array on the front of the palm in a 3x3 matrix.	- Warning signals - Spatial obstacle location signal. - Direction conveyance to the user.	- Warning signal vibrates all vibrators. - Spatial location of an obstacle is sent in the particular motor with near, medium and far range to obstacle. - Direction cue vibrates the center motor with two pulses and then vibrates motor of the desired direction.	- No user testing done yet	- No user testing done yet
[226] [227]	Vibrotactile cues for navigating surgeons hand during surgery.	4		- Continuous vibrations based on the amount of off target displacement	- Optical tracking of visual markers on the surgeons hand is translated to vibrotactile cues to give off-center information.	- Subjects were required to move a surgical tool to the target location.	- Subjects react to varying impulse input as required. No quantification provided in the paper.
[228]	Field of view in front of individual who is blind is captured with a camera and translated to vibrotactile cues corresponding to a depth map.	?	No specific information provided. 	- Magnitude of vibration is directly proportional to distance to obstacle. - Frequency of vibration is inversely proportional to the confidence in depth measurement.	The image from the camera is used to determine a depth map of obstacles in front of the user and is translated into vibrotactile cues.	Two obstacle courses were set within the laboratory environment and the participants were required to navigate the course. Course 1:  Course 2:	- 9 participants, 3 blind and 6 with low vision. - Course 1: Travelled the minimal hitting path 65% with their existing navigation aid and increased to 75% with the glove. - Course 2: Travelled the minimal hitting path 65% with their existing navigation aid and decreased to 57% with the glove.

							
[229] [230]	Framework for delivering haptic data along with audio video data from an entertainment perspective. Specifically, adding a haptic layer to the MPEG 4 audio video coding.	76	Vibrotactors are added all over the glove both on top and bottom of the hand. No specific configuration pattern is discussed in the paper.	Custom designed vibration patterns that take into account all the vibrators on the glove.	Manually encoded by entertainment specialists based on the movie and the scene.	No user study.	No user study.
[231]	Using vibrators to convey slip information in a prehensile glove.	5	Fingertips of the five fingers.	Motion sensors (optical motion sensor similar to the one used in an optical mouse) mounted outside the glove on the finger tips measure the slip of an object. The slip information measured as optic flow is conveyed to the vibrator as varying frequency.	Slip motion is proportional to the frequency of vibration.	Users placed the glove on a surface that was laterally pulled from under the glove and the reaction time was measured by asking the participants to press a button with their free hand. Experiment was conducted with bare hands, with a prehensile glove without vibrators and with the slip glove.	12 subjects. <u>Mean reaction time:</u> Bare hand: 0.214s Normal Glove: 1.669s Slip Glove: 0.483s  <u>Percent Failure:</u> Bare hand: 0% Normal Glove: 27.8% Slip Glove: 5.6%

### 1.7.5 Summary:

*Vibraotactile cueing provides immense opportunity to deliver high bandwidth information through the use of somatosensory channels. Unfortunately, not much research exists in the development of delivery devices for assistive aid. Research is needed in designing high bandwidth delivery form factors and determining the right encoding on the haptic signals to allow effective delivery of information.*

### 1.8 Research Questions:

Social interactions are vital for everyday living and it is very important for the development of social learning and social feedback in human interpersonal communication. Most part of this communication happens through the use of visual non-verbal cues that put people who are blind or visually impaired at a disadvantage. While the problem of social interaction assistant remains unattended, couple of existing computer vision and signal processing technologies offers possibility of building such assistive device. To this end, we identify some of the important research questions that need to be addressed towards developing effective social interaction assistant.

1. What non-verbal cues are important from the perspective of an individual who is blind or visually impaired?
2. What assistive technology framework can be developed towards addressing the important social needs of individuals who are blind and visually impaired?
3. How effectively can the non-verbal assistive and rehabilitative cues be identified from state-of-the-art sensors used in developing the above social interaction assistant framework?
  - a. How effectively can social interaction cues be identified from an exocentric perspective using camera as the primary input sensors?
  - b. How effectively can social interaction cues be identified from a egocentric perspective using body motion sensors?
4. How effectively can data be delivered back to the users of the social interaction assistant by using haptic processing technologies?

### 1.9 References:

- [1] A. Perret-Clermont, C. Pontecorvo, L.B. Resnick, T. Zittoun, and B. Burge, *Joining Society: Social Interaction and Learning in Adolescence and Youth*, Cambridge University Press, 2003.
- [2] M.L. Knapp and J.A. Hall, *Nonverbal Communication in Human Interaction*, Harcourt College Pub, 1996.
- [3] C. Segrin and J. Flora, "Poor Social Skills Are a Vulnerability Factor in the Development of Psychosocial Problems.," *Human Communication Research*, vol. 26, 2000, pp. 489-514.
- [4] A. Beck, C. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An Inventory for Measuring Depression," *Archives of General Psychiatry*, vol. 4, Jun. 1961, pp. 571, 561.
- [5] D.W. Russell, "UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure," *Journal of Personality Assessment*, vol. 66, Feb. 1996, pp. 20-40.

- [6] R.E. Riggio, *Social Skills Inventory*, Palo Alto, CA: Consulting Psychologists Press, 1989.
- [7] R.E. Riggio, "Assessment of basic social skills," *Journal of Personality and Social Psychology*, vol. 51, 1986, pp. 649-660.
- [8] R.E. Riggio and J. Zimmermann, "Social skills and interpersonal competence: Influences on social support and social seeking," *Advances in Personal Relationships*, W.H. Jones and D. Perlman, eds., London: Jessica Kingsley, 1991, pp. 133-155.
- [9] D. Magnusson, "An Analysis of Situational Dimensions," *Perceptual and Motor Skills*, vol. 32, 1991, pp. 851-867.
- [10] H.E. Gardner, *Frames Of Mind: The Theory Of Multiple Intelligences*, Basic Books, 1993.
- [11] T. Bradberry and J. Greaves, *Emotional Intelligence 2.0*, TalentSmart, 2009.
- [12] E.L. Thorndike, "Intelligence and its uses," *Harper's Magazine*, vol. 140, 1920, pp. 227-235.
- [13] K. Albrecht, *Social Intelligence: The New Science of Success*, Pfeiffer, 2005.
- [14] G. Matthews, M. Zeidner, and R.D. Roberts, *Science of Emotional Intelligence: Knowns and Unknowns*, Oxford University Press, USA, 2007.
- [15] D. Goleman, *Working with Emotional Intelligence*, Bantam, 2000.
- [16] K.V.[. Petrides, R.[. Pita, and F.[. Kokkinaki, "The location of trait emotional intelligence in personality factor space," *British Journal of Psychology*, vol. 98, May. 2007, pp. 273-289.
- [17] N.K. Humphrey, *Vision in a monkey without striate cortex: a case study*, 1974.
- [18] L. Brothers, "The social brain: A project for integrating primate behavior and neurophysiology in a new domain.," *Concepts in Neuroscience*, vol. 1, 1990a. , pp. 51, 27.
- [19] S. Baron-Cohen, H. Ring, S. Wheelwright, E. Bullmore, M. Brammer, A. Simmons, and S. Williams, "Social intelligence in the normal and autistic brain: An fMRI study," *European Journal of Neuroscience*, vol. 11, 1999, pp. 1898, 1891.
- [20] Brent D. Ruben, *Human communication handbook*, (Rochelle Park, N.J): Hayden Book Co., 1975.
- [21] P. Borkenau, N. Mauer, R. Riemann, F. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence.," *Journal of personality and social psychology*, vol. 86, Apr. 2004, pp. 614, 599.
- [22] R. Brown, *Social Psychology*, New York, NY: Free Press, 1986.
- [23] O. Hargie, *Social Skills in Interpersonal Communication*, Routledge, 1994.
- [24] W.B. Walsh, K.H. Craik, and R.H. Price, *Person-environment psychology*, Routledge, 2000.
- [25] D.T. Kenrick and S.W. MacFarlane, "Ambient Temperature and Horn Honking: A Field Study of the Heat/Aggression Relationship," *Environment and Behavior*, vol. 18, Mar. 1986, pp. 179-191.
- [26] E. Krupat, *People in Cities: The Urban Environment and its Effects*, Cambridge University Press, 1985.
- [27] R. Sommer, *Personal Space: The Behavioral Basis of Design*, Prentice Hall Trade, 1969.
- [28] R. Sommer, *Tight spaces; hard architecture and how to humanize it*, Prentice-Hall, 1974.
- [29] A. Schauss, "The physiological effect of color on the suppression of human aggression," *International Journal of Biosocial Research*, vol. 7, 1985, pp. 55-64.
- [30] P.A. Bottomley and J.R. Doyle, "The interactive effects of colors and products on

- perceptions of brand logo appropriateness,” *Marketing Theory*, vol. 6, Mar. 2006, pp. 63-83.
- [31] T. Farrenkopf and V. Roth, “The University Faculty Office as an Environment,” *Environment and Behavior*, vol. 12, Dec. 1980, pp. 467-77.
  - [32] R.H. Moos, *The Human Context: Environmental Determinants of Behavior*, Krieger Pub Co, 1985.
  - [33] V. Manusov and J.H. Harvey, *Attribution, Communication Behavior, and Close Relationships*, Cambridge University Press, 2001.
  - [34] A.C. North, D.J. Hargreaves, and J. McKendrick, “In-store music affects product choice,” *Nature*, vol. 390, Nov. 1997, p. 132.
  - [35] J. Meer, “The light touch,” *Psychology Today*, vol. 19, 1985, pp. 60-67.
  - [36] D.S. Berry, “Attractive Faces Are not all Created Equal: Joint Effects of Facial Babyishness and Attractiveness on Social Perception,” *Pers Soc Psychol Bull*, vol. 17, Oct. 1991, pp. 523-531.
  - [37] B.H. Johnson, R.H. Nagasawa, and K. Peters, “Clothing Style Differences: Their Effect on the Impression of Sociability,” *Family and Consumer Sciences Research Journal*, vol. 6, Sep. 1977, pp. 58-63.
  - [38] Helen H. Jennings, *Sociometry in group relations*, (Washington): American Council on Education, 1959.
  - [39] L.A. Zebrowitz, *Reading Faces*, Boulder CO: Westview Press, 1997.
  - [40] D.S. Berry and L.Z. McArthur, “Perceiving character in faces: the impact of age-related craniofacial changes on social perception,” *Psychological Bulletin*, vol. 100, Jul. 1986, pp. 3-18.
  - [41] J.B. Cortés and F.M. Gatti, “Physique and self-description of temperament,” *Journal of Consulting Psychology*, vol. 29, Oct. 1965, pp. 432-439.
  - [42] L.A. Tucker, “Physical Attractiveness, Somatotype, and the Male Personality: A Dynamic Interactional Perspective,” *Journal of Clinical Psychology*, vol. 40, 1984, pp. 1226-34.
  - [43] C. Cameron, S. Oskamp, and W. Sparks, “Courtship American Style: Newspaper Ads,” *The Family Coordinator*, vol. 26, Jan. 1977, pp. 27-30.
  - [44] C.L. Ogden, K.M. Flegal, M.D. Carroll, and C.L. Johnson, “Prevalence and Trends in Overweight Among US Children and Adolescents, 1999-2000,” *JAMA*, vol. 288, Oct. 2002, pp. 1728-1732.
  - [45] J.H. Griffin, R. Bonazzi, J.H. Griffin, and R. Bonazzi, *Black Like Me*, Signet, 1996.
  - [46] R. Porter, “Olfaction and human kin recognition,” *Genetica*, vol. 104, Dec. 1998, pp. 259-263.
  - [47] T. Lord and M. Kasprzak, “Identification of self through olfaction,” *Perceptual and motor skills*, vol. 69, 1989, pp. 224, 219.
  - [48] M.J. RUSSELL, “Human olfactory communication,” *Nature*, vol. 260, Apr. 1976, pp. 520-522.
  - [49] N. Barber, “Mustache Fashion Covaries with a Good Marriage Market for Women,” *Journal of Nonverbal Behavior*, vol. 25, Dec. 2001, pp. 261-272.
  - [50] W.E. Hensley, “The effects of attire, location, and sex on aiding behavior: A similarity explanation,” *Journal of Nonverbal Behavior*, vol. 6, 1981, pp. 3-11.
  - [51] N. Joseph, *Uniforms and Nonuniforms: Communication Through Clothing*, Greenwood Press, 1986.
  - [52] T.L. Rosenfeld and T.G. Plax, “Clothing as communication,” *Journal of Communication*,



- vol. 27, pp. 24-31.
- [53] C. Sanders and D.A. Vail, *Customizing the Body: The Art and Culture of Tattooing*, Temple University Press, 2008.
  - [54] P. Ekman, "Nonverbal Communication: Movements with Precise Meanings," 1976.
  - [55] M. Wagner and N. Armstrong, *Field Guide to Gestures: How to Identify and Interpret Virtually Every Gesture Known to Man*, Quirk Books, 2003.
  - [56] D. Efron, *Gesture, Race and Culture*, Walter de Gruyter, Inc., 1972.
  - [57] G.E. Weisfeld and J.M. Beresford, "Erectness of posture as an indicator of dominance or success in humans," *Motivation and Emotion*, vol. 6, Jun. 1982, pp. 113-131.
  - [58] E.C. Grant and J.H. Mackintosh, "A Comparison of the Social Postures of Some Common Laboratory Rodents," *Behaviour*, vol. 21, 1963, pp. 246-259.
  - [59] A. Kleinsmith, P.R.D. Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers*, vol. 18, Dec. 2006, pp. 1371-1389.
  - [60] A. Montagu, *Touching: The Human Significance of the Skin*, Harper Paperbacks, 1986.
  - [61] W.A. Afifi and M.L. Johnson, "The Use and Interpretation of Tie Signs in a Public Setting: Relationship and Sex Differences," *Journal of Social and Personal Relationships*, vol. 16, Feb. 1999, pp. 9-38.
  - [62] "The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research.(infants)," 2006.
  - [63] M.J. Hertenstein, D. Keltner, B. App, A.B. Bulleit, and R. Jaskolta, "Touch communicates distinct emotions," *Emotion*, vol. 6, 2006, pp. 528-533.
  - [64] G. Robles-De-La-Torre, "Principles of haptic perception in virtual environments," *Human Haptic Perception: Basics and Applications*, 2008, pp. 363-379.
  - [65] L.J. Carver and G. Dawson, "Development and neural bases of face recognition in autism," *Molecular Psychiatry*, vol. 7, 2002, pp. S18-S20.
  - [66] W.E. Rinn, "The neuropsychology of facial expression: A review of neurological and psychological mechanisms for producing facial expressions," *Psychological Bulletin*, vol. 95, 1984, pp. 52-77.
  - [67] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.*, Consulting Psychologists Press, 1978.
  - [68] C.E. Izard, "maximally discriminative facial movement coding system," 1979.
  - [69] M. Argyle and M. Cook, *Gaze & Mutual Gaze*, Cambridge University Press, 1976.
  - [70] C.L. Kleinke, "Gaze and eye contact: a research review," *Psychological Bulletin*, vol. 100, Jul. 1986, pp. 78-100.
  - [71] A. Kendon, "Some functions of gaze-direction in social interaction.," *Acta Psychol (Amst)*, vol. 26, 1967, pp. 63, 22.
  - [72] M.S. Mast, "Dominance as Expressed and Inferred Through Speaking Time," *Human Communication Research*, vol. 28, 2002, pp. 420-450.
  - [73] J.B. Bavelas, L. Coates, and T. Johnson, "Listener Responses as a Collaborative Process: The Role of Gaze," *The Journal of Communication*, vol. 52, 2002, pp. 566-580.
  - [74] A.M. van Dulmen, P.F.M. Verhaak, and H.J.G. Bilo, "Shifts in Doctor-Patient Communication during a Series of Outpatient Consultations in Non-Insulin-Dependent Diabetes Mellitus.," *Patient Education and Counseling*, vol. 30, 1997, pp. 227-37.
  - [75] A.M. Glenberg, J.L. Schroeder, and D.A. Robertson, "Averting the gaze disengages the environment and facilitates remembering," *Memory & Cognition*, vol. 26, Jul. 1998, pp.

- 651-658.
- [76] J. Orozco, O. Rudovic, F. Roca, and J. Gonzalez, "Confidence assessment on eyelid and eyebrow expression recognition," *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, 2008, pp. 1-8.
  - [77] D. Jindal-Snape, "Generalization and Maintenance of Social Skills of Children with Visual Impairments: Self-Evaluation and the Role of Feedback," *Journal of Visual Impairment and Blindness*, vol. 98, 2004, pp. 470-483.
  - [78] D. Jindal-Snape, "Use of Feedback from Sighted Peers in Promoting Social Interaction Skills," *Journal of Visual Impairment and Blindness*, vol. 99, Jul. 2005, pp. 1-16.
  - [79] D. Jindal-Snape, "Using self-evaluation procedures to maintain social skills in a child who is blind," *Journal of Visual Impairment and Blindness*, vol. 92, 1998, pp. 362-366.
  - [80] C.G. McGaha and D.C. Farran, "Interactions in an Inclusive Classroom: The Effects of Visual Status and Setting.," *Journal of Visual Impairment & Blindness*, vol. 95, 2001, pp. 80-94.
  - [81] L. Kekelis, *The Development of Social Skills by Blind and Visually Impaired Students: Exploratory Studies and Strategies*, Amer Foundation for the Blind, 1992.
  - [82] T. D'Allura, "Enhancing the Social Interaction Skills of Preschoolers with Visual Impairments.," *Journal of Visual Impairment & Blindness*, vol. 96, 2002, pp. 576-84.
  - [83] J.N. Felps and R.J. Devlin, "Modification of Stereotypic Rocking of a Blind Adult.," *Journal of Visual Impairment and Blindness*, vol. 82, 1988, pp. 107-08.
  - [84] B.B. Blasch, "Blindisms: Treatment by Punishment and Reward in Laboratory and Natural Settings," *Journal of Visual Impairment & Blindness*, 1972, pp. 215-230.
  - [85] S. Raver, "Modification of Head Droop during Conversation in a 3-Year-Old Visually Impaired Child: A Case Study," *Journal of Visual Impairment and Blindness*, vol. 78, 1984, pp. 307-10.
  - [86] R.L. Simpson and And Others, "Modification of Manneristic Behavior in a Blind Child via a Time-Out Procedure," *Education of the Visually Handicapped*, vol. 14, 1982, pp. 50-55.
  - [87] R.L. Ohlsen, "Control of body rocking in the blind through the use of vigorous exercise," *Journal of Instructional Psychology*, vol. 5, 1978, pp. 19-22.
  - [88] D.B. McAdam and C.M. O'Cleirigh, "Self-monitoring and verbal feedback to reduce stereotypic body rocking in a congenitally blind adult," *Re:View*, vol. 24, Winter93. 1993, p. 163.
  - [89] D.B. Shabani, D.A. Wilder, and W.A. Flood, "Reducing stereotypic behavior through discrimination training, differential reinforcement of other behavior, and self-monitoring.," *Behavioral Interventions*, vol. 16, Oct. 2001, pp. 279-286.
  - [90] A.H. Estevis and A.J. Koenig, "A cognitive approach to reducing stereotypic body rocking.," *Re:View*, vol. 26, Fall94. 1994, p. 119.
  - [91] P.J. Schloss and M.A. Smith, "Increasing appropriate behavior through related personal characteristics," *Applied Behavior Analysis in the Classroom*, Boston: Allyn & Bacon, 1994.
  - [92] G. Cartledge, *Teaching Social Skills to Children: Innovative Approaches*, Allyn & Bacon, 1986.
  - [93] S. Raver and P.W. Darsh, "Increasing social skills training for visually impaired children," *Education of the Visually Handicapped*, vol. 19, 1988, pp. 147-155.
  - [94] K. Shinohara and J. Tenenberg, "A blind person's interactions with technology," *Commun.*

- ACM, vol. 52, 2009, pp. 58-66.
- [95] K. Shinohara, "Designing assistive technology for blind users," *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Portland, Oregon, USA: ACM, 2006, pp. 293-294.
  - [96] K. Shinohara and J. Tenenbergs, "Observing Sara: a case study of a blind person's interactions with technology," *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, Tempe, Arizona, USA: ACM, 2007, pp. 171-178.
  - [97] V. Kulyukin, C. Gharpure, J. Nicholson, and G. Osborne, "Robot-assisted wayfinding for the visually impaired in structured indoor environments," *Autonomous Robots*, vol. 21, 2006, pp. 29-41.
  - [98] D. Yuan and R. Manduchi, "Dynamic environment exploration using a virtual white cane," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 243-249 vol. 1.
  - [99] S. ur Rehman, Li Liu, and Haibo Li, "Manifold of Facial Expressions for Tactile Perception," 2007, pp. 239-242.
  - [100] S. Rehman, L. Liu, and H. Li, "Vibrotactile Rendering of Human Emotions on the Manifold of Facial Expressions," *Journal of Multimedia*, vol. 3, 2008, pp. 18-25.
  - [101] A. Teeters, R. Kaliouby, and R. Picard, "Self-Cam: feedback from what would be your social partner," *SIGGRAPH '06: ACM SIGGRAPH 2006 Research posters*, Boston, Massachusetts: ACM, 2006, p. 138.
  - [102] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signals, their function, and automatic analysis: a survey," *Proceedings of the 10th international conference on Multimodal interfaces*, Chania, Crete, Greece: ACM, 2008, pp. 61-68.
  - [103] T. Kim, A. Chang, L. Holland, and A. Pentland, "Meeting Mediator: Enhancing Group Collaboration and Leadership with Sociometric Feedback," San Diego, CA, USA: 2008, pp. 457-466.
  - [104] A. Pentland, *Honest Signals: How They Shape Our World*, The MIT Press, 2008.
  - [105] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: state-of-the-art and future perspectives of an emerging domain," *Proceeding of the 16th ACM international conference on Multimedia*, Vancouver, British Columbia, Canada: ACM, 2008, pp. 1061-1070.
  - [106] R.E. Transon, "Using the feedback band device to control rocking behavior," *Journal of Visual Impairment & Blindness*, vol. 82, 1988, pp. 287 - 289.
  - [107] A. Adam, E. Rivlin, and I. Shimshoni, "Aggregated Dynamic Background Modeling," 2006, pp. 3313-3316.
  - [108] Ming-Hsuan Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, 2002, pp. 34-58.
  - [109] P. Viola and M. Jones, "Robust Real-time Object Detection," *International Journal of Computer Vision*, 2001.
  - [110] Yijun Xiao, N. Werghi, and P. Siebert, "A topological approach for segmenting human body shape," 2003, pp. 82-87.
  - [111] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker, "Detailed Human Shape and Pose from Images," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 8, 1.

- [112] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *Image Processing, IEEE Transactions on*, vol. 14, 2005, pp. 294-307.
- [113] K.E. Ozden and L.V. Gool, "Background Recognition in Dynamic Scenes with Motion Constraints," IEEE Computer Society, 2005, pp. 250-255.
- [114] Y. Ren, C. Chua, and Y. Ho, "Statistical background modeling for non-stationary camera," *Pattern Recogn. Lett.*, vol. 24, 2003, pp. 183-196.
- [115] S. Todorovic and M.C. Nechyba, "Detection of Artificial Structures in Natural-Scene Images Using Dynamic Trees," IEEE Computer Society, 2004, pp. 35-39.
- [116] M. Barnard, M. Matilainen, and J. Heikkila, "Body part segmentation of noisy human silhouette images," 2008, pp. 1189-1192.
- [117] P. Srinivasan and Jianbo Shi, "Bottom-up Recognition and Parsing of the Human Body," 2007, pp. 1-8.
- [118] Xuelong Li, S. Maybank, Shuicheng Yan, Dacheng Tao, and Dong Xu, "Gait Components and Their Application to Gender Recognition," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 38, 2008, pp. 145-155.
- [119] "Person Identification Using Automatic Height and Stride Estimation," IEEE Computer Society, 2002, p. 40377.
- [120] R. Collins, R. Gross, and Jianbo Shi, "Silhouette-based human identification from body shape and gait," *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, 2002, pp. 366-371.
- [121] A. Gallagher and Tsuhan Chen, "Clothing cosegmentation for recognizing people," 2008, pp. 1-8.
- [122] Wei Zhang, Bo Begole, M. Chu, Juan Liu, and Nicholas Yee, "Real-time clothes comparison based on multi-view vision," *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, 2008, pp. 1-10.
- [123] Y. Yacoob and L. Davis, "Detection and analysis of hair," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, 2006, pp. 1164-1169.
- [124] T. Sano and H. Yamamoto, "Human body shape imaging for Japanese kimono design," 2004, pp. 1120-1123 Vol.2.
- [125] Jung-Eun Lee, A. Jain, and Rong Jin, "Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification," *Biometrics Symposium, 2008. BSYM '08*, 2008, pp. 1-8.
- [126] Jingyu Yan and M. Pollefeys, "A Factorization-Based Approach for Articulated Nonrigid Shape, Motion and Kinematic Chain Recovery From Video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, 2008, pp. 865-877.
- [127] Jaewon Sung and Daijin Kim, "Combining Local and Global Motion Estimators for Robust Face Tracking," 2007, pp. 345-350.
- [128] Shan Du and R. Ward, "A Robust Approach for Eye Localization Under Variable Illuminations," 2007, pp. I - 377-I - 380.
- [129] Huchuan Lu and Hui Lin, "Gender Recognition using Adaboosted Feature," 2007, pp. 646-650.
- [130] Haibin Ling, S. Soatto, N. Ramanathan, and D. Jacobs, "A Study of Face Recognition as People Age," 2007, pp. 1-8.
- [131] Xuelong Li, S. Maybank, and Dacheng Tao, "Gender recognition based on local body motions," 2007, pp. 3881-3886.
- [132] F. Matta and J. Dugelay, "A Behavioural Approach to Person Recognition," 2006, pp. 1461-1464.

