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.

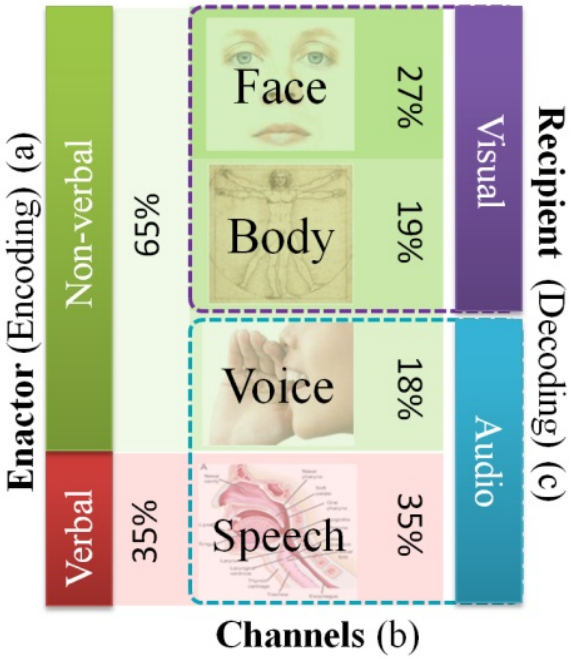


Figure 1.1: Relative importance of a) verbal vs. non-verbal cues, b) four channels of non-verbal cues, and c) visual vs. audio encoding and decoding of bilateral human interpersonal communicative cues.

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

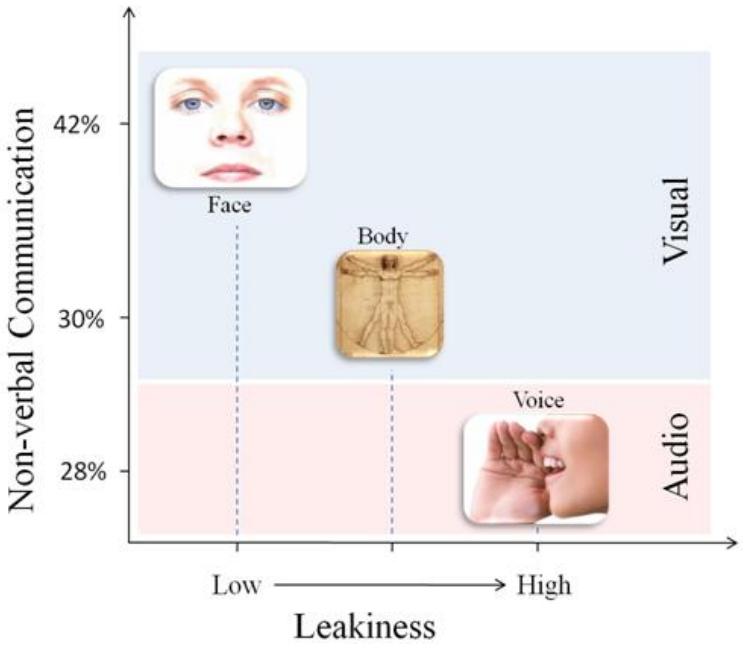


Figure 1.2: Relative communicative information plotted against its leakiness. Speech forms the verbal channel. Face, body and voice form the non-verbal communication channels.

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

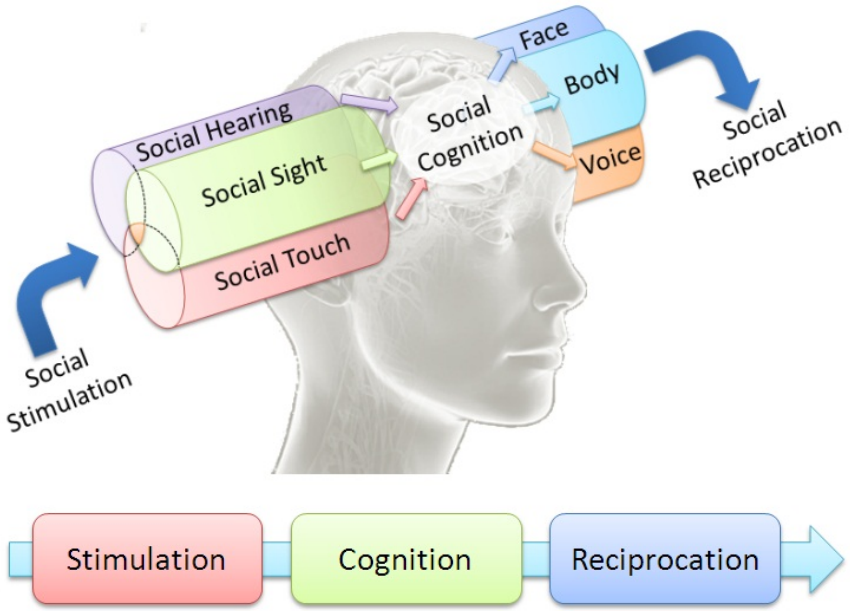


Figure 1.3: Social Situational Awareness.

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].

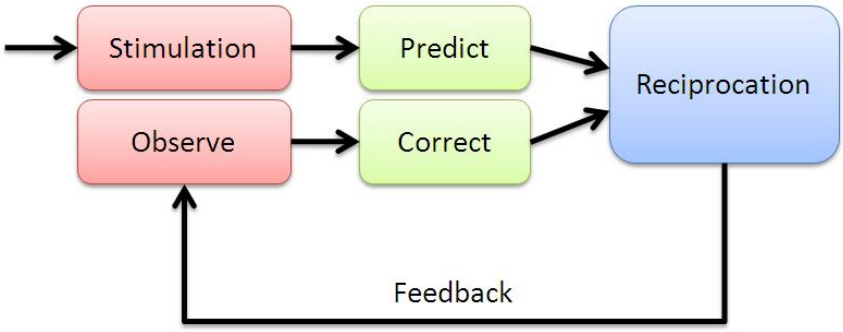


Figure 1.4: Social learning systems with continuous learning feedback loop.

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

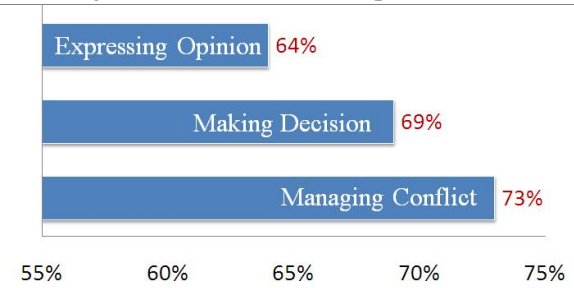
An industry survey [1] of 1592 individuals who collaborated remotely, carried out by RW3 CultureWizard - a company focused on improving international collaborations - reported difficulties similar to what was faced by the individuals who are blind. "Respondents found virtual teams more challenging than 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.

