Analyzing Visual Semantic Processing and Recognizing the Shape Changes in Video

Abhinav Gupta[1], Gary Garcia[2], and Touradj Ebrahimi[2]

[1] Department of Computer Science and Engineering,
Indian Institute of Technology, Kanpur
{abhigupt}@iitk.ac.in
[2] Signal processing Laboratory,
Swiss Federal Institute, Lausanne
{gary.garcia,touradj.ebrahimi}@epfl.ch

Abstract. How does the brain work? How do we recognize things? These are one of the few mystery of the world which has bothered the cognitive researchers for years and continue to haunt them. In this project, we try to analyze the visual semantic processing of the objects of different shapes in the brain. We show the people videos with changing shapes and analyze their brain signals to see the effects of these changes in the brain signals. Then we train computer by machine learning approach to identify the change in videos just by classifying and recognizing these signals. We also try to see the evoked potentials in an event related potential when a person sees a normal video and try to analyze these evoked potentials to find the area of semantic processing.

1 Introduction

Visual semantics refers to that part of semantic memory that deals with knowledge about visual aspects of elements in the world around us. Martin et al. [9] suggested that semantic knowledge about objects is stored close to areas that mediate the perceptual processing of those objects. Furthermore, they proposed that the organization of visual semantic knowledge parallels the organization of perceptual function which is known to be separated into several functional areas that each code for a specific visual property (shape, color, motion, etc) (e.g. [3]).

1.1 Electroencephalogram and Semantics

EEG (electroencephalogram) measures the electrical activity of the brain via electrodes that are placed on the scalp. EEG is able to provide a high resolution temporal image of neuronal processing. A popular way to analyze EEG is to calculate the brain's electrical signal in response to a specific event (e.g. in response to the presentation of a word or a picture, or a manual key-press by a subject). By averaging together the EEG recorded to a large number of events (typically about 100) the EEG which is unrelated to the event is averaged out, leaving the brain's electrical response that is causally related to the stimulus or the event.

This so-called event-related potential or ERP has a long scientific tradition and different ERP effects and components have been mapped for a diverse range of stimuli and conditions. Psycholinguistic research involved with semantics has centered on the N400 component[7] which consists of a large negative wave that reaches its maximum amplitude approximately 400 ms after stimulus onset. The N400 is sensitive to semantic relations between words, but is also found with non-linguistic stimuli such as line-drawings[6] and faces[1], suggesting that the N400 reflects the processing of meaningful stimuli in general. Also the N400 has been used to investigate whether there is a common semantic systems for all meaningful stimuli, or whether separate semantic systems exist (e.g. separate systems for pictorial and verbal information)[4]. However, still little is known about the neural generators of the N400 effect, and more importantly, on how semantic information is distributed and processed in different parts of the brain. Although ERPs hold the potential of providing a high resolution temporal image of neural activity in response to semantic processing, this potential has been largely disregarded by focusing too much on the N400 and neglecting other ERP components. For example, research directed at investigating the visual aspects of semantics should not just center on the N400 but should additionally note the involvement of ERP components in the visual domain. Appropriate in this respect would be to use the extensive knowledge on ERP effects of visual attention, working memory, and mental imagery in relation to semantic processes. Apart from ERP's, direct EEG trials called single trial analysis is also done. The features of these trials are extracted and then these features are used for classifying mental activities [2].

In this project, we try to analyze the visual semantic processing of the objects of different shapes in the brain. We show the people videos with changing shapes and analyze their brain signals to see the effects of these changes in the brain signals. We use standard machine learning approaches to form classes of different shapes and try to classify the signals(single trial EEG's) recorded later into one of the class.

We also try to see the evoked potentials in an event related potential when a person sees a normal video and try to analyze these evoked potentials to find the area of semantic processing.

We discuss in section 2 the protocol of the trial and the kind of movies shown in the trials. Section 3 has been dedicated to all the approaches used in classification of single trial EEG's. We finally conclude in section 5 suggesting some future works.

2 A Trial Session Protocol

In this work, we try to observe the effect of change of shape of visual objects on the signals from the brain. Our each session consists of 16 trails (See Fig.1). In each trial three movies are shown. There is an indefinite gap between two successive trials.

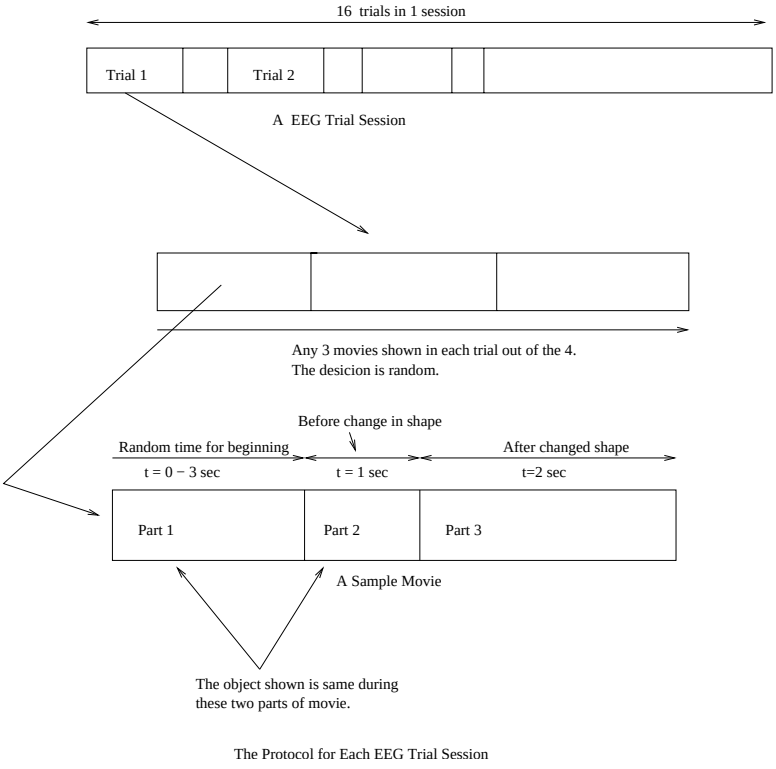
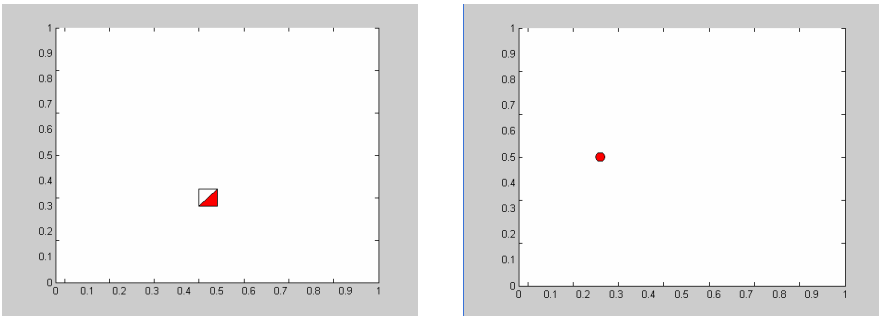


Fig. 1. The Protocol for a Session.

Each movie consists of three parts.In part one frames is shown with either no object or a circle in a sine wave or a square in a sine wave. The duration of part one is random and varies from zero seconds to three seconds. In part two same object remains in the frames and perform same actions but the length is fixed to be one second. In part three, the object changes completely and a object of new shape executes the sine curve motion. The content of the four movies made has been described in Table.1.

Table 1. The Contents of Movies Shown

Movie	Part1	Part2	Part3
1	Blank	Blank	Circle
2	Blank	Blank	Square
3	Circle	Circle	Square
4	Square	Square	Circle



(a) (b)

Fig. 2. Sample frames of movie.

