Project Report

Grammatical facial expression recognition using customized deep neural network architecture

Prepared by:

Mohammed Basil	B150451CS
Nekkanti Lakshmi Sai Pavan	B150810CS
Padavala Bhavani Venkata Krishna	B150775CS
Raavi Prathap Reddy	B150681CS
Vishnumolakala Rohith	B150541CS

Under the guidance of:

Mr. Vinith R

National Institute of Technology Calicut

Data Description:

The automated analysis of facial expressions has been widely used in different research areas, such as biometrics or emotional analysis. Special importance is attached to facial expressions in the area of sign language, since they help to form the grammatical structure of the language and allow for the creation of language disambiguation, and thus are called Grammatical Facial Expressions.

The dataset is composed by eighteen videos recorded using Microsoft Kinect sensor. In each video, a user performs (five times), in front of the sensor, five sentences in Libras (Brazilian Sign Language) that require the use of a grammatical facial expression. By using Microsoft Kinect, we have obtained: (a) an image of each frame, identified by a timestamp; (b) a text file containing one hundred coordinates (x, y, z) of points from eyes, nose, eyebrows, face contour and iris; each line in the file corresponds to points extracted from one frame. The images enabled a manual labelling of each file by a specialist, providing a ground truth for classification.

The dataset is organized in 36 files: 18 data point files and 18 target files, one pair for each video which compose the dataset. The name of the file refers to each video: the letter corresponding to the user (A and B), name of grammatical facial expression and a specification (target or data points).

Attribute Information:

Data points files:

Coordinates x and y are given in pixels.
Coordinates z are given in millimetres.

Label of frame
0 - 7 (x, y, z) - left eye
8 - 15 (x, y, z) - right eye
16 - 25 (x, y, z) - left eyebrow
26 - 35 (x, y, z) - right eyebrow
36 - 47 (x, y, z) - nose
48 - 67 (x, y, z) - mouth
68 - 86 (x, y, z) - face contour
87 (x, y, z) - left iris
88 (x, y, z) - right iris
89 (x, y, z) - nose tip
90 - 94 (x, y, z) - line above left eyebrow
95 - 99 (x, y, z) - line above right eyebrow

The attribute point locations on user face can be found in figure 1.

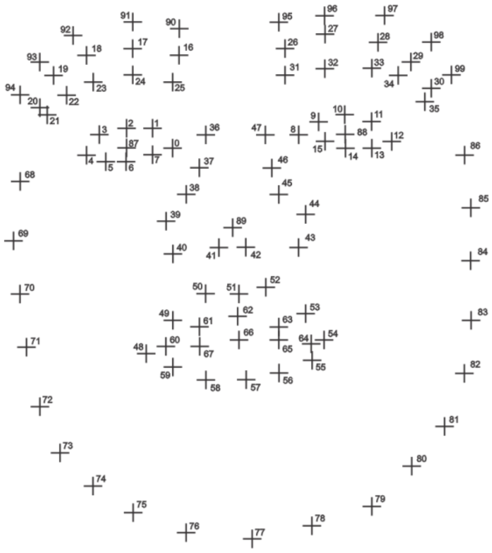

Fig 1.Attribute point locations on user face

Data Pre-processing:

The data points are normalised using the Z-score standardisation. This also makes the learnt model invariant to the location of face in captured frame (i.e. having a different set of attribute numerical values). This does lead to an appreciable increase in model performance. The data points are split into train data and test data in the ratio of 80:20.The class labels y are converted into hot vectors.

Customized (Sparse) Deep Neural Network Architecture:

For this model, customized feed-forward deep network architecture was implemented. It consists of two hidden layers along with the standard input and output layers. The entire customized architecture can be referred to in Figure 2. Here, for each sample (frame) the attribute points standardized X, Y, Z coordinates are fed to a single neuron in the first hidden layer. Thus, 100 neurons present in first hidden layer are tuned to find learning pattern in each of its respective attribute point's coordinates. The space represented by first layer can be expressed as,

