International Journal of Soft Computing 5 (1): 13-18, 2010
ISSN: 1816-9503

A Novel Approach Integrating Geometric and Gabor Wavelet Approaches to Improvise Visual Lipreading

[1]B. Sujatha and [2]T. Santhanam
[1]MTWU-Kodaikanal, Meenakshi College for Women, Chennai, India
[2]Department of Computer Science, D.G. Vaishnav College, Chennai, India

Abstract: Lipreading a perception of speech for listeners with hearing impairment is purely based on observing the lip movements under noisy conditions where visual speech information plays an important role. Lipreading, a visual modality which involves watching the movement of lips constitutes 1/3 of the conveyed message. This study investigates the use of two feature extraction methodologies for recognizing isolated words. The first type is based on a geometric approach which extracts the features like inner height, inner width, outer height and outer width of the lips while the second type is based on a set of Block-Based Gabor-Wavelet co-efficient extracted from each frame. Then these two features are given as input into the Ergodic Hidden Markov Model for recognizing the words.

Key words: Lip reading, visual feature extraction, lip segmentation, Gabor Wavelet Transform model, Height-Width model, Ergodic Hidden Markov model

INTRODUCTION

Human speech perception is a multimodal process (Potamianos *et al*., 2004). Recent studies show that information related with speech includes audio signs and visual signs. Audio sign based systems are not very safe because of the fact that the signs are affected by different noises drastically. Also, processing the extracted audio information of speaker is a complex computing processes and this is a major problem in speech recognition. To increase the performance of speech recognition, visual information is used as supplementary information along with audio information. In some areas like, facial expression, sign language and gesture analysis the lip movements are considered as visual signs.

Lipreading, a visual modality which involves watching the movement of lips constitutes 1/3 of the conveyed message (Terry, 2007). To enhance the level of speech understanding, human listener can use visual cues, such as lip and tongue movements (Rabinar and Juang, 1993). The application automatic lipreading are examined in two contexts. They are speech recognition and analysis of sign language in which speech recognition is the most important components for human communication and Human Computer Interaction.

When working with gray scale images some methods use the histogram to get the lip corner while other methods such as edge-detect operators enhance the image. To get the lip contour directly from the color images Yaling and Minghui (2008) uses different color spaces like RGB and HSV. The change in lip contour is bigger than the one in other area as the mouth area has high edge content. Still the efficiency is curtailed in case of occlusions such as beard images.

Werda *et al*. (2005) presents a method allowing to carry out a spatial-temporal tracking of some Points Of Interest (POI) in the speaker's face and to indicate the different configuration of the mouth through visemes (visual phonemes).

Later on these visemes will be associated to relatively precise physical measures like the spreading of the lips and mouth height in order to establish a correlation between the phoneme and viseme. Successive pictures of a video sequence a re used to track the POI on lip contours. The Freeman Coding of direction is used to track the quick movements of the lips in all directions.

Nguyen and Milgram (2008) proposed an efficient method for lip shape deformation modeled by a statistically deformable model based on Active Shape Model. The algorithm for face detection and lip tracking include neural networks, template matching, Active Shape/Appearance Models and AdaBoosting. For localizing and tracking the mouth, Gaussian Mixture Models or deformable contours (snakes) are used with any of the above framework.

Zhao (2009) presented local spatiotemporal descriptors to represent and recognize isolated phrases

Corresponding Author: B. Sujatha, MTWU-Kodaikanal, Meenakshi College for Women, Chennai, India

based solely on visual input. In an image each pixel produced a binary code by thresholding its neighbor pixel with the value of the center pixel. Appropriate features of lips are extracted by using binary pattern operator which is a gray-scale invariant texture primitive statistics. It shows an excellent performance in the classification of various kinds of textures.

In this study image and shape based approach are used to extract features from lip image frames in order to recognize words.

