Visual Speech Recognition Using Artificial Neural Networking

Group members

Iftekhar Shamsuddoha - 13301053

Nowshin Sharmili - 13301027

Lamia Tasnim - 13301121

Tahsin Ahmed - 13301030

Supervisor:

Dr. Amitabha Chakrabarty

Submission Date:

22/08/2017

AUTHOR'S DECLARATION

We, hereby declare that this thesis is based on the results found by ourselves. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole or in part, has been previously submitted for any degree.

SIGNATURE OF THE AUTHORS:

.....

Iftekhar Shamsuddoha

13301053

.....

Nowshin Sharmili

13301027

.....

Lamia Tasnim

13301121

.....

Tahsin Ahmed

13301030

SIGNATURE OF THE SUPERVISOR:

.....

DR. AMITABHA CHAKRABARTY
ASSISTANT PROFESSOR
DEPT. OF COMPUTER SCIENCE, BRAC UNIVERSITY

ABSTRACT

Automatic Speech Recognition plays an important role in human-computer interaction, which can be applied in various application like crime-fighting and helping the hearing-impaired consists of two domain – Audio Speech Recognition and Visual Speech Recognition. This thesis is based on Recognition of Speech in the visual domain only.

This paper provides a new approach to lip reading Bengali words using a combination of the curvature of the inner and outer lips and Neural Networks. The method uses a more robust a faster algorithm to detect the lip contour than conventional methods used so far.

Processing multiple frames and by collecting the contours, we can predict the Bengali words that are stored inside the database. Our thesis will mainly focus on detecting some specific Bengali words.

ACKNOWLEDGEMENTS

We feel honored in expressing our heartfelt indebtedness and gratitude to our respected supervisor Dr. Amitabha Chakrabarty for guiding me with continuous encouragement, technical suggestions and valuable instructions throughout the thesis work.

Also, many thanks to Nahid Akhter for the luminescence of wisdom and sharing her thesis work with us.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

Chapter 1

INTRODUCTION

When it comes to language processing systems, Lip- Reading is notoriously. Depending on not only the context and knowledge of the language but also visual clues, lip reading has been topic of experiment and research for the past few decades. Being one of the most important and essential breakthroughs, research on visemes has been going on in different languages like French, English and Spanish as well as Bengali. However, the efficiency is debatable.

Although numerous researches have been done on lip-reading and they have all been ground breaking, but very little has been done in Bengali. Because of presence of similar pronunciation techniques for different alphabets, lip- reading in Bengali is particularly challenging and thus avoided by most. The most recent approach has been proposed by the paper "A Viseme Recognition System using Lip Curvature and Neural Networks to Detect Bangla Vowels" where the author has used combination of the curvature of inner and outer lips and Neural Networks.

Inspired by the paper, our thesis is all about increasing the efficiency of the existing techniques and proposing a more effective algorithm.

1.1 Objectives

1. It has been observed that it is possible for trained people to recognize or understand speech by simply looking at the shape of the mouth while speaking. Thus, it should be possible to create an algorithm that uses the power of machine learning to read lips with some amount of effectiveness.

Moreover, as human beings observe the shape of the mouth for reading lips, it was decided to use information on how much the mouth curves while speaking to be used in the machine learning process.

2. It was found that a lot of research has already been done on lip-reading English words and some amount on few other languages like Chinese, Arabic, Hindi and Tibetan. However, no information could be found by the author on lip reading in Bangla. Therefore, development of a system to lip-read Bangla visemes was desirable.

3. Most algorithms found on lip segmentation or contour extraction were iterative, which means using them in tracking videos would lead to a lot of time lag, so it was important to develop an algorithm with a good amount of accuracy that would read lips without the need of any iterative function to allow for quick lip reading.

Research on Lip-reading suggests that the field of Lip-reading is still in its infancy. A completely accurate and efficient lip-reading algorithm is yet to be developed. The objective of this research is to provide its contribution to this developing field with an algorithm that saves time and provides at least an above average accuracy. Contributions of this thesis are in three areas of lip reading. These are Lip Segmentation, Feature Extraction and Viseme Recognition.

1.2 Challenges

1. Since, the basic technique of lip reading includes the recognizing a sequence of shapes made by mouth and matching it specific word, vowel or sequence of words or vowels, it is pretty challenging to begin with. Explicitly speaking, the mouth forms between 10-14 diverse shapes known as visemes. On the other hand, speech contains around 50 individual sounds known as phonemes. As such, more than phoneme can be represented by a single visemes. Thus initiating an commotion, associating an array of visemes with a unique word or sequence of Bengali words is next to impossible. The challenge is choosing the one that the speaker has used.

2. Secondly, in most cases the speaker's lips are obscured or not totally visible. In those cases, the result is not correct or a precise detection is not possible.

3. A more difficult challenge is in recognizing, extracting and categorizing the geometric features of the lips during speech.

4. One is that beards and mustaches can significantly confuse visual speech recognition systems. Consequently, they are more successful with female than male speakers.

Unlike human interpreters, computers cannot grasp or analyze a lot of additional information essential for a perfect reading like context of the conversation, the speaker's body movements and a good knowledge of grammar, idioms and common speech. As such a barrier remains there.

1.3 Applications of Lip Reading

Lip-reading is a way of understanding speech by visually interpreting the movements of another's lips, tongue, and face. Body language, pace of speech, and the monitoring of syllables also play an important task, as well as context.

Lip-reading requires great concentration, and can be difficult if the person you are lip-reading talks fast or covers their mouth. Not all words are easy to lip-read, and it takes time to become proficient.

