Socio-Interpersonal Communications

Second Line

Third Line

by

Sreekar Krishna

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2011 by the
Graduate Supervisory Committee:

Sethuraman Panchanathan, Chair
John Black Jr.
Baoxin Li
Gang Qian
Michelle Shiota

ARIZONA STATE UNIVERSITY

January 2011

ABSTRACT

This is a sample abstract

Your dedication goes here.

ACKNOWLEDGEMENTS

[Enter your text here]

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Human interpersonal interactions are socially driven exchanges of verbal and non-verbal communicative cues. The essence of humans as social animals is very well exemplified in the way humans interact face-to-face with one another. Even in a brief exchange of eye gaze, humans communicate a lot of information about themselves, while assessing a lot about others around them. Though not much is spoken, plenty is always said. We still do not understand the nature of human communication and why face-to-face interactions are so significant for us.

Social interaction refers to any form of mutual communication between two individuals or between an individual and a group [2]. Such communications involve any or all forms of sensory and motor activities as deemed necessary by the participants of the interaction. Social, Behavioral and Developmental Sociologists emphasize that the ability of individuals to effectively control expressive behavior is essential for the social and interpersonal functioning of our society. Such social interactions are the aggregate cause of social behaviors, social actions and social contact that helps not only in effective bilateral communication, but also in forming an efficient feedback driven behavioral learning loop. It is this feedback (termed as social feedback) that children use towards developing good social and communicative skills.

Recent studies in behavioral psychology are furthering our understanding of the importance of social behaviors and social actions in everyday context. Researchers have revealed an unconscious need in humans to mimic and imitate the mannerisms of their interaction partners. An increasing number of experiments have highlighted this need for imitation to be very primeval and that they offer an elegant channel for building trust and confidence between individuals.

1.1 Components of Social Interactions

From a neurological perspective, social interactions result from the complex interplay of cognition, action and perception tasks within the human brain. For example, the simple act of shaking hands involves interactions of sensory, motor and cognitive events. Two individuals who engage in the act of shaking hands have to first make eye contact, exchange emotional desire to interact (this usually happens through a complex set of face and body gestures, such as smile and increased upper body movements), determine the exact distance between themselves, move appropriately towards each other maintaining Proxemics (interpersonal distance) that are befitting of their cultural setting, engage in shaking hands, and finally, move apart assuming a conversational distance which is invariably wider than the hand shake distance. Verbal exchanges may occur before, during or after the hand shake itself. This example shows the need for sensory (visual senses of face and bodily actions, auditory verbal exchange etc.), perceptual (understanding expressions, distance between individuals etc.), and cognitive (recognizing the desire to interact, engaging in verbal communication etc.) exchange during social interactions. Further, though social interactions display such complex interplay, they have been studied in the human communication literature under two important categories [3], namely,

- *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.

While the spoken language plays an important role in communication, speech accounts for only 35% of the interpersonal exchanges. Nearly 65% of all information communication happens through non-verbal cues [4]. 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) [5]. A closer look at the various non-verbal communication modes can highlight the importance of the multi-modality of social exchanges (See Figure 1.1).

Non-verbal communication 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. In everyday social interactions, people communicate so effortlessly through both verbal and non-verbal cues that they are not cognizant of the complex interplay of their voice, face and body in establishing a smooth communication channel.

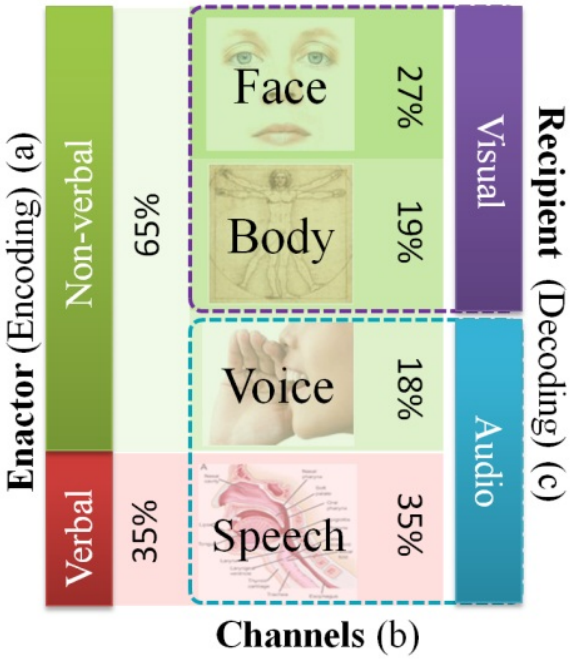


Figure 1.1: Relative importance of a) verbal vs. non-verbal cues, b) four channels of non-verbal cues, and c) visual vs. audio encoding and decoding of bilateral human interpersonal communicative cues.

Social Sight and Social Hearing

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 [6] 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 1.2 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.

Social Touch

Apart from visual and auditory channels of social stimulation, humans increasingly rely on social touch during interpersonal interactions. For example, hand shake represents an important aspect of social communication conveying confidence, trust, dominance and other important personal and professional skills [7]. Social touch has also been studied by psychologists in the context of emotional gratification. Wetzel [8] demonstrated patron gratification effects through tipping behavior when waitresses touched their patrons. Similar studies have revealed the importance of social touch and how conscious decision making is connected deeply with the human affect system. In the recent years social touch has gained a lot of interest in the area enriching remote interactions [9] [10] to help better understand an

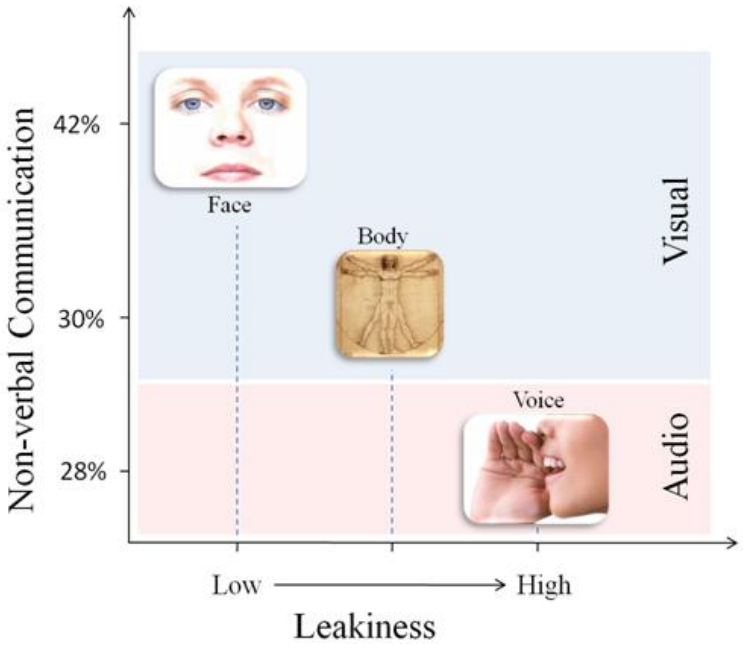


Figure 1.2: Relative communicative information plotted against its leakiness. Speech forms the verbal channel. Face, body and voice form the non-verbal communication channels.

individual's social awareness and social presence. In the next section, we describe the term *Social Situational Awareness* as seen pertinent to this report and emphasize the importance of any individual being aware of his/her social situational awareness.

1.2 Social Situational Awareness

We refer to the term Social Situational Awareness (SSA) as the ability of individuals to receive the visual, auditory and touch based non-verbal cues and respond appropriately through their voice, face and/or body (touch and gestures). Figure 1.3 represents the concept of consuming social cues and reacting accordingly to the needs of social interaction. Social cognition bridges stimulation and reciprocation and allows individuals to interpret and react to the non-verbal cues.

The Transactional Communication Model [11] suggests that during any face-to-face interaction, the interpretation of the social stimulation and the corresponding social response are under the control of various factors including the culture, physical and emotional

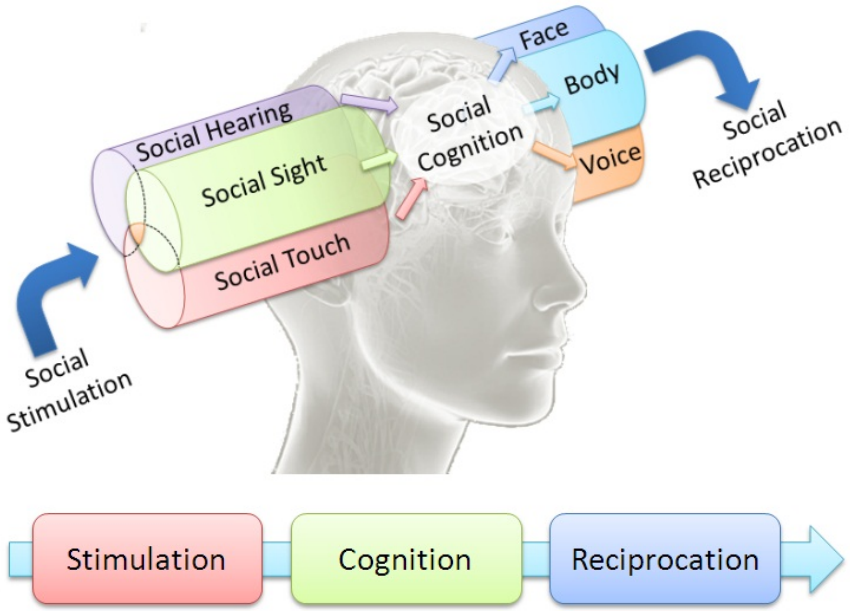


Figure 1.3: Social Situational Awareness.

state, experience, memory, expectation, self concept and attitude of the individuals involved in the interaction. In order to effectively cognize and react to the social stimulation, it is necessary that individuals be able to receive and synthesize these above factors. Enriching social situational awareness then represents the ability of a mediator (telecommunication technology for remote interactions; social assistive technologies for the disabled population) to allow the social cognition of an individual to have access to the above mentioned factors and thereby evoking appropriate social reciprocation.

Social Situational Awareness in Everyday Social Interactions

SSA in Dyadic Interactions

Human communication theories have studied dyadic or bilateral interaction between individuals as the basis of most communication models. Theories of leadership, conflict and trust base their findings on dyadic interaction primitives where the importance of the various non-verbal cues is heightened due to the one-on-one nature of dyadic interactions. Eye contact, head gestures (nod and shake), body posture (conveying dominance or submissive-

ness), social touch (hand shake, shoulder pat, hug, etc.), facial expressions and mannerisms (smile, surprise, inquiry, etc.), eye gestures (threatened gaze, inquisitive gaze, etc.) are some of the parameters that are studied closely in dyadic understanding of human bilateral communication [12]. Enriching SSA in dyadic communication thus focuses on appropriate extraction and delivery of communicator's face, body and voice based behaviors to a remote participant or to a person who is disabled.

SSA in Group Interactions

Group dynamics refer to the interactions between members of a team assembled together for a common purpose. For example, teams of medical professionals operating on a patient, a professional team meeting for achieving a certain goal, a congressional meeting on regulations, etc. represent groups of individuals with a shared mental model of what needs to be accomplished. Within such groups, communication behaviors play a vital role in determining the dynamics and outcome of the meeting. Zancanaro et. al. [13] and Dong et. al. [14] presented one model of identifying role-play of participants in a group discussion. They identified two distinct categories of roles for the individuals within the group, namely, the socio-emotion roles and the task roles. The socio-emotional roles included the protagonist, attacker, supporter and neutral, and the task roles included the orienteer, seeker, follower and giver. These roles were dependent heavily on the emotional state (affect) of the individuals participating in the group interaction. Good teams are those where individual team members and their leaders are able to compose and coordinate their affect towards a smooth and conflict free group interaction. And effective leaders are those who can read the affect of their group member, make decisions on individual's roles and steer the group towards effective and successful decisions. Inability to access the affective cues of team members has significant consequences to team leaders leading to unresolved conflict situations and underproductive meetings, or in the worst case, the death of a patient. Thus, enriching SSA in group settings correspond to the extraction and delivery of team's interaction dynamics (which are in turn modulated in their mutual and group affect) to a remotely located team member or to a co-located individual who is disabled.

In essence, SSA enrichment technologies provide for a richer interaction experience for individuals involved either in a dyadic or group interaction. It is well established that in teams comprising of good communication strategies a shared mental model towards effective decision is achieved faster with little or no emotional stress on the team members. The lack of social awareness can lead to interactions where individuals are not committed cognitively and find it very difficult to focus their attention on the communication. This is true in the case of remote interactions, disability and situations where doctors, nurses and other medical professionals are operating simultaneously on a patient.

Learning Social Awareness

Figure 1.3 represents a simple unidirectional model of social stimulation and reciprocation. In reality, social awareness is a continuous feedback learning system where individuals are learning through observing, predicting, enacting and correcting themselves. It is this learning mechanism that allows people to adapt easily from one culture to another with ease - here we refer to term culture in very broadly encompassing work culture, social culture in a new environment and culture of a new team, etc. Figure 1.4 shows the continuous feedback loop involved in social learning systems, based on the model of human cognition as proposed by Hawkins [15].

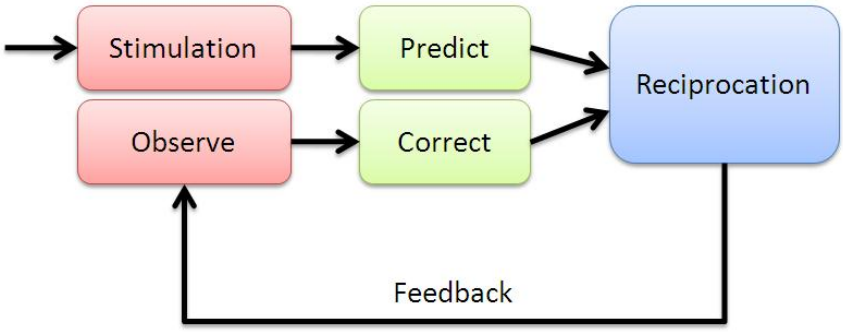


Figure 1.4: Social learning systems with continuous learning feedback loop.

People exposed to everyday social interactions learn social skills from the three different social stimulations (social sight, social hearing and social touch) effortlessly. When

faced with a new environment, individuals exercise their learned social skills to predict what social actions are appropriate in the setting. Once executed, they observe and assess their counterparts to determine if their new behavior is appropriate or not for the new setting. Such learning continues until their social rule set adapts to the new environment. Psychologists have been studying the nature of learning that happens in individuals who move from Western to Eastern cultures and vice versa. Largely, USA and Japan have been the countries of choice based on their economic equality and cultural diversity [16]. In the West, large body movements and excitement in the voice are considered to be typical and to a large part encouraged as a good social skill. Similar attitudes in the East are considered to be inappropriate in professional settings and to a large extent considered indecent. An individual displaying any such inappropriate mannerisms or gestures will receive social feedback from his counterparts (everyone staring at the individual, reduced interaction with the individual, etc.). Thus, social awareness is a learned set of rules about the environment within which the individual is present and this requires continuous monitoring of the various social channels of stimulation. Deprivation of any one of these channels can in turn affect the ability of the individual to learn social actions and responses that are pertinent to a social situation. Thus, enriching SSA not only offers the means for individuals to make appropriate social decisions, but also cognitively trains them towards effective social judgments.

———————————- In this paper, we advocate that the social separation induced by remote interactions in physically separated partners is similar to the social separation resulting from information impoverishment induced by sensory/physical disabilities in co-located interaction partners and propose technologies targeted at enriching 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 individuals non-verbal cueing [4]. These subdivisions include,

(a) The communication environment

(b) The physical characteristics of the communicators

(c) The behaviors of the communicators

Below, these three items are discussed in detail providing a highlevel discussion on the nature of their influence on the non-verbal communication between individuals.

The Communication Environment

The communication environment or surroundings where the interactions are taking place make a huge difference of how humans respond or react [17] [18]. For example, lengthy periods of extreme heat [19] 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 [20]. 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 1.1. **These are some of the well identified factors towards which psychologists and sociologists are working towards.**

Table 1.1: The various factors of the communicator's environment that can affect interpersonal communication.

The Communication Environment	
Familiarity of the environment	[21] [22]
Colors in the environment	[23] [24]
Other people in the environment	See next two subsections.
Architectural Designs	[25]
Objects in the environment	[26]
Sounds	[27] [28]
Lighting	[29]
Temperature	[19]

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 [30] 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 [31]. Table 1.3 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 [32].

Table 1.2: The physical characteristics of a communicator that can affect interpersonal communications.

The Physical Characteristics	
The human facial attractiveness	[30] [33] [34]
Body shape	[35] [36]
Height of a person	[37]
Self image	[38]
Body color	[39]
Body smell	[40] [41] [42]
Body hair	[43]
Clothing	[31] [44]
Personality	[45] [46]
Body decoration or artifacts	[47]

Physical Characteristics that affect interpersonal communication

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.

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 [48] [49].

(b) Speech-related gestures, or illustrators (pointing to a thing, drawing a shape while describing etc) [50].

(c) Punctuation gestures, that emphasize, organize and accent important segments of a communication, like pounding the hand, raising a fist in the air etc.

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 [51], submissiveness and social hierarchy [52]. 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 [53]. 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.

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 [54]. 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 [55]. 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 [56]. Historically, the sense of touch (Haptics Communication [57]) 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 [58].

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 [59]. 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 [60]. 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, Frisen [61] and Izard [62] 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 Frisen 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 Frisen 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.

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.

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 [63]. People use their gaze to convey subtle information that enables smooth verbal interaction which eventually leads to information exchange [64]. From a research perspective, the function of gaze has been classified into four important functional categories [65]. These include

Table 1.3: FACS communicative actions on the human face

1	Inner Brow Raiser	24	Lip Pressor
2	Outer Brow Raiser	25	Lips part
4	Brow Lowerer	26	Jaw Drop
5	Upper Lid Raiser	27	Mouth Stretch
6	Cheek Raiser	28	Lip Suck
7	Lid Tightener	29	Jaw Thrust
9	Nose Wrinkler	30	Jaw Sideways
10	Upper Lip Raiser	31	Jaw Clencher
11	Nasolabial Deepener	32	Lip Bite
12	Lip Corner Puller	33	Cheek Blow
13	Cheek Puffer	34	Cheek Puff
14	Dimpler	35	Cheek Suck
15	Lip Corner Depressor	36	Tongue Bulge
16	Lower Lip Depressor	37	Lip Wipe
17	Chin Raiser	38	Nostril Dilator
18	Lip Puckerer	39	Nostril Compressor
19	Tongue Out	41	Lid Droop
20	Lip stretcher	42	Slit
21	Neck Tightener	43	Eyes Closed
22	Lip Funneler	44	Squint
23	Lip Tightener	45	Blink
		46	Wink

Table 1.4: The role of human eye in interpersonal communications.

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 [66] and eliciting response from communication partners [67].
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 [68].
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 [69]. Thus, cognitive processing is displayed very elegantly by monitoring eye gaze patterns.
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 [70]. 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.

Chapter 2

MOTIVATION

In this chapter we discuss three important problems that highlight the need to communicate social situational awareness to individuals involved in interpersonal interactions.

2.1 Assistive Technology

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 nonverbal 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 [71], 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 . 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 . World Health Organization 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 [72] [73] [74] 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.

While most persons who are blind or visually impaired eventually make accommodations for the lack of visual information, and lead a healthy personal and professional life, the path towards learning effective accommodations could be positively effected through the use of assistive aids. Specifically, children with visual disabilities find it very difficult to learn social skills while growing amongst sighted peers, leading to social isolation and psychological problems [72]. Social disconnect due to visual disability has also been observed at the college level [75] where students start to learn professional skills and independent living skills. Any assistive technology aid that can enrich interpersonal social interactions could prove beneficial for persons who are visual disabled.

2.2 Remote Interactions

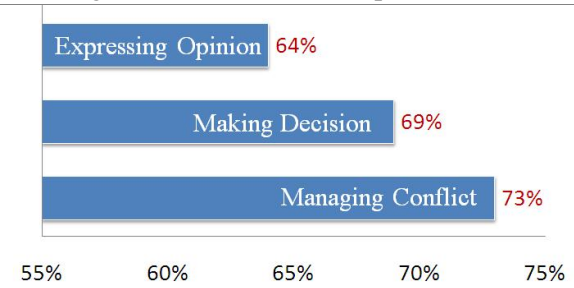
An industry survey [1] of 1592 individuals who collaborated remotely, carried out by RW3 CultureWizard - a company focused on improving international collaborations - reported difficulties similar to what was faced by the individuals who are blind. "Respondents found virtual teams more challenging than face-to-face teams in managing conflict (73%), making decisions (69%), and expressing opinions (64%). The top five challenges faced during virtual team meetings were insufficient time to build relationships (90%), speed of decision making (80%), different leadership styles (77%), method of decision making (76%), and colleagues who do not participate (75%)." These results can be correlated to the need for Social Situational Awareness in group settings, specifically one that can promote leadership and personal understanding of each other as indicated in Section 2.1.2.

Further, when the participants were asked about the personal challenges faced during virtual team meetings, they reported inability to read non-verbal cues (94%), absence of collegiality (85%), difficulty establishing rapport and trust (81%), difficulty seeing the whole picture (77%), reliance on email and telephone (68%), and a sense of isolation (66%)." Delivering non-verbal cues, establishing trust and rapport, and easing isolation are all derivatives of increasing one's social connection to their interaction partners, be it remote or face-to-face. Observing people who are disabled and the way they communicate with their co-located partners, it is possible to derive inspirations for novel social mediation technologies. The following subsection discusses one example of how to develop an evidence-based social situational awareness model based on hand shaking in the blind population as an example of social interaction between participants.

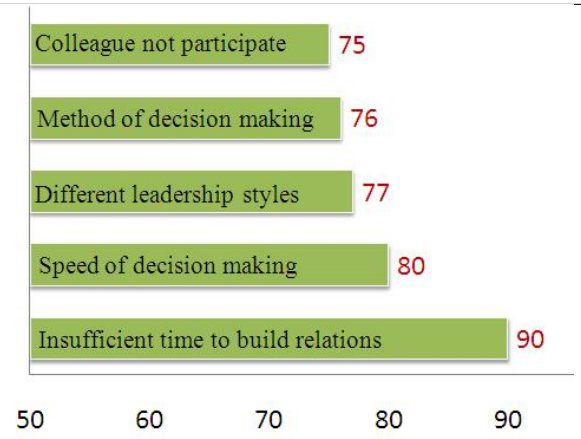
2.3 Medical Teams

Table 2.1: Survey on the challenges of remote interaction [1]

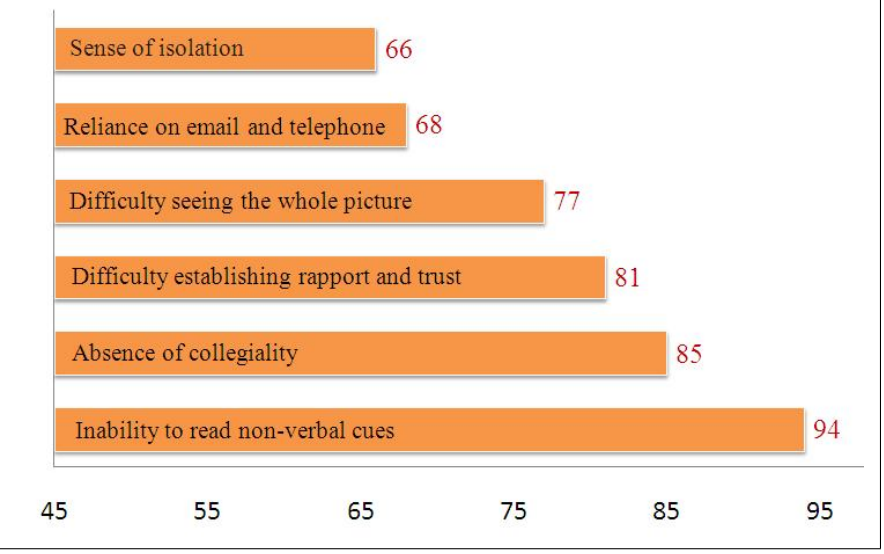
Challenges in virtual teams compared to face-to-face teams



Top five challenges faced during virtual team meetings



Personal challenges during virtual team meetings



Chapter 3

ASSISTIVE TECHNOLOGY DESIGN

Affective Computing research has employed algorithmic framework to quantitatively study both verbal and non-verbal cues displayed by the humans during social communication. Signal streams from various sensors, including visual sensors (e.g. cameras), audio sensors (e.g. microphones) and various physiological sensors (such as EEG, EMG, and galvanic skin resistance sensors) have been used to evaluate human emotional states. A good review of research work in Affective Computing can be found in [76]. This research has enabled a better understanding of human physiological signals, with respect to emotional states, and the results have been used to facilitate human-computer interaction (HCI). In theory, a system that can detect non-verbal social cues could also be used as an assistive device to provide social feedback to people with disabilities. The emphasis here would not be so much on interpreting these cues as on presenting social cue information to the user, and allowing the user to interpret them. However, very little research has been done towards finding intuitive methods for presenting social cue information to humans. [77] developed a haptic chair for presenting facial expression information. It was equipped with vibrotactile actuators on the back of the chair that represented some specific facial feature. Experiments conducted by the researchers showed that people were able to distinguish between six basic emotions. However, this solution had the obvious limitation that the user needed to be sitting in the chair to use the system.

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.

People with disabilities are not always able to perceive or interpret implicit social feedback as a guide to improving their communication competence. However, they might be able to use explicit feedback provided by a technological device. Rana and Picard [78] 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 the technology is being developed to address social interactions, 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.

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.

Vinciarelli et. al. [79] 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 [80] 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 [81] honest signals. Some of his recent works [82] 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, as it is not realistic to assume that everyone who interacts with a person with a disability will wear sensors.

Observation 3: Assistive technology designed towards social assistance should incorporate mechanisms embodied on the user to determine both self and other's social man-

nerism.

In two independent experiments [83] and [84], researchers developed a social feedback device that provides intervention when a person with visual impairment starts to rock their body displaying a stereotypy. [83] 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 [73], 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.

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.

In [84] 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.

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.

3.1 Conceptual Framework

Design principles for social assistive and rehabilitative devices

A device that is developed to facilitate the social interactions of people with sensory, or cognitive disabilities might do so by, (a) detecting social cues during social interactions and delivering that information to the user in real time to enable empathy, or (b) 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.

3.2 Requirements Analysis for a Social Assistive Technology for Individuals who are Blind and Visually Impaired

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 where conducted [1]. 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 [2].

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. Based on these discussions, a list of needs was complied 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

[1] 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.

[2] To quote some candidate's opinion about social assistance technology in a everyday setting:

- Ït 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, Ït would be great to walk into a bar and identify beautiful women.

with children who were visually impaired. She identifies these two needs under the *Social Learning* and *Social Feedback*. While these two important categories were identified, for simplification, 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.

Further, in order to understand the importance of these non-verbal communication primitives 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.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.

3.3 Results from the Online Survey

Average Response

Table 3.1 shows the eight aspects of social interactions that were investigated with the individuals who are blind and visually impaired. The results are 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.

Response on Individual Questions

Figure 3.1 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 Lickert Scale. Each histogram adds up to a total of 28 that corresponds to the 28 participants that took part in the online survey.

Some of the observations from the important histograms include,

- 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).
- 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

Table 3.1: Average Score on the 8 Questions obtained through an Online Survey.

Question No.	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

highlights the importance of facial mannerisms and gestures, which are mostly visual in their decoding.

- 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).

- 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

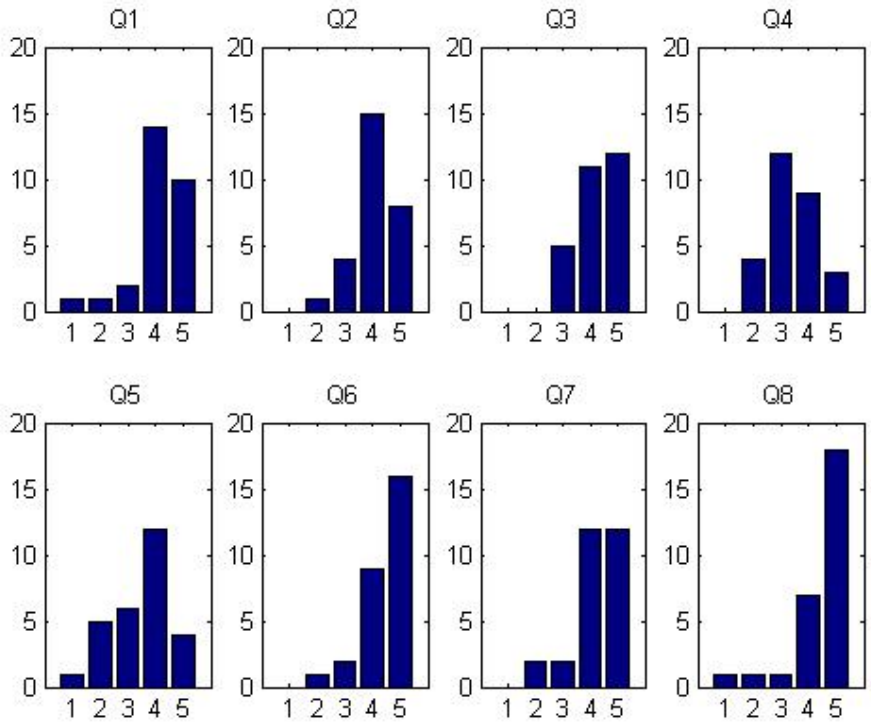


Figure 3.1: Histogram of Responses grouped by Questions

were standing. This can result in individuals addressing an empty space assuming that someone was standing there based on their memory.