3 Classification of Single trial EEG's

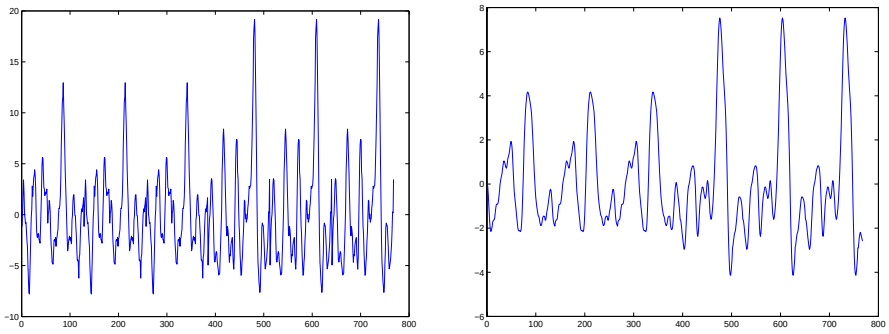
Before using any EEG signal for training or simulation certain preprocessing steps have to be done. We first remove artifacts using frequency domain and k-means algorithm. We also make the spatial mean to be zero.

3.1 Evoked potentials

Most of the research in the past regarding semantics of audio signals and words has been done in relation to evoked potentials. The N400 discovered by [7] has

been used primarily for the research in field of semantics. Since we try to classify semantics of videos based on shapes and we have four classes of videos we extract some features based on evoked potentials, their peak values and the peak latencies.

The first step in this field is extraction of averaged smooth signals. The following figure (Figure.3(a), Fig.3(b) and Fig.4) shows the effect of averaging on EEG's.



(a) (b)

Fig. 3. (a)A sample original signal. (b)The signal after applying a smoothing filter.

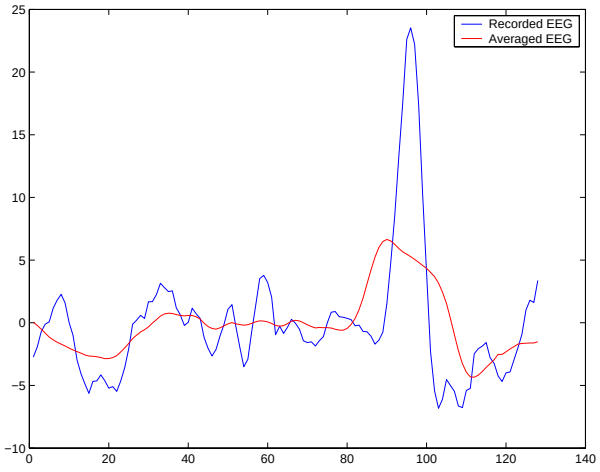


Fig. 4. Comparison between original and averaged signal.

Only the two electrodes O1 and O2 are used for extraction of features as most of the visual information processing has been reported in inferotemporal cortex which lies in the occipital part of the brain[10].
The extracted features are trained on multi-layer perceptron neural network. The numbers of layers are three in the neural network. The input layer has 4 neurons and the output layer has also four neurons. The number of neurons in the hidden layer is computed by the validation process. The graph plotted below (Fig.5) shows the number of neurons in the hidden layer on x-axis and percentage error (false negatives) on the y-axis.

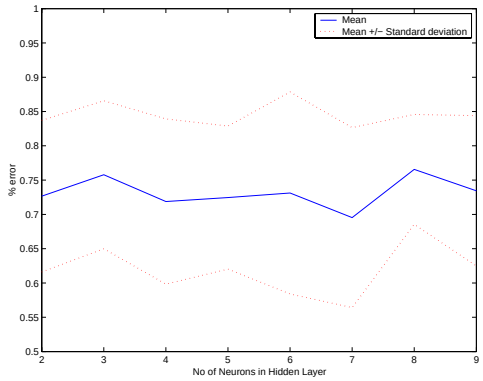


Fig. 5. Validation curve for the Neural Network trained on evoked potential features.

A neural network was finally built having four neurons in input layer, seven neurons in hidden layer and four neurons in output layer (The output being [1 0 0 0] for Movie 1, [0 1 0 0] for Movie 2 etc.). The weights of the neural network has been represented in the Hinton Diagrams[5] shown in Fig.6 and Fig.7.

The simulation was done on a set of forty eight trials and the final results obtained have been shown in the bar-graph below (Fig.8). As the results suggest that the classification is not so good in case of evoked potential features. The results were also compiled on basis of classification in two classes one with movies having final shape as square and a movie with final shape as a circle. The results of this simulation has been plotted below (Fig.9).

3.2 MicroStates Based Approach[8]

A brain microstate is defined as a functional/physiological state of brain during which certain neural computations are performed. It is characterized by spatial distribution of active neuronal generators with time varying intensity. Thus a brain activity or a event related activity can be modeled as time sequence of non overlapping microstates with some duration.
A geometrical interpretation(See Fig.10) of the microstate can be defined. Consider a 2-D plane defined by potential of two electrodes which are used to define a

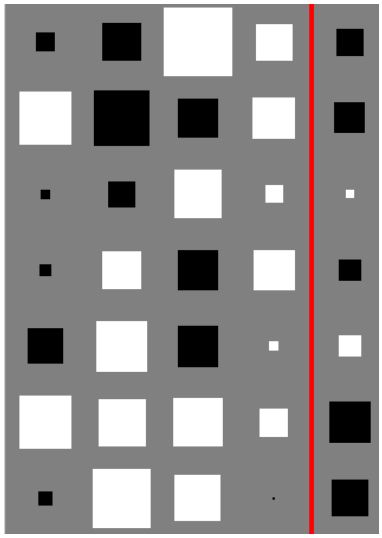


Fig. 6. The hinton diagram for weights from input layer to hidden layer.

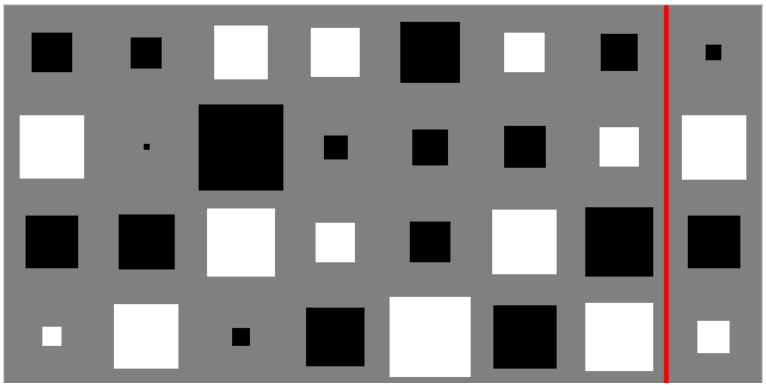


Fig. 7. The hinton diagram for weights from hidden layer to output layer.

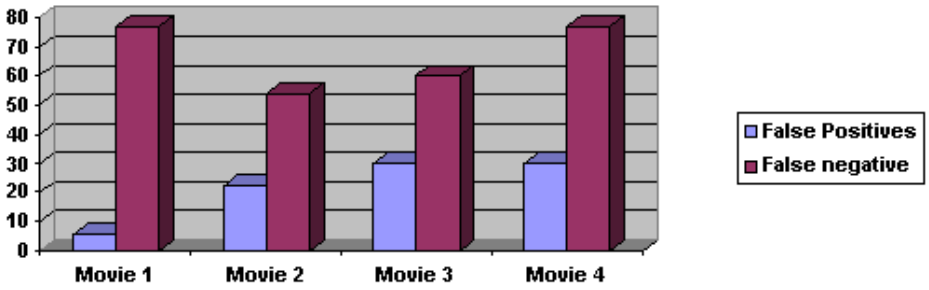


Fig. 8. The performance of neural network for features computed from evoked potentials. Classification was done in four classes.