In noisy environments when it is hard to hear but necessary to listen, lip reading software is very important. Again, in case when situation demands long distance hearing, for instance, to interpret a criminal conservation, the software can play a vital part. Last but not the least, in the lives of people with hearing problems, a lip reading application can benefit all greatly.

1.4 Thesis Contribution

Many algorithms and models have been proposed by researchers for detecting objects and segmenting them. We have done a brief literature review illustrating some of these methods and provided a unique mixture of Active Contour Model and butterfly method which is supposed to outperform all the mentioned algorithms of literature review in many performance measuring parameters.

Chapter 2

LITERATURE REVIEW

2.1 Visual Speech Recognition

Consisting of two domains namely Audio Speech Recognition and Visual Speech Recognition, Automatic speech recognition is a segment of Artificial Intelligence and Neural Networks.

2.1.1 Image Acquisition

A webcam or a camera is utilized to get the video of a man talking such that from his articulates each syllable must be recognized, however video should maintain continuity, with no sound. This articulation of words ought to be taken a couple of times, to choose the perfect radiance, to such an extent that it is simpler to play out the resulting steps. This video is then spared in avi or mpgeav arrange [7]. The procured video is then separated into outlines or a picture arrangement, with the end goal that every video outline is presently a different picture record. This method is finished by utilizing MATLAB's image processing toolbox.

2.1.2 Lip Detection

The time has come to find the lips region in the face, after identifying the face. This should be possible by the Adaboost calculation utilizing Haar highlights. [68], is a machine learning algorithm, which was defined by Freund and Schapire [69]. Since lip area is localized at bottom half of face. So, to diminish the computational flaws and enhance the productivity, the upper portion of confront picture is expelled.

2.2 Visual speech recognition by recurrent neural networks

Pointing out a major flaw of acoustically based speech recognizers, Gihad and Si [31] stated that, with noise the performance of these recognizers depreciate significantly. As such they attempted to develop a program that recognizes speech with visual information of the speaker. In their research they wanted to extract visual speech through image processing in controlled environment. Considering the dynamic nature of speech patterns with respect to time as well as spatial variations in the individual patterns the researchers suggested recurrent neural networks architecture. Trained with no more than feed forward complexity, the recurrent network's desired behavior is based on characterizing a given word by well-defined segments. Implicitly speaking, this technique completes the implementation in two steps. Starting with sequences that are segmented individually, a generalized version of dynamic time warping is used to align the segments of all sequences afterwards. With each traversal, the weights of the distance functions used in the two steps are updated in a way that minimizes a segmentation error. Precisely speaking, the system has been successful in distinguishing between words with common segments tolerates to a great extent variable-duration words of the same class. Although the system has been tested on a few words and received satisfactory results, it is yet to be implemented on Bengali verses.

2.3 Visual Speech Recognition: Lip Segmentation and Mapping

Visual Speech Recognition: Lip Segmentation and Mapping [32] presents an advanced account of exploration performed in the areas of lip segmentation, visual speech recognition, and speaker identification and verification. A useful reference for researchers working in this field, this book contains the latest research results from renowned experts with in-depth discussion on topics such as visual speaker authentication, lip modeling, and systematic evaluation of lip features. This research talks about Discriminative lip motion features, Gesture coding, Lip analysis systems, lip contour extraction from video sequences, lip feature extraction along with lip modeling and segmentation.

2.4 Machine learning technique boosts lip-reading accuracy

A very recent research in East Angelia , UK has been found to be able to interpret mouthed words in a better accuracy that human lip readers[33]. This algorithm is designed in such a way that it doesn't need to know the context. Although it is still in research stage, there are scores of potential applications for technology that could automatically transform visual cues into accurate speech. While looking solely at visual inputs one of the team members, Dr Helen Bear says, "We're looking at… visual cues and saying how do they vary? We know they vary for different people. How are they using them? What's the difference? And can we actually use that knowledge in this particular training method for our model? And we can,"

"The idea behind a machine that can lip read is that the machine itself has got no emotions, it doesn't mind if it gets it right or wrong — it's just trying to learn. So in the paper… I've been showing how we can use those visual confusions to make better phoneme classifiers. So it's a new training method," she adds.

2.5 Lipnet

Recent Discoveries: "A team from the University of Oxford's Department of Computer Science has developed a new artificial-intelligence system called LipNet [34]. As Quartz reported, its system was built on a data set known as GRID, which is made up of well-lit, face-forward clips of people reading three-second sentences. Each sentence is based on a string of words that follow the same pattern.

The team used that data set to train a neural network, similar to the kind often used to perform speech recognition. In this case, though, the neural network identifies variations in mouth shape over time, learning to link that information to an explanation of what's being said. The AI doesn't analyze the footage in snatches but considers the whole thing, enabling it to gain an understanding of context from the sentence being analyzed. That's important, because there are fewer mouth shapes than there are sounds produced by the human voice.

When tested, the system was able to identify 93.4 percent of words correctly. Human lip-reading

volunteers asked to perform the same tasks identified just 52.3 percent of words correctly."