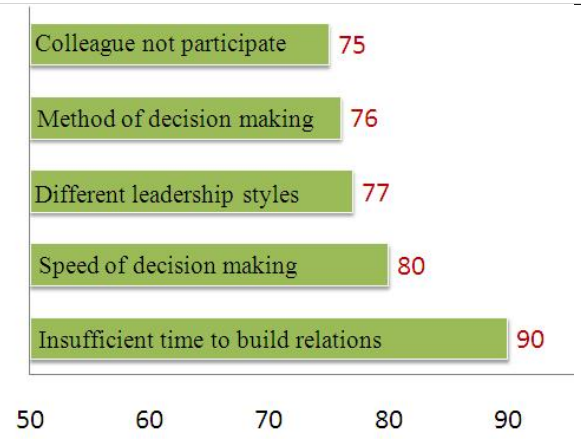
2.3 Medical Teams

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

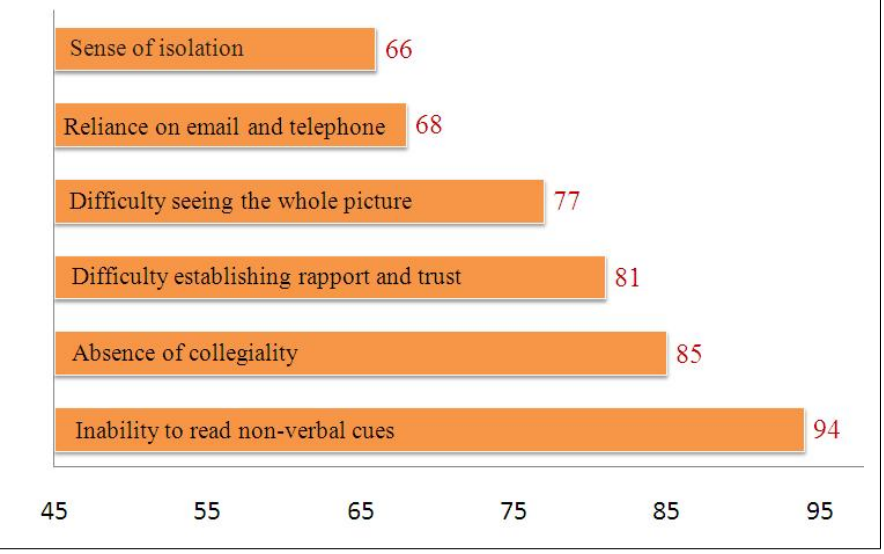
Challenges in virtual teams compared to face-to-face teams



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

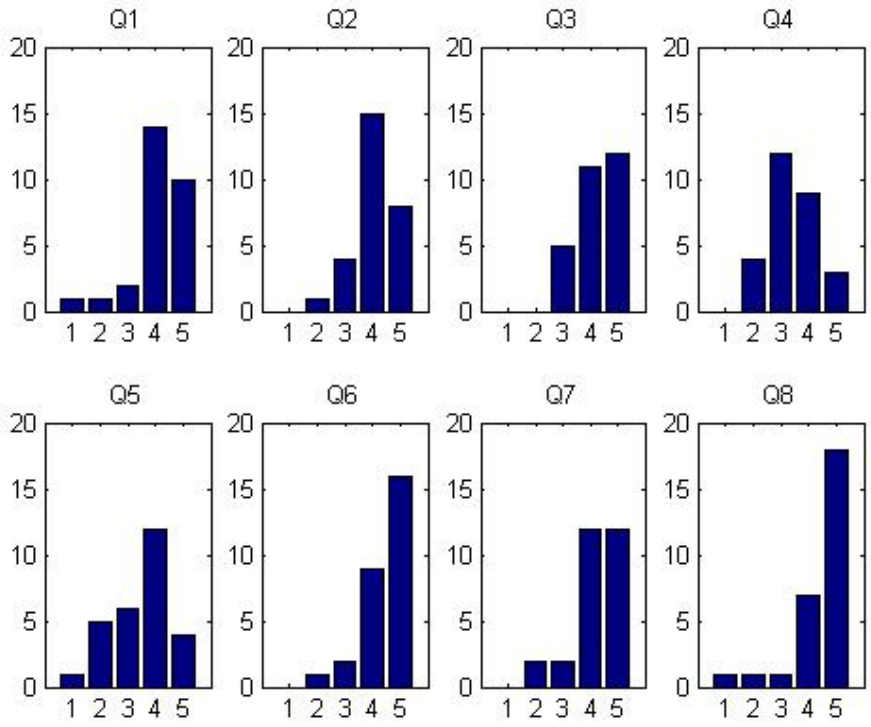


Figure 3.1: Histogram of Responses grouped by Questions

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

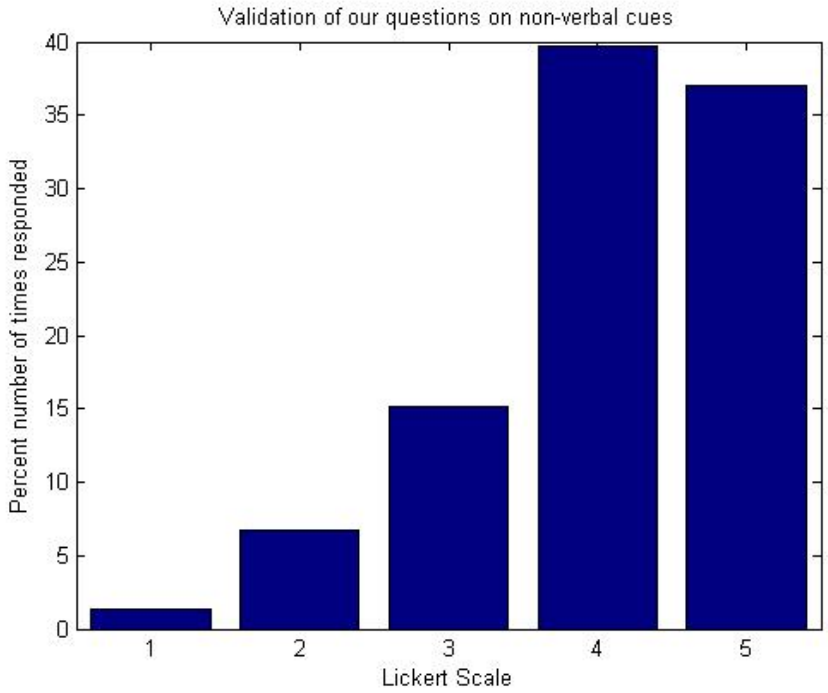


Figure 3.2: Response Ratio

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

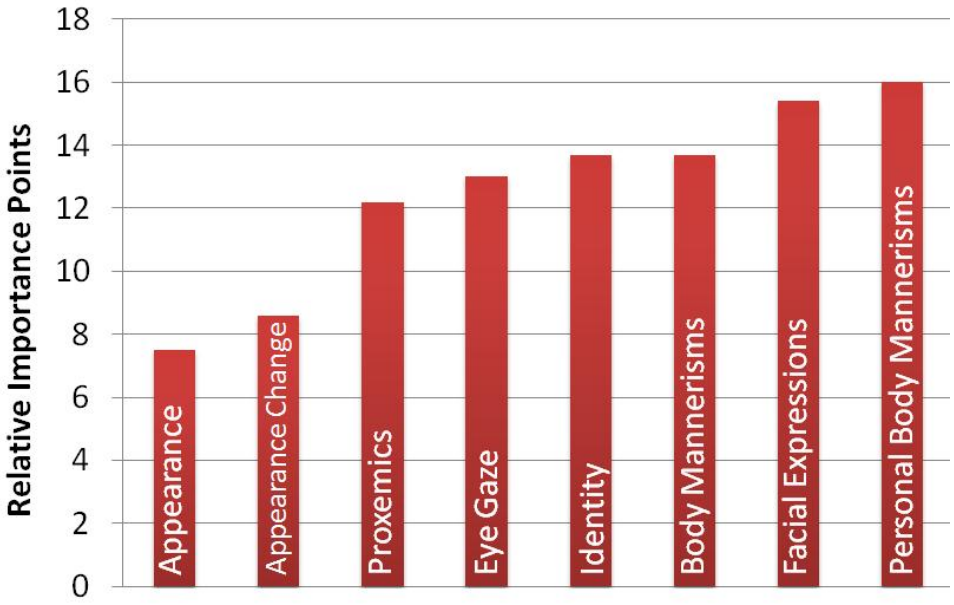


Figure 3.3: Rank average of the 8 questions

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

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 [85]. 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 [86], 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 [87] 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 [88] [89], which are vibratory signals used to communicate abstract messages [89] by changing the dimensions of the signal including frequency, amplitude, location, rhythm, etc. Based on pilot test results, we chose to pursue distinct rhythms over monotonic rhythms as users find it difficult to identify interpersonal distances using monotonic rhythms as the vibratory signal varies smoothly with changes in distance.

