Chapter 4

Exocentric Sensing & Delivery: Proxemics

4.1 Proxemics

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

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

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

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

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

4.2 Conceptual Framework

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

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

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

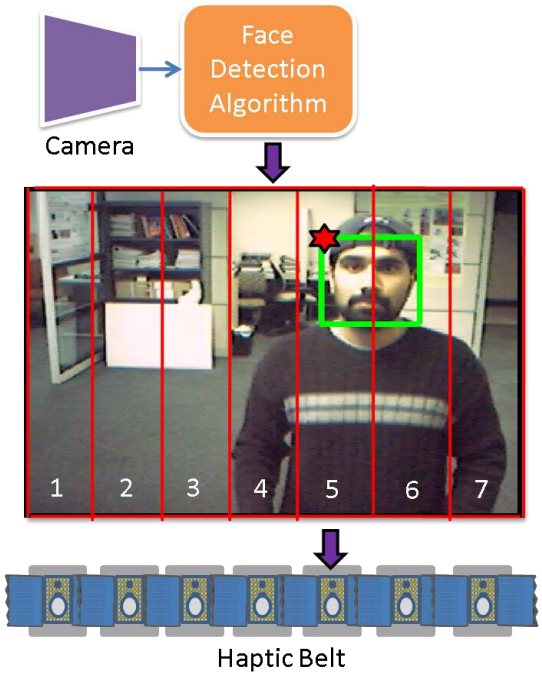

Figure 1: System Architecture for Haptic Belt used as part of the Social Interaction Assistant

4.3 Accurate Face Detection through the Wearable Camera

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

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

Figure 2: An example false face detection

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

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

1) A human skin-tone detector with a dynamic background modeler (**Module 1**), and

2) An evidence-aggregating human face silhouette random field modeler (**Module 2**), which provides a confidence metric on its validation task.

