Evidential Reasoning for Multimodal Fusion in Human Computer Interaction

by

Bakkama Srinath Reddy

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2007

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Bakkama Srinath Reddy

Abstract

Fusion of information from multiple modalities in Human Computer Interfaces (HCI) has gained a lot of attention in recent years, and has far reaching implications in many areas of human-machine interaction. However, a major limitation of current HCI fusion systems is that the fusion process tends to ignore the semantic nature of modalities, which may reinforce, complement or contradict each other over time. Also, most systems are not robust in representing the ambiguity inherent in human gestures. In this work, we investigate an evidential reasoning based approach for intelligent multimodal fusion, and apply this algorithm to a proposed multimodal system consisting of a Hand Gesture sensor and a Brain Computing Interface (BCI). There are three major contributions of this work to the area of human computer interaction. First, we propose an algorithm for reconstruction of the 3D hand pose given a 2D input video. Second, we develop a BCI using Steady State Visually Evoked Potentials, and show how a multimodal system consisting of the two sensors can improve the efficiency and the complexity of the system, while retaining the same levels of accuracy. Finally, we propose a semantic fusion algorithm based on Transferable Belief Models, which can successfully fuse information from these two sensors, to form meaningful concepts and resolve ambiguity. We also analyze this system for robustness under various operating scenarios.

Acknowledgements

I would like to acknowledge the invaluable guidance and encouragement received from my advisor, Dr. Otman A. Basir, for the work in this thesis. His wisdom and enthusiasm in his work has been inspirational for me. I am greatly thankful to Dr. Susan Leat for her guidance and help during the course of this thesis.

I am also highly thankful to the readers of my thesis, Dr. Fakhreddine Karray and Dr. Susan J. Leat.

I would also like to thank the members of the PAMI lab, for sharing ideas and suggestions, and for a great learning experience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Human Computer Interaction

Intelligent machines and devices have been the focus of many researchers for the past few decades. Highly powerful computing devices have been introduced in all walks of life, from industrial robots to entertainment devices to intelligent vehicles. With the growing use of such technology, there is also a growing need for intelligent and more human-like behavior by these devices, to be able to perform in a real-world scenario.

Human Computer Interaction has received attention from researchers in various disciplines due to the above requirements. Considerable work has been done in developing and improving various modalities which can recognize, analyze and understand human thought and intent in a more natural and intelligent manner. These modalities can involve body gestures [1, 2, 3], speech [4, 5], and brain computing [6, 7]. Abstract thoughts and commands can be communicated to such interfaces more easily, which thereafter process them intelligently, using past heuristics and concepts.

It is becoming increasingly apparent that single input modalities cannot completely represent human intent accurately. This is because the meaning conveyed by an individual gesture may vary depending on the context within which it is used. Thus, recognition of human gestures is mostly an ill-posed problem, since there may not be a unique solution (concept) for a detected gesture. To condition or improve the problem, it is imperative to incorporate multiple modalities and to augment such modalities with the context and semantic cues during the understanding stage. In this respect, interfaces for multimodal interaction, which fuse information cues from different input modalities, can allow for a better representation of human intent and facilitate interaction with 'intelligent' computers, e.g a household service robot.

In this work a multimodal system is developed comprising of a Hand Gesture recognition system and a Brain Computing Interface, a multimodal fusion system that can extract information cues from multiple modalities and fuse them for the purpose of gesture/intent disambiguation, is discussed. This work also proposes and discusses the semantic fusion of multiple modalities. The contributions of this work are outlined in the next section.

1.2 Contributions of this work

1. *3D Hand Reconstruction from 2D input video* - An algorithm is proposed for the reconstruction of 3D hand pose based on a 27 Degrees of Freedom (DOF) hand model, given the static 2D view of the hand from an input video. The algorithm is computationally less intensive as compared to other standard hand reconstruction approaches and is suitable for real-world scenarios.

2. *Brain Computer Interface based on Steady State VEPs* - A Brain Computer Interface (BCI) has been developed which allows the user to make selections from a range of flickering options on a computer display. The system analyzes the EEG signals from the visual cortex to extract the Steady State Visual Evoked Potential (SSVEP). The SSVEP has a definite response at the frequency of the visual stimulus. This property has been used to identify the desired choice of the user.

3. *Multimodal fusion of two modalities for disambiguation* - A technique is proposed to improve the performance of the BCI. It is shown that this approach is efficient in a multimodal system utilizing information from hand gestures. The use of hand gestures can help resolve the ambiguity present in the BCI and as a result improve the performance and scope of applications of the system.

4. *Semantic fusion based on evidential reasoning* - A novel approach for semantic fusion of the two modalities is discussed, based on the theory of evidential reasoning. Evidences from different sources can be combined to form extended concepts, based on a pre-defined domain specific knowledge base. It is shown that the proposed algorithm can perform all the roles of multimodal fusion, as will be discussed in the next chapter.

1.3 Organization of the thesis

The thesis is organized as follows:

Chapter 1 discusses Human Computer Interaction and the scope and contribution of this work in the area of Hand Gesture Recognition, Brain Computing and Multimodal Fusion.

Chapter 2 presents the need for multimodal fusion, and discusses the various roles of any comprehensive multimodal system. It gives an overview of the current state of the art in multimodal fusion and also gives an overall view of the proposed multimodal system, the individual hand gesture and brain computing sensors, and the fusion process.

Chapter 3 discusses 3D hand pose reconstruction and the associated issues, and discusses the proposed algorithm for hand pose estimation given a 2D input video. The algorithm is run under varying levels of artificially generated noise to estimate its performance.

Finally, the system is also tested on a few real-world videos to observe its performance.

Chapter 4 presents a Brain Computer Interface utilizing SSVEPs. The system for the generation and acquisition of the response is described, and the experimental and analysis procedure are discussed. The chapter also analyzes the sensitivity of the system parameters. Parameter values for optimum performance are determined.

Chapter 5 discusses basic multimodal fusion for the two modalities described above. It is shown how a fusion system consisting of multiple modalities allows for greater flexibility. The BCI is implemented again using a smaller set of frequency values than the number of options. It is shown how the system can still make a correct decision, using the information from the hand gesture sensor. This also helps improve the information transfer rate of the modalities.

Chapter 6 explains the fundamentals of Transferable Belief Models (TBMs), and introduces a TBM based algorithm for fusing multimodal HCI cues. The fusion algorithm is tested under various operating scenarios in order to estimate the merit of the multimodal fusion. The chapter also discusses how the fusion process can be extended over time to make a scalable and comprehensive multimodal system, capable of understanding complex commands.

Chapter 7 concludes the thesis and discusses some future work worth pursuing to advance multimodal Human Computer Interaction.

Chapter 2

Multimodal System - an Overview

This chapter discusses the state of the art in multimodal fusion, and explains why a more comprehensive fusion algorithm is needed. Also a typical application scenario in Human Computer Interaction is shown and a multimodal fusion system utilizing Hand Gestures and Brain Computing is proposed. The proposed fusion algorithm is briefly discussed.

2.1 Multimodal Fusion

Multimodal Fusion has received attention in the past few years. The major roles of multimodal fusion are ‘Reinforcement’ and ‘Disambiguation’ (or Clarification). For example, humans use body gestures, like facial or hand gestures, along with speech to either reaffirm intent, and/or to clarify meaning. Two exemplary scenarios for the need for multimodal information, during interaction with a household service robot are:

- Commanding the robot to ‘Stand’. We can also use hand gestures to depict the same command. Here the role of the system is *Reinforcement*, since both input modalities imply the same command.
- Commanding the robot to ‘Go’, we can also use our hands or heads to point to the desired destination, since otherwise the robot cannot be sure about the intended destination. In this case, the hand or head gestures clarify the intention of the

user. This is *Disambiguation*, where ambiguity in one modality is resolved using the information from another cue.

Most of the current research in HCI concentrates on only one of the above roles. However, a comprehensive multimodal system should be able to perform both roles simultaneously. Another challenge before such a system is that the sensors may give incorrect or unreliable cues, due to noise or insufficient information. While the improvement of the recognition accuracy needs to be dealt with at the sensor level and after, any fusion system should be tolerant of such limitations to the maximum level possible.

Many researchers have proposed methods addressing the issue of multimodal fusion. A popular approach is the use of a rule based system for semantic fusion [8, 9]. Holzapfel et al. [8] have implemented a multimodal system for fusing speech and 3-D pointing gestures. They have defined a rule-based framework, using Typed Feature Structures for defining the semantic gestures. Fusion is performed based on the n-best lists generated by each of the parsers. In Mehta et al. [9], a rule based system has also been used for fusion in Human Computer interaction. A drawback of this approach is its assumption that the individual sensors can give an unambiguous output. Gupta et al. [10] have used a multichannel parameter fusion algorithm using a weighted mean technique to classify differential brain activities. Wu et al. [11] have addressed the issue of fusion by using independent component analysis, to reduce the overall dimensionality of the feature set. This is followed by a super-kernel fusion technique to fuse the individual classifier outputs.

Johnston and Bangalore [12] have proposed Finite-state models for integrating multimodal input from speech and gestures. However, the model does not address ambiguity in gesture events, and can only disambiguate cues in speech modality. Kaiser et.al. [13] describe an approach to multimodal fusion that accounts for the uncertain nature of information sources. Their system employs 3D gestures, speech and referential agents using Typed Feature Structures for multimodal interpretations. They report good disambiguation rates in an immersive virtual reality environment. An output is generated only if

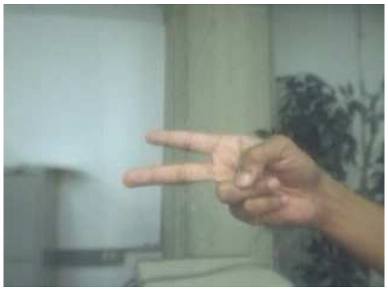

Figure 2.1: 'Victory' or 'Two'? Ambiguity remains even after accurate reconstruction.

the individual modalities produce outputs. This constitutes a limitation in a real-world environment, where it is quite possible that only one modality might be functional at any point in time.

A large body of fusion techniques are based on Bayes' theory. If the *apriori* and conditional probabilities of the input data set are known, then the *posteriori* probabilities can be computed. Chu et al. [14] have used a multistage fusion technique using a modified Bayesian classifier by adjusting the importance of the individual classifiers using exponential weights. Fusion is performed both at the feature and the decision level, thus enabling separability of the classes using minimum number of features. However, the product-combination technique suggested in their method is susceptible to noise. An adaptive Bayes network algorithm has been proposed by Varshney et al. [15] for use in management and fusion of biometric sensors. Lo et al. [16] use a Bayesian fusion technique for localization of a speaker using audio and visual cues. They assign reliability values to individual sensors, which are then considered during fusion. While this approach can take into account the possibility of a faulty sensor, it does not account for ambiguity in the sensor output.

Bayesian approaches work well when the *apriori* and the conditional probabilities are well defined. This might not always be the case, especially at an individual sensor level. Indeed, in most real-world scenarios of gesture recognition, even after accurate tracking and recognition, a single modality may still have ambiguity about the intended gesture, as shown in Figure 2.1. In this case most approaches assign equal probability to each $p(two) = p(victory) = x$. However, a more informative approach would be to assign the probability belief to the combined set $p[\{two, victory\}]$. This representation of uncertainty is difficult using orthodox Bayes' theory. An extension of the Bayes' theory, the Dempster-Shafer evidence theory [17] uses *belief* and *plausibility* values to represent the evidence and their corresponding uncertainty. These values represent how the uncertainty of a hypothesis increases or decreases as more and more evidence becomes available. The advantage of this approach is that we can work with incomplete, ambiguous or conflicting pieces of evidence. The DS-theory is being increasingly used for information fusion [18, 19, 20]. Yang et al. [20], have used DS theory to make timely recommendations on system identification. The algorithm is based on a Valuation-based system that allows for reducing of the computational effort. Ruthven and Lablas in [18] use DS theory for modeling the combination of evidence for use in relevance feedback in document retrieval. The DS-theory has also been used successfully to detect faults in mechanical devices [19].

However, the DS model fails to resolve mutually exclusive belief functions. This handicap was pointed out by Zadeh in 1986 [21]. As an extension of the DS theory, Smets introduced the use of Transferable Belief Models (TBMs) in 1990 [22, 23], which can account for the inconsistency in DS models. TBMs justify the use of DS theory in evidential reasoning, and the use of belief functions to model an individual's belief in the available evidence. It is not just an adaptation of probability theory, but a theory in itself considering both the 'credal' (dealing with 'beliefs') and 'pignistic' (dealing with actual probabilities) values when fusing information. It accepts the possibility that two pieces of evidence may be conflicting, can support or even extend each other.

In this work, a novel approach for multimodal fusion based on the theory of TBMs has been proposed. In this scheme, pieces of evidence from different sources can be combined to form new concepts which were not previously present. This is done based on a pre-defined domain specific knowledge-base or *ontology*, represented using *Conceptual Graphs (CG)*. CGs have been used by many researchers [24, 25] for representing universes of discourse, since they are a convenient way for knowledge representation. Over time, the successfully recognized concepts can be combined to form a coherent statement containing many complex ideas.

In this work, it is shown that the proposed approach is successful in satisfying the requirements of a multimodal HCI system due to the following reasons:

- Accurate representation of ambiguity in the outputs of individual HCI sensors.
- Capable of handling scenarios where one sensor or more may be non-functional.
- Efficient combination of evidences from multiple sensors to form a reliable decision.
- Scalability across time, for generating complex concepts, which allows more human-like interaction.

The proposed approach is employed in a multimodal system consisting of a *Hand Gesture Recognition System* and a *Brain Computing Interface.* This system can be especially useful for people with speech disabilities, helping them to interact with a service robot by means of hand signals and brain signals. Since even the current state of the art in these two fields can only recognize a few basic gestures at best, the fusion system would prove to be quite effective in fusing the information from these two modalities. It can also deal with the large ambiguity present in the outputs of the individual sensors, resolving what would otherwise require very complex recognition systems.

2.2 Application Scenario

The proposed multimodal fusion system consists of a Hand Gesture interface (HGI) and a Brain Computing interface (BCI). The gesture recognizer can understand pointing gestures, numbers and simple commands like sit, stand and eat, etc. The Brain Computing Interface (BCI) is more restricted and can only recognize a few basic concepts like eat, come, bring. The fusion system is designed in the context of a service robot assistant for people with speech disabilities. The robot assistant's task could be to help them in their daily tasks. For example, the user can command the robot to bring some item or perform some chore. These tasks typically require the use of both gestures and thought, since one modality may be inefficient in expressing the intent completely. A few typical interaction scenarios requiring disambiguation and/or reinforcement are presented (the modality used for signifying the gesture is shown in braces):

Example 1 (Reinforcement)
User: 'Stand' (BCI) + Hand gesture for 'Stand' (HGI)
System: The robot stands up

Example 2 (Disambiguation by HGI)
User: 'Come' (BCI) + Pointing Gesture (HGI)
System: Goes to user

Example 3 (Disambiguation by BCI) An object by the door
User: 'Bring' (BCI) + Points in the general direction of object/door (HGI)
System: Brings the vase

As seen from the examples, the fusion system combines information from both modalities, and the concepts from either one of the modalities can be used to reinforce, disambiguate or extend the other. This is unlike previous approaches, where one gesture is used as the reference or the base gesture, while the other simply serves to reinforce it [8, 13]. In the second example, the ambiguous 'Come' command is disambiguated by the Hand gesture recognizer which determines the desired destination. In the third example, when the user points towards the object, the system does not know whether they are pointing to the object or the door. However, combining the information provided by the BCI ('Bring') implies that the user wants the object, since 'Bring Object' is a much stronger and valid concept than 'Bring Door'. Thus fusing the two modalities eliminates ambiguity and also forms a more coherent concept.

2.3 Domain specific knowledge

The application scenario just presented requires the use of a domain specific knowledge-base in the form of Conceptual Graphs [25]. Conceptual Graphs are an effective method for knowledge representation using bipartite graphs [26], based on a combination of existential graphs and semantic networks. In this work simple conceptual graphs have been used without negation or nesting.