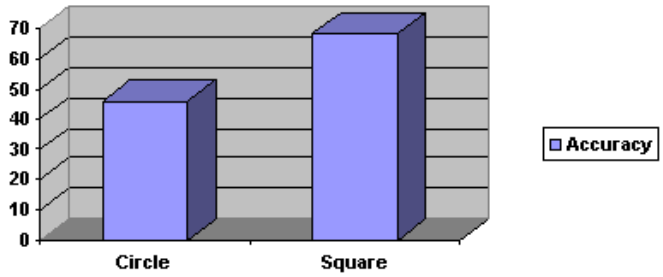


Fig. 9. The performance of neural network for features computed from evoked potentials

microstates. A point in this plane correspond to an instantaneous measurement. A brain microstate is characterized on such a plane as a vector lying at a unit distance from the origin. All the points lying on line from origin to microstate belong to that microstate. The distance from the origin to a point on this line is directly propotional to the strength of neuronal generators corresponding to the microstate.

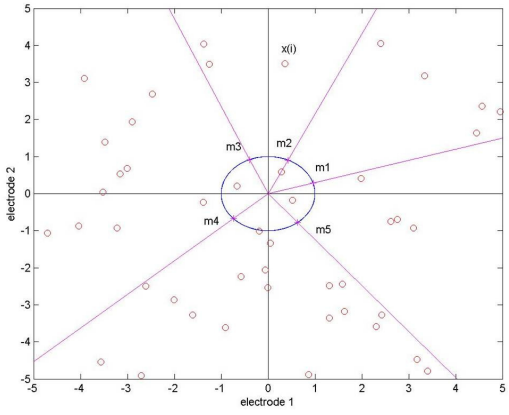


Fig. 10. A Microstate Model

A microstate model for average data can be stated as

$$V_t = \sum_{k=1}^{N_\mu} a_{kt} \Gamma_k \tag{1}$$

where N_μ is the number of microstates, V_t is $N_s x 1$ vector representing a potential measurements at an instant, Γ_k is normalized vector representing $k-th$ microstate and a_{kt} is the $k-th$ microstate intensity at time t. Since microstates should be non overlapping following conditions must be satisfied:

$$a_{k_1 t}.a_{k_2 t} = 0, \forall k_1 \neq k_2, \forall t \tag{2}$$

$$\sum_{k=1}^{N_\mu} a_{kt}^2 \geq 0 \forall t \tag{3}$$

Estimation For given N_μ, the model parameters can be estimates by minimizing

$$\eta = \frac{1}{N_T(N_s-1)} \sum_{t=1}^{N_T} (\|V_t - \sum_{k=1}^{N_\mu} (a_{kt}\Gamma_k)\|)^2 \tag{4}$$

with the restrictions specified in 2 and 3. This is modified version of k-means method.

In first step all microstates are given. The orthogonal distance between measurement vector and microstate is computed by

$$d_{kt}^2 = V_t'.V_t - (V_t'.\Gamma_k)^2 \tag{5}$$

After this each measurement is linked to the microstate with which it has minimum distance. The estimator for a_{kt} is then given by

$$a_{kt} = V_t'.\Gamma_k \tag{6}$$

In the next step labels are given and the function4 is minimized by finding normalized eigen vector

$$S_k = \sum_{t/L_t=k} V_t.V_t' \tag{7}$$

and then calculating new microstate as

$$\Gamma_k = argmax X' S_k X \tag{8}$$

The last two steps are performed alternatively to estimate the microstates.

Validation The error is found by changing N_μ and a validation function is introduced as

$$\epsilon = \frac{1}{N_T(N_s-1)} \sum_{t=1}^{N_T} (\|V_t - \sum_{k=1}^{N_\mu} (a_{kt}\Gamma_k)\|)^2 + N_n \tag{9}$$

where N_n is the number of unused microstates. The N corresponding to the minimum error is the optimal number of microstates. A sample validation curve is shown in Fig.11.

Results The microstates were computed for the four time periods i.e 500-1000 msec, 1000-1500 msec, 1500-2000 msec, 2000-2500 msec. The classification was done using all the four separately. The classification was done on basis of minimum distance with the four classes. However since each class has different number of microstates some are better represented than others. To nullify that effect we divide each distance by penalization constant whose value is determined by number of microstates for the class. The results(true positives) for all the four time periods have been shown below in Fig.12.

As the results show the accuracy in the 1500-2000 millisecond period is much higher and better than others. The false positives and negatives error for this period has been shown in Fig.13.

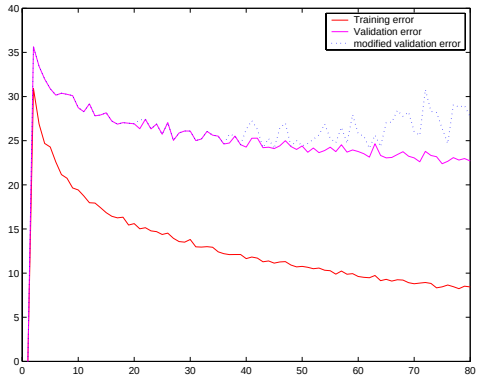


Fig. 11. A sample validation curve.

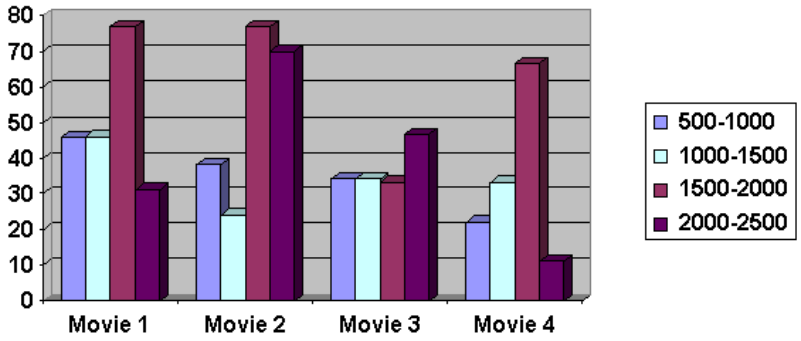


Fig. 12. The Accuracies using the microstates computed in the four periods.

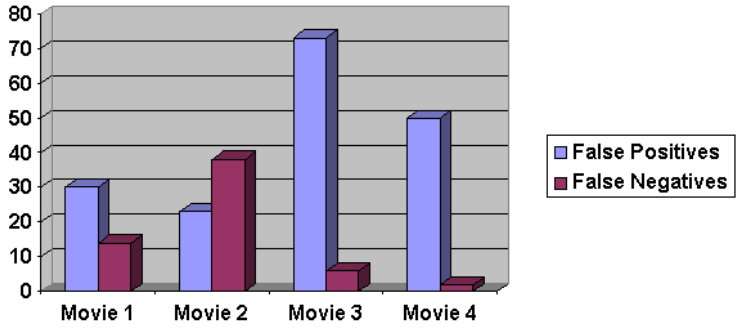


Fig. 13. The Classification errors using microstate model for 1500-2000 milliseconds

However when the results were compiled 14 with respect to the final shape of the object the accuracy improved suggesting a relation between the final shape and the potentials measured. It can be also stated that the microstates computed were dependent on final shapes of object.

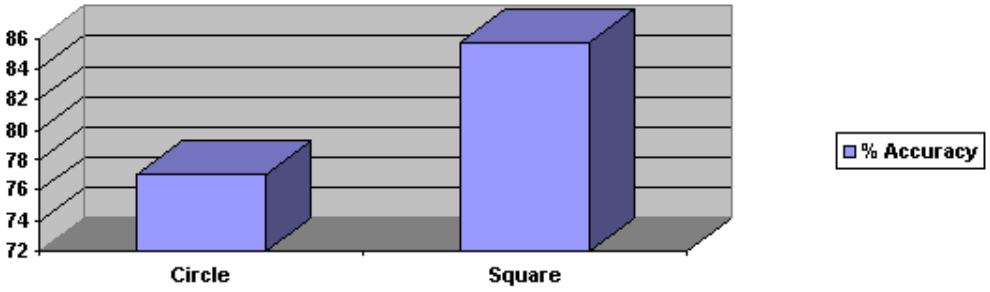


Fig. 14. The classification results for classification based on final shapes using microstate model for 1500-2000 milliseconds

The microstates computed for all the four classes of movies have also been shown in the figures(15, 16, 17, 18) below.