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

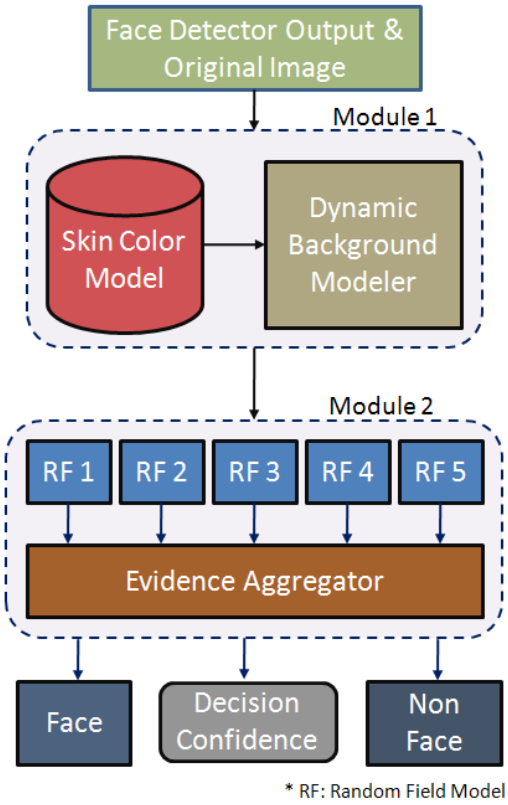

Figure 3: Block Diagram

4.3.1 Face Validation Framework:

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

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

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

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

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

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

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

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

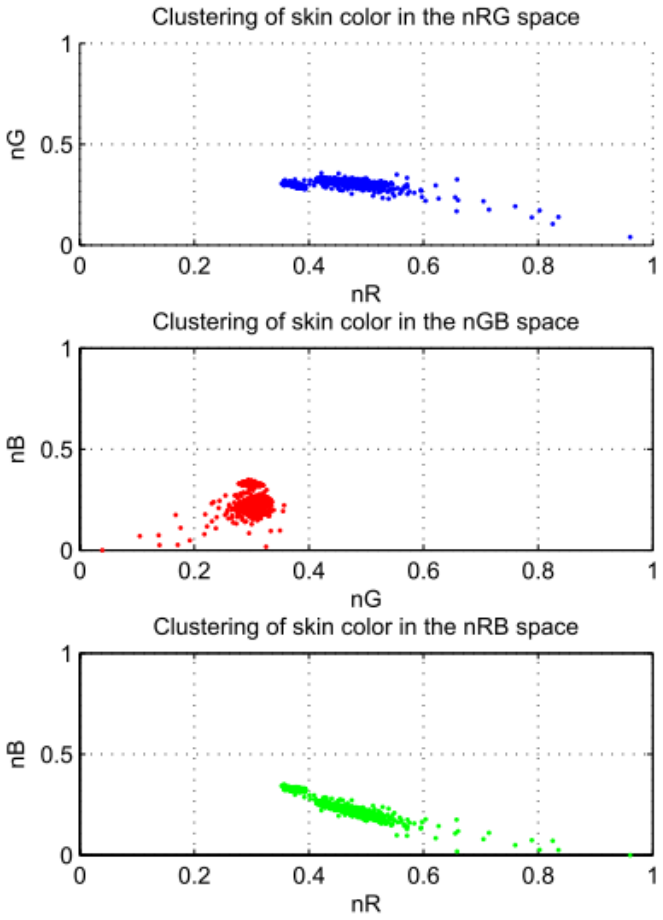

Figure 4: Skin Pixels in nRGB space

In our experiments, we found that skin pixels form a tight cluster when projected on nG and nB space as shown in the Figure 4. The study was based on a skin pixel database, consisting of nearly 150,000 samples, built by randomly sampling skin regions from 1040 face images collected on the web as well as from FERET face database [12]. Further analysis also showed that the cluster formed on the 2D nG-nB space had two prominent density peaks which motivated the modeling of skin pixels with a Bi-modal Gaussian mixture model learnt using Expectation Maximization (EM) with a *k*-means initialization algorithm [13]. The Bi-modal Gaussian mixture model is represented as.

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

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

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

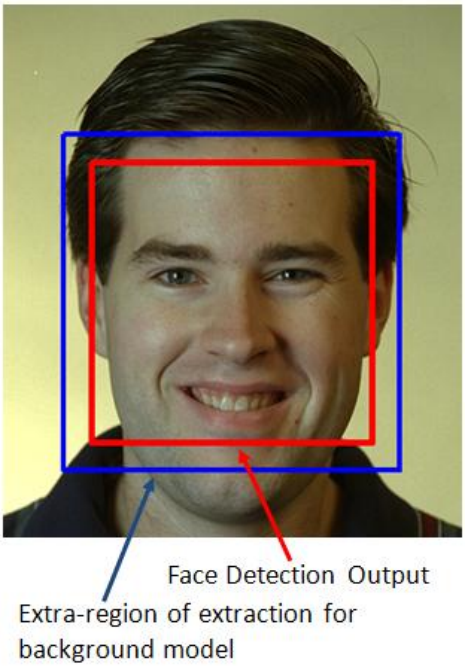

Figure 5: Extra region for background modeling

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

represented as

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

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

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

The skin and non-skin models, $f_{X|X=[nG,nB]}^{skin}(x)$ and $f_{X|X=[R,G,B]}^{non-skin}(x)$ respectively, are used for classifying every pixel in the scaled face image obtained as explained above. Example skin-masks are shown in Figure 6. This example shows two sets of images - one corresponding to a *true* face detection result, and another *false* face detection result.