Conceptual Graphs and semantic structures are becoming increasingly popular in language understanding [24, 26, 27], since they serve as an intermediate step for translation of computer formalism and natural languages. Understanding is performed by comparing the detected input sequence and the conceptual graph, using graph isomorphisms and projection techniques. A comprehensive description of conceptual graphs is provided by Sowa in [26].

An excerpt from a pre-defined domain ontology is shown in Figure 2.2. A few concepts from the knowledge-base and their relational operators are shown. The knowledge base is easily extensible to bring in more concepts and their relation with the existing gesture primitives. Understanding can be achieved by matching a detected gesture sequence to the conceptual graph to identify the most probable command.

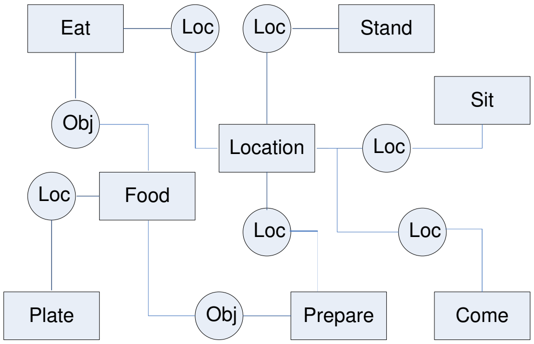

Figure 2.2: An excerpt from the pre-defined domain ontology.

2.4 Architectural View of the multimodal HCI system

The proposed multimodal system is shown in Figure 2.3. The individual inputs from both modalities are detected and pre-processed, and the corresponding features are then extracted and fed into the individual classifiers. In accordance with the Transferable Belief Models and Dempster-Shafer's theory of evidence, the classifiers assign belief values over the set of possible gestures. These belief values, lying between 0 and 1, signify the confidence of the sensor about a particular concept or set of concepts. The system is different from other classification schemes, in the sense that such systems tend to either output only the best matched concept (neural nets), generate n-best lists (typed feature structures), or compute a probability distribution over the set of gestures (Bayesian methods). The requirements and procedure for assigning belief functions will be explained in Chapter 6.

The advantage of using belief functions is that each individual classifier can assign beliefs for combined gestures. In Example three, the Hand Gesture Interface can assign a belief value of 0.3 for {object, door}, implying that it believes with a mass value of 0.3 that the current gesture is either 'object' or 'door', but cannot distinguish between the two. In Chapter 6, it will be shown how this ambiguity can be resolved by the proposed fusion

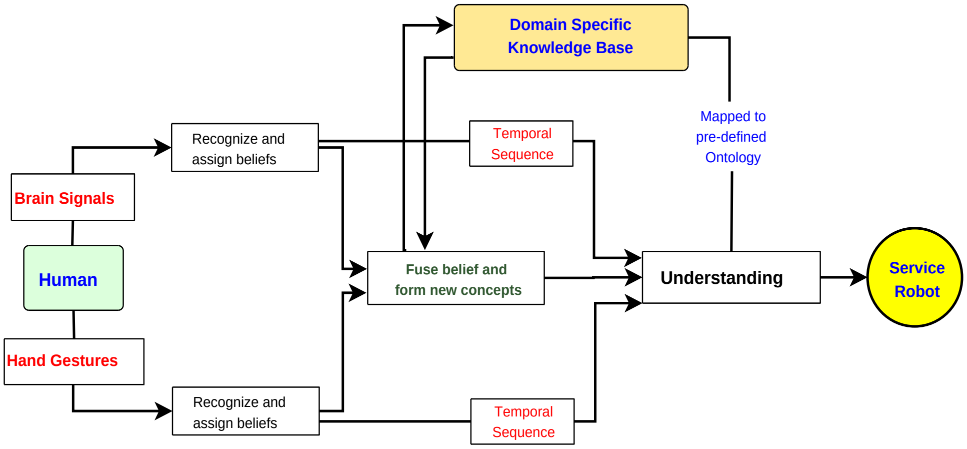

Figure 2.3: The multimodal fusion system comprising of the Brain Computing Interface and the Hand Gesture Recognizer.

approach. The sensor can also assign a belief value ‘Unknown’ state to signify that the sensor has no idea about the gestures, or that there is no input. This would be especially useful since one of the sensors might not register any gesture, while the other might be active. It is shown that the fusion algorithm would provide robust recognition for such cases.

The belief functions of the individual sensors are then input to the fusion system. The fusion process uses a pre-defined ontology of concepts, represented by ‘Conceptual Graphs’ [25, 27]. The gestures are combined to form new extended concepts only if they are deemed valid according to the Concept Graph (CG). In example three presented above, ‘Bring’ and ‘Object’ can be combined to form an extended concept, ‘Bring Object’, since they would be closely related according the conceptual graph for that system. ‘Bring Door’ would not be permitted by the CG, hence it would be discarded.

The final recognized concepts are used in the understanding stage to generate the entire sequence of fused concepts over time. Eventually, the robot should be able to understand extended and complex sentences, like “Bring the object and put it on the table”. It is important to note that this would involve the use of feedback to disambiguate those concepts that were previously recognized in time. The final action is evaluated by the user, who gives a positive (thumbs up) or negative signal (thumbs down), which is then used as feedback for the system to further reinforce itself. If this sentence is held to be valid and is ratified by the user, it is added to the vocabulary for future use, if not already present. Thus the ontology of concepts can also build itself over time. This stage is a topic for further research in multimodal fusion.

Chapter 3

3D Hand Pose Estimation

3.1 Hand Gesture Recognition

Hand Gestures are an integral part of communication between humans, and they can be highly useful in interacting with robotic systems. Many efforts are underway to build robotic vision systems which can understand these human gestures and perform the desired tasks efficiently, with highly encouraging results [2, 28, 29].

The major hurdle before hand gesture recognition is the fact that the human hand is highly articulated. The hand itself has 27 Degrees of Freedom (DOF) [30]. Combined with the movement of the elbow joints and the shoulders, the precise modeling of hand gestures becomes a very difficult task. Another problem with the analysis of hand gestures is the occlusion of one or more fingers, which makes the recognition process even more complicated.

Most approaches to hand gesture recognition use appearance or model-based techniques. Appearance based techniques try to map the detected contour in the current frame, to a previously trained database of gesture templates [2, 28]. While this approach leads to easy implementation, it does not provide specific information about the position and the shape of the hand, and as such is not suitable for manipulative environments, like a Virtual

Reality scenario. The model based techniques [30, 29, 31], on the other hand, attempt to track and estimate the position of the hand in an image sequence, followed by actual reconstruction of the entire hand.

Research efforts on model based reconstruction rely on the use of multiple cameras which take the image of the hand from various viewpoints [29]. Thus, the problem of hand pose estimation reduces to the standard case of reconstruction. Some other efforts use inverse kinematics techniques for computing the possible configurations of the hand. These techniques have provided good results. However, they are computationally expensive, or require more information than might be available, for example, multiple cameras from different angles might not be possible in the real-world environment.

In this work, an attempt has been made to avoid the complex kinematic techniques for some constrained hand poses, where the hand is parallel to the camera. An estimate of the 3D hand pose is made using the information from 2D images only. The human hand is analyzed, and the degrees of freedom reduced from 27 to 12 using the constraints proposed in [30, 31]. The reduction in the DOF is achieved without a significant loss in the quality of the results. Thus the problem of hand gesture estimation can be reduced to a two step process: extraction and segmentation of the feature points of the hand, namely the fingertips, palm, and the wrist, followed by the reconstruction of the hand model. At present, occlusion has not been specifically handled in this work, and it is assumed that an approximate 2D location of the fingertips and the wrist in the frame, is available to the system. In this chapter, firstly, an algorithm is proposed for extracting the hand from the input video frames, and the computing the feature points (fingertips, center of the palm, wrist). The various algorithms used for the above processing are well established techniques. In this work, a reconstruction algorithm is proposed and developed for estimating the hand posture, namely the joint angles for all the fingers, which can then be used for gesture recognition. The preliminary experiments on the proposed algorithm show that the proposed algorithm does indeed give good results.

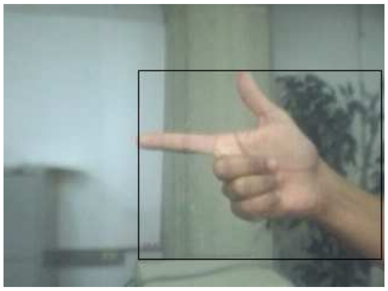

Figure 3.1: Using the bounding box method to detect the hand

3.2 Extraction and Segmentation of the Hand

The first requirement of any vision-based gesture analysis algorithm is the extraction and segmentation of the hand from the input video. For this purpose, a simple but effective segmentation algorithm has been employed for extracting the hand region and the identification of the various feature points of the hand. Each frame of the input video undergoes low level processing operations to extract the hand contour and the locations of the fingertips and the palm. Subsequently, the angles for all the joints are extracted. Once the information about the hand is available, it is possible to reconstruct the hand posture.

3.2.1 Hand Detection

The first step in the extraction process is to detect the hand in the video. This is done using a 'hue-based' color segmentation scheme, which has also been used by Herpers in [32]. Since the hue value of skin is quite invariant to lighting conditions, this is a reliable method for locating the regions of the frame containing the hand. The system is trained on the hue value of the user's skin. This trained value is subsequently used in the identifying and extracting the hand region out of the input video. The use of hue as a detection scheme has proved to be quite efficient and fast in extraction of the region belonging to the hand.

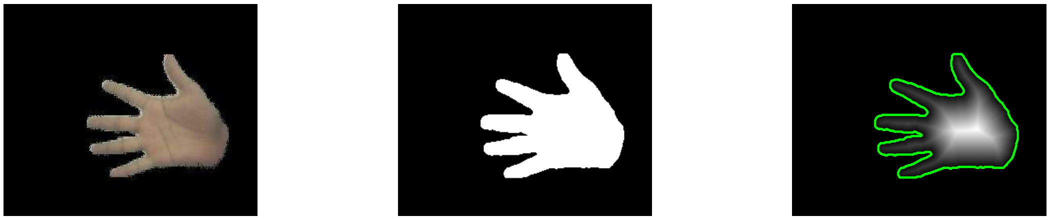

Figure 3.2: Extracting the hand region from the frame, generating a binary image, and computing the distance transform.

Once the possible hand regions have been identified, a 'Bounding Box' approach is used to find the hand region in the video. The frame is divided into equal segments and the segment with the maximum number of skin pixels is chosen. This is the 'Region of Interest' (ROI) for the hand region. Starting with this box, the algorithm begins growing the skin region till it encloses all skin-color pixels 8-connected with this region, as shown in Figure 3.1. All further processing occurs only on this ROI. This leads to a significant performance improvement as compared to processing the entire frame.

3.2.2 Morphological Operations

The ROI is then median filtered to remove any noise and to smoothen the image. Only the region comprising the skin pixels is extracted. The frame is converted to a binary image to reduce the amount of processing later on. To identify the various components of the hand, like fingers, palm etc., the distance of each pixel from the background is required. For this, the Distance Transform as proposed by Morris in [33], is used. The result of the Distance Transform is a gray-level image, where the intensities correspond to the distance of the pixel from the background. Since the Distance Transform is an expensive operation, performing it on only part of the image rather than the whole frame saves a lot of computation time. The results of above steps are shown in Figure 3.2. Using the Distance transform, it is straightforward to find the center of the palm. The palm can be described as the largest circular disk in the ROI [33]. Hence, the pixel with the peak value in the distance transform, or highest distance from the background is the center of the palm.

3.2.3 Locating the feature points

Now that the location of the center of the palm is known, the radius of the palm in the current frame can be computed. Starting with this center, circles with increasing radius are drawn, until a significant portion of the circle lies in the background. This gives the radius of the palm, to be used in the reconstruction of the hand. The palm can be approximated by a circle of this radius around the center of the palm. To find the locations of the fingertips first a contour of the hand is generated, which is achieved by a simple boundary trace algorithm.

Then, the procedure proposed by Malik in [1], is employed to identify the fingertips. For any contour point k, the angle between the vector formed by k and $k + n$, and the one formed by k and $k - n$, is computed. The value n is chosen depending upon the input video. If this angle is less than a certain threshold (in our case 60 degrees), then the point k is either a fingertip or a valley point. The actual fingertips can be easily identified by the fact that all the pixels lying between these three points should belong to the foreground. Once the fingertips are identified we traverse along the distance transform of the image along the line of maximum intensity, till the circle subtended by the palm. This gives the approximate location of the finger joints. Thus, the algorithm obtains information about the important feature points of the hand.

The above features are easier to compute as compared to matching the shape of the contour at that point, using B-splines, or other expensive techniques [1]. The results of the distance transform and the located fingertips are shown in Figure 3.3. One requirement of the algorithm is that the first frame in any input video should be an outstretched hand, with no occlusion. This is because the computed features (length of fingers, radius of the

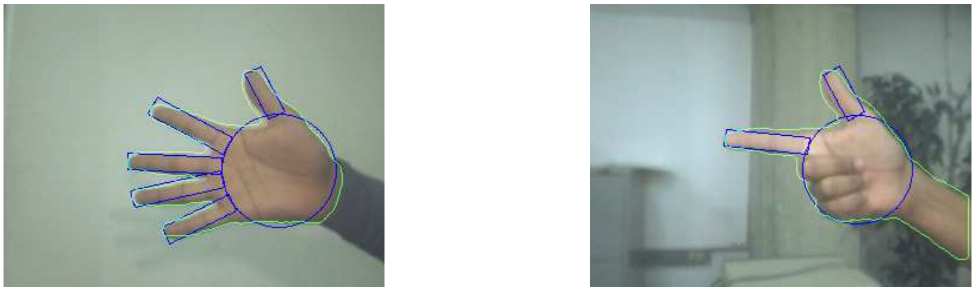

Figure 3.3: The generated cardboard model of the hand for an outstretched hand and gesture ‘one’. The contour of the hand and the rectangular model of the hand can be seen.

palm etc.) from the first frame are used as the reference values for future frames. These features are the lengths of the fingers and the radius of the palm, which are used by the reconstruction algorithm for determining the 3D posture.

3.3 Hand Reconstruction

The hand model used in this thesis has been suggested by Lee and Kunii [30]. The model assumes a 27 DOF hand model. The joints of the human hand are of three kinds: flexion, directive, or spherical, having one DOF (extension/flexion), two DOFs (one for extension/flexion and one for adduction/abduction), or three DOFs (rotation) respectively. The four fingers have four DOFs ($\theta_1 - \theta_4$), and the thumb has five DOFs ($\theta_1 - \theta_5$). The wrist has six Degrees of Freedom for the translation and the rotation respectively, making up the total 27 DOFs. Further description of these joints is given by Lee in [30, 31]. The hand model is shown in Figure 3.4.

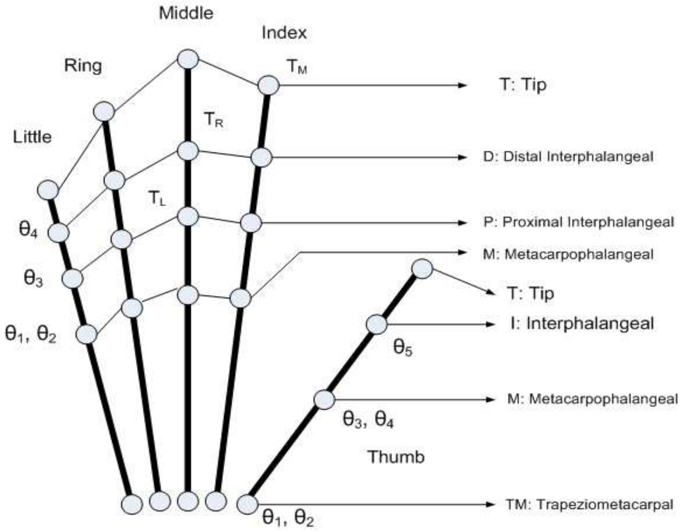

Figure 3.4: The hand model showing the 27 Degrees of Freedom

3.3.1 Constraints in the Hand Model

The large number of degrees of freedom of the hand makes it difficult to generate a hand model in computer vision applications for real world scenarios, where the available information may be insufficient for solving such problems. Thus it becomes necessary to exploit the inter-dependance of the joint angles in normal gestures. Chua et al. [31] have proposed some constraints on the dependencies between the joint angles. These constraints help reduce the degrees of freedom to a solvable system of equations, while causing minimum degradation in performance. The following constraints have been used:

1. It has been assumed in this work that the orientation of the palm plane is always known. The reconstruction of the finger poses also occurs with the palm plane as the reference. To facilitate this, the palm should be *always parallel* to the camera plane. While this imposes a big constraint on the free movement of the hand, there are still many gestures that are possible in spite of this restriction. Future work

would involve the prediction of the movement (translation/rotation) of the palm. This would enable any arbitrary orientation of the palm in the 3D space, allowing many more dynamic gestures. However the remaining constraints and the subsequent reconstruction process for the fingers would remain the same.

