Chapter 5

Exocentric Sensing & Delivery: Facial Expressions

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

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

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

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

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

5.1 Sensing Facial and Head Mannerisms and Expressions:

5.1.1 FaceAPI:

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

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

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

d. The nose (4 points)

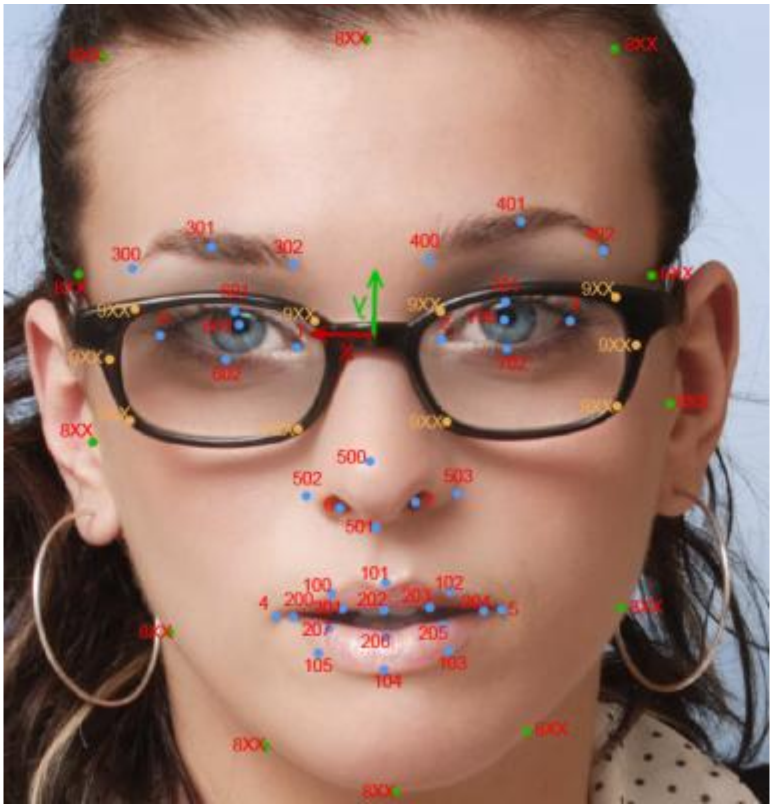


Figure 1: FaceAPI feature points along with head tracking

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

5.1.2 Facial Image Features under Investigation:

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

5.1.2.1 Image features for expression recognition:

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

5.1.2.1.1 LBP:

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

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

Where,

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

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

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

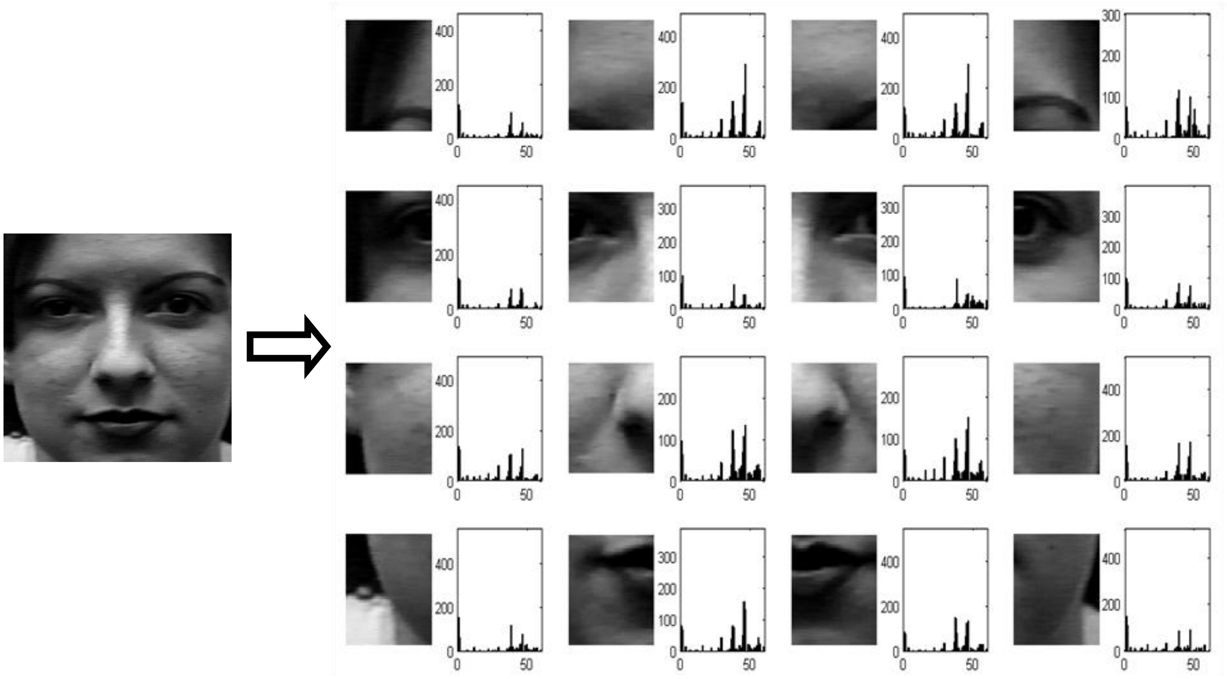


Figure 2: Example histogram of the local binary pattern exteacted over a sample face image by dividing the face into 8x8 non-overlapping windows.

5.1.2.1.2Line Segment Features:

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

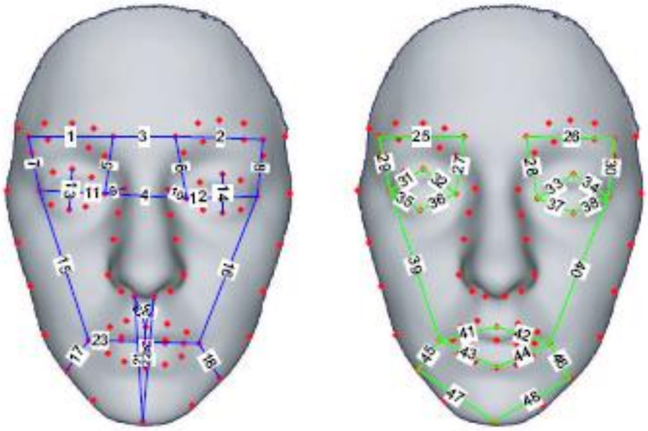


Figure 3: Line segment features, (a) corresponds to the distance features between facial fiducials, (b) corresponds to the orientation of the vectors joining the facial fiducials.

5.2 Importance of Facial Features:

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

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

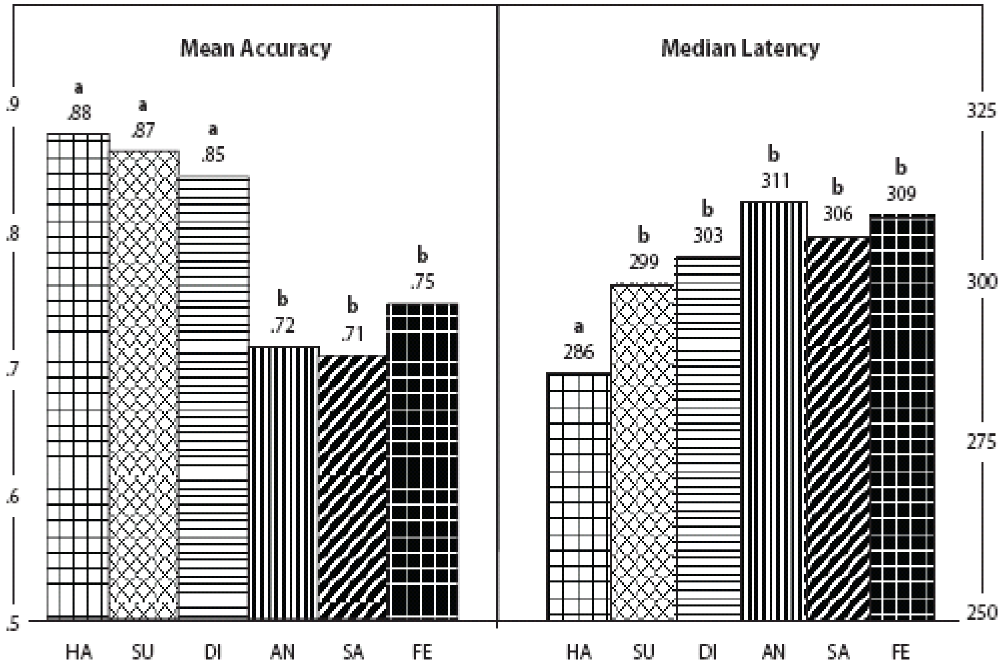


Figure 4: Dwell times in ms for each expression. Source [3]

