Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 58 (2015) 210 – 216

Procedia Computer Science

Sign Language Recognition using Facial Expression

Siddhartha Pratim Das[a], Anjan Kumar Talukdar[b], Kandarpa Kumar Sarma[c]

[ab]*Dept. of Electronics and communication Engineering,Gauhati University, Assam-781014,India*
[c]*Dept. of Electronics and communication Technology,Gauhati University, Assam-781014,India*

Abstract

Vision- based approaches of recognition of sign languages have made spectacular advances in the last few years. These also include many works in the area of speech processing to convert speech to text. A vision-based approach to classify facial gestures (lip movement, eye brow pattern etc.) for communication designed especially for the differently abled persons is a less explored area. In our work , we explore certain approaches to classify facial gestures to enhance its effectiveness and incorporate it to any sign language or vision-based gesture recognition movements for precise decision making. In our work, we have designed a real time system to detect alphabets by recognizing the lip pattern based on texture and shape. The system takes live video input and processes it in real time. Object detector of computer vision toolbox is used to classify the lips from extracted frames of video input. Five consecutive frames are extracted so as to trace the movements caused while speaking a particular syllable. Histogram of oriented gradients (HOG) of extracted lip image is used as features for recognition. The recognizer is designed using Artificial Neural Network (ANN) to recognize four classes viz. the lips movements formed for the four alphabets 'A','B','C','D'. The entire system is modelled and tested for real time performance with a video of 10 frames per second. Experimental results show that the system provides satisfactory performance with recognition rate as high as 90.67%.

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)

Keywords: Feature extraction; Histogram of Oriented Gradients; Viola Jones Algorithm; Artificial Neural Network

1. Introduction

Vision based approaches has always been forming a basis to be used as primary communication method that allows impaired hearing people to communicate with others in their daily life. It is the fundamental communication bridge among the hearing deficient person. Any vision based system or technique involving sign language with hand gestures alone does not yield effective results and hence use of facial gestures have received wide spread acceptance. Among the various types of facial gestures, eye-brow and lip motion are primary ingredients in such a system as these parts generally undergo various stresses and expansion during any expression and hence are the most

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)
doi:10.1016/j.procs.2015.08.056

commonly used facial features. The upper and bottom part of the face conveys messages and descriptors in the language. Grammatical importance is always sought after in these facial movements. The field of vision based system is very vast and the problems faced by recognition system are immense. Further, the system being related to human computer interaction (HCI), it possesses great importance in making such communication real time and effective. Our system mainly focuses on certain designs which are intended to remove some of the deficiencies and limitation observed in any such recognition based system. It focuses on formulation of certain approaches and engaging some of the parameters that are quite extensively used in such recognition systems. Facial expressions are virtually ignored because of their various complexities, interpretation of the character and unevenness of understanding. To reduce the common shortcomings of such a system, we have devised certain methods to overcome this and make the system acclimatize to various conditions. The system that we incorporate also takes into account the substantial and ineffective incorporation of all its features in real time systems thereby making it undergo certain changes and come out with effective output results. Any facial based recognition system substantially incorporated with a vision based hand recognition system could prove to be beneficial for deaf and dumb people. The efficiency of such systems shall be measured in terms of befitting and precise results provided to the user. We generally tend to incorporate the following three main basic steps in such a system which includes acquisition, detection and substantial pattern recognition. For the purpose, we use Viola Jones algorithm followed by Adaboost techniques for better thresholding. Finally to get the desired results, we generate the histograms of all these oriented graphs so as to obtain appropriate states for getting better results during training and tracking. The resembling features or extracted components are then nearly matched and are used for spotting errors or actual rightly detected parts so as to bring out precise and accurate results.

Thus we propose such a machine vision based system that takes video feeds from camera. For our work, we are using a USB Webcam linked into Matlab via the webcam support package to access live video. Video input is taken into the designed system at a speed of 10 frames per second (FPS) and five consecutive frames are considered as belonging to one syllable. Thus the system functions on basis of the assumption that a person speaking will speak two syllables per second. Frames from that video are extracted for the purpose of preprocessing to make it suitable for the classification and recognition process. The first step is to detect the face region for the purpose of extracting desired portion of the image. This is achieved by using object detector of computer vision toolbox in Matlab. Here we have worked on recognizing syllables taking into account only the lip movement. Five consecutive frames are considered to build the features for one particular syllable. The bounding box of the lip is determined to crop the region from the original image for recognition.