2. The thumb is more complicated, since it has more degrees of freedom. The constraints proposed on the thumb [31] are as follows:

$$\begin{aligned} \theta_1 &= 2(\theta_3 - \frac{1}{6}\pi) && (3.1) \\ \theta_2 &= \frac{7}{5}\theta_4 && (3.2) \\ \theta_5 &= k\theta_4 \qquad 0 \leq k \leq 1/2 && (3.3) \end{aligned}$$

   These help reduce the number of degrees of freedom for the thumb to 2. However, it is also assumed that the thumb can move only parallel to the camera plane. That is the flexion/extension angles $(\theta_2, \theta_4, \theta_5)$ are assumed to be zero for now. This restricts the thumb from overlapping the palm.

$$\theta_2 = \theta_4 = \theta_5 = 0 \tag{3.4}$$

3. The dependancy of the joint angles, for the four fingers, between the Distal (D) and Proximal (P) is given as:

$$\theta_4 = \frac{2}{3}\theta_3 \tag{3.5}$$

   Also the joint angles of the P and the M joints are related by the equation:

$$\theta_1 = k\theta_3 \qquad 0 \leq k \leq 1/2 \tag{3.6}$$

   These are widely accepted constraints and have been used by many researchers, as described by Chua in [31]. Together, they reduce the DOFs for each of the fingers from 4 to 2.

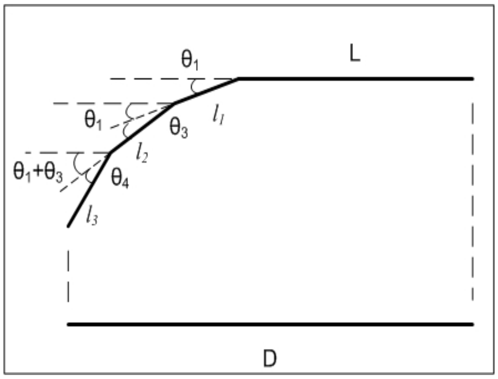

Figure 3.5: The finger model used for computing the joint angles.

4. It has been assumed that there is little adduction/abduction of the fingers. This is an extension of the constraint proposed by Lee [30].

$$\theta_2 = 0 \tag{3.7}$$

This constraint, in conjunction with the constraint in 3, reduces the DOFs for each finger to just 1.

5. The joints in each finger, represented by W, M, P, D, T, are assumed to lie in the same plane, referred to as the 'finger plane' [31]. This is a natural derivation from the constraint in 4, which ignores the abdution/adduction of the fingers. Thus M, P, D are all extenstion/flexion joints.

3.3.2 Reconstruction Process

The reconstruction algorithm approximates the 3D pose of the hand, based on the 2D information available from the input video. As can be seen from Figure 3.5, the actual feature points observed on the input frame are the coordinates of the joints projected onto the X-Y plane. The feature points extracted by the algorithm are the *(x,y)* coordinates of the fingertips, and the location of the wrist. Using this information, and the constraints proposed in the previous section, the hand pose can be reconstructed.

From Figure 3.5, the equation for the measured distance D, and the joint angles is given as:

$$D = L + l_1 cos(\theta_1) + l_2 cos(\theta_1 + \theta_3) + l_3 cos(\theta_1 + \theta_3 + \theta_4) \tag{3.8}$$

Incorporating the constraints of Equations 3.5 and 3.6, the following relation between D and θ_3 is obtained:

$$D = L + l_1 cos(\frac{1}{2}\theta_3) + l_2 cos(\frac{3}{2}\theta_3) + l_3 cos(\frac{13}{6}\theta_3) \tag{3.9}$$

In the above equation, D can be computed every frame (it is the 2D distance between the detected fingertip and the wrist) and L, l_1, l_2, l_3 have been computed beforehand. Thus the equation can be solved for θ_3. Thereafter, θ_1 and θ_4 can also be computed. Once all the joint angles are known, the 3D coordinates of all the joints, W, M, P, D, T can be easily computed.

The preceding analysis occurs parallel to the palm plane, as seen in Figure 3.5. Hence for complete analysis of the 3D pose, it is sufficient to know the 2D locations of the fingertips, and the orientation of the palm in space. A major feature of the algorithm compared to previous approaches [31], is that no other visual aids are used for the detection of the feature points. Thus there is no information about the spatial location of the feature points, unlike other approaches. This is one of the reasons for the requirement that the palm be parallel to the image plane, as mentioned in the constraints. Future work would look into computing the orientation of the palm from the input frame.

3.4 Experimental Results

In this section the reconstruction algorithm is evaluated. To evaluate the performance of the individual reconstruction algorithm, and the performance of the combined system, a two-step experimentation approach has been used:

- First, the reconstruction process is evaluated. For this, an artificial set of gestures is generated. These were exposed to varying levels of gaussian noise, and the resulting feature points (projected to 2D space) were input to the reconstruction module.
- Then, the system is tested on real world images, wherein the performances of both the hand extraction and reconstruction algorithm are tested.

3.4.1 Gesture recognition for an artificial environment

An artificial database of gestures was generated for the preliminary testing of the algorithm. The set of basic gestures used for testing were: *Open Hand, Two, Three, Four, Clenched Fist, Rock and Little finger extended.* These gestures were generated in 3D space and these were projected onto the 2D X-Y plane, to simulate an input 2D video. The resulting feature points were input to the algorithm for reconstruction. The list of feature points input to the system were:

- 2D coordinates (x, y) of the fingertips, in presence of noise.
- 2D coordinates of the wrist.
- Parameters of the hand, like lengths of the fingers and radius of the palm.

The algorithm returns the computed joint angles (in degrees) for each finger. Since all other parameters are known, the 3D hand pose can be reconstructed using this information and the Equation 3.9. The results of the reconstruction process are shown in Table 3.1 and accompanying process is shown in Figure 3.6, the generated 3D gesture, the input gesture (3D gesture projected onto 2D plane), and the final reconstructed 3D gesture. From Figure 3.6, it is seen that the reconstructed hand pose is very close to the original 'two' gesture in shape. However, the actual computed angles for the fingers $(45, 90, 60)$ are different from the actual values used in reconstruction $(72, 90, 72)$. This is because of the constraints used in the reconstruction process. This process will be suitable when the feature vectors of

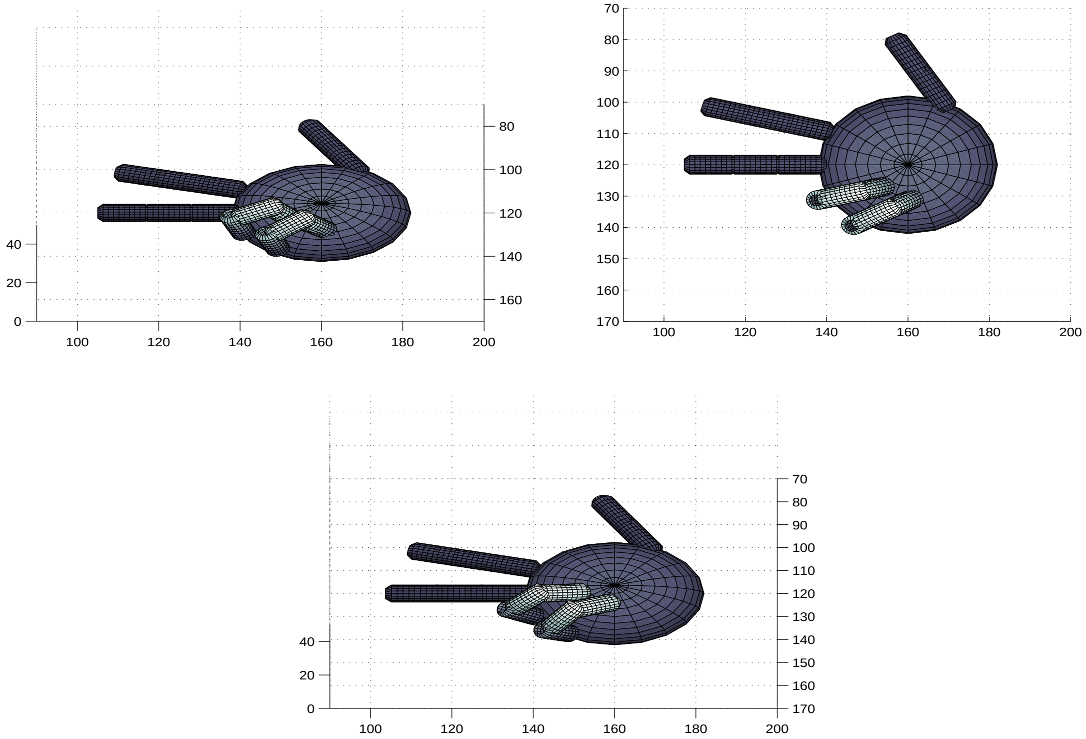

Figure 3.6: Reconstruction results for the ‘two’ gesture. The ‘Original gesture’, Input 2D frame and the reconstructed 3D hand pose

various gestures differ significantly, in that there is no other gesture similar to the one in question.

As a simple test for the robustness of the algorithm, the reconstruction process was evaluated under varying levels of noise. A Back-Propagation Feed Forward Neural Network was trained on the original set of gestures. The feature vector for each gesture consisted of the joint angles for each finger for that gesture. The test frames consisted of the gestures, which were projected onto a 2D plane and were subject to gaussian noise under varying levels of Signal to Noise Ratio (SNR), and then input into the reconstruction algorithm. The reconstructed values for the joint angles formed the test feature vector for each test

Gesture ‘Two’	Original (°)			Reconstructed (°)		
Finger	θ_1	θ_3	θ_4	θ_1	θ_3	θ_4
Index	0	0	0	0	0.001	0
Middle	0	0	0	0	0.001	0
Ring	72	90	72	45	90	60
Little	72	90	72	45	90	60

Table 3.1: Reconstruction results for the artificial gesture ‘two’.

Classification results (in %)					
Gesture	SNR (dB)				
	30	27	25	22	20
Open	93	76	67	70	65
Clenched	100	98	99	91	85
Two	96	88	73	91	75
Three	89	76	67	67	60
Four	97	83	82	90	88
Rock	87	62	44	45	28
Little Finger	98	93	87	55	26

Table 3.2: Results for the classification of gestures

case, and was input to the neural network for classification. From Table 3.2, it can be seen that the classification of the classifier is very good in the absence of noise, and the performance begins to degrade as the Signal to Noise Ratio goes down. In the real world, the measurements of the fingertips and the wrist would invariably contain noise. Thus it is necessary that the prediction and estimation processes be as accurate as possible even under high levels of noise.

3.4.2 Preliminary Experiments on real images

To test the robustness of the algorithm in real-world scenarios, the algorithm was also evaluated on a few real world images. The same set of basic gestures, as the previous section, was captured using an off-the-shelf web camera, and the input video of the gestures was captured at 10 frames per second at 320x240 resolution.

The first step in the estimation of the 3D pose is to extract the hand from the input 2D frame. The extraction process has been explained in Section 3. Following that, the feature points (fingertips, wrist) extracted from the input 2D images were input to the algorithm for reconstruction. The results for a few cases are shown in Figure 3.7. The extracted models from the 2D input video frame are shown on the right column. It is seen that all the non-occluded feature points are detected correctly, while the occluded fingertips have some error in reconstruction. A shortcoming of the algorithm is that it does not perform any prediction technique as yet. Hence it is unable to identify the fingertips that are occluded due to the hand. As a temporary work-around for this problem, it is currently assumed that the fingertips lie on the circle of the palm. This is a reasonable and conservative assumption under current conditions, considering no other information is available about the location of the feature points.

For example, if the middle finger is not detected, as in the ‘one’ gesture, it is assumed to lie on the circumference of the bounding circle of the palm. This location is then used by the algorithm for reconstruction. This assumption results in the reconstructed finger pose

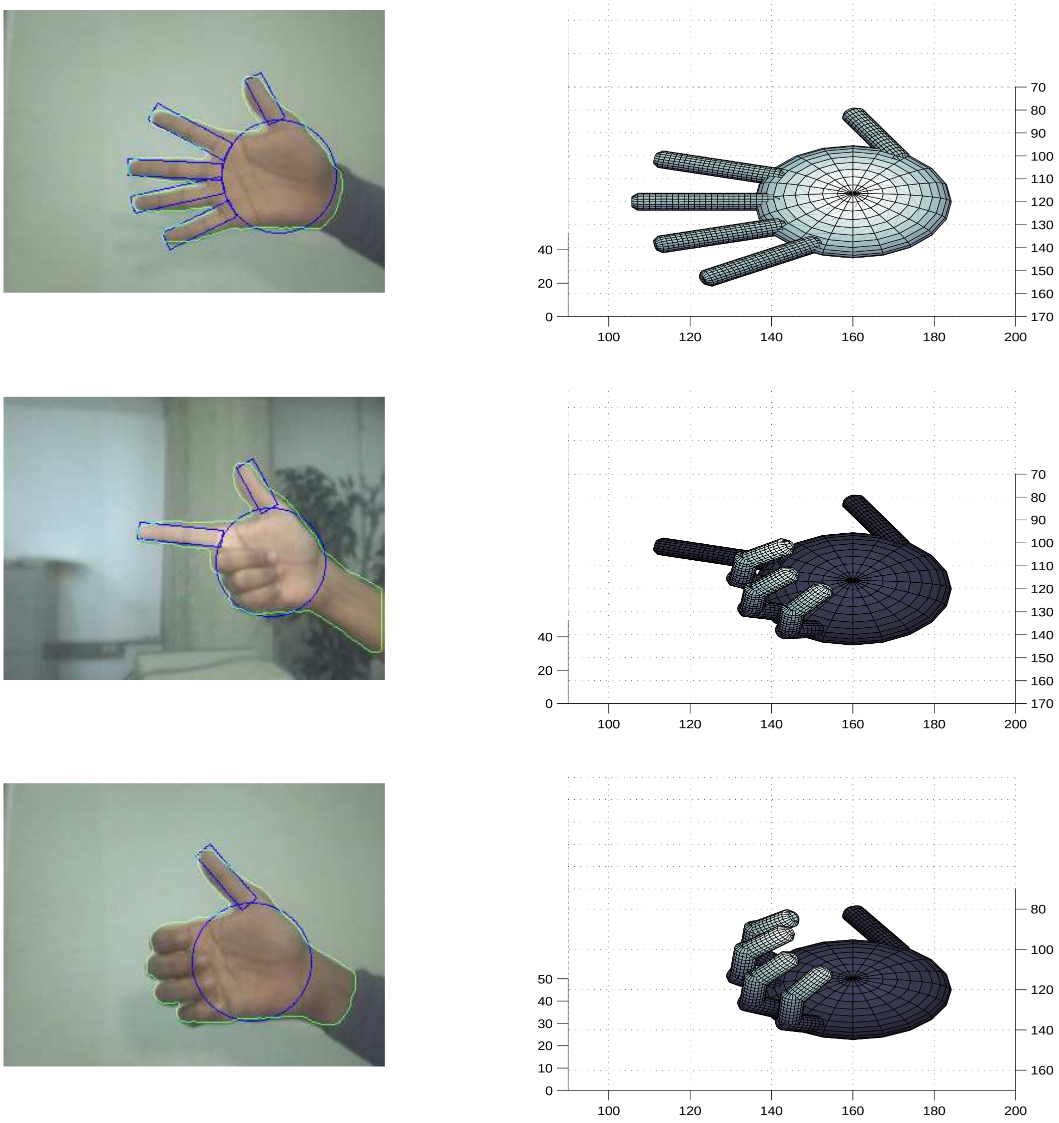

Figure 3.7: Reconstruction results for real world inputs.

to be *underestimated*, that is it always lags behind the actual finger pose, though not by a huge margin. Future work would look into incorporating Motion Prediction techniques, like Kalman or Particle Filtering, into the algorithm. This would enable the algorithm to perform better for real world scenarios.