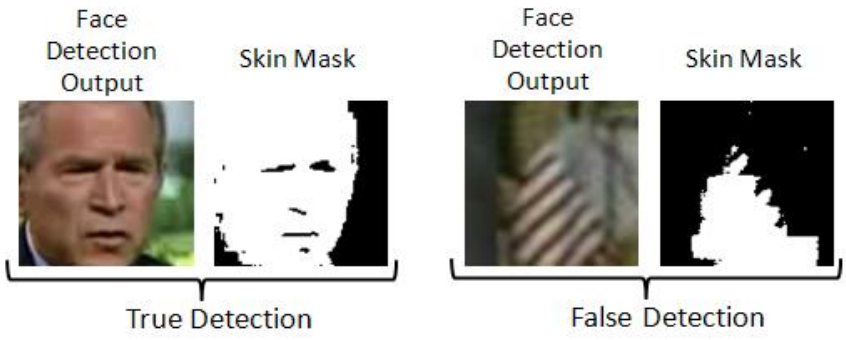

Figure 6: Example of *true* and *false* face detections

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

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

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

4.3.1.2.1 Random Field (RF) Models

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

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

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

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

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

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

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

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

4.3.1.2.2 Pre-processing

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

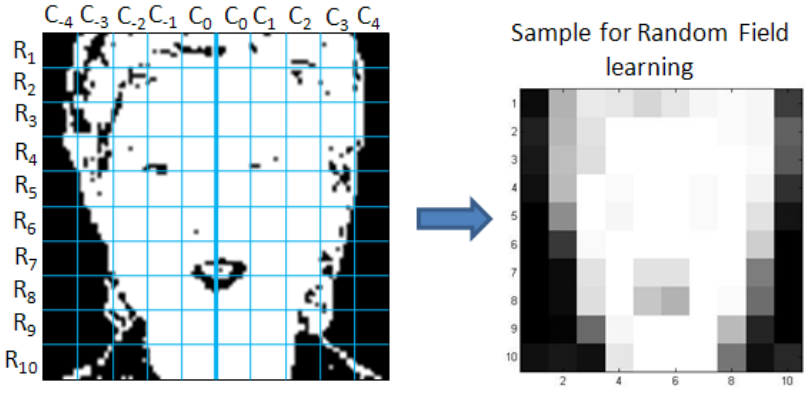

Figure 7: Pre-processing

4.3.1.2.3 The Neighborhood System

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

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

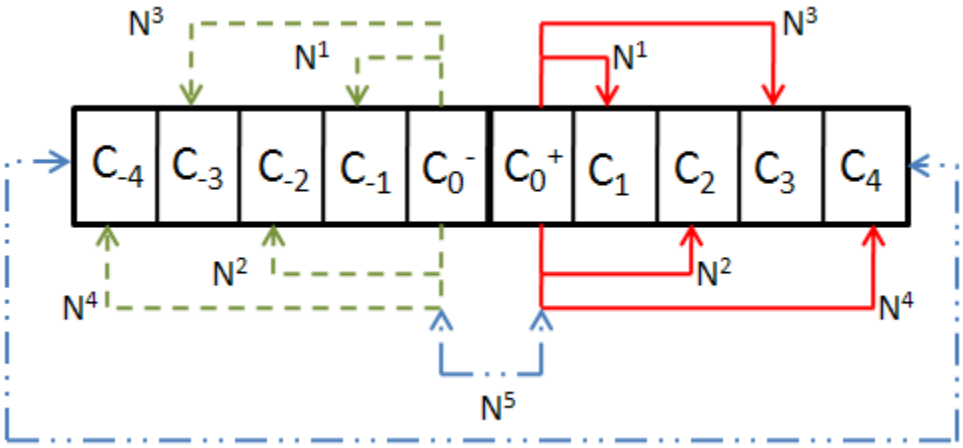

Figure 8: Neighborhood System

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

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

4.3.1.2.4 Local Conditional Probability Density (LCPD)

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

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

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

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

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

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

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

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

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

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

Figure 9 shows the 5 LCPDs learnt over a set of 390 training frontal face images.