- 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.

Response Ratio

Figure 3.2 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

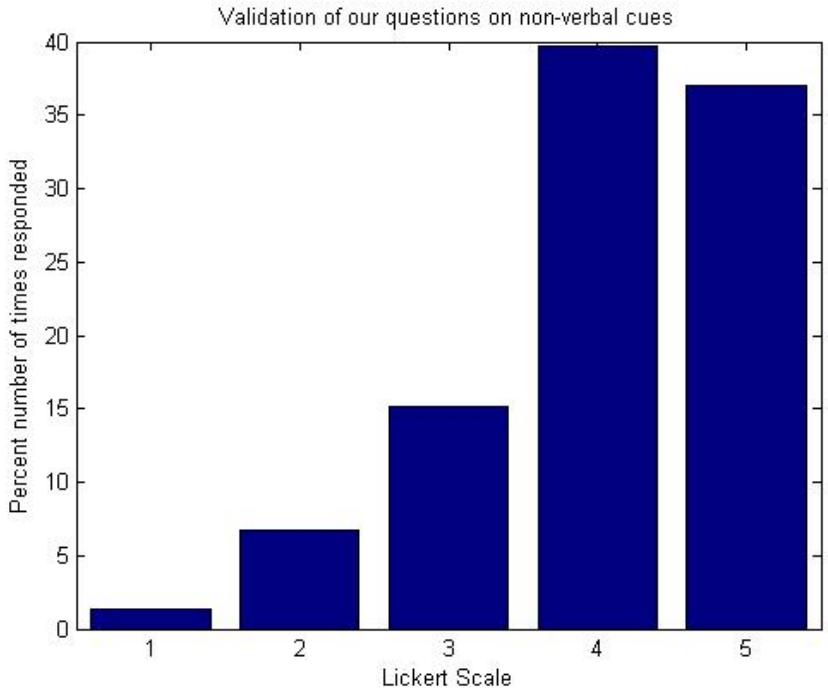


Figure 3.2: Response Ratio

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.

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).

Rank Average Importance Map for Various Non-verbal Cues

As can be seen from Figure 3.2, the questionnaires were biased and the frequency of the responses is not Gaussian. This bias implies that using sample mean of the Lickert Scale

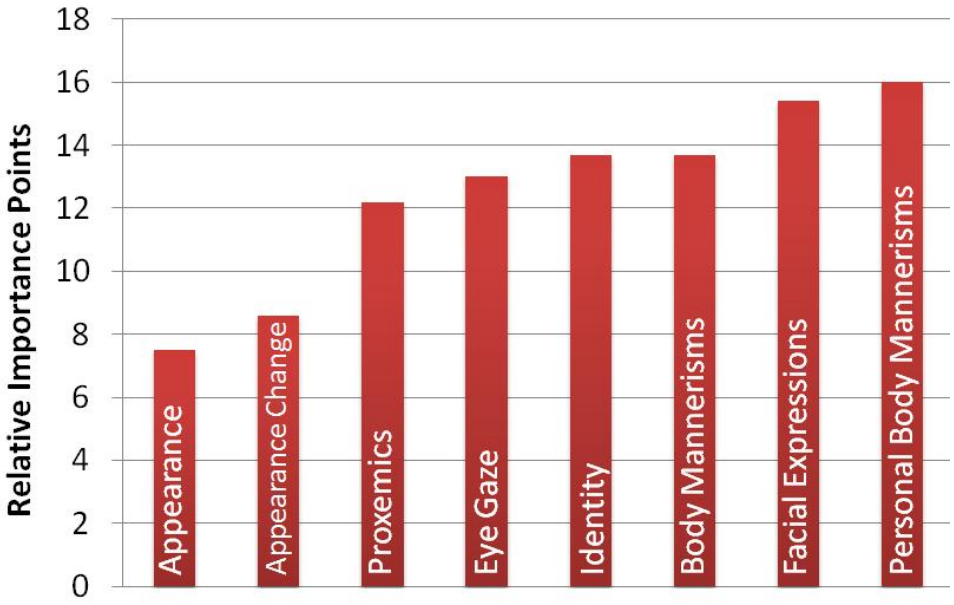


Figure 3.3: Rank average of the 8 questions

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. Please see Appendix A for the algorithm to determine the Rank Average. 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 3.3, 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 3.3 to Table 3.1, 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.

Chapter 4

Detecting Stereotypic Body Mannerisms

STEREOTYPIC behavior refers to any mannerism or utterance that is repetitive and non functional in nature [85] [86]. Stereotypy occurs in a large portion of the general population in various forms, although aggressive forms have been associated with sensory and cognitive disabilities in individuals. For example, individuals who are blind have the tendency to develop body rocking, head weaving, head drooping, and eye poking [87], while, individuals who are deaf have a propensity to develop various repetitive vocal behaviors [88][89]. Cognitive disabilities (both acquired, like brain injury and congenital, like autism and mental retardation) are associated with stereotypic behaviors like body rocking, hand flapping, jumping, and marching in place [90].

Though harmless by itself, Stereotypy can become a hindrance to social interactions and social acceptance [91]. Reference [84] introduces a 21 year old congenitally blind student who has an extreme case of body rocking (both while sitting and standing) that has become an obstruction to his career and an independent vocational evaluation states that a reduction in the student's body rocking was absolutely necessary for any form of employment. Stereotypy is a concerning problem in children, for whom peer acceptance is very important for their healthy growth and development of good social skills [92]. Children with stereotypic behaviors become victims of teasing thereby leading to social isolation, bullying and social segregation leading to negative self esteem. Aggravating these problems, social segregation and isolation have long term psychologically effects on the individual rendering an overall poor social skill set. Studies have shown that poor social skills are a leading cause for psychological problems such as depression, loneliness, and social anxiety [71].

Stereotypy, like any other human behavior, is very person specific. But socio-behavioral studies have shown that there are commonalities in these behaviors and there are broad classifications that can be identified in stereotypy prevalent in the general population. Eichel [93] introduces taxonomy for mannerisms that people with blindness and visual

impairment tend to display. He identifies that body rocking appears on top of the most commonly seen behavior stereotype. A review of literature [94] further supports the claim that body rocking and head related mannerisms, including head weaving and drooping, are distinctive behaviors exhibited by individuals who have sensory or cognitive disabilities. For example, [87] discusses the case of a blind student who has developed extreme body rocking stereotype. The student bends in a 30 degree arc when he is sitting and when standing, places a foot well ahead of him and bends forward in an even greater arc. Such stereotypes can hinder the interactions of these individuals with friends and family, eventually leading to isolation and social inadequacy in their personal and professional life.

4.1 Focus of the chapter

Having identified stereotypic body behaviors to be an important deterrent in social acceptance of individuals with cognitive or physical disabilities, we focus on the possibility of building a rehabilitative and/or assistive technology towards providing feedback to individuals about their stereotypic body behavior. Specifically, we focus on body rocking, as it tops the list of most widely seen stereotypic behaviors. To this end, our research aims to answer three important questions:

1. Is there any evidence of individuals responding to rehabilitation for reducing stereotypic body rocking behavior?
2. If yes, what is the state-of-the-art technology available to detect and notify individuals of their rocking behavior?
3. 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 answer the first question by looking into the immense literature available in behavioral psychology which has been studying behaviors in humans and their response to rehabilitation and assistance. To answer the second and third questions, we focus our attention towards wearable computing solutions that have gained a lot of momentum in the

recent past. In specific, we develop an argument for an inclusive framework that uses state-of-the art motion sensors with effective learning algorithms for detecting stereotype body rocking. As mentioned above, body rocking seems to be the most widely seen stereotype behavior and we use it as a basis for our argument that current level of technology can provide immense opportunities for developing rehabilitative and assistive technology solutions for reducing or controlling stereotypic behaviors.

4.2 Background and Related Work

Foundations for social rehabilitation of behavioral stereotypes

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.

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 [84] issued an annoying tone in the ears of the subject while [94] 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, [95], [95] and [96] 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.

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 [87] 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 [90] 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 [97] 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 [98]. According to [99], feedback is critical to social development because after an individual receives in-

formation 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 [72] [74], 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 [100]. 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.

Need for Assistive or Rehabilitative Technology

The feedback needed for intervention usually comes from people in and around these individuals who have stereotypic behavior. It has been observed that significant others in the environment often fail to give feedback, and even when they do, it is not meaningful or understandable to individuals who need rehabilitation - for example, in case of individuals who are blind or visually impaired, nodding one's head in reply to a question or gesturing [73] would be futile. Meaningful feedback is important, not only for social interaction, but for accurate self-evaluation by individuals. Most times people within the vicinity of individuals with needs fail to offer these crucial feedbacks. Many times, the individuals with needs feel guilty or obligated to ask for help from others in their environment. The ability to augment or replace this significant individual(s) in the environment with a reliable feedback mechanism is the aim and goal of all assistive technology solutions (In an independent online survey conducted by [101], the researchers found that people who are visually impaired would expressed the need for an assistive technology that would provide feedback on their own social mannerisms and offer a potential to improve their social outlook). Focusing on the development of such a technology that effectively detects body rocking and provides feedback to an individual is the goal of this paper. While we focus only on intervention through feedback, in the Future works section we highlight some ideas for extending the proposed framework into self-monitoring tools.

Transon [83] developed a head mounted switching device that would trigger a tone when an individual starts to rock. The device 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 such that the speaker aligns with the ears. 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 [73] and these studies support the above observation that short time feedback is most of the times 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. Thus, any technology developed for behavioral rehabilitation should be small and researchers should target the use over long durations in such a way that the feedback is slowly tapered off over a significantly longer duration of time.

Similar to the pervious experiment, [84] 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 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. Assessing these above technologies, we resort to two important design dimensions in every step of the building of our assistive device.

1. Size and placement of the device: We argue that any assistive device developed for the sake of improving social outlook of an individual should respect the appearance of a person in his/her social circle and should provide a solution that is discrete and non intrusive. We call this the Acceptance dimension.

2. Ability to discriminate rocking from other functional activities: False feedback even over a short period of time could be discouraging for an individual to continue using his/her assistive tool. It is imperative that the device be able to distinguish between the stereotypy from any other form functional activities effectively to keep the motivation of device use high. We call this the Motivation dimension.

The proposed methodology uses these two design dimensions while addressing the need of a new assistive technology.

4.3 Methodology

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 [102]. 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 4.1.

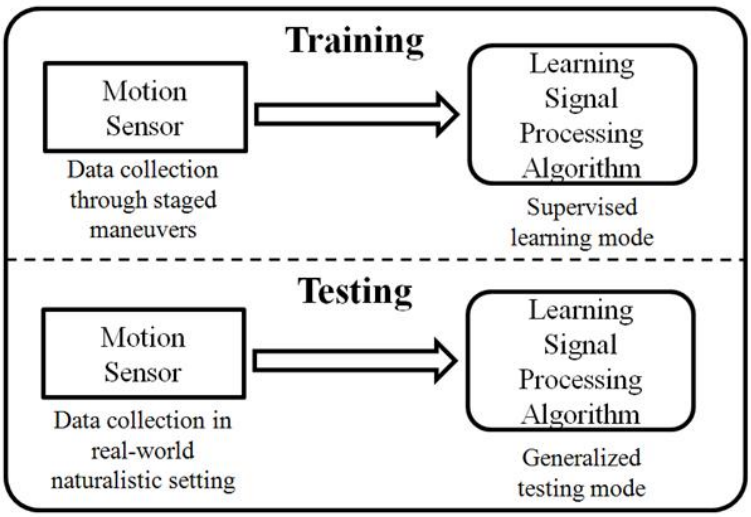


Figure 4.1: Training and testing phases of a typical learning framework found in literature.

Five bi-axial accelerometers are used in [102], 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 [103], 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) [104] with a subject independent accuracy of 64%. Since each dimension of the accelerometer data is similar to audio waveform, popular Hidden Markov Models [105] can be used to learn motion patterns. Reference [106] used HMM to learn the accelerometer data for specific tasks performed by participants and reports a recognition rate of over 90%. In [107], 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 [108] 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. [109] 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, [110] 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.

Based on all these findings, we hypothesize that an accelerometer based motion detector should be capable of capturing body rocking data and should be able to discriminate between rocking and other functional activities. We specifically chose the motion sensor and learning algorithm based on previous work done at our institute with the detection of seven simple body activities [111]. Researchers analyzed the performance of discriminative classifiers like AdaBoost, Support Vector Machines and RLogReg for recognizing these seven different activities and concluded that AdaBoost classifier offered the best recognition rate at 94%. Based on these results, in this paper, we extend the use of AdaBoost learning framework into body rock detection. We discuss the use of two AdaBoost classifiers - the classical AdaBoost [112] and the more recent Modest AdaBoost [113] for detecting and discriminating body rocking effectively. Our focus in the paper is directed towards understanding the generalization capabilities of the two AdaBoost learning models so that false positive rate is reduced while keeping the true positive rate high.

Motion Sensors - Design choice along the "Acceptance" Dimension

In order to keep the motion detector discrete, we have chosen state-of-the-art tri-axial accelerometer package, ZStar III [114], marketed by Freescale Semiconductor. The accelerometer is shown in the inset of Figure 4.2. 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 4.2 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

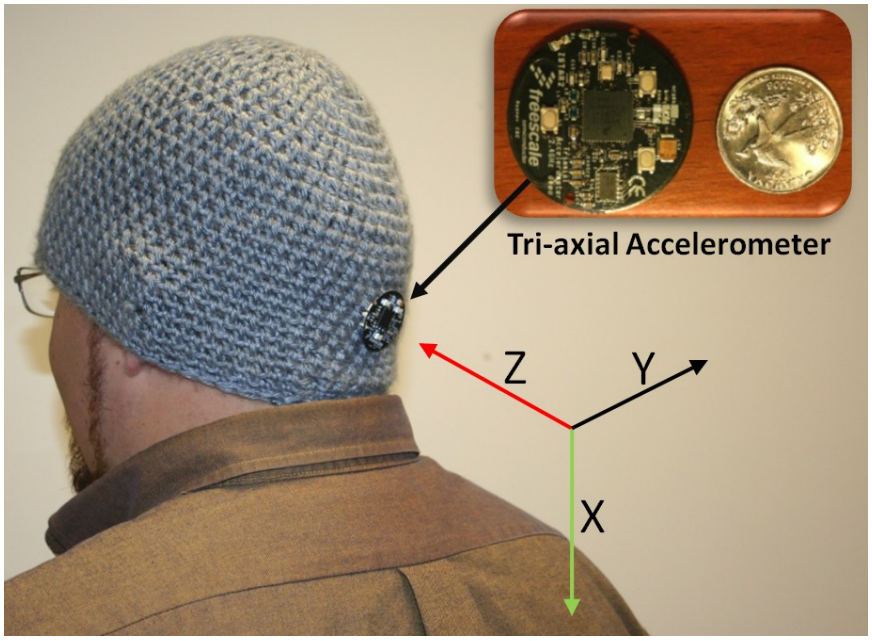


Figure 4.2: The proposed hardware for use in the detection of body rocking stereotypic behavior. The accelerometer, in comparison with a US quarter, is shown in the inset. The three axes marked in the image shows the orientation of the accelerometer as it is placed on the head.

of the device to the effect of an assistive technology thereby maintaining a longer duration feedback based rehabilitation regimen.

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 III-B of the proposed solution were placed on the PDA that could be carried by a user without any extra load. The proposed assistive technology is a planned addition on top of the Social Interaction Assistant proposed in [**?**]. 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

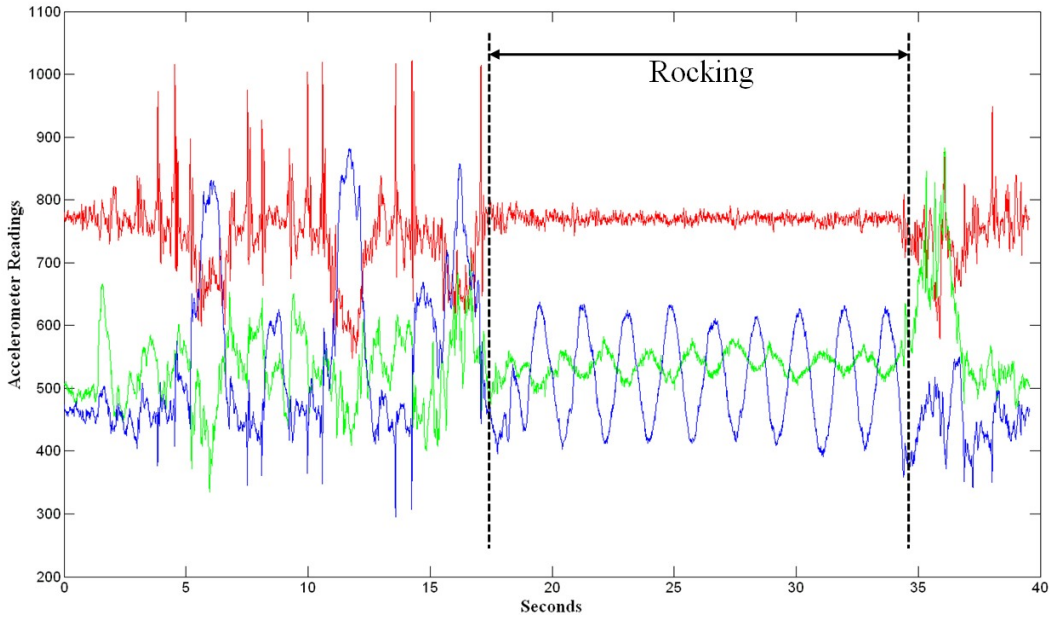


Figure 4.3: Data stream for the tri-axial accelerometer. The three streams correspond to the three axes. The figure shows non-rocking events followed by rocking and then followed by non-rocking.

only their rocking behavior but more complex self-monitoring routines that could allow the user to withdraw from the rocking behavior effectively.

Figure 4.3 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.

Extracting Body Rock Information from Motion Sensor Data - Design choice along the "Motivation" Dimension

As mentioned before, the work presented in this paper builds on top of the work presented in [111] 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 [36].

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.

Group 1 - Popular features used by the motion analysis research community [102] [111]: 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.

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.

Table 4.1: Features for Body Rock Detection: Group 1

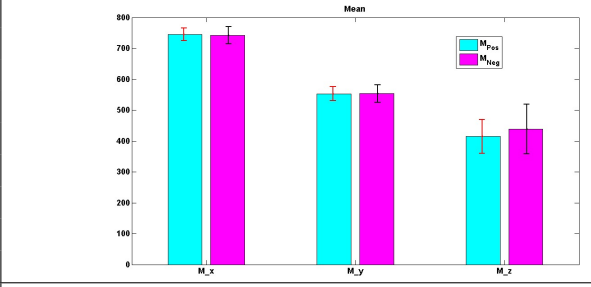
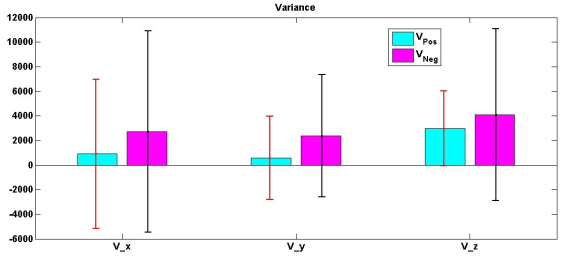
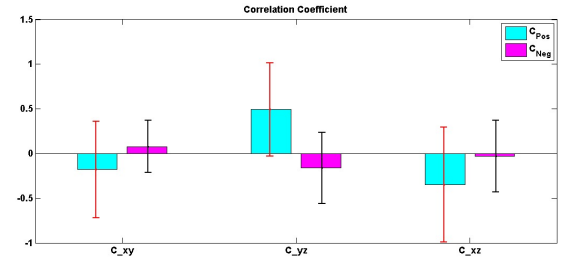
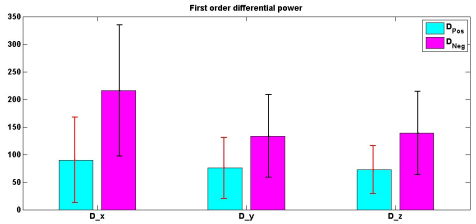
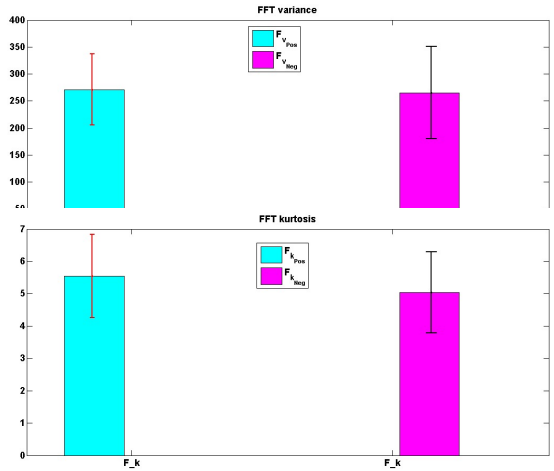
Group 1			
Set 1 **Definition:** Mean on the temporal dimension. **Axes affected:** x, y, z **Number of contributing features:** 3 **Feature Identification Numbers:** 1, 2, 3	$M_x = \frac{1}{N}\sum_{i=1}^{N} x_i$	1. M_x 2. M_y 3. M_z	
Set 2 **Definition:** Variance on the temporal dimension. **Axes affected:** x, y, z **Number of contributing features:** 3 **Feature Identification Numbers:** 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	
Set 3 **Definition:** Cross Correlation between axes. **Axes affected:** x, y, z **Number of contributing features:** 3 **Feature Identification Numbers:** 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}	

Table 4.2: Features for Body Rock Detection: Group 2

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

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.

Thus, the features used in our study can be categorized as belonging to two groups with three sets in Group 1and 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.

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.

(a) Classical 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 [115] 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 mod-

ule 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

$$L(x) = \text{sign}\left[\sum_{i=1}^{N} w_i f_i(x)\right] \tag{4.1}$$

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

$$Error_k = \sum_{i=1}^{M} e^{-y_i . L(x_i)} \tag{4.2}$$

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.

(a) Modest AdaBoost Learning Framework: 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 [113] 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. $P_m^{(+1)} = P_{(D_m)}(y = +1 \cap L(x))$: 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. $P_m^{(-1)} = P_{(D_m)}(y = -1 \cap L(x))$: 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. $\bar{P}_m^{(+1)} = P_{(\bar{D}_m)}(y = +1 \cap L(x))$: 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. $\bar{P}_m^{(-1)} = P_{(\bar{D}_m)}(y = -1 \cap L(x))$: 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

$$f_m = \left(P_m^{(+1)}(1 - \bar{P}_m^{(+1)}) - P_m^{(-1)}(1 - \bar{P}_m^{(-1)}) \right)(x) \tag{4.3}$$

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.

4.4 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.

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:

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.

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.

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.

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.

4.5 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 4.4) and Routine B (discussed under 4.4) were used as labeled positive (rocking) and negative (non-rocking) data for training the AdaBoost classifiers. Data collected under Routine C (discussed in Section**??**) 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 4.4. 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 4.1 and Table 4.2 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 VII, 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 [42].

4.6 Results

Figures 4.4 and Figure 4.5 shows the box plot [116] 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 4.4. 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 = third quartile - 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.

Table 4.6 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

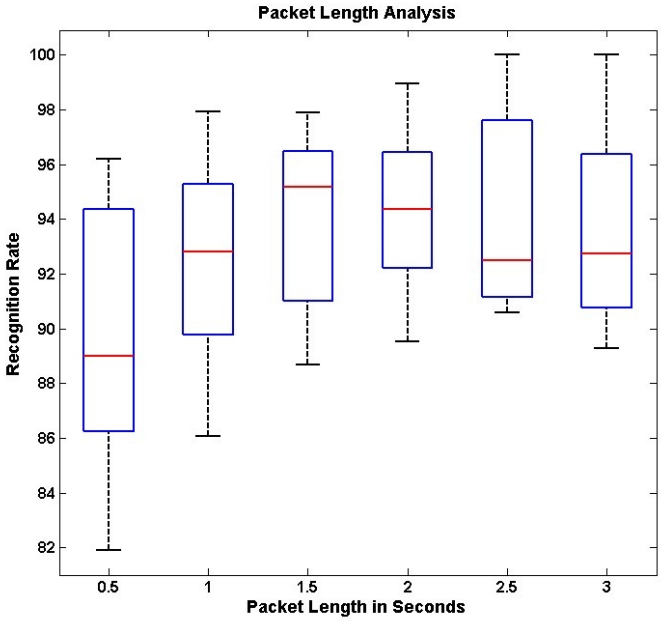


Figure 4.4: Packet length to recognition rate comparison under the classic AdaBoost framework.

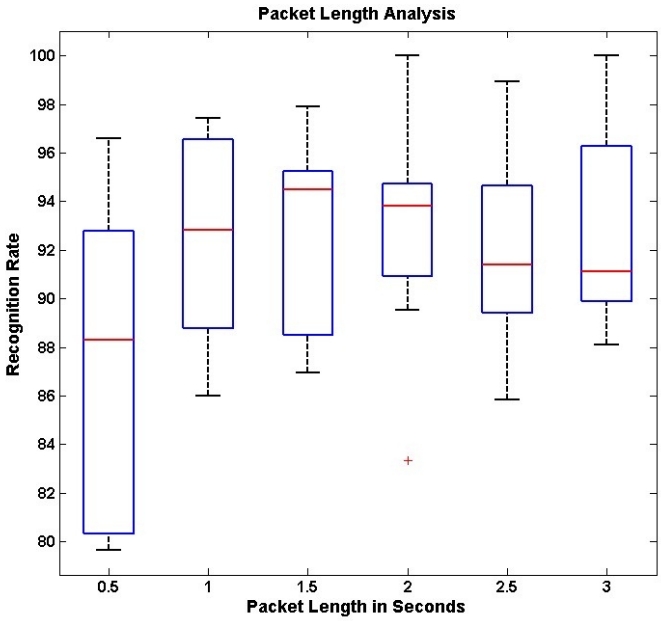


Figure 4.5: Packet length to recognition rate comparison under the Modest AdaBoost framework.

Generalization Capabilities	Classic AdaBoost	Modest AdaBoost
False Positives per Minute[1]	86	44
False Negatives per Minute[1]	20	9

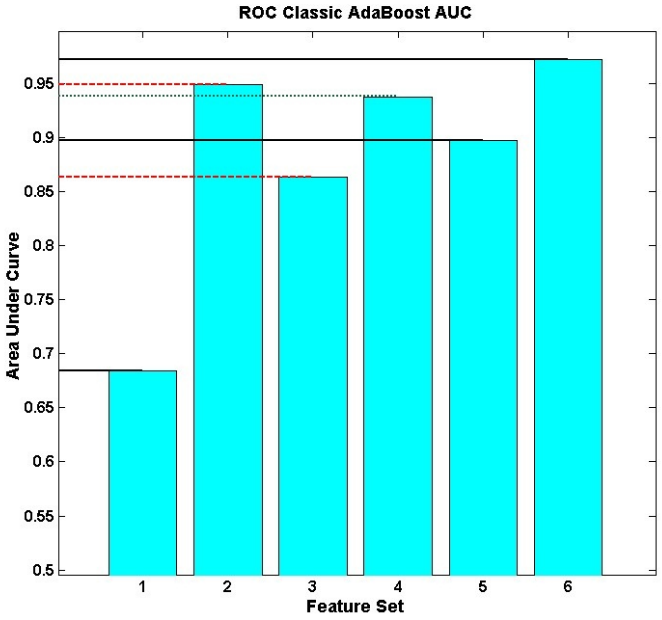
[1] Averaged over 10 minutes

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 trails 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.