3.5 Discussion

In this chapter, a reconstruction algorithm was proposed for the estimation of the 3D hand pose given the 2D input frame of a hand gesture. It is shown that the proposed algorithm performs quite well for static hand postures. The system is tested under varying levels of noise and it is seen that the algorithm functions well for real-world conditions. Also the system is evaluated with a few real-world gestures. It is seen that the reconstruction results are acceptable, given the constraints of the system. Future work for this system would involve incorporating the use of motion prediction techniques for recognition of temporal and spatial motion in all dimensions. Also the system would be evaluated further on real-world gestures under conditions of stress and noise.

The following chapters discuss a Brain Computer Interface, which can be used in conjunction with the Hand Gesture recognizer, to form a robust multimodal system. The proposed algorithm for fusion will also be discussed.

Chapter 4

Brain Computing using Visually Evoked Potentials

4.1 Brain Computing

Brain Computer Interfaces (BCI) have become a very popular area of research in the past few years. BCIs translate brain signals into control signals, without the need for any physical movement. This is done using non-invasive electroencephalography (EEG) from the various cortices in the brain. The two major types of BCIs are ‘spontaneous’ or ‘evoked’, depending on whether the system input is spontaneous or evoked EEG signals [34]. Spontaneous EEG signals are controlled by the user and do not depend on any external stimuli, e.g. muscular movements. These consist of μ-rhythms and slow cortical potentials. Evoked EEG signals are the signals generated in various cortices of the user in response to an external stimulus. These signals can be of different kinds, like P-300, Visually Evoked Potentials, Auditory potentials etc.

In this work, a Brain Computer Interface based on Steady State Visually Evoked Potentials (SSVEP) has been developed. SSVEP is a periodic response generated in the brain, in response to a repetitive patterned stimulus. The frequency of the response matches that of the stimulus, and extends over a narrow bandwidth. SSVEP is a reliable measure of user response, and has been used in many BCI systems for conveying commands or selecting options [35, 36]. The user makes a selection from the options displayed on the screen by concentrating on one of them, due to which the SSVEP shows a maxima at the target frequency. Flash-VEP (FVEP), on the other hand, are short responses to flash or OFF-to-ON or ON-to-OFF stimuli, which occur only once. FVEPs are time and phase locked to the flash onsets of the input stimulus, and thus can be used to detect occurrence of stimuli. Lee et.al. [37] use FVEPs to control the movement of a cursor on the screen.

SSVEP based measures have been used quite successfully in many BCIs over the past few years, mainly owing to high information transfer speeds and robustness. Cheng et al. [38] developed a telephone using SSVEPs. The system has good accuracy levels and is robust under various scenarios. Kelly et al. developed an independent VEP based BCI controlled by spatial attention [35]. The BCIs above have high information transfer rates, are quite robust and accurate, and very little training is required for them. However, it has been observed that low frequency-SSVEPs can cause fatigue and discomfort. A BCI based on high-frequency (21-45 Hz) SSVEP was proposed by Wang [39]. Techniques for reducing the high inter-user variation in information transfer rates were also proposed by them. These include channel selection, stimulus frequency, and trial length.

Kremlacek et al.[40] used damped oscillators to model pattern-reversal VEPs and motion-onset VEPs. An algorithm for accurate implementation of visual stimulators on generic computers using hardware counters has been developed by Jaganathan [41]. The frequency patterns are highly accurate to within 0.1% of the desired frequency.

Most of current research work in SSVEP based interfaces employs spectral analysis for classification of the response. Kelly et al. [42] used both parametric models and spectral analysis in the classification for BCIs. Greco et al. [43] proposed many techniques for filtering and the rejection of artifacts from EEG signals. Kramarenko and Tan [44] investigated the validity of spectral analysis in time-varying EEG signals, and have made some recommendations for the same.

In all the above works, using SSVEPs for brain computing requires the use of different stimulus frequencies for each of the control options. This would become computationally intensive on many systems, as the number of options increases [41]. Also increasing the number of different frequency values implies that the frequencies become closer to each other, or become multiples (harmonics) of each other, causing an increased error rate during the classification stage. Multimodal systems can help offset this issue by combining information from more than one modality. Thus a smaller set of frequencies can be used to show a large number of options with same frequency values. The other modality can be used to disambiguate the intended selection.

In this work, a multimodal system consisting of a Hand Gesture Recognizer and an SSVEP-based Brain Computing Interface is developed. It is shown that a BCI, implemented with a smaller set of frequencies than the number of options, can be disambiguated using hand gestures. The following sections propose a basic BCI system with two options, and analyze the performance and information transfer rates of the BCI under varying operating conditions. The next chapter analyzes the proposed system which requires a lower frequency set than the number of options.

4.2 The Brain Computer Interface

The experimental setup was designed to evaluate the performance of the user in a standard BCI, and compute the optimum values of the operating parameters. The details of the setup and procedure are explained in the following sections.

4.2.1 Experimental Setup

The experimental setup was comprised of a CRT display, for presenting the options on the screen, and a data acquisition system for capturing the EEG signals. Subjects were seated about 56 cm from the CRT display. The refresh rate of the monitor was selected to be 75 Hz. The stimuli consisted of two checkerboard patterns alternating at different frequencies. The checkerboard patterns were chosen since they have been found to produce the most pronounced SSVEP as compared to a simple flicker stimulus [35]. The two patterns were modulated at frequencies of 15 Hz and 25 Hz. These were the two frequencies that generated the maximum SSVEP response in the range 5 - 30 Hz, excluding the alpha band of 8-13 Hz. Also, choosing these frequencies avoided the alpha band of the subjects, lying between 8-13 Hz [35]. The setup is shown in Figure 4.1.

For recording the EEG signals, a Grass bio-potential amplifier was used followed by a National Instruments DAQ system to digitize the signal at sampling rate of 512 Hz. The EEG signals were recorded using a Grass gold electrode from the O1 or O2 positions on the primary visual cortex, based on the international 10-20 system [45]. The Oz position was not used because os higher levels of noise in initial empirical tests. The channel was referenced to the frontal site Fz, with the ground on the left earlobe. The measurement signal was amplified (50k), line filtered at 60 Hz to remove line noise, and then band-pass filtered over 1-100 Hz. This signal was then sampled by the DAQ unit and saved for further processing.

4.2.2 Experimental Procedure

The experiment was conducted to estimate the optimum values for the operating parameters, namely the size of the checkerboard pattern, and the time duration of each trial. Four subjects aged between 24 and 28 participated in the study. This study was approved by the Office of Human Research at the University of Waterloo. All had normal or corrected to normal vision. In the preliminary test, the subjects were instructed to stare at a

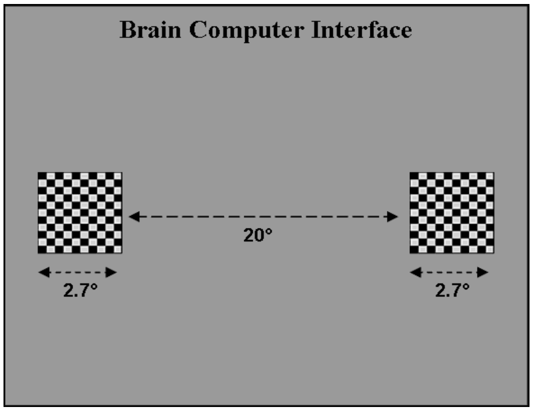

Figure 4.1: The Test Bed (stimulus) for computing operating parameters for the BCI (shown at checkerboard size of 100 pixels)

checkerboard pattern at various frequencies, ranging between 5-30 Hz. The two frequencies generating the maximum SSVEP, 15 and 25 Hz, were selected. The test consisted of two stages. In the first stage, the size of each checkerboard pattern was varied between 20 pixels to 120 pixels. The subjects underwent 10 trials for each size of the checkerboard pattern. In this stage the duration of each trial was 10 sec. In the second stage, the size of the checkerboard patterns was fixed at 120 pixels and the time duration was varied from 1 sec to 20 sec, with 10 trials for each value of time duration.

4.2.3 Feature Extraction and Analysis

In each trial, the subject was instructed to concentrate on one of the stimuli for the duration of the test. The EEG signals were extracted for the entire duration of each trial at a sampling rate of 512 Hz. The signal was filtered using a low pass elliptical filter at 60 Hz to eliminate the high frequency line noise. The time series data was divided into epochs of 512 sample each with a 256 point overlap between successive epochs. These segments denote 1 sec of continuous data with a significant overlap to ensure a smoother spectral analysis.

Three methods for feature analysis were used for classification of the time-series data. In the *first* method, the Fast Fourier Transform (FFT) was computed and squared for each time segment, and a single feature for each segment was computed [35]:

$$\begin{aligned} F(n) &= \log(\frac{X_n(f1)}{X_n(f2)}) \\ X_n &= (FFT(x_n(t)))^2 \end{aligned} \qquad (4.1)$$

where x_n is the nth segment, and $f1$ and $f2$ are the two frequencies of stimulation. It should be noted that this feature is a measure of power spectrum values at the desired frequencies. The set of values $F(n)$ form the feature vector for one trial.

In the *second* method, the autocorrelation function is calculated for each time segment, and then the FFT is computed for the autocorrelation function.

$$\begin{aligned} R^n_{xx}(t) &= E[x(t_0)x(t_0 - t)] \\ Y_n &= (FFT(R^n_{xx}(t))) \\ F(n) &= \log(\frac{Y_n(f1)}{Y_n(f2)}) && (4.2) \\ & && (4.3) \end{aligned}$$

where $R^n_{xx}(t)$ is the autocorrelation function of the nth segment, x_n. This method is more resilient to noise [35], since the autocorrelation of white noise is zero. The set of values $F(n)$ form the feature vector for one trial.

The *third* method involves parametric modeling of the EEG time series data, instead of spectral analysis. Autoregressive (AR) models, proposed by Kelly in [42], have been used to model the time series data. The order of the model was chosen as 5, since this provided the best approximation during initial empirical tests. The coefficients $a_1, a_2, ...$ of the AR model were used to form the feature vector for each trial.

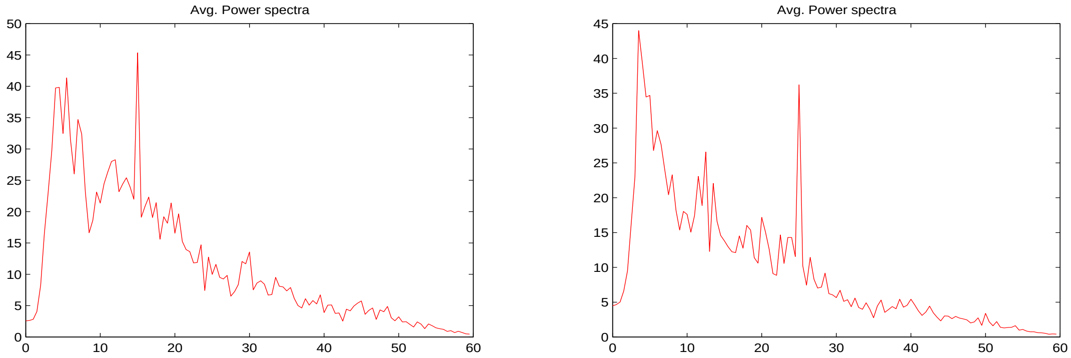

Figure 4.2: Power Spectra for a typical SSVEP response. The left and right figures are for the left and right gaze respectively.

4.2.4 Classification

Linear Discriminant Analysis (LDA) was used as a classifier in this experiment, since it is computationally quite efficient, and suitable for such applications [35]. The performance was evaluated using leave-one out cross-validation scheme [46]. The leave one out cross-validation scheme is a simplified version of the k-fold cross validation scheme. In this method, for n feature vectors, n LDA classifiers were trained with a different feature vector as the testing time, while the other $n-1$ feature vectors were used as the training set. The LDA classification scheme is run using the three feature extraction methods; the FFT, the Autocorrelation function and the Autoregressive models. The results of leave one out test are generally pessimistically biased, since the training has been done on a subsample of the data.

4.3 Results and Discussion

The power spectra for the steady state response for a representative subject are shown in Figure 4.2. It is observed that there is a maxima at 15 Hz for the left gaze, and similarly a maxima at 25 hz for the right gaze. Also there is a lot of contamination of the spectra

Performance for varying checkerboard sizes					
ChkBd (pix)	Subj 1	Subj 2	Subj 3	Subj 4	Avg.
20	50	70	50	40	52.5
30	50	70	60	80	65
40	80	80	80	60	75
50	90	80	80	90	85
70	100	80	90	90	90
100	100	100	100	100	100
120	100	100	100	100	100

Table 4.1: Variation of the accuracy with checkerboard size for Method 1

at the lower frequencies and at the alpha rhythm. Therefore, the classification was done on the basis of the relative magnitudes of the two desired frequency values, as described in Sec 4.2.3. The use of improved data acquisition equipment, along with the use of analysis techniques like Independent Component Analysis, can help in the removal of noise and suppression of the alpha band activity.

4.3.1 Classification Results

The results for classification using FFT are shown in Tables 4.1 and 4.2. From Tables 4.1 and 4.2, the classification accuracies for varying checkerboard size and time durations for all the four subjects are shown. It is seen that the accuracy of the system attains very high levels at a checkerboard size of 70 pixels. The time duration required for accuracy levels of over 90% is only 4 seconds, which makes the BCI very fast compared to other HCI systems. The optimum value for the checkerboard chosen for the future experiments in the next chapter was 100 pixels and 10 sec, to enable a steady settling time, and to assure the system of maximum accuracy possible.

The average results for the classification using the autocorrelation and autoregressive

Performance for varying time duration of the trials					
Duration (secs)	Subj 1	Subj 2	Subj 3	Subj 4	Avg.
2	70	80	70	50	67.5
4	90	100	90	90	92.5
7	100	100	90	100	97.5
10	100	90	100	100	97.5
15	100	100	100	100	100
20	100	100	100	100	100

Table 4.2: Variation of the accuracy with time duration of the trials for Method 1

models are also shown in Tables 4.3 and 4.4. The classification results using the three methods is shown in Figure 4.3. The plots for variation between checkerboard size and accuracy, and the time duration and accuracy, are shown. It can be seen that the method 2 using ACF, slightly outperforms the FFT method, though it comes at a cost of increased complexity. This is probably due to the noise reduction capability of the Autocorrelation function. Method 3 utilizing AR models is the worst performer, with classification accuracies much lower as compared to the other two approaches.

The Receiver Operating Characteristic (ROC) curves have been computed for the FFT and the ACF approaches, at an optimum checkerboard size of 100 pixels. The ROC curves are shown in Figure 4.4. It is seen that the curves are very high, both in terms of sensitivity and specificity. Thus both approaches are highly robust in discriminating between the two classes. It should be noted that the curves are slightly higher than expected because of the classification of the test cases is based only on the relative magnitudes of the frequencies at 15 Hz and 25 Hz.