- [133] M. Hahnel, D. Klunder, and K. Kraiss, "Color and texture features for person recognition," 2004, p. 652.
- [134] U. Saeed, F. Matta, and J. Dugelay, "Person Recognition based on Head and Mouth Dynamics," *Multimedia Signal Processing, 2006 IEEE 8th Workshop on*, 2006, pp. 29-32.
- [135] U. Saeed and J. Dugelay, "Person Recognition Form Video using Facial Mimics," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. I-493-I-496.
- [136] M. Kawade, "Vision-based face understanding technologies and applications," *Micromechatronics and Human Science, 2002. MHS 2002. Proceedings of 2002 International Symposium on*, 2002, pp. 27-32.
- [137] A. Kanaujia, Yuchi Huang, and D. Metaxas, "Emblem Detections by Tracking Facial Features," *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, 2006, p. 108.
- [138] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," 2005, pp. 3437-3443 Vol. 4.
- [139] D. Kulic, W. Takano, and Y. Nakamura, "Combining automated on-line segmentation and incremental clustering for whole body motions," 2008, pp. 2591-2598.
- [140] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, 2000, pp. 831-843.
- [141] J. Hwang, I. Karliga, and H. Cheng, "An automatic three-dimensional human behavior analysis system for video surveillance applications," 2006, p. 4 pp.
- [142] Sangho Park and J. Aggarwal, "Segmentation and tracking of interacting human body parts under occlusion and shadowing," *Motion and Video Computing, 2002. Proceedings. Workshop on*, 2002, pp. 105-111.
- [143] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, 2007, pp. 311-324.
- [144] J. Nunamaker, G. Tsechpenakis, D. Metaxas, M. Adkins, J. Kruse, J. Burgoon, M. Jensen, T. Meservy, D. Twitchell, and A. Deokar, "HMM-Based Deception Recognition from Visual Cues," 2005, pp. 824-827.
- [145] I. Bacivarov, M. Ionita, and P. Corcoran, "Statistical models of appearance for eye tracking and eye-blink detection and measurement," *Consumer Electronics, IEEE Transactions on*, vol. 54, 2008, pp. 1312-1320.
- [146] T. Meservy, M. Jensen, J. Kruse, J. Burgoon, J. Nunamaker, D. Twitchell, G. Tsechpenakis, and D. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *Intelligent Systems, IEEE*, vol. 20, 2005, pp. 36-43.
- [147] T. Funahashi, T. Fujiwara, and H. Koshimizu, "Face and eye tracking for gaze analysis," *Control, Automation and Systems, 2007. ICCAS '07. International Conference on*, 2007, pp. 1337-1341.
- [148] A. Villanueva and R. Cabeza, "A Novel Gaze Estimation System With One Calibration Point," *Systems, Man, and Cybernetics, Part B, IEEE Transactions on*, vol. 38, 2008, pp. 1123-1138.
- [149] A. Fawky, S. Khalil, and M. Elsabrouty, "Eye detection to assist drowsy drivers," 2007, pp. 131-134.
- [150] U. Rajashekar, I. van der Linde, A. Bovik, and L. Cormack, "GAFFE: A Gaze-Attentive

- Fixation Finding Engine,” *Image Processing, IEEE Transactions on*, vol. 17, 2008, pp. 564-573.
- [151] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Jul. 2008.
  - [152] J. Rurainsky and P. Eisert, “Mirror-Based Multi-View Analysis of Facial Motions,” *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 2007, pp. III - 73-III - 76.
  - [153] M. Yeasin, B. Bullo, and R. Sharma, “Recognition of facial expressions and measurement of levels of interest from video,” *Multimedia, IEEE Transactions on*, vol. 8, 2006, pp. 500-508.
  - [154] S. Fukuda, “Detecting Emotions and Dangerous Actions for Better Human-System Team Working,” 2008, pp. 205-206.
  - [155] Yinggang Xie, Zhiliang Wang, Ning Cheng, Guojiang Wang, and M. Nagai, “Facial and Eye Detection and Application in Affective Recognition,” *Control Conference, 2006. CCC 2006. Chinese*, 2006, pp. 1942-1946.
  - [156] Zhihong Zeng, M. Pantic, G. Roisman, and T. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, 2009, pp. 39-58.
  - [157] A.B. Ashraf, S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K.M. Prkachin, and P.E. Solomon, “The painful face - Pain expression recognition using active appearance models,” *Image Vision Comput.*, vol. 27, 2009, pp. 1788-1796.
  - [158] M.S. Bartlett, G. Littlewort, P. Braathen, T.J. Sejnowski, and J.R. Movellan, “A Prototype for Automatic Recognition of Spontaneous Facial Actions,” 2003, pp. 1271-1278.
  - [159] M. Yeasin, B. Bullo, and R. Sharma, “Recognition of facial expressions and measurement of levels of interest from video,” *Multimedia, IEEE Transactions on*, vol. 8, 2006, pp. 500-508.
  - [160] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Fully Automatic Facial Action Recognition in Spontaneous Behavior,” IEEE Computer Society, 2006, pp. 223-230.
  - [161] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, “Facial expression recognition from video sequences: temporal and static modeling,” *Computer Vision and Image Understanding*, vol. 91, Aug. 2003, pp. 160-187.
  - [162] J. Cohn, L. Reed, Z. Ambadar, Jing Xiao, and T. Moriyama, “Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior,” 2004, pp. 610-616 vol.1.
  - [163] R.E. Kaliouby and P. Robinson, “Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures,” IEEE Computer Society, 2004, p. 154.
  - [164] B. Fasel, F. Monay, and D. Gatica-Perez, “Latent semantic analysis of facial action codes for automatic facial expression recognition,” New York, NY, USA: ACM, 2004, pp. 181-188.
  - [165] H. Gunes and M. Piccardi, “Affect recognition from face and body: early fusion vs. late fusion,” 2005, pp. 3443 Vol. 4, 3437.
  - [166] S.V. Ioannou, A.T. Raouzaoui, V.A. Tzouvaras, T.P. Mailis, K.C. Karpouzis, and S.D. Kollias, “Emotion recognition through facial expression analysis based on a neurofuzzy network,” *Neural Netw.*, vol. 18, 2005, pp. 423-435.

- [167] Qiang Ji, P. Lan, and C. Looney, "A probabilistic framework for modeling and real-time monitoring human fatigue," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 36, 2006, pp. 862-875.
- [168] A. Kapoor and R.W. Picard, "Multimodal affect recognition in learning environments," Hilton, Singapore: ACM, 2005, pp. 677-682.
- [169] A. Kapoor, W. Burleson, and R.W. Picard, "Automatic prediction of frustration," *Int. J. Hum.-Comput. Stud.*, vol. 65, 2007, pp. 724-736.
- [170] C. Lee and A. Elgammal, "Facial Expression Analysis Using Nonlinear Decomposable Generative Models," *Analysis and Modelling of Faces and Gestures*, 2005, pp. 17-31.
- [171] G.C. Littlewort, M.S. Bartlett, and K. Lee, "Faces of pain: automated measurement of spontaneous all facial expressions of genuine and posed pain," Nagoya, Aichi, Japan: ACM, 2007, pp. 15-21.
- [172] S. Lucey, A. Bilal, and J. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," *Face Recognition Book*, K. Kurihara, ed., Pro Literatur Verlag, 2007.
- [173] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 36, Apr. 2006, pp. 433-449.
- [174] M. Pantic and L. Rothkrantz, "Case-based reasoning for user-profiled recognition of emotions from face images," 2004, pp. 391-394 Vol.1.
- [175] N. Sebe, M. Lew, I. Cohen, Yafei Sun, T. Gevers, and T. Huang, "Authentic facial expression analysis," 2004, pp. 517-522.
- [176] Yan Tong, Wenhui Liao, and Qiang Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, 2007, pp. 1683-1699.
- [177] M.F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," Nagoya, Aichi, Japan: ACM, 2007, pp. 38-45.
- [178] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," Banff, Alberta, Canada: ACM, 2006, pp. 162-170.
- [179] H. Wang and N. Ahuja, "Facial Expression Decomposition," IEEE Computer Society, 2003, p. 958.
- [180] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution," IEEE Computer Society, 2006, pp. 1399-1406.
- [181] Zhen Wen and T. Huang, "Capturing subtle facial motions in 3D face tracking," 2003, pp. 1343-1350 vol.2.
- [182] J. Whitehill and C.W. Omlin, "Haar Features for FACS AU Recognition," IEEE Computer Society, 2006, pp. 97-101.
- [183] Zhihong Zeng, Yun Fu, Glenn I. Roisman, Zhen Wen, Yuxiao Hu, and Thomas S. Huang, "Spontaneous Emotional Facial Expression Detection," 2006.
- [184] L. Bao and S.S. Intille, "Activity recognition from user-annotated acceleration data," 2004, pp. 1-17.
- [185] J.B. Bussmann and W.L. Martens, "Second International Conference Proceedings, Pervasive Computing," *Behavior Research Methods, Instruments, & Computers*, vol. 33, 2001, pp. 349-356.

- [186] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
- [187] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, 1989, pp. 257-286.
- [188] O. Amft, H. Junker, and G. Troster, "Detection of eating and drinking arm gestures using inertial body-worn sensors," 2005, pp. 160-163.
- [189] G. Chambers, S. Venkatesh, G. West, and H. Bui, "Hierarchical recognition of intentional human gestures for sports video annotation," 2002, pp. 1082-1085 vol.2.
- [190] Seon-Woo Lee and K. Mase, "Activity and location recognition using wearable sensors," *Pervasive Computing, IEEE*, vol. 1, 2002, pp. 24-32.
- [191] F. Foerster, M. Smeja, and J. Fahrenberg, "Detection of posture and motion by accelerometry: a validation in ambulatory monitoring," *Computer in Human Behavior*, vol. 15, 1999, pp. 571-583.
- [192] S. Arteaga, J. Chevalier, A. Coile, A.W. Hill, S. Sali, S. Sudhakhrisnan, and S.H. Kurniawan, "Low-cost accelerometry-based posture monitoring system for stroke survivors," Halifax, Nova Scotia, Canada: ACM, 2008, pp. 243-244.
- [193] J. Scheibert, S. Leurent, A. Prevost, and G. Debregeas, "The Role of Fingerprints in the Coding of Tactile Information Probed with a Biomimetic Sensor," *Science*, Jan. 2009, p. 1166467.
- [194] Y.K. Dillon, J. Haynes, and M. Henneberg, "The relationship of the number of Meissner's corpuscles to dermatoglyphic characters and finger size," *Journal of Anatomy*, vol. 199, 2001, pp. 577-584.
- [195] A. Iggo and A.R. Muir, "The structure and function of a slowly adapting touch corpuscle in hairy skin," *The Journal of Physiology*, vol. 200, Feb. 1969, pp. 763-796.
- [196] H. Hensel, "Functional and structural basis of thermoreception," *Progress in Brain Research*, vol. 43, 1976, pp. 105-118.
- [197] M. Paré, C. Behets, and O. Cornu, "Paucity of presumptive ruffini corpuscles in the index finger pad of humans," *The Journal of Comparative Neurology*, vol. 456, 2003, pp. 260-266.
- [198] A. Nakamura, T. Yamada, A. Goto, T. Kato, K. Ito, Y. Abe, T. Kachi, and R. Kakigi, "Somatosensory homunculus as drawn by MEG," *NeuroImage*, vol. 7, May. 1998, pp. 377-386.
- [199] S. Brave and A. Dahley, "inTouch: a medium for haptic interpersonal communication," *CHI '97 extended abstracts on Human factors in computing systems: looking to the future*, Atlanta, Georgia: ACM, 1997, pp. 363-364.
- [200] C. Dodge, "The bed: a medium for intimate communication," *CHI '97 extended abstracts on Human factors in computing systems: looking to the future*, Atlanta, Georgia: ACM, 1997, pp. 371-372.
- [201] B.J. Fogg, L.D. Cutler, P. Arnold, and C. Eisbach, "HandJive: a device for interpersonal haptic entertainment," *Proceedings of the SIGCHI conference on Human factors in computing systems*, Los Angeles, California, United States: ACM Press/Addison-Wesley Publishing Co., 1998, pp. 57-64.
- [202] O. Morikawa, J. Yamashita, and Y. Fukui, "The sense of physically crossing paths: creating a soft initiation in HyperMirror communication," *CHI '00 extended abstracts on Human factors in computing systems*, The Hague, The Netherlands: ACM, 2000, pp. 183-184.



- [203] K. Dobson, D. boyd, W. Ju, J. Donath, and H. Ishii, "Creating visceral personal and social interactions in mediated spaces," *CHI '01 extended abstracts on Human factors in computing systems*, Seattle, Washington: ACM, 2001, pp. 151-152.
- [204] M.O. Alhalabi, S. Horiguchi, and S. Kunifuji, "An experimental study on the effects of Network delay in Cooperative Shared Haptic Virtual Environment," *Computers & Graphics*, vol. 27, Apr. 2003, pp. 205-213.
- [205] A. Chang, S. O'Modhrain, R. Jacob, E. Gunther, and H. Ishii, "ComTouch: design of a vibrotactile communication device," *DIS '02: Proceedings of the conference on Designing interactive systems*, ACM Press, 2002, pp. 320, 312.
- [206] A. Rovers and H.V. Essen, "HIM: a framework for haptic instant messaging," *CHI '04 extended abstracts on Human factors in computing systems*, Vienna, Austria: ACM, 2004, pp. 1313-1316.
- [207] F.' Mueller, F. Vetere, M.R. Gibbs, J. Kjeldskov, S. Pedell, and S. Howard, "Hug over a distance," *CHI '05 extended abstracts on Human factors in computing systems*, Portland, OR, USA: ACM, 2005, pp. 1673-1676.
- [208] L. Bonanni, C. Vaucelle, J. Lieberman, and O. Zuckerman, "TapTap: a haptic wearable for asynchronous distributed touch therapy," *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, Montr'\{e}al, Qu'\{e}bec, Canada: ACM, 2006, pp. 585, 580.
- [209] J. Werner, R. Wettach, and E. Hornecker, "United-pulse: feeling your partner's pulse," *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, Amsterdam, The Netherlands: ACM, 2008, pp. 535-538.
- [210] S. Brewster and L. Brown, "Tactons: structured tactile messages for non-visual information display," *AUIC '04: Proceedings of the fifth conference on Australasian user interface*, Australian Computer Society, Inc., 2004, pp. 23, 15.
- [211] S. Ertan, C. Lee, A. Willets, H. Tan, and A. Pentland, "A wearable haptic navigation guidance system," *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, 1998, pp. 164-165.
- [212] K. Tsukada and M. Yasumura, "ActiveBelt: Belt-Type Wearable Tactile Display for Directional Navigation," *UbiComp 2004: Ubiquitous Computing*, 2004, pp. 399, 384.
- [213] L. Jones, M. Nakamura, and B. Lockyer, "Development of a tactile vest," *Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2004. HAPTICS '04. Proceedings. 12th International Symposium on*, 2004, pp. 82-89.
- [214] A.H. Rupert, "An instrumentation solution for reducing spatial disorientation mishaps," *IEEE Engineering in Medicine and Biology Magazine: The Quarterly Magazine of the Engineering in Medicine & Biology Society*, vol. 19, Apr. 2000, pp. 71-80.
- [215] R. Traylor and H.Z. Tan, "Development of a Wearable Haptic Display for Situation Awareness in Altered-gravity Environment: Some Initial Findings," *Proceedings of the 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, IEEE Computer Society, 2002, p. 159.
- [216] C. Wall and M. Weinberg, "Balance prostheses for postural control," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 22, 2003, pp. 84-90.
- [217] R.W. Cholewiak, J.C. Brill, and A. Schwab, "Vibrotactile localization on the abdomen: effects of place and space," *Perception & Psychophysics*, vol. 66, Aug. 2004, pp. 970-987.
- [218] J.B.F.V. Erp, H.A.H.C.V. Veen, C. Jansen, and T. Dobbins, "Waypoint navigation with a

- vibrotactile waist belt,” *ACM Trans. Appl. Percept.*, vol. 2, 2005, pp. 106-117.
- [219] R. Lindeman, Y. Yanagida, H. Noma, and K. Hosaka, “Wearable vibrotactile systems for virtual contact and information display,” *Virtual Reality*, vol. 9, Mar. 2006, pp. 203-213.
  - [220] W. Heuten, N. Henze, S. Boll, and M. Pielot, “Tactile wayfinder: a non-visual support system for wayfinding,” *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, Lund, Sweden: ACM, 2008, pp. 172-181.
  - [221] L.A. Jones and K. Ray, “Localization and Pattern Recognition with Tactile Displays,” *Proceedings of the 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, IEEE Computer Society, 2008, pp. 33-39.
  - [222] A. Ferscha, B. Emsenhuber, A. Riener, C. Holzman, M. Hechinger, D. Hochreiter, M. Franz, D. Zeider, M.D.S. Rocha, and C. Klein, “Vibro-tactile space awareness,” 2008.
  - [223] L. Cappelletti, M. Feeri, and G. Nicoletti, “Vibrotactile colour rendering for the visually impaired within the VIDET project,” 1998, pp. 92-96.
  - [224] T. Oron-Gilad, J. Downs, R. Gilson, and P. Hancock, “Vibrotactile Guidance Cues for Target Acquisition,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, 2007, pp. 993-1004.
  - [225] H. Uchiyama, M.A. Covington, and W.D. Potter, “Vibrotactile Glove guidance for semi-autonomous wheelchair operations,” *Proceedings of the 46th Annual Southeast Regional Conference on XX*, Auburn, Alabama: ACM, 2008, pp. 336-339.
  - [226] A. Hein and M. Brell, “conTACT - A Vibrotactile Display for Computer Aided Surgery,” *World Haptics Conference*, Los Alamitos, CA, USA: IEEE Computer Society, 2007, pp. 531-536.
  - [227] M. Brell, D. Roßkamp, and A. Hein, “Fusion of Vibrotactile Signals Used in a Tactile Display in Computer Aided Surgery,” *Haptics: Perception, Devices and Scenarios*, 2008, pp. 383-388.
  - [228] J.S. Zelek, S. Bromley, D. Asmar, and D. Thompson, “A Haptic glove as a tactile-vision sensory substitution for wayfinding,” *Journal of Visual Impairment & Blindness*, vol. 97, 2003, pp. 1-24.
  - [229] J. Cha, Y. Seo, Y. Kim, and J. Ryu, “An Authoring/Editing Framework for Haptic Broadcasting: Passive Haptic Interactions using MPEG-4 BIFS,” *EuroHaptics Conference, 2007 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics 2007. Second Joint*, 2007, pp. 274-279.
  - [230] J. Cha, Y. Ho, Y. Kim, J. Ryu, and I. Oakley, “A Framework for Haptic Broadcasting,” *IEEE MultiMedia*, vol. 16, 2009, pp. 16-27.
  - [231] A.M. Murray, R.L. Klatzky, and P.K. Khosla, “Psychophysical characterization and testbed validation of a wearable vibrotactile glove for telemanipulation,” *Presence: Teleoper. Virtual Environ.*, vol. 12, 2003, pp. 156-182.

## Chapter 2

# Systematic Analysis of non-verbal cue requirements and design of the Social Interaction Assistant

From Research Question 1, it is certain that there is no analysis on the needs of the visually impaired and the blind community about their needs during social interactions. Though research supports the need for social interactions, no efforts have been taken towards determining the specific necessities that this community has. In order to identify the unmet needs of the visually impaired community, two focus groups consisting primarily of people who are blind, as well as disability specialists and parents of students with visual impairment and blindness were conducted<sup>3</sup>. Members of these focus groups who were blind or visually impaired were encouraged to speak freely about their challenges in coping with daily living. During these focus groups, the participants agreed on many issues as being important problems. However, one particular problem - that of engaging freely with their sighted counterparts - was highlighted as a particularly important problem that was not being addressed by technology specialists<sup>4</sup>.

As an example of the type of social disconnect that people who are visually impaired face, consider a simple form of nonverbal communication: glancing at a watch to signal that it is time to wrap up a meeting. The sighted participants might respond to such a glance automatically, without consciously realizing that this visual information is not accessible to a participant who is blind. Similarly, a sighted person asking a question in a group will use gaze direction and eye contact to indicate to whom the question is directed. Without access to this visual cue, people who are blind might be left wondering whether the question was directed to-wards them. They can answer immediately (at the risk of feeling foolish if the question was not directed at them) or they can wait to see if anyone else answers (and risk being thought of as rather slow witted).

While various other examples were cited by individuals during these focus group studies, the inability to access non-verbal cues were considered of highest priority. In this chapter, we introduce a self-report survey that was conducted based on the focus group study results that highlight the various non-verbal cues that are considered important from the perspective of the

---

<sup>3</sup> In order to understand the assistive technology requirements of people who are blind, we conducted two focus group studies (one in Tempe, Arizona USA - 9 participants, and another in Tucson, Arizona USA - 11 participants) which included:

1. students and adult professionals who are blind,
2. parents of individuals who are blind
3. professionals who work in the area of blindness and visual impairments.

There was unanimous agreement among participants that a technology that would help people with visual impairment to recognize people or hear them described would significantly enhance their social life.

<sup>4</sup> To quote some candidates opinion about face recognition technology in a social setting:

- “It would be nice to walk into a room and immediately get to know who are all in front of me before they start a conversation”.
- One young man said, “It would be great to walk into a bar and identify beautiful women”.

user population. Further, with the non-verbal cue priority list determine, the design of a platform that can be used for extracting and delivering these non-verbal cues is presented.

## 2.1 Requirements for a Social Interaction Assistant

Based on the discussions conducted through the two focus groups, a list of needs was compiled that characterized social needs often experienced by people with visual impairments. In doing so, two important aspects of social interaction were identified. These included

1. Access to the non-verbal cues of others during social interactions, and
2. How one is perceived by others during social interactions.

These needs correlated with the psychology studies conducted by Jindal-Snape with children who were visually impaired. She identifies these two needs under the *Social Learning* and *Social Feedback*. As discussed in Chapter 1, Section 1.4, these are the two important aspects of providing assistance and rehabilitation for people who are blind and visually impaired. While these two important categories were identified, for simplicity, the non-verbal cue needs were reduced to 8 aspects of social interactions that focused primarily on the physical characteristics of the interaction partner and the behaviors of the interaction partner. These questions were developed with the help of visually impaired professionals and students:

1. Knowing how many people are standing in front you, and where each person is standing.
2. Knowing where a person is directing his/her attention.
3. Knowing the identities of the people standing in front of you.
4. Knowing something about the appearance of the people standing in front of you.
5. Knowing whether the physical appearance of a person who you know has changed since the last time you encountered him/her.
6. Knowing the facial expressions of the person standing in front of you.
7. Knowing the hand gestures and body motions of the person standing in front of you.
8. Knowing whether your personal mannerisms do not fit the behavioral norms and expectations of the sighted people with whom you will be interacting.

While these 8 aspects of social interaction were important from the perspective of enriching social interactions of the people who are blind or visually impaired, it was not sufficient to just identify them, but it is important to determine the relative importance of these needs with respect to each other. To this end, an online survey was carried out to determine a self-report importance map of the various non-verbal cues. This list of questions included both the importance from the perspective of allowing access to the non-verbal cues of the interaction partner (for enabling Social Learning), while also focusing on the personal body mannerism (for enabling Social Feedback) of the individual.

## 2.2 Online Survey

The online survey was anonymously completed by 28 people, of whom 16 were blind, 9 had low vision, and 3 were sighted specialists in the area of visual impairment and vocational training. The online survey consisted of eight questions that corresponded to the previously identified list of needs. Respondents answered each question using a five-point Likert scale, the metrics being

- (1) Strongly disagree,
- (2) Disagree,

- (3) Neutral,
- (4) Agree, and
- (5) Strongly agree

The survey can be analyzed as having 3 groups (individuals who are blind, individuals with visual impairment and specialists with 20/20 vision) and 8 question groups each corresponding to the 8 aspects of social interactions that were identified from our focus group.

## 2.3 Results:

### 2.3.1 Mean Score Table:

Table 1 shows the eight aspects of social interactions, sorted by descending importance, as indicated by the survey respondents (the question numbers correspond to the need listed in the previous section). The mean score is the average of the respondents on the 5 point scale that was used to capture the opinions. A score closer to 5 implies that the respondents strongly agree with a certain question and that they consider inaccessibility to that particular non-verbal cue to be important deterrent to their social interactions. On the other hand, a score closer to 1 represents the respondent did not consider the access to a specific non-verbal cue to be important during their social interactions.

Need	The Question	Mean Score
8.	I would like to know if any of my personal mannerisms might interfere with my social interactions with others.	4.5
6.	I would like to know what facial expressions others are displaying while I am interacting with them.	4.4
3.	When I am standing in a group of people, I would like to know the names of the people around me.	4.3
7.	I would like to know what gestures or other body motions people are using while I am interacting with them.	4.2
1.	When I am standing in a group of people, I would like to know how many people there are, and where each person is.	4.1
2.	When I am standing in a group of people, I would like to know which way each person is facing, and which way they are looking.	4.0
5.	I would like to know if the appearance of others has changed (such as the addition of glasses or a new hair-do) since I last saw them.	3.5
4.	When I am communicating with other people, I would like to know what others look like.	3.4

Table 1: Results of the online survey

### 2.3.2 Histogram of Responses:

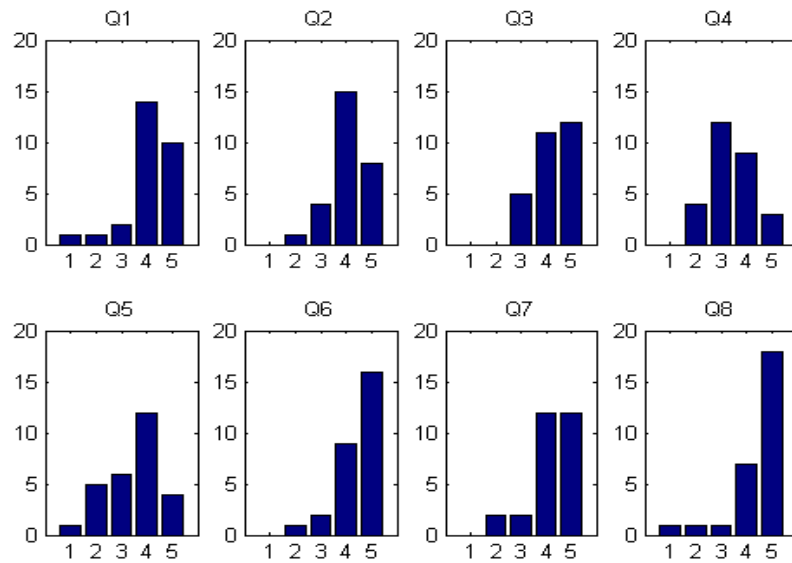


Figure 2: Histogram of Responses grouped by Questions

Figure 2 shows the histogram of responses for the 8 Questions that were asked as part of the survey. Each subplot refers to a single question and shows the number of times users responded to that particular question with answers from 1 to 5 on the Likert Scale. Each histogram adds up to a total of 28 that corresponds to the 28 participants that took part in the online survey.

### 2.3.3 Box Plot Analysis:

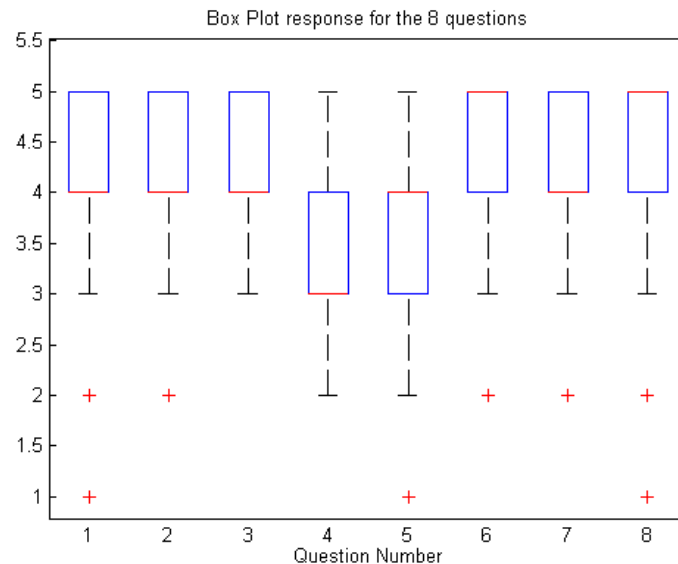


Figure 3: Box Plot of user responses for the 8 questions on the survey

The Box Plot of the 8 question responses is shown in Figure 3. The median values of the responses are shown as red lines for each of the 8 questions. While the blue box shows the enclosure for all responses between the 75 percentile and the 25 percentile points. Since the responses were on an integer scale of 1 through 5, the median coincides with the upper or lower 25 percentile. The whisker corresponds to the upper and lower limit of the values represented under that particular question. The plus marks represent any outliers under each question.

Outliers are identified based on whether they are outside the 3 sigma (variance) from the mean value. Note that the median value for questions 6 and 8 are at 5, median value of 4 for questions 1, 2, 3, 5 and 7, and median of 3 for question 4. Historically, Lickert Scale data has been analyzed using Box Plot analysis as the plot captures all the descriptive statistics of minimum value, maximum value, median, variance, and the inter quartile range that encapsulates the 50 percentile of the data around the mean.