Fig. 1 & Fig. 2 shows the different lip patterns formed for two different English alphabets ‘A’, ‘B’. Fig. 1(a), 1(b), 1(c), 1(d) & 1(e) shows five consecutive frames for alphabet ‘A’ and fig. 2(a), 2(b), 2(c), 2(d) & 2(e) shows five consecutive frames for alphabet ‘B’.

Features of extracted frames are extracted using Histogram of Oriented Gradients (HOG) because of its simple implementation and low computational time, since real time application is our main criteria. The recognition process is based on Artificial Neural Networks (ANN). The cropped image of the lips is recognized using a trained ANN. The ANN is trained with cropped images of lips from a video feed. The images were cropped manually for the purpose of feature extraction and training.

Fig. 1. Lip pattern formed for alphabet ‘A’

Fig. 2. Lip pattern formed for alphabet ‘B’

This work primarily concerns itself on recognizing and classifying four alphabets. To make the system more efficient for hearing deficient people to communicate others in real life, several other features needs to be incorporated for efficient functioning.

2. Theoretical Background

2.1. HOG

HOG are explained as descriptors used in vision based technologies which are used to describe the intensity gradients or directions of edges based within a stipulated shape for object detection.[1] The technique follows the repetition of orientation of gradient operators in localized regions and subjective portions of any image. HOG and its associate principle works on developing grids based on density into small square sized cells which are equally spaced and computes out individually the histogram of gradient features or edge orientations for each of its graded pixels and the summation of all these features sums up the descriptor. However in HOG transformation better normalized contrasting features using intensity measurements are used for precise results and accurate measurements.[2, 3]

2.2. ANN

ANN is an interconnected group of statistical algorithms combined together in nodal systems having multiple layers mainly an unknown input layer, a hidden layers and an output layer interconnected by modified weights. They are mainly used in pattern recognition as they are very much adaptive in nature. It consist of sets having parameters which have weights adaptive and are implemented by a learning algorithm and are also used in nonlinear function measurement of respective inputs. ANN performs better in static conditions as compared to dynamic surroundings as they change over sequentially and more rigorously and single ANN will not perform training with great accuracy.[4, 5]

2.3. Viola-Jones Algorithm

It's a framework of object detection to give a real time output. This process involves detecting of any facial features by running a sub window across. This algorithm is highly capable of functioning in a robust environment which means that it should detect all visible faces in any conceivable image. It involves rescaling the detector using a pre-determined scale invariant detector which transforms each pixel in an image as the sum of its pixels to the top and to its left. The next step that follows is computation of sum within the specified rectangle as proposed by the scale invariant detector followed by boosting techniques. Under normal circumstances for a 24x24 detection region, the number of possible rectangle features is 180,000.[6]

3. System Model & Methodology

The proposed system is shown in Fig 3. The following section discusses the methodology associated with each block.

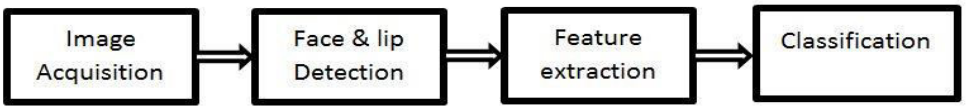

Fig. 3. Block diagram of the proposed system

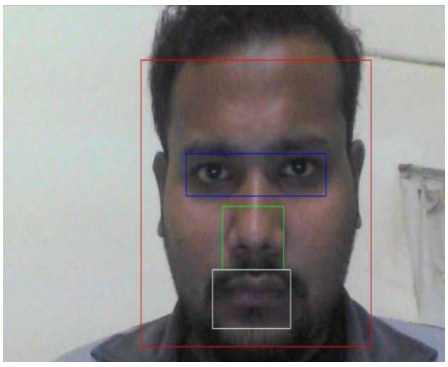

Fig. 4. Object detector output for face

Fig. 5. Object detector output for lips

3.1. Image Acquisition

Input images are directly loaded into system using the image acquisition toolbox. Video stream is taken input at 10 FPS. The system is so designed so as to process and recognize 5 frames as one syllable in 0.5 secs. Frames from the input video stream are extracted and processed for object recognition and feature extraction.