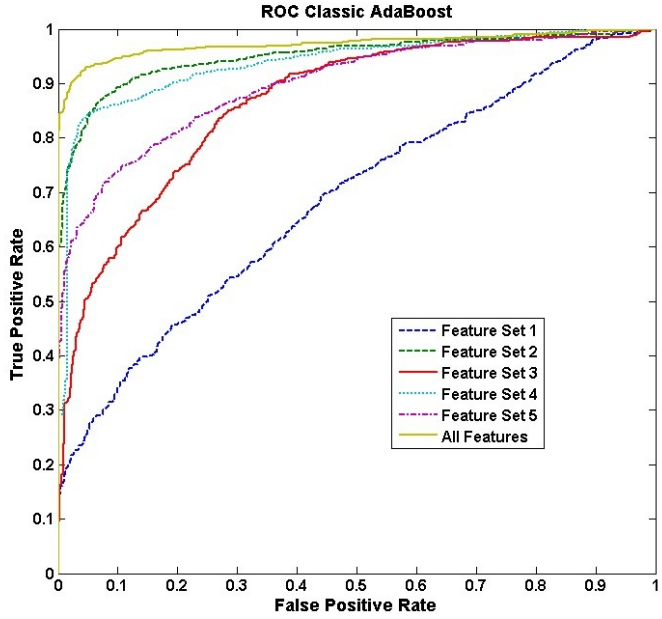
Figure 4.6 and Figure 4.7 shows the piecewise analysis of the classic AdaBoost and Modest AdaBoost frameworks. Subfigure (a) shows the performance of each feature set considered one at a time in detecting body rocking; feature set 6 corresponds to the use of all 14 features together. For example, column 1 in Figure 4.6(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) [117] 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 4.6(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.

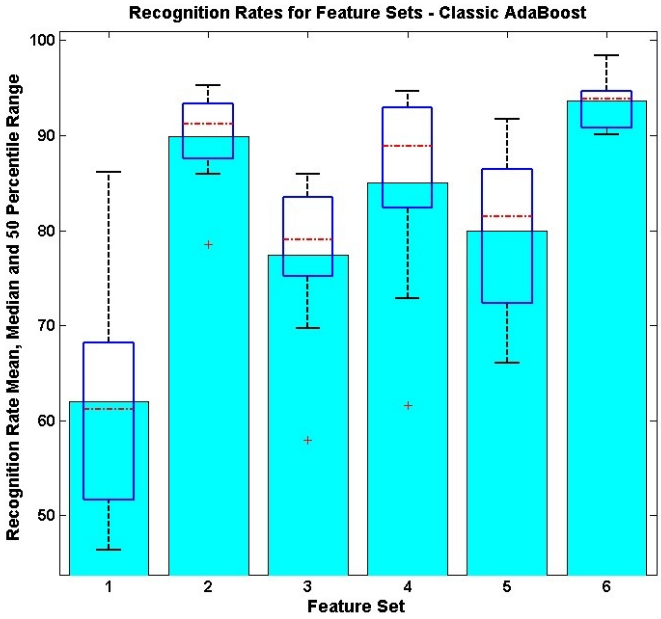
Subfigure (c) is a derivate of the ROC curves in subfigure (b). Each bar in the graph is representing the area under the corresponding curve (AUC) in (b). An area of 1 represents an ideal classifier with no false positive or false negatives, while an area of 0.5 represents


ROC Classic AdaBoost AUC
Area Under Curve
Feature Set
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1
2
3
4
5
6

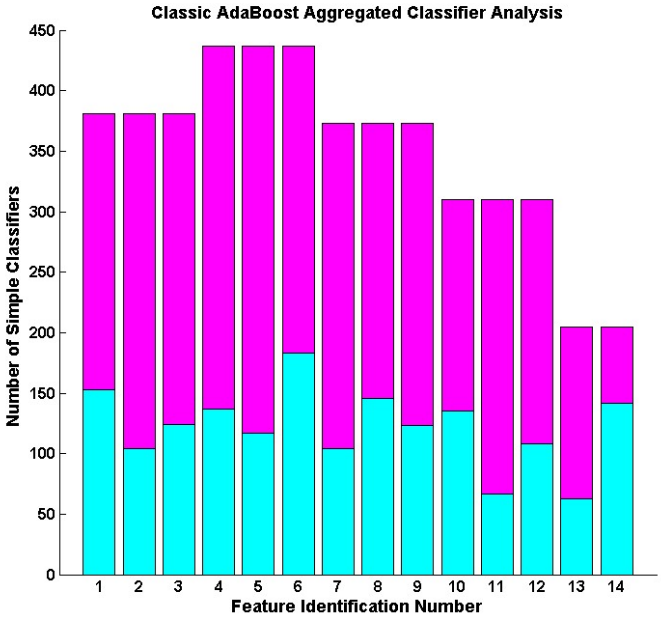
(a)


ROC Classic AdaBoost
True Positive Rate
False Positive Rate
Feature Set 1
Feature Set 2
Feature Set 3
Feature Set 4
Feature Set 5
All Features
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(b)

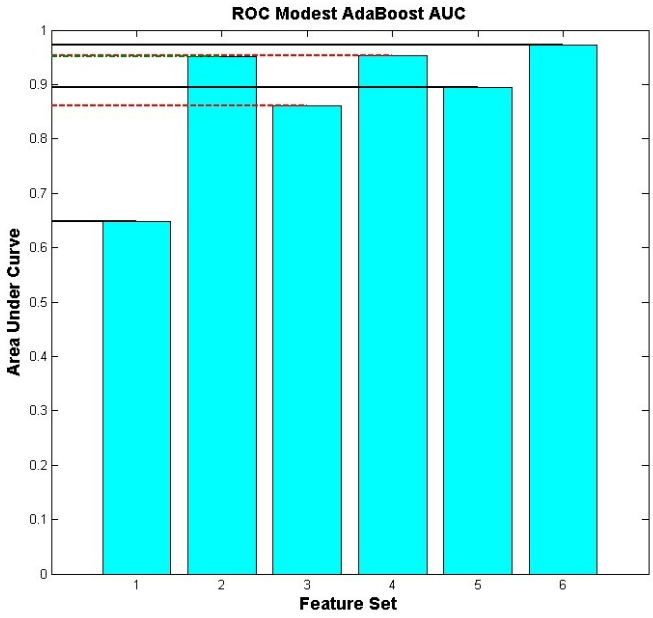


(c)

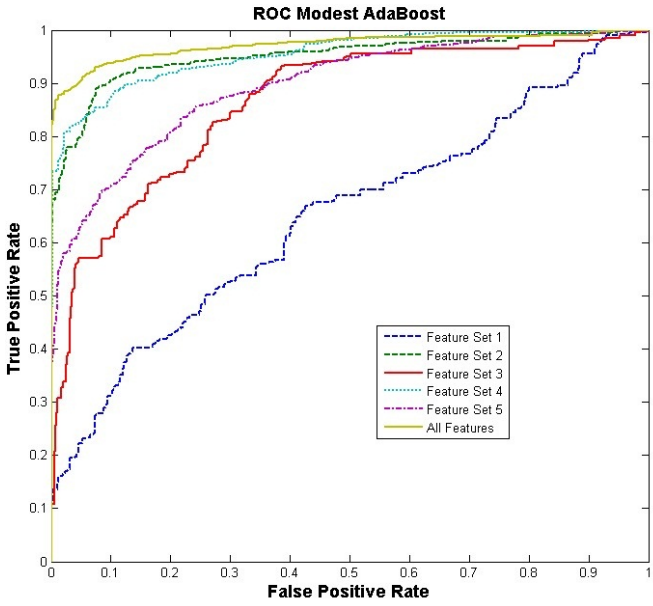


(d)

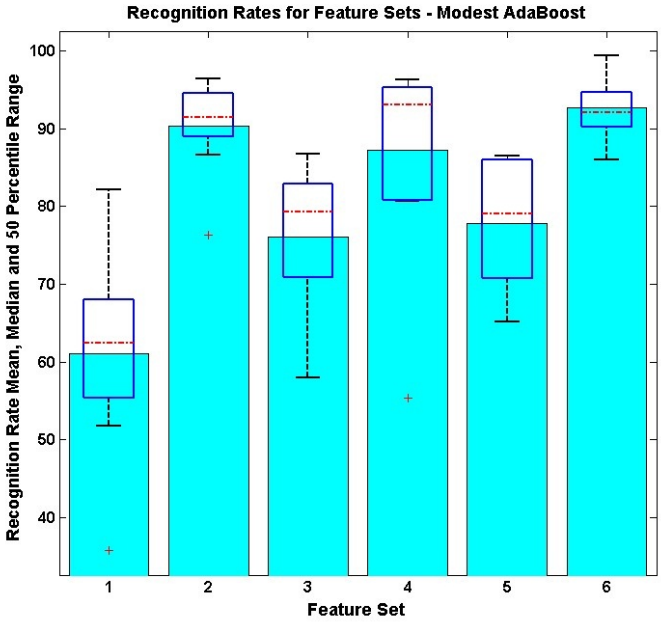
Figure 4.6: Piecewise performance analysis of the classic AdaBoost classifier framework; (a) Recognition rates under use of individual feature sets; (b) The Receiver Operating Characteristics (ROC) under the use of individual feature sets; (c) Area under the curve (AUC) for each feature set as estimated 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 Tables 4.1 and 4.2. Set 6 represents a set containing all 14 features from Tables 4.1 and 4.2.


ROC Modest AdaBoost AUC
Area Under Curve
Feature Set
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
2
3
4
5
6

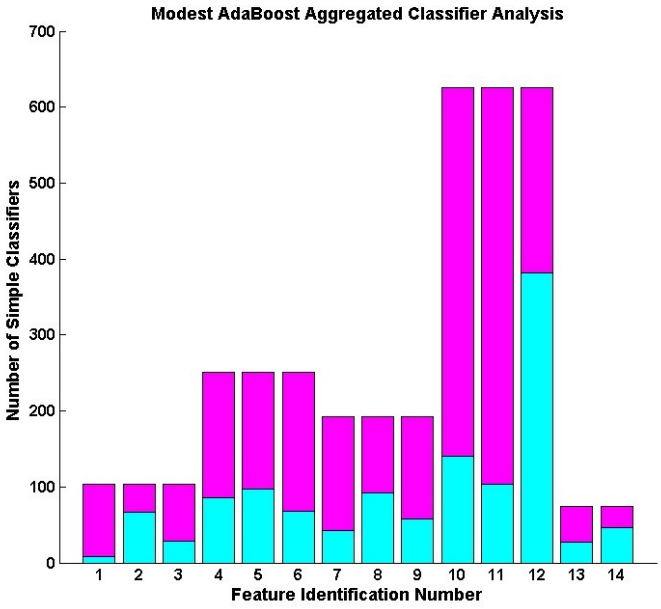
(a)


ROC Modest AdaBoost
True Positive Rate
False Positive Rate
Feature Set 1
Feature Set 2
Feature Set 3
Feature Set 4
Feature Set 5
All Features
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(b)



(c)



(d)

Figure 4.7: 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 Tables 4.1 and 4.2. Set 6 represents a set containing all 14 features from Tables 4.1 and 4.2.

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 form 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 4.7(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.

4.7 Discussion of Results

Before discussing the results of the experiments conducted on the accelerometer data, we step back to the first research question that we identified in Section 4.1, Is there any evidence of individuals responding to rehabilitation for reducing stereotypic body rocking behavior? From the Psychology background work presented in Section 4.2, we believe that there is enough evidence that individuals with sensory or cognitive impairment respond to rehabilitation through assistive devices. Specifically, the experiments highlighted in Section 4.2 support the claim that body rocking can be decreased by providing immediate feedback to the individual.

Regarding the second research question, what is the state-of-the-art technology available to detect and notify individuals of their rocking behavior? We 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. Our solution to this research question caters to the Acceptance design dimension that we identified in Section 4.2 1.

Focusing on the third research question, 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 4.6 to prove the efficiency of our proposed method in detecting body rocking and distinguishing it from other non-rocking behavior.

Packet Length, and Detection Efficiency

From Figure 4.4 and Figure 4.7, 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 [118]), 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. Combing 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

Generalization Capabilities

From Figure 4.4 and Figure 4.7, it is very difficult to distinguish any performance benefits between classic AdaBoost and Modest AdaBoost. But analyzing Table 4.6, 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 IV - A 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 4.4). 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 4.4) 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 4.6 (b) with Figure 4.7 (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 4.6 (d) with Figure 4.7 (d) it can be seen that Modest AdaBoost distributed it 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 4.1 Row 2 and Table 4.2 Row 1 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 4.7 (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 4.3, 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.

4.8 Conclusion

In this chapter, we have addressed the topic 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 [119] 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 [3], but we have not yet tested it for its feedback capabilities.

Chapter 5

Person-Specific Face Recognition

face recognition has the potential for recognizing people at a distance, without their knowledge or cooperation. For decades, banking, retail, commercial, and industrial buildings have been populated with surveillance cameras that capture video streams of all people passing through critical areas. More recently, as a result of threats to public safety, some public places (such as in Glasgow and London) have been heavily populated with video surveillance cameras. On average, a person moving through London is captured on video over 5 times a day. This offers an unprecedented basis for developing and testing face recognition as a biometric for security and surveillance.

Given this great potential, it is not surprising that many private corporations have attempted to develop and deploy face recognition systems, as an adjunct to existing video security and surveillance systems. However, the performance of these systems has been disappointing. Depending on how such a system is adjusted, miscreants might easily pass through the system undetected, or innocent people might be incessantly inconvenienced by false alarms.

One of the most difficult problems that face recognition researchers encounter in surveillance applications is that face databases of miscreants typically contain only frontal and profile views of each person's face, with no intermediate views. Surveillance videos captured of the same person with the same camera in the same lighting conditions might have face images that look quite different, due to pose angle variations, making it very difficult to compare captured face images to those in a database. Combine this problem with the fact that miscreants are highly motivated to disguise their identity, and the fact that face databases often contains thousands of faces, and the problem seems insurmountable.

Given all of these complicating factors, it is premature to rely upon face recognition systems for detecting miscreants in public places. On the other hand, the use of face recognition in controlled access applications (where users are highly motivated to cooper-

ate, and where face database images can be both captured and tested with the same camera under the same illumination conditions) is certainly within the limitations of current face recognition algorithms.

Employing face recognition to facilitate social interactions

However, there is a real-world application for face recognition that is moderately challenging, but still potentially within the realm of possibility. When people who are blind enter a room, they might find it awkward to initiate social interactions because they don't know how many people are in the room, who those people are, or where they are standing. A robust, wearable face recognition device could solve this problem.

This problem is simplified considerably by the fact that, on a day-to-day basis most people encounter a limited number of people whom they need to recognize. It is further simplified by the fact that people typically don't attempt to disguise their appearance in social situations. When a new person is encountered, the system could employ face detection to extract and save a sequence of face images captured during a conversation. This would provide a wide variety of facial expressions and pose angles, that could be stored in a database, and used for training a face recognition algorithm.

As people use such an assistive device over an extended period of time, they will learn both its abilities and its limitations. Conjectural information from the system can then be combined with the user's other sensory abilities (especially hearing) to jointly ascertain the identity of the person. This synergy between the user and the system relaxes some of the stringent requirements normally placed on face recognition systems.

However, such an assistive technology application still poses some significant challenges for researchers. One problem is the extreme variety of in lighting conditions encountered during normal daily activities. While there are standards for indoor office lighting that tend to provide diffuse and adequate lighting, lighting in other public places might vary considerably. For example, large windows can significantly alter lighting conditions, and incandescent lighting is much more yellow than florescent lighting. Outdoor lighting can

be quite harsh in full sunlight, and much more blue and diffuse in shadows. A person who is blind might not be aware of extreme lighting conditions, so the system would need to either (1) be tolerant of extreme variations or (2) recruit the user to ameliorate those extreme conditions.

In summary, the development of an assistive face recognition system for people who are blind provides a more tractable problem for face recognition researchers than security and surveillance applications. It imposes a somewhat less stringent set of requirements because (1) the number of people to be recognized is generally smaller, (2) facial disguise is not a serious concern, (3) multiple pose angles and facial expressions of a person can be captured as training images, and (4) the person recognition process can be a collaborative process between the system and the user.

In an attempt to provide such an assistive face recognition system, we have developed a new methodology for face recognition that detects and extracts unique features on a person's face, and then uses those features to recognize that person. Contrast this with conventional face recognition algorithms that might avoid the use of a few distinguishing features because that approach might make the system very vulnerable to disguise.

5.1 Face Recognition in Humans

For decades, scientists in various research areas have studied how humans recognize faces. Developmental psychologists have studied how human infants start to recognize faces, cognitive psychologists have studied how adolescents and adults perform face recognition; neuroscientists have studied the visual pathways and cortical regions used for recognizing faces, and neuropsychologists have attempted to integrate knowledge from neurobiological studies with face recognition research. Computer vision researchers are relatively new to this area, and have attempted to develop face recognition algorithms using image processing methods. Only recently have computer vision researchers been motivated to better understand the process by which humans recognize faces, in order to use that knowledge to develop robust computational models. Their new interest has lead to more inter-disciplinary

face recognition research, which will likely aid our understanding of face recognition.

New studies have shown that humans, to a large extent, rely on both the featural and configural information in face images to recognize faces [120]. Featural information provides details about the various facial features, such as the shape and size of the nose, the eyes, and the chin. Configural information defines the locations of the facial features, with respect to each other. Psychologists Vicki Bruce and Andrew Young [121] agree with this dual representation, saying that humans create a view-centric description of a human face by relying upon feature-by-feature perceptual input, which is then combined into a structural model of the face.

Sadar et al [122] showed that characteristic facial features are important for recognizing famous faces. For example, when they erased eye-brows from famous people's faces, face recognition by human participants was adversely affected. Young [123] showed that human participants were confused when asked to recognize faces that combined facial features from different famous faces. These studies suggest that the details of facial features are important in the recognition of faces.

However, [124] showed that the relative locations of the facial features was also very important for the recognition of faces. They collected face images of famous personalities, and then changed the aspect ratio of those images, such that the height was greatly compressed, while the width was emphasized. Surprisingly, all the resulting face images were still recognizable, despite their contorted appearance, as long as the relative locations of the features were maintained within the distorted image. This study suggests that humans can flexibly use the configural information when recognizing faces.

Another important area of research in the human perception of faces has been in understanding the medical condition of face blindness, called *prosopagnosia*. People with prosopagnosia are unable to recognize faces including their own. Until recently it was assumed that prosopagnosia was acquired often as a result of a localized stroke. However new evidence suggests that a substantial portion of the general population have a congenital form of prosopagnosia [125]. Kennerknecht et al [126] conducted a survey of 789 stu-

dents in 2006 which showed that 17 (2.5%) suffered from congenital prosopagnosia. These students went about their daily life without realizing their disorder in face recognition.

Other studies at the Perception research centers at Harvard and Univ College of London have shown that prosopagnosics recognize people using unique personal characteristics, such as hair style, gait, clothing, and voice. These findings suggest that the detection of unique personal characteristics might provide a basis for face recognition systems to better recognize people. Since current methods of face recognition have met with only limited success, it makes sense to explore the use of this alternative approach.

Research in Own-Race Bias (ORB) in face recognition [127] has also revealed some interesting results regarding human face recognition capabilities. David Turk et al. found that, when humans are presented with new objects or new faces, they initially learn to recognize those objects and faces based on their distinctive features. Then, as familiarity increases, they incorporate configural information, moving towards holistic recognition. This study suggests that distinctive features are important during the initial stages of face recognition, and that configural information subsequently provides additional useful information.

Distinctive facial features can take many different forms. For example, after a first encounter with a person who has a handlebar moustache, we readily recognize that person by the presence of his distinctive feature. Similarly, a person with a large black mole on her face will be remembered by first-time acquaintances by that feature. Given the current limited understanding of how humans recognize faces, it makes sense to use these observations as the basis for a new approach to face recognition.

The research described in this chapter is based on the approach of identifying distinctive facial features that can be used to distinguish each person's face from other faces in a face database. In recognition of the role played by configural information in the later stages of face recognition, it also takes into account the location of these features with respect to each other. The results of our research suggest that this approach can be very effective for distinguishing one person's face from other faces.

5.2 Our Approach to Face Recognition

Having introduced the potential for using characteristic person-specific features for face recognition, we now turn our attention towards the development of a method for discovering such features, and for using them to index face images. Then we propose a novel methodology for face recognition, using person-specific feature extraction and representation. For each person in a face database, a learning algorithm discovers a set of distinguishing features (each feature consisting of a unique local image characteristic, and a corresponding face location) that are unique to that person. This set of characteristic facial features can then be compared to the normalized face image of any person, to determine the presence or absence of those features. Because a unique set of features is used to identify each person in the database, this method effectively employs a different feature space for each person, unlike other face recognition algorithms that assign all of the face images in the database to a locality in a shared feature space. Face recognition is then accomplished by a sequence of steps, in which query face images is mapped into a locality within the feature space of each person in the database, and its position is compared to the cluster of points in that space that represents that person. The feature space in which the query face images are closest to the cluster is used to identify the query face images.

Having introduced the conceptual theory behind a person-specific characteristic feature extraction approach to face recognition, we now propose in the subsequent sections a method for detecting and extracting such features from face images, and for constructing a feature space that is unique to each person in the database.

5.3 Feature Extractors

What is a Feature?

The task of face recognition is inherently a multi-class classification problem. For every face image X, there is an associated label y that is the name of the class, i.e. the name of the person depicted in the image. While X represents the image of the person, there is no inherent constraint on whether the image is a color RGB, HUV or YCbCr image, or a

gray-scale image with a gray-scale range of 0 to 255, or even spectral representation that is extracted from the face image using Fourier transform or Wavelets. Irrespective of the image representation, the basis vectors spanning that representation are called features. The feature space spanned by these basis vectors is partitioned by the decision boundaries that ultimately define the different classes in the multi-class problem of face recognition. In this work, we choose a particular set Gabor filters as feature detectors, and each of those feature detectors for each person in the database, and that set of Gabor filters spans a unique feature space for that person.

Gabor Features

Gabor filters are a family of functions (sometimes called Gabor Wavelets) that are derived from a mother kernel (a Gabor Function) by varying the parameters of the kernel. As with any wavelet filters, the Gabor filters extract local spatial frequency content from the underlying image. Gabor Filters specifically capture the spatial location and spatial orientation of the intensity variations in the image underneath the filter's location. By varying the spatial frequency and the spatial scope of the filters, it is possible to extract a Gabor coefficient that partially describes the nature of the image underneath it. The coefficients obtained by filtering a locality in a face image with a set of different Gabor Filters are called Gabor Features.

Use of Gabor Filters in Face Recognition

Gabor filters have been widely used to represent the receptive field sensitivity of simple cell feature detectors in the human primary visual cortex. Recognizing this fact, Gabor features have been widely used by face recognition researchers. Over the last few years, the extensive use of Gabor wavelets as generators of feature spaces for face recognition, has led to objective studies of the strength of Gabor features for this application. For example, Shan et al [Shan2004] reviewed the strength of Gabor features for face recognition using an evaluation method that combined both alignment precision and recognition accuracy. Their experiments confirmed that Gabor features are robust to image variations caused by

the imprecision of facial feature localization. As indicated by Gkberk et al [128], several studies have concentrated on examining the importance of the Gabor kernel parameters for face analysis. These include: the weighting of Gabor kernel-based features using the simplex algorithm for face recognition [129], the extraction of facial subgraphs for head pose estimation [130], the analysis of Gabor kernels using univariate statistical techniques for discriminative region finding [131], the weighting of elastic graph nodes using quadratic optimization for authentication [132], the use of principal component analysis (PCA) to determine the importance of Gabor features [133], boosting Gabor features [134] and Gabor frequency/orientation selection using genetic algorithms [135].

A relevant work on Gabor Filters for face recognition that is closely related to the research presented here is by Wiskott and von der Malsburg [136]. Their work [137] [138] [139] [140], [136] proposes a framework for face recognition that is based on modeling human face images as labeled graph. Termed *Elastic Bunch Graph Matching* (EBGM), the technique has become a cornerstone in face recognition research. Each node of the graph is represented by a group of Gabor filters/wavelets (called "jets") which are used to model the intensity variations around their locations. The edges of the graph are used to model the relative location of the various jets. Since the jets represent the underlying image characteristics, it is desirable to place them on fiducial points on the face. This is achieved by *manually* marking the locations of the facial fiducial points using a small set of controlled graphs that represent "general face knowledge", which represents an average geometry for the human face. In our work, a genetic algorithm is used to obtain the spatial location of the fiducial points. Besides automating the process of locating these points, our work identifies spatial locations on the face image that are unique to every single person, rather than relying on an average geometry.

Closely following the work of Wiskott et. al., Lyons et. al. [141] proposed a technique that uses Gabor Filter coefficients extracted at 1) automatically located rectangular grid points or 2) manually selected image feature points. These coefficients are then used to bin face images based on sex, race and expression. The technique relies on a combined

Principal Component Analysis (PCA) dimensionality reduction and Linear Discriminant Analysis (LDA) classification over the extracted Gabor coefficients, to achieve a pooling of images. While the classification task is not related directly to *identifying* individuals from face images, this technique also demonstrates the ability of Gabor Filters to extract features that can encode subtle variations on facial images, providing a basis for face identification.

Gabor Filters

Mathematically, Gabor Filters can be defined as follows:

$$\Psi_{\omega,\theta}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \cdot G_\theta(x,y) \cdot S_{\omega,\theta}(x,y) \tag{5.1}$$

$$G_\theta(x,y) = \exp\left\{-\left(\frac{(x\cos\theta + y\sin\theta)^2}{2\sigma_x^2} + \frac{(-x\sin\theta + y\cos\theta)^2}{2\sigma_y^2}\right)\right\} \tag{5.2}$$

$$S_{\omega,\theta}(x,y) = \left[\exp\{i(\omega x\cos\theta + \omega y\sin\theta)\} - \exp\left\{-\frac{\omega^2\sigma^2}{2}\right\}\right] \tag{5.3}$$

where,

- $G_\theta(x,y)$ represents a Gaussian Function.
- $S_{\omega,\theta}(x,y)$ represents a Sinusoid Function.
- (x,y) is the spatial location where the filter is centered with respect to the image axis.
- ω is the frequency parameter of a 2D Sinusoid.
- σ_{dir}^2 represents the variance of the Gaussian (and thus the filter) along the specified direction. *dir* can either be x or y. The variance controls the region around the center where the filter has influence.

From the definition of Gabor filters, as given in Equation 5.1, it is seen that the filters are generated by multiplying two components: a Gaussian Function $G_\theta(x,y)$ (Equation 5.2) and a Sinusoid $S_{\omega,\theta}(x,y)$ (Equation 5.3). The following discussions detail the two components of Equation 5.1.

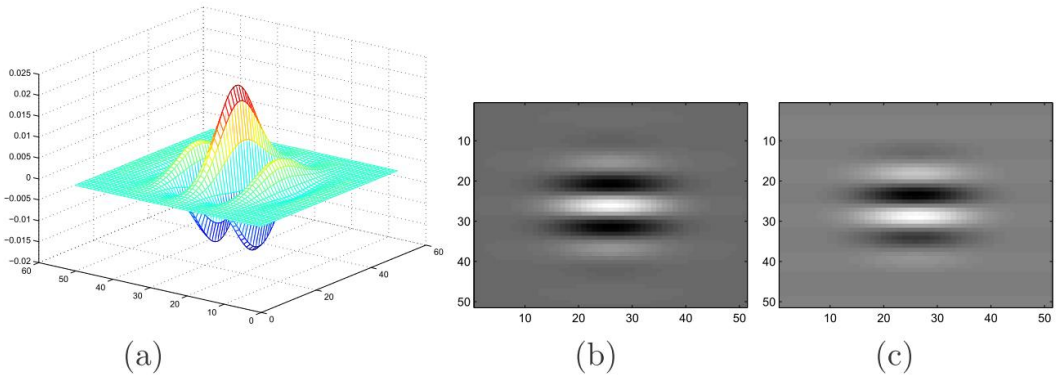


Figure 5.1: (a) 3D representation of a Gaussian mask; $\sigma_x = 10$, $\sigma_y = 15$ and $\theta = 0$
(b)Image of the Gaussian mask $\sigma_x = 10$, $\sigma_y = 15$ and $\theta = 0$

Gaussian Function

The 2D Gaussian function defines the spatial spread of the Gabor filter. This spread is defined by the variance parameters of the Gaussian, along the x and y direction together with the orientation parameter θ. Figure 5.1(a) shows a 3D representation of the Gaussian mask generated with $\sigma_x = 10$ and $\sigma_y = 15$ and rotation angle $\theta = 0$. The image in Figure 5.1(b) shows the region of spatial influence of an elliptical mask on an image, where the variance in the x direction is larger than the variance in the y direction.

Typically the Gaussian filter has the same variance along both the x and y directions, that is $\sigma_x = \sigma_y = \sigma$. Under such conditions the rotation parameter θ does not play any role as the spread will be circular.