3.3 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It finds two bases, one for each variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. In other words, it finds the two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. The dimensionality of these new bases is equal to or less than the smallest dimensionality of the two variables.
Mathematically, Canonical correlation analysis can be defined as the problem of finding two sets of basis vectors, one for X and the other for Y, such that the correlations between the projections of the variables onto these basis vectors are mutually maximized.
Consider the linear combinations $X = X^T \widehat{w_x}$ and $Y = Y^T \widehat{w_y}$ of the two variables respectively. This means that the function to be maximized is

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}} \tag{10}$$

The maximum of ρ with respect to w_x and w_y is the maximum canonical correlation.

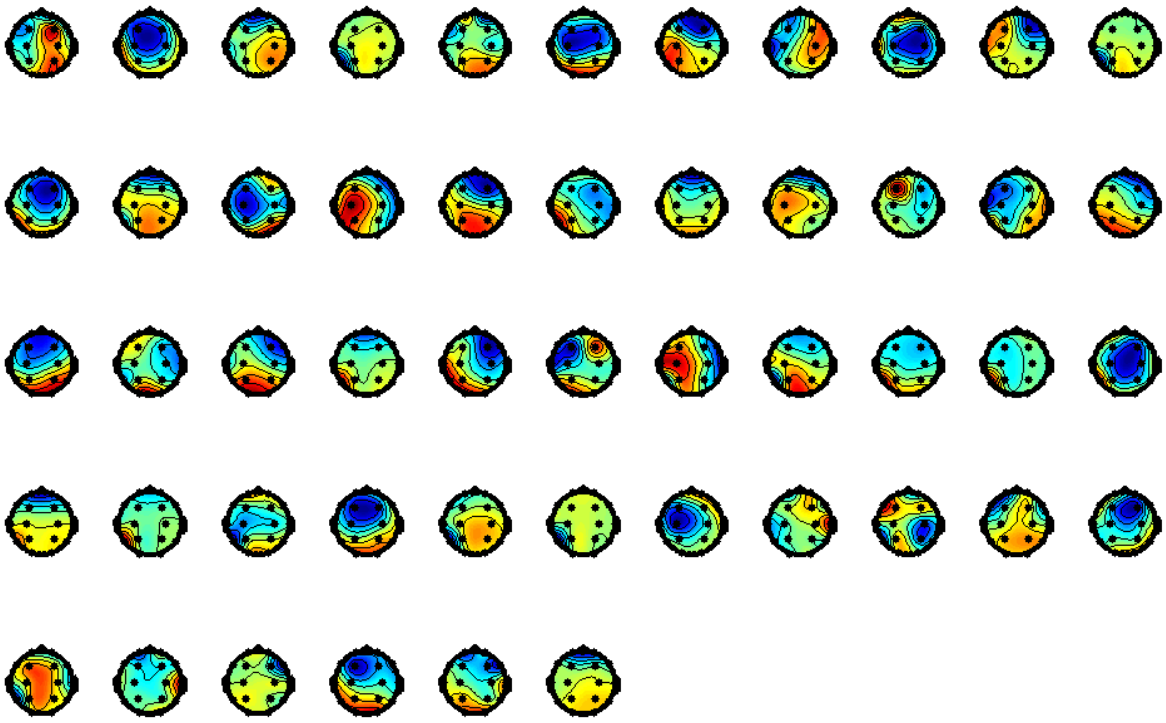


Fig. 15. Microstates for Movie1

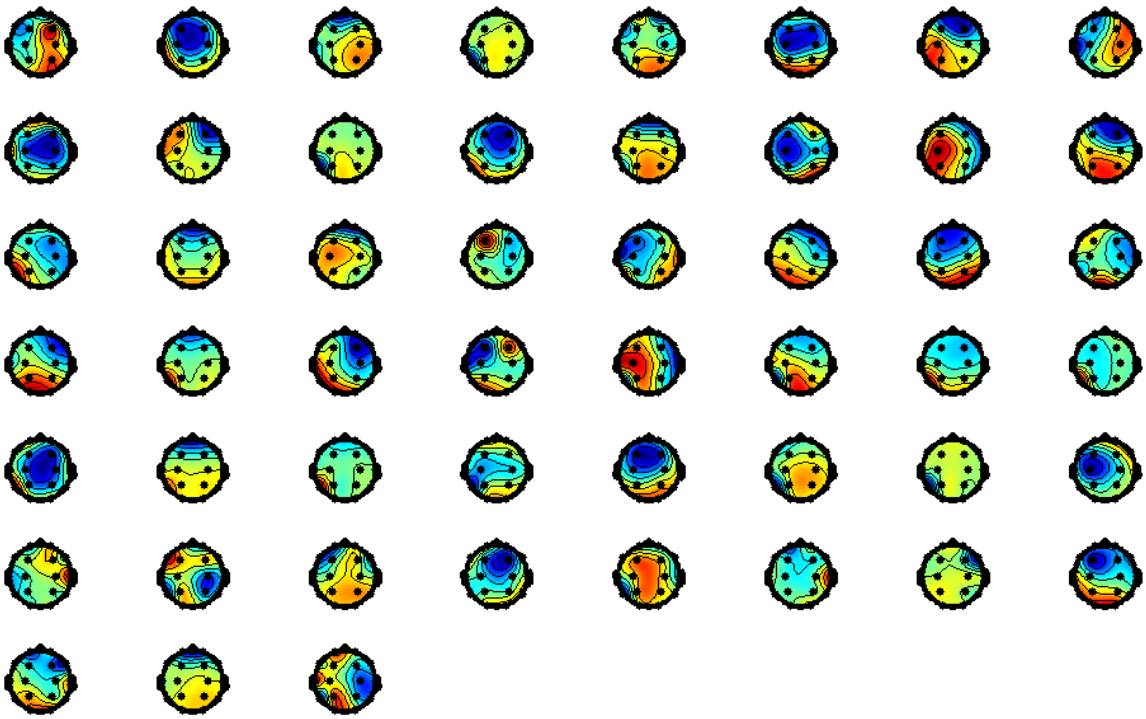


Fig. 16. Microstates for Movie2

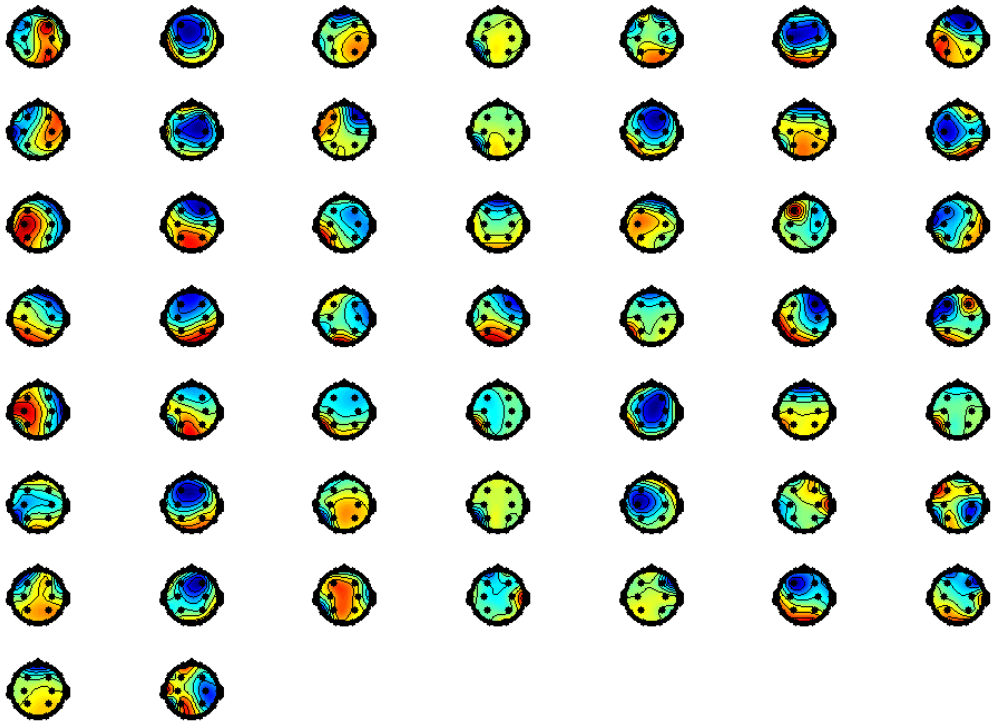


Fig. 17. Microstates for Movie3

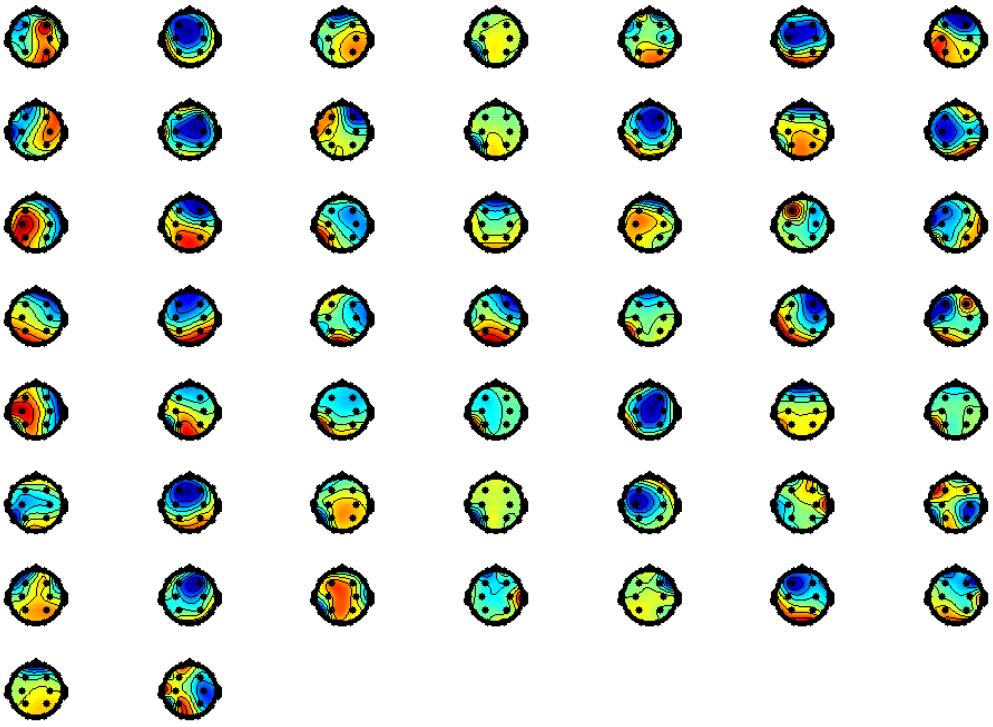


Fig. 18. Microstates for Movie4

Calculation of Canonical Correlation Consider two random variables x and y with zero mean. The covariance matrix is given by

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} \tag{11}$$

The canonical correlations between x and y can be found by solving the eigenvalue equations equations.

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} = \rho^2 \widehat{w_x} \tag{12}$$

$$C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} = \rho^2 \widehat{w_y} \tag{13}$$

where the eigenvalues ρ^2 are the squared canonical correlations and the eigenvectors $\widehat{w_x}$ and $\widehat{w_y}$ are the normalized canonical correlation basis vectors.