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

5.3 Delivering Facial Mannerisms and Expressions:

5.3.1 Design Considerations:

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

5.3.2 Construction of the Vibrotactile Glove:

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

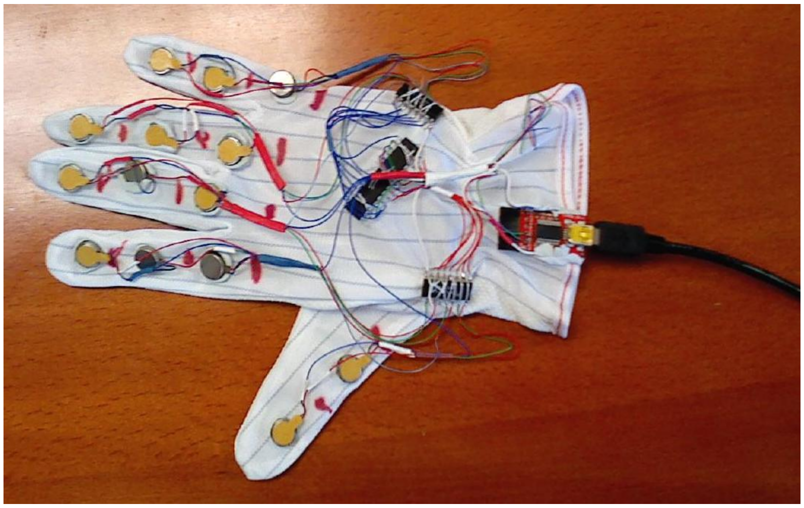


Figure 5: Haptic Glove: The figure shows a glove made out of stretchable material with 14 motors on the back of the glove with each motor corresponding to one phalange of the 5 digits. A microcontroller, two motor drivers and 1 USB controller (4 ICs) are also integrated on to the back of the glove with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

Currently, the device only delivers the 6 basic expressions (Smile, Anger, Disgust, Surprise, Sad and Fear) along with indications of when the face reaches neutral expression. In future, we plan to encode the dynamic motion of the human facial features into vibrotactile patterns. This would allow indiscriminate access to the facial movements of the interaction counterpart.

5.3.3 Mapping for facial expressions:

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

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

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

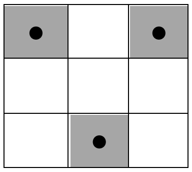
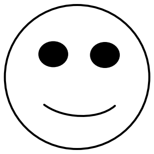
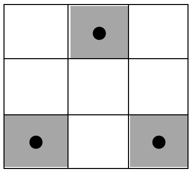
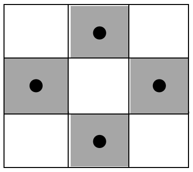
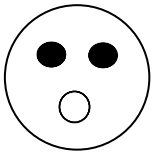
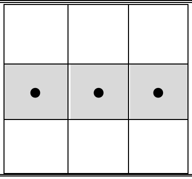
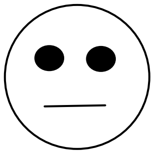
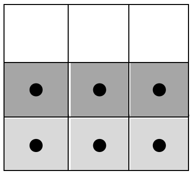
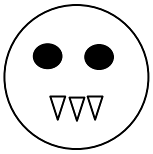
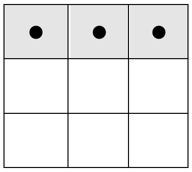
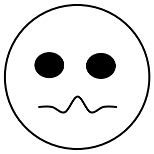
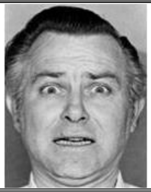
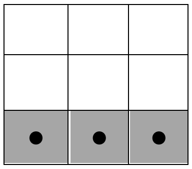
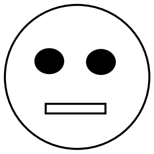
Happy			
Sad			
Surprise			
Neutral			
Angry			
Fear			
Disgust			