### 2.3.4 Response Ratio:

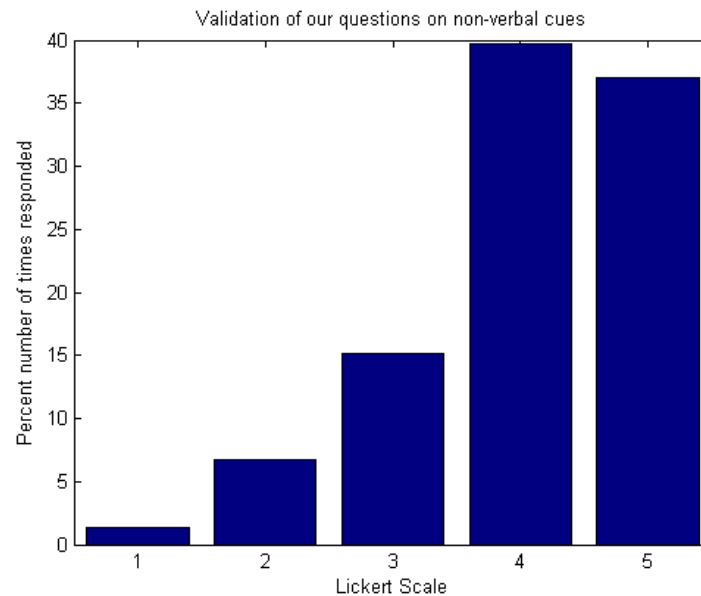


Figure 4: Response Ratio

Figure 4 shows the number of times the respondents chose to answer the 8 questions with their agreement or disagreement. The y-axis has been normalized to 100 points. The graph shows that respondents chose to answer the most by agreeing (Likert Scale 4) with the 8 questions. Followed closely behind was the strong agreement (Likert Scale 5) with the questions asked in the survey. The respondents chose to answer the least through strong disagreement (Likert Scale 1) to what was asked in the survey.

### 2.3.5 Rank Average and F-score:

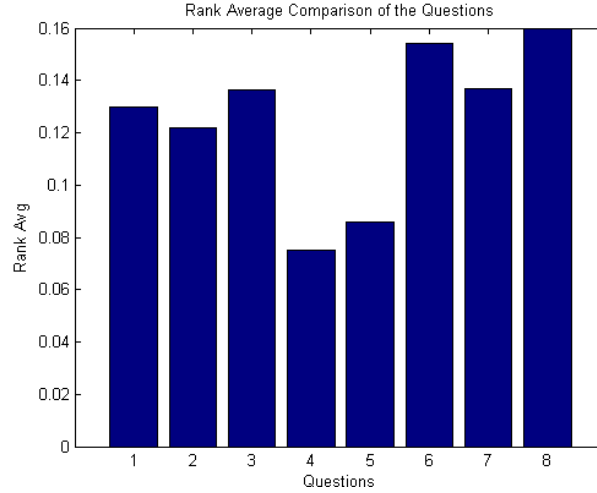


Figure 5: Rank Average of responses for the 8 questions asked in the online survey.

As can be seen from Figure 4, the questionnaires were biased and the frequency of the responses is not Gaussian. This bias implies that using sample mean of the Lickert Scale responses will immediately show the same bias. This is due to the Gaussian iid assumption that is made while extracting the mean for the answers. In order to overcome this non-Gaussianity, we resort to non-parametric mean for the responses. Rank average of the responses is estimated instead of the typical mean of the responses for each of the question. The procedure for estimation of the rank average is shown below:

1. Rank all data from all groups (question) together; i.e., rank the data from 1 to  $N$  ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied. Let this rank be referred to as  $r_{ji}$ , where  $i$  represents the group (question) and  $j$  represents the individual element.
2. Rank Average for each group is then given as

$$\bar{r}_i = \frac{\sum_{j \in G_i} r_{ji}}{n_i}$$

Where,

$\bar{r}_i$  is the average rank of the group (question)  $G_i$  with the cardinality  $n_i$ .

Further,

$$N = \sum_{i=1}^8 n_i$$

Since no assumptions on the distribution of the response are made, unlike the mean, the rank average gives a non-parametric method for comparing the responses of the individuals. The ranks can be either assigned ascending or descending with respect to the responses, i.e. rank 1 could mean all responses that were answered with strongly disagree (numeral 1), or rank 1 could mean all responses that were answered with strongly agree (numeral 5). In the Figure 5, we have assigned rank 1 to strongly disagree. This is for the sake of visual convenience. Thus, higher the average rank, higher is that group's response from the respondents. Comparing Figure 5 to Table 1, it can be seen that the same ordering of priority can be seen through mean and rank average.



But the mean tends to show very little variation between responses due to the bias that is present in the questions. On the other hand the rank average provides a good comparison scale.

### 2.3.6 Average Response per Group:

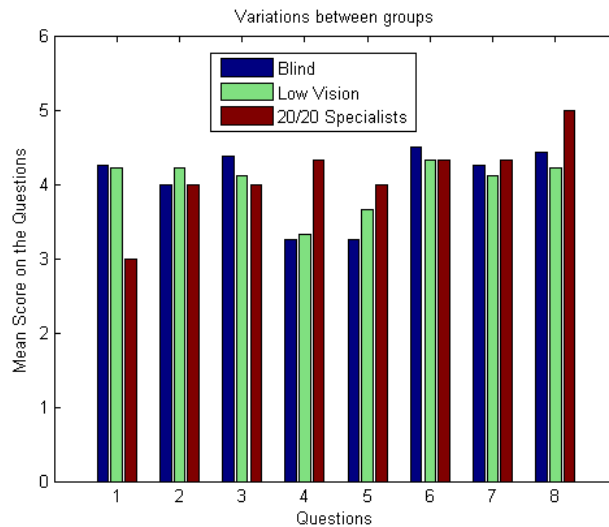


Figure 6: Average responses of the 3 user groups (Low Vision, Blind and 20/20 Specialists)  
Figure 6 shows the average responses for each question based on the group to which they belong. Based on whether the respondents belong to blindness, low vision or the sighted specialists group, the average of all the responses is plotted. In most of the questions, all three groups seem to be responding similarly, but in question 1 and 4 there is a significant deviation of the sighted respondent from the visually impaired respondents.

## 2.4 Analysis of the survey responses

### 2.4.1 Histogram of the responses:

A histogram of responses is shown in Figure 1. From the average score of the participants who took the online survey, it can be seen that,

1. Respondents are highly concerned about how their body mannerisms are perceived by their sighted peers (based on the response to Question 8 on the survey).
2. Facial expressions form the most important visual non-verbal cue that individuals who are blind or visually impaired feel they do not have access to (based on Question 6 on the survey). This correlates with the studies into non-verbal communication that highlights the importance of facial mannerisms and gestures, which are mostly visual in their decoding.
3. Followed by facial expressions, body mannerisms seem to be of higher importance for individuals who are blind and visually impaired (based on Question 3 of the survey). This can be correlated to the table shown Chapter 1, Section 1.2.1, where body follows the face in terms of displaying non-verbal cues.
4. The responses to questions 7, 1 and 2 suggest that respondents would like to know the identities of the people with whom they are communicating, relative location of these

people and whether their attentions are focused on the respondent. This corresponds to knowing the position of their interaction partners when they are involved in a bilateral or group communication. People tend to move around, especially when they are standing, causing people who are blind to lose their bearing on where people were standing. This can result in individuals addressing an empty space assuming that someone was standing there based on their memory.

5. The responses to questions 4 and 5 indicate that there was a wide variation in respondents' interest in (4) knowing the physical appearance of people with whom they are communicating and (5) knowing about changes in the physical appearance of people with whom they are communicating. Many respondents indicated moderate, little, or no interest in either of these areas.

#### **2.4.2 Box Plot Analysis:**

The Box Plot analysis reiterates the fact that Question 8 and 6 carries the highest response, with the respondents wanting to know their own body mannerism and how it was affecting the social interactions. This was immediately followed by the facial expressions of the interaction partners. Thus, self assessment in social interactions was of prime importance to these individuals.

#### **2.4.3 Response Ration - Questionnaire Bias:**

As described earlier, the 8 questions corresponding to the social needs of the individuals were identified from the focus group survey that was conducted. Thus, the questions presented in the online survey questions were biased towards the needs of everyday social interactions of individuals who are blind and visually impaired. Thus, the implicit assumption while preparing this survey itself is that most of these items have been identified as being important and that only a priority scale needs to be extracted. This implicit assumption is immediately brought out by looking at the frequency with which the respondents answer with their agreement (Likert Scale 4) and strong agreement (Likert Scale 5).

#### **2.4.4 Rank Average Response:**

Taking into account the bias that is present in the questionnaire, the rank average response for all the 8 questions indicate that the inferences that are derived from the mean analysis in Table 1 and the box plot analysis in Figure 3 are consistent and that individuals who are blind and visually impaired do consider that own body mannerism to be of utmost importance when they are involved in a social interaction. Further, facial expressions follow their egocentric body behavior as being the next most important aspect of their social interactions. The rank average correlates with mean analysis even after the questionnaire bias is removed.

#### **2.4.5 Average Response per Group:**

Finally, from Figure 6 it is seen that there is some response difference between the visually impaired (including blind) population and the sighted specialist population that were presented with the same set of questions. Though the results between populations seem consistent, for question 1 and question 4, there seems to be disagreement between the sighted and the visually impaired populations. This could be because of the smaller sample size of sighted specialists that took the survey and could purely be due to outliers. Further investigation is needed into this issue.

## 2.5 Summary:

In this study we generate a prioritized list of social cues that are considered important by people who are blind and visually impaired. Created from two focus group studies of blind and visually impaired individuals, the list of social cues are then assessed through a self-reported ranking scheme that provides the much needed prioritized social needs list.

It can be seen from this list that the people who are visually impaired and blind consider their own personal body mannerisms to be of highest importance when they are involved in social interactions. Any feedback that can be given to them about their body mannerisms will aid in their social learning. Followed closely behind their own mannerisms, the facial expressions of their interaction partners seem to be of highest importance when it comes to visual non-verbal cues.

Following this discussion of the various social needs of individuals who are blind and visually impaired, the Social Interaction Assistant platform itself is introduced below.

## 2.6 Alternative Sensing Platforms for a Social Interaction Assistant

Having determined the requirements for a Social Interaction Assistant, a potential platform for the Social Interaction Assistant is considered next. In Section 1.5.1 of Chapter 1, we discussed some of the important observations we have made on the needs for an social interaction assistive technology. The needs are listed here again for clarity.

A device that is developed to facilitate the social interactions of people with sensory, or cognitive disabilities might do so by (1) detecting social cues during social interactions and delivering that information to the user in real time to enable them to engage in social interactions, or (2) detecting the user's stereotypic behaviors during social interactions and communicating that information to the user in real time to provide social feedback. The first device might be classified as an assistive technology, while the second might be classified as a rehabilitative technology. Ideally, such a device would be based on the following design principles:

*Design principle 1:* The device should be portable and wearable so that it can be used in any social situation, and without any restriction on the user's everyday life.

*Design principle 2:* The device should employ sensors and personal signaling devices that are unobtrusive, and do not become a social distraction.

*Design principle 3:* The device should include sensors that can detect the social mannerisms of both the user and other people with whom the user might communicate.

*Design principle 4:* The device should be comfortable enough to be worn repeatedly for extended periods of time, to allow it to be used effectively for rehabilitation.

*Design principle 5:* The device should be able to reliably distinguish between the user's problematic stereotypic mannerisms and normal functional movements, to ensure that it will be worn long enough to achieve rehabilitation.

### 2.6.1 Concept Social Interaction Assistant Prototypes:

#### 2.6.1.1 Concept 1:

A wearable video camera in a clip-on device, and a small audio emitter device that could be worn on the ear without obstructing normal hearing. Both of these devices would be connected to a compact computing element such as an Ultra-mobile PC (UMPC) (Fig. 7).

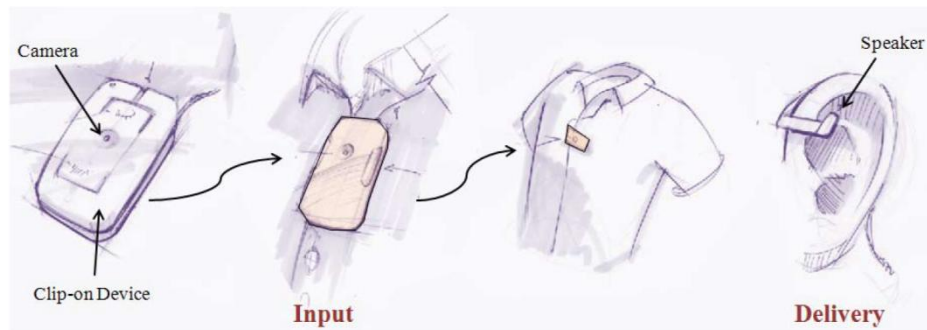


Figure 7: A clip on camera and small speaker

### 2.6.1.2 Concept 2:

A tiny, ear-mounted video camera and sound emitter (inspired by Bluetooth headsets) mounted on a small device that communicates with a UMPC (Fig. 8).



Figure 8: A ear-mounted video camera and speaker

### 2.6.1.3 Concept 3:

A tiny video camera and a sound emitter mounted unobtrusively in a pair of glasses - both of which are attached to a UMPC (Fig. 9).

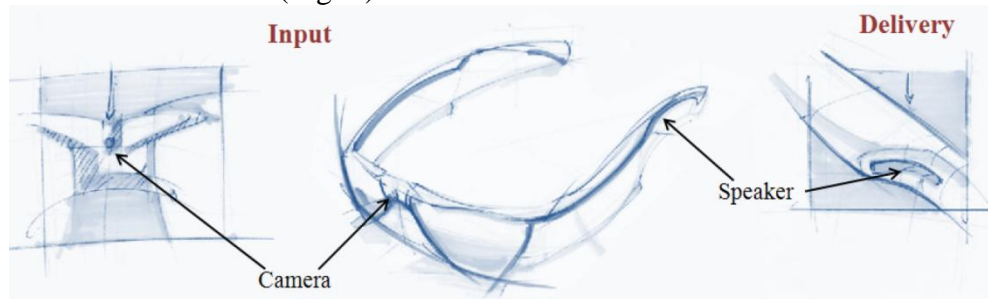


Figure 9: A tiny video camera and speaker on a pair of glasses.

## 2.6.2 Social Interaction Assistant Prototype:

Having analyzed the requirements and considering the various components of the sensing and delivery technology, we resorted to Concept 3 and incorporated the important aspects of egocentric and exocentric sensing into the prototype system.

### 2.6.2.1 System Architecture

The system level architecture of the proposed social interaction assistant is shown in the Figure 10. The sensor suite consists of:

1. A visual sensor (1.4 Megapixel camera),
2. A motion sensor ( $\pm 12g$  accelerometer), and
3. A 5-button clicker, which serves as a user interface.

The social interaction assistant software (implemented on a Windows Operating System PDA) uses these sensors to collect information about the various social and behavioral mannerisms of the user and participants in the vicinity of the user.

Interpretations of the social interactions generated by custom algorithms are communicated to the user through an actuator suite, consisting of:

1. A haptic belt, and
2. A pair of ear phones.

The haptic belt encodes information in the form of vibrotactile cues, while the ear phones provide short audio cues. As future extensions to this project, in Chapter 5, we introduce the Haptic Glove and a new interface for communicating facial affect.

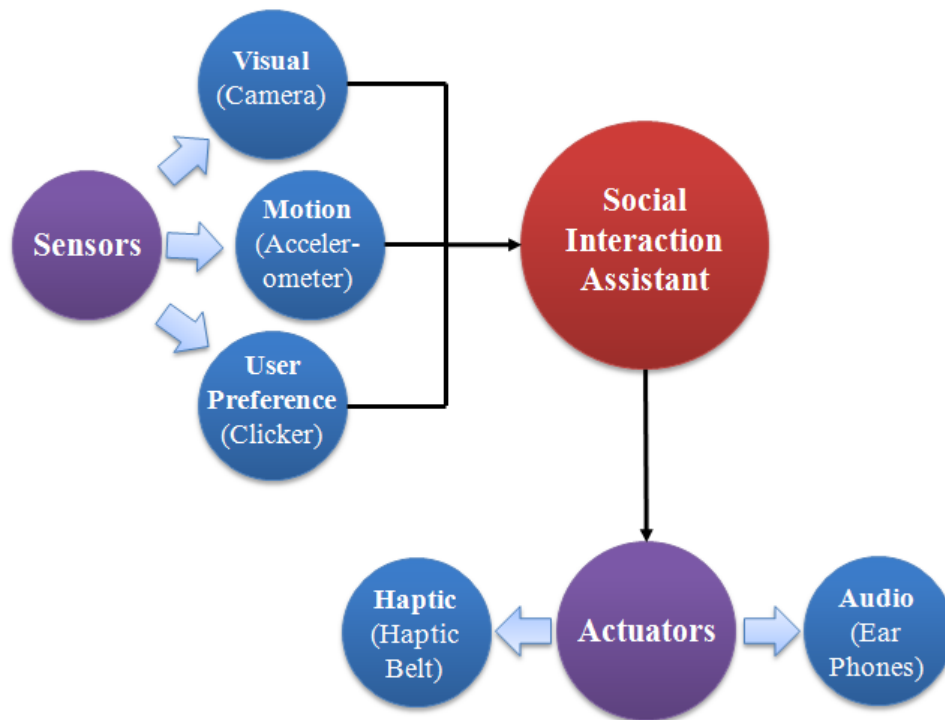


Figure 10: System level architecture of the Social Interaction Assistant

### 2.6.2.2 Prototype System:

Figure 11 shows the implementation of the proposed Social Interaction Assistant. A tiny video camera is placed unobtrusively on a pair of glasses, and a tiny state-of-the-art accelerometer is placed unobtrusively in a hat, and is used to monitor the user's body mannerisms – particularly those related to head and body movement. (Most communicative gestures are encoded in



movements of head and the most widely occurring and problematic stereotypic body mannerisms are done with the head.) The accelerometer operates on a coin battery that allows for uninterrupted operation for over 4 hours. The user uses the 5-button clicker to control what types of information are delivered by the system. The haptic belt can be worn under the clothing, and the earphones are worn discretely under their hat. Thus, the proposed design of the assistive technology is (1) wearable, (2) portable, (3) unobtrusive, (4) self and other sensing, and (5) can be worn by the user for extended periods of time.



Figure 11: the implementation of the Social Interaction Assistant

### 2.6.3 The Haptic Belt:

While most other components of the Social Interaction Assistant are sensors that are already well explored in the areas of signal processing, pattern recognition and machine learning, the haptic belt as an actuator is novel contribution from the work that was done towards Social Interaction Assistant. The details of the belt are given below for the sake of completeness.

Figure 12 shows the specifics of the implementation. The belt is wireless with 16 vibrotactile actuators that encircle the waist of the user. The wireless connection between the belt and the computer provides the desired portability and limited cumbr upon which the rest of the system is developed. The wireless haptic belt consists of a hierarchical microcontroller design with a main controller (Haptic Belt Controller) for PC or PDA communication and overall system maintenance, and auxiliary controller (Tactor Controller) for monitoring each vibration motor. While the main controller provides the user interface to access the tactors on the belt, the

auxiliary controllers ensure fine control of amplitude (perceived level of vibration intensity) and timing of vibration for each motor. This multilayer architecture caters to the important functional requirements of scalability, reconfigurability and portability. Any number of tactor modules, up to a maximum of 128, can be added to the belt without changing the firmware on the main controller (although we limited our implementation to 16 tactors or less). The functionality of the belt is exposed through an application programming interface, and can be leveraged through a command line (terminal control) or a graphical user interface for belt configuration and activation.

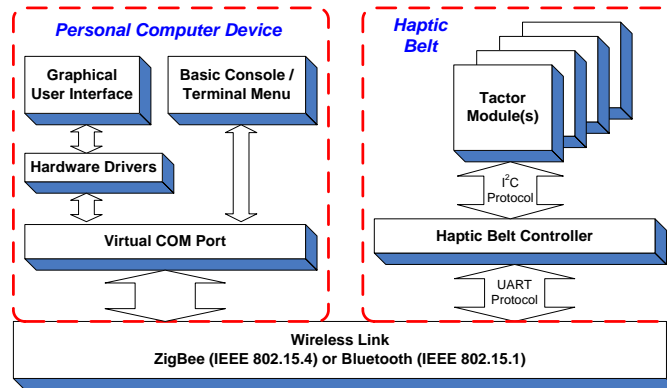


Figure 12. High-level system block diagram.

The entire system is powered by a slim 3.7V lithium-ion battery with higher per cell voltage (3.7V) and high power density (100-160 Wh/kg) when compared to Ni-Cd (1.2V at 40–60 Wh/kg) or Ni-Mh (1.2V at 30-80 Wh/Kg) batteries. The power is distributed using two of the haptic belt's four bus wires. The remaining two bus wires act as the data and clock lines of a standard I<sup>2</sup>C bus on which all 16 tactor modules listen to the main controller for specific commands on the amplitude and timing of vibration.

### 2.6.3.1 Hardware

#### *Belt Form Factor*

The belt harness and electronic system enclosures, shown in Figure 13, ultimately determine wearability. The belt harness, easily adjustable to any waist size, was constructed from 1.5 inch flat nylon webbing with quick connect acetyl plastic buckles. Likewise, the Serpac model C-2 electronic enclosure with a pocket clip was selected as an inexpensive commercial off-the-shelf (COTS) low-profile enclosure for the tactor modules. The pocket clips and bus connectors allow tactors to be easily repositioned, added or removed. This design was chosen over a Velcro based implementation for several reasons: to achieve better adaptability to different waist sizes; to hold tactors very close to the body during use; and robustness and rigidity for real-world use. Moreover, this design is lightweight, comfortable, silent and physically discreet as the control box can fit inside a pant pocket or attach to the belt and status LEDs can be turned off during use.



Figure 13. Haptic belt harness and tactor modules.

### Tactor Module

The tactor module houses a controller which drives the vibration motor. An ATtiny88 Atmel microcontroller forms the core of the tactor module with a small design footprint and onboard oscillator. The pulse-width modulation (PWM) unit on the controller is used to change the amplitude of vibration (by varying the duty cycle) and also generate different vibrotactile patterns and rhythms. Six pins of the ATtiny88 were configured to read a DIP switch setting that allows automatic configuration of its data communication bus address upon cycling the power. This eliminates the need to reprogram all tactor modules for different applications/uses, thus providing plug-and-play functionality.

The circuit diagram of an individual tactor module is shown in Figure 14. A coin-type shaftless vibration motor, Precision Microdrives 312-101, forms the vibrator with a rotational speed of 150Hz and a nominal vibration of 0.9g. The motor is switched with a low-side NUD3105 MOSFET inductive load driver, which has internal back emf protection built into its circuitry. The use of a MOSFET allows for lower gate current (less than 1mA) and even less leakage current when compared to a BJT transistor. LEDs visible on the outside of the module are provided for debug purposes.

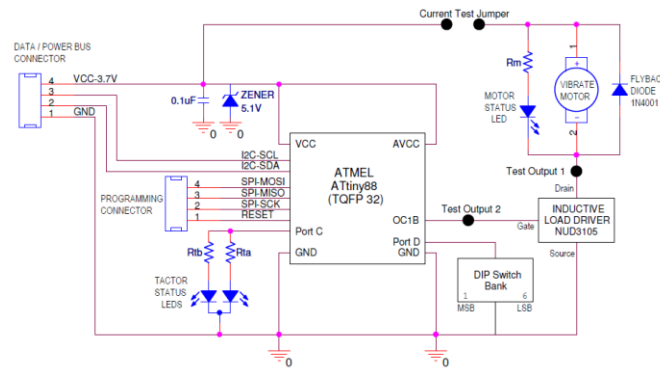


Figure 14. Tactor module schematic.

### Main Controller

Since all control and data communication of the haptic belt flows through the main controller, it must handle several mechanisms of communication including UART, Wireless, and I<sup>2</sup>C data bus communication. The main controller must also have enough on-chip memory to store belt configurations and console debug menus. We chose a specific implementation of the popular Arduino Open Source platform, called Funnel I/O. The board is based on Atmel ATmega168 microcontroller, a fully functional 8-bit controller with 16KB of Flash memory. The Funnel I/O supports all of the capabilities of the ATmega168 with a 1:1 pad to pin ratio for input/output. The board already has a power switch, reset button, status LEDs, lithium ion charging circuitry through a miniUSB connector, battery connection, and headers prewired for a plug-and-play XBee wireless module. The PCB is fairly small in size, which meets our form factor requirements.



### *Wireless Module*

We chose a self-encapsulated COTS wireless module with small form factor and an integrated chip antenna. Digi's XBee ZNet module was selected given that the Funnel I/O controller board can integrate with it without any additional design and the supported mesh network is forward looking. The XBee is a plug-and-play ZigBee wireless protocol module that fully supports the IEEE 802.15.4 sensor mesh network standard, and offers data transfer rates of 250 kbps with a range of up to 133 feet indoors. Similarly, a self-encapsulated Bluetooth module RN-41 from Roving Networks, using the IEEE 802.15.1 protocol and with similar range to the Xbee, was selected for an alternate wireless interface because of its ubiquitousness and the module was easily modified to fit within the Funnel's Xbee port.

### **2.6.3.2 Software**

#### *Firmware*

The architecture of the haptic belt's real-time embedded firmware fulfills several purposes. It controls vibration amplitude, timing and location, from which vibrotactile spatio-temporal patterns can be created. Up to five rhythm patterns, four amplitudes and the last in-use mode configuration can be stored for later use. Additionally, the firmware controls all belt logic including inter-module communication including the PC-wireless link, on-chip memory, tactor modules on the data bus, and provides a basic console/terminal menu that allows direct interaction with the belt configuration through a serial communication link (wireless or RS-232). With only limited memory space (16KB for the ATmega168 or 8KB for the ATtiny88), the firmware architecture had to be carefully engineered to provide the necessary functionality and ease of use while maintaining real-time performance. A simple command set structure similar to Hayes AT commands are used to minimize transmissions on the interconnect bus, and allows the 16 tactor modules to be sequentially switched on or off with a granularity of a few microseconds. The firmware was designed using the C language, and the open-source Arduino and Atmel's AVR libraries. The firmware provides four primary user modes to create a new belt configuration, query the current configurations, test vibrotactile patterns, and activate "in-use" mode. There are several levels of configuration available that allow users the flexibility of creating different vibrotactile spatio-temporal patterns. The current configuration settings along with all programmed vibrotactile patterns are stored in non-volatile memory to maintain a readiness state and ease of use.

#### *Graphical User Interface*

The graphical controls, written using the C# language and .NET components, allow easy configuration of complex vibrotactile rhythm patterns using text inputs and drop-down menu selections (Figure 15). Users can also specify tactor module locations, and query the wireless haptic belt for its current configuration. The software also provides utilities for creating spatio-temporal patterns using specified tactor modules and rhythms.

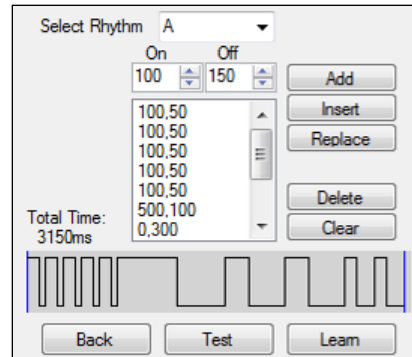


Figure 15. Graphical interface controls for vibrotactile rhythms.

While the details of the hardware design of the haptic belt are shown here, the usability aspect of it is not discussed here. Details of the application of the haptic belt as a part of Social Interaction Assistant will be provided in Chapter 4, where the problem of delivering Proxemics information as a non-verbal cue is discussed from the sensing and the delivery perspective.

## 2.7 Summary:

In this chapter a detailed analysis of the important social cues for people who are blind and visually impaired were analyzed through an online survey and the two most important needs of feedback on one's own body mannerism and the importance of facial expressions in social interpersonal communication were established.