Calculation of feature vectors and Results We used Canonical correlation analysis for extracting the features out of EEG signals. We noted that the signal change due to the sudden change in shape of the object. Thus we thought of extracting a feature based on some relation between the signals before the change and after the change. Hence we find the relationship between the signals before and after the change in shapes of object.

A neural network was trained with 16 neurons in the input layer, 4 neurons in output layer and the 34 neurons in the hidden layer. The optimal number of neurons in the hidden layer was obtained by changing the number of neurons in the hidden layer and obtaining the validation curve for the neural network. The validation curve for classification in four classes and classification in two classes(Square and Circle) are shown in Fig.19 and Fig.20.

The network for four class problem has been represented by hinton diagrams shown in Fig.21 and Fig.22. The results of the final simulation are also shown in the bar graphs below in Fig.23 and Fig.24.Note in the 1st graph errors has been plotted while in 2nd graph accuracy has been plotted.

3.4 Spatial and Frequency Filters

Let X_1 be the covariance of a trial in the target and X_2 be the covariance of a non-target trial. We diagonalize the matrix $X_1 + X_2$ as

$$X_1 + X_2 = U^T D_c U \tag{14}$$

Let us consider another matrix

$$Y_1 = D^{\frac{1}{2}} U X_1 U^T D^{\frac{-1}{2}} \tag{15}$$

Now it can be proved that $Y_1 + Y_2 = I$. When we diagonalize the matrix

$$Y_1 = V^T D_V V \tag{16}$$

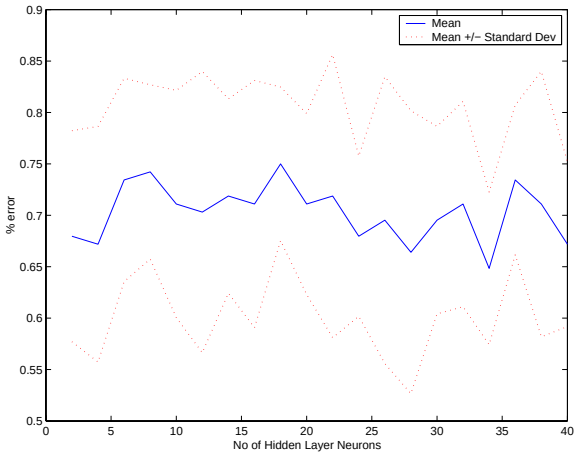


Fig. 19. Validation curve for four class problem.

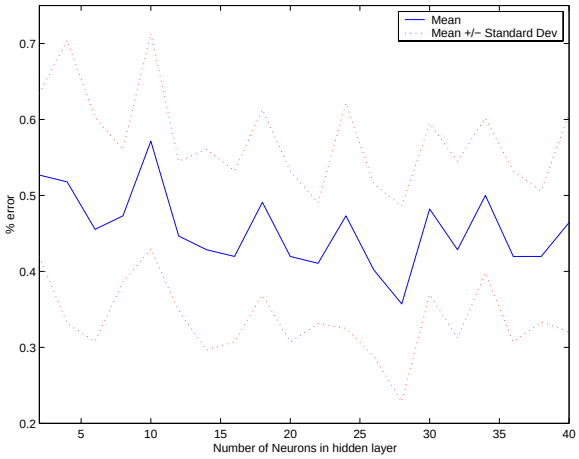


Fig. 20. Validation curve for two class problem.

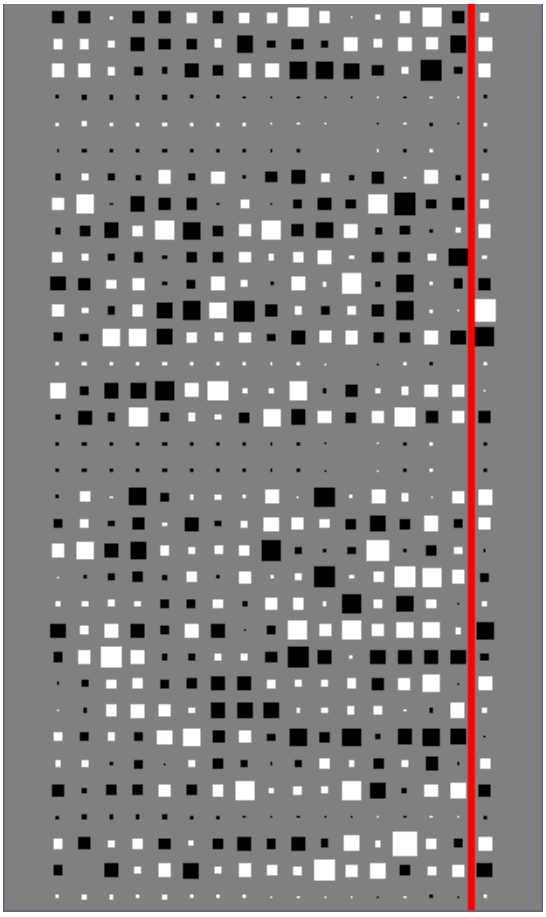


Fig. 21. The hinton diagram for weights from input layer to hidden layer.