Sinusoid

The 2D complex Sinusoid defined by Equation 5.3 generates the two Sinusoidal components of the Gabor filters which (when applied to an image) extracts the local frequency content of the intensity variations in the signal. The complex Sinusoid has two components (the real and the imaginary parts) which are two 2D sinusoids that are phase shifted by $\frac{\pi}{2}$ radians. Figure 5.2(a) shows the 3D representation of a Sinusoidal signal (either real or

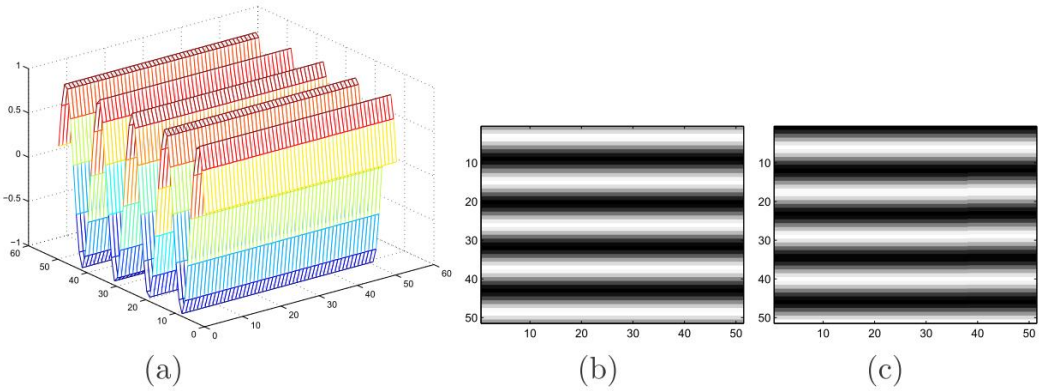


Figure 5.2: (a)3D representation of a Sinusoid $S_{\omega,\theta}$
(b)Image representation of the real part of the complex Sinusoid $\Re\{S_{\omega,\theta}\}$
(c)Image representation of the imaginary part of complex Sinusoid $\Im\{S_{\omega,\theta}\}$

imaginary) at $\omega = 0.554$ radians and $\theta = 0$ radians, while Figure 5.2(b) and 5.2(c) show an image of the real and imaginary parts of the same complex Sinusoid, respectively. It can be seen that the two filters are similar, except for the π radian phase shift.

Multiplying the Gaussian and the sinusoid generates the complex Gabor filter, as defined in Equation 5.1. If $\sigma_x = \sigma_y = \sigma$, then the real and imaginary parts of this complex filter can be described as follows.

$$\Re\{\Psi_{\omega,\theta}(x,y)\} = \frac{1}{2\pi\sigma^2} \cdot G_\theta(x,y) \cdot \Re\{S_{\omega,\theta}(x,y)\} \tag{5.4}$$

$$\Im\{\Psi_{\omega,\theta}(x,y)\} = \frac{1}{2\pi\sigma^2} \cdot G_\theta(x,y) \cdot \Im\{S_{\omega,\theta}(x,y)\} \tag{5.5}$$

Figure 5.3(a) shows the 3D representation of a Gabor filter (either real or imaginary) at $\omega = 0.554$ radians, $\theta = 0$ radians, and $\sigma = 10$ and Figure 5.3(b) and 5.3(c) show an image with the real and imaginary parts of the complex filter.

In order to extract a Gabor feature at a location (x,y) of an image I, the real and imaginary parts of the filter are applied separately to the same location in the image, and a magnitude is computed from the two results. Thus, the Gabor filter coefficient at a location

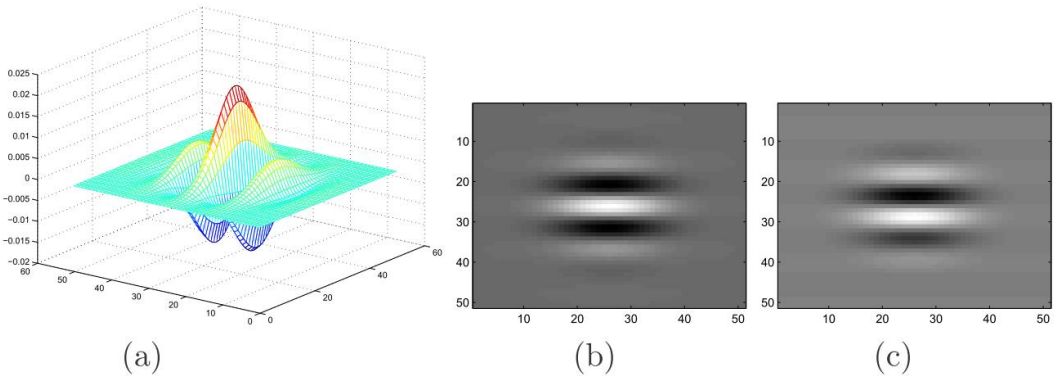


Figure 5.3: (a)3D representation of a Gabor filter $\Psi_{\omega,\theta}$
(b)Image representation of the real part of Gabor filter $\Re\{\Psi_{\omega,\theta}\}$
(c)Image representation of the imaginary part of Gabor filter $\Im\{\Psi_{\omega,\theta}\}$

(x,y) in an image I with a Gabor filter $\Psi_{\omega,\theta}$ is given by

$$C_{\Psi}(x,y) = \sqrt{(I(x,y) * \Re\{\Psi_{\omega,\theta}(x,y)\})^2 + (I(x,y) * \Im\{\Psi_{\omega,\theta}(x,y)\})^2} \tag{5.6}$$

In our experiments, a *Gabor filter bank* was created by varying three parameters of $\Psi_{\omega,\theta}$: (1) the frequency parameter ω, (2) the orientation parameter θ, and (3) the variance parameter σ. We chose five values for each of these parameters thereby generating 125 different Gabor filters.

- $\omega = \left(2^{(-f+2)/2} \cdot \pi\right)$ where, $f = \{0,1,2,3,4\}$
- $\theta = \left(\frac{\pi}{2} \cdot \frac{1}{5} \cdot t\right)$ where, $t = \{0,1.25,2.5,3.75,5\}$
- $\sigma = \{5,10,15,20,25\}$

5.4 The Learning Algorithm

The proposed method uses the above described Gabor filters to find distinguishing features (and corresponding feature locations) within a face image. That is, for each person in the database, the algorithm finds a set of Gabor filters which, when applied at their corresponding (x,y) locations within the image will produce coefficients that are unique for that individual. This means that all of the 125 Gabor filters in the filter bank are applied at each and

every location of each of the individual's face images, and then tested for their ability to distinguish every individual. Given a 128×128 face image, there will be $128 \times 128 \times 125 \times n$ filter coefficients that will be generated per face image per person, where n is the number of characteristic features to be extracted for each person. This must be computed for every person in the training set, which further increases the search space. To search such a vast space of parameter values (the size of the Gaussian mask, the frequency of the complex sinusoid, the orientation of the entire Gabor filter, and the (x,y) location where the filter is placed) it is important that some scheme for effective search be incorporated into the system. To this end, we have chosen Genetic Algorithms to conduct the search. For each person in the training set, all of the face images that depict to that person are indexed as positives, while all of the other face images in the database are indexed as negatives. Dedicated Genetic Algorithm based search is conducted with these positive and negative images, with the aim of finding a set of Gabor filters and filter locations that distinguish all the positives from the negatives.

Genetic Algorithms

When the parameter space is vast (as it is in our case) a Genetic Algorithm (GA) searches for the optimum solution by randomly picking parameter sets and evolving newer ones from the best performers. This happens over many generations, hopefully resulting in the optimum set of parameters. To start the search, the GA generates a random set of *parents*. Each parent is characterized by the presence of a *chromosome*. The chromosome internally encodes all the parameters that are used by the parent to perform the intended operation. In our case, the intended operation is face recognition. The parent uses the parameters that are found in its chromosome to derive the Gabor features on the positive and negative images.

Based on the ability of these features to distinguish a face from all others in the database, the parent is ranked within its population. This rank is also referred to as the *fitness of the parent*. The ranking of all the parents, based on their fitness, marks the end of a generation, and a new generation needs to be created. New generations are formed based on three important aspects of GAs, *Retention*, *Cross Over* and *Mutation*. A portion

of the newer generation is derived from the older generation, using the above mentioned methods, and the rest of the new generation is created randomly, maintaining the same overall number of parents between generations. Once a new population has been formed, the process of ranking parents occurs (as explained earlier) and a new generation is born out of that ranking. This iterative process continues until the parents in a certain generation are fit enough to achieve the given task (with the desired amount of success) or until the desired number of generations have evolved.

Use of Genetic Algorithms in Face Recognition

GAs have been used in face recognition to search for optimal sets of features from a pool of potentially useful features that have been extracted from the face images. Liu et al [142] used a GA along with Kernel Principal Component Analysis (KPCA) for face recognition. In their approach, KPCA was first used to extract facial image features. After feature extraction using the KPCA, GAs were employed to select the optimal feature subset for recognition - or more precisely the optimal non-linear components. Xu et al [143] used GAs along with Independent Component Analysis to recognize faces. After obtaining all the independent components using the Fast ICA algorithm, a genetic algorithm was introduced to select optimal independent components.

Wong and Lam [144] proposed an approach for reliable face detection using genetic algorithms with eigenfaces. After histogram normalization of face images and computation of eigenfaces, the 'k' most significant eigenfaces were selected for the computation of the fitness function. The fitness function was based on the distance between the projection of a test image and that of the training-set face images. Since GAs are computationally intensive, the search space for possible face regions was limited to possible eye regions alone.

Karungaru et al [145] performed face recognition using template matching. Template matching was performed using a genetic algorithm to automatically test several positions around the target, and to adjust the size of the template as the matching process

progressed. The template was a symmetrical T-shaped region between the eyes, which covered the eyes, nose and mouth.

Ozkan [146] used genetic algorithms for feature selection in face recognition. In this work, the Scale Invariant Feature Transform (SIFT) [147] was used to extract features. Since SIFT was originally designed for object recognition in general, genetic algorithms were used to identify SIFT features, which are more suitable to face recognition.

Huang and Weschler [148] developed an approach to identify eye location in face images using navigational routines, which were automated by learning and evolution using genetic algorithms. Specifically, eye localization was divided into two steps: (i) the derivation of the saliency attention map, and (ii) the possible classification of salient locations as eye regions. The saliency map was derived using a consensus between navigation routines that were encoded as finite state automata (FSA) exploring the facial landscape and evolved using genetic algorithms (GAs). The classification stage was concerned with the optimal selection of features and the derivation of decision trees for confirmation of eye classification using genetic algorithms.

Sun and Yin [149] applied genetic algorithms for feature selection in 3D face recognition. An individual face model was created from a generic model and two views of a face. Genetic algorithms were used to select optimal features from a feature space composed of geometrical structures, the labeled curvature types of each vertex in the individualized 3D model.

Sun et al [150] approached the problem of gender classification using a genetic algorithm to select features. A genetic algorithm was used to select a subset of features from a low-dimensional representation, which was obtained by applying PCA and removing eigenvectors that did not seem to encode information about gender.

As is evident from these citations, many feature-based approaches towards face recognition use genetic algorithms for feature selection. However, these approaches employ a single feature space derived from a set of face images. We believe that it is more

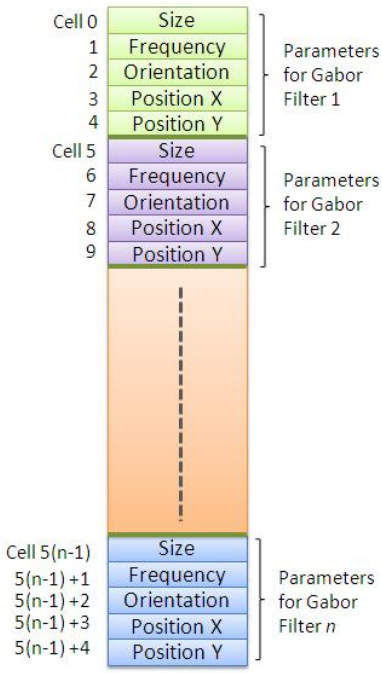


Figure 5.4: A typical chromosome used in the proposed method.

effective to employ aimed at extracting person-specific features, and that an effective way to do this is by using genetic algorithms. As observed by [127], humans initially learn to recognize faces based on person-specific characteristic features. This suggests that better recognition performance might be achieved by representing each person's face in a person-specific feature space that is learned using GAs.

The following paragraphs describe how we employed GAs to solve the problem of finding person-specific Gabor features aimed at face recognition.

The Chromosome

Each parent per generation encodes the parameters of a set of Gabor filters in the form of a chromosome. In our implementation, each Gabor filter is represented by five parameters. If there are n Gabor filters, parameters for all of these filters are encoded into the chromosome in a serial manner, as shown in Figure 5.4. Thus the length of the chromosome is $5n$. The number of Gabor filters being used per face image determines the length of the

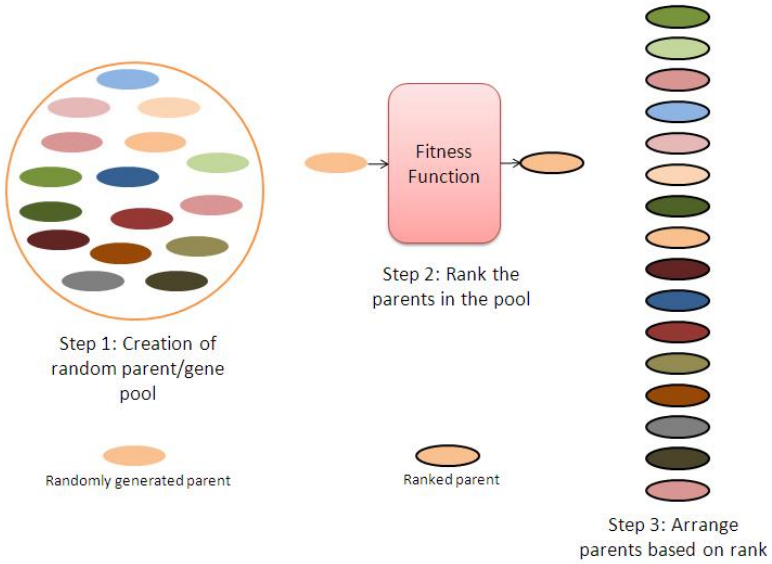


Figure 5.5: Stages in the creation of the first generation of parents

chromosome. As shown in Figure 5.4, each parameter in the chromosome is encoded as a gene. The boundaries of these genes defines the regions where the chromosome undergoes both the crossover and mutation. The genes can be considered as the primary element of the parent responsible in the evolution.

Creation of the first generation

Figure 5.5 depicts the first generation of parents, which are created randomly. Each parent's chromosome is filled randomly with parameter values where, each parameter value is within the allowed range for that parameter. Thus, in our experiment, each parent potentially has the parameters needed for it to perform face recognition using Gabor filters for feature extraction.

Once these parents are created, each parent in the gene pool is evaluated based on its capacity to perform face recognition. To this end, a fitness function is defined, which takes into account the ability of each parent to distinguish an individual from all others based on the most distinguishing features on the individual's face.

This fitness function also takes into account the similarity of the extracted features,

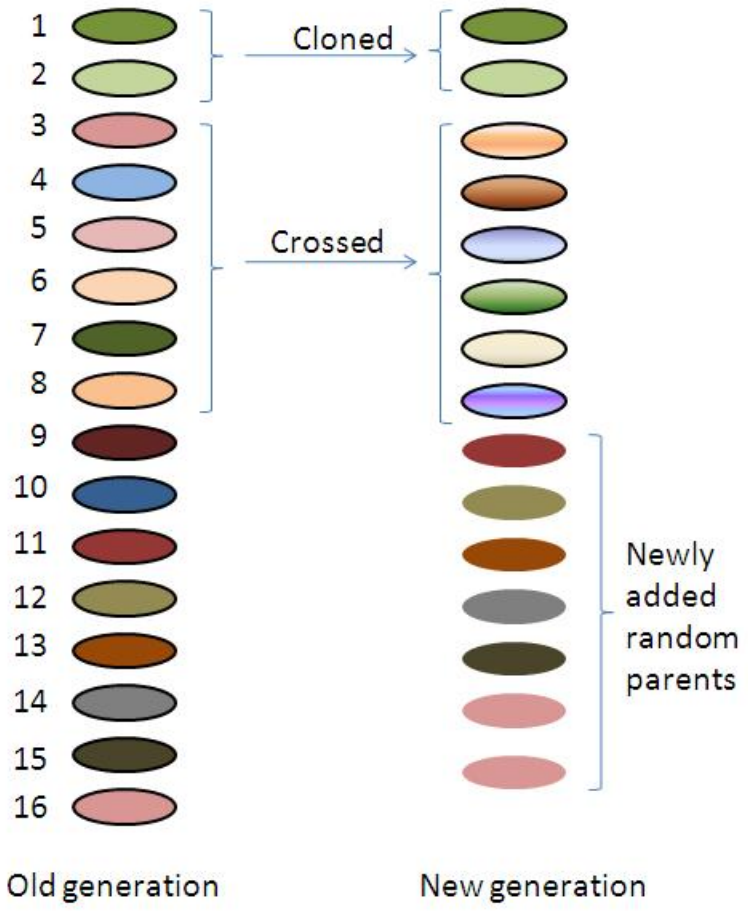


Figure 5.6: Deriving newer parents from the current generation

and discourages the selection of features that are highly correlated with each other. This ensures that the face images will be searched for multiple distinguishing characteristics. Subsection 5.5 explains in detail the fitness function used in our experiments. The parents with the best fitness are ranked higher, and have the highest probability of being picked for using genetics the next generation. At the end of the rank ordering process, the parents are arranged in a descending order, based on their fitness. This rank ordering determines the probability of each parent being used to create the subsequent generation. If a parent has a higher fitness, it will have a higher probability of being cloned into the next generation, or of otherwise being involved in reproduction.

Creation of the newer generations

The newer generations are created from the older population using *clones*, *mutants*, and *crossovers* of the fittest parents. To better search for the optimal parameter set, new random parents are created every generation. This reduces the likelihood that the algorithm will get stuck in a local minimum in the search space.

Figure 5.6 shows crossover creates a newer generation, using the fittest parents from the older generation.

The number of offsprings created from mutation, cloning, and crossover are determined by parameters of the Genetic algorithm. The number of clones, mutants, and corssovers are controlled by the following parameters:

1. *Cloning Rate* This parameter controls the number of parents from the previous generation that will be retained without undergoing any changes in their genetic structure.

2. *Crossover Rate* This parameter controls the number of offsprings that will be born from crossing the parents from the previous generation.

3. *Mutation Rate* This parameter determines how many of the crossed offsprings will then be mutated.

4. *Cloning Distribution Variance* After determining the number of offsprings be to cloned, the index of the parents for cloning are chosen using a normal distribution random number generator, with the mean zero and variance equal to this parameter. Since the parents from the previous generation have been rank ordered in descending order of fitness, the zeroth parent will be the top performer (which coincides with the mean of the random number generator, and has the highest probability of getting picked).

5. *Crossover Distribution Variance* This parameter (which is similar to the Cloning Distribution Variance) is used to choose the index of the parents who will undergo

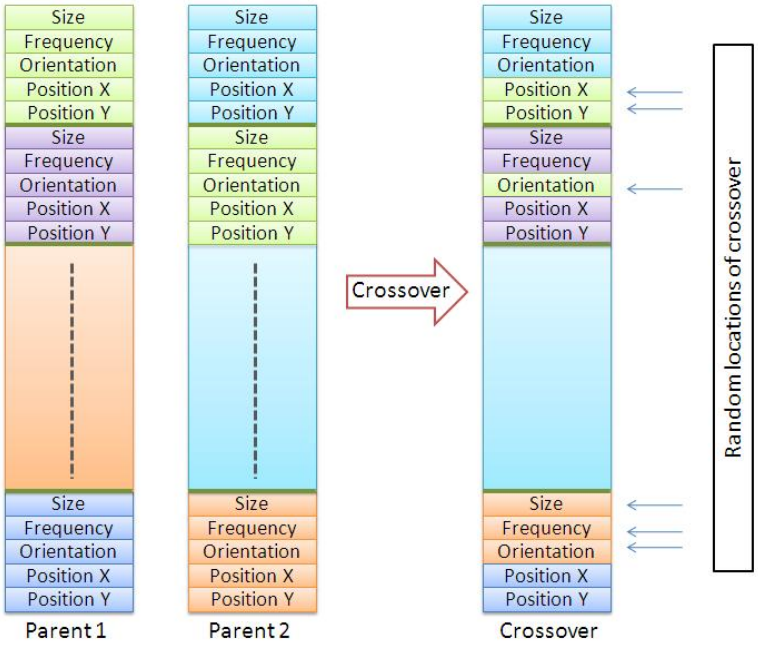


Figure 5.7: Typical crossing of two parents to create an offspring

Crossover.

Crossover

As discussed earlier, the parents for crossover are selected by a random number generator. Between these parents, the points of crossover are determined by choosing locations of crossover randomly. As seen in the Figure 5.7, these locations are arbitrary gene boundary locations and at these locations the gene content from the two parents gets mixed. The offspring thus created now contains parts of the genes coming from the contributing parents. The motivation for this step is the fact that, as more and more generations pass, the fittest parents undergoing crossover will already contain the better sets of parameters, and their crossing might bring together the better sets of parameter values from both the parents.

Mutation

In addition to the process of crossover at gene boundaries in the chromosome, the values of some parameters within the genes might be changed randomly. This is illustrated in the

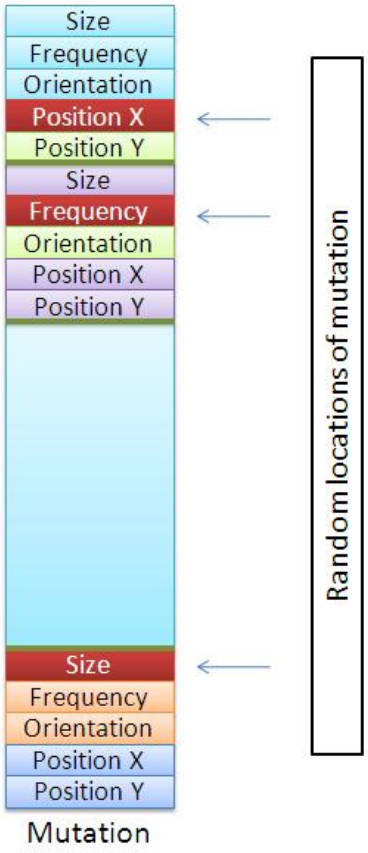


Figure 5.8: Mutation of a newly created offspring

Figure 5.8. Such mutations help in exploring the local parameter space more thoroughly. Mutations can be seen as small perturbations to the larger search that explores the vast parameter space, searching for the global minima.

5.5 Methodology

Most feature-based face recognition methods use feature detectors that are not tailored specifically for face recognition, and they make no attempt to selectively choose feature detectors based specifically on their usefulness for face recognition. The method described in this paper uses Gabor wavelets as feature detectors, but evaluates the usefulness of each particular feature detector (and a corresponding (x,y) location) for distinguishing between the faces within our face database. Given the very large number of possible Gabor feature detectors and locations, we use a Genetic Algorithm (GA) to explore the space of possibil-

ities, with a fitness function that propagates parents with a higher ability to distinguish between the faces in the database. By selecting the Gabor feature detectors and locations that are most useful for distinguishing each person from all of the other people in the database, we define a unique (i.e. person-specific) feature space for each person.

The FacePix (30) Database

All experiments were conducted with face images from the FacePix (30) database [151]. FacePix(30) was compiled to contain face images with pose and illumination angles annotated in 1 degree increments. Figure 5.9 shows the apparatus that is used for capturing the face images. A video camera and a spotlight are mounted on separate annular rings, which rotate independently around a subject seated in the center. Angle markings on the rings are captured simultaneously with the face image in a video sequence, from which the required frames are extracted.

Figure 5.9: The data capture setup for FacePix(30)

This database has face images of 30 people across a spectrum of pose and illu-

mination angles. For each person in the database, there are three sets of images. (1) The *pose angle set* contains face images of each person at pose angles from +90 to 90 (2) The *no-ambient-light set* contains frontal face images with a spotlight placed at angles ranging from +90 to -90 with no ambient light, and (3) The *ambient-light set* contains frontal face images with a spot light placed at angles placed at angels from +90 to -90 in the presence of ambient light. Thus, for each person, there are three face images available for every angle, over a range of 180 degrees. Figure 5.10 provides two examples extracted from the database, showing pose angles and illumination angles ranging from -90 to +90 in steps of 10. For earlier work using images from this database, please refer [152]. Work is currently in progress to make this database publicly available.

Figure 5.10: Sample face images with varying pose and illumination from the FacePix(30) database

We selected at random two images out of each set of three frontal (0) (Figure 5.11) images for training, and used the remaining image for testing. The genetic algorithms used the training images to find a set of Gabor feature detectors that were able to distinguish each persons face from all of the other people in the training set. These feature detectors were then used to recognize the test images.

In order to evaluate the performance of our system, we used the same set of training and testing images with face classification algorithm based on low-dimensional representation of face images extracted through Principal Component Analysis [153]. Specifically, the performance of the implementation of PCA-based face recognition followed by [154]

(a) (b) (c)

Figure 5.11: Sample frontal images of one person from the FacePix(30) Database

was used in our experiments.

The Gabor Features

Each Gabor feature corresponds to a particular Gabor wavelet (i.e. a particular special frequency, a particular orientation, and a particular Gaussian-defined spatial extent) applied to a particular (x, y) location within a normalized face image. (Given that 125 different Gabor filters were generated, by varying ω, σ and θ in 5 steps each, and given that each face image contained $128 \times 128 = 16,384$ pixels, there was a pool of $125 \times 16384 = 2,048,000$ potential Gabor features to choose from.) We used an N-dimensional vector to represent each person's face in the database, where N represents the predetermined number of Gabor features that the Genetic Algorithm selected from this pool. Figure 5.12 shows an example face image, marked with 5 locations where Gabor features will be extracted (i.e. N = 5). Given any normalized face image, real number Gabor features are extracted at these locations using Equation 5.6. This process can be envisioned as a projection of a 16,384-dimensional face image onto an N dimensional subspace, where each dimension is represented by a single Gabor feature detector.

Thus, the objective of the proposed methodology is to extract an N dimensional real-valued person-specific feature vector to characterize each person in the database. The N (x, y) locations (and the spatial frequency and spatial extent parameters of the N Gabor wavelets used at these locations) are chosen by a GA, with a fitness function that takes into

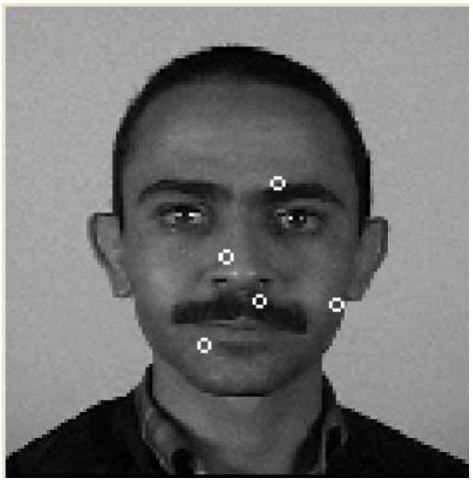


Figure 5.12: A face image marked with 5 locations where unique Gabor features were extracted

account the ability of each Gabor feature detector to distinguish one face from all the other faces in the database.

The Genetic Algorithm

Every GA is controlled in its progress through generations with a few control parameters such as,

- the number of generations of evolution (n_g)
- the number of parents per generation (n_p)
- the number of parents cloned per generation (n_c)
- the number of parents generated through cross over (n_{co})
- the number of mutations in every generation (n_m)

In our experiments, the GA used the following empirically-chosen GA parameters: $n_g = 50$, $n_p = 100$, $n_c = 6$, $n_{co} = 35$ and $n_m = 5$.

The Fitness Function

The fitness function of a genetic algorithm determines the nature of the search conducted over the parameter space. For face recognition applications, the fitness function is the capacity of a parent to classify the individuals accurately. In our proposed method, the fitness function needs to take both the Gabor features and the corresponding feature locations into consideration when evaluating face classification. We define here a fitness function that has two components to it. One determines the capacity of the parent to isolate an individual's face image from the others in the database, and the other evaluates whether the feature is redundant with other extracted features (i.e. whether a feature detector produces coefficients that are highly correlated with the coefficients produced by another feature detector.) Thus the fitness F can be defined as

$$F = w_D D - w_C C \tag{5.7}$$

where D is the distance measure weighted by w_D, and C represents the correlation measure which measure the similarity between the coefficients that have been extracted. The correlation measure C is weighted by the factor w_C.

If a parent extracts features from a face image that distinguish one individual from all the others very well (compared to the other parents within the same generation) then the distance measure D will be the largest for that parent, making its fitness F large. If the correlation between the extracted features is small, C will be small, which also makes the fitness F large. Thus, the correlation measure serves as a *penalty* for extracting the same feature from the face image multiple times, even though that particular feature might be the best distinguishing feature on that face.