2.6 Performance Analysis of Automatic Lip Reading Based on Inter-Frame Filtering

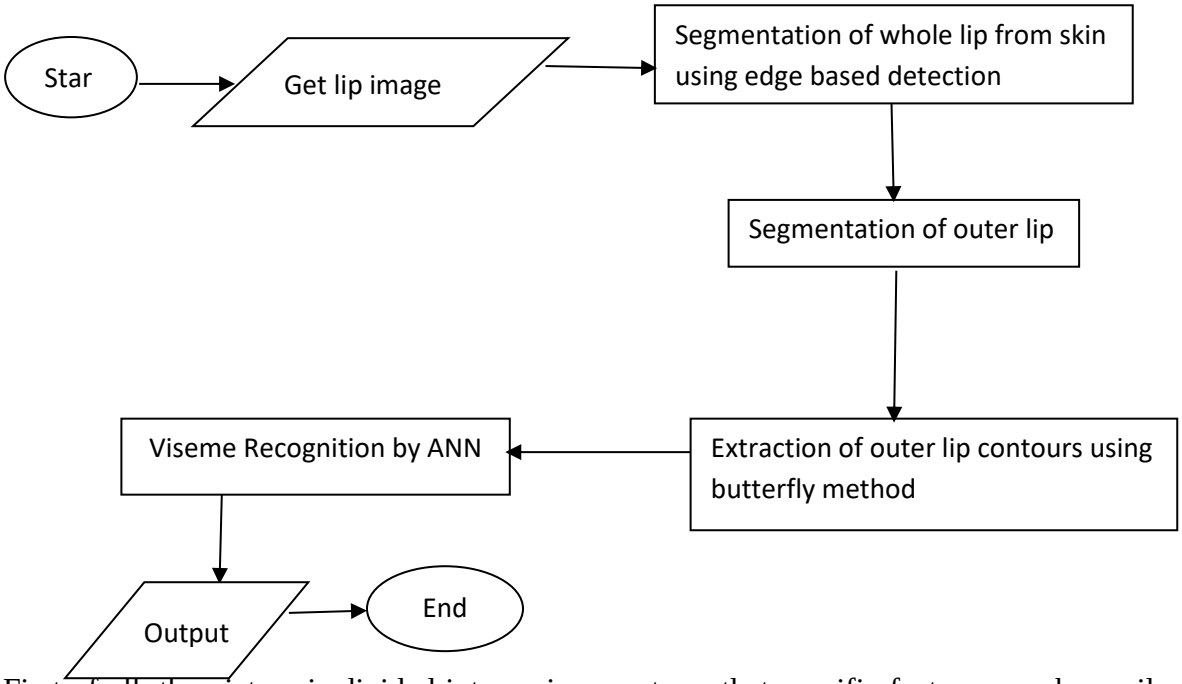
In noisy environments, Automatic lip reading is considered as a complementary method of automated speech recognition. Among the automated lip reading algorithms, the image transform based lip-reading (ITLR) is one of the most competitive. However the performance of this algorithm varies greatly with the variation of lighting. Mr. Kim [35] has proposed a very interesting inter frame filtering method known as RASTA. Being intended to be used for rejecting Stationary and white noise in speech signal processing, RASTA approach is used in the paper. Adding it in ITLR and analyzing the performance has been the min ambition. This algorithm proposes 2 merging techniques of integration and in the pre-integration process, inter frame filtering is done after the image transform process. Moreover, the effectiveness of high pass filtering and band-pass filtering has been evaluated as well.

WORKING PROPOSAL

We shall begin to explain in this chapter the method we came up with for the viseme recognition system.

3.1 Active Contour

3.1.1 Workflow:

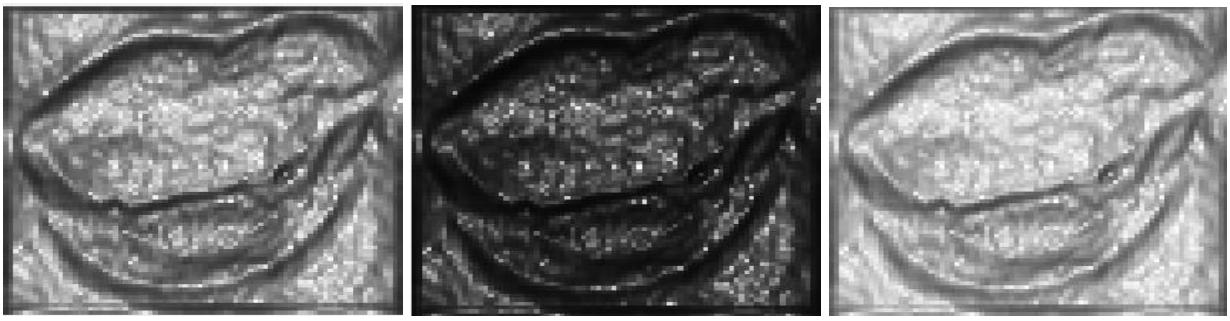


First of all, the picture is divided into various parts so that specific features can be easily made distinguishable and thus the software can work with that data to single out those features to make

a contour around them. This image segmentation is used to divide specific data so that it easier to make an outline around the lips. Active contour is a method to find closed contour around objects and uses energy minimization for the image segmentation. Edge-based active contour shown in "Edge- Based Active Contour With Level-Set Implementation" [70] searches for the edges of objects in the image. Stronger edges induce higher gradients, therefore energy functional can be said to be the inverse of the image gradient.

$$\Phi = \frac{1}{1 + \lambda \|\nabla I\|} \dots\dots\dots (1)$$

If the gradient is small, the energy functional Φ is one, and if the gradient is large, the energy functional is small, and thus the gradient indicates where the edge is.