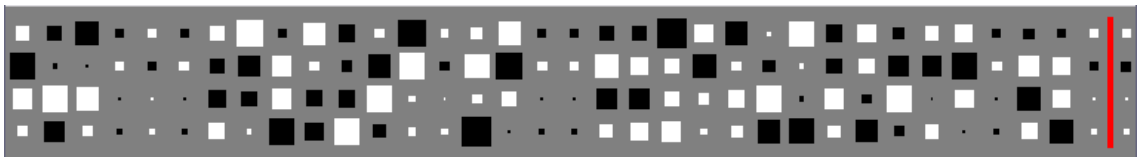


Fig. 22. The hinton diagram for weights from hidden layer to output layer.

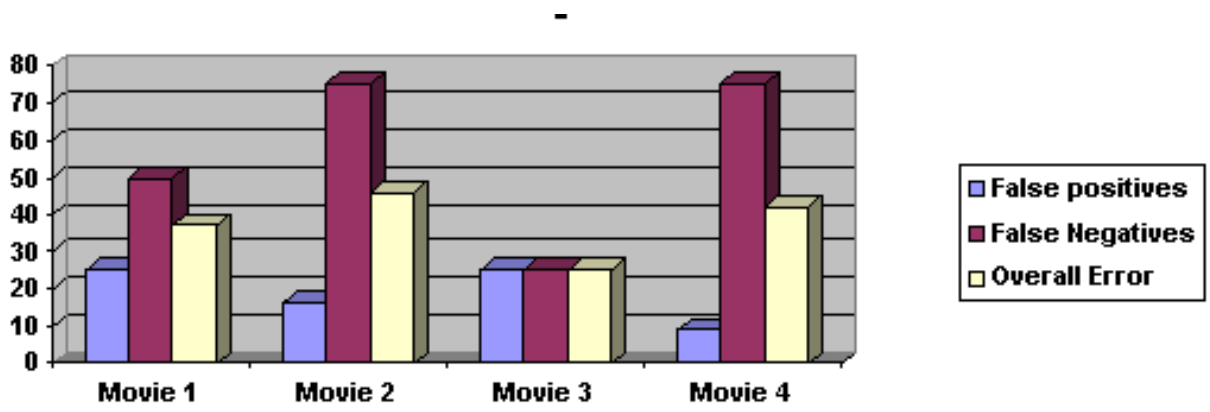


Fig. 23. The results of simulation of four class problem.

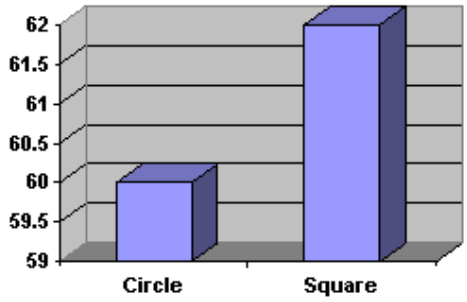


Fig. 24. The accuracy result of simulation of two class problem.

Computing D_v (whose variance is the feature in our case) we get

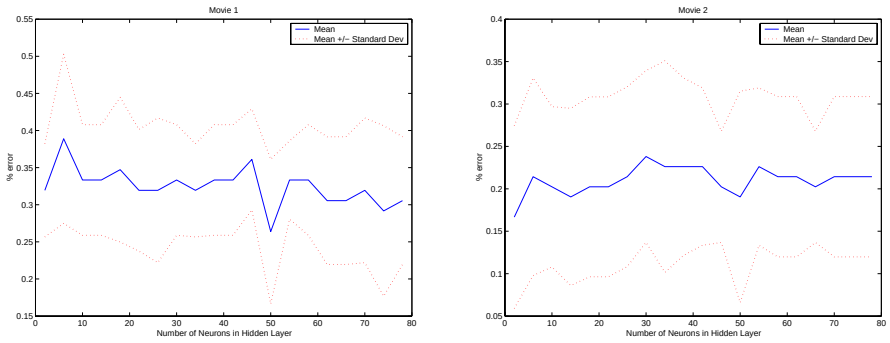
$$D_v = VY_1V^T = VD^{\frac{-1}{2}}UX_1U^TD^{\frac{-1}{2}}V^T \tag{17}$$

This $VD^{\frac{-1}{2}}U$ becomes the projection vector which can be used to project EEG trails. However all the components are not necessary and hence to reduce dimensionality we can use the 4 components (The first two discriminating the target and last two discriminating the non-target).

However this filter does not look at the frequency. So instead of applying the filter on the original signal we first divide the signal into 10 frequency bands and then apply this feature separately on each band to get 4X10 = 40 features.

We applied this approach to extract features because with this approach we can reduce the coarse of dimensionality. The frequency bands introduces also introduced frequency based features.

For our problem we had four classes so four projection vectors. We created four neural networks for separating each class from the rest of mental activities. Each neural network had 40 neurons in input layer and 1 neurons in output layer. The number of neurons in hidden layer was obtained by validation curves shown in Figures25(a), 25(b), 26(a), 26(b).



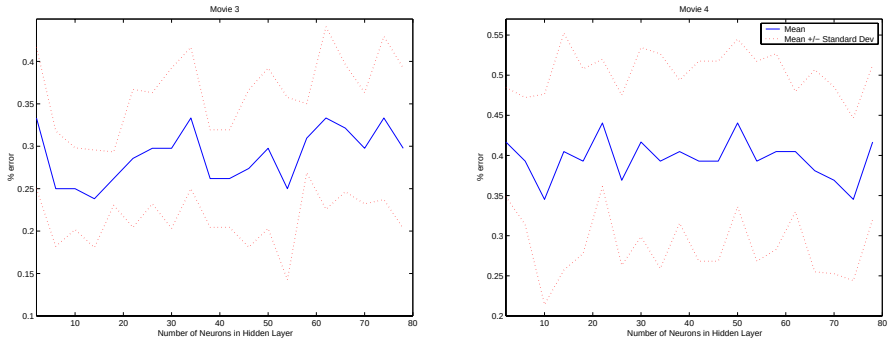
(a) (b)

Fig. 25. (a)The validation curve for Movie 1. (b)The validation curve for Movie 2.

However to obtain labels for each trial we compare the results of the neural network and the one with the highest value is considered to be the label. The results of separating from rest of the mental activities and classification is shown in Figures 27 and 28 respectively.

3.5 Event Related Potentials

We also looked at the ERP's as discussed in section 1. We showed the people normal videos and ask them to understand semantics of the video. The potentials



(a) (b)

Fig. 26. (a)The validation curve for Movie 3. (b)The validation curve for Movie4.

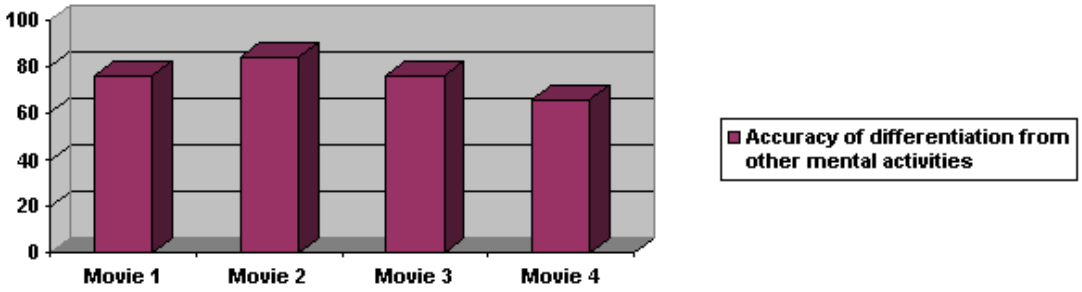


Fig. 27. The results of differentiation from rest of the activities.