Following the discussion of the important social needs, we have introduced a novel assistive technology framework that is capable of sensing and delivering some of the important non-verbal visual cues to people who are blind or visually impaired. From the perspective of an assistive technology, we only introduced the system side of the development. More on the usability will be discussed in the successive chapters.

## Chapter 3

# Egocentric Sensing of body movements

As explained in Chapter 2, people who are blind and visually impaired consider their body mannerisms to be an important component of their social interactions with their sighted counterpart. They are specifically worried about the display of any stereotypic body mannerisms that could potentially become a social distraction and hindrance for them. In Chapter 1, a detailed discussion of such stereotypic body mannerisms is presented and one specific quality of stereotypic body rocking was discussed in detail as body rocking is the most widely studied stereotypy in visually impaired population. In this chapter we introduce the concept of using egocentric motion sensing towards detection such stereotypic behaviors. While the discussion is limited to body rocking, similar concepts can be used to extend the rehabilitation to other body mannerisms.

Recently, human activity detection and recognition using motion sensors have taken a front seat in technology and behavioral research. This is due to the availability of micro mechanized electronic systems (MEMS) that have started to implement complex mechanical systems at a micro scale on integrated circuit chips. These offers advantages like reliability, cheaper cost of production, smaller form factor and above all extremely precise measurement with least or no maintenance. One such sensor is the accelerometer that is capable of measuring the effect of gravity on three perpendicular axes. When mounted on any moving object, the opposing motion (opposing gravity) of the entity allows these sensors to measure the speed and direction of motion. Integrating the magnitude and orientation information over time it is possible to accurately measure the exact motion pattern of the moving entity. These accelerometers have been used by researchers to track motion activity in almost every joint of the human body. Researchers have used single, double or triple orthogonal axis accelerometers to detect various activities of humans. They all follow the same underlying supervised learning architecture with difference in learning algorithm used. A simplified representation of the same is shown in Figure 1.

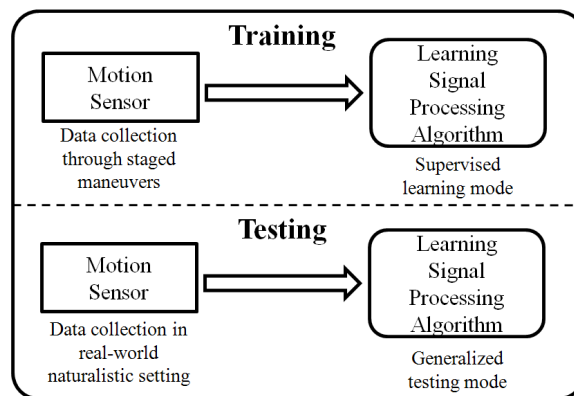


Figure. 1: Training and testing phases of a typical learning framework found in literature. In a similar framework, we use these motion sensors that are placed on strategically located places on the human body to detect body rocking behaviors.

### 3.1 The Hardware:

In our application, in order to keep the motion detector discrete, we have chosen state-of-the-art tri-axial accelerometer package, ZStar III, marketed by Freescale Semiconductor. The accelerometer is shown in the inset of Figure 2. The device (including a coin battery as a power source) is an inch in diameter and less than eighth of an inch in thickness thereby allowing an elegant integration into everyday clothing. Figure 2 shows the typical use of the accelerometer in the proposed application for detecting body rocking. The accelerometer has a very high sensitivity with protection against excessive g-force damage. The sensors wirelessly connect to a PDA and/or cell phone through IEEE 802.15.4 (ZigBee) wireless standards. The use of low power consumption electronics for both acceleration sensing and wireless communication allows this device to work for hours at length on a single coin battery. Further, the advanced sleep mode implementations allow the device to stay at nano watt power mode during non-operation. The proposed solution allows for prolonged use of the device to the effect of an assistive technology thereby maintaining a longer duration feedback based rehabilitation regimen.

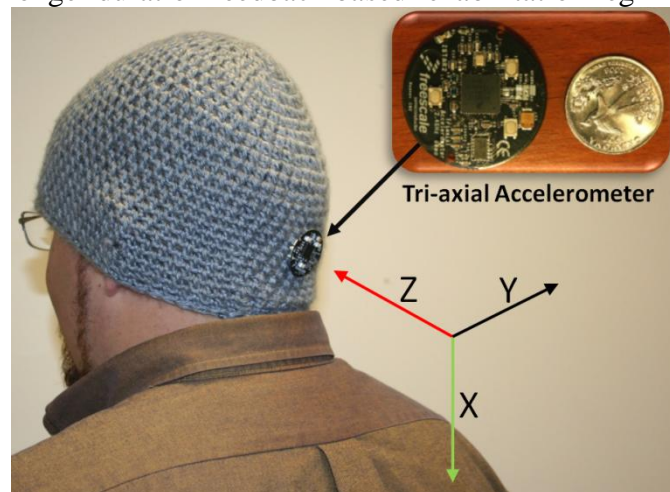


Figure 2. The proposed hardware for use in the detection of body rocking stereotypic behavior.

The accelerometer, in comparison with a US quarter, is shown in the inset. The three axes marked in the image shows the orientation of the accelerometer as it is placed on the head. The processing element for the current study was a Windows Mobile Operating System based PDA running on a 400Mhz XScale processor. The software components (described in detail in Section 3.2 of the proposed solution) were placed on the PDA that could be carried by a user without any extra load. The software component implementation is generic to be ported to most modern cell phones that possess enough processing power, but is always underutilized for its capacity. The feedback (an audio tone) is currently being provided through a Bluetooth headset that is paired with the processing element. The choice of this feedback device was again based on the idea that Bluetooth headset has everyday acceptance among the masses and is no longer seen as a social distraction. In future, we plan to explore the use of delivery modalities that transcends the typical visual and audio medium. We intend to use haptic cues to inform the participant not only their rocking behavior but more complex self-monitoring routines that could allow the user to withdraw from the rocking behavior effectively.

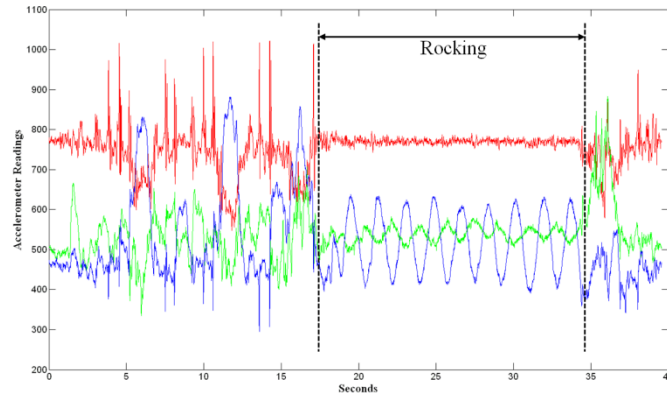


Figure 3. Data stream for the tri-axial accelerometer. The three streams correspond to the three axes. The figure shows non-rocking events followed by rocking and then followed by non-rocking.

Figure 3 shows a typical data stream collected from the accelerometer shown in Figure 2 during rocking and non-rocking functional behavior. The three data streams correspond to the three axis of the accelerometer each sampled at 100 Hz. It can be seen that the data stream under rocking conditions are visually distinguishable when compared to non-rocking functional movements. The following section highlights our choice of learning framework and features we extracted from these data stream in order to achieve reliable rocking and non-rocking discrimination.

## 3.2 Extracting Body Rock Information from Motion Sensor Data

The work presented in this paper builds on top of the work presented in [1] where the authors use two accelerometers placed one at the ankle and the other on the thigh to distinguish between simple activities like walking, running, standing etc. They proved the use of an aggregated AdaBoost classifier system that was built out of simple linear classifiers to achieve activity recognition. Unfortunately, the work does not provide any assessment on the generalization capabilities of their aggregate classifier. We extend their work into the problem of body rock detection using only one accelerometer placed on the back of the person's head. Below, we discuss the various features that we extract from the accelerometer data and introduce the variant of AdaBoost that generalizes on its training set very well (termed Modest AdaBoost). We show results of our experiments and discuss our reasoning to believe how the new AdaBoost framework is able to generalize on body rocking data when compared to classical AdaBoost used by [1].

### 3.2.1 Features:

Since we are using a tri-axial accelerometer, we obtain three orthogonal axis data through rocking and non-rocking events. In order to capture the temporal variation in the acceleration data, we accumulate the input stream on each axis for a fixed duration  $T$  seconds and all features are extracted on this packet of acceleration data. As a part of the assessment, we determine the best packet length for the task of body rock detection. Further, successive packets are extracted with a fixed duration of overlap between them.

We chose five sets of features that were extracted on the three axes of accelerometer data. For the sake of clarity, we cluster these sets into two groups based on whether they were chosen due to popular use in the accelerometer data processing community or due to the author's insights into the body rocking data.

### 3.2.1.1 Group 1 – Popular features used by the motion analysis research community [2] [1]:

We choose the following three sets each of which were applied on all three axes of acceleration data, henceforth referred to as x, y, z axis data.

1. Mean of x, y, z data over the duration of packet.
2. Variance of x, y, z data over the duration of packet.
3. Correlation between the three axes (x-y, y-z and z-x) over the duration of packet.

### 3.2.1.2 Group 2 – Authors insights into body rocking data:

Inspecting the accelerometer data shown in Figure 3, it can be seen that the Z axis changes from random signal pattern to more of a sinusoidal pattern when the individual's behavior transitions from non-rocking to rocking. Thus we choose two sets of features which we hope would capture this non-sinusoid to sinusoid transition between events. These features include

4. The first order differential power on all three axes – Sinusoidal signals change gradually over time such that the averaged sum square energy in the temporal first order differential of the signal should be less when compared to a random signal where the first order differential can have very high variations and hence higher power.
5. Fourier Transform variance and kurtosis on the Z-axis only – An effective way to capture power distribution of a signal into sinusoids is by using Fourier Transform. We hypothesize that the non-sinusoid to sinusoid transitions can be captured by quantitatively measuring the power spread spectrum of the Z-axis accelerometer data. We model the power spread to be a Gaussian and extract the variance and kurtosis (peaking) of the spread to determine if there is rocking or not.

Thus, the features used in our study can be categorized as belonging to two groups with three sets in Group 1 and two sets in Group 2. Each set has varying number of features based on what parameter the set is extracting from the temporal accelerometer data. Based on the descriptions above, the entire feature set has a total of 14 features. We identify each of these by their respective Feature Identification Numbers. Table 1 shows the two groups and the different sets under the group with typical values of these features under rocking and non-rocking behavior.

Group 1			
<b>Set 1</b> <b>Definition:</b> Mean on the temporal dimension. <b>Axes affected:</b> x, y, z <b>Number of contributing features:</b> 3 <b>Feature Identification Numbers:</b> 1, 2, 3	$M_x = \frac{1}{N} \sum_{i=1}^N x_i$	1. $M_x$	
		2. $M_y$	
		3. $M_z$	
<b>Set 2</b> <b>Definition:</b> Variance on the temporal dimension <b>Axes affected:</b> x, y, z <b>Number of contributing features:</b> 3 <b>Feature Identification Numbers:</b> 4, 5, 6	$V_x = \frac{1}{N-1} \sum_{i=1}^N (x_i - M_x)^2$	4. $V_x$	
		5. $V_y$	
		6. $V_z$	
<b>Set 3</b> <b>Definition:</b> Cross correlation between axes <b>Axes affected:</b> x, y, z <b>Number of contributing features:</b> 3 <b>Feature Identification Numbers:</b> 7, 8, 9	$C_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - M_x)(y_i - M_y)$	7. $C_{xy}$	
		8. $C_{yz}$	
		9. $C_{xz}$	
		14. $F_{kz}$	

Table 1: Group 1 Features. Most popularly used features for extracting motion information. The figures shown in the last column plots mean values of data from positive rocking samples and negative rocking samples as bars. The variance on the same is shown as vertical error lines around the mean. The lighter (blue when viewed in color) shaded bar are values from the positive class, whereas the darker (pink when viewed in color) bar are values from the negative class.



Group 2			
<b>Set 4</b>  <b>Definition:</b> First order differential power. <b>Axes affected:</b> x, y, z <b>Number of contributing features:</b> 3 <b>Feature Identification Numbers:</b> 10, 11, 12	$D_x = \sqrt{\sum_{i=2}^N (x_i - x_{i-1})^2}$	10. $D_x$	
		11. $D_y$	
		12. $D_z$	
<b>Set 5</b>  <b>Definition:</b> Gaussian fit power spread spectrum – Variance and Kurtosis <b>Axes affected:</b> z <b>Number of contributing features:</b> 2 <b>Feature Identification Numbers:</b> 13, 14	<p>If, <math>Freq_k = \left\{-\frac{\gamma}{2}, \dots, 0, \dots, \frac{\gamma}{2}\right\}</math>, <math>\gamma</math> is the sampling frequency, and</p> $X_k = \sum_{i=1}^N x_n e^{\frac{2\pi i k n}{N}}, k = \{1, \dots, N\},$ <p>then</p> <p><b>FFT Variance:</b></p> $F_v = \sum_{i=1}^N X_k (Freq_k)^2$ <p><b>FFT Kurtosis:</b></p> $F_k = \sum_{i=1}^N X_k (Freq_k)^4$	13. $F_{v_z}$	
		14. $F_{k_z}$	

Table 2: Group 2 features that were specifically extracted based on the nature of body rocking data. The figures shown in the last column plots mean values of data from positive rocking samples and negative rocking samples as bars. The variance on the same is shown as vertical error lines around the mean. The lighter (blue when viewed in color) shaded bar are values from the positive class, whereas the darker (pink when viewed in color) bar are values from the negative class.

### 3.2.2 Learning Algorithm:

As discussed in introduction of this section, we compare the performance of two AdaBoost learning frameworks to determine which one can generalize the best on the training data. The two algorithms are introduced briefly below. For further details, the reader is referred to appropriate references provided within the subsections.

#### 3.2.2.1 Classic AdaBoost Learning Framework:

AdaBoost learns any classification problem by working with a set of weak classifiers. Weak classifiers are those classifiers that use simple decision steps to categorize data into one of two pools – positives or negatives (In all our experiments, we used a three level decision tree [3] as the simple classifier). AdaBoost proceeds by ranking the labeled training data as being simple to complex based on how many weak classifiers are needed to learn each of the examples. The



process continues on an iterative manner until all the training examples are learnt or till the allowed number of learning cycles are exhausted. Let,  $X$  be the input to a learning algorithm, in our case the features extracted as explained in the previous step, and  $Y$  be the label of what class the data belongs to, in our case,  $Y = \{1, -1\}$  implying {rocking, non-rocking}, respectively. Values at each dimension of input  $X$  can be considered to characterize the incoming data in some manner and the task of the learning algorithm is to learn these representational values of the input dimensions that allow the algorithm to distinguish between rocking and non-rocking. AdaBoost does this learning by using a large set of simple (weak) learners (or classifiers) that act on each of the dimension of the input data with the determined goal of distinguishing rocking from non-rocking. The final decision of the complete learning module is a combined opinion of all the simple learners that make up the system. The beauty of AdaBoost implementation is that the human intervention into the learning process stops at identifying what simple (weak) learners to use and what feature pool to operate on. Selection of number of weak learners, selection of input dimension on which the weak learners have to act, and the confidence to place on the decision of each of the weak learner is all determined by the algorithm during the training phase. Once the algorithm is trained, the final learnt rocking/non-rocking classifier can be represented as

$$L(x) = \text{sign} \left[ \sum_{i=1}^N w_i f_i(x) \right]$$

where,

$x$ : An instance of all possible rocking patterns  $X$ .

$L$ : The final learnt classifier that can distinguish input  $x$  as rocking or non-rocking.

$f$ : The simple (weak) learner.

$N$ : The total number of weak learners that make up the complete learner  $L$ .

$w$ : Weight associated with each weak learners output. This corresponds to the confidence placed in each weak learner by the Boosted system.

From a learning perspective, in each step of the iterative learning, the AdaBoost algorithm implements a greedy optimization to pick a set of weak learners that minimize exponential classification error of the picked simple classifiers as shown below

$$Error_k = \sum_{i=1}^M e^{-y_i L(x_i)}$$

where,

$y$ : Label of the input instance  $x$

$M$ : Total number of examples in the training set

$k$ : Learning iteration number

Further, based on each iterative step, a distribution ( $D_m$ ) is created over the training set examples to represent their complexity (difficulty to learn). For example, in a given iteration, an example that could be solved is assigned a lower distribution weight while, a sample that was not learnt in that iteration step is assigned a higher weight. The lower weight on the learnt example implies that this example will be stressed less in the next learning iteration while all other examples which could not be solved will become the focus for picking new weak learners. Moving from one iteration to the next, all the weak learners from the past  $k$  iterations are added into a pool of selected weak learners leading up to the final classifier  $L$ .

### 3.2.2.1 Modest AdaBoost

All learning algorithms, including AdaBoost suffer from the problem of over fitting or over learning. This is due to the fact that training sample sets of positives and negatives can never be representative of all the possible samples that the algorithm will face in its operational life span. Since the learning is limited to a restricted set of examples, there is always the problem of over fitting into this small set and thereby loosing the ability to generalize their learnt knowledge to all other possible examples. To this end, many alternatives have been proposed to AdaBoost that will allow the algorithm to generalize better. We introduce Modest AdaBoost [4] which was recently proposed towards better generalization capabilities and has been shown to be powerful on various machine learning datasets. Unlike the classic AdaBoost where the distribution penalizes only examples that are not learnt in the previous iteration, Modest AdaBoost penalizes for examples that are not learnt and also examples that are learnt very well (over fitting). This is done by projecting all the examples in the training pool on to four separate distributions,

1.  $P_m^{+1} = P_{D_m}(y = +1 \cap L(x)) \rightarrow$  Probability of the learner, as measured on  $D_m$ , predicting an input instance  $x$  correctly as being rocking when the label also represents it to be rocking.
2.  $P_m^{-1} = P_{D_m}(y = -1 \cap L(x)) \rightarrow$  Probability of the learner, as measured on  $D_m$ , predicting an input instance  $x$  correctly as being non-rocking when the label also represents it to be non-rocking.
3.  $\bar{P}_m^{+1} = P_{\bar{D}_m}(y = +1 \cap L(x)) \rightarrow$  Probability of the learner, as measured in the inverse distribution ( $\bar{D}_m$ ), predicting an input instance  $x$  correctly as being rocking when the label also represents it to be rocking.
4.  $\bar{P}_m^{-1} = P_{\bar{D}_m}(y = -1 \cap L(x)) \rightarrow$  Probability of the learner, measured in the inverse distribution ( $\bar{D}_m$ ), predicting an input instance  $x$  correctly as being rocking when the label also represents it to be rocking.

Conditions 1 and 2 penalize the classifier on examples that are not learnt during a training iteration, whereas 3 and 4 penalize examples that are already learnt in the previous iteration which was learnt again in the current iteration. Combining these four measures as

$$f_m = (P_m^{+1}(1 - \bar{P}_m^{+1}) - P_m^{-1}(1 - \bar{P}_m^{-1}))(x)$$

provides a means for penalizing the learner for not classifying an example and also for over fitting an example. This provides a means for modest learning of the final combined classifier  $L$ .

We hypothesize that the choice of a learning algorithm that generalizes well will provide the opportunity to allow better non-rocking detection thereby hopefully increasing discrimination ability for the assistive device. This would directly reflect upon the motivation of the user to get feedback only when he/she is rocking and not performing other functional activities.

## 3.3 Data Collection

Two separate data collections were carried out, one in a controlled setting while the other in a more uncontrolled naturalistic everyday research laboratory setting. The controlled setting data collection was used for training and lab testing the device, whereas the uncontrolled naturalistic setting was used to determine how well the learning algorithm was able to generalize when used for an extended period of time as an assistive tool.

### 3.3.1 Controlled Data Collection:

Data was collected on ten participants who did not have any known stereotype rocking behavior. The goal of the experiments was to collect data for training the system to differentiate rocking

from non-rocking behavior. To this end, we devised three separate data collection routines where the subjects were required to do rocking and non-rocking tasks as naturally as possible. The details of the routines are as follows:

### **3.3.1.1 Routine A:**

Rocking data: Participants were allowed to choose from a rocking chair or a stool or sitting on the ground, so they could rock as comfortably and naturally as possible. We found some cultural preferences to the way people choose to rock. The subjects were asked to rock for a total of 20 complete cycles.

### **3.3.1.2 Routine B:**

Non-rocking data: The participants were asked to do activities that did not involve rocking. They moved around the experimental setup reading posters, operating computers, interacting with everyday office equipments and included some functional body motions similar to rocking like, stooping down to pick up objects, rapidly bending down to pick up objects etc. Data was collected for a total of 30 seconds.

### **3.3.1.3 Routine C:**

Test data: Since rocking can happen at any given instance, we collected data where subjects did various activities and interspersed them randomly with rocking. The goal is to determine how fast and accurately our system can detect such rocking occurrences. In all of these data streams, rocking instances were manually identified and marked for the sake of ground truth. Figure 3 shows the combination of rocking and non-rocking activities by the participants. It can be noticed that there is a clear demarcation between the two activity zones.

## **3.3.2 Uncontrolled Data Collection:**

The uncontrolled data was collected towards testing the generalization capabilities of the learnt system. To this end, the body rock detection system was worn by the primary author during everyday laboratory activities. Body rock detection was provided as a feedback through a pair of headphones in the form of an audio beep. Five trails of four separate ten minute data collections were done. Two of the four were done with classic AdaBoost whereas the other two were done with Modest AdaBoost. Further, under each of these two classifiers, one data collection measured how many false positives were detected, whereas the second data collection counted how many rocking actions went undetected. During all these data collection the researcher counted the number of false positive or false negatives using a handheld thumb counter. This experiment was conducted purely to test the generalization capability of the learnt classifier.

## **3.4 Experiments**

Experiments were carried out for comparing the performance of the classic AdaBoost framework with Modest AdaBoost for the specific tasks of determining

- a. The length of a temporal packet of data needed to effectively distinguish rocking from non-rocking.
- b. The accuracy with which the two classifiers can distinguish between rocking from non-rocking.
- c. The generalization capabilities of the two classifier systems.

To this end the rocking samples collected in Routine A (discussed under Section 3.3.1.1) and Routine B (discussed under Section 3.3.1.2) were used as labeled positive (rocking) and negative (non-rocking) data for training the AdaBoost classifiers. Data collected under Routine C (discussed in Section 3.3.1.3) were used for testing the learnt classifiers. The results from this analysis were used for determining a. and b. above. We varied the packet length on the data stream and determined the recognition rate on the test data. While the packet length was varied, a constant overlap was maintained between successive packets. This overlap was determined empirically to be 0.5 seconds or 50 samples (@ 100 Hz sampling rate). With the ground truth already provided for the test set, we were able to determine the accuracy of the two classifiers.

To determine c., we resorted to using the data collected in Section 3.3.2. The primary author of the paper used the device to collect false positive and false negative data in order to determine how well the classifiers generalized on the training data. Further, we analyzed the working of the two classifiers in a piece wise manner by breaking down the features into individual sets (Sets 1 through 5 as identified in Table 1 and Set 6 that included all 14 features) and understanding the functional ability of the classifiers under individual feature sets. This allowed for an in-depth analysis of the workings of the two classifiers. In Section 3.6, we discuss the generalization capability of the two classifiers by heuristic analysis of the piecewise operational modes.

All our experiments were carried out with the aid of the AdaBoost Matlab library developed by Graphics and Media Lab at the Dept. of Computer Science at Moscow State University [5].

### 3.5 Results

Figures 4 and Figure 5 shows the box plot [6] of packet length ( $T$  secs) versus recognition rate for classic AdaBoost and Modest AdaBoost frameworks, respectively. The abscissa represents the length of the data stream (in seconds) used for the analysis, while the ordinate represents the recognition rate. Training and testing were all carried out on the data collected as depicted in Section 3.3.1. The horizontal line inside the box represents the median (second quartile) of recognition rates over the ten subject's data. The lower end of box presents the first quartile (25 percentile) and the upper end of the box represents the third quartile (75 percentile). Thus the box surrounds the center 50 percentile ranges of recognition results. This box is also called the Inter-Quartile Range ( $IQR = \text{third quartile} - \text{first quartile}$ ). The dotted extremity represents the minimum and maximum recognition rate under a certain packet length among the ten subjects. Any outlier (an outlier is greater than 1.5  $IQR$  from the median in any direction) is marked by an asterisk.

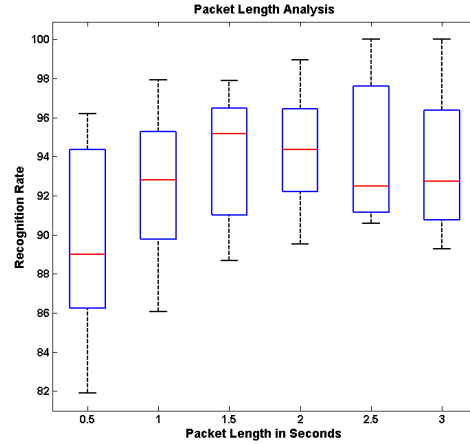


Figure. 4. Packet length to recognition rate comparison under the classic AdaBoost framework.

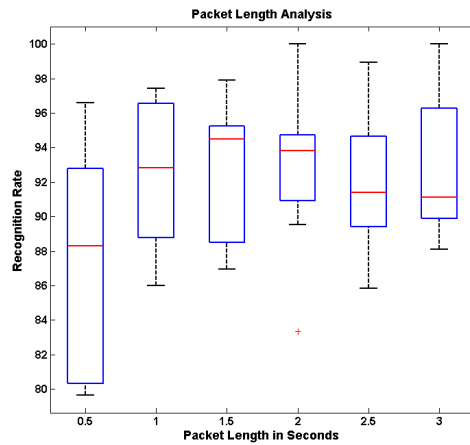


Figure. 5. Packet length to recognition rate comparison under the Modest AdaBoost framework.

Table 2 presents the results from the experiment carried out to determine the generalization capabilities of the two classifiers. The entries in the table are counts as measured by the researchers of the number of false positives and false negatives counted manually while using the device for body rock detection and feedback. Five trials were carried out of 10 minutes each for determining these numbers. False positives represent the number of times the device falsely gave feedback when the user was not involved in rocking. It is important that this rate be minimal as too many false feedbacks would be discouraging for the user to continue using the assistive aid. The false negative represents the number of times the device did not detect that the user was rocking. This metric could be correlated to the failure of the device to perform its functional task.

TABLE 2  
EXPERIMENTS ON NATURALISTIC DATA

Generalization Capabilities	Classic AdaBoost	Modest AdaBoost
<b>False Positives per Minute</b> – Number of false feedback in ten minute <sup>†</sup>	86	44
<b>False Negatives per Minute</b> – Number of times rocking was not detected in ten minute <sup>†</sup>	20	9

<sup>†</sup> These numbers were averaged over 5 trails of 10 minute each
--

Figure 6 and Figure 7 shows the piecewise analysis of the classic AdaBoost and Modest AdaBoost frameworks. Subfigure (a) shows the performance of each feature set considered one at a time in detecting body rocking; feature set 6 corresponds to the use of all 14 features together. For example, column 1 in Figure 6 (a) represents the recognition performance using only temporal mean along x, y and z axis tested on all ten subjects. The bar graph in (a) shows the mean performance rate while the superimposed box plot shows the performance at first, second and third quartile as discussed earlier.

Subfigure (b) represents the Receiver Operating Characteristics (ROC) [7] for the same six feature sets as in subfigure (a). ROC is plotted a false positive rate (FPR) versus true positive rate (TPR). The better the performance, the curve moves towards the (1,1) co-ordinate. For example, in Figure 6 (b) Set 6 with all features is performing better than feature set 1 as Set 6 curve is closer towards (1,1) while the feature set 1 curve is almost along the diagonal of the plot. The diagonal of the ROC plot represents a recognition rate of 50% i.e. random pick.