MATERIALS AND METHODS

Lip segmentation: Lip boundary extraction is an important problem in lipreading. Lip texture information is more valuable than using the lip boundary information in Audio-Visual Speech Recognition (Ozgur *et al*., 2008). The main objective of automatic and semiautomatic methods for the extraction of visual indices necessary to recognize visual speech. Lip area is segmented from the face image using the given lip segmentation algorithm and it is shown in Fig. 1.

Steps involved in lip segmentation technique

Step 1: Input the lip image frames one by one.

Step 2: Convert the given image into gray scale for preprocessing.

Step 3: Create the signed distance map by masking using initial mask.

Step 4: Get the narrow band of the above image.

Step 5: Find the interior and exterior mean to get the curvature.

Step 6: Find the gradient descent from the curvature image to minimize the energy.

Step 7: For re-initialization of level set of the image the sussman filter (Zhao, 2009) technique was used.

Step 8: Crop the segmented lip area for feature extraction techniques.

Feature extraction: The key idea of the image pattern is to map from a large, complex problem space into a small, simple feature space.

Every type of application uses a different kind of mapping, as mapping into the feature space is also the hard part for any pattern. For any lipreading system, the features of lip are extracted by Geometry based approach and Image based approach. In first method, the mouth images are analyzed to extract the parameter describing the geometry and shape of the mouth. This approach extracts the features like inner height, inner width, outer height and outer width of the lips from each frame (Fig. 2).

Fig. 1: Lip segmentation technique

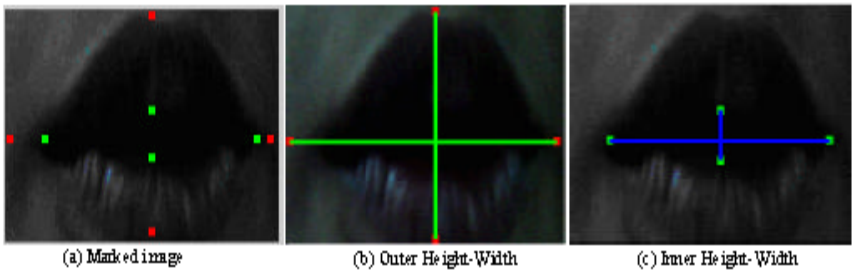


Fig. 2: Height-width model (geometric based) approach

In the second approach the features are extracted by directly operating on the pixel intensities of the image by transforming the two-dimensional mouth image into a feature vector.

Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT) are some techniques used in this type of approaches. The features extracted are having high dimensionality of which 85-95% of the transformed coefficients of energy are used for further processing. So, the approaches when considered separately yielded lesser accuracy (52 and 59.29%) whereas when combined produced better accuracy of 66.83%.

Height-width model: The geometric or shape based approach called HW model is used for feature extraction. Based on the changes in lip shape the geometric features are extracted. The input for this technique is the output of the lip segmentation algorithm. For feature extraction, static image frames for each word is used. These features vary according to the persons lip thickness, mouth openness, visibility of teeth and tongue.

From the segmented lip image, the features such as the Innerlip Heigh (IH), Innerlip Width (IW), Outerlip Height (OH) and Outerlip Width (OW) are manipulated. For each frame four related features are listed in Table 1 along with the frame names for example 1-2 frame means for digit '1' 2nd frame (Fig. 3).

Figure 3 shows the feature space of the words namely 'one', 'two', 'three' and 'four'. The high discrimination of the features can be clearly seen from this column chart.