(a) (b) (c)

Figure 3.1 Image (a) is the converted grayscale version of the input image. Image (b) is the energy functional λ=1. Image (c) is the energy functional λ=0.1

Then the total energy along the curve is defined as:

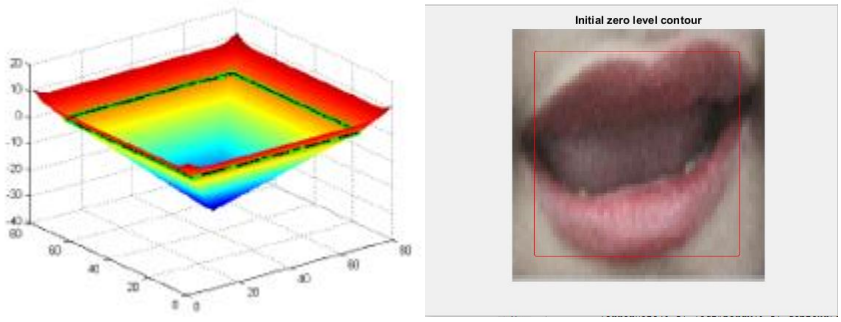
$$E(C) = \int_C \Phi \, ds = \int_0^L \Phi \, ds \dots\dots\dots (2)$$

where C is the region along the curve, s is the arc length, and L is the length of the curve.

$$C_t = \Phi K N - (\nabla \Phi . N) + \alpha \Phi N \dots\dots\dots (3)$$

3.2 Level-Set Implementation

This approach evolves the surface ψ instead of the curve C.



(A) (B)

Figure 3.2 Image (A) is the initial level set function ψ. The green and black line outlines the zero-level set. Image (B) displays the zero-level set as the curve on the input image.

The change in the surface lets new regions to appear and disappear. The equation which is used in "Edge- Based Active Contour With Level-Set Implementation" [70] for the curve evolution is:

$$\Psi_t = \left(\hat{\Phi}K + \nabla\hat{\Phi}.N - \alpha\hat{\Phi}\right)\|\nabla\Psi\| \dots\dots\dots(4)$$

$$K = \nabla.\left(\frac{\nabla\Psi}{\|\nabla\Psi\|}\right) \dots\dots\dots(5)$$

$$N = -\frac{\nabla\Psi}{\|\nabla\Psi\|} \dots\dots\dots(6)$$

$$\Psi_t = \hat{\Phi}\|\nabla\Psi\|\nabla.\left(\frac{\nabla\Psi}{\|\nabla\Psi\|}\right) + \nabla\hat{\Phi}.\nabla\Psi - \alpha\hat{\phi}\|\nabla\Psi\| \dots\dots\dots(7)$$

3.3 Algorithm in pseudo Code

Shown in "Edge- Based Active Contour With Level-Set Implementation" [70] is the pseudo code for energy evolution for lip detection.

Set ψ = ψ0

for k = 1 : #iterations

Compute current K using central difference

Compute current ψ , ψy using upwind difference

Compute current $\|\nabla\psi\|$ using upwind entropy

Set idx as indices for the zero-level set (narrow band where $\psi \approx 0$)

Set ψ(idx) using implicit level set method

Set Δt = 0.5 / max(ψt)

Evolve ψ near zero-level set by $\psi(t + \Delta t) = \psi(t) + \Delta t * \psi t$

Reinitialize ψ to the signed distance function to its zero level set.

End

3.4 Butterfly Method

The mask of the segmented image of the lip is taken which is then reduced using canny edge detection. Then the left and right most pixel points of the mask are found. Then the midpoint of the two points is used to find the boundary region which is 20% of the distance from the midpoint. Afterwards the lowest pixel is selected from within the boundary region, which is called the dipping point.

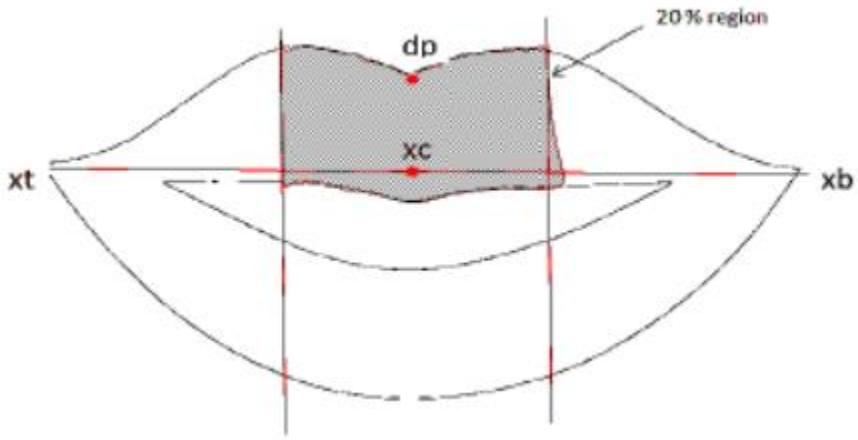


Fig.3.3 Detection of dipping point

Then 8 points are taken from each side of the dipping point and another 16 points are taken from the lower lip. These 32 points are then used to train the built-in Artificial Neural Network in MATLAB so that it can recognize which letter the input image matches with the most.

Upper right lip:

$$a_1x^2 + b_1x + C_1 = 0 \dots \dots \dots (8)$$

Upper Left lip:

$$a_2x^2 + b_2x + C_2 = 0 \dots \dots \dots (9)$$

Lower Lip:

$$a_3x^2 + b_3x + C_3 = 0 \dots \dots \dots (10)$$

As a result, the training image coefficients are compared to the coefficients of the input images to find out what letter is being said in the image.