The correlation between coefficients was used instead of spatial separation to counter the problem of similar features being extracted, because the Gabor filters might not be able to represent the underlying image characteristic completely. If there are some large image

features on the face (such as beard) that require multiple Gabor features within a certain spatial locality. Setting a hard lower limit on this spatial separation might lead to insufficient representation of that large image feature, in terms of the Gabor filters.

Consider a parent searching for a unique set of M Gabor filters to distinguish one individual's face from all other faces. Let this set of filters be referred to as S. Thus, $S = \{G_1, G_2, \cdots, G_M\}$ where, G_m represents the m^{th} Gabor filter.

If the set all individuals in the database is referred to as $I = \{i_1, i_2, \cdots, i_j\}$ with J number of individuals, then for every individual i in I a set S_i has to be extracted. To achieve this, all the images in the database depicting individual i are marked as positives, and the ones not depicting that individual are marked as negatives. Let the set of positive images be referred to as P_i (with L number of images) and the set of negatives be referred to as N (with K number of images). Thus, $S_i = \{G_{1i}, G_{2,i}, \cdots, G_{mi}\}$, $P_i = \{p_{1i}, p_{2i}, \cdots, p_{li}\}$ and $N_i = \{n_{1i}, n_{2i}, \cdots, n_{ki}\}$ are the sets of Gabor filters, positive images and negatives images set respectively for the individual i.

- **The Distance Measure** D

  A parent trying to recognize an individual i with a Gabor filter set S_i can be thought of as a transformation that projects all of the face images from the image space to a M-dimensional space, where the dimensions are defined by the M Gabor filters in the set S_i. Thus, all of the images in the two sets P_i and N_i can be considered as points on this M-dimensional space. Since the goal of the genetic algorithm is to find the set S_i which best distinguishes the individual i from others, in our method we search for the M dimensional space (defined by a parent) that best separates the points formed by the sets P_i and N_i. Figure 5.13 is an illustration of hypothetical set of face images projected on a 2 dimensional space defined by a set of 2 Gabor filters $S_i = \{G_0, G_1\}$. As shown in the figure, the measure D is the minimum of all the Euclidian distances between every positive and negative points.

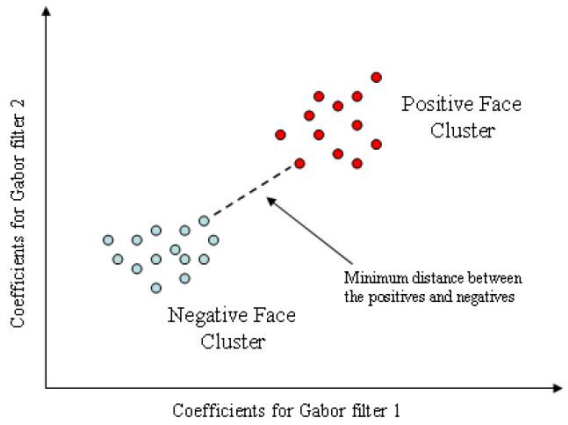


Figure 5.13: Distance Measure D for the fitness function

Thus, D can be defined as follow:

$$D = \min_{\forall\ l,k} [\delta_M (\phi_M(p_{li}), \phi_M(n_{ki}))] \tag{5.8}$$

where,

$\delta_M(A,B) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_m-b_m)^2}$ is the M-dimensional Euclidian distance between A and B. a_x and b_x corresponds the x^{th}-coordinate of A and B respectively

$\phi_M(X)$ is the transformation function that projects image X from the image space to the M-dimensional space defined by the set of Gabor filters.

- **The Correlation Measure** C

  In the proposed method, in addition to having every parent selecting the Gabor filter set S_i that can best distinguish the individual i from all the others in the database, it is necessary to ensure that this set of Gabor filters does not include filters that extract identical image features. If there were no such constraint, the algorithm might find one very distinguishing image feature on the face image and, over generations of evolution, all of its Gabor filters might converge to this one image feature. To avoid this, the correlation measure C determines the correlation between the image features extracted at all the locations pointed to by the chromosome. To test for correlations

between the Gabor features at the different spatial locations, we use the entire set of 125 Gabor filters to thoroughly characterize the textural context at these locations.

Assuming that there are M Gabor features that we are looking for on the face image of individual i, let $(x_m, y_m), m = 1, 2, \ldots, M$ be the M points that have been selected genetically in the chromosome. To find the correlations of the image features extracted at each of these points, the N Gabor filters $G_i, i = 1, 2, \ldots, N$ are used to characterize each of the points. Let the coefficients of such a characterization be represented by a matrix A. Thus, matrix A is $M \times N$ in dimension, where the rows correspond to the M locations and $N = 125$ refers to the Gabor filter coefficients. Thus,

$$A = \begin{bmatrix} g_{(1,1)} & g_{(1,2)} & \cdots & g_{(1,N)} \\ g_{(2,1)} & g_{(2,2)} & \cdots & g_{(2,N)} \\ \vdots & \vdots & \vdots & \vdots \\ g_{(m,1)} & g_{(m,2)} & \cdots & g_{(m,N)} \end{bmatrix} \tag{5.9}$$

where, $g_{(m,n)}$ is the coefficient obtained by applying the n^{th} Gabor filter to the image at the point (x_m, y_m).

The Correlation measure can now be defined in terms of matrix A as follows

$$C = \log\left(\det\left(diag(B)\right)\right) - \log\left(\det(B)\right) \tag{5.10}$$

where, $diag(B)$ returns the diagonal matrix corresponding to B, and B is the covariance matrix defined by $B = \frac{1}{N-1}(AA^T)$.

Examining the Equation 5.10, it can be seen that the first log term gets closer to the second log term when the off diagonal elements of B reduces. The diagonal elements of the matrix B corresponds to the variance of the M image locations, whereas the off diagonal elements correspond to the covariance between pairs if locations. Thus, as the covariance between the image points decreases, the value of the overall correlation parameter decreases.

- **Normalization of D and C**

  In order to have an equal representation of both the Distance measure D and the Correlation term C in the fitness function, it is necessary to normalize the range of values that they can take. For each generation, before the fitness values are used to rank the parents, parameters D and C are normalized to range between 0 and 1.

$$D_{norm} = \frac{D - D_{Min}}{D_{Max} - D_{Min}} \tag{5.11}$$

$$C_{norm} = \frac{C - C_{Min}}{C_{Max} - C_{Min}} \tag{5.12}$$

  where, the Max represents the maximum value of D or C in a single generation across all the parents and Min refers to the minimum value.

- **Weighting factors w_D and w_C**

  The influence of the two components of the fitness function are controlled by the weighting factors w_D and w_C. We used the relation $w_C = 1 - w_D$ to control the two parameters simultaneously. With this relationship, a value of $w_D \approx 1$ will subdue the effect of the Correlation measure, causing the genetic algorithm to choose the Gabor filters on the most prominent image feature alone. On the other hand, $w_D \approx 0$ will subdue the Distance measure, deviating the genetic algorithm from the main goal of face recognition. Thus an optimal value for the weight w_D has to be estimated empirically, to suit the face image database in question.

5.6 Results

To evaluate the relative importance of the two terms (D and C) in the fitness function, we ran the proposed algorithm on the training set several times with 5 feature detectors per chromosome, while changing the weighting factors in the fitness function for each run, setting w_D to 0, .25, .50, .75, and 1.00, and computing $w_C = (1 - w_D)$. Figure 5.14 shows the recognition rate achieved in each case.

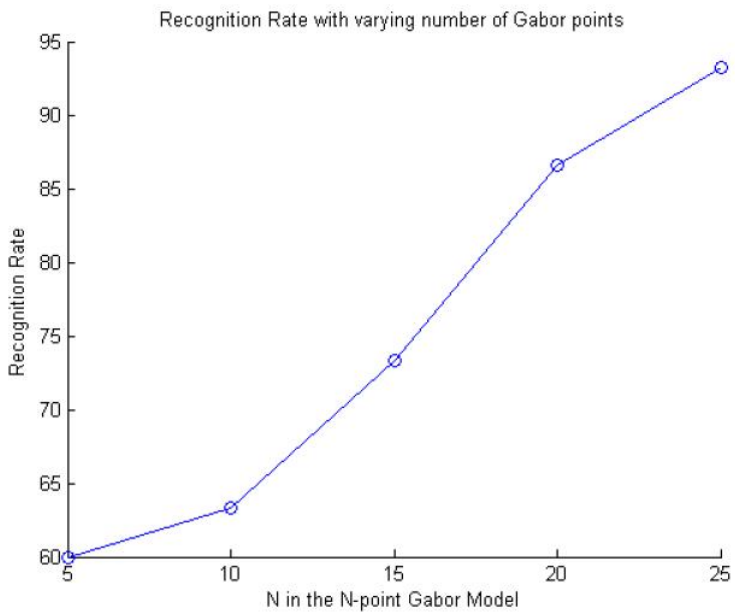


Figure 5.14: The recognition rate versus the number Gabor feature detectors

We then ran the proposed algorithm on the training set 5 times, while changing the number of Gabor feature detectors per parent chromosome for each run to 5, 10, 15, 20, and 25. In all the trials, w_D=0.5. Figure 5.15 shows the recognition rate achieved in each case.

Discussion of Results

Figure 5.14 shows that the recognition rate of the proposed algorithm when trained with 5, 10, 15, 20, and 25 Gabor feature detectors increases monotonically, as the number of Gabor feature detectors (N) is increased. This can be attributed to the fact that increasing the number of Gabor features essentially increases the number of dimensions for the Gabor feature detector space, allowing for greater spacing between the positive and the negative clusters.

Figure 5.15 shows that for N = 5 the recognition rate was optimal when the distance measure D and the correlation measure C were weighted equally, in computing the fitness function F. The dip in the recognition rate for $w_D = 0.75$ and $w_D = 1.0$ indicates the significance of using the correlation factor C in the fitness function. The penalty introduced

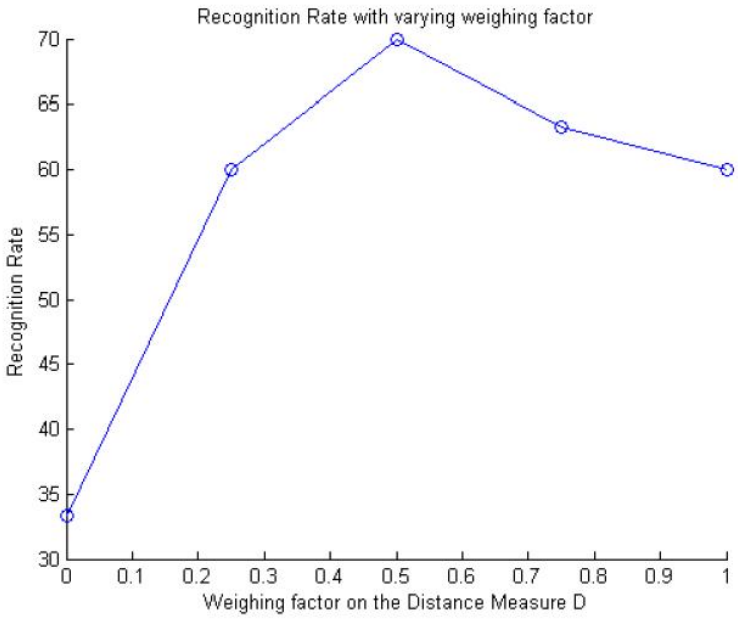


Figure 5.15: Recognition rate with varying w_D

by C ensures that the GA searches for Gabor features with different textural patterns. If no such penalty were be imposed, the GA might select Gabor features that are clustered on one salient facial feature, such as a mole.

The best recognition results for the proposed algorithm (93.3%) were obtained with 25 Gabor feature detectors. The best recognition performance for the PCA algorithm was reached at about 15 components, and flattened out beyond that point, providing a recognition rate for the same set of faces that was less than 83.3%. This indicates that, for the face images used in this experiment (which included substantial illumination variations) the proposed method performed substantially better than the PCA algorithm.

Person-specific feature extraction

When the FacePix(30) face database was built, all but one person were captured without eyeglasses or a hat. Figures 5.16(a) and 5.16(b) show the results of extracting 10 and 20 distinguishing features from that person's face images. The important things to note about these results are:

1. At least half of the extracted Gabor features (8 of the 10) and (10 of the 20) are located on (or near) the eyeglasses.

2. As the number of Gabor features was increased from 10 to 20, more Gabor features are seen toward the boundaries of the images. This is due to the fact that the genetic algorithm chooses Gabor feature locations based on a Gaussian probability distribution that is centered over the image, and decreases toward the boundaries of the images.

(a) (b)

Figure 5.16: 10 and 20 person-specific features extracted for a particular individual in the database

These results suggest that person-specific feature extraction might be useful for face recognition in small face databases, such as those typical of a social interaction assistance device for people who are blind.

5.7 Conclusions and Future Work

As mentioned earlier, the proposed person-specific approach to evolutionary feature selection in face images is well-suited for applications such as those that enhance social interaction for people who are blind, because people do not generally disguise their appearance in normal social situations, and even when some significant change occurs (such as a man shaving off his beard) the system can continue to evolve as it captures new images with each encounter.

A wearable social interaction assistant prototype has been implemented using a pair of eyeglasses equipped with a tiny unobtrusive video camera in the nose bridge [155] and is shown in Section XXXX, Figure XXXX. The analog video output from this camera is passed through a video digitizer, and the resulting digital stream is then fed into a portable laptop computer. A video stream is captured of any person standing in front of the eyeglasses. A face detection algorithm, based on Adaboost [156], is then used to identify the frames of the video where a face is present, and to localize that face within that frame. This detected face is then cropped and compared to indexed faces in a face database.

The performance of the proposed approach for identifying person-specific features relies, to a large extent, on obtaining near-frontal views of faces. To offset this limitation, there is ongoing work [157] to perform person-independent head pose estimation on the face images obtained from this platform. It is expected that this will help us select face images from the video stream with near-frontal views, which will improve the performance of our algorithm in identifying person-specific features.

Another factor that limits the performance of our algorithm is illumination variations in the captured images. Especially problematic are variations between outdoor-indoor and day-night settings. (Of course, this limitation is not unique to our algorithm.) As a strategy to provide additional light unobtrusively under adverse lighting conditions, we are employing infra-red LED illuminators in conjunction with an infrared-sensitive camera.

In summary, while there have been many different feature-based approaches to face recognition over the last two decades of research, we have proposed a novel methodology based on the discovery and extraction of person-specific characteristic features to improve face recognition performance for small face databases. This approach is aimed at facilitating social interaction in casual settings. The use of Gabor features, in tandem with a genetic algorithm to discover characteristic person-specific features has been inspired by the human visual system and is based on knowledge that has been developed about the process by which humans recognize faces. We believe that more needs to be learnt about human face recognition, and that as more is learnt, the knowledge can be put to use to

develop more robust face recognition algorithms.

Chapter 6

EXOCENTRIC SENSING

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 [158]. The following list describes the American proxemic distances; note that such distances vary with culture and environment.

1. Intimate Distance (Close Phase): 0-6 inches
2. Intimate Distance (Far Phase): 6-18 inches
3. Personal Distance (Close Phase): 1.5-2.5 feet
4. Personal Distance (Far Phase): 2.5-4 feet
5. Social Distance (Close Phase): 4-7 feet
6. Social Distance (Far Phase): 7-12 feet
7. Public Distance (Close Phase): 12-25 feet
8. 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 [159], 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 [160] 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 [161] [162], which are vibratory signals used to communicate abstract messages [162] 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.

6.1 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 [163] 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. **********

6.2 Accurate Face Detection

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 [164] 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 [**?**] and [165].

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 6.1 shows an example where a face detection algorithm detects two faces - one true and the other false.

Figure 6.1: 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 paper 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 6.2 shows the functional flow of data through the two modules in the proposed framework. The details of the statistical models and their learning will be presented later in the paper, which is organized as follows. Section 2 reviews some of the earlier research. Section 3 introduces the proposed framework, with details on the learning process. Section 4 discusses the experiments carried out to test the proposed framework. Section 5 presents the results while Section 6 discusses them. Section 7 concludes the paper and discusses future work.

6.3 Related Work

As mentioned earlier, the problem of face detection validation has not been treated methodically before, though the problem has been handled by many as an integral component of face detection algorithms. All the past work in this area can be broadly characterized into two groups: a) Low level image feature models mostly based on skin color such as [166], [167] and [168], and b) High level facial feature models such as [169], [170] and [171].

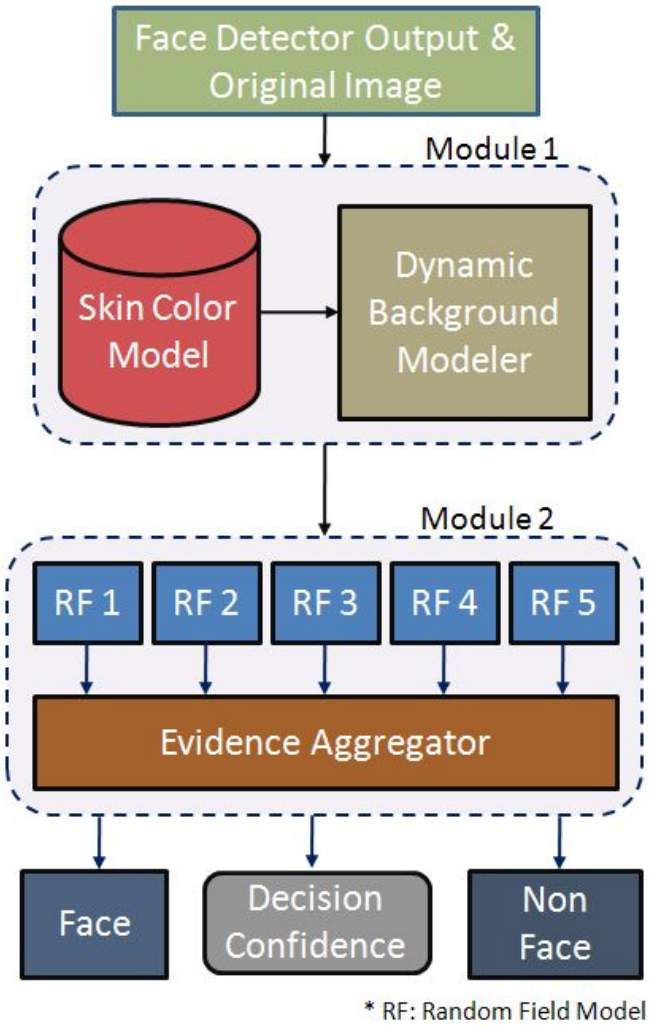


Figure 6.2: Block diagram.

The low level skin color based approaches try to reduce computational complexity by first identifying skin color in images so that search can be reduced. Most of the times, simple geometrical properties of the retained skin regions are used to determine if the region is a face. Such simplification of faces into trivial geometrical structures results in false detections. The facial feature based methods achieve face detection by individually identifying the integral components of a face image such as eyes, nose, etc. Though these schemes could be robust, the associated computational load is high. Interested readers could find more related references in [165] and [**?**]. The framework proposed in this paper uses statistically learnt knowledge about human faces to overcome computational complexity

thereby augmenting face validation to existing face detection algorithms seamlessly.

6.4 Proposed Framework

As shown in Figure 6.2, 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 constucting 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 (*t*rue) face detection outputs. The results of each random field model within the aggregate are then combined, using rules of Dempster-Shafer Theory of Evidence [172]. This *c*ombining 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.

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:

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:

$$X^{nRGB}_{i|i\in\{r,g,b\}} = \frac{X_i}{\left(\sum_{\forall_{i|i\in\{r,g,b\}}} X_i\right)} \tag{6.1}$$

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:

$$X^{nRGB}_{i|i\in\{nR,nG,nB\}} = 1 - \left(\sum_{\forall k|(k\in\{nR,nG,nB\},k\neq i)} X_k \right) \tag{6.2}$$

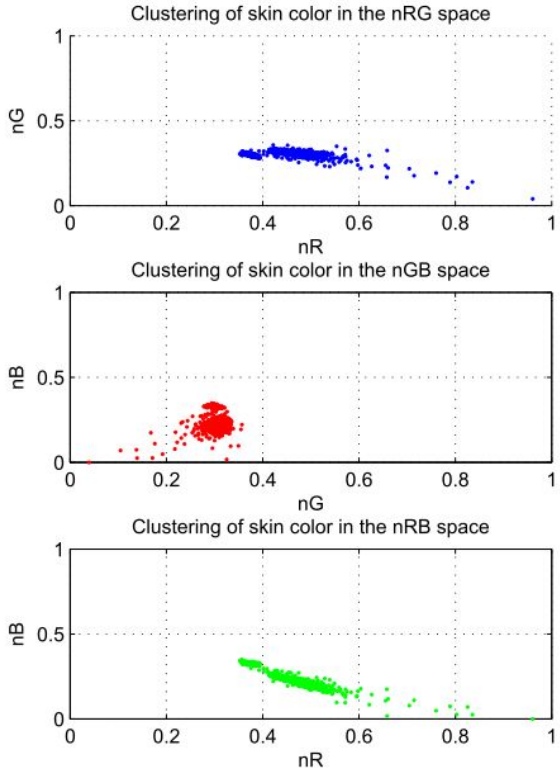


Figure 6.3: Skin pixels in nRGB space.

In our experiments, we found that skin pixels form a tight cluster when projected on nG and nB space as shown in the Figure 6.3. 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 [173]. 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 [174]. The Bi-modal Gaussian mixture model is represented as.

$$f^{skin}_{X|X=[nG,nB]}(x) = w_1 f_{Y_1}(x;\Theta_1 = [\mu_1,\Sigma_1]) +$$

$$w_2 f_{Y_2}(x; \Theta_2 = [\mu_2, \Sigma_2]) \tag{6.3}$$

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 6.4. Since the size of the face detector output varies from image to image, it is necessary to normalize the size. This is done by downsampling 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.

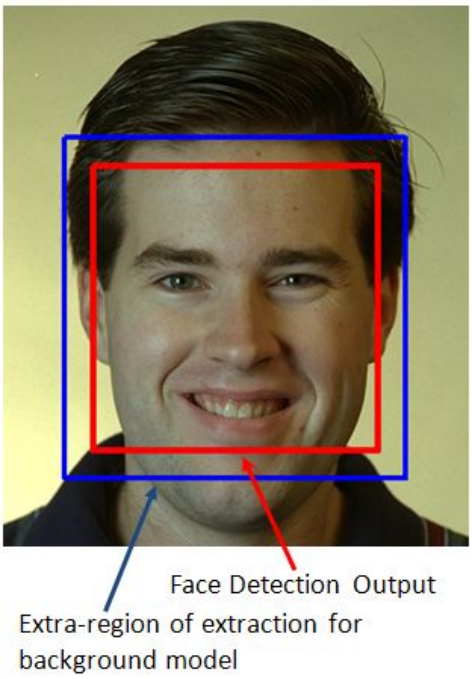


Figure 6.4: Extra region for background modeling.

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.

$$f^{non-skin}_{X|X=[R,G,B]}(x) = \sum_{i=1}^{m} w(i) f_{Y_i}(x; \Theta_i = [\mu_i, \Sigma_i]) \tag{6.4}$$

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.

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

The skin and non-skin models, $f^{skin}_{X|X=[nG,nB]}(x)$ and $f^{non-skin}_{X|X=[R,G,B]}(x)$ respectively, are used for classifying every pixel in the scaled face image obtained as explained in the Section 6.4. Example skin-masks are shown in Figure 6.5. This example shows two sets of images - one corresponding to a *t*rue face detection result, and another *f*alse face detection result.

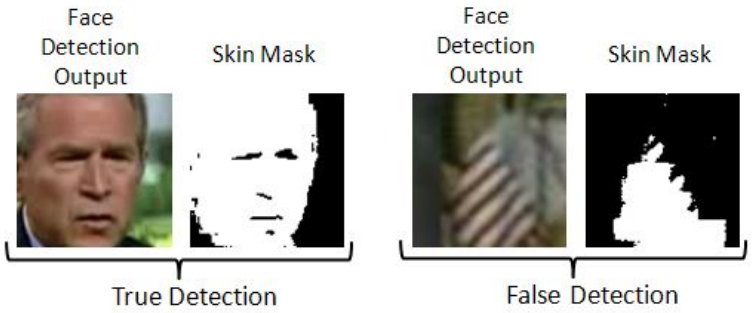


Figure 6.5: Example of *true* and *false* face detection.

The structural analysis through Random Field models explained in the next section will describe the design concepts that will help distinguish between *t*rue and *f*alse face detections shown in Figure 6.5.

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

In order to validate the skin region extracted as explained in Section 6.4, we build statistical models from examples of faces. We developed statistical learners inspired by Markov Random Fields (MRF) to capture the variations possible in *t*rue skin masks (face silhouette). The following subsections describes MRF models and the variant we created for our experiments.

Random Field (RF) Models

In this work, we used a minor variant of MRFs to learn the structure of a *t*rue 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 [175].

In an MRF, the sites in a set, $\mathscr{S}$, are related to one another via a neighborhood system, which is defined as $\mathscr{N} = \{\mathscr{N}_i, i \in \mathscr{S}\}$, where $\mathscr{N}_i$ is the set of sites neighboring i, $i \notin \mathscr{N}_i$ and $i \in \mathscr{N}_j \iff j \in \mathscr{N}_i$.

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

$$P(\mathbf{x}) > 0, \forall \mathbf{x} \in \mathscr{X} \tag{6.5}$$

$$P(x_i | x_{\mathscr{S}-\{i\}}) = P(x_i | x_{\mathscr{N}_i}) \tag{6.6}$$

where, $P(x_i | x_{\mathscr{S}-\{i\}})$ represents a Local Conditional Probability Density function defined over the neighborhood $\mathscr{N}$. The variant of MRF that we created for our experiments relaxed the constraints imposed by MRFs on $\mathscr{N}$. Typically, MRFs requires that sites in set $\mathscr{S}$ 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, Equation 6.6 could be alternatively seen as a set $P(\mathbf{x}) = \{P^1(\mathbf{x}), \ldots, P^5(\mathbf{x})\}$, each having their own neighborhood system $\mathscr{N}^k = \{\mathscr{N}^1, \mathscr{N}^2, \ldots, \mathscr{N}^5\}$, such that

$$P^k(x_i | x_{\mathscr{S}-\{i\}}) = P(x_i | x_{\mathscr{N}_i^k}) \tag{6.7}$$

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 6.6. 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.

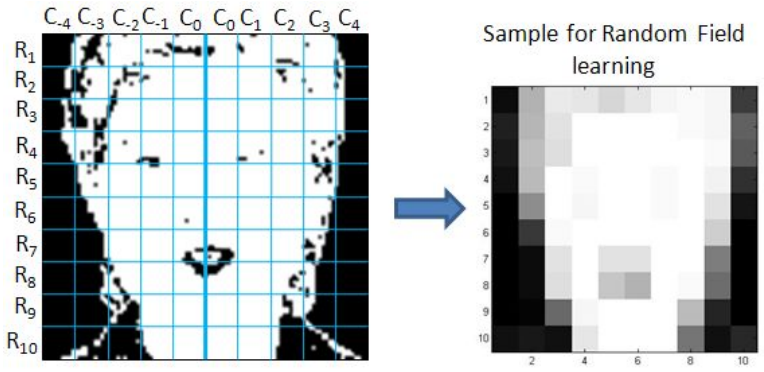


Figure 6.6: Pre-processing.

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 [176] 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 6.6. 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 6.6 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 any other $R_{i|i\neq 1}$. Thus, analysis of Local

Conditional Probability could be restricted to single R_i at a time, as shown in Figure 6.7.

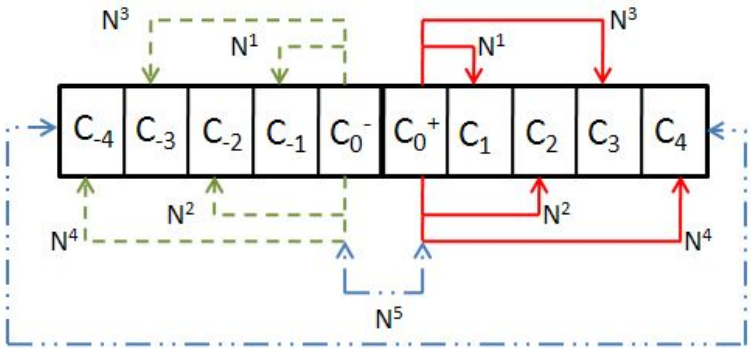


Figure 6.7: Neighborhood System.