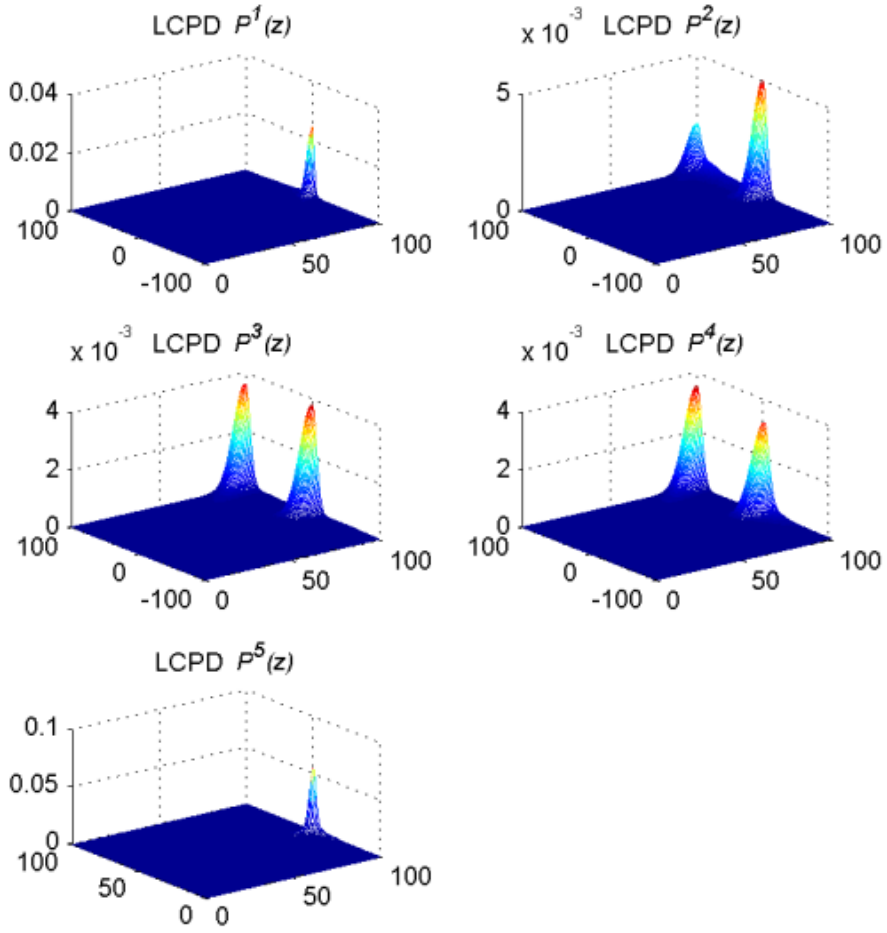

Figure 9: Frontal face Local Conditional Probability Density (LCPD) Models

4.3.1.2.5 Human Face Pose

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

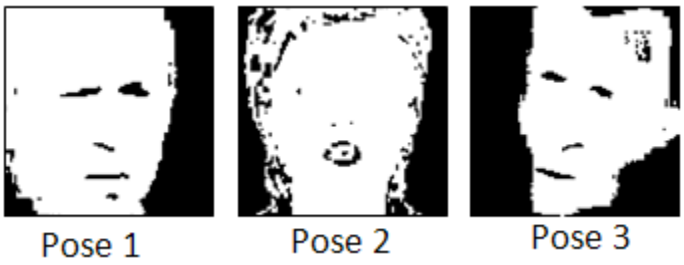

Figure 10: Skin-region masks.

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

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

4.3.1.3 Combining Evidence

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

$$\prod_{\forall z \in H^k_m} \ln\left(l^k_m(z)\right) = \sum_{\forall z \in H^k_m} L^k_m(z)$$