Phoneme	Example	Viseme
অ	অজগর	অ
আ	আম	আ
ই	ইঁদুর	ই
ঈ	ঈদ	ই
উ	উত্তর	উ
ঊ	ঊষা	উ
এ	এক	এ
ঐ	ঐক্য	ঐ
ও	ওল	ও

ঔ	ঔষধ	ঔ

Table 4.0: Phoneme -Viseme chart

3.5 Advantages over other methods

There are many advantages of this proposed technique. First and foremost, it improves upon some of the drawbacks of the existing methods of contour extraction. Adding robustness and accuracy to image-based algorithms, it it extracts the curvature of the lips and so, the results are independent of the size or quality of the picture, illumination or mouth rotation.

As for model-based methods like ACM, ASM and AAM, the proposed method does away with the need to initially add manual landmarks to the image as well as the need to train the contour-extractor. This saves a lot of processing time, memory resources and the possibility of wrong initialization by the user. This makes the proposed method ideal to be used on low performance machines and simple smartphones. Moreover, the results show 12-20% less error and more accuracy compared to the result obtained by [1] in the research.

RESULTS AND ANALYSIS

In this chapter, the results of our proposed method will be explained thoroughly and these results will be compared with others.

Our experiment had three parts –Lip Segmentation, contour extraction and viseme recognition

4.1 Lip Segmentation

Edges produce higher gradient which show below how the boundaries have darker pixels than the rest.

The gradient (Fig 4.2) of the original image (Fig 4.1) was calculated using the formula below. A constant of 1 is added in the denominator to avoid division by 0.

$$\Phi = \frac{1}{1 + \lambda\|\nabla I\|} \dots\dots\dots (11)$$

Λ is added to the gradient function to control the effect of the energy function.

Fig 4.1 Original Image

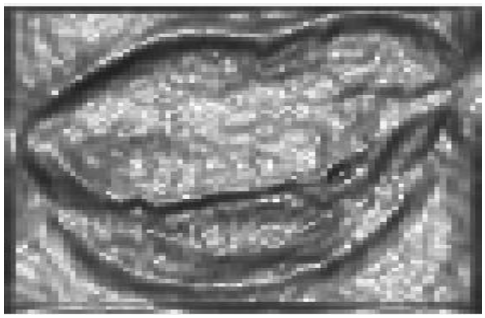


Fig 4.2 Image Gradient

Active contour explores the use of energy minimization as a framework to perform image segmentation. There are two popular variants in active contour: edge-based and region-based. In our project, we have chosen to use edge-based one.

Energy Function:

$$E(C) = \int_C \Phi \, ds = \int_0^L \Phi \, ds \, \ldots\ldots\ldots (2)$$

For the curve evolution, the energy function is minimized using derivation:

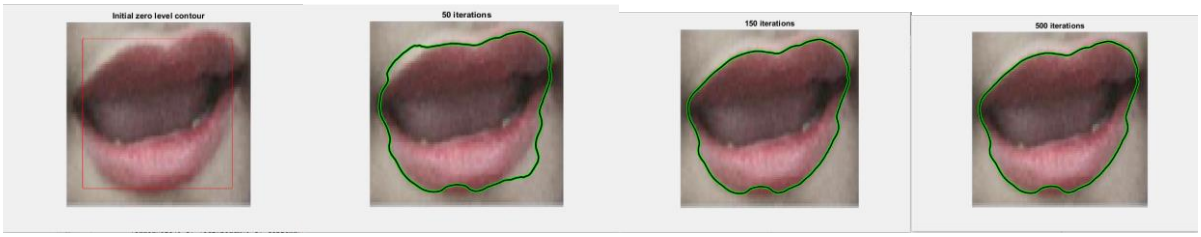
$$\frac{dE}{dt} = \frac{d}{dt}\int_0^L \Phi ds \, \ldots\ldots\ldots (12)$$

$$C_t = \Phi KN - (\nabla\Phi \, . N)N + \alpha\Phi N \, \ldots\ldots\ldots (13)$$

where N is the inward normal and K is the curvature.

If the initial curve lies inside the edge, in order to make the curve evolve outward, an inflationary term α needs to be added.

$$C_t = \Phi KN - (\nabla\Phi \, . N)N + \alpha\Phi N \, \ldots\ldots\ldots (13)$$



0 iteration 50 iteration 150 iteration 500 iteration

Fig 4.3: Sequence of the curve shrinking and converging on the edge of the lip.

4.2 Energy Curve

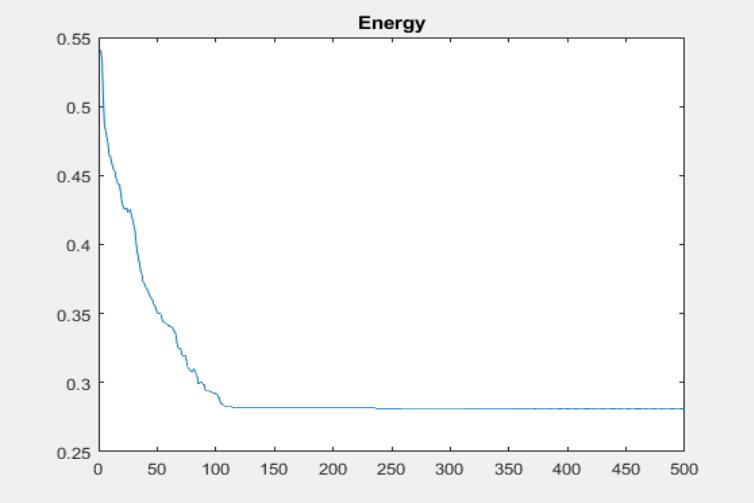


Fig 4.4: Energy Curve

One simple technique to evaluate the accuracy of the implementation is to plot the total energy on the curve over iteration. Since the objective is to minimize that energy through gradient descent, the energy should be decreasing until the curve reaches the boundary. From the figure, it is observed that the energy quickly decreased in the first 150 iterations; then the energy gradually approaches the energy limit.