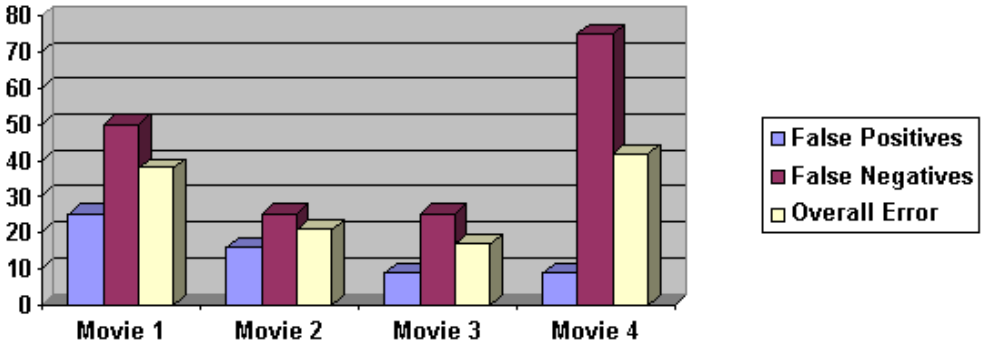


Fig. 28. The results of classification in one of four classes.

in different part of the scalps have been plotted in figures(29,30, 31, 32) below. On analyzing them we found that there is always a cycle which continues after first 500 milliseconds. But they subside and most of the mental activity stops around 4500 milliseconds which seem to suggest that the person has understood video.

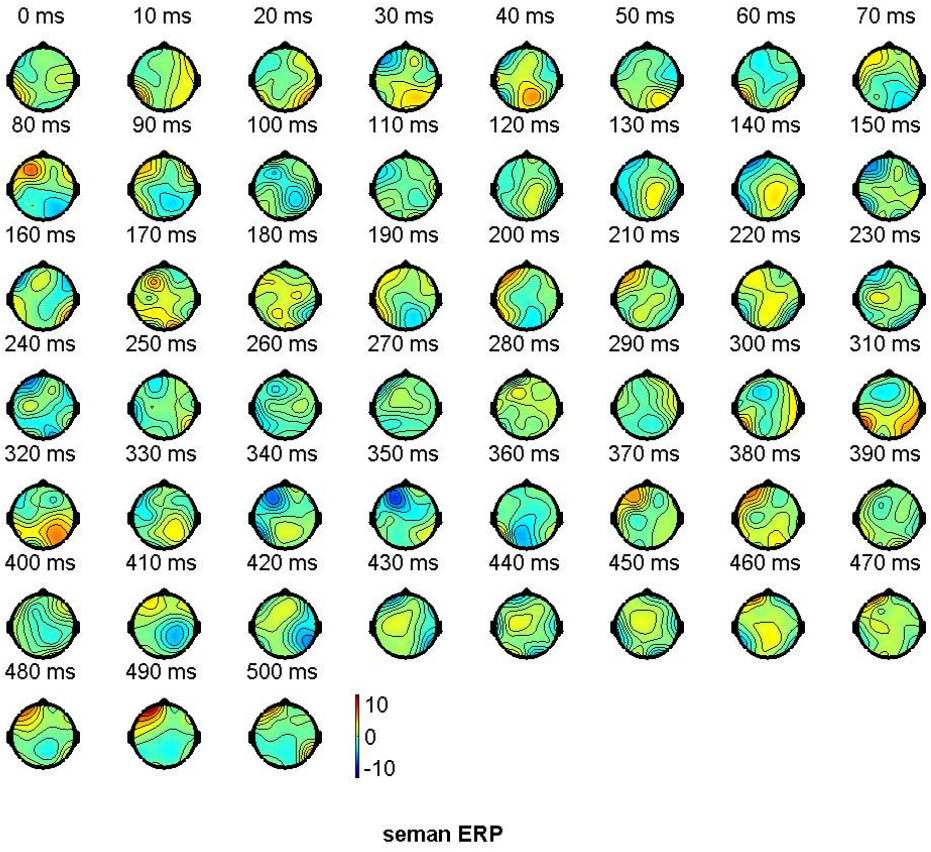


Fig. 29. The potential on scalp from 0-500 ms.

The other notable features are the P300 and N400 at 320 and 420 milliseconds respectively. Most of the activity has been found in occipital region of the brain after we remove the artifact component by Independent Component Analysis.

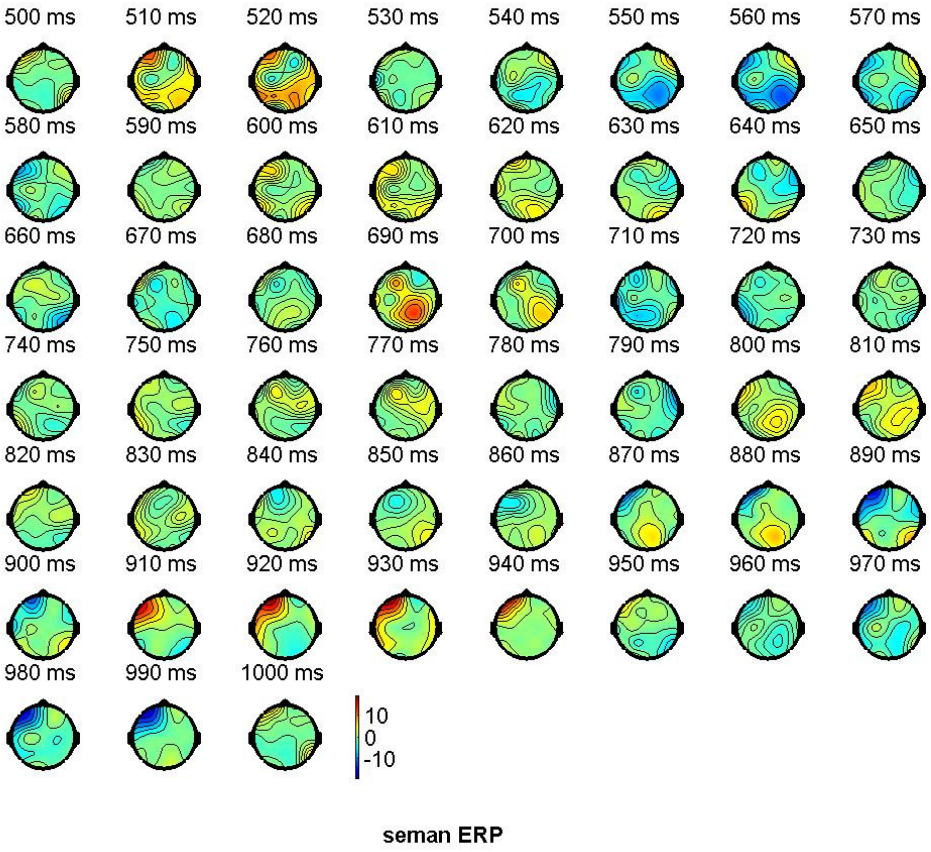


Fig. 30. The potential on scalp from 500-1000 ms.