Subfigure (c) is a derivate of the ROC curves in subfigure (b). Each bar in the graph is representing the area under the corresponding curve (AUC) in (b). An area of 1 represents an ideal classifier with no false positive or false negatives, while an area of 0.5 represents randomness in the classifier output. AUC can be used to immediately determine the curve with the best performance.

Subfigure (d) is an understanding of how the aggregated AdaBoost classifier is built. As discussed above, AdaBoost classifier uses a collection of simple classifiers to achieve the final classifier. We plotted the number of times a particular feature is being used by the aggregate classifier. Further, the features are grouped into 5 sets corresponding to the five feature sets identified in Table 1. Columns belonging to the same set have the same top count which corresponds to the total simple classifiers used from that set. Each column within the set represents how many classifiers are used on each feature within that set. The count on the individual feature is represented by the bottom color along each column. For example, consider set 4 in Figure 7 (d), features with identification number 10, 11 and 12 form this set (corresponding to the first order differential power from x, y and z axis of the accelerometer data) and have a top count of 646 simple classifiers. Within the group, the z axis differential power dominates the other two by having a count of 374.

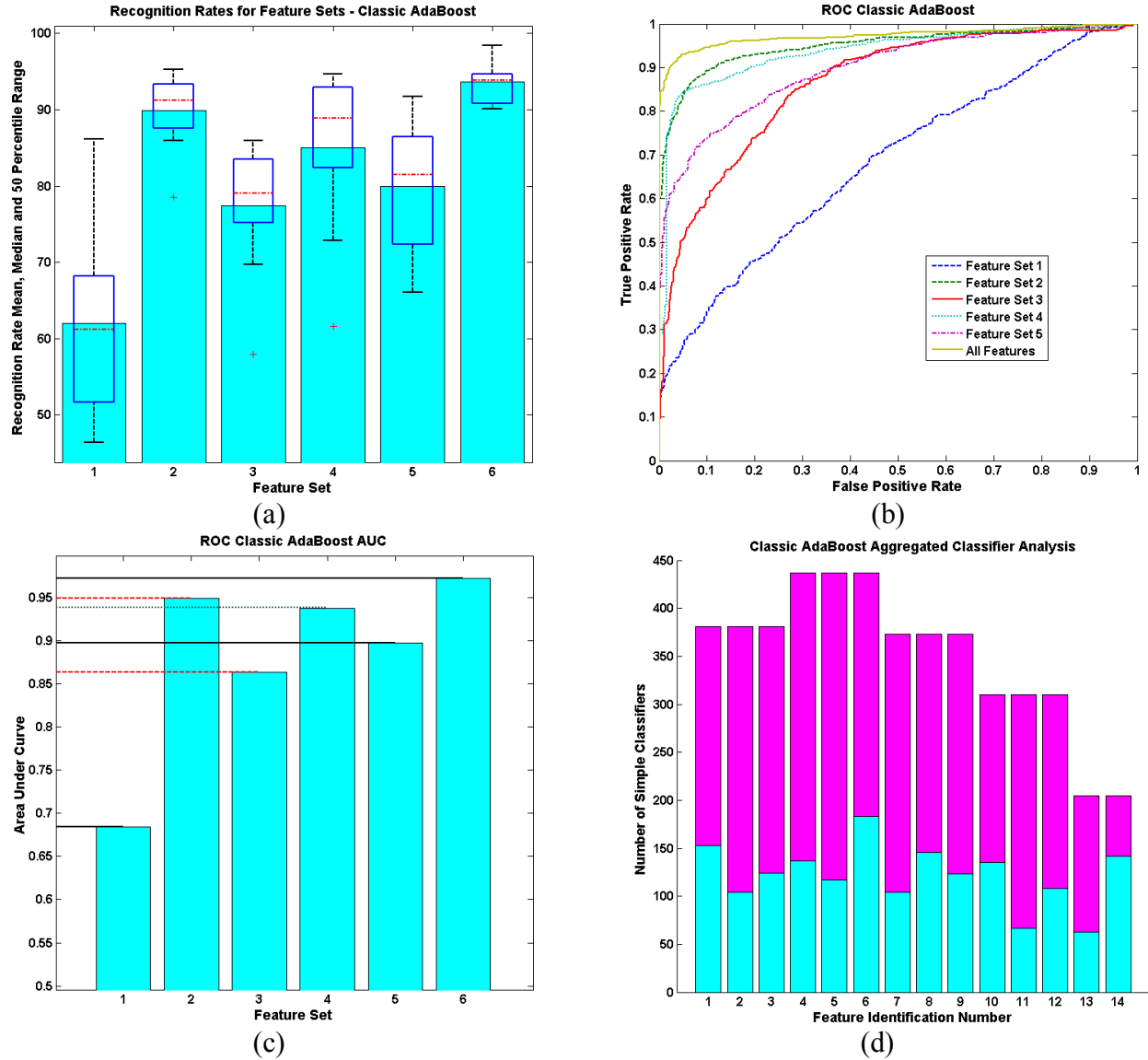
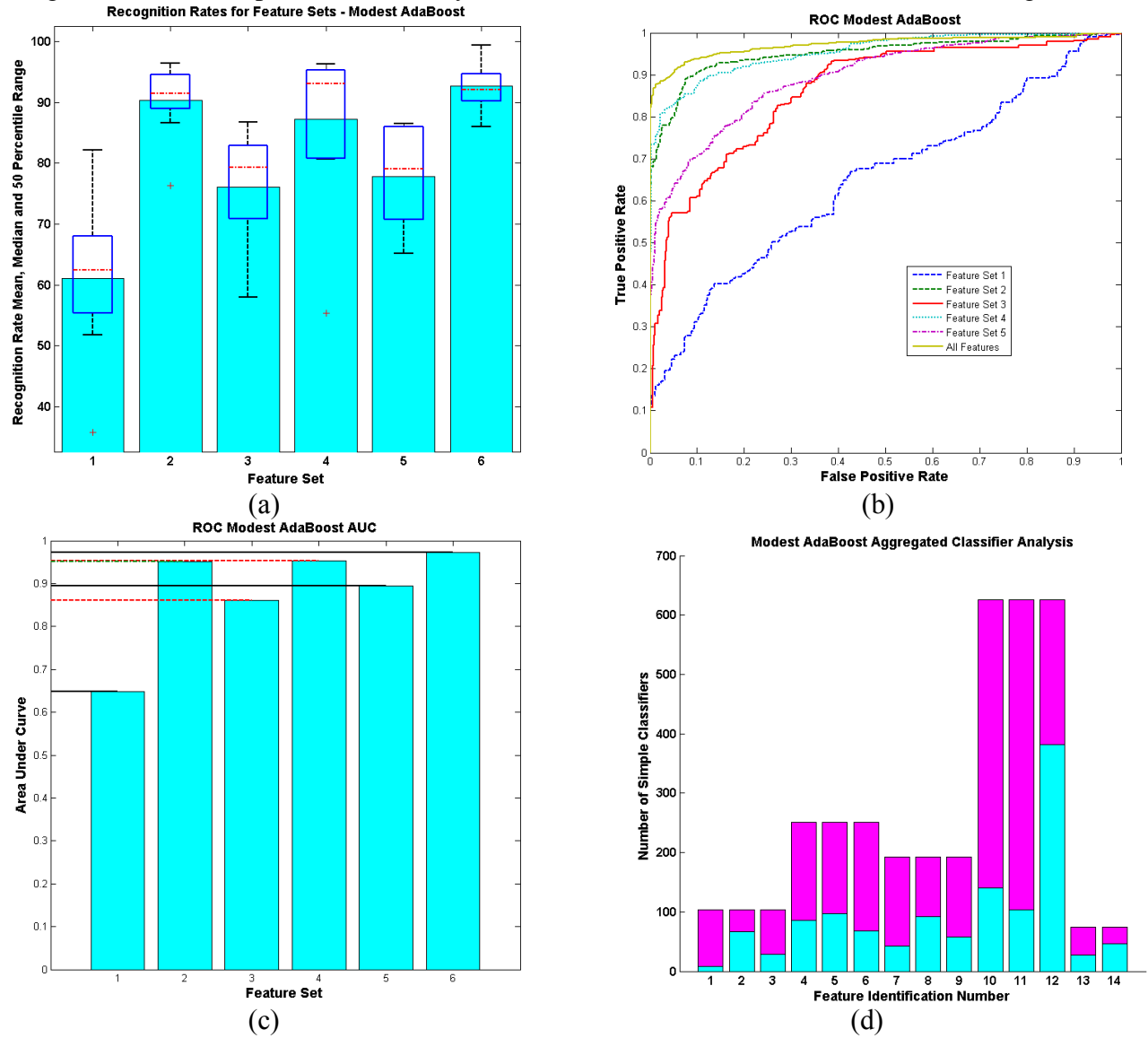


Figure 6: Piecewise performance analysis of the classic AdaBoost classifier framework. (a) Recognition rates under use of individual feature sets. (b) The Receiver Operating Characteristics (ROC) under the use of individual feature sets. (c) Area under the curve (AUC) for each feature set as estimated from the ROC. (d) The number of simple classifiers used by the aggregated AdaBoost classifier. Each set and each feature representation in the classifier pool are separately marked. In all the graphs Set 1 through 5 are as explained by Table 1. Set 6 represents a set containing all 14 features from Table 1.

Figure 7. Piecewise performance analysis of the classic AdaBoost framework. (a) Recognition



rates under use of individual feature sets. (b) The Receiver Operating Characteristics (ROC) under the use of individual feature sets. (c) Area under the curve (AUC) for each feature set as estimated from the ROC. (d) The number of simple classifiers used by the aggregated AdaBoost classifier. Each set and each feature representation in the classifier pool are separately marked. In all the graphs Set 1 through 5 are as explained by Table 1. Set 6 represents a set containing all 14 features from Table 1.

### 3.6 Discussion of Results

Regarding the research question, what is the state-of-the-art technology available to detect and notify individuals of their rocking behavior? We have identified the state-of-the-art motion



sensor that is small enough in form factor to become part of one's everyday clothing. Further, we designed this device to be discrete so that the user does not feel any intrusion into their everyday activities. The software can be run on any mobile processing device that the user already carries like a cell phone or PDA. This allows the users to use the device without carrying any additional load.

Focusing on the research question of, Is it possible to build a device that detects body rocking condition and how well can it distinguish body rocking from other functional activities of daily living? We turn our attention to the various results presented in Section 3.5 to prove the efficiency of our proposed method in detecting body rocking and distinguishing it from other non-rocking behavior.

### 3.6.1 Packet Length, and Detection Efficiency

From Figure 4 and Figure 5, it is evident that the recognition rates for the two classifiers are comparable and the median recognition rate ranges from 89% to 95%. Based on these numbers, the best performance was achieved at a sample length of 1.5 seconds or 150 samples per packet. Packet length of 150 samples has the highest recognition rate on both the classifiers. Comparing this packet with the 2 seconds packet length or 200 samples per packet, we notice that the 2 seconds packet is very close behind and it has a smaller 1.5 IQR box. Thus, the variance in the recognition rates between 10 subjects is lesser in the 200 samples packet length, implying that the results are more consistent. Further, we noticed that the average natural rocking motion of the 10 subjects was around 27 rocks a minute (i.e. 27 rocks in 60 seconds or 2.22 seconds per rock; this is supported by results from [8]), which implies that a latency of 2 seconds was the closest to the time duration of a single rocking action. As mentioned earlier, all experiments were carried out with an overlap 0.5 seconds or 50 samples between successive packets. Combining these two results, we have

1. Optimum Packet Length: 2 seconds or 200 samples with 0.5 seconds or 50 samples overlap between packets.
2. Best Detection Rate: @ 2 seconds packet length  $\approx$  94% under both classifiers

### 3.6.2 Generalization Capabilities

From Figure 4 and Figure 5, it is very difficult to distinguish any performance benefits between classic AdaBoost and Modest AdaBoost. But analyzing Table 2, we can notice a dramatic difference in the performance of the Modest AdaBoost when compared to classic AdaBoost. The number of false positives is down from 86 to 44 over a ten minute period. That is, the user receives nearly half less number of false feedback with Modest AdaBoost framework when compared to the classic AdaBoost. This was not evident in the detection tests that were carried out with data collected from Routine C (Section 3.3.1.3). We asked the question of why there is an increased performance in Modest AdaBoost and why there is a discrepancy between the test results from Routine C and the naturalistic data capture (Section 3.3.2). The answer to these questions lies in the generalization capabilities of the two classifiers. We noticed that most of the false feedback provided by classic AdaBoost occurred while the user was sitting and not rocking. In hind sight, we realized a slight discrepancy in our non-rocking (negative class) data collection. While capturing data under Routine B (as explained in Section 3.3.1.2.) the participants were asked to perform various tasks that did not involve rocking to use as negative training set. We realized that most of the participants performed tasks that involved some form of

walking or standing activities while they did no activity that involved sitting and not rocking. Thus, just sitting activity was a non-rocking event that was not represented in the training data set. We hypothesize that classic AdaBoost over trained on the non-rocking data while Modest AdaBoost, which is penalized for learning the training set very well, had a better generalization. Extending this heuristic analysis to a more formal analysis, we look at the piecewise performance of the two classifiers. Comparing the ROC curves from Figure 6 (b) with Figure 7 (b), it can be seen that feature set 2 – Variance and feature set 4 – First Order Differential Power performed the best following Set 2 - All features set. Now comparing Figure 6 (d) with Figure 7 (d) it can be seen that Modest AdaBoost distributed its simple classifiers such that there were more classifiers representing the two feature sets 2 and 4. On the other hand, the classic AdaBoost's distribution of simple classifiers is unexplainable as feature set 1 – Mean – seems to have received more representation than set 4. Mean had the worst performance as an individual feature set as can be verified by the ROC curve that comes closest to the diagonal on the plot hinting that the performance is barely above random guess. Contrasting this with Modest AdaBoost selection, Mean is in the bottom two sets among the five feature sets. This bad performance of Mean as a feature set can be understood by looking at the graph shown in the first row and last column of Table 1. It can be seen that the Mean acceleration values between rocking and non-rocking are not significantly different. Table 1, Rows 2 and 4 highlights the capabilities of Variance and First Order Differential Power in distinguishing rocking from non-rocking. This is further confirmed by the ROC graph.

Feature 4 having the highest distribution of simple classifiers under Modest AdaBoost (Figure 7 (d)), within this feature set we can see that the highest number of simple classifier is assigned to feature 12 which corresponds to First Order Differential Power on z axis. As can be verified from Figure 3, the best distinguishing character between non-rocking and rocking patterns seems to be the transformation of a random signal pattern on z-axis to a deterministic sinusoidal waveform. If this can be the true identity of the rocking data stream, feature 12 would capture it in the best possible manner by measuring the power in the first order differential of the temporal signal. Using this feature as the most reliant feature would provide a good basis to support the final classifier selected by Modest AdaBoost.

We are now ready to answer the last research question stating that the use of approximately 2 seconds (or 200 samples @ 100 Hz sampling rate) packet length used with a learning framework biased towards generalized learning (like Modest AdaBoost) would be a good assistive technology solution for detecting and giving feedback towards stereotypic body rocking. We can extend the same argument to other body mannerisms that involve any form of repetitive body part movement.

### 3.7 Summary:

In this chapter, we have addressed the topic of detecting stereotypic body mannerisms, specifically body rocking, and propose a technology solution for providing an assistive technology that may reduce or control body rocking. We have discussed the hardware and software components of the proposed system in detail and offer a thorough analysis on the learning framework that provides generalization benefits to allow this framework to be extended to detection of any body mannerism. Investigations are in progress to determine how incoming samples of acceleration data can be labeled automatically by the system based on the AdaBoost classifier's classification confidence metrics. This would provide opportunity for self-learning [9] modes where the device can readily understand and learn data points that were not available

in the training set. Combining such self-learning into a generalized learner would provide immense opportunities for not only body mannerism detection, but for solving future data mining problems where typical lab setting training data collection would just not be sufficient to train a robust classifier.

From the assistive technology perspective, we plan to integrate a well planned self-monitoring as a part of the proposed device. We are exploring the broad area of human communication to determine the best cognitive self-correction techniques that could augment the proposed solution. We have implemented a rudimentary form of real-time body rock counter, as discussed in [10], but we have not yet tested it for its feedback capabilities.

### 3.8 References:

- [1] N. Krishnan and S. Panchanathan, "Analysis of low resolution accelerometer data for continuous human activity recognition," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 3337-3340.
- [2] L. Bao and S.S. Intille, "Activity recognition from user-annotated acceleration data," 2004, pp. 1-17.
- [3] Y. Yuan and M.J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets Syst.*, vol. 69, 1995, pp. 125-139.
- [4] A. Vezhnevets and V. Vezhnevets, "Modest AdaBoost - Teaching AdaBoost to Generalize Better," Novosibirsk Akademgorodok, Russia: 2005.
- [5] A. Vezhnevets and V. Vezhnevets, *GML AdaBoost Matlab Toolbox 0.3*, Graphics and Media Lab, Moscow State University, 2007.
- [6] Y. Benjamini, "Opening the Box of a Boxplot," *The American Statistician*, vol. 42, Nov. 1988, pp. 257-262.
- [7] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, 2006, pp. 861-874.
- [8] K.M. Newell, T. Incledon, and J.W. Bodfish, "Variability of stereotypic body-rocking in adults with mental retardation," *American Journal on Mental Retardation*, vol. 104, May. 1999, pp. 279 - 88.
- [9] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng, "Self-taught learning: transfer learning from unlabeled data," *Proceedings of the 24th international conference on Machine learning*, Corvallis, Oregon: ACM, 2007, pp. 759-766.
- [10] D.B. McAdam and C.M. O'Cleirigh, "Self-monitoring and verbal feedback to reduce stereotypic body rocking in a congenitally blind adult," *Re:View*, vol. 24, Winter93. 1993, p. 163.

# Chapter 4

---

## Exocentric Sensing & Delivery: Proxemics

---

### 4.1 Proxemics

In behavioral psychology, influences of interpersonal distances on social interactions between people have been studied for over four decades. The term *proxemics*, coined by Edward T. Hall, describes influence of interpersonal distances in animal and man [1]. The following list describes the American proxemic distances; note that such distances vary with culture and environment.

- Intimate Distance (Close Phase): 0-6 inches
- Intimate Distance (Far Phase): 6-18 inches
- Personal Distance (Close Phase): 1.5-2.5 feet
- Personal Distance (Far Phase): 2.5-4 feet
- Social Distance (Close Phase): 4-7 feet
- Social Distance (Far Phase): 7-12 feet
- Public Distance (Close Phase): 12-25 feet
- Public Distance (Far Phase): 25 feet or more

Proxemics plays a very important role in interpersonal communication, but people who are blind and visually impaired do not have access to this information. In [2], Ram and Sharf introduced The People Sensor: an electronic travel aid, for individuals who are blind, designed to help detect and localize people and objects in front of the user. The distance between the user and an obstacle is found using ultrasonic sensors and communicated through the rate of short vibratory pulses, where the rate is inversely proportional to distance. However, the researchers did not do any user testing to determine the usefulness of their technology. Similar to this system, our technology uses the haptic belt described in Chapter 2 for delivering the proxemics information to an individual who is blind or visually impaired.

Tactile rhythms delivered using a vibrotactile belt were used in [3] to convey distance information during waypoint navigation. Time between vibratory pulses was varied using one of two schemes: monotonic (rate is inversely proportional to distance) or three-phase-model (three distinct rhythms mapped to three distances). Distinct tactile rhythms are promising for use with multidimensional tactons [4] [5], which are vibratory signals used to communicate abstract messages [5] by changing the dimensions of the signal including frequency, amplitude, location, rhythm, etc. Based on pilot test results, we chose to pursue distinct rhythms over monotonic rhythms as users find it difficult to identify interpersonal distances using monotonic rhythms as the vibratory signal varies smoothly with changes in distance.

From the sensing perspective we resort to the camera that is on the user's glasses and through the use of computer vision technology, face detection, we extract non-verbal cues for social

interaction, including the number of people in the user's visual field, where people are located relative to the user, coarse information related to gaze direction (pose estimation algorithms could be used to extract finer estimates of pose), and the approximate distance of the person from the user based on the size of the face image.

## 4.2 Conceptual Framework

As shown in Figure 1, the output of the face detection process (indicated by a green rectangle on the image) provided by the Social Interaction Assistant is directly coupled with the haptic belt. Every frame in the video sequence captured by the Social Interaction Assistant is divided into 7 regions. After face detection, the region to which the top-left corner of the face detection output belongs is identified (as shown by the star in Figure 3). This region directly corresponds to the tactor on the belt that needs to be activated to indicate the direction of the person with respect to the user. To this end, a control byte is used to communicate between the software and the hardware components of the system. Regions 1 through 7 are coded into 7 bits on the parallel port of a PC. Depending on the location of the face image, the corresponding bit is set to 1. The software also controls the duration of the vibration by using timers. The duration of a vibration indicates the distance between the user and the person in his or her visual field. The longer the vibration, the closer the people are, which is estimated by the face image size determined during the face detection process.

An overall perspective of the system and its process flow is given below. When a user encounters a person in his or her field of view, the face is detected and recognized (if the person is not in the face database, the user can add it). The delivery of information comprises two steps: Firstly, the identity of the person is audibly communicated to the user (we are currently investigating the use of tactons [6][7] to convey identities through touch, but this is part of future work). Secondly, the location of the person is conveyed through a vibrotactile cue in the haptic belt, where the location of the vibration indicates the direction of the person and the duration of vibration indicates the distance between the person and the user. Based on user preference, this information can be repeatedly conveyed with every captured frame, or just when the direction or distance of the person has changed. The presence of multiple people in the visual field is not problematic as long as faces are not occluded and can be detected and recognized by the Social Interaction Assistant. We are currently investigating how to effectively and efficiently communicate non-verbal communication cues when the user is interacting with more than one person.

In this chapter we introduce the sensing and the delivery end of the system that can deliver proxemics information to an individual who is blind or visually impaired. From the sensing end, we describe a face detection methodology that is capable of identifying exact boundaries of the face region through which we model the distance of the interaction partner from the person who is using the device. From the delivery end, we describe user tests that were conducted to determine the use of tactons for conveying direction and distance information.

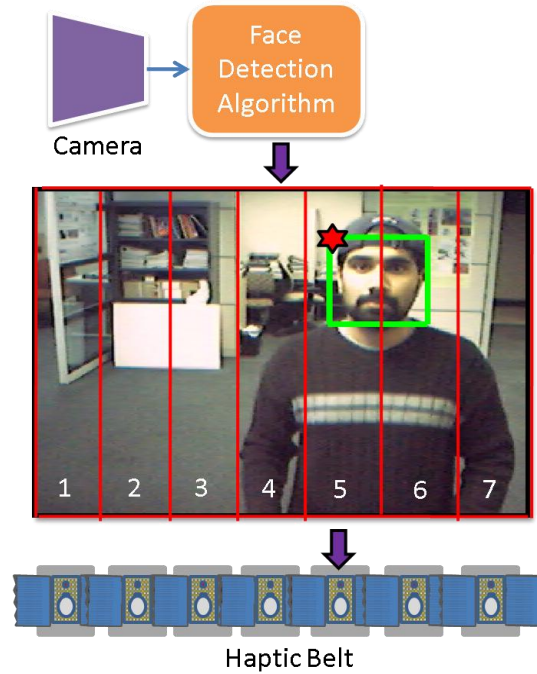


Figure 1: System Architecture for Haptic Belt used as part of the Social Interaction Assistant

### 4.3 Accurate Face Detection through the Wearable Camera

Face detection has become an important first step towards solving plethora of other computer vision problems like face recognition, face tracking, pose estimation, intent monitoring and other face related processing. Over the years many researchers have come up with algorithms that have over time, become very effective in detecting faces in complex backgrounds. Currently, the most popular face detection algorithm is the Viola-Jones [8] face detection algorithm whose popularity is boosted of by its availability in the open source computer vision library, OpenCV. Other popular face detection algorithms are identified in [9] and [10].

Most face detection algorithms learn faces by modeling the intensity distributions in upright face images. These algorithms tend to respond to face-like intensity distributions in image regions that do not depict any face as they are not contextually aware of the presence or absence of a human face. These spurious responses make the results unsuitable for further processing that requires accurate face images as inputs, such as the ones mentioned above. Figure 2 shows an example where a face detection algorithm detects two faces - one true and the other false.





Figure 2: An example false face detection

The problem of false face detection has motivated some researchers to develop heuristic approaches aimed for validating the face detection results. Most of these heuristics integrate primitive context into the problem by searching for skin tone in the output subimages. However, this simple approach often fails to distinguish faces from non-faces, because face detectors often fail to center the cropping box precisely around the detected face. This produces a significant patch of skin colored pixels, but only a partial face. This centering problem can be dealt with by extracting the skin colored regions and comparing their shape to an ellipse. While such heuristics are simple, and somewhat effective, their validation is not reliable enough to meet the needs of higher level face processing tasks. Further, they do not provide a confidence metric for their validation.

This section treats the problem of face detection validation in a systematic manner, and proposes a learning framework that incorporates both contextual and structural knowledge of human faces. A face validation filter is designed by combining two statistical modelers,

- 1) A human skin-tone detector with a dynamic background modeler (**Module 1**), and
- 2) An evidence-aggregating human face silhouette random field modeler (**Module 2**), which provides a confidence metric on its validation task.

The block diagram in Figure 3 shows the functional flow of data through the two modules in the proposed framework.

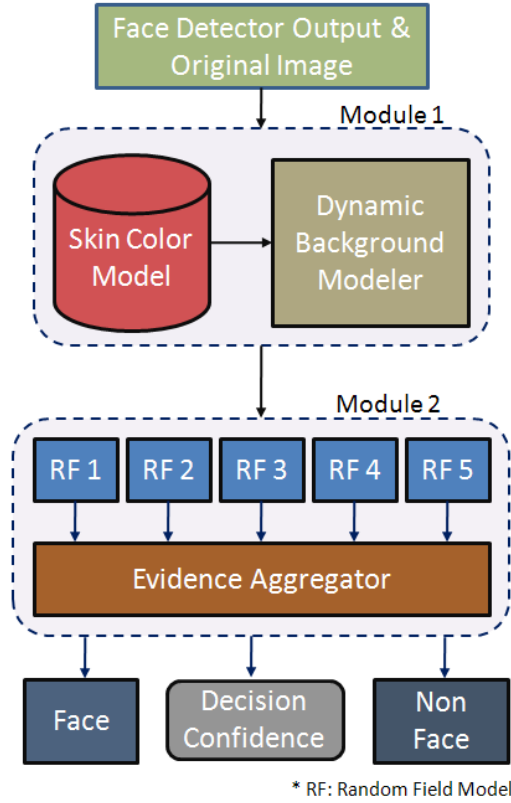


Figure 3: Block Diagram

#### 4.3.1 Face Validation Framework:

As shown in Figure 3, the framework essentially has two statistically learnt models, Module 1 and Module 2 that are cascaded to form the face detection validation filter. The output from a face detector is sent to Module 1, which distinguishes the skin pixels in the face region from the background pixels, thereby constructing a skin region mask. This skin region mask then becomes the input to Module 2, which is essentially an aggregate of random field models learnt from manually labeled (*true*) face detection outputs. The results of each random field model within the aggregate are then combined, using rules of Dempster-Shafer Theory of Evidence [11]. This *combining of evidence* provides a metric for the belief (i.e. confidence) of the system in its final validation. The two modules are detailed in the following subsections.