4.3 Contour Length

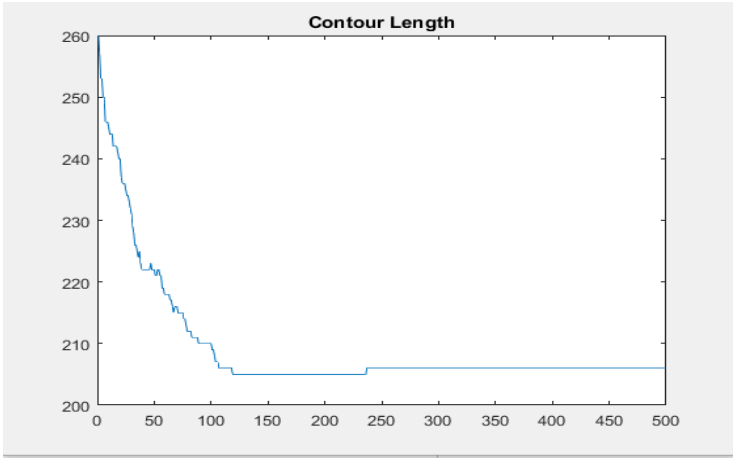


Fig 4.5: Contour Length

The contour length keeps decreasing as it shrinks to the boundary. Again, it can be seen that the contour length becomes quite stable as it approached 150 iterations where it almost reached the boundary position.

4.4 Butterfly Method in work

The dip separates the upper left and right lip. The left and right most pixel points are found. The midpoint of these two pixels is calculated and 20% boundary region from this midpoint is saved. The lowest pixel in this region is the dip of the lip. As the shapes can be considered to be a two degree parametric equations, we can simply use 'polyval' in matlab to find the coefficients of these equations by giving several coordinates of each shape.

Upper right lip: $a_1x^2 + b_1x + c_1$ ... ...(8)

Upper Left lip: $a_2x^2 + b_2x + c_2$ ... ...(9)

Lower Lip: $a_3x^2 + b_3x + c_3$ ... ...(10)

For pattern recognition, various machine learning tools are available, like Support Vector machines, KNN classifier and WEKA. For our experiments, we have used the Artificial Neural Networks tool available with the MATLAB package.

4.5 Observing Error Rate

λ	1	0.7	0.5	0.3	**0.2**	0.1	0.05	0.01
Error Rate(%)	1.0156	1.0312	0.9219	0.5625	**0.5313**	0.5469	0.5625	1.1875
Contour Length	185	187	187	185	184	184	185	181
Energy Limit	0.0226	0.0316	0.0428	0.06	0.0866	0.158	0.2713	0.6466
Iterations	365	296	326	499	395	327	304	189

Table 4.5a: Error Rate vs λ (α=0)

α	0	0.005	0.01	0.015	0.02
Error Rate(%)	0.5313	0.5313	0.5313	0.7969	0.8594
Contour Length	184	184	185	185	185
Energy Limit	0.0866	0.0866	0.0866	0.096	0.0969
Iterations	395	428	473	371	316

Table 4.5b: Error Rate vs α (λ=0.2)

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Proposing neural network as a multiclass pattern classifier, this thesis has attempted to identify visemes of Bangla vowels being spoken. Focusing on the video feature only, the success of the viseme classification appears to depend upon factors like choice of lip-localization and contour-finding algorithm, the choice of features extracted and the pattern recognition system used.

We have used a technique involving combination of conversion to different color spaces like HSV, CIELAB and CIELUV to localize the lip's inner and outer curves. This technique is apparently faster and more memory-efficient than using Active Shape Models and Active Appearance Models.

5.2 Future Work

This research has a lot of potential for future work. Among various segments, we have conducted a droplet amidst sea. Huge research is yet to be conducted with audio and video, with Bengali consonants as well as Bengali words.

REFERENCES

[1] Akhter, N. and Chakrabarty, A. (2016) A Survey-based study on lip segmentation techniques for lip-reading Applications, International Conference on Advanced Information and Communications Technology (ICAICT), June 2016.

[2] Naz, B. and Rahim, S. (2011) B Audio-Visual Speech Recognition Development Era; From Snakes To Neural Network: A Survey Based Study. Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition Vol. 2, No. 1, 2011.

[3] Miyaki, T., Sughara et al. (2006) Active Contour Model with Splitting Characteristics for Multiple Area Extractions and its Hardware Realization. SICE-ICASE International Joint Conference 2006.

[4] Kim, J., Choi S. and Park S.(2002) Performance Analysis of Automatic Lip Reading Based on Inter-Frame Filtering, Dept. of Information & Communications Engineering, Dongshin University, Naju, Korea, http://www.dtic.mil/dtic/tr/fulltext/u2/p014024.pdf

[5] Kass, M. et al. (1987) Snakes: Active contour models. International Journal of Computer Vision, pp 321-331.

[6] Lomax, N.,(2016), Machine learning technique boosts lip-reading accuracy, https://techcrunch.com/2016/03/24/tech-to-read-my-lips/

[7] Mei, L.P. (2014) Interpretation Of Alphabets By Images Of Lips Movement For Native Language. Universiti of Teknologi, Malaysia, 2014.