4.3.2 Information Transfer Rate

An objective measure for BCI performance is the information transfer rate, or the bit-rate, proposed by Wolpaw [34]. For a trial with N possible options, P being the probability of a

Performance for varying checkerboard sizes			
ChkBd (pix)	FFT	ACF	AR models
20	52.5	57.5	52.5
30	65	75	67.5
40	75	77.5	65
50	85	82.5	80
70	90	97.5	77.5
100	100	100	75
120	100	100	80

Table 4.3: Classification accuracies for varying checkerboard sizes for the three methods

Performance for varying Test Durations			
Duration (secs)	FFT	ACF	AR models
2	67.5	70	52.5
4	92.5	92.5	82.5
7	97.5	100	85
10	97.5	97.5	85
15	100	100	90
20	100	100	92.5

Table 4.4: Classification accuracies for varying test duration for the three methods

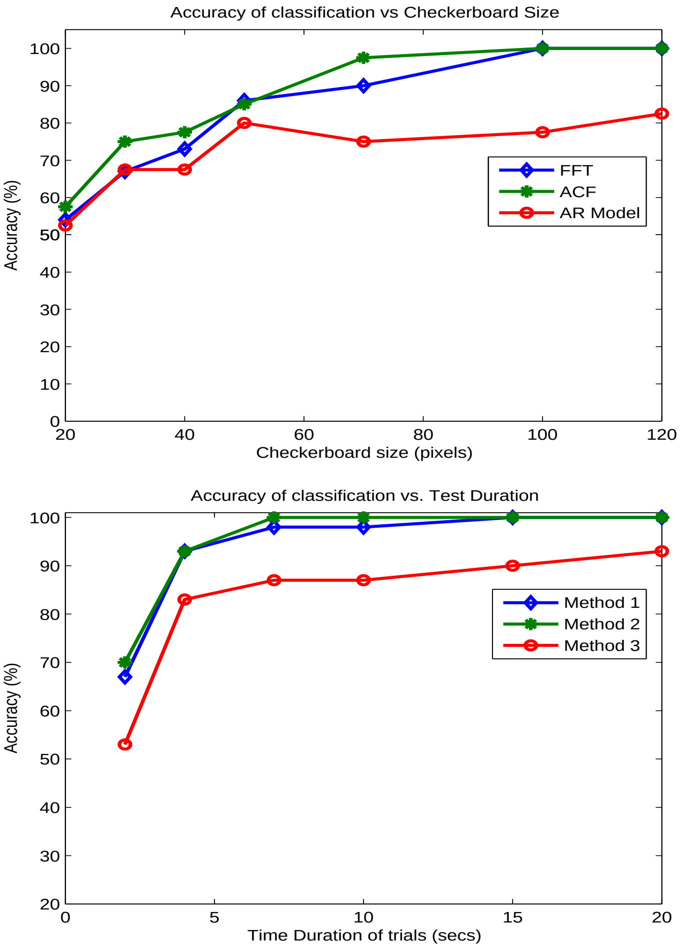

Figure 4.3: The variation of the accuracy levels with checkerboard size and time duration.

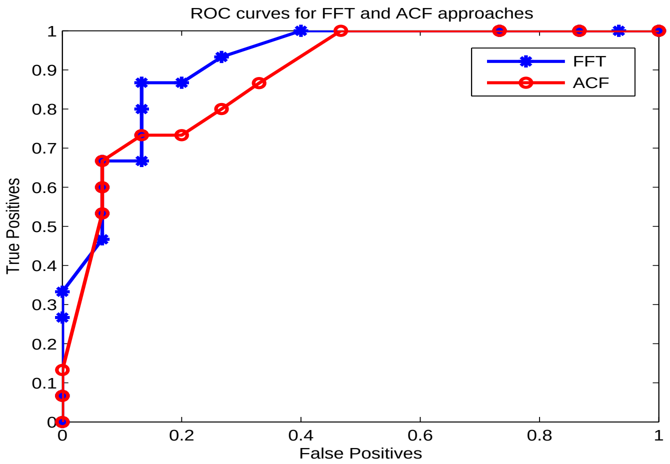

Figure 4.4: ROC curves for FFT and ACF methods

correct decision of the chosen option, and assuming each error to have the same probability, the bit-rate is defined as:

$$\text{Bits per Symbol} = log_2 N + P log_2 P + (1-P) log_2 \frac{1-P}{N-1} \tag{4.4}$$

$$\text{Bit Rate} = \text{Bits per symbol} * \text{symbols per minute}$$

To estimate the optimum operating parameter for maximum performance vs accuracy, the information transfer rates for the different values of the time durations for the trials is computed for the results of Method 1. From Figure 4.5, it can be seen that the most optimum operating parameter is at trials of 4 seconds duration each, where the information transfer rate is 9.54 bits/min. This is comparable to the information transfer rates computed by Kelly[35], or Wang [39]. It is also consistent with the earlier observation from the results that the accuracy of the system is already very high at 4 seconds.

4.3.3 Discussion

The results of the BCI developed in this work are highly encouraging, and show that it has very high rates of classification. There are a few observations about the EEG time series

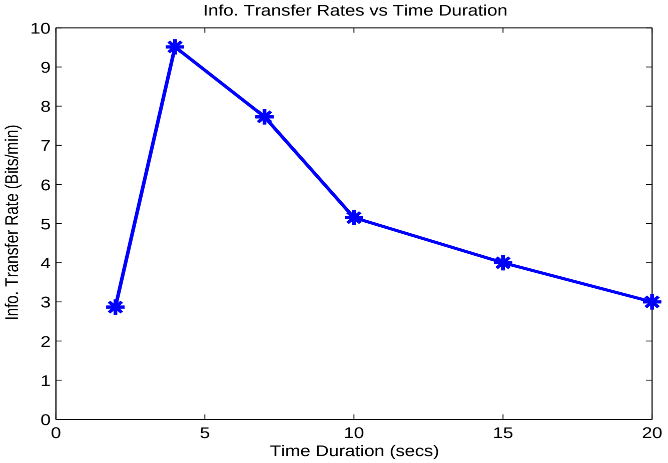

Figure 4.5: Variation of Information Transfer Rate with time duration

and analysis approach:

- The SSVEP response to a 15 Hz stimulus is stronger than the response to a 25 Hz stimulus. Thus a trained classifier has been used to determine a suitable threshold. This is also because the EEG signal in the absence of stimulus tends to be concentrated in the alpha band (8-13 Hz) and decreases at higher frequencies.

- The alpha rhythm of the subjects caused significant contamination of signals in many test cases. This is more prominent in the trials with longer time durations, due to arousal effects [35]. This is also a reason why the classification has been done on the basis of relative magnitudes of the stimulus frequencies only in this work. The effect of the alpha can be significantly reduced by the use of an comprehensive data acquisition system, where the EEG signals would be measured at multiple locations on the brain. Also the use of improved signal processing techniques like Independent Component Analysis and Entropy based techniques for rejection of artifacts [43] can improve the noise reduction of the system and allow for better characterization of the signals.

- The best operating parameters of the BCI are found to be 70 pixels for the checkerboard size and 4 seconds for the duration of the trials, for best performance at highest possible information transfer rate. In the next section, the BCI will be implemented with multiple stimuli at a lower set of values. The operating parameters are chosen as 100 pixels and 10 seconds, to minimize the effect of the stimulus size and the duration of the trial on the experiment. Also the use of hand gesture systems in improving the efficiency and information transfer rates of the BCI will be discussed.

Chapter 5

Multimodal System with Disambiguation

In this chapter, a multimodal system consisting of hand gestures and brain computing is proposed. It is shown how the performance of the SSVEP-based BCI developed in the previous chapter can be improved to allow for a greater number of options while using a lower set of frequency values. Disambiguation in such cases can be performed using information from the hand gestures.

5.1 Need for Disambiguation

In the previous chapter, the Brain Computer Interface was developed with each option flickering at a different frequency. Most systems developed today use the same concept of having a different frequency for each option [36, 38, 39, 41]. An issue with this approach is that it might become impractical when the system has to display a huge number of options to the user. Using a different value for the frequency can cause many problems with efficiency and implementation:

- The frequency values can become too close, making it harder for the system to identify the correct option from the EEG response.

- Care would have to taken such that the set of frequencies does not contain any values that are multiples of each other. This is because a SSVEP response causes a maxima at the frequency of stimulation and all successive even harmonics.
- The display of the system would become more computationally intensive, because of having to support a huge set of frequencies.
- The recognition algorithm would become more complex since it would have to become more accurate and robust due to the above reasons.

An efficient solution would be to use a multimodal system which can alleviate the demands on the brain computer interface, while retaining the robustness and performance. The proposed system implements a Brain Computer Interface, where the set of frequencies used for display is less than the number of options displayed on the screen. This implies that more than one option would give the same SSVEP response. Disambiguation between these options can be done using information from the Hand Gesture system. It has been assumed for now that the Hand Gesture System is accurate and always gives a correct result. It can be seen that the overall accuracy for the system will be bounded by the accuracies of the individual recognition systems.

5.2 The Brain Computer Interface revisited

The experimental setup is similar to the one described in Section 4.2. The refresh rate was again fixed at 75 Hz. The display in this experiment consisted of 4 flickering options in 2 rows, with the options in the top row having frequencies 15 Hz and 25 Hz. The options in the bottom row have the same values for the frequencies. It is expected that the SSVEP response would still show a maxima at the frequency of the desired option, though Disambiguation would be required to identify the correct row. This can be accomplished by a hand pose recognition system which identifies two basic gestures, denoting the row number. In this work, the hand gestures used to denote the top row and the bottom row are ‘One’ and ‘Two’. The setup for the BCI is shown in Figure 5.1.

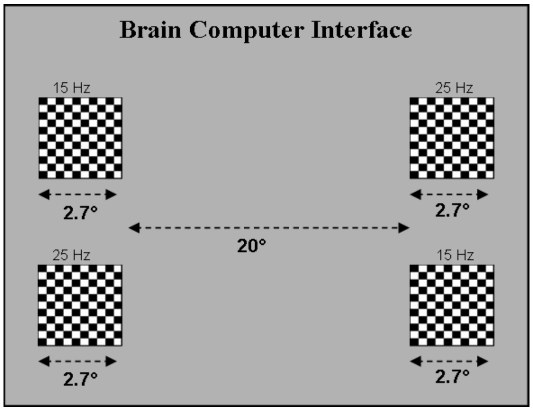

Figure 5.1: BCI displaying multiple options at a lower set of frequency values

5.2.1 Experimental Procedure and Display

Seven subjects aged between 22 and 28 participated in the study. All had normal or corrected to normal vision. The subjects were asked to concentrate on a chosen option for the duration of the trials. The two frequency values chosen were 15 and 25 Hz, and the size of checkerboard was fixed at 100 pixels and the duration of the trial was set at 10 sec. The distance between the various checkerboard options was varied in three steps to cover the inter-stimulus subtended angles of approximately 9°, 13°, 17° and 20°. There were 20 trials for each subtended angle.

EEG signals were acquired using the same setup as in Section 4.2. The signals were acquired at 512 Hz, followed by low-pass filtering, and spectral analysis using time epochs of 512 samples each, with an overlap of 256 samples. The feature extracted for each epoch was computed:

$$\begin{aligned} F(n) &= \log(\frac{X_n(f1)}{X_n(f2)}) \\ X_n &= (FFT(x_n(t)))^2 \end{aligned} \tag{5.1}$$

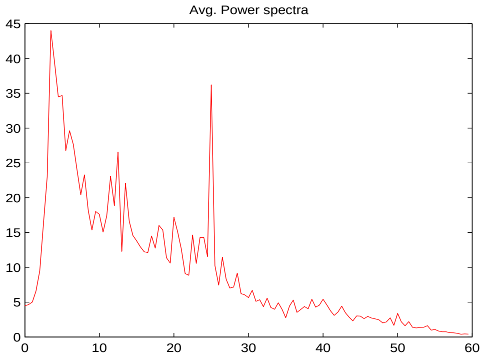

Figure 5.2: SSVEP response for stimulus at 25 Hz

where x_n is the nth segment, and $f1$ and $f2$ are the two frequencies of stimulation. All the computed values constituted the feature vector for this trial. Linear Discriminant Analysis was used as the classifier, using the leave-one-out cross-validation scheme, as described in the previous chapter.

The power spectrum of a representative SSVEP response is shown in Figure 5.2. It can be seen that there is a maxima at 25 Hz. Thus it is known that the subject chose one of the options which was flickering at the steady state frequency of 25 Hz. Hence the performance of the classifier is analyzed on whether it can identify the correct frequency value, and not the exact chosen option. The classification accuracies for the BCI are shown in Table 5.1. It can be seen that the classification results are quite high, with an average classification rate of over 90%. Also the variation of the accuracy of the system while varying the distance between the displayed options was studied. It is seen from Figure 5.3 that the accuracy remains very high even when the options are quite close to each other.

Performance for varying inter checkerboard distance				
Subject	Subtended angle between checkerboards (degrees)			
	9	13	17	20
Sv	70	75	90	90
Ak	95	95	95	100
Ad	90	100	100	100
Rj	100	95	95	100
Sr	95	100	100	100
Ab	90	95	90	100
Nv	75	75	80	90
Average	87.85	90.71	92.85	97.14

Table 5.1: Variation of accuracy with inter checkerboard distance

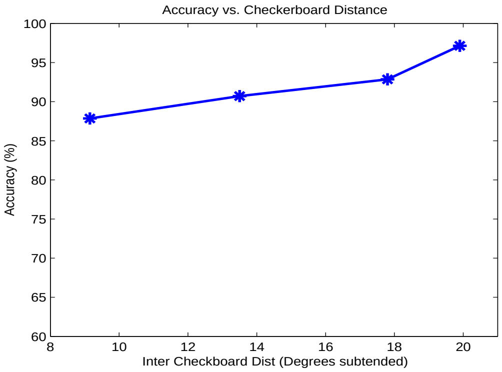

Figure 5.3: Variation of accuracy with inter checkerboard distance

5.2.2 Disambiguation using hand gestures

As outlined in the previous subsection, the classifier is very accurate in identifying the correct frequency value, but cannot distinguish between the various options. For example, consider a BCI system employing multiple gestures at each frequency value:

$$\begin{aligned} Options &= \{a, b, c, d\} \\ freq(a, c) &= 15\text{Hz} \\ freq(b, d) &= 25\text{Hz} \end{aligned} \tag{5.2}$$

If the subject concentrates on option A, the SSVEP response shows a maxima at 15 Hz. The LDA classifier can identify the frequency 15Hz accurately. There is ambiguity between the options A and C, which can be easily resolved using hand gestures. For example, the current system uses an off-line method for disambiguation, whereby the subjects can use hand gesture after the trial to indicate which row they were looking at. They can select the hand gesture ‘One’ to denote the top row, or the gesture ‘two’ to signify the bottom row. The two gestures are shown in Figure 5.4. Currently it has been assumed that the Hand gesture recognition system is totally accurate. This is due to the fact that the system is currently not real-time. Future implementation would involve making a complete multimodal system where the brain computing and hand gestures run in parallel.

Since, the hand gestures are assumed to be completely accurate, the accuracy of the multimodal system remains very high in theory. It can be seen that the actual performance of the multimodal system will be bounded by the individual performances of the BCI and the Hand Gesture recognizer. This is because the two modalities are independent of each other. Thus the observations and concepts in one of the modalities does not affect the other. To derive an upper bound on the actual performance of the system, the Bayesian Classification theory is employed [47]. The probability of a concept C is given as:

$$P(C) = P(c_1c_2|m_1m_2) \tag{5.3}$$

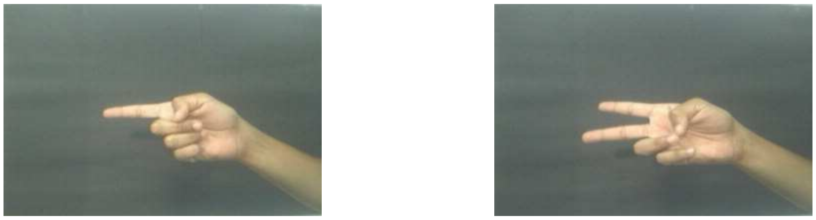

Figure 5.4: Hand Gestures used for disambiguation

where c_1, c_2 are the concepts (options) for the two modalities (in this work, BCI and Hand Gestures), and m_1, m_2 are the observations for the two two modalities in any trial. The two modalities are independent of each other, because selecting either of the gestures 'One' or 'Two' does not impact the SSVEP response, and vice versa. Hence the above equation reduces to:

$$P(C) = P(c_1|m_1)P(c_2|m_2) \tag{5.4}$$

This is simply the product of the performances of the individual modalities, since $P(c_1|m_1)$ is the probability of concept c_1, given observation m_1. The Hand Gesture recognizer discussed in Chapter 3 has accuracy levels of over 80% for the set of basic concepts. Thus such a system used in conjunction with the BCI just proposed would still have a very high rate of recognition.