##### 4.3.1.1 Module 1: Human Skin Tone Detector with Dynamic Background Modeler

Most of the skin tone detectors used for human skin color classification use prior knowledge, which is provided in the form of a parametric or non-parametric model of skin samples that are extracted from images - either manually, or through a semiautomated process. In this paper we employ such an a priori model, in combination with a dynamic background modeler, so that the skin vs. non-skin boundary is accurately determined. Accurate skin region extraction is essential for Module 2, as it validates images based on their structural properties. The two functional components of Module 1 are:



#### 4.3.1.1.1 a-priori Bi-modal Gaussian Mixture Model for Human Skin Classification

A normalized RGB color space has been a popular choice among researchers for parametric modeling of human skin color. The normalized RGB (typically represented as nRGB) of a pixel  $X$  with  $X_r$ ,  $X_g$ ,  $X_b$  as its red, green and blue components respectively, is defined as:

$$X_{i|i \in \{r,g,b\}}^{nRGB} = \frac{X_i}{\left( \sum_{\forall i|i \in \{r,g,b\}} X_i \right)}$$

Normalized RGB space has the advantage that only two of the three components, nR, nG or nB, is required at any one time to describe the color. The third component can be derived from the other two as:

$$X_{i|i \in \{nR,nG,nB\}}^{nRGB} = 1 - \left( \sum_{\forall k|k \in \{nR,nG,nB\}, k \neq i} X_k \right)$$

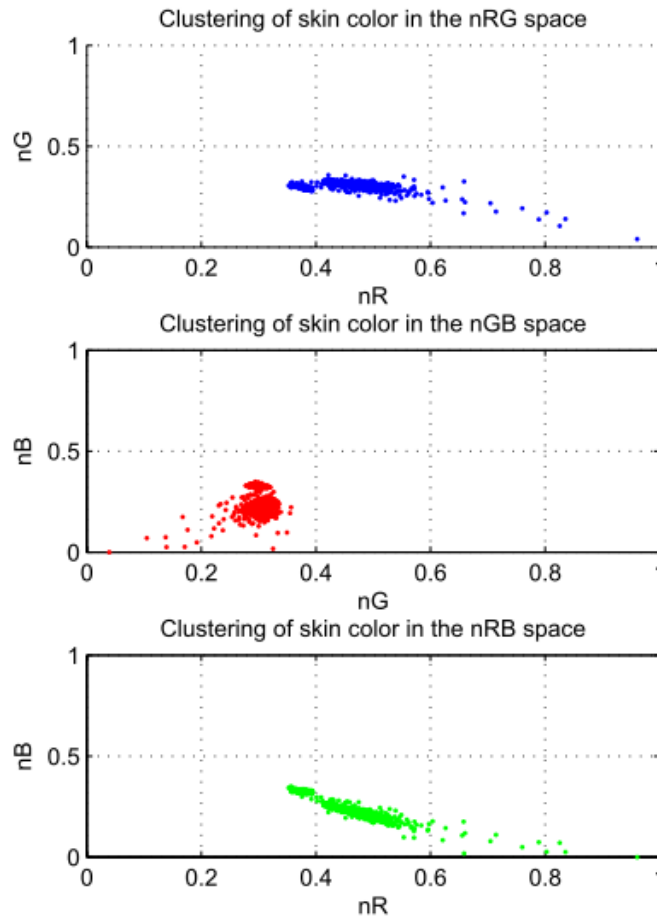


Figure 4: Skin Pixels in nRGB space

In our experiments, we found that skin pixels form a tight cluster when projected on nG and nB space as shown in the Figure 4. The study was based on a skin pixel database, consisting of nearly 150,000 samples, built by randomly sampling skin regions from 1040 face images collected on the web as well as from FERET face database [12]. Further analysis also showed that the cluster formed on the 2D nG-nB space had two prominent density peaks which motivated the modeling of skin pixels with a Bi-modal Gaussian mixture model learnt using

Expectation Maximization (EM) with a  $k$ -means initialization algorithm [13]. The Bi-modal Gaussian mixture model is represented as.

$$f_{X|X=[nG,nB]}^{skin}(x) = w_1 f_{Y_1}(x; \Theta_1 = [\mu_1, \Sigma_1]) + w_2 f_{Y_2}(x; \Theta_2 = [\mu_2, \Sigma_2])$$

#### 4.3.1.1.2 Dynamically Learnt Multi-modal Gaussian Model for Background Pixel Classification

As mentioned earlier, classification of regions into face or non-face requires accurate skin vs. non-skin classification. In order to achieve this, we learn the background color surrounding each face detector output dynamically. To this end we extract an extra region of the original image around the face detector's output, as shown in Figure 5. Since the size of the face detector output varies from image to image, it is necessary to normalize the size. This is done by down sampling the size of the original image to produce a face detector output region containing 90x90 pixels. The extra region pixels surrounding the face are then extracted from the 100x100 region around this 90x90 normalized face region.

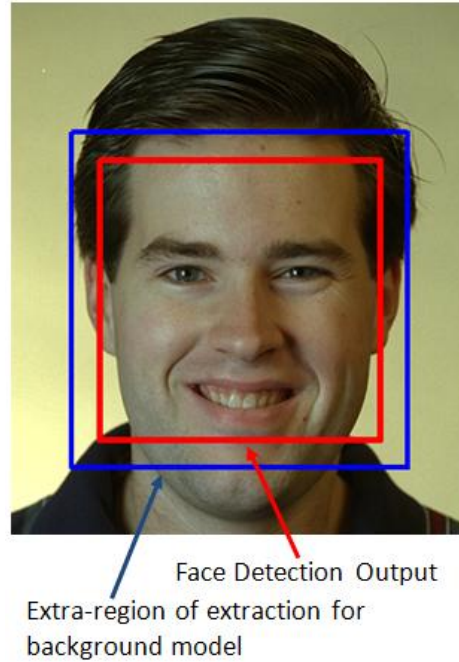


Figure 5: Extra region for background modeling

Once the outer pixels are extracted, a Multi-modal Gaussian Mixture is trained using EM with  $k$ -means initialization, similar to the earlier case with skin pixel model. The resultant model can be represented as

$$f_{X|X=[R,G,B]}^{non-skin}(x) = \sum_{i=1}^m w_i f_{Y_i}(x; \Theta_i = [\mu_i, \Sigma_i])$$

where,  $m$  is the number of mixtures in the model. We found empirically that a value of  $m=2$  or  $m=3$  modeled the backgrounds with sufficient accuracy.

#### 4.3.1.1.3 Skin and Background Classification using the learnt Multi-modal Gaussian Models

The skin and non-skin models,  $f_{X|X=[nG,nB]}^{skin}(x)$  and  $f_{X|X=[R,G,B]}^{non-skin}(x)$  respectively, are used for classifying every pixel in the scaled face image obtained as explained above. Example skin-

masks are shown in Figure 6. This example shows two sets of images - one corresponding to a *true* face detection result, and another *false* face detection result.

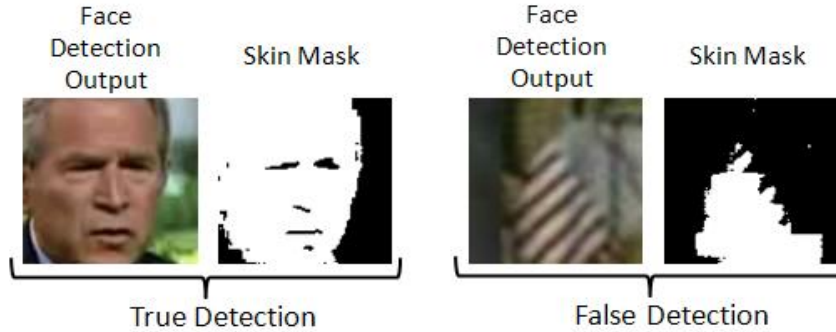


Figure 6: Example of *true* and *false* face detections

The structural analysis through Random Field models explained in the next section will describe the design concepts that will help distinguish between *true* and *false* face detections shown in Figure 6.

#### 4.3.1.2 Module 2: Evidence-Aggregating Human Face Silhouette Random Field Modeler

In order to validate the extracted skin region, we build statistical models from examples of faces. We developed statistical learners inspired by Markov Random Fields (MRF) to capture the variations possible in *true* skin masks (face silhouette). The following subsections describes MRF models and the variant we created for our experiments.

##### 4.3.1.2.1 Random Field (RF) Models

In this work, we used a minor variant of MRFs to learn the structure of a *true* face skin mask. MRFs encompass a class of probabilistic image analysis techniques that rely on modeling the intensity variations and interactions among the image pixels. MRFs have been widely used in low level image processing including, image reconstruction, texture classification and image segmentation [14].

In an MRF, the sites in a set,  $S$ , are related to one another via a neighborhood system, which is defined as  $N = \{N_i, i \in S\}$ , where  $N_i$  is the set of sites neighboring  $i$ ,  $i \neq N_i$  and  $i \in N_j \Leftrightarrow j \in N_i$ .

A random field  $X$  said to be an MRF on  $S$  with respect to a neighborhood system  $N$ , if and only if,

$$P(x) > 0, \forall x \in X$$

$$P(x_i | x_{S-\{i\}}) = P(x_i | x_{N_i})$$

where,  $P(x_i | x_{S-\{i\}})$  represents a Local Conditional Probability Density function defined over the neighborhood  $N$ . The variant of MRF that we created for our experiments relaxed the constraints imposed by MRFs on  $N$ . Typically, MRFs requires that sites in set  $S$  be contiguous neighbors. The relaxation in our case allows for distant sites to be grouped into the same model.

We empirically found out that modeling the skin-region validation problem into one single RF gave poor results. We devised 5 unique RF models with an Dempster-Shafer Evidence aggregating framework that could not only validate the face detection outputs, but also provide a metric of confidence. Thus,  $P(x_i | x_{S-\{i\}})$  could be alternatively seen as a set  $P(x) = \{P^1(x), \dots, P^5(x)\}$ , each having their own neighborhood system  $N^k = \{N^1, \dots, N^5\}$ , such that

$$P^k(x_i|x_{S-\{i\}}) = P(x_i|x_{N_i^k})$$

#### 4.3.1.2.2 Pre-processing

As described earlier, each face detector output is normalized and expanded to produce a 100x100 pixel image, from which a binary skin mask is generated. A morphological opening and closing operation is then performed on the skin mask (to eliminate isolated skin pixels), and the mask is then partitioned into one hundred 10x10 blocks, as shown in Figure 7. The number of mask pixels (which represent skin pixels) are counted in each block, and a 10x10 matrix is constructed, where each element of this matrix could contain a number between 0 and 100. This 10x10 matrix is then used as the basis for determining whether the face detector output is indeed a face.

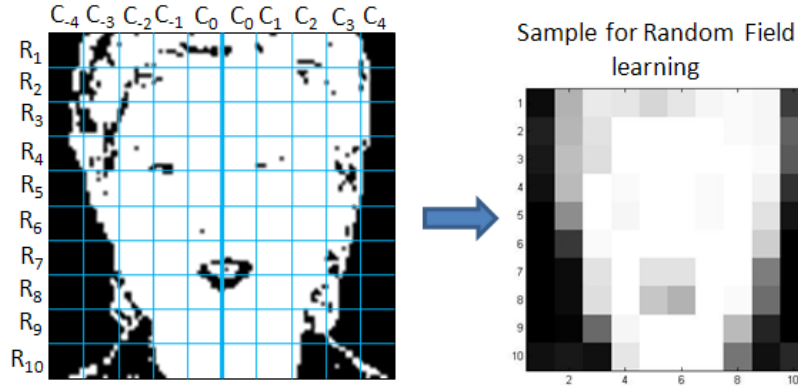


Figure 7: Pre-processing

#### 4.3.1.2.3 The Neighborhood System

The determination of whether the face detector output is actually a face is based on heuristics that are derived from anthropological human face models [15] and through our own statistical analysis. These include:

1. Human faces are horizontally symmetrical (i.e. along any row of blocks  $R_i$ ) about a central vertical line joining the nose bridge, the tip of the nose and the chin cleft, as shown in Figure 7. In particular, our analysis of a large set of frontal face images showed that the counts of skin pixels in the 10 blocks that form each row in Figure 7 were roughly symmetrical across this central line.
2. The variations along the verticals ( $C_i$ s) are negligible enough that in building a Local Conditional Probability Density function, each  $R_i$  can be considered independent of the other. That is, for example, modeling variations of  $C_0$  w.r.t  $C_1$  on  $R_1$  is similar to modeling variations of  $C_0$  w.r.t  $C_1$  on  $R_{i|i \neq 1}$ . Thus, analysis of Local Conditional Probability could be restricted to single  $R_i$  at a time, as shown in Figure 8.

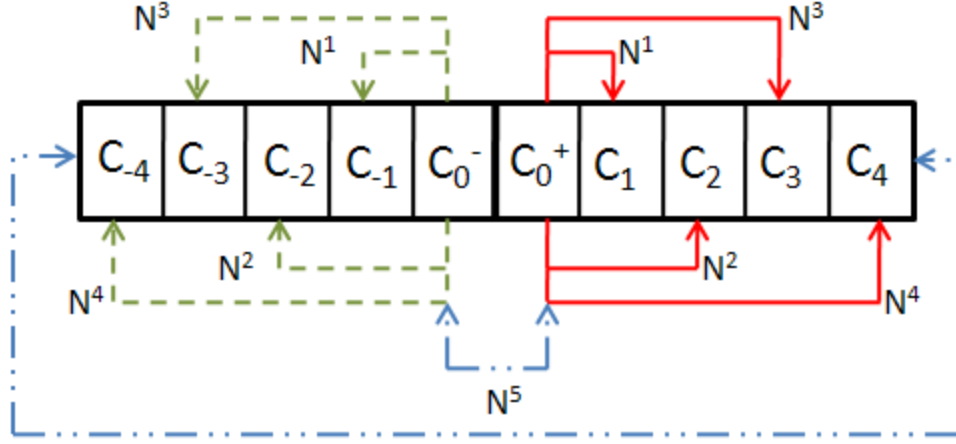


Figure 8: Neighborhood System

The different neighborhood systems  $N^k$ , used in the RF models,  $P^k(x|x_{N^k})$ , can be defined as (Refer Figure 8):

$$N^k = \{C_j | j \in \{|k|, 0^-, 0^+\}\}$$

#### 4.3.1.2.4 Local Conditional Probability Density (LCPD)

To model the variations on the skin-region mask, we choose to build 2D histogram for each of the 5 RF over their unique neighborhood system. The design of the dimensions were such that they captured the various structural properties of *true* skin masks. The two dimensions (represented in a histogram pool  $H^k$  with individual element of the pool,  $z$ , can be defined as:

$$H^{k \in \{1,2,3,4\}} = \{z\}, \text{ where } z = [x_{C_{0^\pm}}, \delta(x_{C_{0^\pm}}, x_{C_{k^\pm}})], \forall R_j$$

$$H^{k=5} = \{z\}, \text{ where } z = [\mu(x_{C_{0^-}}, x_{C_{0^+}}), \mu(x_{C_{4^-}}, x_{C_{4^+}})], \forall R_j$$

Where,  $x_{C_k}$  is the count of skin pixels in the block  $C_k$ . The two functions  $\delta(\dots)$  and  $\mu(\dots)$  are defined as

$$\delta(x_{C_{0^\pm}}, x_{C_{k^\pm}}) = \begin{cases} x_{C_{0^+}} - x_{C_{k^+}}, & i > 0 \\ x_{C_{0^-}} - x_{C_{k^-}}, & i < 0 \end{cases}$$

$$\mu(a, b) = \frac{a + b}{2}$$

In order to estimate the LCPD on these 5 histogram pools, we use Parzen Window Density Estimation (PWDE) technique, similar to [16], with a 2D Gaussian window. Thus, each of LCPD can now be defined as

$$P^k(z) = \frac{1}{(2\pi)^{\frac{d}{2}} n h_{opt}^d} \sum_{j=1}^n e^{\left[ -\frac{1}{2h_{opt}^2} (z - H_j^k)^T \Sigma^{-1} (z - H_j^k) \right]}$$

where,  $n$  is the number of samples in the histogram pool  $H^k$ ,  $d$  is number of dimensions (in our case 2),  $\Sigma$  and  $h_{opt}$  are the covariance matrix over  $H^k$  and the optimal window width, respectively, defined as:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad h_{opt} = \frac{\sigma_1 + \sigma_2}{2} \left\{ \frac{4}{n(2d+1)} \right\}^{\frac{1}{d+4}}$$

Figure 9 shows the LCPDs learnt over a set of 390 training frontal face images.

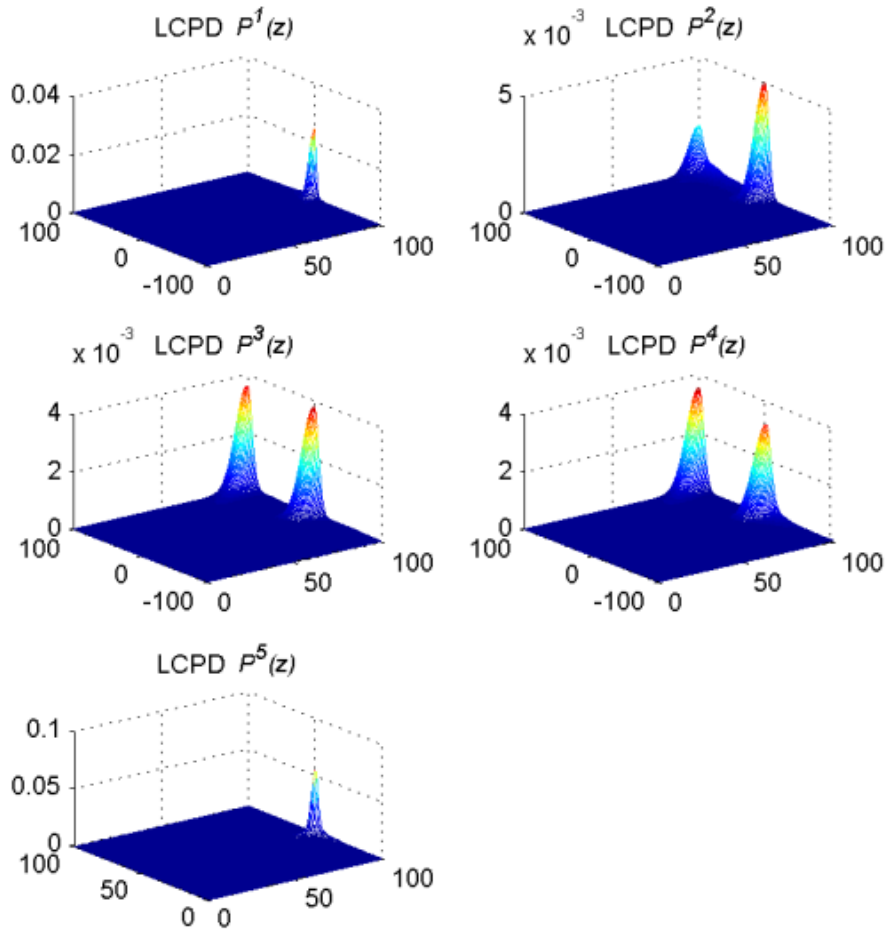


Figure 9: Frontal face Local Conditional Probability Density (LCPD) Models

#### 4.3.1.2.5 Human Face Pose

During our studies we discovered that the structure of the skin-region varies based on the pose of detected face as shown in Figure 10. Combining face examples from different pose into one set of RFs seemed to dilute the LCPDs and hence the discriminating capability. This motivated us to design three different sets of RFs, one for each pose. This was accomplished by grouping *true* face detections into three piles, Turned right (*r*), Facing front (*f*), and, Turned Left (*l*).

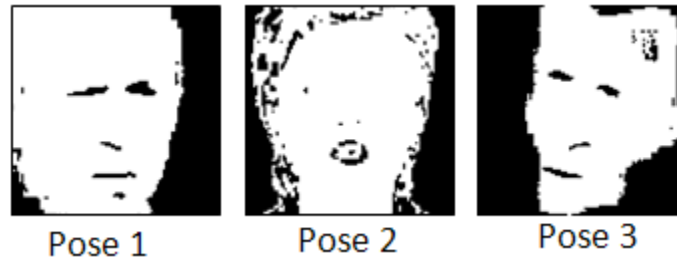


Figure 10: Skin-region masks.

Thus, the final set of LCPDs could be described by the super set.

$$P(z) = \left\{ P_m^k | k=\{1,\dots,5\} \right\}_{m=\{r,f,l\}}(z)$$

#### 4.3.1.3 Combining Evidence

Given any test face detection output,  $z$  is extracted and projected on the LCPD set  $P(z)$  to get a set of likelihoods  $l_m^k$ . As in the case of any likelihood analysis, we combined the joint likelihood of multiple projections using log-likelihood function,  $L_m^k = \ln(l_m^k)$ , such that,

$$\prod_{\forall z \in H_m^k} \ln(l_m^k(z)) = \sum_{\forall z \in H_m^k} L_m^k(z)$$

Given these log-likelihood values, one can set hard thresholds on each one of them to validate a face subimage discretely as *true* or *false*. We incorporated a piece-wise linear decision model (soft threshold) instead of a hard threshold on the acceptance of a face subimage. This is illustrated in the Figure 11. Each LCPD  $P^k(z)$  was provided with an upper and lower threshold of acceptance and rejection respectively. The upper and lower bounds were obtained by observing  $P^k(z)$  for the three face poses  $P_{r,f,l}^k(z)$ . Thus, any log-likelihood values lesser than the lower threshold ( $L_L$ ) would result in a decision against the test input (Probability 0), while any log-likelihood value greater than the upper threshold ( $L_U$ ) would be a certain accept (probability 1). Anything in between would be assigned a probability of acceptance. In order to combine the decisions from the five LCPD  $P^k(z)$ , we resort to Dempster-Shafer Theory of Evidence.

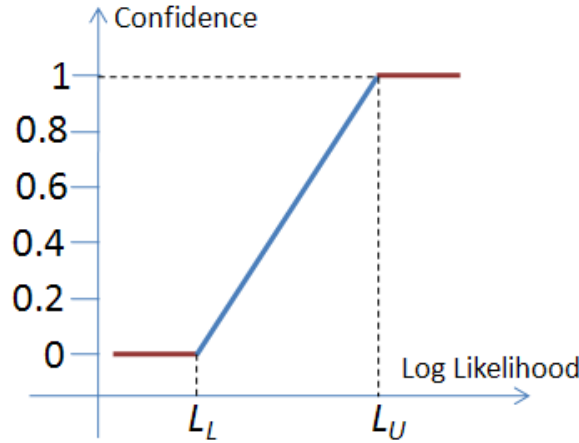


Figure 11: Soft threshold

##### 4.3.1.3.1 Dempster-Shafer Theory of Evidence (DST)

The Dempster-Shafer theory is a mathematical theory of evidence [11] which is a generalization of probability theory with probabilities assigned to sets rather than single entities.

If  $X$  is an universal set with power set,  $P(x)$  (Power set is the set of all possible sub-sets of  $X$ , including the empty set  $\emptyset$ ), then the theory of evidence assigns a belief mass to each subset of the power set through a function called the basic belief assignment (BBA),  $m: P(X) \rightarrow [0, 1]$ , when it complies with the two axioms.

- a)  $m(\emptyset) = 0$ , and
- b)  $\sum_{A \in P(X)} m(A) = 1$ .

The mass,  $m(A)$ , of a given member of the power set expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to  $A$  and to no particular subset of  $A$ . In our case,  $m(A)$  correlates to the probability assigned by each of LCPDs towards the subimage being a face or not.



The true use of DST in our application becomes clear with the *rules of combining evidences* which was proposed as an immediate extension of DST. According to the rule, the combined mass (evidence) of any two expert's opinions,  $m_1$  and  $m_2$ , can be represented as:

$$m_{1,2}(A) = \frac{1}{1-K} \sum_{B \cap C = A, A \neq \emptyset} m_1(B)m_2(C)$$

Where,

$$K = \sum_{B \cup C = \emptyset} m_1(B)m_2(C)$$

is a measure of the conflict in the experts opinions. The normalization factor,  $(1 - K)$ , has the effect of completely ignoring conflict and attributing any mass associated with conflict to a null set.

The 5 LCPDs, were considered as experts towards voting on the test input as a face or non-face. In order to use these mapped, we normalized evidences generated by the experts to map between  $[0,1]$ , and any conflict of opinions were added into the conflict factor,  $K$ . For the sake of clarity, we show an example of combining two expert opinions in Figure 12. The same idea could be extended to multiple experts.

		Expert 1's opinion	
		Face $m_1(B)$	Non-Face $m_1(C)$
Expert 2's Opinion	Face $m_2(B)$	Opinion Intersect $[m_1(B) * m_2(B)]$ (Sum in Numerator)	Opinion Conflict $[m_1(C) * m_2(B)]$ (Sum into $K$ )
	Non-face $m_2(C)$	Opinion Conflict $[m_1(B) * m_2(C)]$ (Sum into $K$ )	Opinion Intersect $[m_1(C) * m_2(C)]$ (Sum in Numerator)

Figure 12: An example of combining evidence from two experts under DST.

#### 4.3.1.4 Coarse Pose estimation

Since the RF models were biased with pose information, we also investigated the possibility of determining the pose of the face based on the evidences obtained from the LCPDs. We noticed that the LCPDs  $P^3(z)$ ,  $P^4(z)$  and  $P^5(z)$  were capable of not only discriminating faces from non-faces, but were also capable of voting towards one of 3 pose classes, Looking right, Frontal, and Looking Left along with a confidence metric.

#### 4.3.2 Experiments

In all our experiments, Viola-Jones face detection algorithm \cite{viola\_robust\_2004} was used for extracting face subimages. The proposed face validation filter was tested on two face image data sets,

1. The FERET Color Face Database, and
2. An in-house face image database created from interview videos of famous personalities.

In order to prepare the data for processing, face detection was performed on all the images in both the data sets. The number of face detections do not directly correlate to the number of unique face images as there are plenty of false detections. We manually identified each and every face detection to be *true* or *false* so that ground truth could be established. The details of this manual labeling is shown below:



1. FERET

- Number of actual face images: 14,051
- Number of faces detected using Viola-Jones algorithm: 6,208
- Number of *true* detections: 4,420
- Number of *false* detections: 1,788 (28.8%)

2. In-house database

- Number of actual face images: 2,597
- Number of faces detected using Viola-Jones algorithm: 2,324
- Number of *true* detections: 2,074
- Number of *false* detections: 250 (10.7%)

### 4.3.3 Results

In order to compare the performance of the proposed face validation filter, we defined four parameters:

1. **Number of false detections (NFD)**

$$\text{NFD} = \text{Count of false detections}$$

2. **False detection rate (FDR):**

$$\text{FDR} = (\# \text{ of false detections}) / (\text{Total } \# \text{ of face detections}) \times 100$$

3. **Precision (P)**

$$P = (\# \text{ of true detections}) / (\# \text{ of true detections} + \# \text{ of false detections})$$

4. **Capacity (C)**

$$C = (\# \text{ of true detections}) / (\# \text{ of actual faces in database}) - \text{FDR}$$

	Before Validation	After Validation
NFD	1,788	208
FDR	28.8%	3.35%
P	0.7120	0.9551
C	0.026	0.281

Table 1: Face detection validation results on FERET database

	Before Validation	After Validation
NFD	250	2
FDR	10.76%	0.01%
P	0.892	0.999
C	0.691	0.798

Table 1: Face detection validation results on the in-house face database

As explained above, the framework was extensible to perform coarse pose estimation. Figure 13 shows the result of passing two frames of a video sequence as input the face validation filter. The frames were extracted from a video of the same individual exhibiting arbitrary facial motion. The frames were 0.55 seconds apart. As can be noticed, the head pose is slightly different between the two frames. The pose estimation results are shown below the two frames.