From the sensing perspective we resort to the camera that is on the user's glasses and through the use of computer vision technology, face detection, we extract non-verbal cues for social interaction, including the number of people in the user's visual field, where people are located relative to the user, coarse information related to gaze direction (pose estimation algorithms could be used to extract finer estimates of pose), and the approximate distance of the person from the user based on the size of the face image.

4.1 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 [90] 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. **********

4.2 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 [91] 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 [92].

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

Figure 4.1: 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 4.2 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.

4.3 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 [93], [94] and [95], and b) High level facial feature models such as [96], [97] and [98].

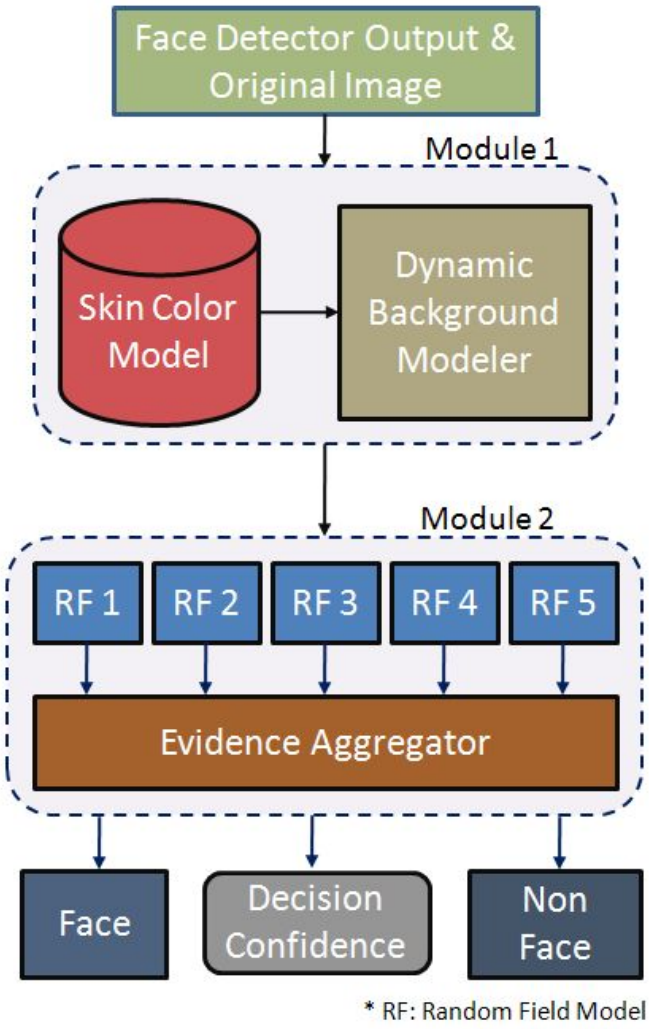


Figure 4.2: Block diagram.

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 [92] and [**?**]. The framework proposed in this paper uses statistically learnt knowledge about human faces to overcome computational complex-

ity thereby augmenting face validation to existing face detection algorithms seamlessly.

4.4 Proposed Framework

As shown in Figure 4.2, 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 [99]. 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:

$$X^{nRGB}_{i|i\in\{r,g,b\}} = \frac{X_i}{\left(\sum\limits_{\forall_i|i\in\{r,g,b\}} X_i\right)} \tag{4.1}$$

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:

$$X^{nRGB}_{i|i\in\{nR,nG,nB\}} = 1 - \left(\sum_{\forall k|(k\in\{nR,nG,nB\},k\neq i)} X_k \right) \tag{4.2}$$

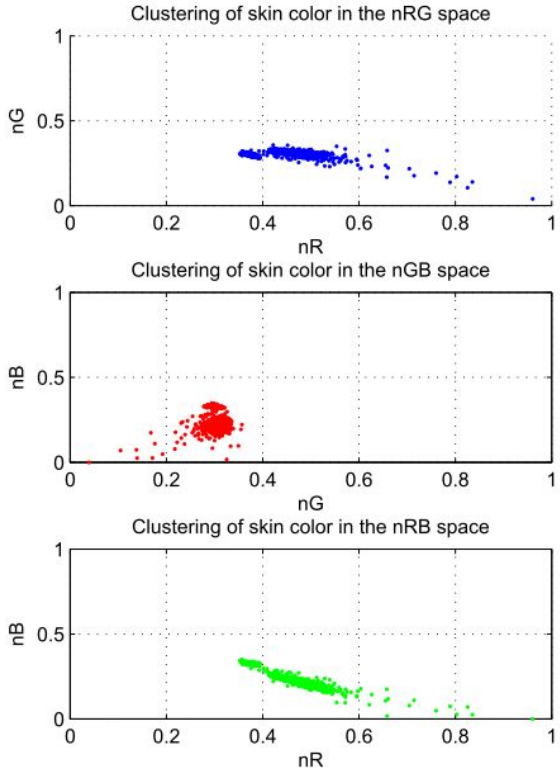


Figure 4.3: Skin pixels in nRGB space.

In our experiments, we found that skin pixels form a tight cluster when projected on nG and nB space as shown in the Figure 4.3. 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 [100]. 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 [101]. The Bi-modal Gaussian mixture model is represented as.

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

$$w_2 f_{Y_2}(x; \Theta_2 = [\mu_2, \Sigma_2]) \tag{4.3}$$

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 4.4. 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.

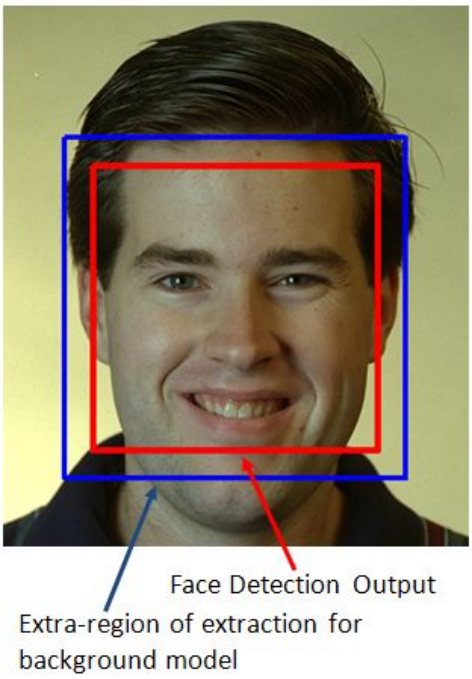


Figure 4.4: Extra region for background modeling.

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.

$$f^{non-skin}_{X|X=[R,G,B]}(x) = \sum_{i=1}^{m} w(i) f_{Y_i}(x; \Theta_i = [\mu_i, \Sigma_i]) \tag{4.4}$$

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 4.4. Example skin-masks are shown in Figure 4.5. This example shows two sets of images - one corresponding to a *t*rue face detection result, and another *f*alse face detection result.

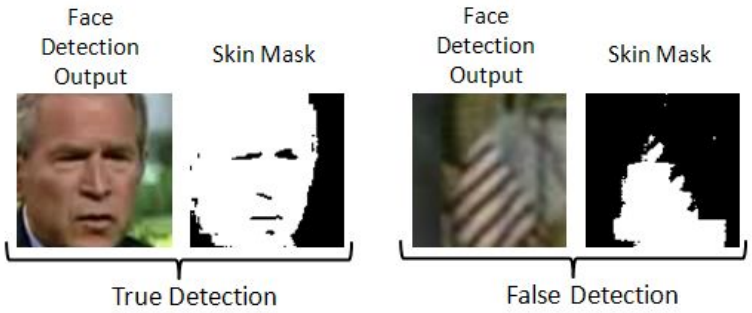


Figure 4.5: Example of *true* and *false* face detection.