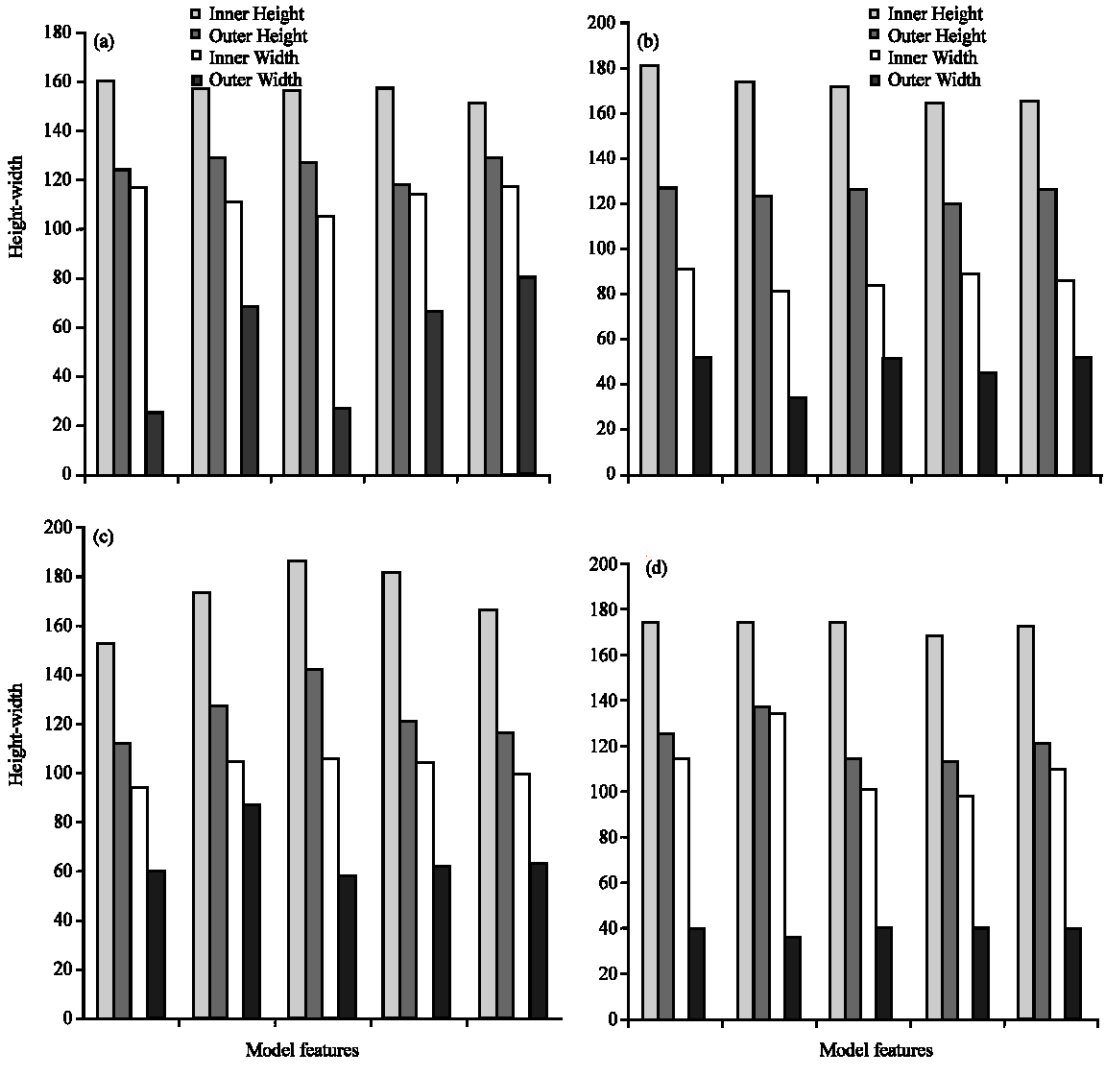
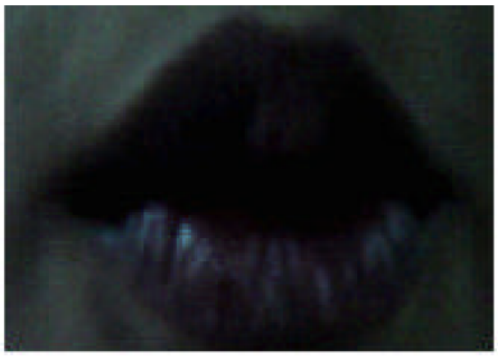


Fig. 3: Comparison chart for HW model features. a) One, b) Two, c) Three and d) Four

Table 1: Extracted features using Height-Width Model

Frame name	Inner Height (IH)	Inner Width (IW)	Outer Height (OH)	Outer Width (OW)
1-2	25	124	117	160
1-3	68	129	111	157
1-5	27	127	105	156
1-7	66	118	114	157
1-9	80	129	117	151



(a) Segmented Lip image

(b) Mouth openness

Fig. 4: Gabor Wavelet Transform (image based) approach

Gabor Wavelet Transform model: Gabor wavelets are widely used for analyzing images and computer vision.