The different neighborhood systems $\mathscr{N}^k$, used in the RF models, $P^k(x|x_{\mathscr{N}^k})$, can be defined as (Refer Figure 6.7):

$$\mathscr{N}^k = \left\{C_{j|j\in\{|k|,0^-,0^+\}}\right\} \tag{6.8}$$

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 *t*rue skin masks. The two dimensions (represented in a histogram pool $\mathbf{H}^k$) with individual element of the pool, $\mathbf{z}$, can be defined as:

- $\mathbf{H}^{k|k=\{1,2,3,4\}} = \{\mathbf{z}\}$, where,

$$\mathbf{z} = [x_{C_{0^\pm}}, \delta(x_{C_{0^\pm}}, x_{C_{\pm k}})], \forall R_j \tag{6.9}$$

- $\mathbf{H}^{k=5} = \{\mathbf{z}\}$, where,

$$\mathbf{z} = [\mu(x_{C_{0^+}}, x_{C_{0^-}}), \mu(x_{C_{-4}}, x_{C_{+4}})], \forall R_j \tag{6.10}$$

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

$$\delta(x_{C_{0^\pm}}, x_{C_{\pm i}}) = \begin{cases} x_{C_{0^+}} - x_{C_{+i}}, & i > 0 \\ x_{C_{-i}} - x_{C_{0^-}}, & i < 0 \end{cases} \tag{6.11}$$

$$\mu(a,b) = \frac{a+b}{2} \tag{6.12}$$

In order to estimate the LCPD on these 5 histogram pools, we use Parzen Window Density Estimation (PWDE) technique, similar to [177], with a 2D Gaussian window. Thus, each of LCPD can now be defined as

$$P^k(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{d}{2}} n h_{opt}^d} \sum_{j=1}^{n} \exp \left[-\frac{1}{2h_{opt}^2} \left(\mathbf{z} - \mathbf{H}_j^k \right)^T \Sigma^{-1} \left(\mathbf{z} - \mathbf{H}_j^k \right) \right] \tag{6.13}$$

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

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

Figure 6.8 shows the 5 LCPDs learnt over a set of 390 training frontal face images.

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 6.9. 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 *t*rue face detections into three piles, Turned right (r), Facing front (f), and, Turned Left (l).

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

$$P(\mathbf{z}) = \left\{ P_{m|m=\{r,f,l\}}^{k|k=\{1,\ldots,5\}}(\mathbf{z}) \right\} \tag{6.14}$$

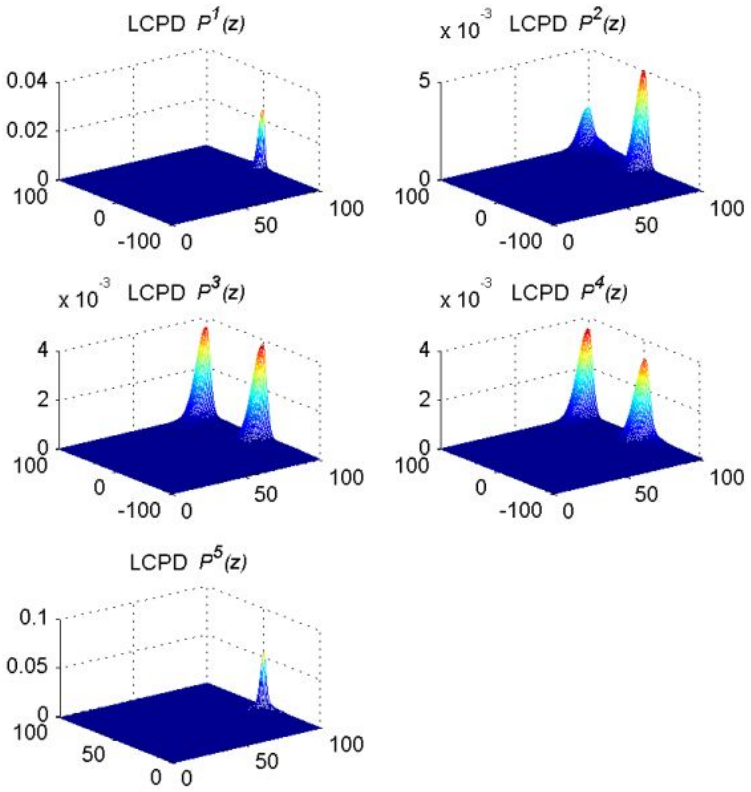


Figure 6.8: Frontal face Local Conditional Probability Density (LCPD) models.

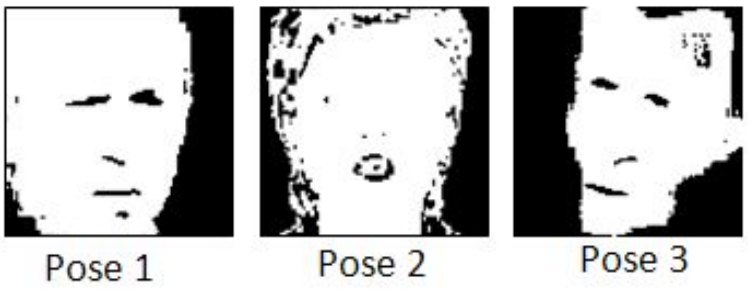


Figure 6.9: Skin-region masks.

Combining Evidence

Given any test face detection output, $\mathbf{z}$ is extracted (as described in Equation 6.9 and 6.10) and projected on the LCPD set $P(\mathbf{z})$ 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, $L_m^k = \ln\left(l_m^k\right)$, such that,

$$\prod_{\forall \mathbf{z} \in \mathbf{H}_m^k} \ln\left(l_m^k(\mathbf{z})\right) = \sum_{\forall \mathbf{z} \in \mathbf{H}_m^k} L_m^k(\mathbf{z}) \tag{6.15}$$

Given these log-likelihood values, one can set hard thresholds on each one of them to validate a face subimage discretely as *t*rue or *f*alse. 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 6.10. Each LCPD $P^k(\mathbf{z})$ was provided with an upper and lower threshold of acceptance and rejection respectively. The upper and lower bounds were obtained by observing $P^k(\mathbf{z})$ for the three face poses $P^k_{r,f,l}(\mathbf{z})$. 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 that 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 $P^k(\mathbf{Z})$, we resort to

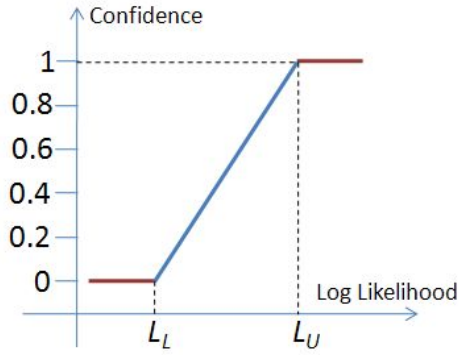


Figure 6.10: Soft threshold.

Dempster-Shafer Theory of Evidence.

Dempster-Shafer Theory of Evidence (DST)

The Dempster-Shafer theory is a mathematical theory of evidence [172] 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, $\mathbf{P}(X)$ (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),

$m : \mathbf{P}(X) \rightarrow [0,1]$, when it complies with the two axioms. a) $m(\emptyset) = 0$ and b) $\sum_{\mathbf{A} \in \mathbf{P}(X)} m(\mathbf{A}) = 1$. 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 *r*ules 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:

$$m_{1,2}(A) = \frac{1}{1-K} \sum_{B \cap C = A, A \neq \emptyset} m_1(B) m_2(C) \tag{6.16}$$

where,

$$K = \sum_{B \cup C = \emptyset} m_1(B) m_2(C) \tag{6.17}$$

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, $P^k(\mathbf{z})$, were considered as experts towards voting on the test input as a face or non-face. In order to use these mapped values in Equation 6.16 - 6.17, 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 6.11. The same idea could be extended to multiple experts.

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 $P^3(\mathbf{z})$, $P^4(\mathbf{z})$ and $P^5(\mathbf{z})$ 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. Due to space

		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 6.11: An example of combining evidence from two experts under Dempster-Shafer Theory.

constraints, the procedure is not explained in detail, but it is similar to what was followed for face versus non-face discrimination as explained in Section 6.4.

6.5 Experiments

In all our experiments, Viola-Jones face detection algorithm [164] was used for extracting face subimages. The proposed face validation filter was tested on two face image data sets, 1. *T*he FERET Color Face Database, and 2. *A*n 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 *t*rue or *f*alse 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 *t*rue detections: $4,420$
   - Number of *f*alse 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 *t*rue detections: $2,074$
- Number of *f*alse detections: 250 (10.7 %)

6.6 Results

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

1. Number of false detections (NFD)

$$\text{NFD} = \text{Count of false detections} \tag{6.18}$$

2. False detection rate (FDR):

$$\text{FDR} = \frac{\text{\# of false detections}}{\text{Total \# of face detections}} \text{x}100 \tag{6.19}$$

3. Precision (P)

$$\text{P} = \frac{\text{\# of true detections}}{\text{\# of true detections} + \text{\# of false detections}} \tag{6.20}$$

4. Capacity (C)

$$\text{C} = \left(\frac{\text{\# of true detections}}{\text{\# of actual faces in database}} \right) - \text{FDR} \tag{6.21}$$

Table 6.1: Face detection validation results on FERET database.

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

Table 6.2: Face detection validation results on the in-house face database.

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

As explained in Section 6.4, the framework was extensible to perform coarse pose estimation. Figure 6.12 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.

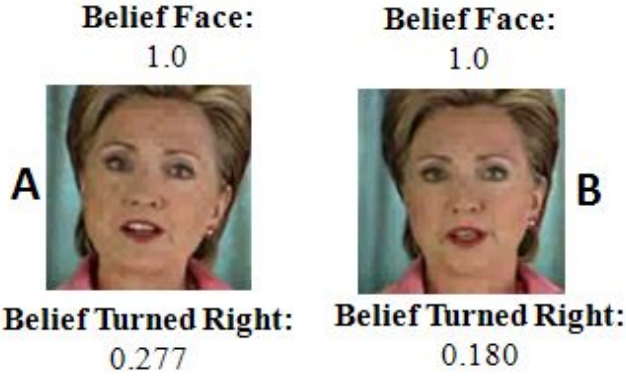


Figure 6.12: Coarse pose estimation.

6.7 Discussion of Results

Performance analysis of the proposed face validation filter can be understood through the four parameters defined in Section 6.6. **N**FB and **F**DR 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 6.1 and 6.2, 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 *t*rue 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 *t*rue 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 6.1 and 6.2 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 *t*rue face detection ability.

Finally, Figure 6.12 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.

Chapter 7

EXOCENTRIC SENSING: ACCURATE TRACKING OF PEOPLE

The problem of person localization in general is very broad in its scope and wide varieties of challenges such as variations in articulation, scale, clothing, partial appearances, occlusions, etc make this a complex problem. Narrowing the focus, this paper targets person localization in real world video sequences captured from the wearable camera of the Social Interaction Assistant. Specifically, we focus on the task of localizing a person who is approaching the user to initiate a social interaction or just conversation. In this context, the problem of person localization can be constrained to the cases where the person of interest is facing the user.

Figure 7.1: Person of interest at a short distance from camera

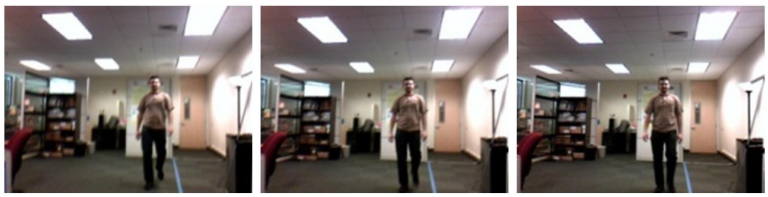


Figure 7.2: Person of interest at a large distance from camera

When such a person of interest is in close proximity, his/her presence can be detected by analyzing the incoming video stream for facial features (Figure 7.1). But when such a person is approaching the user from a distance, the size of the facial region in the video appears to be extremely small. In this case, relying on facial features alone would not suffice and there is a need to analyze the data for full body features (Figure 7.2). In this work, we have concentrated on improving the effectiveness of the SIA by applying computer vision techniques to robustly localize people using full body features. Follow-

ing section discusses some of the critical issues that are evident when performing person localization from the wearable camera setup of the SIA

7.1 Challenges in Person Localization from a wearable camera platform

A number of factors associated with the background, object, camera/object motion, etc. determine the complexity of the problem of person localization from a wearable camera platform. Following is a descriptive discussion of the imminent challenges that we encountered while processing the data using the SIA.

Background Properties

When the Social Interaction Assistant is used in natural settings, it is highly possible that there are objects in the background which move, thus causing the background to be dynamic. Also, there are bound to be regions in the background whose image features are highly similar to that of the person, thus leading to a cluttered background. Due to these factors, the problem of distinguishing the person of interest from the background becomes highly challenging in this context. Figures 7.3 and 7.4 illustrate the contrast in the data due to the nature of the background.

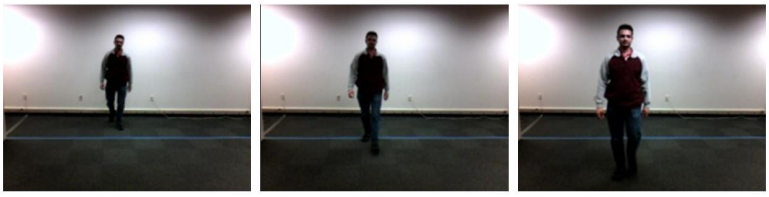


Figure 7.3: Simple Background

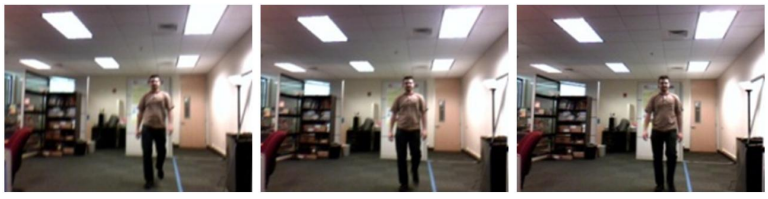


Figure 7.4: Complex Background

Object Properties

As we are interested in person localization, it can be clearly seen that the object is non-rigid in nature as there are appearance changes that occur throughout the sequence of images. Further, significant scale changes and deformities in the structure can also be observed. Also, when analyzing video frames of persons approaching the user, the basic image features in various sub-regions of the object vary vastly. For example, the image features from the facial region are considerably different from that of the torso region. Tracking detected persons from one frame to another will require individualized tracking of each region to maintain confidence. This non-homogeneity of the object poses a major hurdle while applying localization algorithms and has not been studied much in the literature. Figure 7.5 shows the simplicity of the data when these problems are not present, while Figure 7.6 highlights complex data formulations in a typical interaction scenario.

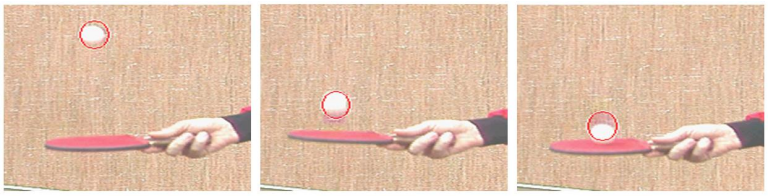


Figure 7.5: Rigid, Homogeneous Object

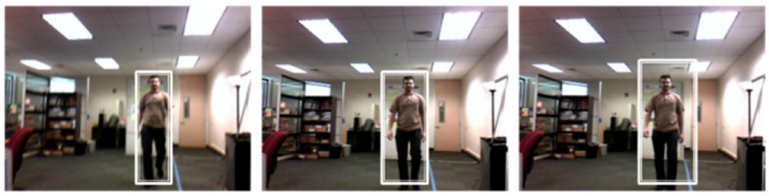


Figure 7.6: Non-Rigid, Deformable, Non-Homogeneous Object

Object/Camera Motion

Traditionally, most computer vision applications use a static camera where strong assumptions of motion continuity and temporal redundancy can be made. But in our problem, as it is very natural for users to move their head continuously, the mobile nature of the platform causes abrupt motion in the image space (Compare Figure 7.7 and Figure 7.9). This is

similar to the problem of working with low frame rate videos or the cases where the object exhibits abrupt movements. Recently, there has been an increase of interest in dealing with this issue in computer vision research [178] [179] [180] [181]. Some important applications which are required to meet real-time constraints, such as teleconferencing over low bandwidth networks, and cameras on low-power embedded systems, along with those which deal with abrupt object and camera motion like sports applications are becoming common place [181]. Though solutions have been suggested, person localization through low frame rate moving cameras still remains an active research topic.

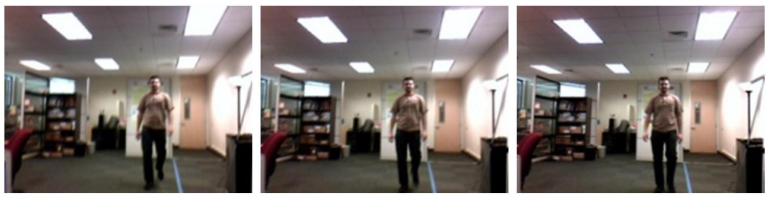


Figure 7.7: Static Camera

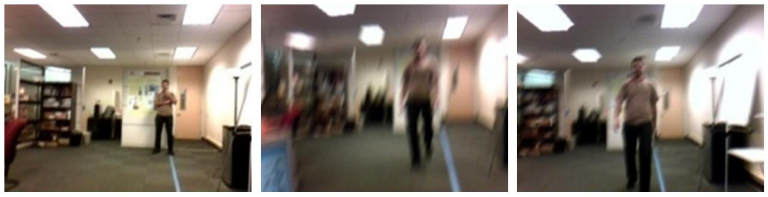


Figure 7.8: Mobile Camera

Other Important Factors Affecting Effective Person Tracking

As the SIA is intended to be used in uncontrolled environments, changing illumination conditions need to be taken into account. Further, partial occlusions, self occlusions, in-plane and out-of-plane rotations, pose changes, blur and various other factors can complicate the nature of the data. See Figure 7.9 for example situations where various factors can affect the video quality.

Given the nature of this problem, in this chapter we focus on the problem of robust localization of a single person approaching a user of the SIA using full-body features. Issues arising due to cluttered background along with object and camera motion have been handled towards providing robustness. In the following section we discuss some of the important

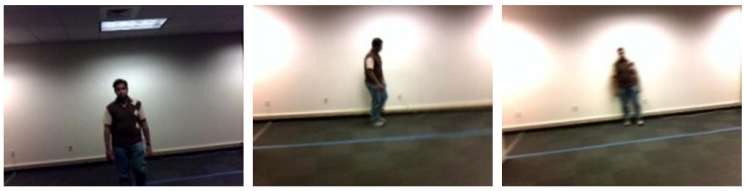


Figure 7.9: Changing Illumination, Pose Change and Blur

related work in the computer vision literature.

7.2 Related Computer Vision Work in Person Localization and Tracking

Historically, two distinct approaches have been used for searching and localizing objects in videos. On one hand, there are detection algorithms which focus on locating an object in every frame using specific spatial features which are fine tuned for the object of interest. For example, haar-based rectangular features [156] and histograms of oriented gradients [182] can develop detectors that are very specific to objects in videos. On the other hand, there are tracking algorithms which trail an object using generic image features, once it is located, by exploiting the temporal redundancy in videos. Examples of features used by tracking algorithms include color histograms [183] and edge orientation histograms [184].

Detection Algorithms

As mentioned previously, detection algorithms exploit the specific, distinctive features of an object and apply learning algorithms to detect a general class of objects. They use information related to the relative feature positions, invariant structural features, characteristic patterns and appearances to locate objects within the gallery image. But, when the object is complex, like a person, it becomes difficult for these algorithms to achieve generality thereby failing even under minute non-rigidity. A number of human factors such as variations in articulation, pose, clothing, scale and partial occlusions make this problem very challenging.

When assumptions about the background cannot be made, learning algorithms which take advantage of the relative positions of body parts are used to build classifiers.

The kind of low-level features generally used in this context are gradient strengths and gradient orientations [185] [182], , entropy and haar-like features. Some of the well-known higher level descriptors are histogram of oriented gradients [182] and covariance features [186]. Efforts have been made to make these descriptors scale invariant as well.

In order to make these algorithms real-time, researchers have popularly resorted to two kinds of approaches. One category includes part-based approach such as Implicit Shape Models [178] and constellation models [187] which place emphasis on detecting parts of the object before integrating, while the other category of algorithms tries to search for relevant descriptors for the whole object in a cascaded manner[188]. Shape-based Chamfer matching [189] is a popular technique used in multiple ways for person detection as the silhouette gives a strong indication of the presence of a person. In recent times, Chamfer matching has been used extensively by the person detection and localization community. It has been applied with hierarchically arranged templates to obtain the initial candidate detection blocks so that they can be analyzed further by techniques such as segmentation, neural networks, etc. It has also been used as a validation tool to overcome ambiguities in detection results obtained by the Implicit Shape Model technique [190].

Tracking Algorithms

Assuming that there is temporal object redundancy in the incoming videos, many algorithms have been proposed to track objects over frames and build confidence as they go. Generally they make the simplifying assumption that the properties of the object depend only on its properties in the previous frame, i.e. the evolution of the object is a Markovian process of first order. Based on these assumptions, a number of deterministic as well as stochastic algorithms have been developed.

Deterministic algorithms usually apply iterative approaches to find the best estimate of the object in a particular image in the video sequence [188]. Optimal solutions based on various similarity measures between the object template and regions in the current image, such as sum of squared differences (SSD), histogram-based distances, distances in

eigenspace and other low dimensional projected spaces and conformity to particular object models, have been explored [188]. Mean Shift is a popular, efficient optimization-based tracking algorithm which has been widely used.

Stochastic algorithms use the state space approach of modeling dynamic systems and formulate tracking as a problem of probabilistic state estimation using noisy measurements [191]. In the context of visual object tracking, it is the problem of probabilistically estimating the object's properties such as its location, scale and orientation by efficiently looking for appropriate image features of the object. Most of these stochastic algorithms perform Bayesian filtering at each step for tracking, i.e. they predict the probable state distribution based on all the available information and then update their estimate according to the new observations. Kalman filtering is one such algorithm which fixes the type of the underlying system to be linear with Gaussian noise distributions and analytically gives an optimal estimate based on this assumption. As most tracking scenarios do not fit into this linear-Gaussian model and as analytic solutions for non-linear, non-Gaussian systems are not feasible, approximations to the underlying distribution are widely used from both parametric and non-parametric perspective.

Sequential monte-carlo based Particle Filtering techniques have gained a lot of attention recently. These techniques approximate the state distribution of the tracked object using a finite set of weighted samples using various features of the system. For visual object tracking, a number of features have been used to build different kinds of observation models, each of which have their own advantages and disadvantages. Color histograms [183], contours [192], appearance models, intensity gradients [193], region covariance, texture, edge-orientation histograms, haar-like rectangular features [188], to name a few. Apart from the kind of observation models used, this technique allows for variations in the filtering process itself. A lot of work has gone into adapting this algorithm to better perform in the context of visual object tracking.

While both the areas of detection and tracking have been explored extensively, there is an impending need to address some of the issues faced by low frame rate visual tracking

of objects. Especially in the case of SIA, person localization in low frame rate video is of utmost importance. In this paper, we have attempted to modify the color histogram comparison based particle filtering algorithm to handle the complexities that occur mobile camera on the Social Interaction Assistant.

7.3 Conceptual Framework

As discussed in the previous section, detection and tracking offer distinctive advantages and disadvantages when it comes to localizing objects. In the case of SIA, thorough object detection is not possible in every frame due to the lack of computational power (on a wearable platform computing platform) and tracking is not always efficient due to the movement of the camera and the object's (interaction partner's) independent motion. Though there are clear advantages in applying these techniques individually, the strengths of both these approaches need to be combined in order to tackle the challenges posed by the complex setting of the SIA. In the past, a few researchers have approached the problem of tracking in low frame rate or abrupt videos by interjecting a standard particle filtering algorithm with independent object detectors [194]. In our experience, the Social Interaction Assistant offers a weak temporal redundancy in most cases. We exploit this information trickle between frames to get an approximate estimate of the object location by incorporating a deterministic object search while avoiding the explicit use of pre-trained detectors. Due to the flexibility in the design, particle filtering algorithms provide a good platform to address the issues arising due to complex data. These algorithms give an estimate of an object's position by discretely building the underlying distribution which determines the object's properties. But, real-time constraints impose limits on the number of particles and the strength of the observation models that can be used. This generally causes the final estimate to be noisy when conventional particle filtering approaches are applied. Unless the choice of the particles and the observation models fit the underlying data well, the estimate is likely to drift away as the tracking progresses. To mitigate these problems faced in the use of the SIA, we propose a new particle filtering framework that gets an initial estimate of the person's location by spreading particles over a reasonably large area and then successively corrects

the position though a deterministic search in a reduced search space. Termed as Structured Mode Searching Particle Filter (SMSPF), the algorithm uses color histogram comparison in the particle filtering framework at each step to get an initial estimate which is then corrected by applying a structured search based on gradient features and chamfer matching. The details of this algorithm are described in the next section.

7.4 STRUCTURED MODE SEARCHING PARTICLE FILTER

Assuming that an independent person detection algorithm can initialize this tracking algorithm with the initial estimate of the person location, this particle filtering framework focuses on tracking a single person under the following circumstances, namely

- Image region with the person is non-rigid and non-homogeneous
- Image region with the person exhibits significant scale changes
- Image region with the person exhibits abrupt motions of small magnitude in the image space due to the movement of the camera.
- Background is cluttered.

The algorithm progresses by implementing two steps on each frame of the incoming video stream. In the first step (Figure 7.10), an approximate estimate of the person region is obtained by applying a color histogram based particle filtering step over a large search space. This is followed by a refining second step (Figure 7.11) where the estimate is corrected by applying a structured search based on gradient features and Chamfer matching. These two steps have been described in detail below.

Step 1: Particle Filtering Step

In the context of SIA, as the person of interest can exhibit abrupt motion changes in the image space, it is extremely difficult to model the placement of the person in the current image based on the previous frame's information alone. When such data is modeled in the Bayesian filtering based particle filtering framework, the state of each particle's position

Figure 7.10: SMSPF - Step 1

becomes independent of its state in the previous step. Thus, the prior distribution can be considered to be a uniform random distribution over the support region of the image.

$$p\left(x_t^i | x_{t-1}^i = p(x_t^i)\right) \tag{7.1}$$

As it is essential for particle filtering algorithm to choose a good set of particles, it would be useful to pick a good portion of them near the estimate in the previous step. By approximating this previous estimate to be equivalent to a measurement of the image region with the person in the current step, the proposal distribution of each particle can be chosen

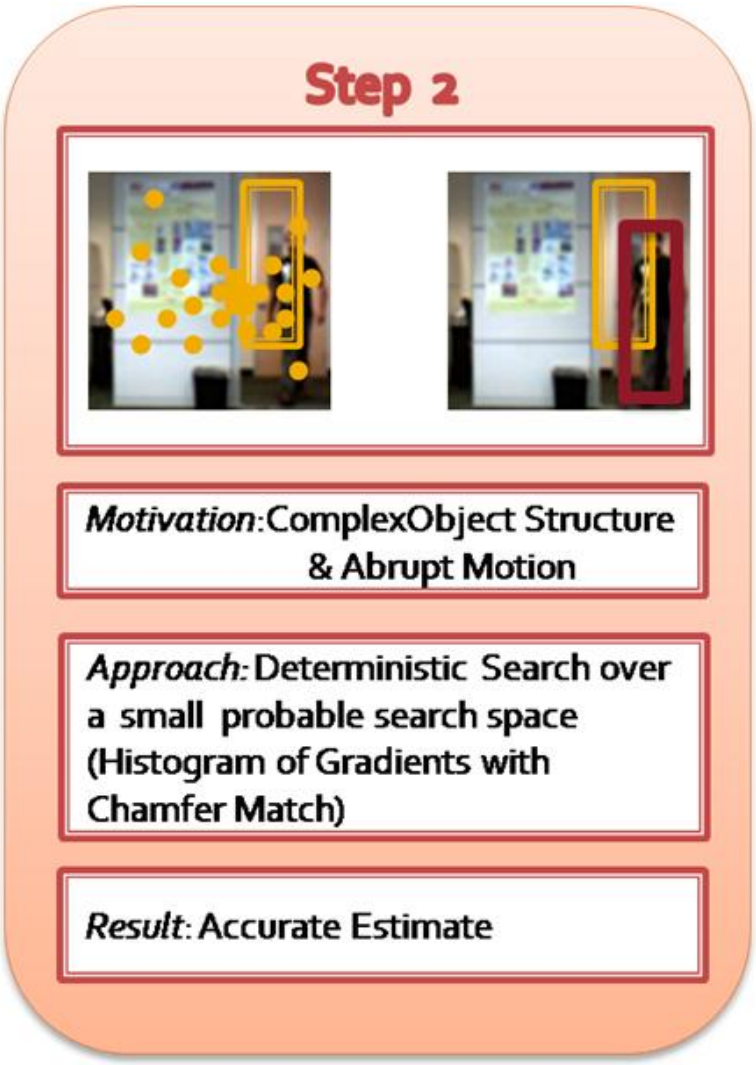


Figure 7.11: SMSPF - Step 2

to be dependent only on the current measurement

$$q\left(x_t^i | x_{t-1}^i Z_t\right) = q\left(x_t^i | Z_t\right) \tag{7.2}$$

Though the propagation of information through particles is lost by making such an assumption, it gives a better sampling of the underlying system. We employ a large variance Gaussian with its mean centered at the previous estimate for successive frame particle propagation. By using such a set of particles, a larger area is covered, thus accounting for abrupt motion changes and a good portion of them are picked near the previous esti-

Figure 7.12: Structured Search

mate, thus exploiting the weak temporal redundancy. As in [183], we have employed this technique using HSV color histogram comparison to get likelihoods at each of the particle locations. Since intensity is separated from chrominance in this color space, it is reasonably insensitive to illumination changes. We use an 8x8x4 HSV binning thereby allowing lesser sensitivity to changes in V when compared to chrominance. The histograms are compared using the well-known Bhattacharyya Similarity Coefficient which guarantees near optimality and scale invariance.