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 4.5.

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

In order to validate the skin region extracted as explained in Section 4.4, 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 [102].

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,

$$P(\mathbf{x}) > 0, \forall \mathbf{x} \in \mathscr{X} \tag{4.5}$$

$$P(x_i | x_{\mathscr{S}-\{i\}}) = P(x_i | x_{\mathscr{N}_i}) \tag{4.6}$$

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 4.6 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

$$P^k(x_i | x_{\mathscr{S}-\{i\}}) = P(x_i | x_{\mathscr{N}_i^k}) \tag{4.7}$$

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 4.6. 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.

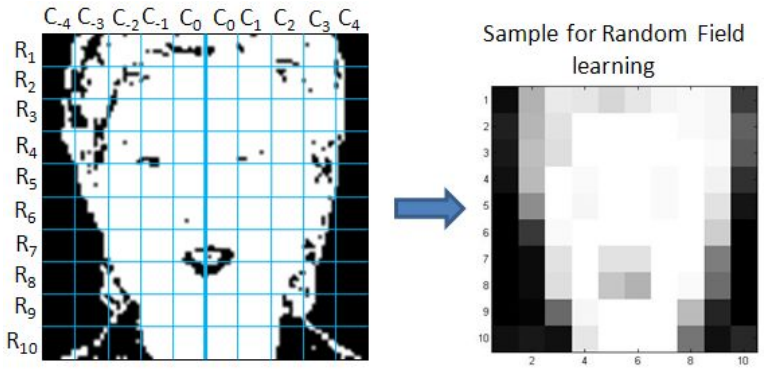


Figure 4.6: Pre-processing.

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 [103] 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 4.6. 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 4.6 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 4.7.

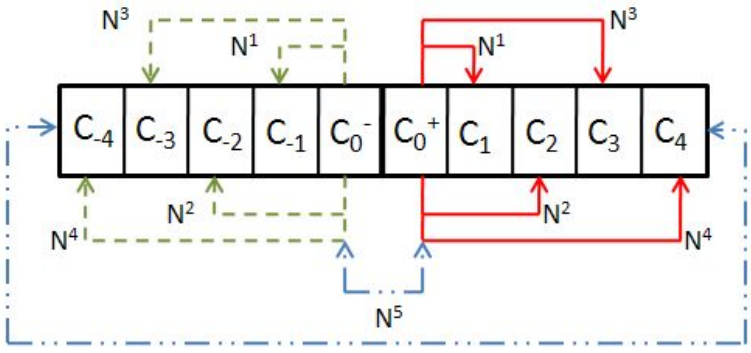


Figure 4.7: Neighborhood System.

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 4.7):

$$\mathscr{N}^k = \left\{C_{j|j\in\{|k|,0^-,0^+\}}\right\} \tag{4.8}$$

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,

$$\mathbf{z} = [x_{C_{0^\pm}}, \delta(x_{C_{0^\pm}}, x_{C_{\pm k}})], \forall R_j \tag{4.9}$$

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

$$\mathbf{z} = [\mu(x_{C_{0^+}}, x_{C_{0^-}}), \mu(x_{C_{-4}}, x_{C_{+4}})], \forall R_j \tag{4.10}$$

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

$$\delta(x_{C_{0^\pm}}, x_{C_{\pm i}}) = \begin{cases} x_{C_{0^+}} - x_{C_{+i}}, & i > 0 \\ x_{C_{-i}} - x_{C_{0^-}}, & i < 0 \end{cases} \tag{4.11}$$

$$\mu(a,b) = \frac{a+b}{2} \tag{4.12}$$

In order to estimate the LCPD on these 5 histogram pools, we use Parzen Window Density Estimation (PWDE) technique, similar to [104], with a 2D Gaussian window. Thus, each of LCPD can now be defined as

$$P^k(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{d}{2}} n h_{opt}^d} \sum_{j=1}^{n} \exp \left[-\frac{1}{2h_{opt}^2} \left(\mathbf{z} - \mathbf{H}_j^k \right)^T \Sigma^{-1} \left(\mathbf{z} - \mathbf{H}_j^k \right) \right] \tag{4.13}$$

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 4.8 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 4.9. 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.

$$P(\mathbf{z}) = \left\{ P_{m|m=\{r,f,l\}}^{k|k=\{1,\ldots,5\}}(\mathbf{z}) \right\} \tag{4.14}$$

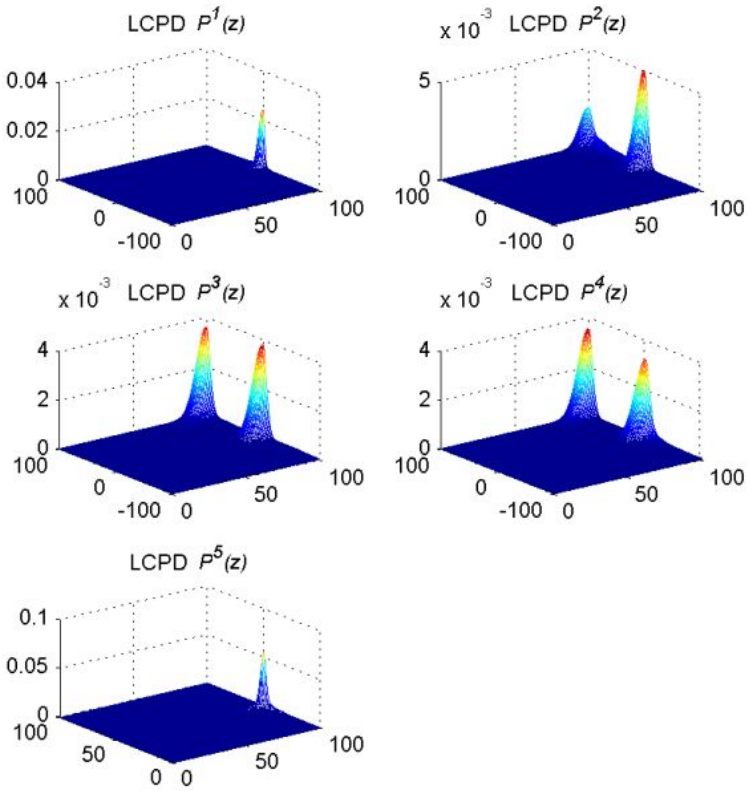


Figure 4.8: Frontal face Local Conditional Probability Density (LCPD) models.

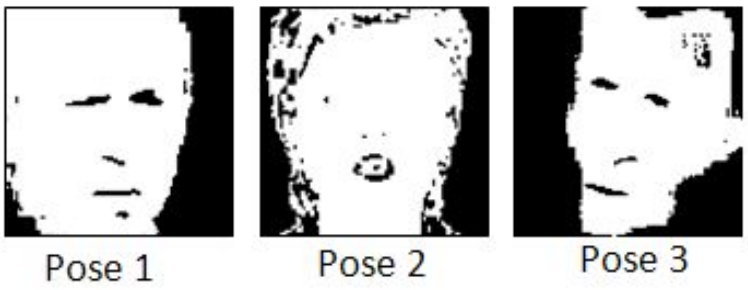


Figure 4.9: Skin-region masks.