The Gabor wavelet transform provides a way to analyze images and elaborate an as a frame for understanding the orientation and spatial frequency selective properties of simple neurons. Figure 4 the GWT is one of the image based approach and it is used for dimensionality reduction is also called as Block-Based Gabor-Wavelet Transform. Daugman defines that simple cells in the visual cortex can be modeled by Gabor functions. A Gabor kernel function is the product of an elliptical Gaussian envelops and a complex plane wave to determine the scale and direction of Gabor function (Sahoolizadeh *et al*., 2008):

$$\psi_{k,d}(x,y)=\frac{\|k\|}{\sigma^2}e^{-\frac{\|k\|^2\|z\|^2}{2\sigma^2}}\cdot\left[e^{ikx}-e^{-\frac{\sigma^2}{2}}\right] \quad (1)$$

Table 2: Extracted features using GWT Model

Frame name	Area	Bounding rectangle	Center point
1-2	9958	[59 50 163 120]	[149 104]
1-3	9217	[60 50 160 114]	[148 102]
1-5	9211	[60 55 159 108]	[147 106]
1-7	12023	[58 53 160 117]	[143 113]
1-9	12413	[55 57 154 120]	[140 119]

where, k is wavelength and orientation of the kernel $\psi_{k,d}$ (x, y) in image coordinates Gabor transform of an image is obtained by convolving the Gabor kernels with the image (Zhang and Yao, 2008).

The advantages of GWT are invariant to some degree to affine deformation and homogeneous changes in image brightness (Nguyen and Milgram, 2008). The optimal parameters of each Gabor-Wavelet reflect the underlying image structure.

The segmented lip image is used in this appearance or image based approach using Gabor Wavelet Transform techniques.

The main advantage of the wavelet image is to extract the inner features easily by using horizontal, vertical and diagonal features of the wavelet transform. In this study GWT technique is used as a segmentation filter to get the correct mouth openness for calculating the features.

From the extracted lip image the center point, bounding rectangle and the area of the segmented image features are calculated and shown in Table 2.

Figure 5 highlights the area dimensionality of the mouth openness in considering the words ‘one, ’two’, ‘three’ and ‘four’. The disparity to the values adds to the recognition of the words.

Ergodic Hidden Markov model: Use of traditional methods such as Hidden Markov Model, Support Vector Machine and Artificial Neural Networks are difficult for articulation and pronunciation variations between inter-speaker and intra-speaker.

To model these variations add more mixtures envisioning to corresponding the extra Gaussian modes (Terry, 2007). HMM was revised as Ergodic HMM and it defines each verbal unit as a state, then someone may visit any state from any other state in finite time. Important consequence of this model is that the individual verbal units are still modeled independently.

For each frame ‘t’ and each state ‘j’ of the Ergodic HMM, the accumulated log-probability is calculated by Viterbi search. By combining the accumulated log-probability L_e (t, j) is calculated by Viterbi search. By