With the above step alone, due to the small number of particles which are spread widely across the image, we can get an approximate location of the person. When such an estimate partially overlaps with the desired person region, the best match occurs between the intersection of the estimate and the actual person region as shown in Figure 7.12. But, it is not trivial to detect this partial presence due to the existence of background clutter. To handle this problem, we introduce a second step which uses efficient image feature representations of the desired person object and employs an efficient search around the estimate to accurately localize the person object.

Step 2: Structured Search

As the estimate obtained using widely spread particles gives the approximate location of the object, the search for the image block with a person in it can be restricted to a region

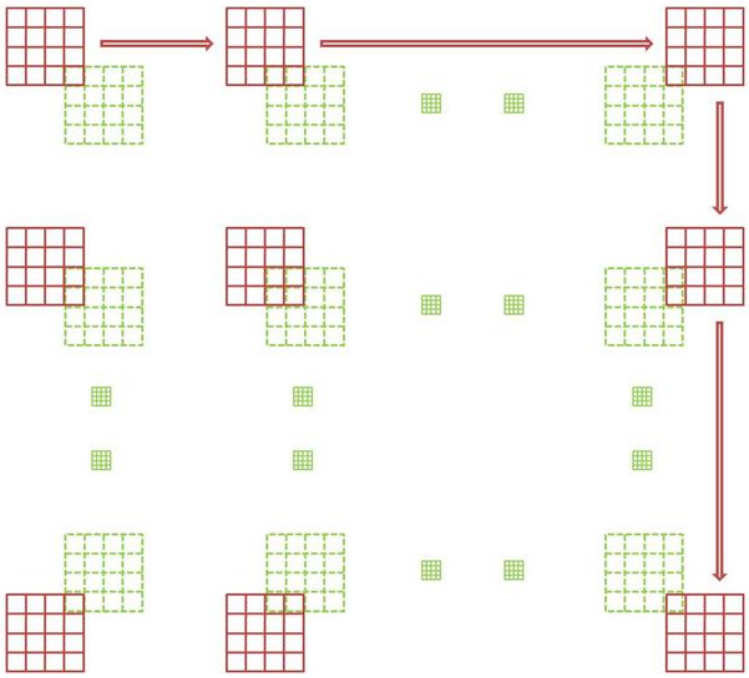


Figure 7.13: Sliding window of the Structured Search (Green: Estimate; Red: Sliding window).

around it. We have employed a grid-based approach to discretely search for the object of interest (a person) instead of checking at every pixel. By dividing the estimate into an m x n grid and sliding a window along the bins of the grid as shown in Figure 7.13, the search space can be restricted to a region close to the estimate. By finding the location which gives the best match with the person template, we can localize the person in the video sequence with better accuracy.

If this search is performed based on scale-invariant features, then it can be extended to identify scale changes as well. In order to achieve search over scale, the estimate and the sliding window need to be divided into different number of bins. If the search is performed using smaller number of bins as compared to the estimate, then shrinking of the object can be identified while searching with higher number of bins can account for dilation of the object. For example, if a (m-1) x (n-1) grid is used with the sliding window while a m x n grid is used with the estimate, then the best match will find a shrink in the object size. Similarly if an m x n grid sliding window is used with a (m-1) x (n-1) estimate grid, then

dilations can be detected. It can be seen that this search is characterized by the number of bins m x n into which the sliding window and the estimate are divided. Based on the nature of the problem, the number of bins and the amount of sweep across scale and space can be adjusted. Currently, these parameters are being set manually, but the structured search framework can be extended to include online algorithms which can adapt the number of grid bins based on the evolution of the object.

If the object of interest was simple, then the best match across space and scale could be obtained by using simple feature matching techniques. But, due to the complex nature of the data, strong confidence is required while searching for the person region across scale. To this end, we propose to perform the structured search by analyzing the internal features of the person region as well as the external boundary/silhouette features and aggregating the confidence obtained from these two measures to refine the person location estimate in the image (Figure 7.14)

In literature, gradient based features have been widely used for person detection and tracking problems and their applicability has been strongly established by various algorithms like Histogram of Oriented Gradients (HoGs) [182]. Following this principle, we have used the Edge Orientation Histogram (EOH) features [184] in order to obtain the internal content information measure. For this purpose, a gradient histogram template (GHT) is initially built using a generic template image of a walking/standing person. This GHT is then compared with the gradient histogram of each structured search block using the Bhattacharyya histogram comparison as in [183] in order to find the block with the best internal confidence. In our implementation, orientations are computed using the Sobel operator and the gradients are then binned into 9 discrete bins. These features were extracted using the integral histogram concept [195] to facilitate computationally efficient searching.

Similarly, in order to obtain the boundary confidence measure, a generic person silhouette template (GPT) (as shown in Figure 7.14) is used to perform a modified Chamfer match on each of the search blocks. In general, Chamfer matching is used to search for a particular contour model in an edge map by building a distance transformed image of the

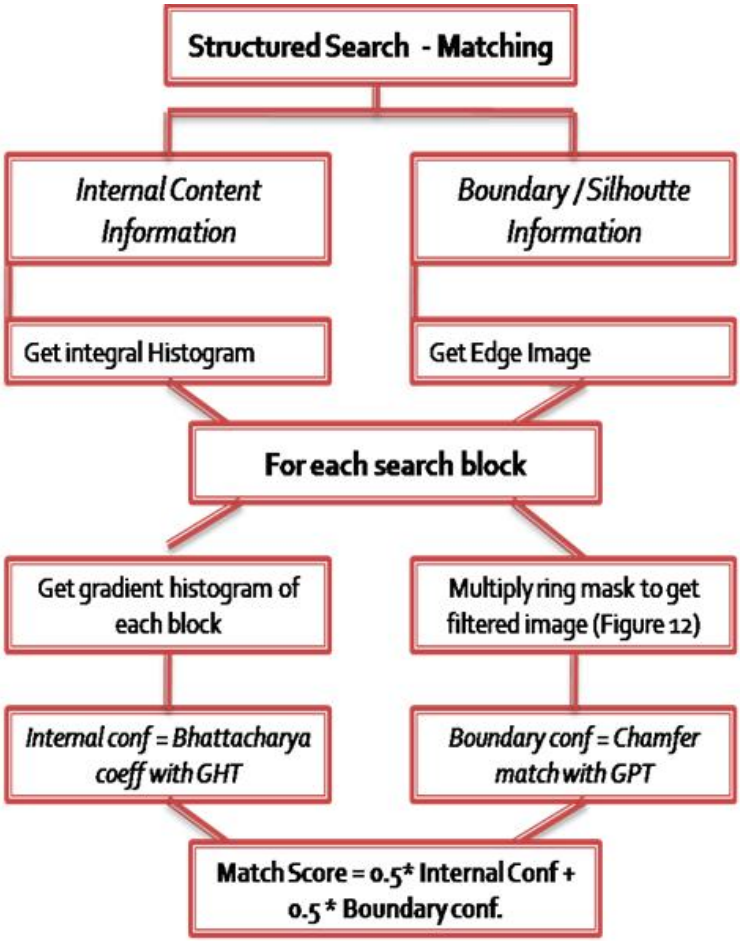


Figure 7.14: Structured Search Matching Technique

edge map. Each pixel value in a distance transformed image is proportional to the distance to its nearest edge pixel. In order to compare the edge map to the contour map, we convolve the edge image with the contour map. If the contour completely overlaps with the matching edge region, we get a chamfer match value of zero. Based on how different the edge map is to the template contour, the chamfer match score will increase and move towards 1. A chamfer match score of 1 implies a very bad match.

While the theory of chamfer matching offers elegant search score, in reality, especially with clutter within the object's silhouette, it is very difficult to get an exact match score. In SIA, since the data is very noisy and complex, certain modifications need to be made with the Chamfer matching algorithm in order achieve good performance. The

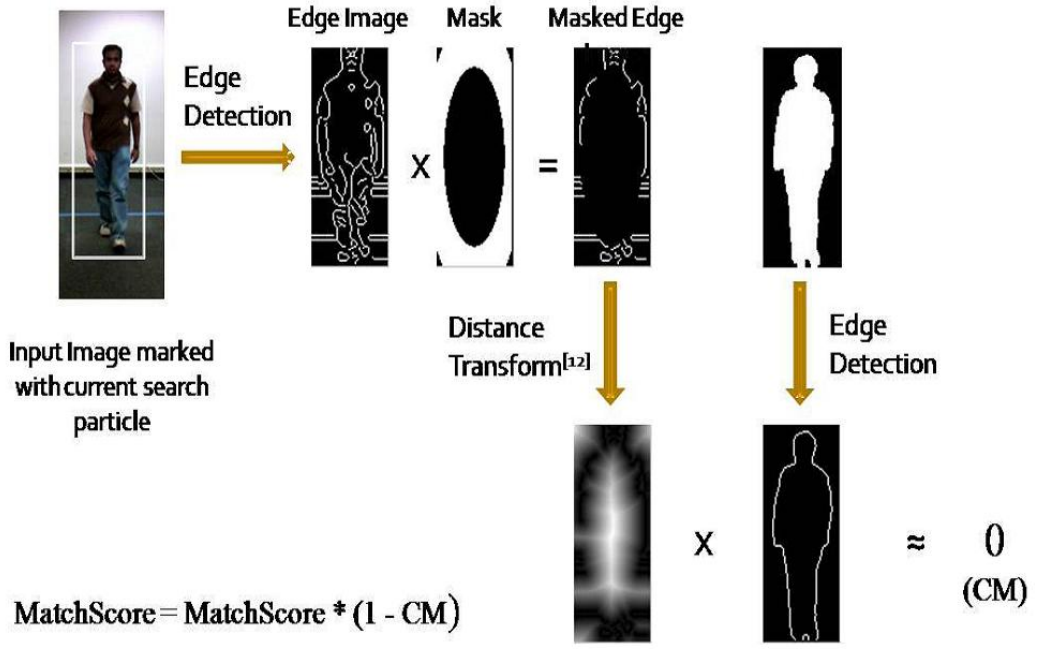


Figure 7.15: Incorporating Chamfer Matching into Structured Search

following section details a modified Chamfer match algorithm introduced in this work.

Chamfer Matching in Structured Search

As discussed above, Chamfer matching gives a measure of confidence on the presence of the person within an image based on silhouette information. We have incorporated this confidence into the structured search in order to detect the precise location of the person around the particle filter estimate. An edge map of the image under consideration is first obtained which is then divided into (m x n) windows in accordance with the structured search and an elliptical ring mask is then applied to each of these windows as shown in Figure 7.15. This mask is applied so as to eliminate the edges that arise due to clothing and background thereby emphasizing the silhouette edges which are likely to appear in the ring region if a window is precisely placed on the object perimeter. A distance transformed image of the window is then obtained using the masked edges.

By applying the modified chamfer matching (with a generic person contour resized to the current particle filter estimate), a confidence number in locating the desired object

within the image region can be obtained. Similar to the Chamfer matching as before, a value close to 0 indicates a strong confidence of the presence of a person and vice versa. As 1 is the maximum value that can be obtained by the chamfer match, this measure can be incorporated into the match score of the structured search using the following equation.

$$\text{BoundaryConf} = (1 - \text{ChamferMatch}) \tag{7.3}$$

The standard form of Chamfer Matching gives a continuous measure of confidence in locating an object in an edge map. But, in our case, when the elliptical ring mask is used to filter out the noisy edges in each search block, this nature of Chamfer match is lost. Since the primary goal of the structured search is to find a single best matching location of the person, it is more advantageous to use the filter mask at the cost of losing this continuous nature of the chamfer match. Further, as it is very likely that the person region is close to the approximate estimate obtained from the first step, one of the search windows of the structured search is bound to capture the entire person object thus resulting in a good match score.

From the above discussion, it can be seen that combining the knowledge about the internal structure of the person region with the silhouette information results in a greater confidence in the SMSPF algorithm. Further, using such complementary features in the structured search robustly corrects the approximate estimate obtained from the particle filtering step while handling various problems associated with search across scale.

7.5 Experiments and Datasets

Datasets

The performance of the structured mode searching particle filter (SMSPF) has been tested using three datasets where a single person faces the camera while approaching it. There are significant scale changes in each of these sequences. Further, non-rigidity and deformability of the person region can also be clearly observed. Different scenarios with varying degrees of complexity of the background and camera movement have been considered. Following

(a) SMSPF Results on a sequence from Dataset1

(b) SMSPF Results on a sequence from Dataset 2

(c) SMSPF Results on a sequence from Dataset 3

Figure 7.16: SMSPF Results

is a brief description of these datasets.

(a) *DataSet* [1]: Plain Background; Static Camera; 320x240 resolution

(b) *DataSet* [2]: Slightly cluttered Background; Static Camera; 320x240 resolution

(c) *DataSet* [3]: Cluttered Background; Mobile Camera; 320x240 resolution

Figure 7.16 shows the sample results on each of the datasets used.

Evaluation Metrics

In order to test the robustness of this algorithm and the applicability in complex situations, its performance has been compared with the Color Particle Filtering algorithm [189]. Assuming that a detection algorithm can detect persons in at least some frames, the image

[1]Collected at CUbiC

[2]CASIA Gait Dataset B with subject approaching the camera [4]

[3]Collected at CUbiC

region containing the person in each of the test sequences has been manually set. The following two criteria have been used to evaluate their performance,

- Area Overlap (A0)
- Distance between Centroids (DC)

Manually labeled rectangular regions around the person in the image have been used as the ground truth. Suppose gTruth_i is the ground truth in the i^{th} frame and track_i is the rectangular region output by a tracking algorithm, then the area overlap criterion is defined as follows

$$AO\left(\text{gTruth}_i, \text{track}_i\right) = \frac{Area\left(\text{gTruth}_i \cap \text{track}_i\right)}{AO\left(\text{gTruth}_i \cup \text{track}_i\right)} \tag{7.4}$$

The average area overlap can be computed for each data sequence as

$$AvgAOR = \frac{1}{N}\sum_{i=1}^{N} AO \tag{7.5}$$

Similar to [3], we use Object Tracking Error (OTE) which is the average distance between the centroid of the ground truth bounding box and the centroid of the result given by a tracking algorithm

$$OTE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(Centroid_{gTruth_i} - Centroid_{Truth_i}\right)} \tag{7.6}$$

In order to evaluate the performance of these algorithms using a single metric which encodes information from both area overlap and the distance between centroids, we have used a measure termed as the Tracking Evaluation Measure (TEM) which is the harmonic mean of the average area overlap fraction (AvgAOR) and a non-linear mapping of the Object tracking error (OTE).

$$TEM = 2 * \frac{AvgAOR.e^{-k.OTE}}{AvgAOR + e^{-k.OTE}} \tag{7.7}$$

where k is a constant which exponentially penalizes the cases where the distance between centroids is large.

7.6 Results

Particle Filtering has been widely used to handle complex scenarios by maintaining multiple hypotheses. As mentioned in [192], in order to handle abrupt motion changes, it is essential that the particles are widely spread while tracking. Following this principle, we have compared the performance of color particle filter (PF) [189] and the structured mode searching particle filter (SMSPF) by using a 2-D Gaussian with large variance as the system model. The position of the person and its scale have been included in the state vector. In order to compensate for the computational cost of structured search, only 50 particles were used for the SMSPF algorithm while 100 particles were used for the PF algorithm. A 10x10 grid with a sweep of 8 steps along the spatial dimension and 3 steps along the scale dimension were incorporated in the structured search.

Figure 7.17 and Figure 7.18 illustrate the comparison of the area overlap ratio and the distance between centroids at each frame of an example sequence. The sample frames are shown beside the tracking results. From Figure 7.17(a), it is evident that the SMSPF algorithm (red) shows a significant improvement over the color particle filter algorithm (green). Here, the area overlap ratio using SMSPF is much closer to 1 in most of the frames while the color particle filter drifts away causing this measure to be closer to 0. The distance between centroids measure also indicates a greater precision of the SMSPF algorithm as seen in Figure 7.18(a) where the distance between centroids using color particle filter is much higher than that with SMSPF(≈ 0).

Figure 7.19, Figure 7.20 and Figure 7.21 show the Tracking Evaluation Measure (TEM) for Datasets 1, 2 and 3. In majority of the cases, the SMSPF algorithm outperforms the color particle filtering algorithm with a higher TEM score.

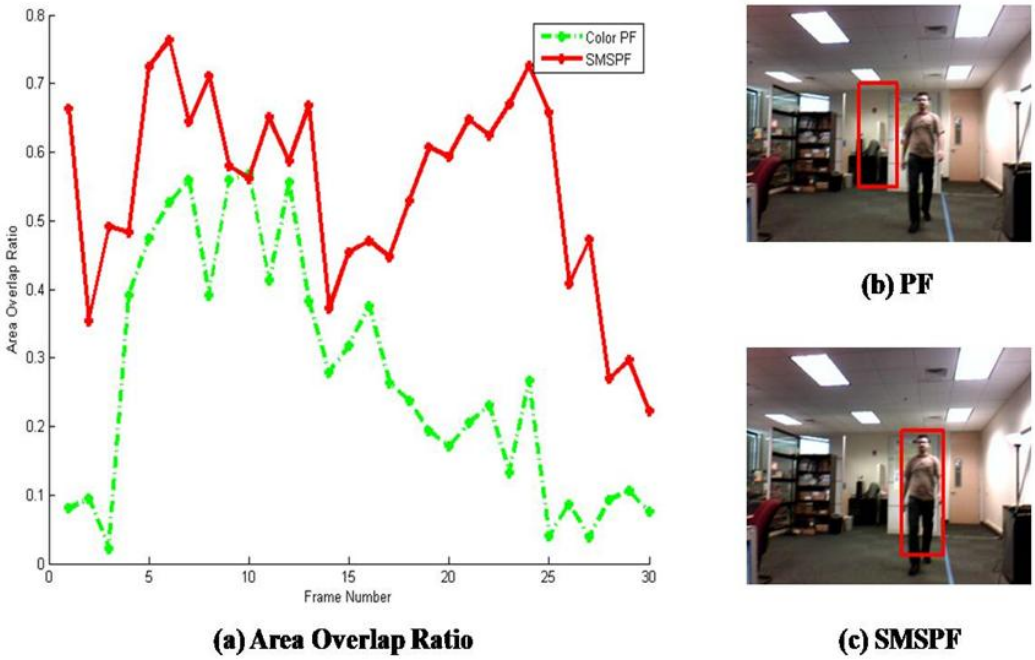


Figure 7.17: AO (Dotted Line: Color PF; Solid Line: SMSPF)

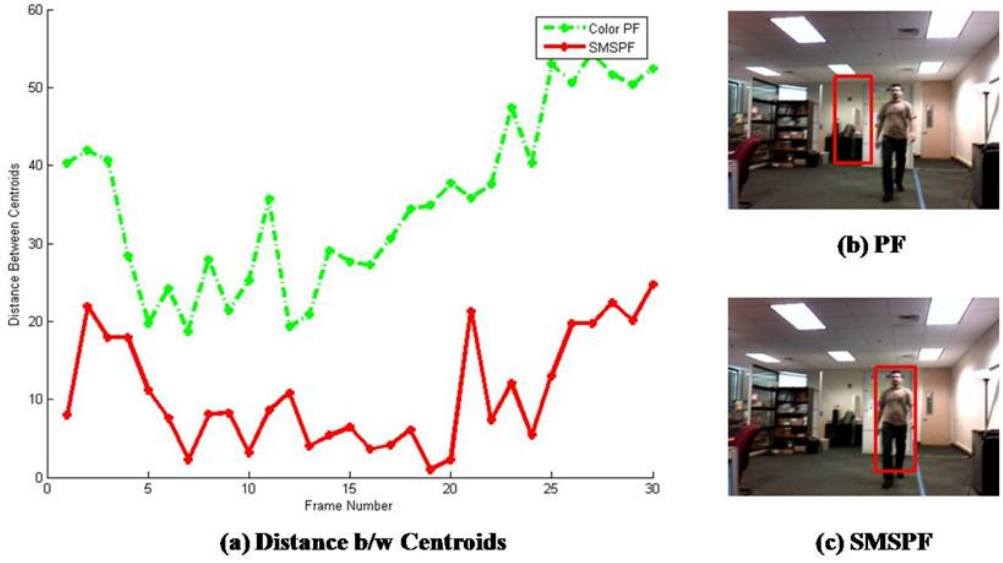


Figure 7.18: DC(Dotted Line: Color PF; Solid Line: SMSPF)

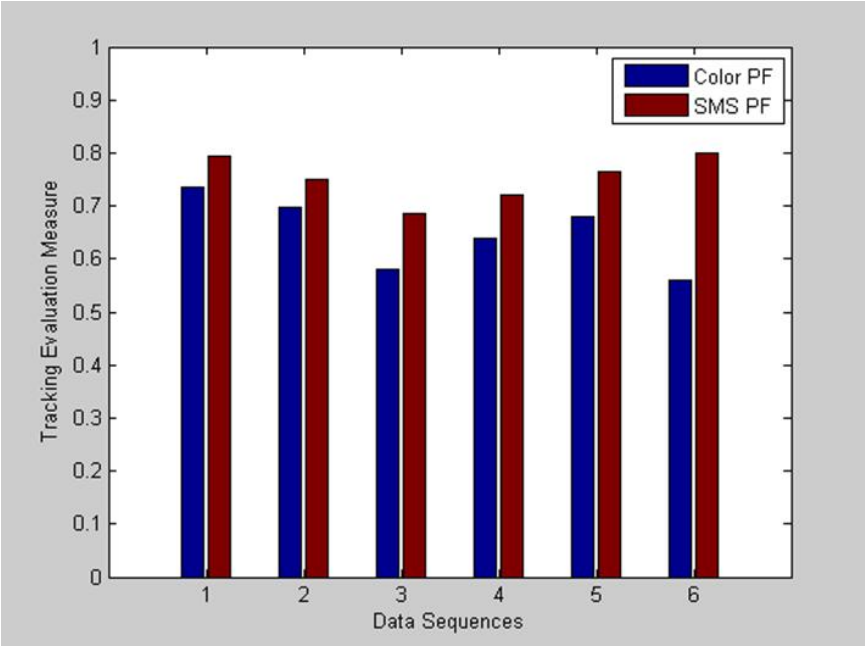


Figure 7.19: Evaluation Measure for DataSet 1

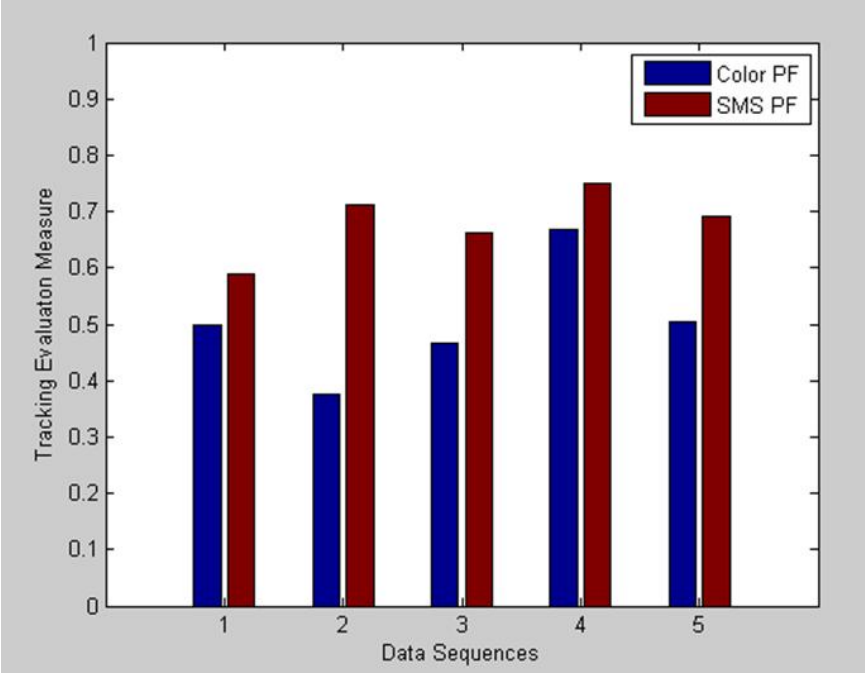


Figure 7.20: Evaluation Measure for DataSet 2

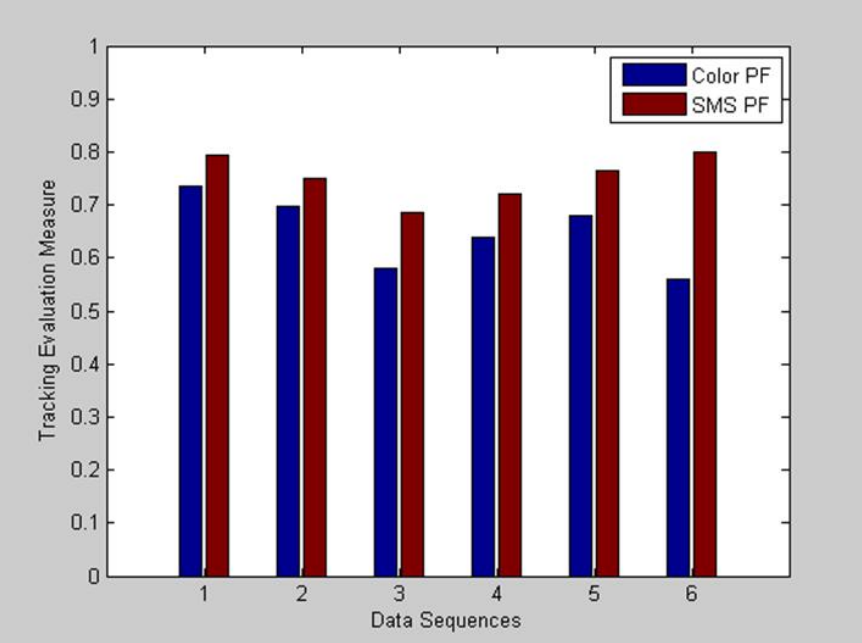


Figure 7.21: Evaluation Measure for DataSet 3

The results presented as a comparison between Color PF and SMSPF shows that incorporating a deterministic structured search into the stochastic particle filtering framework improves the person tracking performance in complex scenarios. The SMSPF algorithm strikes a balance between specificity and generality offered by detection and tracking algorithms as discussed in Section 2. It uses specific structure-aware features in the search in order to handle non-homogeneity of the object and the cluttered nature of the background. On the other hand, generality is maintained by using simple, global features in the particle filtering framework so as to handle non-rigidity and deformability of the object. The clear advantage of using the structured search can be observed on the complex Dataset 3 which encompasses most of the challenges generally encountered while using the Social Interaction Assistant.

REFERENCES

[1] C. Solomon, "The challenges of working in virtual teams: Virtual teams survey report 2010," tech. rep., RW3 CultureWizard, New York, NY, 2010.

[2] R. E. Riggio, "Assessment of basic social skills," *Journal of Personality and Social Psychology*, vol. 51, no. 3, pp. 649–660, 1986.

[3] B. D. Ruben, *Human communication handbook*. (Rochelle Park, N.J): Hayden Book Co., 1975.

[4] M. L. Knapp and J. A. Hall, *Nonverbal Communication in Human Interaction*. Harcourt College Pub, 4th ed., Nov. 1996.

[5] 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, no. 4, pp. 614, 599, 2004.

[6] R. Brown, *Social Psychology*. New York, NY: Free Press, 1986.

[7] J. Burgoon, D. Buller, J. Hale, and M. Turck, "Relational messages associated with nonverbal behaviors," *Human Communication Research*, vol. 10, no. 3, pp. 351–378, 1984.