[8] Wang, S.L. et al (2007) Robust lip region segmentation for lip images with complex background. Science Direct, pp 3481 – 3491, 2007.

[9] Rabi G. and Lu S., "Visual speech recognition by recurrent neural networks", J. Electron. Imaging. 7(1), 61-69 (Jan 01, 1998). ; http://dx.doi.org/10.1117/1.482627

[10] Saini, N. and Singh, H. (2015) Comparison of two different approaches for multiple face detection in color images. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 3, Issue 1, pp 2321-2004.

[11] Zurada, J.M. (1992) Introduction to Artificial Neural Systems, 1992 edition, pp 1-21.

[12] Md. Khalilur Rahman (2005) Neural Network using MATLAB. BRACU, Dhaka, Bangladesh, Powerpoint Presentation, 2005.

[13] Hashimoto, M., Kinoshita, H. and Sakai, Y. (1994) An Object Extraction Method Using Sampled Active Contour Model. IEICE Trans. D-II, Vol.J77-D-II, No.11, pp.2171-2178, 1994.

[14] Sughara, K., Shinchi, T. and Konishi, R. (1997) Active Contour Model with Vibration Factor. IEICE Trans. DII, Vol.J80-D-II, No.12, pp. 3232-3235.

[15] Bregler C. and Konig, Y. (1994) Eigenlips for robust speech recognition. Proceedings of ICASSP94, Adelaide, Australia, pp. 669–672, April 19-22, 1994.

[16] Turk, M. and Pentland, A. (1991) Eigenfaces for Recognition. Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

[17] Gurbuz, S., Patterson, E.K., et al. Lip-Reading from Parametric Lip Contours for Audio-Visual Speech Recognition. Department of Electrical and Computer Engineering, Cemson University, Clemson, SC 29634, USA.

[18] Speech Recognition. Wikipedia, 2015. https://en.wikipedia.org/wiki/Speech_recognition

[19] Lee, C.H. et al. (1999) Automatic Speech and Speaker Recognition-Advanced Topics , Springer, third Edition.

[20] Petajan, E.D. (1984) Automatic lipreading to enhance speech recognition. Proceedings of Global Telecomm. Conf., Atlanta, GA, 1984, pp. 265–272.

[21] Beale, R. and Finaly,J. (1992) Neural networks and pattern recognition in human-computer interaction . Neural networks and pattern recognition in human-computer interaction, pp. 460.

[22] Lippmann, R.P. (1990) Review of Neural Networks for Speech Recognition, Readings in Speech Recognition . Waibel and Morgan Kaufmann Publishers, pp. 374-392, 1990.

[23] Stork, D.G. , Wolf, G. and Levinet, E. (1992) Neural network lipreading system for improved speech recognition. IJCNN, 1992.

[24] Yuhas, B.P., Goldstein, M. H., et al. (1989) Integration of acoustic and visual speech signals using neural networks. IEEE Communications Magazine, 1989.

[25] Kabre, H. (1997) Robustness of a chaotic modal neural recognition network applied to audio-visual speech. Neural Networks for Signal Processing, pp. 607 - 616, September 1997.

[26] Cappe , O. and Moulines, E. (2005) Inference in Hidden Markov Models. Springer, pp 42, 2005.

[27] Govind. Introduction to Hidden Markov Models. Lecture 12, CEDAR, Buffalo (Powerpoint Presentation)

[28] Hulbert, A. Poggio, T. (1998) Synthesizing a Color Algorithm from Examples. Science, vol. 239, pp. 482-485 ,1998.

[29] Canzlerm, U., Dziurzyk, T. (2002) Extraction of Non Manual Features for Video based Sign Language Recognition. Proceedings of IAPR Workshop, pp. 318-321, 2002.

[30] Leung, S.-H., Wang, S-L., Lau, W.-H. (2004) Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. IIEEE Transactions on Image Processing, vol.13, no.1, pp. 51-62, 2004.

[31] Lucey, S., Sridharan, S. and Chandran, V. (2002) Adaptive mouth segmentation using chromatic features. Pattern Recognition Lett, vol. 23, pp. 1293-1302, 2002.

[32] Lucey, S., Sridharan, S. and Chandran, V. (2000) Initialised eigenlip estimator for fast lip tracking using linear regression. Proceedings of the 15th International Conference on Pattern Recognition, vol.3, pp.178-181, 2000.

[33] Nefian, A. et al. (2002) A coupled HMM for audio-visual speech recognition. Proceedings of ICASSP, pp. 2013–2016, 2002.

[34] Guan, Y.-P. (2008) Automatic extraction of lips based on multi-scale wavelet edge detection. IET Computer Vision, vol.2, no.1, pp.23-33, 2008.

[35] Eveno, N., Caplier, A., Coulon, P. (2004) Accurate and quasi-automatic lip tracking. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pp. 706 – 715, 2004.

[36] Kaucic, R., Dalton, B., Blake, A.(1996) Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications. Proceedings of the 4th European Conference on Computer Vision, vol. II, 1996.

[37] Cootes, T. F. 9(2004) Statistical Models of Appearance for Computer Vision. Technical report, University of Manchester , 2004.

[38] Padilla, R., Costa Filho C. F. F. and Costa M. G. F. (2012) Evaluation of Haar Cascade Classifiers Designed for Face Detection. World Academy of Science, Engineering & Technology, Issue 64, pp 362, 2012.

[39] Sagheer A., Tsuruta, N. Taniguchi, R. Arabic Lip Reading System: A combination of Hypercolumn Neural Network Model with Hidden Markov Model.