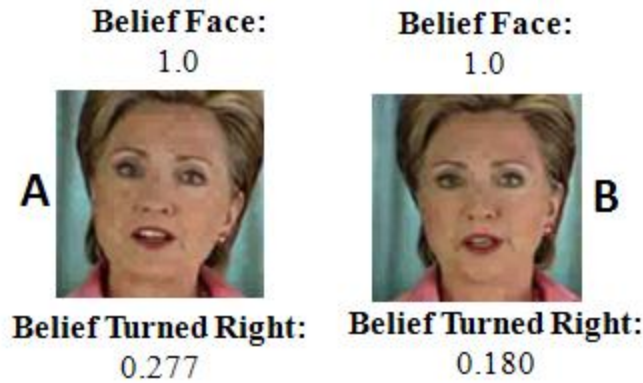


Figure 13: Coarse pose estimation

### Discussion of Results

Performance analysis of the proposed face validation filter can be understood through the four parameters defined in the previous section. NFB and FDR are direct measurements of the number of mistakes (naming non-faces as faces) made by the face detection algorithm on the two data sets. As can be verified from Table 1 and 2, there is a significant reduction in the false detections through the introduction of the filter.

The precision parameter,  $P$ , can be perceived as the probability that a face detection result retrieved at random will truly contain a face. It can be seen that the precision of the system drastically improves with the introduction of the face validation filter thereby assuring a *true* face subimage at the output.

The capacity parameter,  $C$ , measures the relative difference between face detection and false detection rates of a face detection system. Alternately,  $C$  can be considered to measure the net *true* face detection ability of any algorithm on a specific face data set.  $C$  ranges from -1 to 1. -1 when none of the faces in the database are detected with all reported detections being wrong. 1 when all the faces in the database are detected with no false detections. It can be seen from Tables 1 and 2 that the capacity of the face detection system, when combined with face validation filter, is significantly higher and moves towards 1. One can thus infer that the combined system has better *true* face detection ability.

Finally, Figure 13 shows the coarse pose estimation results. The two frames in the figure shows cases when the face is slightly turned right, with one (A) turned more right than the other (B). The face validation filter verifies that the faces are actually turned right and the belief values represent a scale on the amount of rotation. Since we did not do any specific mapping of the belief values to pose angle, we could not confirm quantitatively how accurate the pose estimations were. Through visual consort, one can verify that the labeling is meaningful.

### 4.4 Delivering Proxemics Information

As discussed earlier, the distance and the direction information extracted from the face detection algorithm were then conveyed to the user of the Social Interaction Assistant through the haptic belt. While the direction information is directly mapped to a vibrator, the distance information as encoded in the form of a varying temporal rhythm. The two experiments below represent this mapping from the perspective of conveying distance and direction information.

#### 4.4.1 Delivering Direction Data - *Localization of Vibrotactile Cues*

Prior work [17] showed that reasonable localization accuracy—between 80% to 100% accuracy depending upon tactor location—was possible with a belt design similar to what we used. Our experiment is similar, but offers a few variations to verify the results obtained in [17].

##### 4.4.1.1 Subjects:

10 subjects (8 males and 2 females), of ages between 24 and 59, participated in this experiment. One of the subjects was blind; the rest were sighted. Subjects had no known deficits related to their tactile sense of the waist area. Further, no subjects had prior experience with haptic belts, but all subjects had some exposure to vibrotactile cues (e.g., vibrations of a cell phone).

##### 4.4.1.2 Apparatus:

The haptic belt described in Section III was used for this experiment. Vibratory signals were 600 ms in length, and had a frequency and intensity well within the range of human perception. In contrast to [17], cues are longer—600 ms compared to 200 ms—and we do not use headphones to mask subtle vibration noise, nor do we randomly vary intensity with each cue; the reason for these changes is that we are mostly concerned with how the belt as a complete system accomplishes non-verbal communication, rather than the spatial acuity of the waist. Hence, if a specific intensity of vibration feels different around the waist, and some vibrations can be heard, and if these cues help in tactor localization, then this redundant information should only add to the usability of the system.

##### 4.4.1.3 Procedure:

Subjects put on the haptic belt over their shirt and around their waist such that the middle tactor (#4) was centered at their navel, and the endpoint tactors (#1 and #7) were at their left and right sides, respectively. As the belt has LEDs that light up to indicate tactor activation (used for testing the belt), subjects were instructed to not look down at the belt any time during the experiment. Next, subjects were familiarized with tactor numbering: the experimenter activated tactors in order from #1 to #7, and spoke aloud the number of the activated tactor. This process was repeated twice for each subject.

The training phase involved 35 trials where each tactor was randomly activated 5 times (with approximately 5 seconds between tactor activations) and subjects had to identify the number of each activated tactor. A visual guide was provided for subjects to help recall tactor numbers; this guide was a white board with a drawing of a semicircle (the belt) and the numbers 1 through 7 (tactors) on the belt. Feedback was given during the training phase to correct wrong guesses. The testing phase was similar to the training phase, but involved 70 trials where each tactor was randomly activated 10 times, and feedback was not provided. Subjects stood during the entire experiment.

##### 4.4.1.4 Results:

The localization accuracy for each tactor (number of times identified correctly out of the total number of times activated) was averaged across subjects and is shown in Figure 4 (indicated by the dots centered within each error bar), where error bars indicate 95% confidence intervals. The overall localization accuracy across tactors and subjects was  $(92.1 \pm 7.0)\%$ .

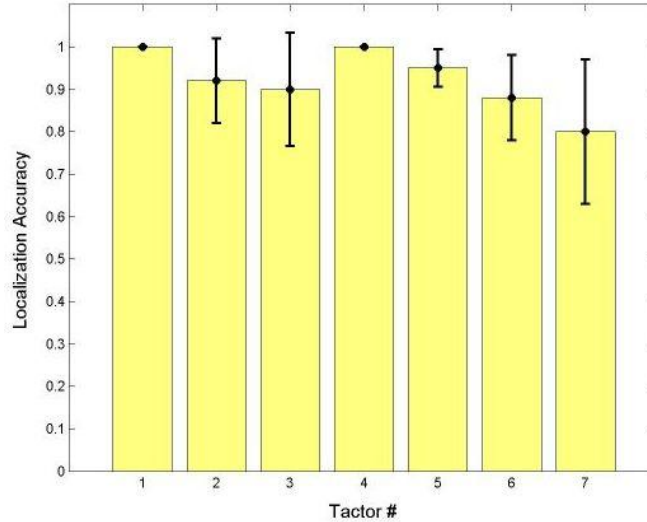


Figure 14: Experiment 1 Results: Mean Localization Accuracy for each Tactor, Averaged across Subjects, with 95% Confidence Intervals

#### 4.4.1.5 Discussion:

An overall localization accuracy of  $(92.1 \pm 7.0)\%$  (an improvement over that of [17]) is promising and shows that our prototype haptic belt can be reliably used to indicate the direction of someone in the user's visual field. Moreover, 100% of misclassifications were off by a single tactor location; hence, even when users made a mistake in localizing an activated tactor, they still had a very good idea of the general direction of someone in their visual field.

We hypothesize that the increase in accuracy is largely due to greater cue duration (600 ms as opposed to the 200 ms used in [17]); it is well known that larger cue durations make localization easier [18]. Moreover, redundant information provided by the belt, such as subtle audible cues when tactors are activated, could have helped as well. Subjects found tactors closer to the midline easier to localize, which agrees with the results found in the literature where spatial acuity improves near the sagittal plane [18] [17] given that spatial acuity is better at anatomical reference points—in this case, the navel.

It is hypothesized in [17] that the tactors at the end of the semicircle, which rest at the sides of the torso, act as landmarks and are easier to localize; but in our experiments, we noticed that tactor #1 could be localized more accurately than tactor #7, as shown in Figure 14.

### 4.4.2 Delivering Distance Data

#### 4.4.2.1 Tactile Rhythm Design

The tactile rhythms used in our experiments were motivated by results reported in [19], where just noticeable differences of vibrotactile duration were assessed. Subjects perceived pulses of duration below 100ms as a poke or nudge. Between 100ms to 2000ms, the just noticeable difference is an increasing curvilinear function of duration; although between 100ms to 500ms, the function is approximately linear. Based on these results, Geldard [19] recommended three durations, specifically 100ms, 300ms and 500ms, for accurate identification by subjects.

We conducted pilot studies to determine rhythm patterns that are convenient for users to identify vibratory rhythms. Through use of a vibrotactile belt, we evaluated use of five rhythms, each 10 seconds in length: 50ms vibrotactile pulses separated by pauses of length 50ms, 100ms, 300ms, 500ms and 1000ms. Subjects found rhythms with pauses of 100ms, 300ms and 500ms difficult to

discriminate between. Based on these findings, we selected the four rhythms depicted in figure 1; this design includes more separation of pauses within 100ms to 500ms, and a small increase of 1000ms to 1200ms (much longer durations may be too time consuming for communication [19]). In the Social Interaction Assistant, these four tactile rhythms are mapped to interpersonal distances corresponding to intimate, personal (close phase), personal (far phase) and social (close phase) space respectively.

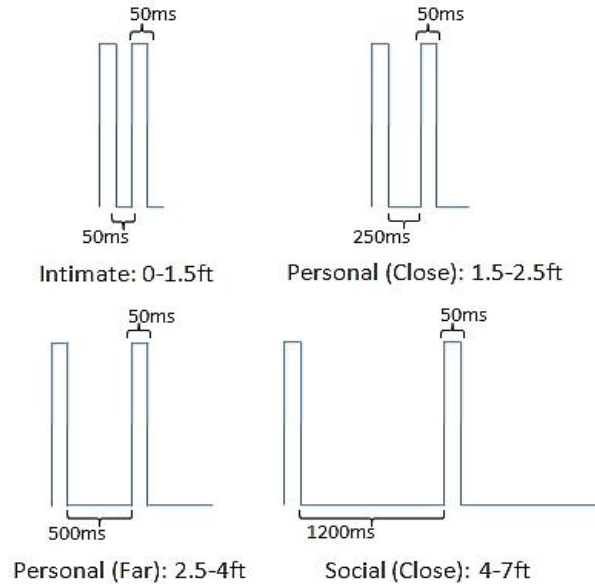


Figure 15. The on/off timing values of the four tactile rhythm designs, and corresponding distances, used in the experiment.

#### 4.4.2.2 Experiment

##### 4.4.2.2.1 Aim:

The aim of this experiment is to evaluate participants' performance identifying the tactile rhythms of figure 1 as they relate to interpersonal distances. Moreover, to ensure that the proposed tactile rhythms do not hamper subjects' ability to localize vibrations, as evaluated in previous work [20] to convey directions, we evaluate how well subjects can identify both cues as conveyed through tactons.

##### 4.4.2.2.2 Hypotheses:

- Subjects will achieve at least 90% accuracy at identification of tactile rhythms;
- Subjects will achieve at least 90% accuracy at identification of vibration locations;
- Subjects will achieve at least 80% accuracy at identification of complete tactons;
- Subjects' ability to localize vibrations will depend on the location of the vibration motor (tactor) around the waist;
- Subjects' ability to identify tactile rhythms will depend on the type of rhythm; and
- Subjects' ability to localize vibrations will not depend on rhythm type, and vice versa.

#### 4.4.2.2.3 Subjects:

11 males and 4 females of ages 22 to 60 (avg. 32) participated; one subject is visually impaired.

#### 4.4.2.2.4 Apparatus:

An elastic vibrotactile belt [20] was used for this experiment. The design of the belt was based on the experiments of Cholewiak, et al. [17]. The belt consists of 7 tactors equidistantly placed in a semi-circle with the first, fourth and seventh tactor at the user's left side, navel, and right side, respectively. Each tactor consists of a pancake motor of diameter 10mm and length of 3.4mm, and operates at 170Hz.

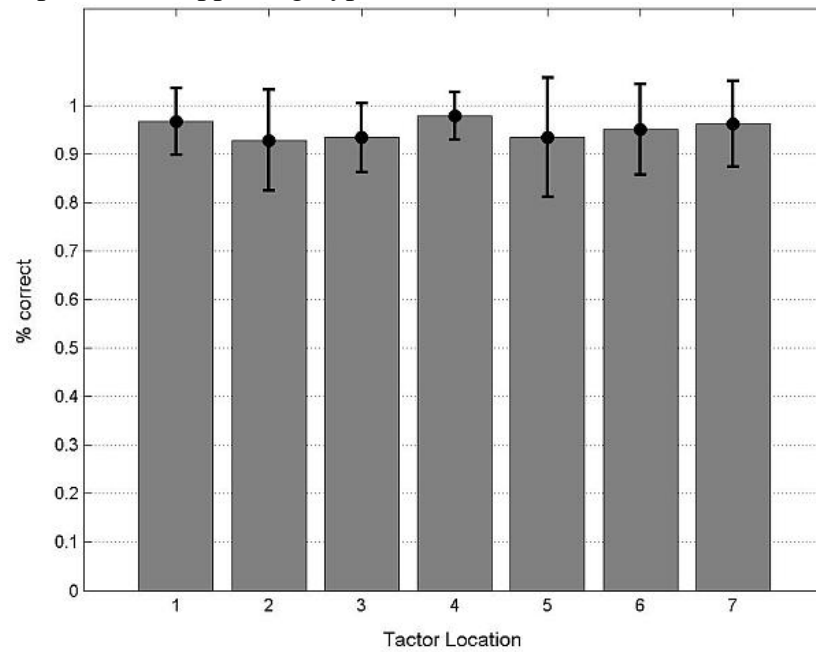
#### 4.4.2.2.5 Procedure:

Subjects wore the belt underneath their clothing and sat during the entire experiment. Subjects had access to visual guides—a semi-circle with tactors #1-7 drawn and interpersonal distances labeled as rhythms #1-4—to recall tactor and rhythm numbers, respectively. First, subjects were familiarized with vibration location as it pertains to direction. Each tactor was vibrated for 3 seconds, and the tactor number was called out by the experimenter. Next, subjects were familiarized with tactile rhythms. Each rhythm was presented for 7 seconds through the fourth tactor at the navel, and the rhythm number was called out by the experimenter. Next, subjects began the training phase where they were asked to identify the direction (through the location of the activated tactor) and distance (through the type of rhythm) indicated by each tacton. All 28 tactons (4 tactile rhythms at 7 different locations/tactors) were randomly presented for 10 seconds each. Subjects were encouraged to respond before the 10 seconds ended. Subjects had to achieve a recognition accuracy of 80% or more on each tacton dimension to proceed immediately to the testing phase; otherwise, the training phase was repeated (only 6 subjects had to repeat training, and all passed on the second try). The experimenter corrected wrong guesses and confirmed correct guesses. The testing phase was similar to the training phase, except no feedback was provided by the experimenter concerning right or wrong guesses, and each of the 28 tactons was randomly presented 3 times for a total of 84 trials.

#### 4.4.2.2.6 Results:

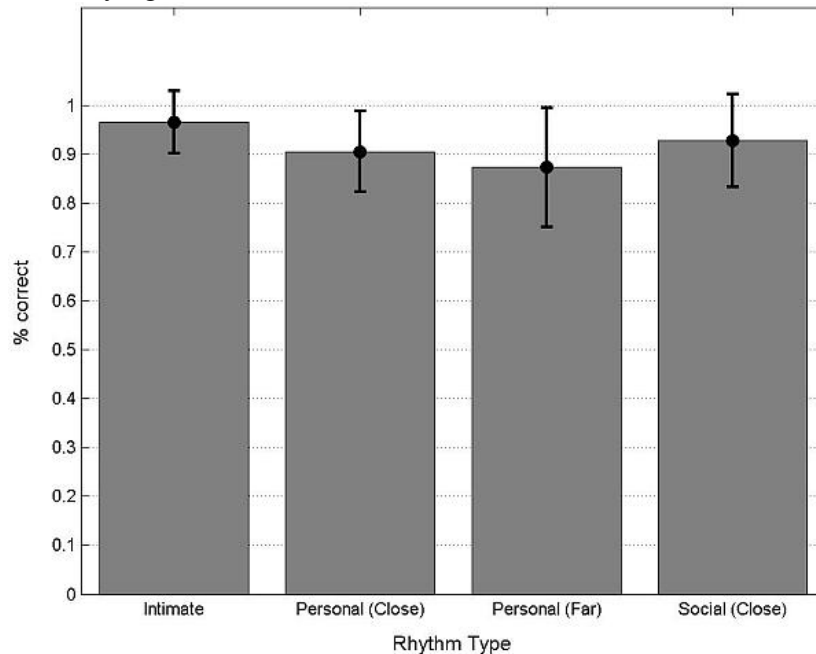
The overall recognition accuracy follows: location (mean: 95%, SD: 4%), rhythm (mean: 91.7%, SD: 5.7%) and both (mean: 87%, SD: 8.5%). These results support hypotheses (1)-(3), and show that, overall, subjects had little difficulty in recognizing rhythms and locations as they pertain to distance and direction, respectively. Feedback from participants after the experiment further supported this. From herein, reported ANOVA results are from a two-way ANOVA on complete tacton recognition accuracy through location and rhythm. The overall recognition accuracy of each tactor location is shown in figure 2. Subjects felt that the vibrations of tactor #1 (left side), #4 (navel) and #7 (right side), were easier to localize compared to tactor #2, #3, #5 and #6. This result is easy to explain as spatial acuity is better at anatomical reference points [17]. Although figure 2 does show a very small difference between recognition accuracies, which supports what subjects reported, there was no significant difference between recognition accuracy of tactor locations [ $F(6,1232)=1.96$ ,  $p=0.068$ ], hence hypothesis (4) cannot be accepted. The overall recognition accuracy of rhythms is shown in figure 3. Subjects felt that rhythm #2 (personal-close) and #3 (personal-far) were more difficult to identify than rhythm #1 (intimate) or #4 (social-close), which is supported by figure 3. A significant difference between recognition accuracy of rhythms [ $F(3,1232) = 5.70$ ,  $p=0.001$ ] supported hypothesis (5). No interaction was

found between location and rhythm for recognition accuracy of complete tactons [ $F(18,1232)=0.91$ ,  $p=0.569$ ], supporting hypothesis (6).



**Figure 15.** Overall direction recognition accuracy of each tactor location with standard deviations.

After the experiment, subjects filled out 10-level Likert scales—1 (lowest) to 10 (highest). Subjects rated their ability to localize vibrations (mean: 8.4), identify rhythms (mean: 7.4), intuitiveness of location to convey direction (mean: 9.7) and intuitiveness of rhythm to convey distance (mean: 8.9). Overall, subjects felt that they could accurately identify the proposed tactons, although identifying direction was easier than distance, and both schemes were intuitive.



**Figure 16.** Overall distance recognition accuracy of each rhythm type with standard deviations.



## Summary:

## References:

- [1] E.T. Hall, *The Hidden Dimension*, Anchor, 1990.
- [2] S. Ram and J. Sharf, "The People Sensor: A Mobility Aid for the Visually Impaired," *iswc*, vol. 00, 1998.
- [3] J.B.F.V. Erp, H.A.H.C.V. Veen, C. Jansen, and T. Dobbins, "Waypoint navigation with a vibrotactile waist belt," *ACM Transactions on Applied Perception*, vol. 2, 2005, pp. 106-117.
- [4] P. Barralon, G. Ng, G. Dumont, S.K.W. Schwarz, and M. Ansermino, "Development and evaluation of multidimensional tactons for a wearable tactile display," *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*, Singapore: ACM, 2007, pp. 186-189.
- [5] L. Brown, S. Brewster, and H. Purchase, "A first investigation into the effectiveness of Tactons," *Eurohaptics Conference, 2005 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2005. World Haptics 2005. First Joint*, 2005, pp. 167-176.
- [6] S. Brewster and L. Brown, "Tactons: structured tactile messages for non-visual information display," *AUIC '04: Proceedings of the fifth conference on Australasian user interface*, Australian Computer Society, Inc., 2004, pp. 23, 15.
- [7] P. Viola and M. Jones, "Robust Real-time Object Detection," *International Journal of Computer Vision*, 2001.
- [8] P. Viola and M. Jones, "Robust Real-time Object Detection," *International Journal of Computer Vision*, 2001.
- [9] E. Hjelmås and B.K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding*, vol. 83, Sep. 2001, pp. 236-274.
- [10] M. Yang, D.J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, 2002, pp. 34-58.
- [11] K. Sentz and S. Ferson, *Combination of evidence in dempster-shafer theory*, Sandia National Laboratories, 2002.
- [12] P. Phillips, Hyeonjoon Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, 2000, pp. 1090-1104.
- [13] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," 1997.
- [14] P. Perez, "Markov Random Fields and images," *CWI Quaterly*, vol. 11, 1998, pp. 413-437.
- [15] M. Vezjak, "An anthropological model for automatic recognition of the male human face," pp. 380, 363.
- [16] R. Paget, I.D. Longstaff, and B. Lovell, "Texture classification using nonparametric markov random fields," 1997, pp. 67-70.
- [17] R.W. CHOLEWIAK, J.C. BRILL, and A. SCHWAB, "Vibrotactile localization on the abdomen: Effects of place and space," *Perception & Psychophysics*, vol. 66, 2004, pp. 970-987.
- [18] J.B.V. Erp, "Vibrotactile Spatial Acuity on the Torso: Effects of Location and Timing



Parameters,” *World Haptics Conference*, Los Alamitos, CA, USA: IEEE Computer Society, 2005, pp. 80-85.

- [19] F.A. Geldard, “Adventures in tactile literacy.,” *American Psychologist*. Vol. 12(3), vol. 12, Mar. 1957, pp. 115-124.
- [20] T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan, “Using a Haptic Belt to convey Non-Verbal communication cues during Social Interactions to Individuals who are Blind.,” 2008.

# Chapter 5

---

## Exocentric Sensing & Delivery: Facial Expressions

---

As described in Chapter 2, in the survey conducted towards understanding the non-verbal cue needs for people who are blind and visually impaired, they emphasized on the lack of access to facial expressions and mannerisms of their interaction partners. This is supported by the argument that most part of the non-verbal cues occur through visual facial mannerisms as described in Section 1.2.1 of Chapter 1. The face encodes a lot of information that is both communicative and expressive in nature. Unfortunately, the face is a very complex data generator and the encodings on the face are not very context sensitive and individualistic in nature. Evolving computing technologies have been focused on developing solutions towards understanding the nature of facial mannerisms and gestures, but most of this multi-modal affective interaction research has been focused on the development of sensors and algorithms that understand user's emotional state in a human-machine interaction scenario. These interactions are mostly unilateral in nature and directed primarily towards the machine interpreting the user's emotional state. That is, the machines become the primary consumers of the affective cues. But from the perspective of an assistive technology affect interactions have to be augmentations that enrich human-human interpersonal interaction, where the machines not only interpret communicator's affective state, but also delivers affect information through novel affect actuators to a social interaction recipient.

As mentioned before most affect information is causal in nature and understanding what the expression or mannerism means requires an understanding of context when it is happening and the situation in which the communication is occurring. Our understanding of the cognitive models within the human brain that allows for the processing of complex facial expressions and emotions is very naïve. Computational models developed towards understanding context are very simplistic and performs nominally even under very well controlled laboratory conditions. Contrary to such a setting, assistive technologies provide some respite to the complexities by having the cognitive abilities of the user of the technology to make decisions. That is, while human computer interfaces need to mimic sensing, cognition and delivery, assistive technologies for people who are blind have to look at sensing and delivery alone and piggy back on human cognition. This requires precise sensing of the facial and head movements while delivering as much information back to the user as possible through technologies that do not overload the user with information but provides just the right level of information to allow them to cognitively process this information.

Thus, the focus of this chapter is on the *precise sensing* and *proficient delivery* of facial mannerisms and gestures of interaction partners to the user of the Social Interaction Assistant who is blind or visually impaired. To this end, the two important aspects of sense and delivery will be handled simultaneously to meet the goal of delivering dynamic facial and head movement information to the user of the social interaction assistant.

*From the sensing perspective, current ongoing experiments in tracking of facial expressions and mannerisms will be described in detail with identified areas that need special attention.*

From the delivery perspective, the latest in haptic interface will be introduced as a means of conveying facial and head mannerisms. Details on the experiments that have been carried out and the ones that need to be conducted will be illustrated.

## 5.1 Sensing Facial and Head Mannerisms and Expressions:

### 5.1.1 FaceAPI:

Going back to Chapter 2, the Social Interaction Assistant is built around the concept of the user carrying a tiny camera on the nose bridge of a pair of glasses. Thus, when they are involved in a bilateral conversation, the camera is looking out into the real-world and picking up the facial and head movements of the interaction partner. If it is possible to achieve real-time tracking of the head and facial features, one can try to deliver the same to the user of the social interaction assistant. To this end, we start with a real-time head and face tracking software sold by Seeing Machines Inc, called FaceAPI. The software provides us with 3D tracking of 38 facial fiducials while offering 28 points that define the face boundary. The software uses a face model fitting which allows 3D data to be provided using just a single camera. Further, once the camera and the lens on the Social Interaction Assistant are fixed, the electro optical image capture system can be calibrated so the FaceAPI software offers real-world depth data with the use of one single camera.

The software is capable of real-time tracking of the human face and can provide all the points mentioned above in 3D space referenced to a (0,0,0) that lies on the nose bridge of the human head that is being tracked. See Figure 1. The important real-time data of the interaction partner that can be extracted from the software includes,

1. Precise head position in the 3D world (Pose of the person w.r.t to the user).
2. Precise head movement through frames (Head based communicative gestures like head nod, head shake etc).
3. Approximate face mask area. (Points marked 8XX in the Figure 1).
4. Approximate tracking of 38 facial fiducials including.
  - a. The eye brow (6 points)
  - b. The eyes (not eye lids) (10 points)
  - c. Lips (Upper and Lower) (16 points)
  - d. The nose (4 points)

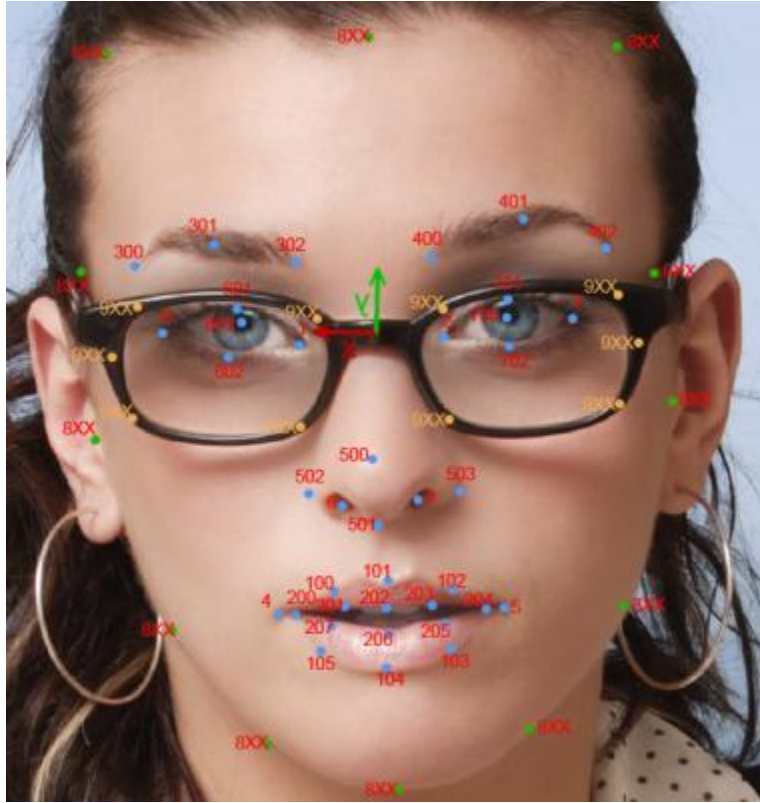


Figure 1: FaceAPI feature points along with head tracking

The greatest advantage of the FaceAPI software is the registration that it offers. The facial feature tracking is not very accurate due to the fact that the software uses models internally to fit the current appearance of the face image. But the registration offered by the software is very accurate and able to determine the exact location of the features themselves in every successive frame. We will use this registration ability to determine the various transformations that are happening on the face and head of the interaction partner.