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

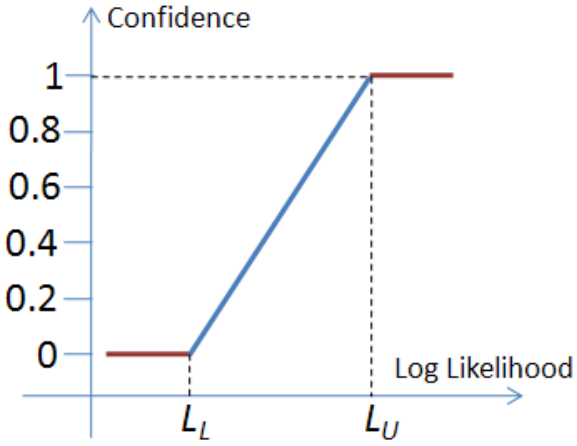

Figure 11: Soft threshold

4.3.1.3.1 Dempster-Shafer Theory of Evidence (DST)

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

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

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

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

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

$$m_{1,2}(A) = \frac{1}{1-K} \sum_{B \cap C = A, A \neq 0} m_1(B) m_2(C)$$

Where,

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

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

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

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

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

4.3.1.4 Coarse Pose estimation

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

4.3.2 Experiments

In all our experiments, Viola-Jones face detection algorithm \cite{viola_robust_2004} was used for extracting face subimages. The proposed face validation filter was tested on two face image data sets,

1. The FERET Color Face Database, and

2. An in-house face image database created from interview videos of famous personalities.

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

1. FERET

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

2. In-house database

- Number of actual face images: 2,597
- Number of faces detected using Viola-Jones algorithm: 2,324

- Number of *true* detections: 2,074
- Number of *false* detections: 250 (10.7%)

4.3.3 Results

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

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

   NFD = Count of false detections

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

   FDR = (# of false detections) / (Total # of face detections) x100

3. **Precision (P)**

   P = (# of true detections) /(# of true detections + # of false detections)

4. **Capacity (C)**

   C = (# of true detections) / (# of actual faces in database) – FDR

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

Table 1: Face detection validation results on FERET database

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

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

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

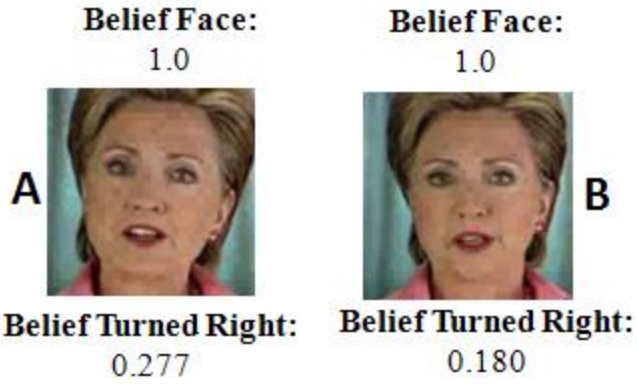

Figure 13: Coarse pose estimation

Discussion of Results

Performance analysis of the proposed face validation filter can be understood through the four parameters defined in the previous section. NFB and FDR are direct measurements of the number of mistakes (naming non-faces as faces) made by the face detection algorithm on the two data sets. As can be verified from Table 1 and2, there is a significant reduction in the false detections through the introduction of the filter.

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

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

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

4.4 Delivering Proxemics Information

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

4.4.1 Delivering Direction Data - *Localization of Vibrotactile Cues*

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

4.4.1.1 Subjects:

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

4.4.1.2 Apparatus:

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

4.4.1.3 Procedure:

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

The training phase involved 35 trials where each tactor was randomly activated 5 times (with approximately 5 seconds between tactor activations) and subjects had to identify the number of each activated tactor. A visual guide was provided for subjects to help recall tactor numbers; this guide was a white board with a drawing of a semicircle (the belt) and the numbers 1 through 7

(tactors) on the belt. Feedback was given during the training phase to correct wrong guesses. The testing phase was similar to the training phase, but involved 70 trials where each tactor was randomly activated 10 times, and feedback was not provided. Subjects stood during the entire experiment.