[40] Kalbkhani, H. and Amirani, M.C. (2012) An Efficient Algorithm for Lip Segmentation in Color Face Images Based on Local Information. J. World Electrical Engineering Technology vol. 1(1), pp 12-16, 2012.

[41] Badura, S., Mokrys, M. (2015) Feature extraction for automatic lips reading system for isolated vowels. The 4th International Virtual Scientific Conference on Informatics and Management Sciences, March 23, 2015.

[42] KerenYu, Jiang, X. and Bunke, H. Sentence Lipreading Using Hidden Markov Model with Integrated Grammar. Department of Computer Science, University of Bern, Switzerland.

[43] Alan C. Bovik (2009) Essential Guide to Video Processing, Academic Press, pp 720, 2009.

[44] Ukai, N. et al. GA Based Informative Feature for Lip Reading. Department of Information Science, Gifu University, Japan.

[45] Mattehews, I. et al. A comparison of Active Shape Model and Scale Decomposition Based features for Visual Speech Recognition. School of Information Systems, University of East Anglia, Norwich, UK.

[46] Ahmad B.A. Hassanat. Visual Speech Recognition. IT Department, Mutah University, Jordan.

[47] Werda, S., Mahdi, W. and Hamadou, A.B. (2007) Lip Localization and Viseme Classification for Visual Speech Recognition. International Journal of Computing & Information Sciences, vol. 5, No.1, pp. 62-75, April 2007.

[48] Image Processing: Morphology-Based Segmentation using MATLAB with program code. www.code2learn.com/2011/06/morphology-based-segmentation.html.

[49] Stillittano, S., Girondel, V. and Caplier, A. (2013) Lip Contour Segmentation and Tracking compliant with lip reading application constraints. Machine Vision and Applications, vol. 24, Issue 1, pp. 1-18, January 2013.

[50] Chen, Q.C. et al. (2006) An Inner Contour Based Lip Moving Feature Extraction Method for Chinese Speech. International Conference on Machine Learning and Cybernetics, August 2006.

[51] Kang, S.H., Song, S.H., Lee, S.H. (2012) Identification of Butterfly Species with a single Neural Network System. Journal of Asia-Pacific Entomology, v. 15(3), pp. 431-435, September 2012.

[52] Butt, W.R. and Lombardi, L. (2013) Comparisons of Visual Features Extraction Towards Automatic Lip Reading. 5th International Conference on Education and New Learning Technologies, Barcelona, Spain, March 2013.

[53] Mishra, A.N. and Chandra, M. (2013) Hindi Phoneme-Viseme Recognition from Continuous Speech. International Journal of Signal and Imaging Systems Engineering, Volume 6 , No. 3, pp. 164-171, January, 2013.

[54] Liew, A.W.C., Wang, S. Visual Speech Recognition: Lip Segmentation and Mapping, Medical Information Science Reference.

[55] Luo, X. (2006) Algorithms for Face and Facial Feature Detection. Master of Science Thesis, Tampere University of Technology.

[56] Kumar, V., Agarwal, A. and Mittal, K. (2011) Tutorial: Introduction to Emotion Recognition for Digital Images. HAL Archives-Ouverte, February, 2001. https://hal.inria.fr/inria-00561918

[57] Zhang, D. et al. (2013) The Lip Position Analysis of the main consonant /w/ in Tibetan Xiahe Dialect. International Workshop on Computer Science in Sports, 2013.

[58] Lip Reading. Wikipedia, 2016. https://en.wikipedia.org/wiki/Lip_reading

[59] Feng, W. (2012) A Novel Lips Detection method Combined Adaboost Algorithm and Camshift Algorithm. The Second International Conference on Computer Application and System Modeling, 2012.

[60] Hoai, B.L., Hoai, V.T. and Ngoc, T.N. Lip Detection In Video Using Adaboost and Kalman Filtering.Http://tapchibcvt.gov.vn/files/_layouts/biznews/uploads/file/Uploaded/hai/Kalman%20Filtering.pdf

[61] Toshio M. (2006) Active Contour Model with Splitting Characteristics for Multiple Area Extractions and its Hardware Realization. SICE-ICASE International Joint Conference, pp. 5723-5726, 18-21 October, 2006.

[62] Contrast Enhancement Techniques, Mathworks website, http://www.mathworks.com/help/images/examples/contrast-enhancement-techniques.html

[63] Discrete Cosine Transform, Mathworks website, http://www.mathworks.com/help/images/discrete-cosine-transform.html

[64] Active Shape Model, Wikipedia, 2016. https://en.wikipedia.org/wiki/Active_shape_model

[65] Lewis, T. and Powers, D.M.W., Lip Feature Extraction Using Red Exclusion. http://crpit.com/confpapers/CRPITV2Lewis.pdf

[66] Ahmad, N. A Motion Based Approach for Audio-Visual Automatic Speech Recognition (2011), Thesis paper, Department of Electronics and Electrical Engineering, Loughborough University, U.K., May 2011.

[67] Canny Edge Detector algorithm Matlab Codes, http://robotics.eecs.berkeley.edu/~sastry/ee20/cacode.html

[68] XINGHAN_Master_Thesis] Adaptive Boosting (AdaBoost) [http://en.wikipedia.org/wiki/AdaBoost

[69] Y. Freund, "An adaptive version of the boost by majority algorithm", in COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers, 1999.

[70] Zhao, S., "Edge-Based Active Contour with Level-Set Implementation", 27 April, 2014.