5.2.3 Information Transfer Rate

The Table 5.1 gives the accuracy levels of the various subjects for the BCI implemented with 4 options using two frequencies. The advantage of this system is improved performance and efficiency at similar levels of accuracy. As discussed in Chapter 4, an objective measure for BCI performance is the information transfer rate, or the bit-rate, proposed in [34]. For a trial with N possible options, P being the probability of correct selection of the chosen option, and assuming each error to have the same probability, the bit-rate is defined as:

$$\text{Bits per Symbol} = log_2 N + Plog_2 P + (1-P)log_2 \frac{1-P}{N-1} \tag{5.5}$$
$$\text{Bit Rate} = \text{Bits per symbol} * \text{symbols per minute}$$

In the present system, only one symbol is identified per trial and each trials is of length 10 seconds. Given that the average accuracy of the system is 95%, the bit-rate is 11.339 bits/min. This is still higher than the bit-rate for the immersive gaming system developed by Kelly et al. [35]. Using a more realistic value for 4 seconds for each trial with average accuracy of 93%, as analyzed in the previous chapter, a bit-rate of 28.34 bits/min can be achieved, which is much higher than other reported bit-rates.

5.3 Concept based reasoning

In the previous section, the improved Brain Computing Interface has been proposed, which utilizes disambiguation from Hand gestures to improve the bit-rate of the BCI, while still performing at high accuracy levels. However, in this interface, while the Hand Gestures are mainly used for disambiguation, they can be used to provide more information, while performing disambiguation.

As a simple example for the above system, consider a multimodal system where the BCI has four options: *Start, Coffee, Tea, Stop*, while the Hand Gesture Interface can recognize the numbers *One, Two, Three, Four, Five*. The above proposed BCI can be easily implemented to accommodate this system in the following manner:

$$\begin{aligned} Options &= \{Start, Coffee, Tea, End\} \\ Freq(Start, Coffee) &= 15 \text{ Hz} \\ Freq(Tea, End) &= 25 \text{ Hz} \end{aligned} \tag{5.6}$$

The user can concentrate on any one of the four choices, and use the hands to provide supplemental information if necessary. For example, while commanding a robot for ‘Two Coffees’, the user would have to concentrate on the stimulus for the choice ‘Coffee’ and use the hand gesture for number two. The individual BCI sensor will be able to identify the stimulus frequency and that the concept is either *Coffee* or *Start*. This result can be combined with the information from the hand gesture which can recognize the concept *Two*, to disambiguate between *Coffee* or *Start*, and also generate the command sequence *Two Coffees*.

The above described process of combining information and concepts from multiple sensors can be extended to include a huge set of concepts linked together to form a knowledge base specific to the domain of real-world usage. This process, called data or knowledge fusion, is a widely researched area. In the next chapter, evidential reasoning will be explored as a fusion mechanism for the combination of concepts to reinforce or disambiguate each other.

Chapter 6

Evidential Reasoning in Multimodal Fusion

Multimodal interfaces allow for a better representation of human intent and can help in better understanding of the same. As discussed in the second chapter, the major roles of multimodal fusion are ‘Reinforcement’ and ‘Disambiguation’. In this chapter, the proposed fusion algorithm is presented. Evidences from different sources can be combined to form extended concepts, based on pre-defined domain specific knowledge base. The theory of evidential reasoning is presented, followed by the proposed fusion algorithm based on Transferable Belief Models. It is shown that the proposed algorithm can perform both the roles of multimodal fusion, as discussed previously.

6.1 Traditional Evidence Theory in Reasoning

The proposed approach is based on Transferable Belief Models (TBM), developed by Smets [22]. TBMs extend the Dempster-Shafer (DS) Theory of evidential reasoning. The notable feature of TBMs is the use of belief functions for representing a sensor's belief in an event. A belief in a proposition is not necessarily the same as the ‘probability’ of that event occurring. It is a representation of the confidence of the sensor about the event. The biggest advantage of this approach is that it allows for a comprehensive representation of uncertainty, as will be demonstrated in the following discussion. The basic notions of traditional TBM are discussed below.

6.1.1 Frame of Discernment

Let Θ be a finite set of hypotheses or gesture primitives in the ontology. This is referred to as the frame of discernment. The power set of Θ is denoted by $\Omega(\Theta)$. For example, if the set of gesture primitives is a, b, c, then

$$\Theta = \{a, b, c\}$$

$$\Omega(\Theta) = \{\phi, a, b, c, \{a, b\}, \{b, c\}, \{c, a\}, \{a, b, c\}\}$$

6.1.2 Mass Functions and Belief Values

A basic belief assignment (bba) is defined as:

$$m : \Omega(\Theta) \rightarrow [0, 1] \qquad \sum_{A: A \subset \Omega} m(A) = 1$$

This term can also be called the *basic belief mass*, or the *basic probability assignment.* The value $m(A)$ represents the belief that supports A, implying that the gesture is A, but does not support any specific subset of A, since there is no further evidence for any particular subset. If the sensor generates some belief that supports A, then $m(A) > 0$, but $m(B) = 0, B \subseteq A$. For example, the sensor may assign some belief value, say 0.3, to $\{a, b\}$, but 0 to the individual concepts a and b, if no further information is available regarding the two concepts.

The bba $m(A)$ itself does not represent the total belief in the gesture being in A. This is because the bba $m(B)$, $B \subseteq A$, also supports the gesture being in set A. Similarly, it can be seen that all the subsets of A contribute to the total belief in A. Thus the total belief in A is given by:

$$bel(A) = \sum_{\phi \neq B \subseteq A} m(B) \qquad A \subseteq \Omega, A \neq \phi \tag{6.1}$$

The mass function $m(\phi)$ is not included in A, since it does not explicitly support A, and also supports $\bar{A}$. Another useful function is the *plausibility function*, defined as:

$$\begin{aligned} pl(A) &= bel(\Omega) - bel(\bar{A}) \qquad \forall A \subseteq \Omega & (6.2) \\ pl(A) &= \sum_{X:X \subseteq A} m(X) \qquad \forall A \subseteq \Omega & (6.3) \end{aligned}$$

The plausibility values specify the maximum amount of support that can be given to a gesture. Thus the belief value of A can be transferred to some subset of A, if some new evidence becomes available.

The advantage of the DS and TBM models is that non-singleton subsets of Ω can also be assigned belief values. It can also be seen that if $m(A) = 0 \ \forall A \subset \Omega, |A| > 1$, then the TBM reduces to the standard Bayesian probability distribution. Indeed, this feature of TBM is important during the actual decision making process.

The important difference between TBM and traditional DS-theory is that TBMs do not have the restriction, $m(\phi) = 0$, which is required by the DS-theory. Quantifying ϕ is important in the proposed fusion system since the individual sensors might not register any gesture most of the time and hence the event of no gesture being detected should be represented as well.

Belief intervals, showing the interval $[bel(A), pls(A)]$, are also frequently used to demonstrate the amount of uncertainty in a concept. They show the difference between the amount of guaranteed belief and amount of possible support that should be given to an hypothesis. Some of the common intervals are shown in Table 6.1.

Interval	Interpretation
$[0, 0]$	No support at all to this hypothesis
$[1, 1]$	Total support to the hypothesis
$[0, 1]$	Absolute uncertainty in this hypothesis
$[0.3, 1]$	Tending to support the hypothesis
$[0, 0.6]$	Tending to disprove the hypothesis

Table 6.1: Various possible Belief Intervals

6.1.3 Evidence Combination

In the standard TBM and DS-theory, the confidence values input from the two sensors are combined according to the Dempster's rule of combination. If m_1 and m_2 are two basic belief assignments on Ω, then:

$$m(C) = \frac{1}{1-K} \sum_{A \cap B = C} m_1(A) m_2(B) \tag{6.4}$$

where

$$K = \sum_{A \cap B = \phi} m_1(A) m_2(B) > 0 \tag{6.5}$$

where K is the mass value associated with the null set, i.e. $m(\phi)$. It represents the amount of conflict between the various sources of evidence. The larger the value of K, the more the conflict. The above determined rule of combination can be extended to any number of sources of evidence. For n sources of evidence, the amount of conflict and the combined mass values are given by:

$$\begin{aligned} K &= \sum_{\cap_{i=1}^{n} E_i = \phi} m_1(E_1) m_2(E_2) \ldots m_n(E_n) && (6.6) \\ m(A) &= (m_1 \oplus m_2 \ldots \oplus m_n)(A) \\ &= \frac{1}{1-K} \sum_{\cap_{i=1}^{n} E_i = A} m_1(E_1) \ldots m_n(E_n) && (6.7) \end{aligned}$$

This property of easy scalability is one of the advantages of evidence theory. Any number of sources of evidence can be added to the multimodal system at a later stage, without a major change in the basic algorithm. Indeed, this feature makes evidential reasoning an attractive proposition for multimodal systems in Human Computer Interaction.

6.2 Combination of evidence in the proposed algorithm

In the traditional evidence theory, the final fused output mass function will be over individual concepts only. For generation of extended concepts, which is one of the goals of this work, the combination of evidence would have to be treated differently.

For example, let $m_1(A)$ and $m_2(B)$ be two mass values of evidence about some hypothesis ω. Using the traditional rule of combination, the new belief and plausibility values for $A, B, A \cap B$, can be obtained. Using these values, a decision can be made about ω. This is an example of 'Reinforcement' using fusion techniques, which has already been used by some researchers [19, 22].

The other case of multi-sensor fusion which should be handled is 'Disambiguation' and clarification. The system should be able to form an extended concept by combining individually identified concepts. This can occur when the two sensors recognize different concepts, which should be combined to form an extended gesture. Using the example in the previous sections, if A = 'Come' and B = 'Here', the desired output would be AB = 'Come Here'. However the previous rule of combination is not able to combine the two concepts, since using Equation 6.4 results in belief values for A and B individually, but not taken together. Hence the TBM theory has to be modified to account for this case also. This procedure requires the use of conceptual graphs as described in [24].

Let the Conceptual Graph be represented as a directed graph (V, E), where V is the set of vertices in the graph (representing various concept primitives), and $E \subseteq V \times V$ is the set of edges linking the various concepts. Then the modified *frame of discernment* is defined as:

- $A \in \Theta$
- $((A \times V \cup V \times A) \cap E) \in \Theta$

As an example, if the domain of conceptual graph consists of only two gestures Come' (a), 'Here' (b) , related as: $Come \rightarrow Here$ then the frame of discernment Θ and the power set $\Omega(\Theta)$ are:

$$\Theta = \{a, b, ab\}$$

$$\Omega(\Theta) = \{\phi, \{a\}, \{b\}, \{ab\}, \{a, b\}, \{b, ab\}, \{a, ab\}, \{a, b, ab\}\}$$

In the following discussion, concept $A = \{a, b\}$ implies that the gesture is either a or b. $A = \{ab\}$ implies the gesture is the combined concept tuple (a, b). $A = \{\phi\}$ implies that either the gesture can be anything from the *frame of discernment* or that it has not registered any input.

Based on the above discussion, the rule for combination can now be proposed. The combined piece of evidence ab is represented as a tuple (a, b), since generation of combined evidences can be seen as a set multiplication operation $\{a\} \times \{b\}$. Thus, the combination process for multiple sources of evidence, is defined as:

$$m(C) = \sum_{D=C} m_1(E_1)m_2(E_2)\ldots m_n(E_n) \tag{6.8}$$

where

$$D \subseteq ((E_1 \times E_2 \times \ldots E_n) \cap \Omega(\Theta))$$

The proposed multimodal system consists of two recognition sensors. Hence, the desired combination of two sources of evidence A and B simplifies to:

$$m(C) = \sum_{D=C} m_1(A)m_2(B) \qquad D \subseteq (A \times B) \cap \Omega(\Theta) \tag{6.9}$$

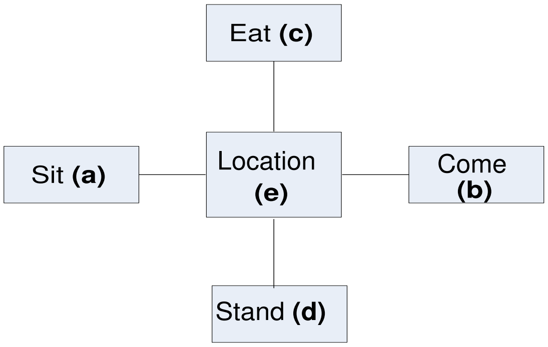

Figure 6.1: A simple conceptual graph relating some basic commands.

In both Equations 6.8 and 6.9, $D = ((A \times B) \cap \Omega(\Theta))$ selects only those combinations of concepts, out of the set product $A \times B$, that are permitted by the conceptual graph. To elucidate this process, an elementary ontology of concepts related by the conceptual graph in Figure 6.1 is shown. Assume that the two sensors A and B return confidence or mass values over concept(s) from the power set of the frame of discernment. The belief that nothing has been detected is represented by the '*' symbol. Thus, if the sensors A and B return belief distributions over the following subsets of $\Omega(\Theta)$:

A: {a}, {b}, {c}, {a,b},{*}
B: {d}, {a}, {e}, {d,a}, {a,e}, {*}

Before proceeding, the difference in the interpretation of the subset $\{a, b\}$ and the tuple (a, b) should be noted. $m(\{a, b\})$ is the mass value attached to the belief that the gesture is either a or b. $m(\{(a, b)\})$ or $m(\{ab\})$ is the mass value attached to the belief on the combined gesture ab.

Combination of the sensors' outputs gives us the following concept table, as shown in Table 6.2. It can be seen that new subsets and tuples have been formed in the fused output, while other concepts have been reduced or eliminated altogether. For example if sensor 1 believes, with some confidence, that the concept is $\{a\}$ and sensor 2 has some belief in

Sensor 1	Sensor 2					
	d	a	e	d,a	a,e	*
a	ϕ	a	ae	a	a,ae	a
b	ϕ	ϕ	be	ϕ	be	b
c	ϕ	ϕ	ce	ϕ	ϕ	c
a,b	ϕ	a	ae,be	a	a,ae,be	a,b
*	d	a	e	d,a	a,e	*

Table 6.2: Fusion of two sensors to form an extended concept table

the concept being $\{e\}$, then there is also some belief about the extended gesture sequence $\{ae\}$, along with the standard beliefs about the individual gestures $\{a\}$ and $\{e\}$.

To demonstrate the robustness of this combination rule, the row corresponding to concept $\{a\}$ from Sensor 1 is traversed, and the combination with various evidences from Sensor 2 is analyzed. Fusing $\{a\}$ with $\{d\}$ results in the null set, since they do not form a valid concept. This is an example of complete conflict. Considering both $\{a\}$ and $\{a\}$ is an example of perfect concord. Combining $\{a\}$ with $\{e\}$ forms a valid extended concept $\{ae\}$ according to the concept graph in Figure 6.1. The belief accorded to $\{ae\}$ from this fusion is the product of the individual mass values, $m_1(a)$ and $m_2(e)$.

Combining $\{d, a\}$ with $\{a\}$, or $\{d, a\} \times \{a\}$, results in only $\{a\}$, since $\{ad\}$ is not a valid concept according to the *CG*. Sensor 2 has ambiguity about the gesture being $\{d\}$ or $\{a\}$. If sensor 1 lays belief in the concept $\{a\}$, a multimodal system composed of the two sensors should lay more belief on the concept $\{a\}$. Thus, the first stage of disambiguation is performed by the fusion process itself. In the next case, $\{e, a\} \times \{a\}$ results in $\{ae, a\}$. This also agrees with the fusion system, since the combined concepts $\{a\}$ and $\{ae\}$ are both valid, so the system still has ambiguity as to whether the user intended to signal $\{a\}$ or $\{ae\}$, and it should lay the same belief to both concepts. Finally, the combination of $\{a\}$ with ϕ, results in only the concept $\{a\}$, since it implies that sensor 2 does not detect anything.

A few important observations about the proposed combination rule:

- Combination of a concept with the null set results in that concept itself. This stems from the definition of an *empty product* in Set Theory, which is defined as 1. Hence, $A \times \phi = A$.

- The order of the gestures has not been stressed. That is, the concept $\{ab\}$ is the same as $\{ba\}$

- The combination of a concept with itself results in the concept itself. This is equivalent to 'reinforcement'. Thus $\{a\} \times \{a\}$ results in $\{a\}$, which has the same semantic interpretation as the product result $\{(a, a)\}$.