Combining Evidence

Given any test face detection output, **z** is extracted (as described in Equation 4.9 and 4.10) 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,

$$\prod_{\forall \mathbf{z} \in \mathbf{H}_m^k} \ln\left(l_m^k(\mathbf{z})\right) = \sum_{\forall \mathbf{z} \in \mathbf{H}_m^k} L_m^k(\mathbf{z}) \tag{4.15}$$

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 4.10. 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

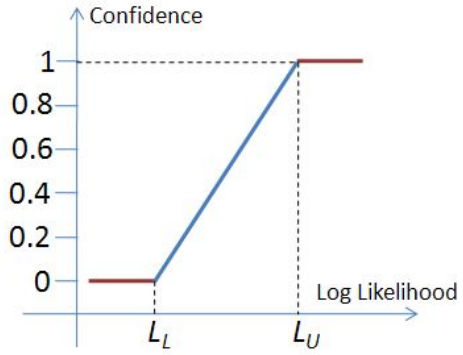


Figure 4.10: Soft threshold.

Dempster-Shafer Theory of Evidence.

Dempster-Shafer Theory of Evidence (DST)

The Dempster-Shafer theory is a mathematical theory of evidence [99] 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:

$$m_{1,2}(A) = \frac{1}{1-K} \sum_{B \cap C = A, A \neq \emptyset} m_1(B) m_2(C) \tag{4.16}$$

where,

$$K = \sum_{B \cup C = \emptyset} m_1(B) m_2(C) \tag{4.17}$$

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 4.16 - 4.17, 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 4.11. 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 4.11: 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 4.4.

4.5 Experiments

In all our experiments, Viola-Jones face detection algorithm [91] 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 %)

4.6 Results

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

1. Number of false detections (NFD)

$$\text{NFD} = \text{Count of false detections} \tag{4.18}$$

2. False detection rate (FDR):

$$\text{FDR} = \frac{\text{\# of false detections}}{\text{Total \# of face detections}} \text{x}100 \tag{4.19}$$

3. Precision (P)

$$\text{P} = \frac{\text{\# of true detections}}{\text{\# of true detections} + \text{\# of false detections}} \tag{4.20}$$

4. Capacity (C)

$$\text{C} = \left(\frac{\text{\# of true detections}}{\text{\# of actual faces in database}} \right) - \text{FDR} \tag{4.21}$$

Table 4.1: 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 4.2: 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 4.4, the framework was extensible to perform coarse pose estimation. Figure 4.12 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.

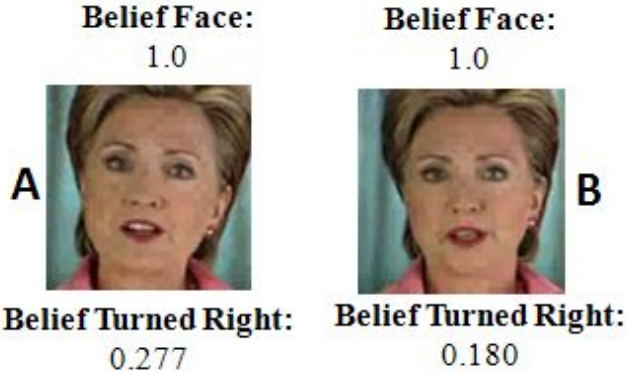


Figure 4.12: Coarse pose estimation.

4.7 Discussion of Results

Performance analysis of the proposed face validation filter can be understood through the four parameters defined in Section 4.6. **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 4.1 and 4.2, 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 4.1 and 4.2 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 4.12 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 5

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 5.1: Person of interest at a short distance from camera

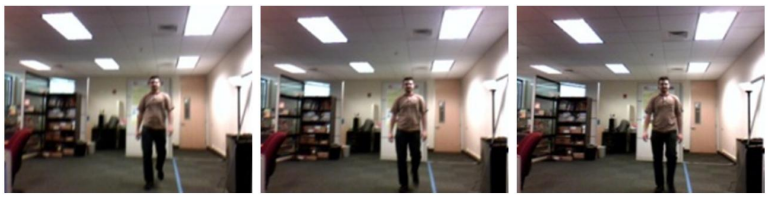


Figure 5.2: Person of interest at a large distance from camera

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 5.1). 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 5.2). 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

5.1 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 5.3 and 5.4 illustrate the contrast in the data due to the nature of the background.

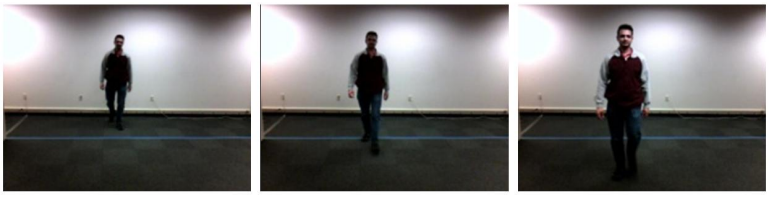


Figure 5.3: Simple Background

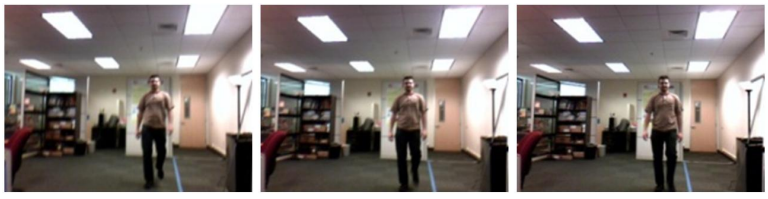


Figure 5.4: Complex Background

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 5.5 shows the simplicity of the data when these problems are not present, while Figure 5.6 highlights complex data formulations in a typical interaction scenario.

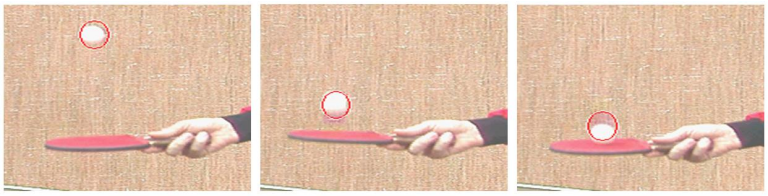


Figure 5.5: Rigid, Homogeneous Object

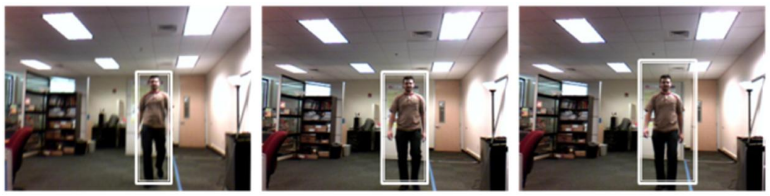


Figure 5.6: Non-Rigid, Deformable, Non-Homogeneous Object

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 5.7 and Figure 5.9). 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 [5] [6-8]. 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 [8]. Though solutions have been suggested, person localization through low frame rate moving cameras still remains an active research topic.