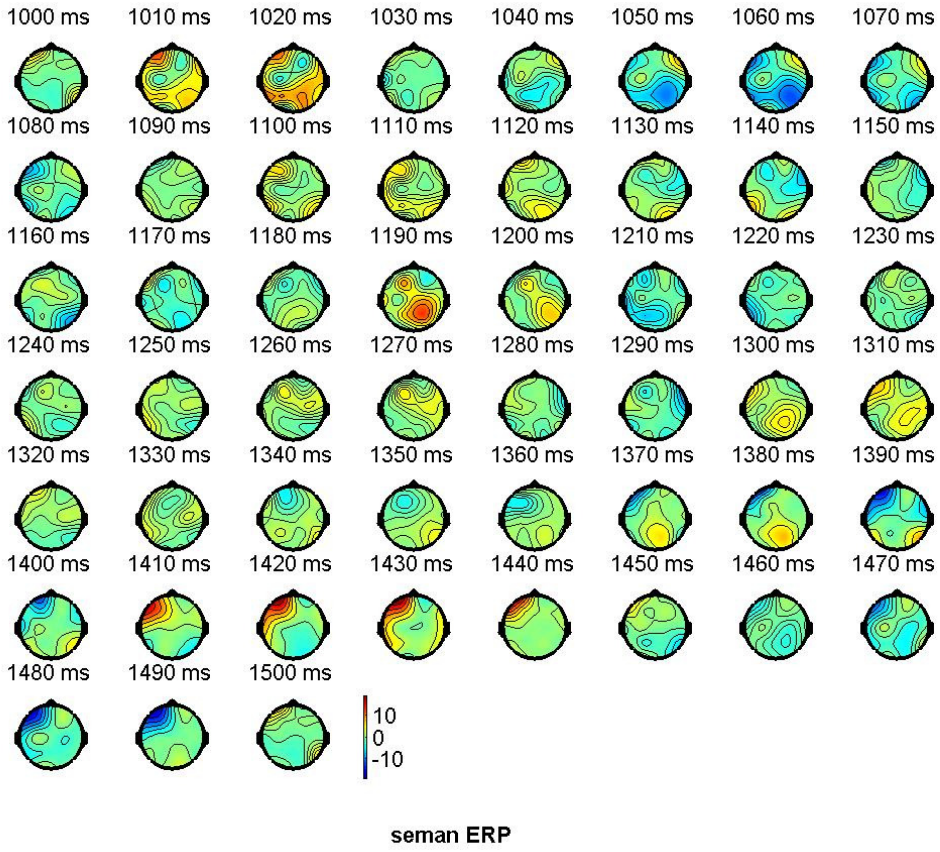


Fig. 31. The potential on scalp from 500-1000 ms.

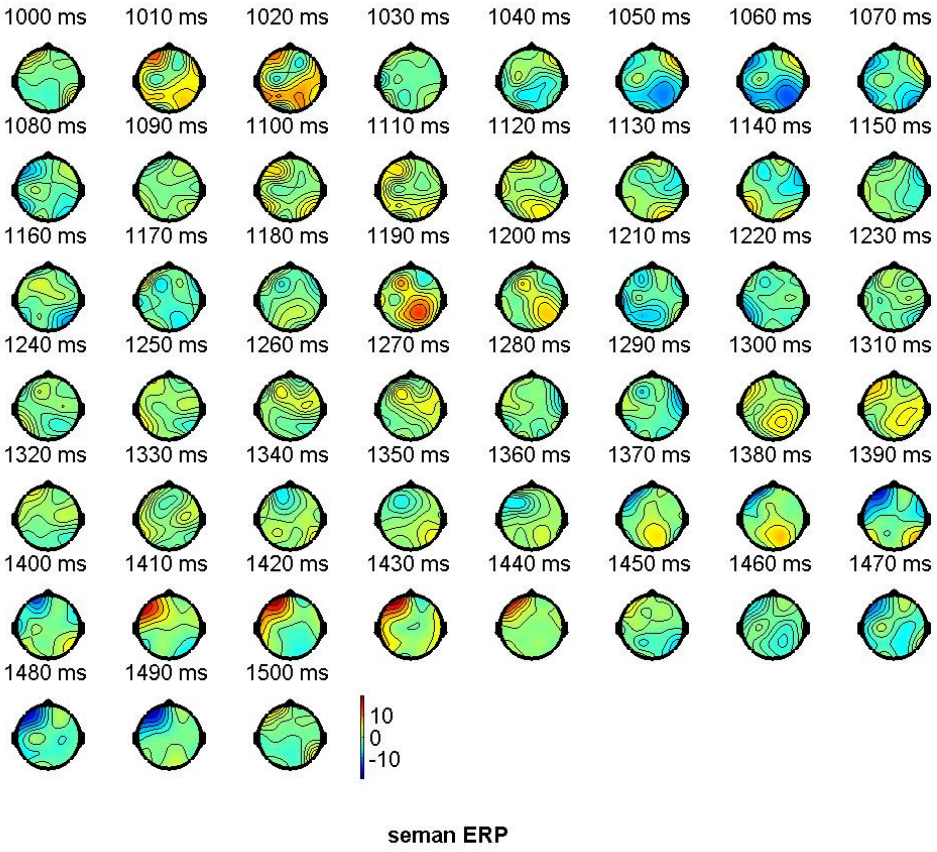


Fig. 32. The potential on scalp from 500-1000 ms.

4 Conclusion and Future works

The comparison of classification schemes is shown in the Fig33. As the figure suggests the microstates approach has best result of all the approaches. However there is a scope of improvement in the microstate approach. Since it is a time sequence of microstates, Hidden Markov Models can be used to increase the accuracy. In spatial filters approach the results are encouraging if we use to

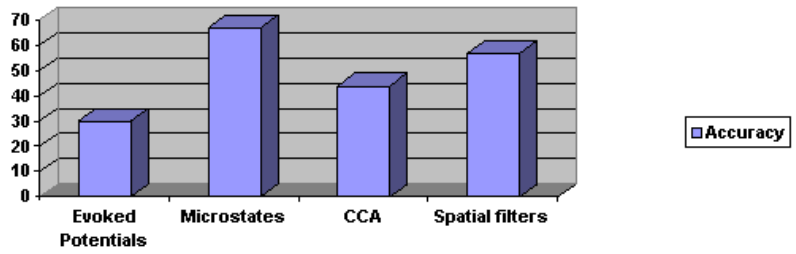


Fig. 33. The results of classification in four classes using all the techniques.

separate a class from others but when we compare the results with each other and classify by giving labels the results are not good. A better approach to compare the simulated results of the neural networks should be worked out.Another approach can be to train neural nets on 40x4 = 160 features, but it requires more number of trials.
CCA and Evoked potentials do not seem to be good features but they can be combined with other approaches to improve the results.

References

1. S.E Barrett and M.D. Rugg: *Event-related potentials and the semantic matching of pictures.* Brain and Cognition, 1990, Vol.14, 201212.
2. Touradj Ebrahimin, Jean-Marc Vesin and Gary Garcia:*Brain-computer Interface in Multimedia Communication* IEEE Signal Processing Magzine, 2003, 14-19.
3. D.J Felleman and D.C Van Essen: *Distributed hierarchical processing in the primate cerebral cortex.* Cerebral Cortex, 1991, Vol 1, 147.
4. K.D. Federmeier and M. Kutas: *Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 2001, Vol.27(1), 202224.
5. Netlab: *Netlab neural network software.* www.ncrg.aston.ac.uk/netlab/
6. P.J. Holcomb and W.B. McPherson: *Event related potentials reflect semantic priming in an object decision task.* Brain and Cognition, 1994, Vol.24, 259 276.
7. M. Kutas and S.A Hillyard: *Reading senseless sentences: Brain potentials reflect semantic incongruity.* Science, 1980, Vol.207, 203205. 16211625.

8. RD. Pascual-Marqui, CM Michel and D. Lehmann: *Segmentation of Brain Electrical Activity into Microstates: Model Estimation and Validation*, IEEE Transactions on Biomedical signal processing, 1995, Vol.42(7).
9. A. Martin, C.L. Wiggs, L.G Ungerleider and J.V. Haxby:*Discrete cortical regions associated with knowledge of color and knowledge of action.* Science, 1995, vol. 270, 102105.
10. K. Tanaka: *Inferotemporal cortex and object vision* Annu. Rev. Neurosci,1996, 19, 109-139.