3.2. Face and lips detection

As mentioned in section 1 object detector based on viola jones algorithm of the computer vision toolbox of Matlab is used for detecting the face and in particular the lip portion of the input video stream. The face detector is first run to detect the region of interest (ROI) i.e. the face and then the algorithm is run again to detect the lips portion. Once the lips portion is detected the bounding box is cropped and the features are extracted for the purpose of classification.

3.3. Feature Extraction

As mentioned in section 1, HOG is used for extracting features for training the ANN. Table 1 gives the parameters considered for extracting features. Fig. 4. and Fig. 5. show the orientation bins obtained for the two alphabets ‘A’ and ‘B’. The features obtained from the five images corresponding to one syllable is collectively used as the features for that particular syllable and used for the training purpose.

Several research works exist that applies HOG for the purpose of facial feature detection. The proposed system uses five consecutive frames at a spacing of 0.1 second and feature extracted from each frame are combined to represent the alphabet which increases recognition efficiency

Parameter	Value
Resizing of frame	50X80
Cell Size	[8X8]
Number of cell in block	4
Number of overlapping cells in adjacent block	1
Number of orientation histogram bins	3

Table 1. Parameters considered for HOG

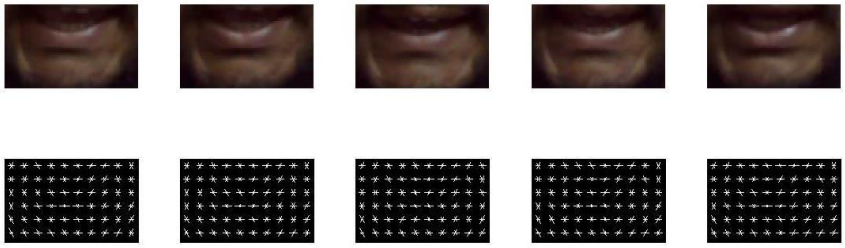

Fig. 6. Features for ‘A’

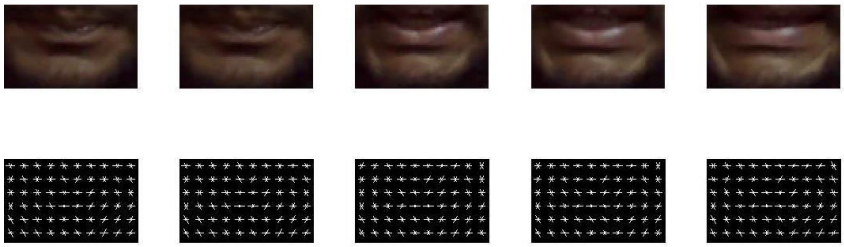

Fig. 7. Features for ‘B’

Alphabet	No of images used for training
A	73
B	65
C	88
D	70

Table 2. No of images used for training ANN

3.4. Classification

The classifier is based on ANN. The ANN is designed using the lip images as explained in section1. Table 2 gives the number of training images used per class for training the ANN. Here the number of images indicates the number of consecutive images considered. As for example the syllable ‘A’ was trained using 365 images which mean 73 samples each constituting 5 consecutive frames. Thus the table indicates the number of samples used for each syllable.

Training higher number of variables presents a difficulty as lip patterns for several pairs of alphabets are almost same and hence requires another approach of cumulating other gestures like eye brows, eyes etc.

4. Experimental Results

The ANN was trained as explained in section 3.4. For testing the network initially prerecorded video samples were considered. The proposed system is tested and the results obtained are tabulated in Table 3. The total number of samples tested is 43 of which 39 were correctly recognized i.e. an accuracy of 90.67 %.

Alphabet	No of video samples tested	No of samples correctly recognized	Recognition rate
A	12	10	83.33 %
B	10	10	100 %
C	10	10	100 %
D	11	9	81.81 %
--------	Total no. of samples tested	Total no. of samples Recognized	Net Recognition rate
	43	39	90.67 %

Table 3. No of images used for testing ANN

5. Conclusion

The development and testing of certain vision based face detection algorithms has resulted in the design of accurate systems that give precise results. In this paper, a vision based recognition system is tested in real environment. The proposed method has been verified with a running video stream from a standard camera and satisfactory collection of data has been observed. With the desired output, the system is able to track relevant details and approximation of the original video feed. The derived results are found to be in order to the expected outcomes of the proposed approach which establishes the effectiveness of the system. Further this proposed method can also be expanded to consider other facial features such as nose or eye brows and even cheeks and now can be used efficiently along with any hand gesture based recognition system for extended applications.