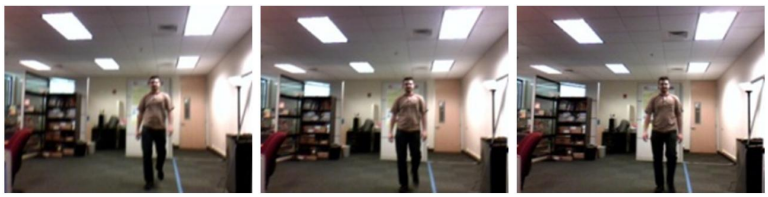


Figure 5.7: Static Camera

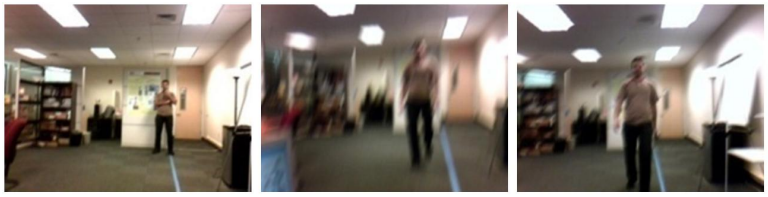


Figure 5.8: Mobile Camera

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 5.9 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

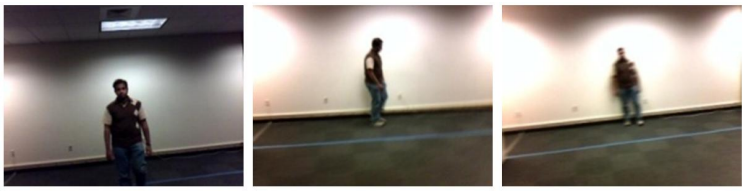


Figure 5.9: Changing Illumination, Pose Change and Blur

related work in the computer vision literature.

5.2 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 [9] and histograms of oriented gradients [10] 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 [11] and edge orientation histograms [12].

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 [13,10], , entropy and haar-like features. Some of the well-known higher level descriptors are histogram of oriented gradients [10] and covariance features [14]. 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 [5] and constellation models [15] 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[16]. Shape-based Chamfer matching [25] 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 [18].

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 [16]. 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 [16]. 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 [20]. 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[11], contours[21], appearance models, intensity gradients[22], region covariance, texture, edge-orientation histograms, haar-like rectangular features [16] , 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.

5.3 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 [23]. 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.

5.4 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 5.10), 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 5.11) 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 5.10: 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.

$$p\left(x_t^i | x_{t-1}^i = p(x_t^i)\right) \tag{5.1}$$

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

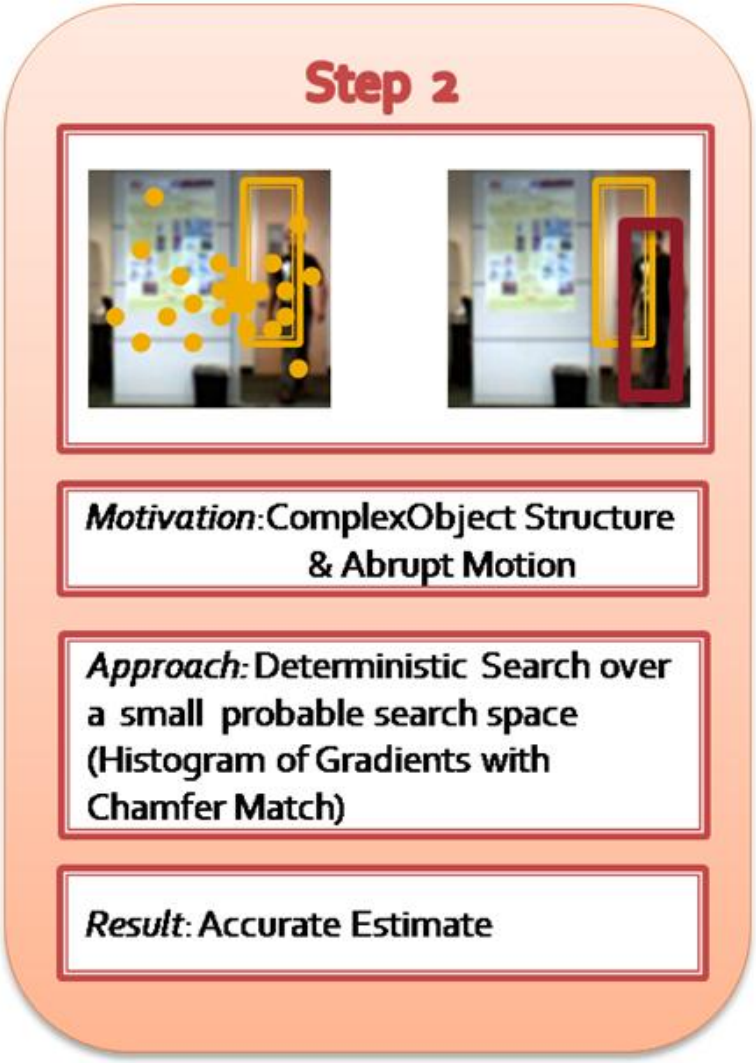


Figure 5.11: SMSPF - Step 2

to be dependent only on the current measurement

$$q\left(x_t^i | x_{t-1}^i Z_t\right) = q\left(x_t^i | Z_t\right) \tag{5.2}$$

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 estimate,

Figure 5.12: Structured Search

thus exploiting the weak temporal redundancy. As in [11], 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 5.12. 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

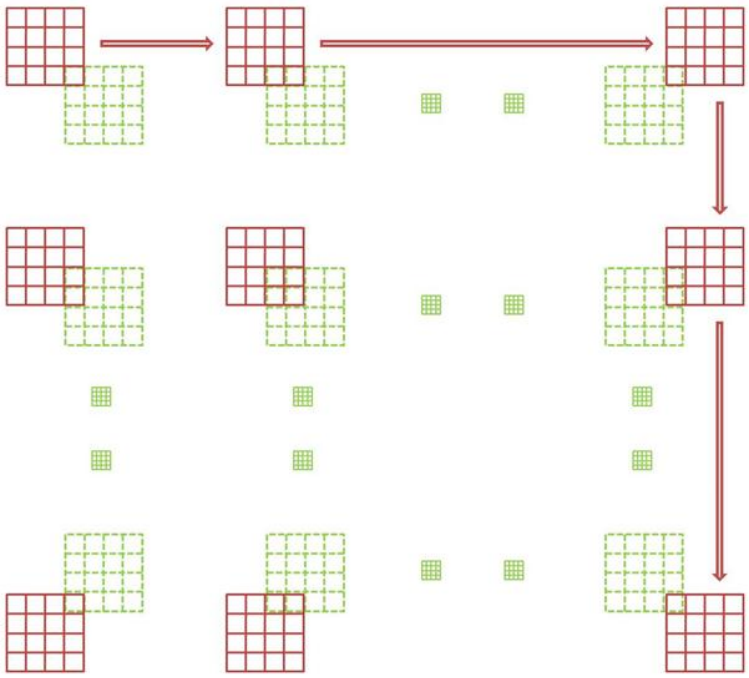


Figure 5.13: Sliding window of the Structured Search (Green: Estimate; Red: Sliding window).

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 5.13, 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 5.14)

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) [10]. Following this principle, we have used the Edge Orientation Histogram (EOH) features [12] 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 [11] 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 [27] to facilitate computationally efficient searching.