[8] C. Wetzel, "The midas touch: The effects of interpersonal touch on restaurant tipping," *Personality and Social Psychology Bulletin*, vol. 10, no. 4, pp. 512–517, 1984.

[9] A. Haans and W. IJsselsteijn, "Mediated social touch: a review of current research and future directions," *Virtual Real.*, vol. 9, no. 2, pp. 149–159, 2006.

[10] J. Bailenson and N. Yee, "Virtual interpersonal touch: Haptic interaction and copresence in collaborative virtual environments," *Multimedia Tools and Applications*, vol. 37, pp. 5–14, Mar. 2008.

[11] A. J. Sameroff and M. J. Chandler, "Reproductive risk and the continuum of caretaker casualty," in *Review of Child Development Research* (F. D. Horowitz, ed.), vol. 4, Chicago: University of Chicago Press, 1975.

[12] U. Altmann, R. Hermkes, and L. Alisch, "Analysis of nonverbal involvement in dyadic interactions," in *Verbal and Nonverbal Communication Behaviours*, pp. 37–50, 2007.

[13] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," (Banff, Alberta, Canada), pp. 28–34, ACM, 2006.

[14] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proceedings of the 9th international conference on Multimodal interfaces*, (Nagoya, Aichi, Japan), pp. 271–278, ACM, 2007.

[15] J. Hawkins and S. Blakeslee, *On Intelligence*. Times Books, adapted ed., Oct. 2004.

[16] E. Rogers, W. Hart, and Y. Miike, "Edward t. hall and the history of intercultural communication: The united states and japan," *Keio Communication Review*, vol. 24, pp. 26, 3, 2002.

[17] O. Hargie, *Social Skills in Interpersonal Communication*. Routledge, 3 ed., June 1994.

[18] W. B. Walsh, K. H. Craik, and R. H. Price, *Person-environment psychology*. Routledge, 2000.

[19] 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, pp. 179–191, Mar. 1986.

[20] E. Krupat, *People in Cities: The Urban Environment and its Effects*. Cambridge University Press, Sept. 1985.

[21] R. Sommer, *Personal Space: The Behavioral Basis of Design*. Prentice Hall Trade, 6th printing ed., June 1969.

[22] R. Sommer, *Tight spaces; hard architecture and how to humanize it*. Prentice-Hall, 1974.

[23] A. Schauss, "The psysiological effect of color on the suppression of human aggression," *International Journal of Biosocial Research*, vol. 7, pp. 55–64, 1985.

[24] P. A. Bottomley and J. R. Doyle, "The interactive effects of colors and products on perceptions of brand logo appropriateness," *Marketing Theory*, vol. 6, pp. 63–83, Mar. 2006.

[25] T. Farrenkopf and V. Roth, "The university faculty office as an environment.," *Environment and Behavior*, vol. 12, pp. 467–77, Dec. 1980.

[26] R. H. Moos, *The Human Context: Environmental Determinants of Behavior*. Krieger Pub Co, June 1985.

[27] V. Manusov and J. H. Harvey, *Attribution, Communication Behavior, and Close Relationships*. Cambridge University Press, 1 ed., Jan. 2001.

[28] A. C. North, D. J. Hargreaves, and J. McKendrick, "In-store music affects product choice," *Nature*, vol. 390, p. 132, Nov. 1997.

[29] J. Meer, "The light touch," *Psychology Today*, vol. 19, pp. 60–67, 1985.

[30] 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, pp. 523–531, Oct. 1991.

[31] 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, pp. 58–63, Sept. 1977.

[32] H. H. Jennings, *Sociometry in group relations*. (Washington): American Council on Education, 105 p. ed., 1959.

[33] L. A. Zebrowitz, *Reading Faces*. Boulder CO: Westview Press, 1997.

[34] 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, pp. 3–18, July 1986. PMID: 3526376.

[35] J. B. Corts and F. M. Gatti, "Physique and self-description of temperament," *Journal of Consulting Psychology*, vol. 29, pp. 432–439, Oct. 1965. PMID: 5827516.

[36] L. A. Tucker, "Physical attractiveness, somatotype, and the male personality: A dynamic interactional perspective.," *Journal of Clinical Psychology*, vol. 40, no. 5, pp. 1226–34, 1984.

[37] C. Cameron, S. Oskamp, and W. Sparks, "Courtship american style: Newspaper ads," *The Family Coordinator*, vol. 26, pp. 27–30, Jan. 1977. ArticleType: primary_article / Full publication date: Jan., 1977 / Copyright 1977 National Council on Family Relations.

[38] 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, pp. 1728–1732, Oct. 2002.

[39] J. H. Griffin, R. Bonazzi, J. H. Griffin, and R. Bonazzi, *Black Like Me*. Signet, 35th anniversary ed., Nov. 1996.

[40] R. Porter, "Olfaction and human kin recognition," *Genetica*, vol. 104, pp. 259–263, Dec. 1998.

[41] T. Lord and M. Kasprzak, "Identification of self through olfaction.," *Perceptual and motor skills*, vol. 69, no. 1, pp. 224, 219, 1989.

[42] M. J. RUSSELL, "Human olfactory communication," *Nature*, vol. 260, pp. 520–522, Apr. 1976.

[43] N. Barber, "Mustache fashion covaries with a good marriage market for women," *Journal of Nonverbal Behavior*, vol. 25, pp. 261–272, Dec. 2001.

[44] W. E. Hensley, "The effects of attire, location, and sex on aiding behavior: A similarity explanation," *Journal of Nonverbal Behavior*, vol. 6, no. 1, pp. 3–11, 1981.

[45] N. Joseph, *Uniforms and Nonuniforms: Communication Through Clothing*. Greenwood Press, Nov. 1986.

[46] T. L. Rosenfeld and T. G. Plax, "Clothing as communication," *Journal of Communication*, vol. 27, pp. 24–31.

[47] C. Sanders and D. A. Vail, *Customizing the Body: The Art and Culture of Tattooing*. Temple University Press, Mar. 2008.

[48] P. Ekman, "Nonverbal communication: Movements with precise meanings," 1976.

[49] M. Wagner and N. Armstrong, *Field Guide to Gestures: How to Identify and Interpret Virtually Every Gesture Known to Man*. Quirk Books, July 2003.

[50] D. Efron, *Gesture, Race and Culture*. Walter de Gruyter, Inc., Oct. 1972.

[51] G. E. Weisfeld and J. M. Beresford, "Erectness of posture as an indicator of dominance or success in humans," *Motivation and Emotion*, vol. 6, pp. 113–131, June 1982.

[52] E. C. Grant and J. H. Mackintosh, "A comparison of the social postures of some common laboratory rodents," *Behaviour*, vol. 21, no. 3/4, pp. 246–259, 1963. ArticleType: primary_article / Full publication date: 1963 / Copyright 1963 BRILL.

[53] A. Kleinsmith, P. R. D. Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers*, vol. 18, pp. 1371–1389, Dec. 2006.

[54] A. Montagu, *Touching: The Human Significance of the Skin*. Harper Paperbacks, 3 ed., Sept. 1986.

[55] 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, pp. 9–38, Feb. 1999.

[56] M. J. Hertenstein, J. M. Verkamp, A. M. Kerestes, and R. M. Holmes, "The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research," *Genetic, Social, and General Psychology Monographs*, vol. 132, pp. 5–94, Feb. 2006. PMID: 17345871.

[57] M. J. Hertenstein, D. Keltner, B. App, A. B. Bulleit, and R. Jaskolta, "Touch communicates distinct emotions," *Emotion*, vol. 6, no. 3, pp. 528–533, 2006.

[58] G. Robles-De-La-Torre, "Principles of haptic perception in virtual environments," in *Human Haptic Perception: Basics and Applications*, pp. 363–379, 2008.

[59] L. J. Carver and G. Dawson, "Development and neural bases of face recognition in autism," *Molecular Psychiatry*, vol. 7, no. s2, pp. S18–S20, 2002.

[60] W. E. Rinn, "The neuropsychology of facial expression: A review of neurological and psychological mechanisms for producing facial expressions," *Psychological Bulletin*, vol. 95, pp. 52–77, 1984.

[61] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, 1978.

[62] C. E. Izard, *The maximally discriminative facial movement coding system*. Instructional Resources Center, University of Delaware, revised ed., 1983.

[63] M. Argyle and M. Cook, *Gaze & Mutual Gaze*. Cambridge University Press, Jan. 1976.

[64] C. L. Kleinke, "Gaze and eye contact: a research review," *Psychological Bulletin*, vol. 100, pp. 78–100, July 1986. PMID: 3526377.

[65] A. Kendon, "Some functions of gaze-direction in social interaction.," *Acta Psychol (Amst)*, vol. 26, no. 1, pp. 63, 22, 1967.

[66] M. S. Mast, "Dominance as expressed and inferred through speaking time," *Human Communication Research*, vol. 28, no. 3, pp. 420–450, 2002.

[67] J. B. Bavelas, L. Coates, and T. Johnson, "Listener responses as a collaborative process: The role of gaze," *The Journal of Communication*, vol. 52, no. 3, pp. 566–580, 2002.

[68] 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, no. 3, pp. 227–37, 1997.

[69] A. M. Glenberg, J. L. Schroeder, and D. A. Robertson, "Averting the gaze disengages the environment and facilitates remembering," *Memory & Cognition*, vol. 26, pp. 651–658, July 1998. PMID: 9701957.

[70] J. Orozco, O. Rudovic, F. Roca, and J. Gonzalez, "Confidence assessment on eyelid and eyebrow expression recognition," in *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pp. 1–8, 2008.

[71] C. Segrin and J. Flora, "Poor social skills are a vulnerability factor in the development of psychosocial problems.," *Human Communication Research*, vol. 26, no. 3, pp. 489–514, 2000.

[72] 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 & Blindness*, vol. 98, pp. 470–483, 2004.

[73] D. Jindal-Snape, "Use of feedback from sighted peers in promoting social interaction skills," *Journal of Visual Impairment and Blindness*, vol. 99, pp. 1–16, July 2005.

[74] 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, pp. 362–366, 1998.

[75] K. Shinohara and J. Tenenberg, "A blind person's interactions with technology," *Commun. ACM*, vol. 52, no. 8, pp. 58–66, 2009.

[76] 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*, vol. 31, no. 1, pp. 39–58, 2009.

[77] S. ur Rehman, L. Liu, and H. Li, "Manifold of facial expressions for tactile perception," pp. 239–242, 2007.

[78] A. Teeters, R. Kaliouby, and R. Picard, "Self-Cam: feedback from what would be your social partner," in *SIGGRAPH '06: ACM SIGGRAPH 2006 Research posters*, (Boston, Massachusetts), p. 138, ACM, 2006.

[79] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signals, their function, and automatic analysis: a survey," in *Proceedings of the 10th international conference on Multimodal interfaces*, (Chania, Crete, Greece), pp. 61–68, ACM, 2008.

[80] T. Kim, A. Chang, L. Holland, and A. Pentland, "Meeting mediator: Enhancing group collaboration and leadership with sociometric feedback," (San Diego, CA, USA), pp. 457–466, 2008.

[81] A. Pentland, *Honest Signals: How They Shape Our World*. The MIT Press, Oct. 2008.

[82] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: state-of-the-art and future perspectives of an emerging domain," in *Proceeding of the 16th ACM international conference on Multimedia*, (Vancouver, British Columbia, Canada), pp. 1061–1070, ACM, 2008.

[83] R. E. Transon, "Using the feedback band device to control rocking behavior," *Journal of Visual Impairment & Blindness*, vol. 82, pp. 287 – 289, 1988.

[84] J. N. Felps and R. J. Devlin, "Modification of stereotypic rocking of a blind adult.," *Journal of Visual Impairment and Blindness*, vol. 82, no. 3, pp. 107–08, 1988.

[85] T. J. Thompson, S. M. Pearcey, J. W. Bodfish, T. W. Crawford, and M. H. Lewis, "Stereotyped movement disorder in an adult following acquired brain injury: effect of environmental stimulation," *Behavioral Interventions*, vol. 10, pp. 79–85, Apr. 1995.

[86] J. Jankovic, "Stereotypies," in *Moverment Disorders* (C. D. Marsden and S. Fahn, eds.), vol. 3, pp. 503–517, London: Butterworth-Heinemann, 1994.

[87] 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, no. 4, p. 163, 1993.

[88] R. S. Reivich and I. A. Rothrock, "Behavior problems of deaf children and adolescents: A Factor-Analytic study," *J Speech Hear Res*, vol. 15, pp. 93–104, Mar. 1972.

[89] E. Haag, W. Huber, R. Hndgen, U. Stiller, and K. Willmes, "Repetitive verbal behavior in severe aphasia," *Der Nervenarzt*, vol. 56, pp. 543–52, Oct. 1985. PMID: 2415840.

[90] 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, pp. 279–286, Oct. 2001.

[91] R. L. Loftin, S. L. Odom, and J. F. Lantz, "Social interaction and repetitive motor behaviors.," *Journal of Autism & Developmental Disorders*, vol. 38, pp. 1124–1135, July 2008.

[92] G. Yu, Y. Zhang, and R. Yan, "Loneliness, peer acceptance, and family functioning of chinese children with learning disabilities: Characteristics and relationships," *Psychology in the Schools*, vol. 42, no. 3, pp. 325–331, 2005.

[93] V. J. Eichel, "A taxonomy for mannerisms of blind children.," *Journal of Visual Impairment and Blindness*, vol. 73, pp. 167–78, May 1979.

[94] B. B. Blasch, "Blindisms: Treatment by punishment and reward in laboratory and natural settings," *Journal of Visual Impairment & Blindness*, pp. 215–230, 1972.

[95] 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, no. 7, pp. 307–10, 1984.

[96] R. L. Ohlsen, "Control of body rocking in the blind through the use of vigorous exercise," *Journal of Instructional Psychology*, vol. 5, pp. 19–22, 1978.

[97] A. H. Estevis and A. J. Koenig, "A cognitive approach to reducing stereotypic body rocking.," *Re:View*, vol. 26, no. 3, p. 119, 1994.

[98] P. J. Schloss and M. A. Smith, "Increasing appropriate behavior through related personal characteristics," in *Applied Behavior Analysis in the Classroom*, Boston: Allyn & Bacon, 1994.

[99] G. Cartledge, *Teaching Social Skills to Children: Innovative Approaches*. Allyn & Bacon, 2 ed., June 1986.

[100] S. Raver and P. W. Darsh, "Increasing social skills training for visually impaired children," *Education of the Visually Handicapped*, vol. 19, pp. 147–155, 1988.

[101] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *Workshop on Computer Vision Applications for the Visually Impaired (CVAVI 08), European Conference on Computer Vision ECCV 2008*, (Marseille, France), Oct. 2008.

[102] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," pp. 1–17, 2004.

[103] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data," *American Association for Artificial Intelligence*, 2005.

[104] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 ed., Mar. 2000.

[105] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[106] O. Amft, H. Junker, and G. Troster, "Detection of eating and drinking arm gestures using inertial body-worn sensors," pp. 160–163, 2005.

[107] G. Chambers, S. Venkatesh, G. West, and H. Bui, "Hierarchical recognition of intentional human gestures for sports video annotation," vol. 2, pp. 1082–1085 vol.2, 2002.

[108] S. Lee and K. Mase, "Activity and location recognition using wearable sensors," *Pervasive Computing, IEEE*, vol. 1, no. 3, pp. 24–32, 2002.

[109] F. Foerster, M. Smeja, and J. Fahrenberg, "Detection of posture and motion by accelerometry: a validation in amulatory monitoring," *Computer in Human Behavior*, vol. 15, pp. 571–583, 1999.

[110] 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), pp. 243–244, ACM, 2008.

[111] N. Krishnan and S. Panchanathan, "Analysis of low resolution accelerometer data for continuous human activity recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3337–3340, 2008.

[112] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[113] A. Vezhnevets and V. Vezhnevets, "Modest AdaBoost - teaching AdaBoost to generalize better," (Novosibirsk Akademgorodok, Russia), 2005.

[114] "DRM103 designer reference manual," Tech. Rep. DRM103 Rev. 1, Freescale Semiconductor, Aug. 2008.

[115] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets Syst.*, vol. 69, no. 2, pp. 125–139, 1995.

[116] Y. Benjamini, "Opening the box of a boxplot," *The American Statistician*, vol. 42, pp. 257–262, Nov. 1988. ArticleType: primary_article / Full publication date: Nov., 1988 / Copyright 1988 American Statistical Association.

[117] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[118] 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, pp. 279 – 88, May 1999.

[119] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*, (Corvalis, Oregon), pp. 759–766, ACM, 2007.

[120] A. Schwaninger, C. C. Carbon, and H. Leder, "Expert face processing: Specilization and constraints," 2003.

[121] V. Bruce and A. Young, *In the Eye of the Beholder: The Science of Face Perception*. Oxford University Press, 2006.

[122] J. Sadr, I. Jarudi, and P. Shinha, "The role of eyebrows in face recognition," *Perception*, vol. 32, pp. 285–93, 2003.

[123] A. W. Young, D. Hellawell, and D. C. Hay, "Configurational information in face perception," *Perception*, vol. 16, pp. 747–59, 1987.

[124] P. Sinha, B. J. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: 19 results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–62, 2006.

[125] H. McConachie, "Developmental prosopagnosia: A single case report," *Cortex*, vol. 12, pp. 76–82, 1976.

[126] I. Kennerknecht, T. Gruter, B. Welling, S. Wentzek, J. Horst, S. Edwards, and M. Gruter, "First report of prevalence of non-syndromic hereditary prosopagnosia (hpa)," *American Journal of Medical Genetics, Part A*, vol. 140, pp. 1617–22, 2006.

[127] D. J. Turk, T. C. Handy, and M. S. Gazzaniga, "Can perceptual expertise account for the own-race bias in face recognition? a split brain study," *Cognitive Neuropsychology*, vol. 22, no. 7, pp. 877–83, 2005.

[128] B. Gkberk, M. O. Irfanoglu, L. Akarun, and E. Alpaydin, "Learning the best subset of local features for face recognition," *Pattern Recognition*, vol. 40, no. 5, p. 1520, 2007.

[129] L. Wiskott, "Phantom faces for face analysis," *Pattern Recognition*, vol. 30, no. 6, p. 837, 1997.

[130] N. Kruger, M. Potzsch, and C. Malsburg, "Determination of face position and pose with a learned representation based on labelled graphs," *Image Vision Computing*, vol. 15, p. 665, 1997.

[131] P. Kalocsai, C. Malsburg, and J. Horn, "Face recognition by statistical analysis of feature detectors," *Image Vision Computing*, vol. 18, no. 4, p. 273, 2000.

[132] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, p. 735, 2001.

[133] D. H. Liu, K. M. Lam, and L. S. Shen, "Optimal sampling of gabor features for face recognition," *Pattern Recognition Letters*, vol. 25, no. 2, p. 267, 2004.

[134] P. Yang, S. Shan, W. Gao, S. Li, and D. Zang, "Face recognition using ada-boosted gabor features," in *Proceedings of the 16th International Conference on Face and Gesture Recognition*, 2004.

[135] X. Wang and H. Oi, "Face recognition using optimal non-orthogonal wavelet basis evaluated by information complexity," in *Proceedings of the 16th International Conference on Pattern Recognition*, p. 164, 2002.

[136] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," (Heidelberg), pp. 456–463, Springer-Verlag, 1997.

[137] C. von der Malsburg, "Nervous structures with dynamical links," *Ber. Bunsenges. Phys. Chem.*, vol. 89, pp. 703–710, 1985.

[138] E. Bienenstock and christoph von der Malsburg, "A neural network for invariant pattern recognition," *Europhysics Letters*, vol. 4, pp. 121–126, 1987.

[139] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, *Face Recognition and Gender Determination*. 1995. Published: International Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[140] L. Wiskott and C. von der Malsburg, *Face Recognition by Dynamic Link Matching*. http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96/: The UTCS Neural Networks Research Group, Austin, TX, 1996.

[141] M. Lyons, M. Lyons, J. Budynek, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 1357–1362, 1999.

[142] Y. Liu and Chongqing, "Face recognition using kernel principal component analysis and genetic algorithms," in *12th IEEE Workshop on Neural Networks for Signal Processing*, p. 337, Sep 2002 2002.

[143] Y. Xu, B. Li, and B. Wang, "Face recognition by fast independent component analysis and genetic algorithm," in *Fourth International Conference on Computer and Information*, p. 194, 14-16 Sep 2004 2004.

[144] K. Wong and K. Lam, "A reliable approach for human face detection using genetic algorithm," in *IEEE International Symposium on Circuits and Systems*, vol. 4, p. 499, 1999.

[145] S. Karungaru, M. Fukumi, and N. Akamatsu, "Face recognition using genetic algorithm based template matching," in *International Symposium on Communications and Information Technologies*, October 26- 29,2004 2004.

[146] D. Ozkan, "Feature selection for face recognition using a genetic algorithm," tech. rep., 2006.

[147] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 20, pp. 91–110, 2003.

[148] J. Huang and H. Wechsler, "Eye location using genetic algorithm," in *2nd International Conference on Audio and Video-Based Biometric Person Authentication*, 1999.

[149] Y. Sun and L. Yin, "A genetic algorithm based feature selection approach for 3d face recognition," in *Biometrics Consortium Conference*, September 19-21, 2005 2005.

[150] Z. Sun, X. Yuan, G. Bebis, and S. Louis, "Neural-network-based gender classification using genetic eigen-feature extraction," 2002.

[151] J. Black, M. Gargesha, K. Kahol, and S. Panchanathan, "A framework for performance evaluation of face recognition algorithms," *ITCOM, Internet Multimedia Systems II, Boston*, July 2002.

[152] G. Little, S. Krishna, J. Black, and S. Panchanathan, "A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose and illumination angle," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Philadelphia, USA), pp. 89–92, 2005.

[153] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, vol. 4, p. 519, 1987.

[154] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

[155] S. Krishna, G. Little, J. Black, and S. Panchanathan, "A wearable face recognition system for individuals with visual impairments," in *Assets '05: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, (New York, NY, USA), pp. 106–113, ACM Press, 2005.

[156] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision - to appear*, 2002.

[157] V. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: A framework for person-independent head pose estimation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07)*, (Minneapolis, USA), June 2007.

[158] E. T. Hall, *The Hidden Dimension*. Anchor, Oct. 1990.

[159] S. Ram and J. Sharf, "The people sensor: a mobility aid for the visually impaired," pp. 166–167, 1998.

[160] 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, no. 2, pp. 106–117, 2005.

[161] P. Barralon, G. Ng, G. Dumont, S. K. W. Schwarz, and M. Ansermino, "Development and evaluation of multidimensional tactons for a wearable tactile display," in *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*, (Singapore), pp. 186–189, ACM, 2007.

[162] L. Brown, S. Brewster, and H. Purchase, "A first investigation into the effectiveness of tactons," in *Eurohaptics Conference, 2005 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2005. World Haptics 2005. First Joint*, pp. 167–176, 2005.

[163] S. Brewster and L. Brown, "Tactons: structured tactile messages for non-visual information display," in *AUIC '04: Proceedings of the fifth conference on Australasian user interface*, pp. 23, 15, Australian Computer Society, Inc., 2004.

[164] P. Viola and M. J. Jones, "Robust Real-Time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[165] M. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[166] A. Hadid and M. Pietikainen, "A hybrid approach to face detection under unconstrained environments," *18th International Conference on Pattern Recognition*, vol. 1, pp. 227–230, 2006.

[167] I. Naseem and M. Deriche, "Robust human face detection in complex color images," *IEEE International Conference on Image Processing*, vol. 2, pp. 338–41, 2005.

[168] M. Wimmer, B. Radig, and M. Beetz, "A person and context specific approach for skin color classification," *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 2, pp. 39–42, 2006.

[169] M. B. Hmid and Y. B. Jemaa, "Fuzzy classification, image segmentation and shape analysis for human face detection," *8th International Conference on Signal Processing*, vol. 4, 2006.

[170] U. Tariq, H. Jamal, M. Shahid, and M. Malik, "Face detection in color images, a robust and fast statistical approach," *Proceedings of INMIC 2004. 8th International Multitopic Conference*, pp. 73–78, 2004.

[171] Y.-W. Wu and X.-Y. Ai, "Face detection in color images using adaboost algorithm based on skin color information," *International Workshop on Knowledge Discovery and Data Mining*, pp. 339–342, 2008.

[172] K. Sentz and S. Ferson, "Combination of evidence in dempster-shafer theory," tech. rep., Sandia National Laboratories, 2002.

[173] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi, "The feret evaluation methodology for face-recognition algorithms," *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, p. 137, 1997.

[174] J. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," (Berkeley CA), International Computer Science Institute, U.C. Berkeley, April, 1998.

[175] P. Perez, "Markov random fields and images," *CWI Quaterly*, vol. 11, pp. 413–437, 1998.

[176] M. Vezjak and M. Stephancic, "An anthropological model for automatic recognition of the male human face," *Annals of Human Biology*, vol. 21, pp. 363–380, 1994.

[177] R. Paget, I. D. Longstaff, and B. Lovell, "Texture classification using nonparametric markov random fields," vol. 1, pp. 67–70, 1997.

[178] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," *IN ECCV WORKSHOP ON STATISTICAL LEARNING IN COMPUTER VISION*, pp. 17—32, 2004.

[179] F. Porikli and O. Tuzel, "Object tracking in Low-Frame-Rate video," *SPIE Image and Video Communications and Processing*, vol. 5685, pp. 72–79, Mar. 2005.

[180] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, 2007.

[181] J. Kwon and K. M. Lee, "Tracking of abrupt motion using wang-landau monte carlo estimation," in *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, (Berlin, Heidelberg), pp. 387–400, Springer-Verlag, 2008.

[182] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IN CVPR*, vol. 1, pp. 886—893, 2005.

[183] K. Nummiaro, E. Koller-Meier, and L. V. Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, pp. 99–110, Jan. 2003.

[184] F. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," *IN PROC. IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION*, vol. 1, pp. 829—836, 2005.

[185] Q. Zhu, Q. Zhu, S. Avidan, S. Avidan, M. chen Yeh, M. chen Yeh, K. ting Cheng, and K. ting Cheng, "Fast human detection using a cascade of histograms of oriented gradients," *IN CVPR06*, vol. 2006, pp. 1491—1498, 2006.

[186] F. Porikli, P. Meer, O. Tuzel, and O. Tuzel, "P.: Human detection via classification on riemannian manifolds," *IN PROC. OF THE IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION*, pp. 1—8, 2007.

[187] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, pp. II–264–II–271 vol.2, 2003.

[188] C. Yang, R. Duraiswami, and L. Davis, "Fast multiple object tracking via a hierarchical particle filter," *IN: INTERNATIONAL CONFERENCE ON COMPUTER VISION*, vol. 1, pp. 212—219, 2005.

[189] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," 2006.

[190] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," *IN CVPR*, vol. 1, pp. 878—885, 2005.

[191] M. S. Arulampalam, S. Maskell, and N. Gordon, "A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 50, pp. 174—188, 2002.

[192] M. Isard and A. Blake, "CONDENSATION - conditional density propagation for visual tracking," *INTERNATIONAL JOURNAL OF COMPUTER VISION*, vol. 29, pp. 5—28, 1998.

[193] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 232—237, 1998.

[194] K. Okuma, A. Taleghani, N. D. Freitas, O. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," *IN ECCV*, vol. 1, pp. 28—39, 2004.

[195] F. Crow, "Summed-area tables for texture mapping," in *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pp. 207–212, ACM, 1984.

Appendix A

ALGORITHM FOR ESTIMATING RANK AVERAGE OF GROUPS

While analyzing the responses of participants to the online survey, the participants responses for each question are represented as entries $x_{i,q}$, where, i represents the i^{th} participant and q represents the q^{th} question. $i = 1,\ldots,N$ are the N participants who responded on the survey, and $q = 1,\ldots,Q$ are the Q questions. In the survey presented in Chapter XXX, $N = 28$ and $Q = 8$.

Procedure

Input: Each participants response is considered as an entry e_m into a pool $E = \{x_{i,q}\}$, where, $m = 1,\ldots,M$, and $M = N\text{x}Q$.

Ouput: The rank average for the Q groups (questions), $\bar{R}_m$.

Steps:

1. Group $e_n \in E$ removing all group affiliations.
2. Order the entries from 1 to M and assign a rank r_{iq}.
3. Assign any tied values the average of the ranks they would have received had they not been tied.
4. Rank Average for each group is then given as

$$\bar{R}_m = \frac{\sum\limits_{i \in Q_m, q=m} r_{iq}}{n_m} \tag{A.1}$$

   Where, Q_m represents the group m with the cardinality n_m.

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 groups.

This LaTeX document was generated using the Graduate College Format Advising tool. Please turn a copy of this page in when you submit your document to Graduate College format advising. You may discard this page once you have printed your final document. DO NOT TURN THIS PAGE IN WITH YOUR FINAL DOCUMENT!