Acknowledgement

The authors are thankful to the Ministry of Communication and Information Technology, Govt. of India for facilitating the research.

References

1. Dalal N, Triggs B. *Histograms of Oriented Gradients for Human Detection*. IEEE Xplore Digital Library.
2. Dalal N. Lecture notes on *Histogram of Oriented Gradients (HOG) for Object Detection*. Joint work with Triggs B, Schmid C.
3. Deniz O, Bueno G, Salido J and Torre F. De la. *Face recognition using Histograms of Oriented Gradients, Pattern Recognition*. Letters 32 (2011) 15981603
4. Jurafsky D, Martin J H. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics*. 2nd edition, Prentice-Hall, 2009.
5. Mudoi D, Kashyap P A. *Vision Based Data Extraction of Vehicles in Traffic*. IEEE Xplore Digital Library.
6. Jensen O H. *Implementing the Viola-Jones Face Detection Algorithm*. Kongens Lyngby 2008 IMM-M.Sc.-2008-93
7. Aran O, Keskin C, Akarun L. *Sign Language Tutoring Tool*. EUSIPCO'05, Antalya, September 2005.
8. Lee L, Tsai, Shieh J. *Applied the Back Propagation Neural Network to Predict Long-Term Tidal Level*. Asian Journal of Information 5(4):396 401, 2006
9. Nixon M S, Aguado A S. *Feature Extraction and Image Processing*. Second edition 2008
10. Zaki M M , Shaheen S I. *Sign Language Recognition Using a Combination of New Vision Based Features*. Pattern Recogn. Lett., vol. 32, no. 4, pp.572-577, March, 201
11. Moghadas K, Gholizadeh S. *A New Wavelet Back Propagation Neural Networks for Structural Dynamic Analysis.* Engineering Letters, 16:1, EL 16 1 03, 19 February 2008
12. Zaki M M, Shaheen S I. *Sign Language Recognition Using a Combination of New Vision Based Features*. Pattern Recogn. Lett., vol. 32, no.4, pp.572-577, March, 201
13. Chuang Z J, Wu C H, Chen W S. *Movement Epenthesis Generation Using NURBSBased Spatial Interpolation*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 11, pp. 1313 - 1323, November, 2006.
14. Yang H D, Lee S W. *Robust Sign Language Recognition with Hierarchical Conditional Random Fields"*. 20th International Conference on Pattern Recognition (ICPR), pp. 2202-2205, Istanbul, August, 2010

15. Duda R O, Hart P E, Stork D G. *Pattern Classification*. 2nd ed., Wiley India, Indian Reprint, New Delhi, 2009
16. Lu Y C. *Background Substraction based Segmentation using Object Motion Feedback.* First International Conference on Robot Vision and Signal Processing(RVSP), pp 224-227, Kaohsiung ,2011.
17. Berggren K, Gregersson P. *Camera Focus Controlled by Face Detection on GPU*. Lund University
18. Boots P J H M, Van Schenk Brill D. *Object Recognition by Contour Matching.* Fontys University of Professional Education, IPA Research Centre Eindhoven, Netherlands
19. Wang S B, Quattoni A, Morency L P, Demirdjian D, Darrell T. *Hidden Conditional Random Fields for Gesture Recognition.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp.1521-1527, 2006.
20. Rabiner L. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of the IEEE, vol. 77, no.2, pp. 257-286, 1989
21. Nazerfard E, Das B, Cook D J, Holder L B. *Conditional Random Fields for Activity Recognition in Smart Environments.* International Symposium on Human Informatics (SIGHIT), 2010
22 Neustaedter C. *An Evaluation of Optical Flow using Lucas and Kanades Algorithm.* Canada, 2002.
23 Lee H.K. and Kim J.H. *An HMM-Based Threshold Model Approach for Gesture Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 961-973, October, 1999.
24. Chetverikov D. *A Simple and Efficient Algorithm for Detection of High Curvature Points in Planar Curves.* Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, vol. 2756, pp. 746-753, 2003.
25. Bartlett M S, Lades H M, Sejnowski T J. *Independent Component Representations for Face Recognition.* Proc. SPIE Symp. Electronic Imaging: Science and Technology: Human Vision and Electronic Imaging III,
26. Donato G, Bartlett M S, Hager J C, Ekman P, Sejnowski T J. *Classifying Facial Actions*. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp.974-989, Oct. 1999