Similarly, in order to obtain the boundary confidence measure, a generic person silhouette template (GPT) (as shown in Figure 13) 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

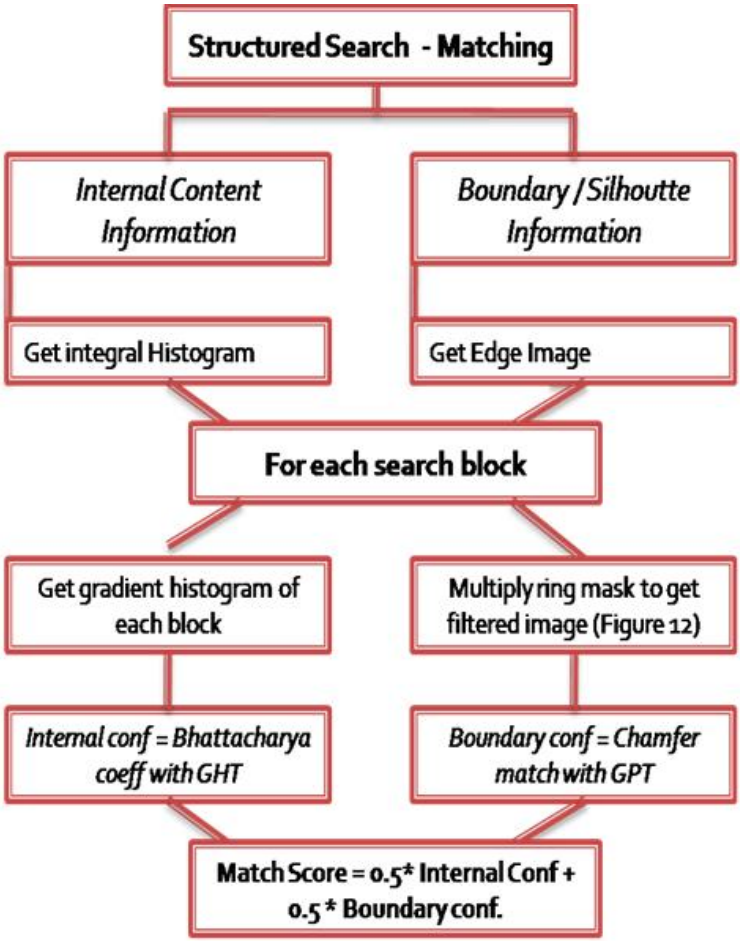


Figure 5.14: Structured Search Matching Technique

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

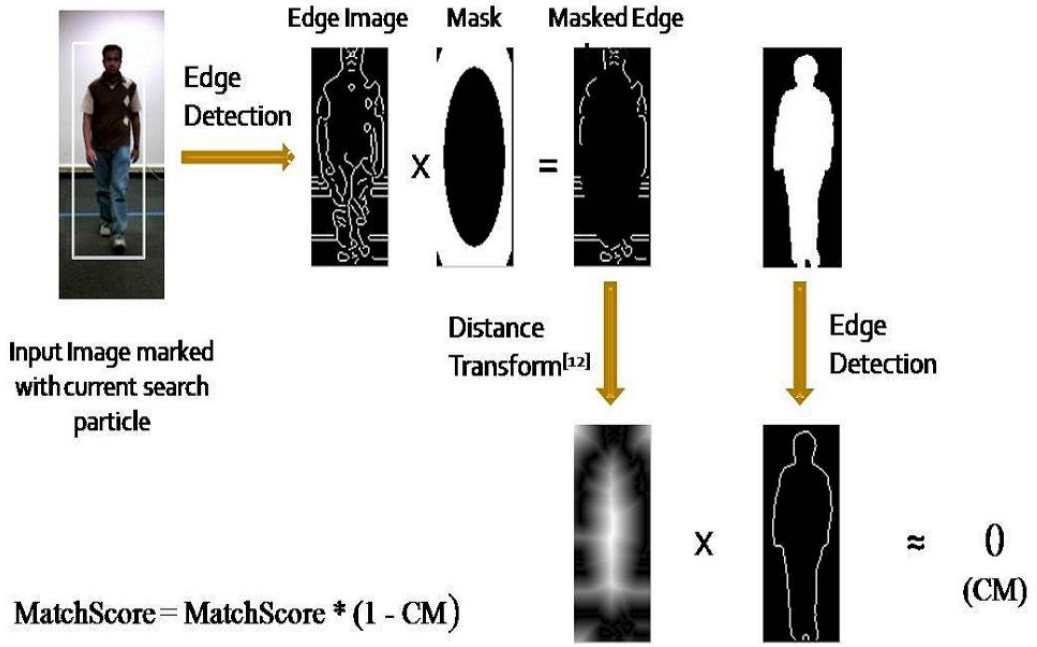


Figure 5.15: Incorporating Chamfer Matching into Structured Search

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 5.15. 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.

$$\text{BoundaryConf} = (1 - \text{ChamferMatch}) \tag{5.3}$$

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.

5.5 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 5.16: 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 5.16 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 [25]. Assuming that a detection algorithm can detect persons in at least some frames, the image region

[1]Collected at CUbiC
[2]CASIA Gait Dataset B with subject approaching the camera [4]
[3]Collected at CUbiC

containing the person in each of the test sequences has been manually set. The following two criteria have been used to evaluate their performance [3].

- 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

$$AO(\text{gTruth}_i, \text{track}_i) = \frac{Area(\text{gTruth}_i \cap \text{track}_i)}{AO(\text{gTruth}_i \cup \text{track}_i)} \tag{5.4}$$

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

$$AvgAOR = \frac{1}{N}\sum_{i=1}^{N} AO \tag{5.5}$$

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

$$OTE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(Centroid_{gTruth_i} - Centroid_{Truth_i}\right)} \tag{5.6}$$

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).

$$TEM = 2 * \frac{AvgAOR.e^{-k.OTE}}{AvgAOR + e^{-k.OTE}} \tag{5.7}$$

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

5.6 Results

Particle Filtering has been widely used to handle complex scenarios by maintaining multiple hypotheses. As mentioned in [21], 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) [25] 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 5.17 and Figure 5.18 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 5.17(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 5.18(a) where the distance between centroids using color particle filter is much higher than that with SMSPF(≈ 0).

Figure 5.19, Figure 5.20 and Figure 5.21 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.

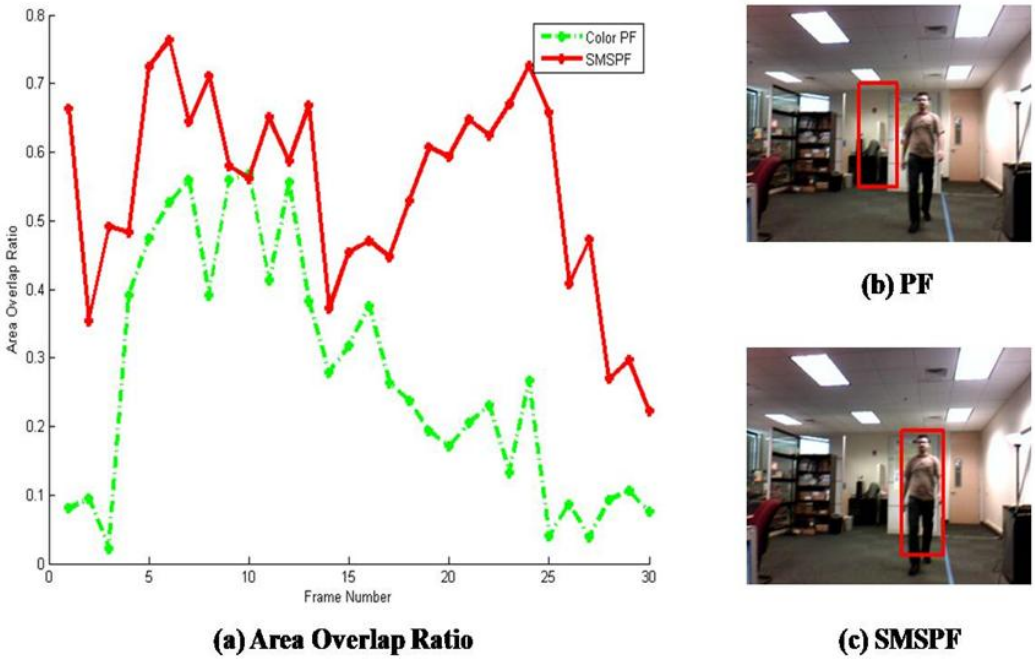


Figure 5.17: AO (Dotted Line: Color PF; Solid Line: SMSPF)