### 5.1.2 Facial Image Features under Investigation:

As explained in Section 1.6.2 of Chapter 1, the facial expression analysis research is an active area within the computing community and has been working on various technologies for real-time tracking and extraction of facial features. In the past, we have conducted various experiments regarding facial expression analysis and it has been focused on distinct classification of the expression into one of six basis facial expressions, but the current exploration is not based on the classification into basic expression, but is more focused on determining the exact movement of the facial muscles while also achieving classification so that the information delivered to the user through the vibrotactile glove (introduced later in this chapter) can deliver the information of what facial movements were observed and what classification can be interpreted through the same. The goal of this expression analysis is to capture both the movement and classification information and deliver it to the user. To this end, we propose to work with image features that are computationally inexpensive to extract from the face which provides facial movement information which will be conveyed to the glove and also used for classifying the expression. We are still investigating how the movement information will be encoded on the glove such that the expression classification from the machine's end will be delivered subtly to the user while allowing the user to make own judgments about the expression.

### 5.1.2.1 Image features for expression recognition:

We are currently investigating two low level features for the extraction of movement information and also used for classification of the facial expression. The recognition rates have not been very effective, but we are investigating on using video to improve the efficiency. These two features include,

#### 5.1.2.1.1 LBP:

The original LBP operator, introduced by Ojala *et al.* [1], labels the pixels of an image by thresholding a 3x3 neighborhood of each pixel with the center value and considering the results as a binary number. Formally, given a pixel at  $(x_c, y_c)$ , the resulting LBP can be expressed in the decimal form as

$$LBP(X_c, Y_c) = \sum_{n=0}^7 s(i_n - i_c) 2^n$$

Where,

$$s(.) = \begin{cases} 1, & \text{if positive} \\ 0, & \text{otherwise} \end{cases}$$

where  $n$  runs over the 8 neighbors of the central pixel,  $i_c$  and  $i_n$  are the gray-level values of the central pixel and the surrounding pixel. The advantage of using LBP is in the low computational overhead when compared to some of the spectral features like Gabors and other wavelets (even including Harr). We may investigate the use of Harr and simple wavelets of that nature, but at the current point in time, we proposed to use simpler image level features.

Figure 2 below shows the LBP extraction on a face image. The face is divided into 8x8 non-overlapping blocks and histogram of the LBP features is extracted. We are currently working on classification of these features using kernel based learning techniques. Experiments with Support Vector Machine (SVM), AdaBoosted SVM and Kernel Discriminant Analysis (KDA) have been promising, but does not reach classification numbers suggested in the research papers. This discrepancy is due to the fact that most researchers work with fixed databases and the algorithms developed on these datasets seem to specialize the learning for that specific dataset. This will not work in our current application as generalization is very important for delivering data to the vibrotactile glove. To this end, we also propose to study the important regions of the human face from where features can be extracted. We propose to do this with an eye tracker as explained in the next section.

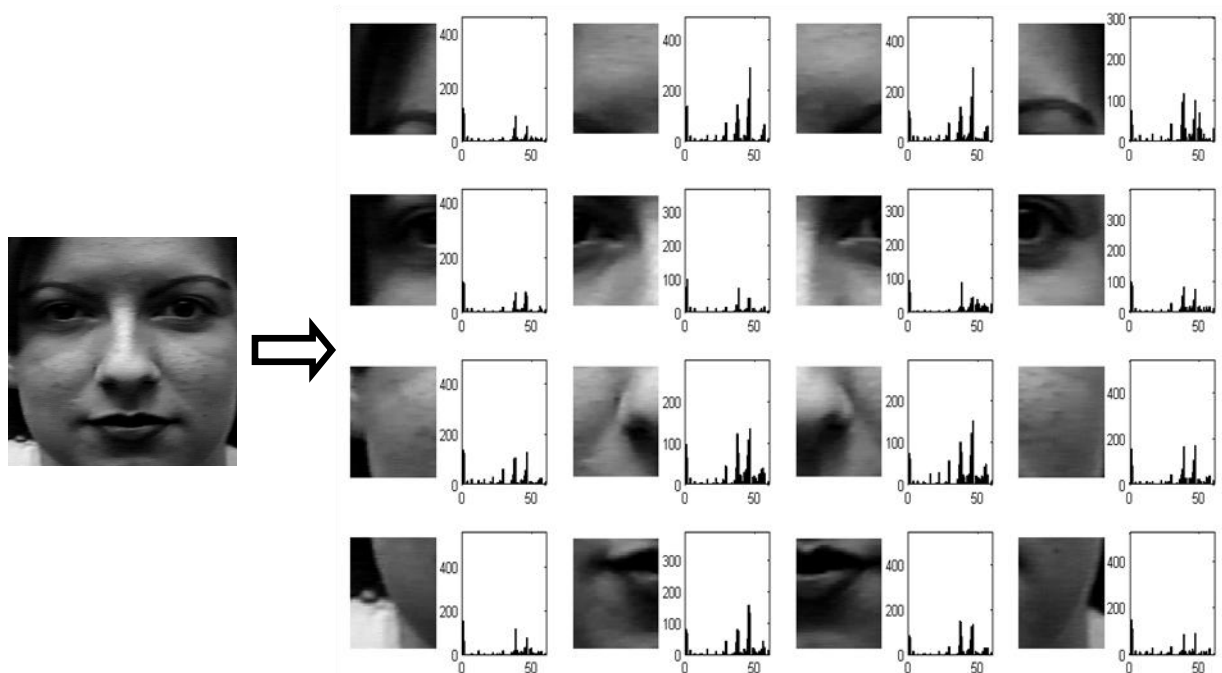


Figure 2: Example histogram of the local binary pattern exteacted over a sample face image by dividing the face into 8x8 non-overlapping windows.

#### 5.1.2.1.2 Line Segment Features:

Proposed originally in [2], the line segment features offer a means of extracting facial movements by monitoring the important facial fiducials. This allows the extraction of local facial feature information and the movement data precisely. Figure 3 shows the two important sets of data that is extracted from facial images. These include the distance information (subfigure (a)) between various facial fiducials and the orientation of the vectors (subfigure (b)) joining these fiducials. We are still experimenting with these features and plan on using them from videos of facial expressions in order to deliver movement information to the glove.

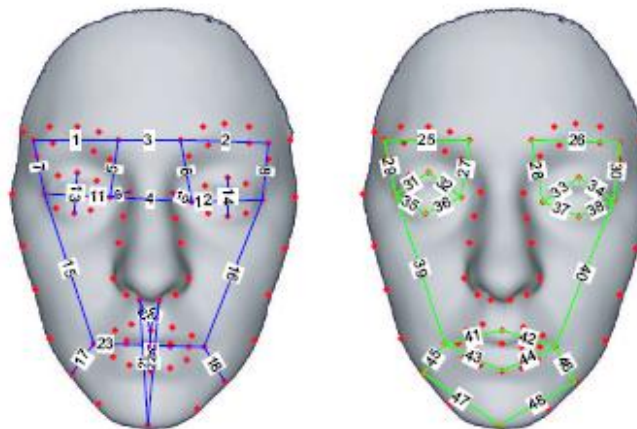


Figure 3: Line segment features, (a) corresponds to the distance features between facial fiducials, (b) corresponds to the orientation of the vectors joining the facial fiducials.



## 5.2 Importance of Facial Features:

Most of the computer vision algorithms treat the facial expression recognition problem as determining features that can classify the data into various bins. But when it comes to real-time expression conveyance, it is not just sufficient to consider the classification problem, but requires the analysis of the data from the perspective of pure motion patterns also. We need to convey the motion patterns and the machine's classification information to the user.

Most of the computer vision algorithms report very high detection accuracies on certain facial expression datasets. These reporting, while promising of the computer vision capabilities, are questionable in their use due to one simple fact that when the same images are presented to humans, they are not able to classify the expressions anywhere close to the algorithms. Further investigation of the questionable images by FACS experts reveal that the mimicked (or posed) expressions sometime do not match any prototype face that experts study. Thus, some of the image processing techniques are really over fitting the training data to achieve very high levels of accuracy. From this perspective we will study the problem of extracting the facial features by using human eye gaze data information. This will allow us to determine the important regions of the face and provide a realistic estimate on what the recognition accuracies are with humans. Further, we will investigate these facial expressions with FACS experts to make sure that we are able to validate any gaze data obtained from the general population into the expert group.

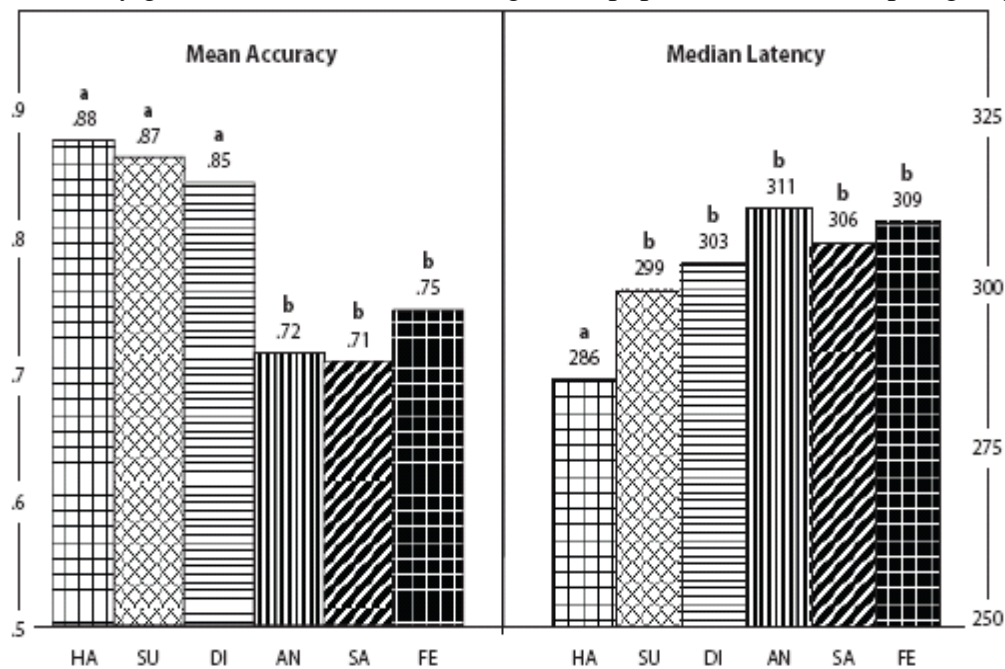


Figure 4: Dwell times in ms for each expression. Source [3]

Another important aspect of the eye gaze analysis will be the extraction of the duration information for facial expressions. There exists some psychology literature on the duration of facial expressions, but there is no comprehensive analysis of this information from a real-world interaction video perspective. We plan on using data of social interactions collected on our experimental social interaction assistant platform in the eye gaze experiments.

## 5.3 Delivering Facial Mannerisms and Expressions:

### 5.3.1 Design Considerations:

People who are blind rely on their auditory senses to understand and comprehend the environment around them. As described in detail in Chapter 1 Section 1.5.1, assistive technologies that use audio cues to deliver information back to a user can cause sensory overload leading to the rejection of any benefits that a device might offer. Especially during social interactions and bilateral conversations, it is imperative that any device should not hinder the primary sensory channel of the user. As shown in Section 1.7.1 of Chapter 1, Haptics offers a high-bandwidth channel for delivering information. As seen from the human homunculus, the hands form a perfect region to deliver this high bandwidth data. Of the various dimensions of somatosensory perception of the human skin, we choose to work with vibrators that can actuate the Meissner's Corpuscles or the Pacinian Corpuscles thereby allowing amplitude, frequency and rhythm as the primary dimensions to work with. A detailed background work on the use of vibrotactile actuations to convey information through the human hands can be found in Section 1.7.4 of Chapter 1. Here we describe in detail the construction of the vibrotactile glove and the mappings used for conveying facial expressions.

### 5.3.2 Construction of the Vibrotactile Glove:

In order to achieve the vibrotactile cueing, we have used shaft less vibration motors that incorporate off-centered mass to create vibrations. The proposed vibrotactile glove was built on top of a stretchable material glove that could fit most hand shapes. The glove has 14 tactors (vibration motors) mounted on the back of the fingers, one per phalange. The 14 motors correspond to the 14 phalanges (3 each on the index finger, middle finger, ring finger and the pinky with 2 on the thumb) on the human hand. A controller is also integrated on the glove to allow control of the motor's vibration (magnitude, duration and temporal rhythm) through the USB port of a PC.

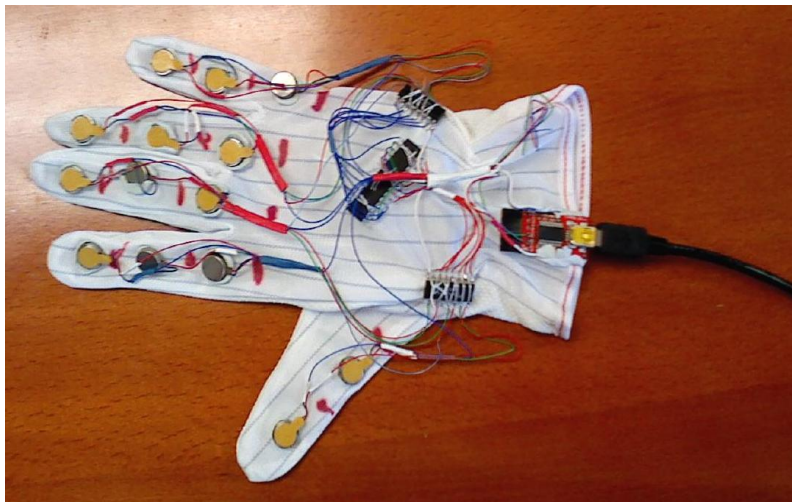


Figure 5: Haptic Glove: The figure shows a glove made out of stretchable material with 14 motors on the back of the glove with each motor corresponding to one phalange of the 5 digits. A microcontroller, two motor drivers and 1 USB controller (4 ICs) are also integrated on to the back of the glove with an ultra thin flexible USB cable leaving the glove.

Currently, the device only delivers the 6 basic expressions (Smile, Anger, Disgust, Surprise, Sad and Fear) along with indications of when the face reaches neutral expression. In future, we plan



to encode the dynamic motion of the human facial features into vibrotactile patterns. This would allow indiscriminate access to the facial movements of the interaction counterpart.

### 5.3.3 Mapping for facial expressions:

In order to encode the 6 basic expressions and neutral facial posture into haptic cues, we resorted to popular emoticon representations of these basic expressions. For example, smile is popularly represented by a smiley which was translated to a vibratory pattern of index finger top phalange, followed by middle finger bottom, followed by ring finger top phalange. The entire vibration sequence was completed within 750 milliseconds (The duration was arrived at after careful pilot studies with participants). The table below gives the vibration finger and phalange location in comma separated sequence for all 7 facial expression postures.

<b>Expression</b>	Comma separated vibration sequence. All sequences are 750ms long First letter indicates the finger – I for index, M for middle and R for Ring Second letter indicates the phalange – T for top, M for middle and B for bottom
<b>Smile</b>	IT, MB, RT
<b>Sad</b>	IB, MT, RB
<b>Surprise</b>	MT, IM, MB, RM, MT
<b>Anger</b>	IM, IB, MM, MB, RM, RB
<b>Neutral</b>	IM, MM, RM
<b>Disgust</b>	RB, MB, IB
<b>Fear</b>	IT, MT, RT, MT, IT, MT, RT

Table 1: Excitation sequence for the various vibrators on the vibrotactile glove.

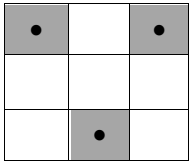


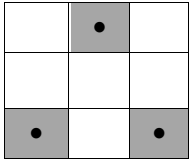


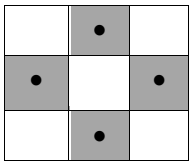


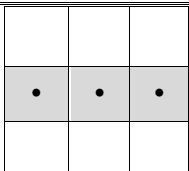
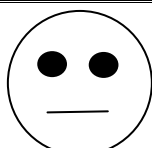
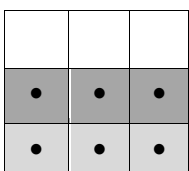


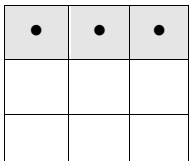


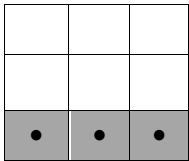


<b>Happy</b>			
<b>Sad</b>			
<b>Surprise</b>			
<b>Neutral</b>			
<b>Angry</b>			
<b>Fear</b>			
<b>Disgust</b>			

Table 2: Excitation table for the six basic expressions and the neutral face through the vibrotactile glove.

### 5.3.4 Experiments:

The above expressions were conveyed to 12 participants one of whom is blind. The participants were trained on the expressions until they were able to recognize all the expressions without any mistake after which 70 stimulations (10 trials of each expression) were presented sequentially with 5 seconds gap between each for the user to respond. The table below represents the results as a 7x7 confusion matrix where each cell entry corresponds to how many times (on average) users when given the row expression as stimulation responded with the column expression as

their answer. Following this average number, separated by a comma is the average time taken for answering.

### 5.3.5 Preliminary Results:

#### 5.3.5.1 Confusion Matrix:

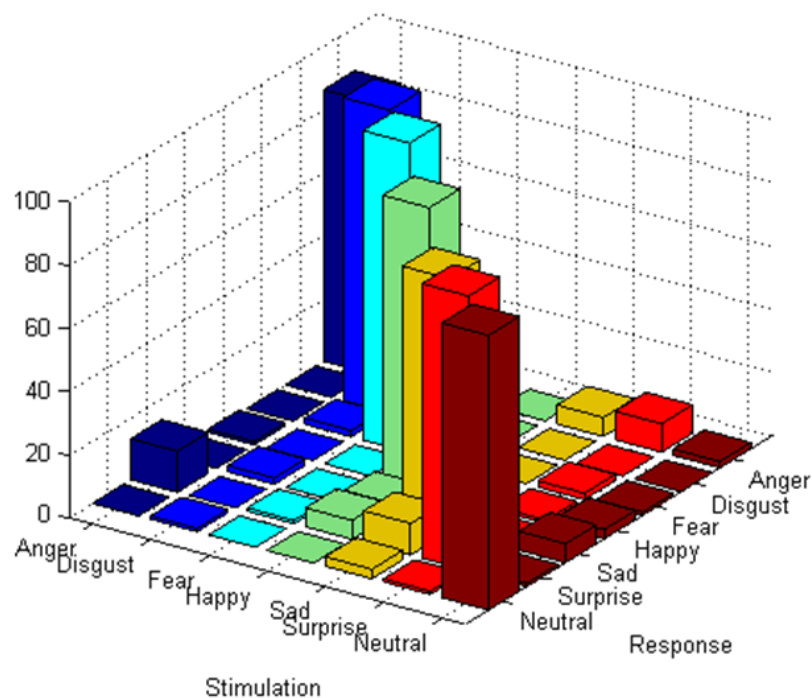


Figure 6: The 3D histogram plot of the confusion matrix. Each color represents a particular stimulation and the corresponding response from the users. This allows for the analysis of where confusion occurred in the delivered data.

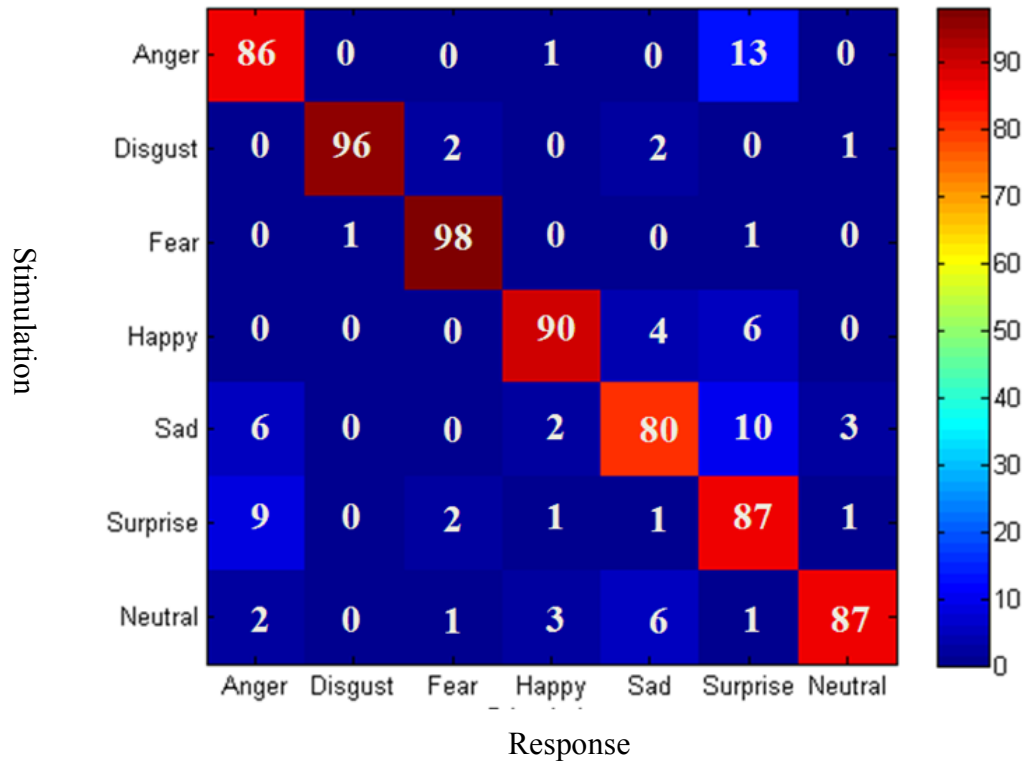


Figure 7: Confusion matrix of responses. Rows represent the stimulation provided to the users and the columns represent the response provided by the user. Each cell has two numbers. The first number represents the percentage recognition of a specific stimulation and a corresponding response. Ideally this matrix should have 100% recognition along the diagonal and zero off-diagonal.

By analyzing the confusion matrix, we can see that some of the design choices in delivering facial expressions were overlapping. This resulted in the confusion of some of the expressions like Anger with Surprise (Row 1, Column 6), Surprise with Anger (Row 6, Column 1), Sad with Surprise (Row 5 Column 6) etc. We are investigating how we can derive the importance maps for the vibration patterns on the glove as the interface. Once the importance maps for the various spatio-temporal patterns are extracted, we will be able to provide an automated process of delivering the facial movement data to the most appropriate region of the glove.

### 5.3.5.2 Response Time:

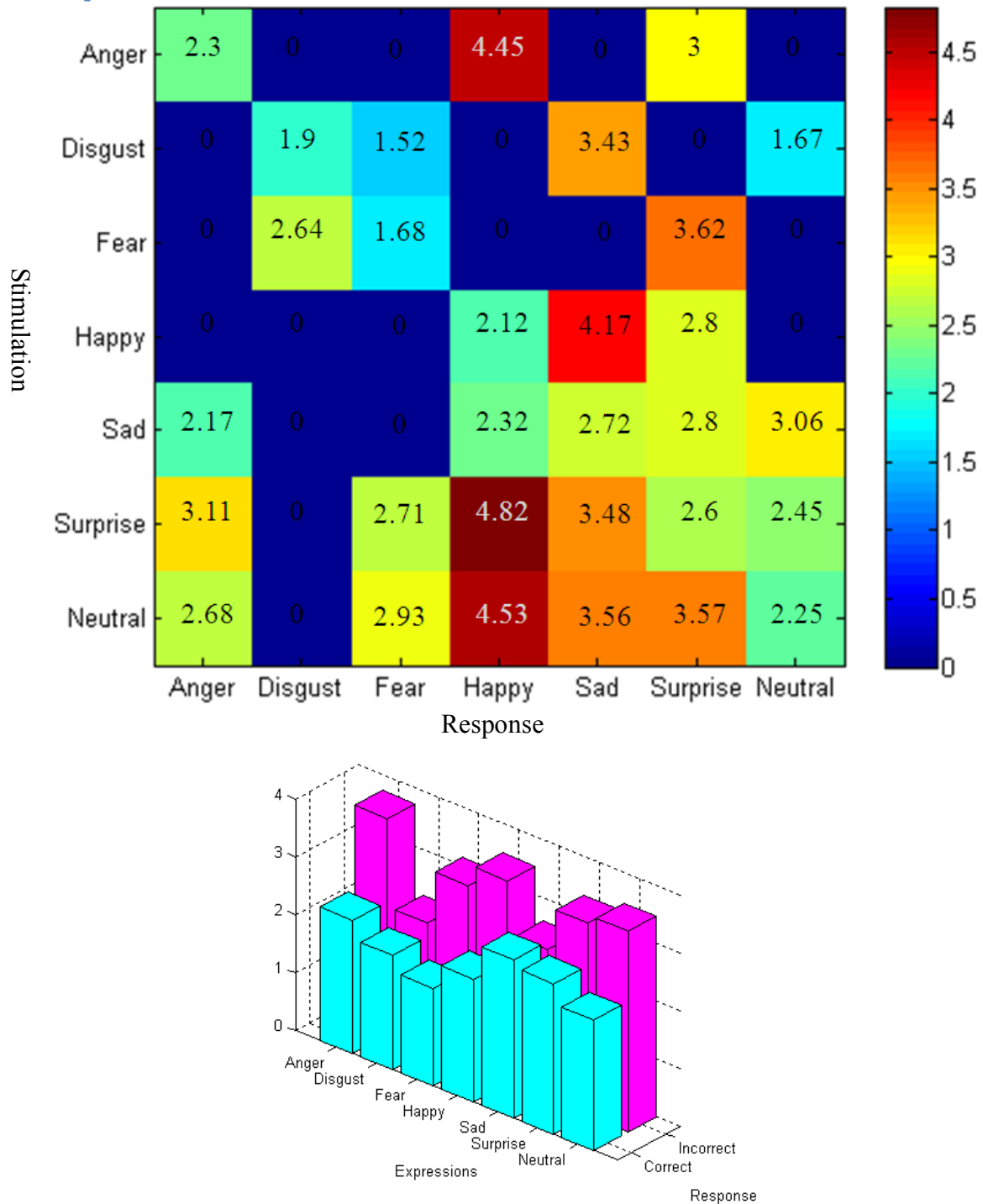


Figure 8: (a) Rows represent the stimulation provided to the users and the columns represent the response provided by the user. Each cell represents the response time for each stimulation and the corresponding response. Ideally, the time has to be as low as possible. (b) Represents the times taken when the users were able to recognize the expressions, in comparison with the data when the users were not able to recognize the expression.

### 5.3.6 Proposed work:

With the vibrotactile glove proven to deliver information to the users about basic facial expression, we will investigate how the same can be used for conveying real-time facial movement information. We will use the facial movement patterns extracted from the videos and correlate it with the facial importance maps that are extracted by using eye tracking information and the combined result will be mapped to the vibrotactile glove. Experiments will be conducted to determine the user's ability to understand the expressions that the interaction partners are displaying.

Since the mapping will be complex patterns of vibrations, we intend to conduct at least two on the same test subject at least one week apart to determine the retention on these vibrotactile patterns. Any results obtained in terms of the retention information will be used to modify the vibrotactile patterns so that the final mappings are intuitive yet informative in nature.

### 5.4 References:

- [1] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, 2002, pp. 971-987.
- [2] Hao Tang and T. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, 2008, pp. 1-6.
- [3] M.G. Calvo and L. Nummenmaa, "Eye-movement assessment of the time course in facial expression recognition: Neurophysiological implications," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 9, Dec. 2009, pp. 398-411.