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

5.3.4 Experiments:

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

5.3.5 Preliminary Results:

5.3.5.1 Confusion Matrix:

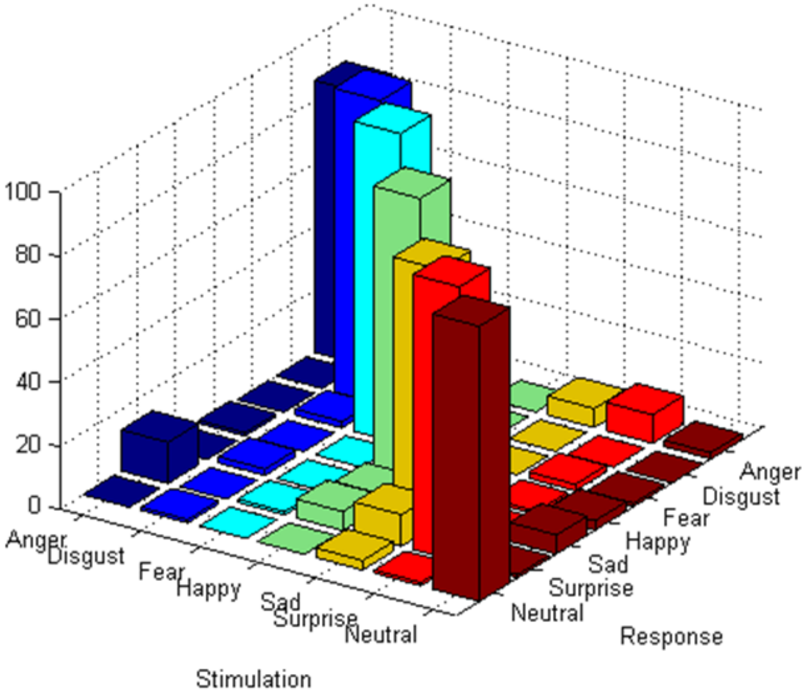


Figure 6: The 3D histogram plot of the confusion matrix. Each color represents a particular stimulation and the corresponding response from the users. This allows for the analysis of where confusion occurred in the delivered data.

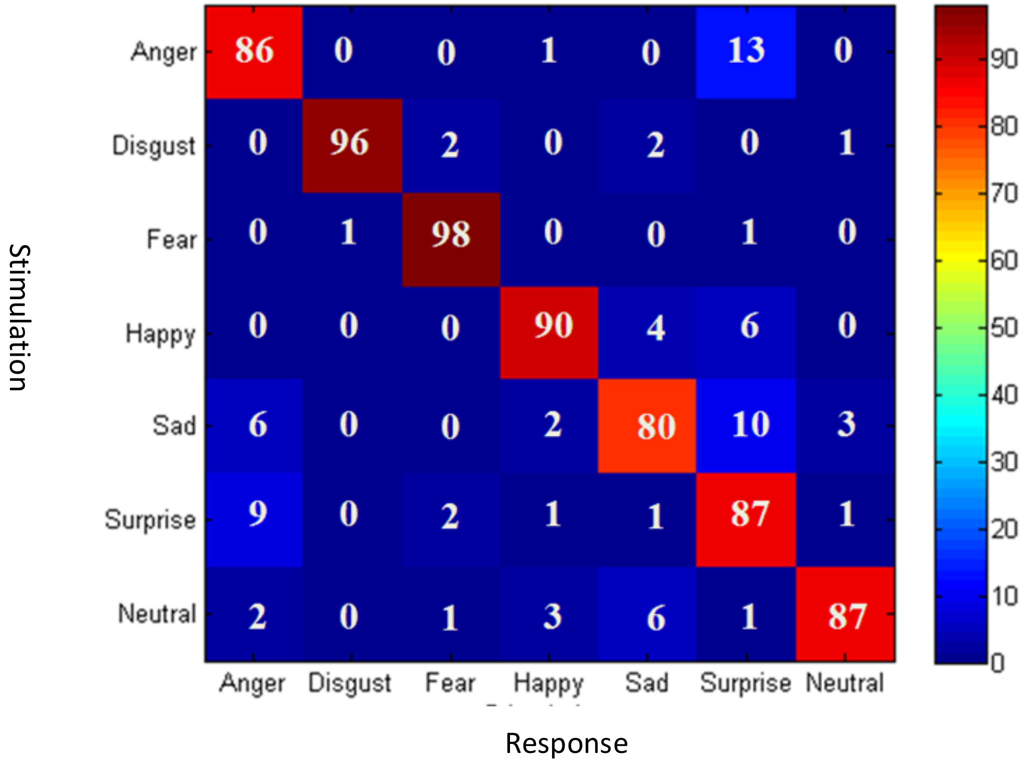


Figure 7: Confusion matrix of responses. Rows represent the stimulation provided to the users and the columns represent the response provided by the user. Each cell has two numbers. The first number represents the percentage recognition of a specific stimulation and a corresponding response. Ideally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