- No normalization is done by the conflict factor 'K'. If two conflicting pieces of evidence dispute each other, the possibility of no valid gesture being present should also be represented. This can also be explained due to the 'open world' [23] model, where the detected gesture may lie outside the frame of discernment. If the open world model is assumed, then $m(\phi)$ need not be zero, since the null set also has a semantic meaning.

- The system is highly scalable with respect to the number of input sensors. Adding one more sensor requires little modification. The frame of discernment would have to be modified to incorporate the possibility of combining three gestures at any instant. The proposed rule of combination can easily incorporate any number of sensors, as given by Equation 6.8.

6.3 Decision making

While the combination was done over 'credal' or belief values, the actual decision making is done based on 'pignistic' or probability values, since the decision should be justified by probabilities [23]. Also, the final decisions are made on the individual concepts from the frame of discernment Θ, i.e. on the focal elements, rather than Ω, which may contain non-singleton sets. Thus the fused mass distribution has to be converted to a probability distribution over the original frame of discernment.

This transformation is called the *Pignistic Transformation*, originally proposed by Smets [23]. In the pignistic transform, the set of subsets from the power set $\Omega(\Theta)$ is mapped back to the the original frame of discernment Θ. It is defined as:

$$P(\omega) = \sum_{A:\omega\in A\subseteq\Omega} \frac{m(A)}{|A|(1-m(\phi))} \tag{6.10}$$

where $|A|$ is the number of elements of Ω in A. It is easy to show that $P(\omega)$ is a valid probability assignment on Ω. In this work the normalization by $(1-m(\phi))$ is not performed and the formula changes to Equation 6.11. This is done to preserve the value of θ. Using this transformation a probability distribution on the original frame of discernment is obtained which is followed by the actual decision making stage, based on highest probability value.

$$P(\omega) = \sum_{A:\omega\in A\subseteq\Omega} \frac{m(A)}{|A|} \tag{6.11}$$

6.3.1 Necessity of $m(\phi)$

It may happen quite often that the highest probability after the pignistic transformation is for the null set. This especially occurs when information from the two sources is conflicting. For example, if Sensor 1 lays complete belief in evidence A (belief value 1) and sensor 2 lays complete belief in evidence B (belief 1 again), then this is an example of complete conflict, which cannot be resolved by fusion approach as long as the frame of discernment is assumed to be a closed world. An example of this was given by Zadeh in 1986 [21].

Resolving this issue is possible only if:

- The reliability of the individual sensors is questioned. If it is possible to assign confidence values to the sensors themselves, these can be incorporated into the system to resolve any such issue. However, it is not a permanent solution since causes of conflict are still possible.

- The alternative argument is according to the open-world assumption at the *credal* level. It is possible that the actual concept might not lie within our knowledge of discourse. Thus the one or both the sensors can assign high belief to the 'null set' and the final probability values can also possibly be highest for the null set.

6.3.2 Evaluating the performance of the fusion system

The performance of the fusion process can be evaluated by the comparing the uncertainty of the system before and after the fusion process. A perfectly ideal system should have no uncertainty after the fusion process. However, that is rarely, if ever, true. The common ways of measuring uncertainty are *nonspecificity* and *conflict*, described in [48]. Nonspecificity measures the uncertainty in discriminating between various possible solutions, while Conflict measures the amount of disagreement between sources of evidence. There have been many measures of uncertainty proposed which compute these measures for various evidences. These are discussed by Harmenac [48]. In this work the measure proposed by Pal et al. [49] has been used where the uncertainty in a distribution is given by

$$E = \sum_{A \in \Theta} m(A) log_2 \frac{|A|}{m(A)} \tag{6.12}$$

where A is the subset of concepts over which the belief value is given. Since in the proposed approach the pignistic transformation has already been computed, $|A| = 1$, the formula reduces to the classical Shannon entropy:

$$E = \sum_{A \in \Theta} m(A) log_2 \frac{1}{m(A)} \tag{6.13}$$

The performance of the algorithm can be evaluated by computing the reduction in uncertainty for the system. This requires computation of the entropies of the system before and after the fusion process:

- After the fusion process the probabilities of the various possible concepts are specified by the pignistic values. Thus the uncertainty in the system is due to these possible concepts. The total amount of entropy, or uncertainty remaining in the system, after the decision making process, *Fused Entropy*, is:

$$E_{fused} = \sum_{C \in \theta} m(C) log_2 \frac{1}{m(C)} \tag{6.14}$$

- Before the fusion process the total uncertainty in the system is determined by the belief functions of the two sensors. Each of the Sensors assigns belief values for its own input modality. For the multimodal system discussed so far, Sensor 1 assigns belief values for a possible hand gesture, while Sensor 2 assigns belief values for a possible brain computing gesture. Thus the system has two unknown concepts (random variables), X and Y, for which the belief functions (basic probability assignments) have been specified. Hence the total uncertainty in the system is the *Joint entropy* of the two variables, and is given by:

$$E_{initial} = E(X) + E(Y|X) \tag{6.15}$$

  or

$$E_{initial} = \sum_{x} p(x) log_2 \frac{1}{p(x)} + \sum_{x,y} p(y|x) log_2 \frac{1}{p(y|x)}$$

  In the case of the proposed multimodal system the sensors 1 and 2 assign their beliefs independent of each other, i.e. the basic probability assignment of a sensor is made on the basis of information available to itself only. Hence, the computation of the total entropy reduces to a sum of the individual entropies.

$$E_{initial} = E(X) + E(Y) \tag{6.16}$$

Using the above formulae the performance of the fusion algorithm in resolving the ambiguity of the multimodal system can be evaluated. The next section analyzes the proposed algorithm under various operating scenarios and also the performance of the proposed algorithm.

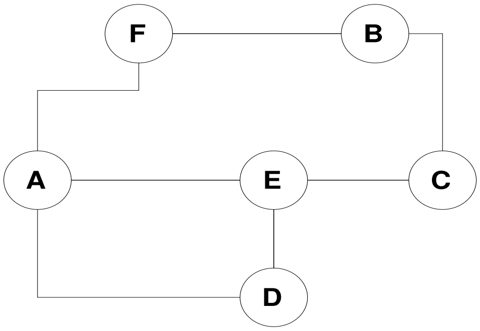

Figure 6.2: The Conceptual Graph used for the various test cases

6.4 Case Study

In this section the proposed approach is evaluated using synthetically generated test cases. Different test cases, which represent the various operating conditions intended to be handled in the multimodal system, have been simulated. It is seen that the proposed system can resolve all of the cases. The approach is compared with traditional evidence theory for both cases of reinforcement and disambiguation. The test conditions represent the following cases:

- One of the sensors non-functional
- Conflict between the sensors
- Reinforcement, when sensors are in harmony
- Disambiguation, when one of the sensors has ambiguity. Both cases are shown, when disambiguation is successful and when it is not.

6.4.1 Combination of evidence

The domain of knowledge, represented by the Conceptual Graph (CG) shown in Figure 6.2, contains some basic gestures which are related to each other as shown. Different possible

Sensor 1		Sensor 2		Fused System			
concept	m(A)	concept	m(B)	Fused	m	bel	pls
a	0.7	*	1	a	0.7	0.7	0.7
*	0.3			*	0.3	0.3	0.3

Table 6.3: Only one sensor registers an input.

Sensor 1		Sensor 2		Fused System			
concept	m(A)	concept	m(B)	Fused	m	bel	pls
a	0.7	b	0.8	a	0.14	0.14	0.14
*	0.3	*	0.2	b	0.24	0.24	0.24
				*	0.62	0.62	0.62

Table 6.4: Resolving a conflict between the two sensors.

test cases from the multimodal system, consisting of the hand gesture recognizer (Sensor A) and the Brain Computing Interface (Sensor B), were input to the system.

1. **Only Sensor 1 is functional -** This situation can occur if sensor 2 is faulty or does not register any gesture at present. This case is represented in Table 6.3, where sensor 2 is not functional at present. Sensor 1 assigns mass values to the concepts it has some belief in (a mass value of 0.7 for the concept $\{a\}$ and a belief of 0.3 that the concept cannot be recognized). Sensor 2 just assigns mass value of 1 to the 'null set', since it is nonfunctional, or does not detect any input at all. It is expected that the decision will be based on sensor 1 only, since sensor 2 does not register any input. The fusion algorithm also returns the same results, with the final fused values having the same mass distribution as that of sensor 1.

2. **Outputs of the sensors are disjoint** This represents the case where the outputs of the sensors are in conflict, as shown in Table 6.4. Sensor 1 lays belief in the concept $\{a\}$, while sensor 2 lays belief on the concept $\{b\}$. This leads to a conflict as these

Sensor 1		Sensor 2		Fused System			
concept	m(A)	concept	m(B)	Fused	m	bel	pls
a	0.7	e	0.8	a	0.14	0.14	0.14
*	0.3	*	0.2	e	0.24	0.24	0.24
				ae	0.56	0.56	0.56
				*	0.06	0.06	0.06

Table 6.5: Both Sensors are highly coherent and can be combined to form an extended concept

concepts are mutually exclusive and cannot be concatenated to form an extended concept. After fusion, it is seen that the highest mass value is actually given to ϕ, since both sensors are in high conflict. At present, the decision is made in favor of $\{b\}$, which has the next highest pignistic value.

3. **The Outputs are highly coherent**. This is the ideal scenario, shown in Table 6.5 where the outputs of the sensors are related perfectly according to the concept graph. Sensor 1 has high belief in the concept $\{a\}$, while sensor 2 has a very high belief in concept $\{e\}$. Since these two form a very valid extended concept, according to the specified ontology *CG*, they can be combined to form the concept $\{ae\}$. Thus, the highest confidence should be given to $\{ae\}$. As is seen from Table 6.5, the highest belief is assigned to the combined gesture $\{ae\}$. Thus the case of coherent sensors is handled well.

4. **Ambiguity in Sensor 2 resolved by Sensor 1**. In this case, sensor 2 has a lot of ambiguity between the concepts $\{c\}$ and $\{f\}$, as shown in Table 6.6. The decision whether the concept is $\{c\}$ or $\{f\}$ cannot be made on the basis of sensor 2 alone. After fusion, it is seen that sensor 1 can help perform disambiguation, since $\{af\}$ is a valid concept while $\{ac\}$ is not allowed by the concept graph. Performing the

Sensor 1	Sensor 2			
	f	c	c,f	*
	(0.3)	(0.3)	(0.2)	(0.2)
a (0.5)	af	ϕ	af	a
d,e (0.1)	ϕ	ec	ec	d,e
b,c,f (0.2)	f,bf	c,bc	c,bc,bf,f	b,c,f
* (0.2)	f	c	c,f	*

Fused Results				Pignistic Values	
Concept	mass	bel	pls	A	P(A)
af	0.25	0.25	0.25	**af**	**0.25**
a	0.1	0.1	0.1	a	0.1
ec	0.05	0.05	0.05	ec	0.05
d,e	0.02	0.02	0.02	d	0.01
f,bf	0.06	0.12	0.24	e	0.01
c,bc	0.06	0.12	0.24	bf	0.04
c,bc,bf,f	0.04	0.32	0.36	f	0.133
b,c,f	0.04	0.2	0.36	c	0.133
f	0.06	0.06	0.24	bc	0.04
c	0.06	0.06	0.24	b	0.013
c,f	0.04	0.16	0.36	*	0.22
*	0.22	0.22	0.22		

Table 6.6: Fusion of two sensors to form an extended concept table, using our combination rule. Sensor 1 helps disambiguate sensor 2

Sensors		Fused Results				Pignistic Values	
A	B	Concept	mass	bel	pls	C	P(C)
a(0.1)	a(0.65)	a	0.173	0.173	0.210	a	0.191
e(0.35)	c(0.1)	ad, ae, af	0.015	0.470	0.510	ad	0.007
f(0.35)	d,e,f(0.15)	af	0.227	0.227	0.283	ae	0.234
a,f(0.05)	*(0.1)	f	0.087	0.087	0.123	**af**	**0.251**
*(0.15)		ea	0.227	0.227	0.250	ec	0.035
		ec	0.035	0.035	0.035	ed	0.026
		ed, e	0.052	0.087	0.110	e	0.069
		e	0.035	0.035	0.110	f	0.099
		a, af	0.033	0.433	0.460	c	0.015
		ad, ae, af, f	0.007	0.565	0.625	d	0.007
		a, f	0.005	0.265	0.328	*	0.065
		c	0.015	0.015	0.015		
		d, e, f	0.022	0.145	0.210		
		*	0.065	0.065	0.065		

Table 6.7: Ambiguity even after fusion. The maximum belief is for subset {f,ad,ae,af}.

pignistic transformation, the highest probability is assigned to the concept $\{af\}$, as expected. Thus the fusion process can help resolve ambiguity in individual sensors and result in a more coherent and extended concept.

5. **Ambiguity even after fusion**- Shown in Table 6.7, there is high ambiguity between the concepts $\{e\}$ and $\{f\}$ in sensor 1. The problem is heightened by the fact that the concept with the highest belief in sensor 2, $\{a\}$, forms valid concepts with both $\{e\}$ and $\{f\}$. In this case it is expected that there will be high ambiguity between possible concepts $\{ae\}$ and $\{af\}$. It can be seen from the algorithm that, after applying the pignistic transform, the concept $\{af\}$ has the highest probability value, though only marginally greater than $\{ae\}$. Hence this concept is chosen as the output though the

difference in the probability values is much less. Though there is still ambiguity in the system, it can be seen that the overall uncertainty in the system has decreased after the fusion process, since belief in the other candidate concepts has been further reduced.

6.4.2 Evaluating performance of the system

The performance of the fusion algorithm for the above cases can be evaluated using the *Entropy measure*, discussed in the previous section and given by the Eqns. 6.14 and 6.16. The computed Entropy of the system, before and after fusion, for the above test cases is shown in Table 6.8.

It is seen that the entropy is reduced for the cases 2, 4 and 5, where there was a lot of ambiguity in the system before fusion. The entropy of the systems decreases considerably after fusion for these cases, allowing for a more informed and better decision. The uncertainty remains the same for test cases 1 and 3 even after fusion. This is because in the test case 1, only one sensor is functional, Sensor 2 does not register any input. Hence the entropy should remain the same even after fusion, since no further information has been added to the system by sensor 2. For test case 3, the sensors are highly coherent, and all possible combinations of the two sensors result in valid concepts. Thus the entropy remains the same. However, the final system fuses the two inputs to form an extended concept, which has more semantic information than the two individual gestures.

In all of the above cases, the fusion system is able to resolve the ambiguity between the concepts satisfactorily. An interesting observation from the results is that the null set ϕ usually gets a very high mass value. While this has been simply ignored in most other works [23], it has been preserved until the final decision making process. This is because of the open-world model. Since both the modalities consist of only a small set of gestures out of the immense set of gestures that humans use in day to day activities, it is the possible that a gesture may not be present currently in the database. Also since both sensors are

Test Case	$E_{initial}$	E_{fused}
1	0.6108	0.6108
2	1.1112	0.9141
3	1.1112	1.1112
4	2.5867	2.0042
5	2.4244	1.9412

Table 6.8: The entropy of the system before and after fusion.

highly susceptible to noise, often they might not register any input or may fail to detect the gesture.

6.4.3 Comparison with traditional DS-theory

The same set of cases was also tested with traditional Dempster theory of combination. The results are presented in Table 6.9. The traditional Dempster rule of combination does not fuse the first case of no input, since it does not allow $m(\phi)$. There is no distinction made for cases 2 and 3, and in both cases, the highest valued concept is output as the fused concept, whereas the proposed approach makes use of the fact that the concepts can also be combined to form extended concepts, while at the same time also preserving some belief about the individual concepts also. In test case 4, the ambiguity between concepts c and f is still present, even after combination. For test case 5, though the final belief on a is high, the system has not combined the two sensors to form the extended concepts af and ae.

As can be seen from the results, the proposed rule for combination is better at combination of gesture primitives from multiple sensors and forming extended concepts. This is encouraging, since it shows that the proposed algorithm performs favorably as compared to other inference and evidential approaches, like DS-theory and Bayesian approaches.