4.4.1.4 Results:

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

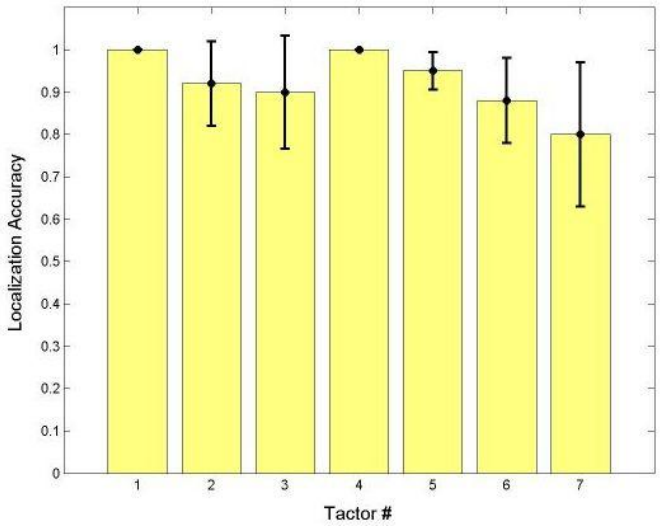

Figure 14: Experiment 1 Results: Mean Localization Accuracy for each Tactor, Averaged across Subjects, with 95% Confidence Intervals

4.4.1.5 Discussion:

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

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

acuity improves near the sagittal plane [18] [17] given that spatial acuity is better at anatomical reference points—in this case, the navel.

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

4.4.2 Delivering Distance Data

4.4.2.1 Tactile Rhythm Design

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

We conducted pilot studies to determine rhythm patterns that are convenient for users to identify vibratory rhythms. Through use of a vibrotactile belt, we evaluated use of five rhythms, each 10 seconds in length: 50ms vibrotactile pulses separated by pauses of length 50ms, 100ms, 300ms, 500ms and 1000ms. Subjects found rhythms with pauses of 100ms, 300ms and 500ms difficult to discriminate between. Based on these findings, we selected the four rhythms depicted in figure 1; this design includes more separation of pauses within 100ms to 500ms, and a small increase of 1000ms to 1200ms (much longer durations may be too time consuming for communication [19]). In the Social Interaction Assistant, these four tactile rhythms are mapped to interpersonal distances corresponding to intimate, personal (close phase), personal (far phase) and social (close phase) space respectively.

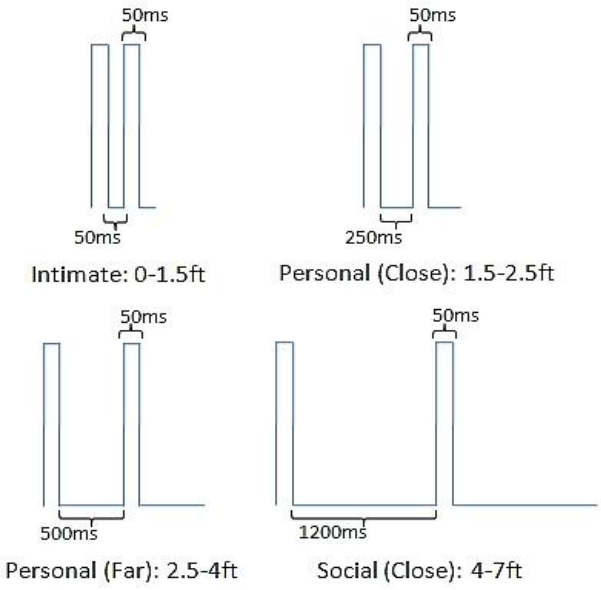

Figure 15. The on/off timing values of the four tactile rhythm designs, and corresponding distances, used in the experiment.

4.4.2.2 Experiment

4.4.2.2.1 Aim:

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

4.4.2.2.2 Hypotheses:

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

4.4.2.2.3Subjects:

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

4.4.2.2.4 Apparatus:

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

4.4.2.2.5 Procedure:

Subjects wore the belt underneath their clothing and sat during the entire experiment. Subjects had access to visual guides—a semi-circle with tactors #1-7 drawn and interpersonal distances labeled as rhythms #1-4—to recall tactor and rhythm numbers, respectively. First, subjects were familiarized with vibration location as it pertains to direction. Each tactor was vibrated for 3 seconds, and the tactor number was called out by the experimenter. Next, subjects were familiarized with tactile rhythms. Each rhythm was presented for 7 seconds through the fourth tactor at the navel, and the rhythm number was called out by the experimenter. Next, subjects began the training phase where they were asked to identify the direction (through the location of the activated tactor) and distance (through the type of rhythm) indicated by each tacton. All 28