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

5.3.5.2 Response Time:

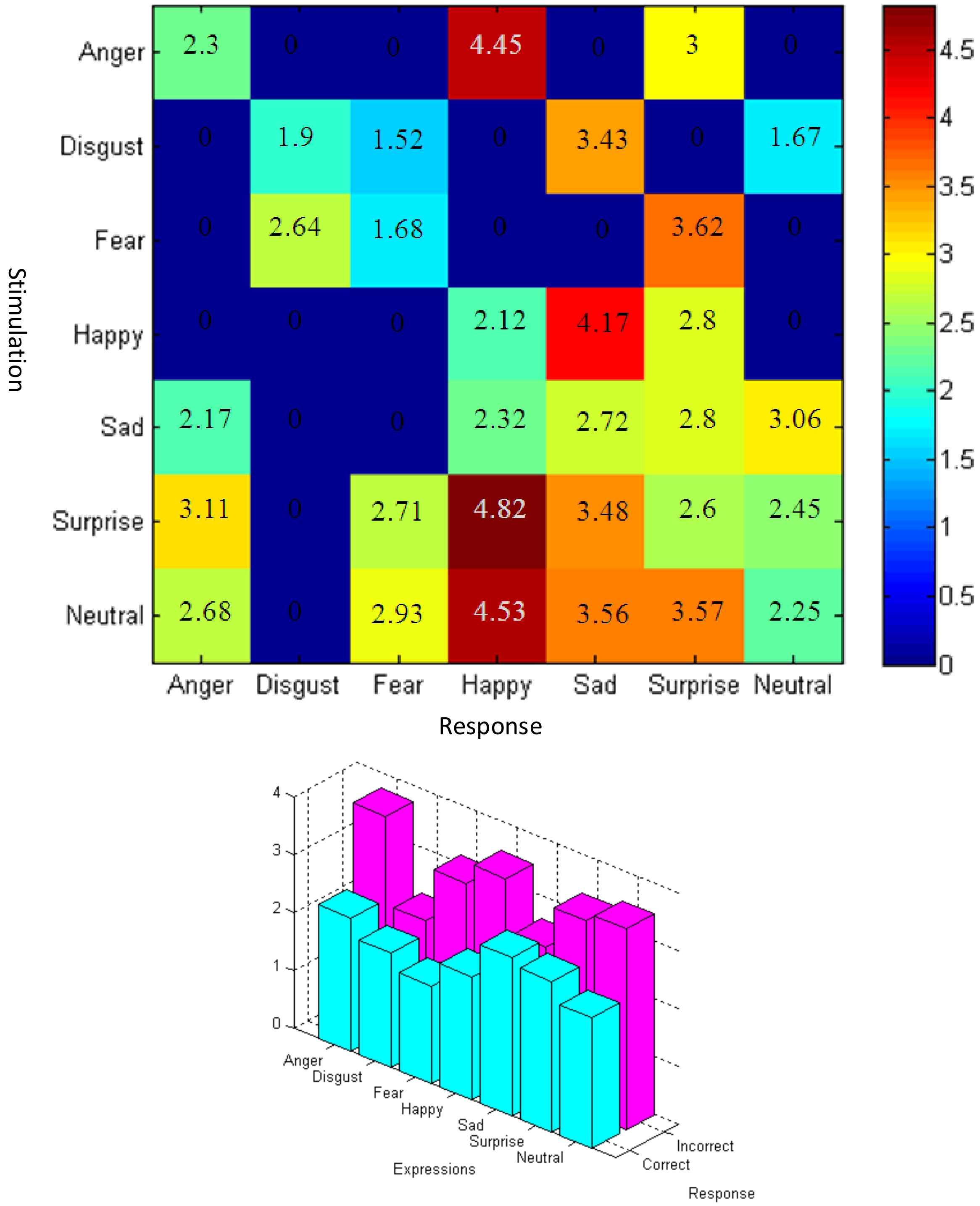


Figure 8: (a) Rows represent the stimulation provided to the users and the columns represent the response provided by the user. Each cell represents the response time for each stimulation and the corresponding response. Ideally, the time has to be as low as possible. (b) Represents the times taken when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

5.3.6 Proposed work:

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

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

5.4 References:

[1] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, 2002, pp. 971-987.

[2] Hao Tang and T. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, 2008, pp. 1-6.

[3] M.G. Calvo and L. Nummenmaa, "Eye-movement assessment of the time course in facial expression recognition: Neurophysiological implications," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 9, Dec. 2009, pp. 398-411.