It has been shown that DS-theory and Bayesian theory follow similar trends in performance in sensor fusion. Bayesian methods give comparable results as dempster shafer

Traditional DS-theory on above test cases				
1	2	3	4	5
-	a 0.37	a 0.37	a 0.208	**a 0.483**
	b 0.63	e 0.63	d 0.021	e 0.22
			e 0.021	f 0.244
			f 0.36	c 0.034
			c 0.36	d 0.016
			b 0.27	

Table 6.9: Comparison with the Dempster Shafer theory.

theory when applied with suitable modifications. This implies that bayesian theory performs well when the DS-theory performs well, and vice versa, with an edge for dempsters theory [50]. Bayesian fusion would give slightly worse results as that of the Dempster-Shafer based approach, and does not perform disambiguation as desired. Thus, the proposed algorithm is more robust at representation and combination of concepts as compared to these approaches.

6.5 Temporal fusion for command sequences

The currently proposed approach can perform multimodal fusion at fixed time stamps. To enable a system based on this approach to function in a real-world scenario, the multimodal system needs to be able to understand complex concepts and command sequences over time, like “Bring object and keep it on the table”.

The proposed solution, currently under research, is based on the approach of Miners et. al. [24], which performs disambiguation of unimodal hand gesture sequences using conceptual graphs. All the recognized gesture primitives over time form the set of concept sequences. Understanding is achieved by finding the closest match between a subset of the existing knowledge base and the given set of concepts. This technique was shown to improve the probability of identifying the intent of communication.

Since the proposed approach uses multiple modalities, the combined concepts at each time stamp are combined to form the set of concept primitives. The sequence of such primitives over time period is fed into the system. Along with this sequence, the individual gesture sequences from the BCI and the hand gesture interface are also fed into the system to further improve the disambiguation process. The system has already been described in chapter 2.

The advantage of preserving $m(\phi)$ would appear in the temporal fusion process. At any time stamp, if there is a lot of conflict in the system even after static fusion, the system can output $\{\phi\}$ at that time stamp. The unknown gesture sequence can be disambiguated during the temporal fusion of the gesture sequence. This proposal is also very intuitive, since humans understand natural language by matching a whole sequence of concepts to their knowledge base. Even if we are not able to understand word of a sentence, we are usually able to understand the entire sentence.

Chapter 7

Conclusions and Future Work

This work focussed on developing a semantic multimodal system utilizing hand gestures and brain computing. Such a system is especially useful for people with physical disabilities, allowing them greater levels of interaction with a service robot. It is possible to interact mostly using brain signals, while minimizing the required physical effort to making a few hand gestures.

In this work, a multimodal system for Human Computer Interaction has been proposed. The initial stages of the system, that of the hand recognition sensor and brain computer interface using steady state visually evoked potentials, have been developed and implemented. An elementary fusion system performing disambiguation was discussed, showing how the use of multiple modalities can improve the process of human computer interaction. Also a novel evidential reasoning based fusion algorithm using conceptual graphs was developed. It was shown that such an algorithm has a lot of potential for semantic fusion.

7.1 Contributions to Human Computer Interaction

The contributions of this work to the area of Human Computer Interaction can be summarized as:

3D hand reconstruction

A hand pose estimation algorithm was proposed to identify the 3D model of a hand gesture, given an input two dimensional video of the gestures. The algorithm proposed is computationally less intensive than some of previous algorithms and suitable for real world scenarios.

Brain Computer Interface using SSVEP

A Brain Computer Interface based on Steady State Visually Evoked Potentials was proposed. It was shown how such a system is highly robust and accurate in identifying user selections from a range of flickering options being displayed on a screen. Various operating parameters were experimented with to identify a suitable operating characteristic for the system. The classification measure used was based on the measure proposed in [35], which gives high rates of classification.

Multimodal system involving disambiguation

The BCI was developed further using the same values of steady state frequencies for more than one option. The use of hand gestures to disambiguate the correct option, allowed for high rates of classification while minimizing the number of frequency values needed on the display. This is of significance, since the set of chosen frequencies is reduced and the frequency values can be now be spaced far apart to minimize chances of overlap.

Multimodal Fusion algorithm based on evidential reasoning

A novel approach for semantic fusion based on evidential reasoning was proposed and it was shown how such an algorithm can help improve multimodal fusion, allowing for both reinforcement and clarification. It was shown that the algorithm is successful in all the various operating scenarios for a multimodal system.

7.2 Future Work

The system presented in this work utilizes information from two modalities and performs semantic fusion. The future work in this system would consist of improving the performance and scope of the hand gesture system and the brain computing interface, extending the fusion algorithm to perform temporal as well as static fusion, and integrating the components into a single stand alone system. The major directions of future research are:

- Extend the hand gesture recognition system to include motion tracking, which will enable improved performance and the recognition of a larger set of gestures. Motion tracking using well established techniques like the Kalman or particle filters would allow for identifying the locations of the fingers even in the presence of occlusion. Thus many gestures involving occlusion of one or more fingers could also be identified and recognized.

- Develop the BCI display to include a much larger set of options. This would require an improved and dedicated system for display and data acquisition. One direction of investigation is the use of multiple electrodes on the scalp to acquire data from many locations on the cortex, and also capture the eye blink motions. EEG signals from multiple electrodes can be processed using time-series analysis and independent component analysis to extract better approximation of the stimulus response. Also, research is needed in faster bit-rate for BCI communication and improved measures for even more accurate classification of the extracted signals.

- Develop the multimodal fusion algorithm to incorporate a larger knowledge base. Also a major research direction is the temporal fusion of concepts to form longer concepts, to better understand natural language. The concepts at each time stamp are combined to form the set of concept primitives. The sequence of such primitives over time period is fed into the system. Along with this sequence, the individual gesture sequences from the BCI and the hand gesture interface will also be fed into the system to further improve the disambiguation process.

- Another interesting possibility is the use of 'Fuzzy conceptual graphs' for sensor fusion. Presently gestures are combined only if they are directly connected in the conceptual graph. Also, even if two concepts are connected, we have no information about the strength of the relations. For example, in normal language, a concept like 'Come Here' is a much stronger concept than 'Come and Dance', and has a more probability of occurrence. Thus, this information could possibly be used during the fusion process, as a weighting factor in the computation of the mass values.

Bibliography

[1] S Malik and J Laszlo. Visual touchpad: a two-handed gestural input device. In *Proc. 6th international conference on Multimodal interfaces (ICMI '04)*, pages 289 – 296, State College, PA, USA, October 2004.

[2] Kang Seong-Pal, M. Tordon, and J. Katupitiya. Curvature based hand shape recognition for a virtual wheelchair control interface. In *Proc. IEEE International Conference on Robotics and Automation, 2004 (ICRA'04)*, volume 2, pages 2049 – 2054, April 26 – May 9 2004.

[3] Yang Liu and Yunde Jia. A robust hand tracking for gesture-based interaction of wearable computers. In *Eighth International Symposium on Wearable Computers, 2004. (ISWC 2004)*, volume 1, pages 22 – 29, October 31 – November 3 2004.

[4] E.C.Paraiso and J.-P.A.Barthes. An intelligent speech interface for personal assistants in rd projects. In *Proc. (IEEE) Ninth International Conference on Computer Supported Cooperative Work in Design, 2005*, volume 2, pages 804 – 809, May 24–26 2005.

[5] A. Green and K.S. Eklundh. Designing for learnability in human-robot communication. *IEEE Trans. on Indus. Engg.*, 50(4):644 – 650, August 2003.

[6] S.Q. Xie, C. Gao, Z.L. Yang, and R.Y. Wang. Computer-brain interface. In *Proc. (IEEE) First International Conference on Neural Interface and Control, 2005*, pages 32 – 36, Wuhan, China, May 26–28, 2005.

[7] S.G. Mason and G.E. Birch. A general framework for brain-computer interface design. *IEEE Trans. on Neural Sys. Rehab. Engg.*, 11(1):70 – 85, March 2003.

[8] H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *Proc. (ACM) 6th International Conference on Multimodal Interfaces, 2004*, State College, PA, USA, October 2004.

[9] A. Corradini, M. Mehta, N.O. Bernsen, J.C. Martin, and S.Abrilian. Multimodal input fusion in human-computer interaction on the example of the on-going nice project. In *Proc. NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, August18 - 29 2003.

[10] L. Gupta, B. Chung, M.D. Srinath, D.L. Molfese, and H. Kook. Multichannel fusion models for the parametric classification of differential brain activity. *IEEE Trans. on Bio Medical Engineering*, 52(11):1869 – 1881, November 2005.

[11] Y.Wu, E.Y.Chang, K.C.-C. Chang, and John R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proc. ACM International Conference on Multimedia (MM)*, pages 572 – 579, New York, USA, October 2004.

[12] S. Bangalore and M. Johnston. Integrating multimodal language processing with speech recognition. In *International Conference on Speech and Language Processing*, volume 2, pages 126 – 129, 2000.

[13] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proc. 5th International Conference on Multimodal Interfaces (ICMI 2003)*, pages 12 – 19, Vancouver, Canada, November 2003.

[14] S.M. Chu, V. Libal, E. Marcheret, C. Neti, and G. Potamianos. Multistage information fusion for audio-visual speech recognition. In *Proc. (IEEE) International Conference*

on Multimedia and Expo, 2004, (ICME '04), volume 3, pages 1651 – 1654, New York, USA, June27 - 30 2004.

[15] K. Veeramachaneni, L.A. Osadciw, and P.K. Varshney. An adaptive multimodal biometric management algorithm. *IEEE Trans. on Systems, Man and Cybernetics (C)*, 35(3):344 – 356, August 2005.

[16] D. Lo, R.A. Goubran, and R.M. Dansereau. Robust joint audio-video talker localization in video conferencing using reliability information-ii: Bayesian network fusion. *IEEE Trans. on Instrument Measurement*, 54(4):1541 – 1547, August 2005.

[17] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, New Jersey, U.S.A, 1976.

[18] I. Ruthven and M. Lalmas. Using dempster-shafer's theory of evidence to combine aspects of information use. *Journal of Intelligent Information Systems, 2002*, 19(3):267 – 302, 2002.

[19] O. Basir and X.Yuan. Engine fault diagnosis based on multi-sensor information fusion using dempster-shafer evidence theory. *Information Fusion*, 2005.

[20] C. Yang and B. Blyth. An evidential reasoning approach to composite combat identification (ccid). In *Proc. (IEEE) Aerospace Conference, 2004*, volume 3, March 6–13, 2004.

[21] Lotfi Zadeh. A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination. *AI Magazine*, 7(2):85 – 90, 1986.

[22] Phillipe Smets. The combination of evidence in transferable belief model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12:447 – 458, 1990.

[23] Phillipe Smets. *Belief Functions and the Transferable Belief Model.* http://ippserv.rug.ac.be/documentation/belief/belief.pdf, Aug 2000.

[24] B. Miners, O.A. Basir, and M. Kamel. Knowledge-based disambiguation of hand gestures. In *Proc. (IEEE) International Conference on Systems, Man and Cybernetics, 2002*, volume 5, UMIST, Manchester, UK, October 6–9, 2002.

[25] M.Chein and M. Mugnier. *Conceptual Graphs are also graphs, Tech. Rep. 95003.* LIRMM,, Universit'e Montpellier II, 1995.

[26] J.F.Sowa. *Knowledge Representation Logical, Philosophical, and Computational Foundations.* Brooks Cole Publishing Co., Pacific Grove CA, 2000.

[27] W.B. Miners, O.A. Basir, and M.S. Kamel. Understanding hand gestures using approximate graph matching. *IEEE Trans. on Systems, Man, and Cybernetics - Part A*, 35(2):239 – 248, March 2005.

[28] R.Herpers, C.Pantofaru, L.Wood, K.Derpanis, D.Topalovic, and J.Tsotsos. Fast hand gesture recognition for real-time teleconferencing application. In *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proc. IEEE ICCV Workshop on*, pages 133 – 140, July13, 2001.

[29] B.Stenger, P.R.S.Mendonça, and R.Cipolla. Model-based hand tracking using an unscented kalman filter. In *Proc. British Machine Vision Conference*, volume 1, pages 63 – 72, September 2001.

[30] Jintae Lee and T.L.Kunii. Model-based analysis of hand posture. *IEEE Trans. on Computer Graphics and Applications*, 15(5):77 – 86, September 1995.

[31] C.S.Chua, H.Guan, and Y.K.Ho. Model based 3-d posture estimation from single 2-d image. *Image and Vision Computing*, 20(3):191 – 202, March 2002.

[32] R.Herpers, K.Derpanis, W.J.MacLean, G.Verghese, M.Jenkin, E.Milios, A.Jepson, and J.K.Tsotsos. Savi: An actively controlled teleconferencing system. *Image and Vision Computing*, pages 793 – 804, 2001.

[33] T.Morris and O.S.Elsehery. Hand segmentation from live video. In *The 2002 Intl. Conference on Imaging Science, Systems, and Technology*, pages 206 – 211, UMIST, Manchester, UK, 2002.

[34] J. R.Wolpaw, N. Birbaumer, D. J.McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):769 – 791, 2005.

[35] S.P.Kelly, E.C.Lalor, C.Finucane, G.McDarby, and R.B.Reilly. Visual spatial attention control in an independent brain-computer interface. *IEEE Trans. on BioMedical Engineering*, 52(9):1588 – 1596, September 2005.

[36] G.Schalk, D.J.McFarland, T.Hinterberger, N.Birbaumer, and J.R.Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Trans. on BioMedical Engineering*, 51(6):1034 – 1043, June 2004.

[37] P.L.Lee, C.H.Wu, J.C.Hsieh, and Y.T.Wu. Visual evoked potential actuated brain computer interface: a brain-actuated cursor system. *Electronics Letters 21st, IEE*, 41(15), July 2005.

[38] Ming Cheng, Xiaorong Gao, Shangkai Gao, and Dingfeng Xu. Design and implementation of a brain-computer interface with high transfer rates. *IEEE Trans. on BioMedical Engineering*, 49(10):1181 – 1186, October 2002.

[39] Yijun Wang, Ruiping Wang, Xiaorong Gao, Bo Hong, and Shangkai Gao. A practical vep-based brain-computer interface. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 14(2):234 – 240, June 2006.

[40] J.Kremlacek, M.Kuba, and J.Holeiki. Model of visually evoked cortical potentials. *Physiological Research*, 51:65 – 71, 2002.

[41] V.Jaganathan, T.M.S.Mukesh, and M.R.Reddy. Design and implementation of high performance visual stimulator for brain computer interfaces. pages 5381 – 5383, September01 - 04 2005.

[42] S.Kelly, D.Burke, P. de Chazal, and R. Reilly. Parametric models and spectral analysis for classification in brain-computer interfaces. In *Proc. 14th International Conference on Digital Signal Processing*, Greece, June 2002.

[43] A.Greco, N.Mammone, F.C.Morabito, and M.Versaci. Kurtosis, renyis entropy and independent component scalp maps for the automatic artifact rejection from eeg data. *International Journal of BioMedical Sciences*, 1(1), 2005.

[44] A.V.Kramarenko and U.Tan. Validity of spectral analysis of evoked potentials in brain research. *Intl. Journal of NeuroScience*, 112:489–499, 2002.

[45] Amer. Electroencephalogr. Soc. Guidelines for standard electrode position nomenclature. *Clin. Neurophysiol.*, 8(2):200 202, 1991.

[46] J.Shao. Linear model selection by cross-validation. *Jrnl. of the American Statistical Association*, 88:486 – 494, 1993.

[47] R. Hanson, J. Stutz, and P. Cheeseman. Bayesian classification with correlation and inheritance. Sydney, Australia, 1991.

[48] D. Harmenac. *Measures of Uncertainty and Information.* http://ensmain.rug.ac.be/ ipp, 1999.

[49] N.R.Pal, J.C.Bezdek, and R.Hemasinha. Uncertainy measures for evidential reasoning ii: A new measure of total uncertainty. *International Journal of Approximate Reasoning*, 8(1):1 – 16, 1993.

[50] J.Braun. Demspter-shafer theory and bayesian reasoning in multisensor data fusion. In *Proc. (SPIE) Sensor Fusion: Architectures, Algorithms and Applications IV*, pages 255 – 266, 2000.