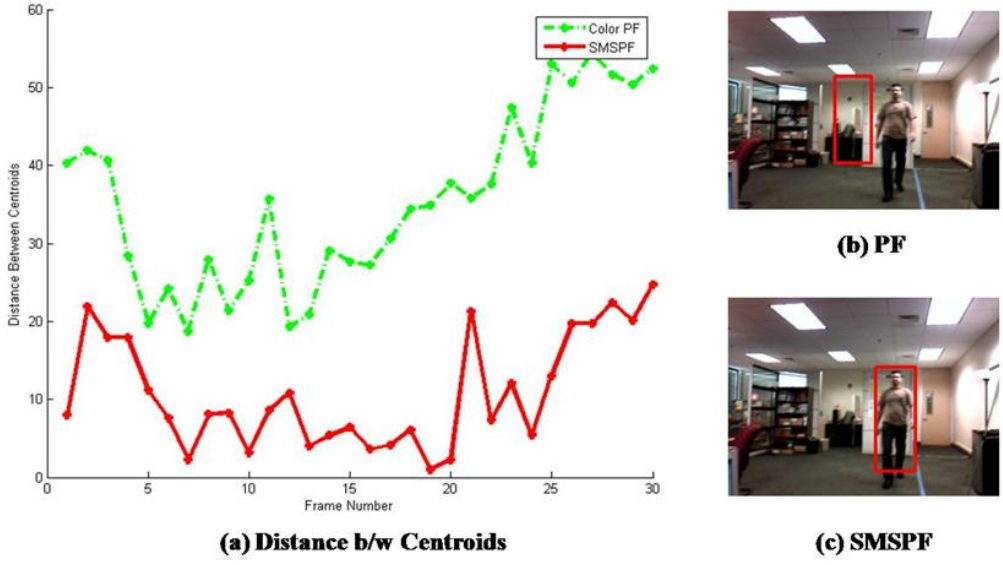


Figure 5.18: DC(Dotted Line: Color PF; Solid Line: SMSPF)

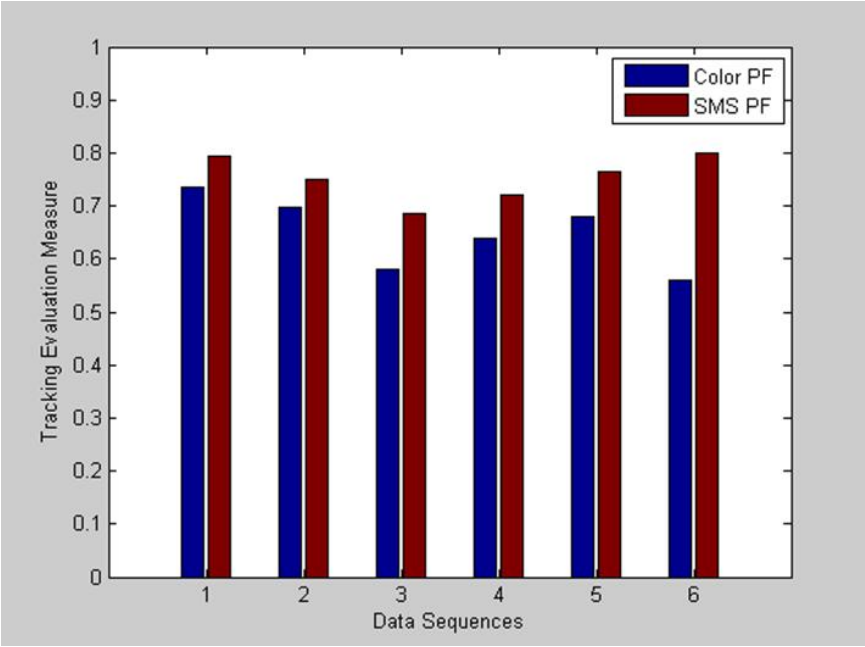


Figure 5.19: Evaluation Measure for DataSet 1

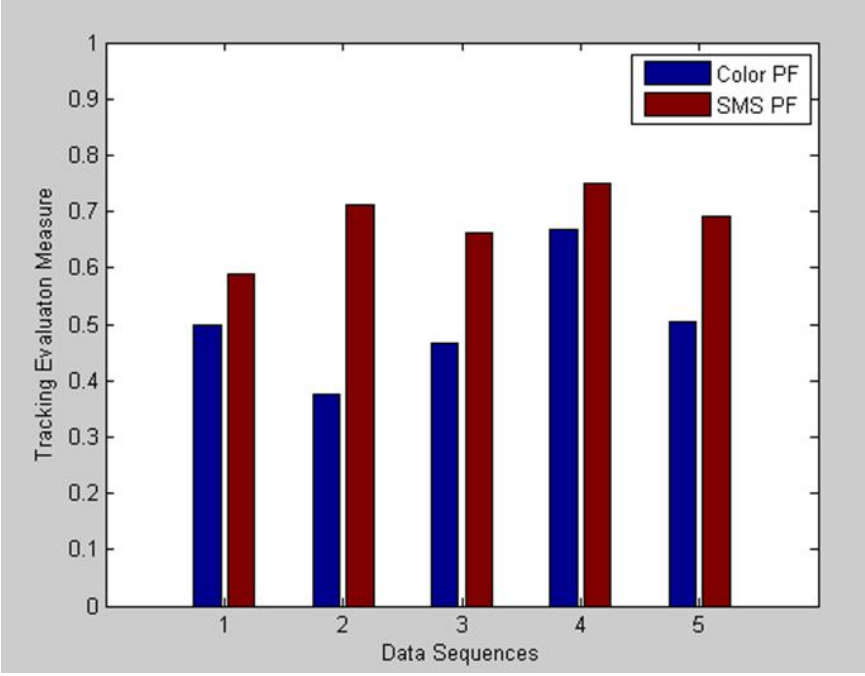


Figure 5.20: Evaluation Measure for DataSet 2

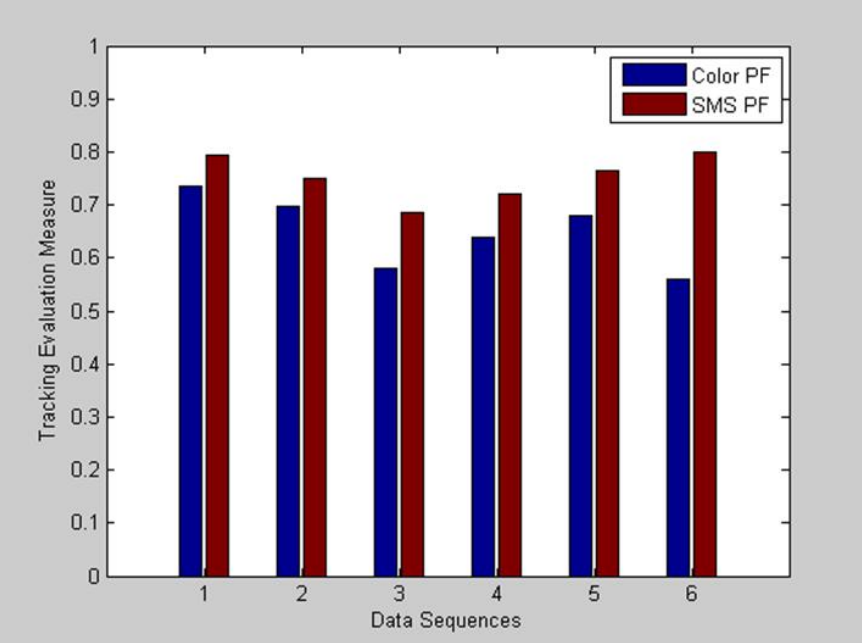


Figure 5.21: Evaluation Measure for DataSet 3

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

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

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

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

[100] 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.

[101] 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.

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

[103] 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.

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

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

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

   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.

Appendix B

INSERT APPENDIX B TITLE HERE

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!