tactons (4 tactile rhythms at 7 different locations/tactors) were randomly presented for 10 seconds each. Subjects were encouraged to respond before the 10 seconds ended. Subjects had to achieve a recognition accuracy of 80% or more on each tacton dimension to proceed immediately to the testing phase; otherwise, the training phase was repeated (only 6 subjects had to repeat training, and all passed on the second try). The experimenter corrected wrong guesses and confirmed correct guesses. The testing phase was similar to the training phase, except no feedback was provided by the experimenter concerning right or wrong guesses, and each of the 28 tactons was randomly presented 3 times for a total of 84 trials.

4.4.2.2.6 Results:

The overall recognition accuracy follows: location (mean: 95%, SD: 4%), rhythm (mean: 91.7%, SD: 5.7%) and both (mean: 87%, SD: 8.5%). These results support hypotheses (1)-(3), and show that, overall, subjects had little difficulty in recognizing rhythms and locations as they pertain to distance and direction, respectively. Feedback from participants after the experiment further supported this. From herein, reported ANOVA results are from a two-way ANOVA on complete tacton recognition accuracy through location and rhythm. The overall recognition accuracy of each tactor location is shown in figure 2. Subjects felt that the vibrations of tactor #1 (left side), #4 (navel) and #7 (right side), were easier to localize compared to tactor #2, #3, #5 and #6. This result is easy to explain as spatial acuity is better at anatomical reference points [17]. Although figure 2 does show a very small difference between recognition accuracies, which supports what subjects reported, there was no significant difference between recognition accuracy of tactor locations [$F(6,1232)=1.96$, $p=0.068$], hence hypothesis (4) cannot be accepted. The overall recognition accuracy of rhythms is shown in figure 3. Subjects felt that rhythm #2 (personal-close) and #3 (personal-far) were more difficult to identify than rhythm #1 (intimate) or #4 (social-close), which is supported by figure 3. A significant difference between recognition accuracy of rhythms [$F(3,1232) =5.70$, $p=0.001$] supported hypothesis (5). No interaction was found between location and rhythm for recognition accuracy of complete tactons [$F(18,1232)=0.91$, $p=0.569$], supporting hypothesis (6).

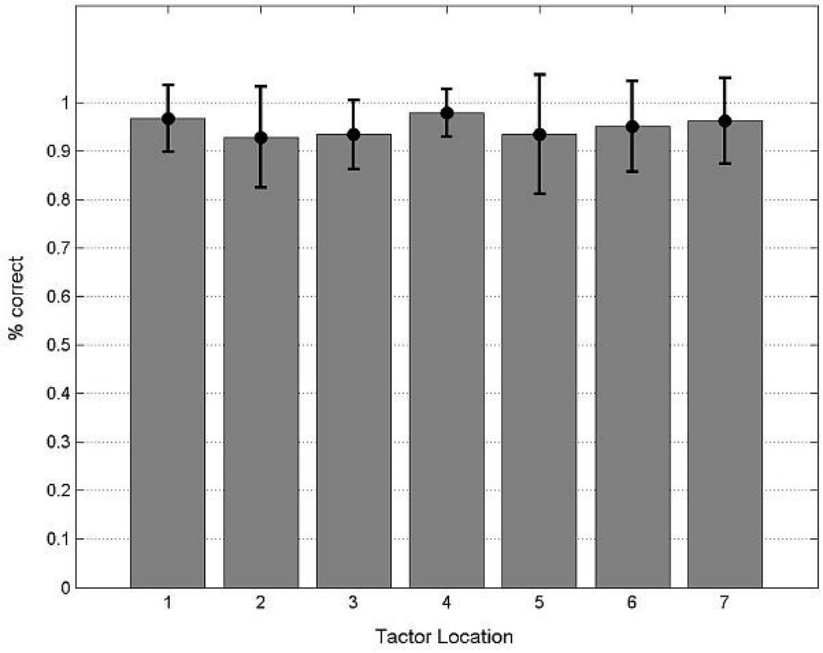

Figure 15. Overall direction recognition accuracy of each tactor location with standard deviations.