H1 ϵ {V0, V1, V2…….Vn ….V99} where Vn ϵ {Xn, Yn, Zn}

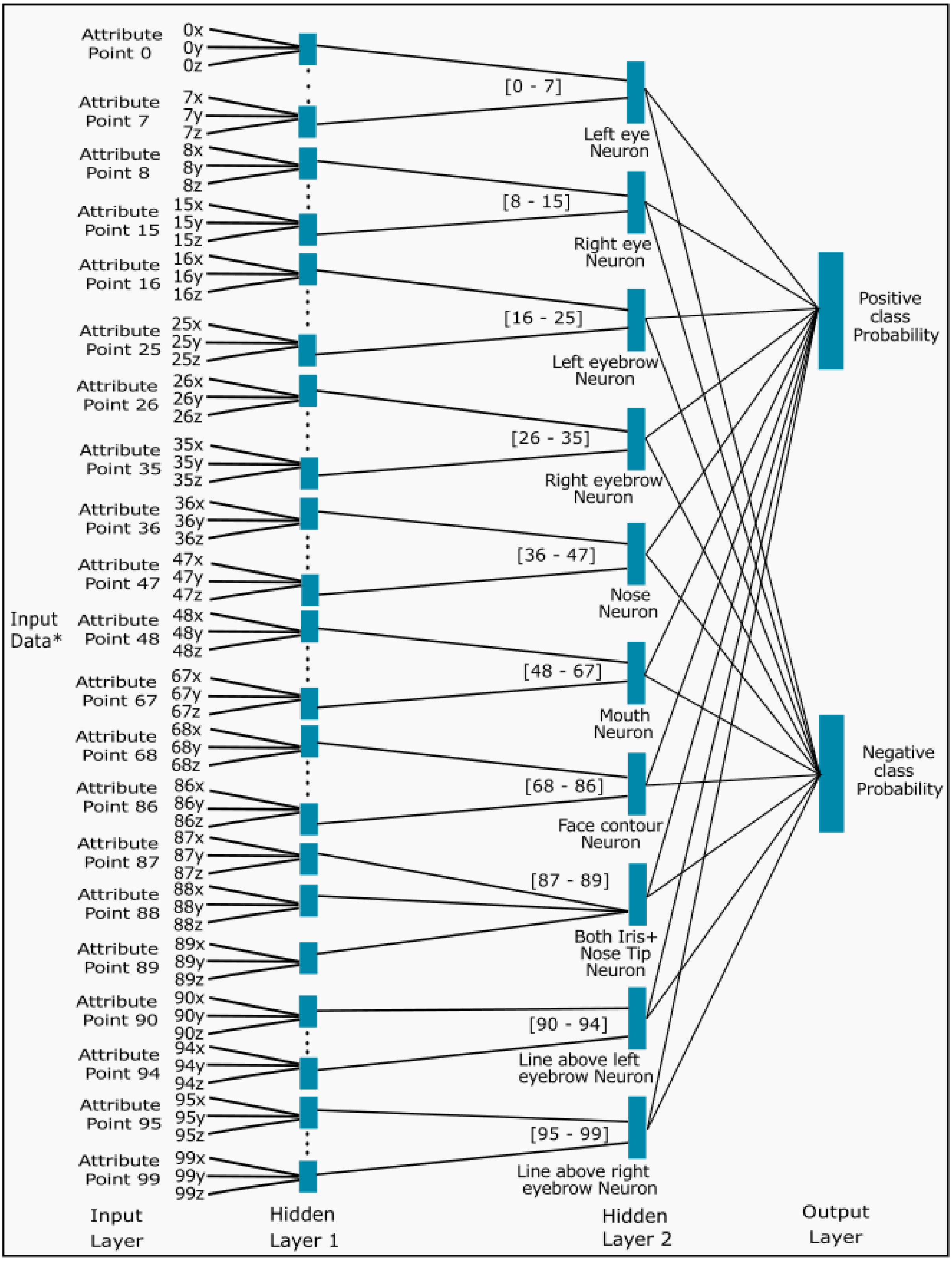

Figure 2

*Normalised Input data points

Subsequently, varied clusters of these neurons are fed to specific neurons in the second hidden layer. As seen from figure 2, certain clusters of attribute points (i.e. first layer neurons) represent specific parts of human face. These respective clusters can be referenced from Table 1

User Face Region	Attribute Point Range
Left Eye	0-7
Right Eye	8-15
Left eyebrow	16-25
Right eyebrow	26-35
Nose	36-47
Mouth	48-67
Face contour	68-86
Left & Right iris +nose trip	87-89
Line above Left eyebrow	90-94
Line above Right eyebrow	95-99

Table 1.Attribute Point groups for different user face regions

Each of the second layer neurons are thus tuned to learn individual patterns in specific face regions, such as left/right eye, nose, mouth etc. respectively. This hidden layer space can be represented as,

H2 ϵ {H10, H11, H12 …H1n…H19}

Where H1n ϵ {V0-V7, V8-V15, V16-25, V26-V35, V36-V47, V48-V67, V87-V89, V90-V94, V95-V99}

Output layer consists of two neurons. The second hidden layer is fully connected to each output neuron. This enables output layer neurons to fully learn patterns from each of the face regions present in H2 space. Each neuron in the architecture has an individual bias weight attached it.

Network Training:

The 'Cross Entropy' (i.e. ∑ y' log(y) where y' is true class and y is predicted class) is used as the cost function, expressed as difference between model predictions and its true output values. For training purpose, Gradient descent optimization algorithm was implemented owing to its faster convergence rate, being computationally efficient and been used to find the weights and biases for each neuron (parameters).

The model learning rate is 0.001. This implementation helps the cost function to reach its minimum value. The train data is used to train and build the model.

Testing:

The model is tested against the test data of data points and accuracy of the test data is noted.

Results:

The accuracy of proposed model on all the markers individually as a binary classification task is shown in Table 2.It demonstrates the accuracies of train data and test data corresponding to user A ,user B and combination of both user A and user B.

Conclusions:

The overall accuracy of proposed method is excellent, such that it can reliably be used for classifying GFEs captured in form of video frames.

S. No	Name	Train Accuracy			Test Accuracy		
		Person A	Person B	Both A & B	Person A	Person B	Both A & B
0	Affirmative	87.986	76.95	55.83	83.568	71.628	55.84
1	Conditional	93.443	81.069	82.2	91.361	80.589	84.1
2	Doubt_question	87.035	83.709	83.0	85.551	86	84.8
3	Emphasis	87.166	82.79	69.68	83.63	79.182	65.63
4	Negative	81.869	82.925	65.76	81.777	80.757	63.09
5	Relative	93.186	88.049	71.5	94.635	88.976	72.7
6	Topics	81.917	85.068	76.6	83.333	87.397	78.7
7	Wh_question	86.868	87.382	59.3	89.535	88.346	60.6
8	Yn_question	88.129	83.525	61.6	85.971	78.736	60.5

Table 2