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

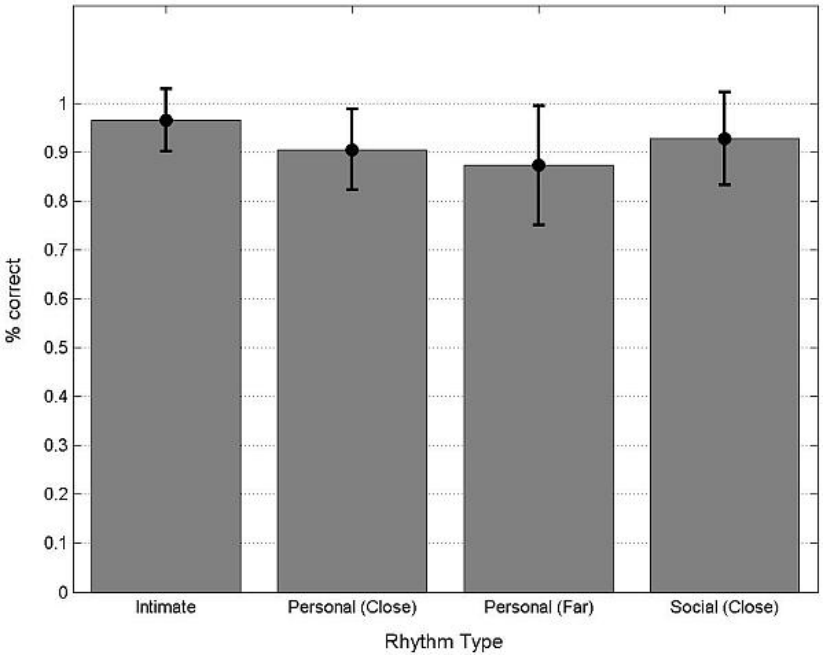

Figure 16. Overall distance recognition accuracy of each rhythm type with standard deviations.

Summary:

References:

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

[2] S. Ram and J. Sharf, "The People Sensor: A Mobility Aid for the Visually Impaired," *iswc*, vol. 00, 1998.

[3] J.B.F.V. Erp, H.A.H.C.V. Veen, C. Jansen, and T. Dobbins, "Waypoint navigation with a vibrotactile waist belt," *ACM Transactions on Applied Perception*, vol. 2, 2005, pp. 106-117.

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

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

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

[7] P. Viola and M. Jones, "Robust Real-time Object Detection," *International Journal of Computer Vision*, 2001.

[8] P. Viola and M. Jones, "Robust Real-time Object Detection," *International Journal of Computer Vision*, 2001.

[9] E. Hjelmås and B.K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding*, vol. 83, Sep. 2001, pp. 236-274.

[10] M. Yang, D.J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, 2002, pp. 34-58.

[11] K. Sentz and S. Ferson, *Combination of evidence in dempster-shafer theory*, Sandia National Laboratories, 2002.

[12] P. Phillips, Hyeonjoon Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, 2000, pp. 1090-1104.

[13] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," 1997.

[14] P. Perez, "Markov Random Fields and images," *CWI Quaterly*, vol. 11, 1998, pp. 413-437.

[15] M. Vezjak, "An anthropological model for automatic recognition of the male human face," pp. 380, 363.

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

[17] R.W. CHOLEWIAK, J.C. BRILL, and A. SCHWAB, "Vibrotactile localization on the abdomen: Effects of place and space," *Perception & Psychophysics*, vol. 66, 2004, pp. 970-987.

[18] J.B.V. Erp, "Vibrotactile Spatial Acuity on the Torso: Effects of Location and Timing

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

[19] F.A. Geldard, "Adventures in tactile literacy.," *American Psychologist. Vol. 12(3)*, vol. 12, Mar. 1957, pp. 115-124.

[20] T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan, "Using a Haptic Belt to convey Non-Verbal communication cues during Social Interactions to Individuals who are Blind.," 2008.