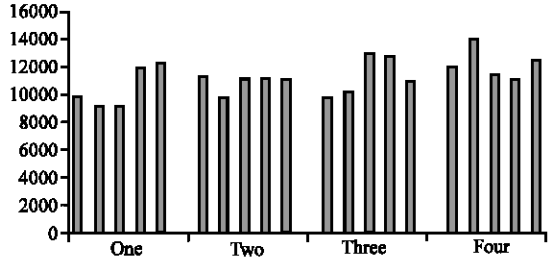


Fig. 5: Area of segmented image features using GWT approach

combining the accumulate log-likelihood $L_e(t, W_{m, s_m})$ and $L_e(t, j)$ yet another score for W_m terminating at frame 't' is defined:

$$S_e(t, W_m) = \{L_e(t, W_{m,s_m}) - L_e(t', W_{m',s_{m'}})\} - \{\max_j L_e(t, j) - \max_j L_e(t', j)\} \quad (2)$$

Where the first term on the right-hand side is identical with $S_e(t, W_m)$. It is expected that the second term

$$\{\max_j L_e(t, j) - \max_j L_e(t', j)\} \quad (3)$$

is a close approximation to Log Probability. This scoring method is called the Ergodic Method. Every state of the model can be reached in a single step from every other state in an Ergodic HMM (Ghayoori *et al*., 2005). This model seems to be the best model for any language because it assumes that every word in any language is reachable from every other. This enables the Ergodic model to adapt itself to any new words that are added to the language.

The Ergodic approach to speech modeling also leads to speech analysis with fixed temporal granularity but fine resolution as opposed to variable granularity with coarse resolution. This higher resolution approach directly allows for the explicit modeling of speech dynamics, while the traditional approach requires extra features.

The purpose of an Ergodic HMM is exploited to normalize the log-likelihood of a hypothesis. For each frame and each state the accumulated log-probability is calculated by the Viterbi search (Ozeki, 1996). A wide range of applications of Ergodic method are enumerated: word-spotting, rejection of out-of-vocabulary words, rejection of misrecognition, continuous speech recognition, text-to-speech system are some of the key areas. In this system only two dimensional static images are used for feature extraction and recognition. The features available in Table 1 and 2 the Ergodic HMM was implemented as a recognizer. In this model, for initialization the random data are used and after that the combined features are loaded for recognition of a word.

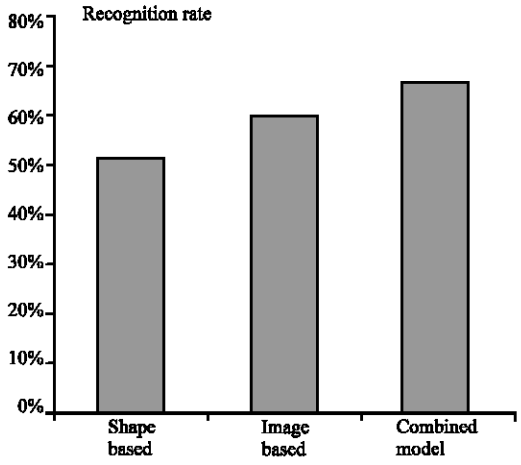


Fig. 6: Comparison of accuracy (%)

Table 3: Recognition rate

Model	Values (%)
Shape based (Bagai *et al*., 2009)	52.00
Image based (Ghayoori *et al*., 2005)	59.29
Proposed model	6.83

RESULTS AND DISCUSSION

A single feature is not sufficient for speech recognition. To increase the robustness as well as accuracy, two or more features are combined together. The database used for the experiments is a mugshot database created by recoding Audio Video Interface (AVI) files and it is converted to frames for further processing. Database consists of five speakers speaking in a controlled, studio environment. Additionally each speaker took part in ten recording sessions, thus giving one hundred instances of each word which are used for the models discussed in this study. After combining the features of both GWT and HW model, using Ergodic HMM the recognition rate will be increasing. The combination of GWT and HW model features results in an improved accuracy of 66.83% which is shown in the Fig. 6.

The comparative analysis of the proposed method is shown in Table 3. It is very evident from the Table 3 that the accuracy of the proposed model performs reasonably better than the existing approaches reported in (Ghayoori *et al*., 2005; Bagai *et al*., 2009).

RECOMMENDATIONS

The scope of the research can be further extended by considering the digits other than one, two, three and four and also to explore new techniques for feature extraction.

CONCLUSION

In this study the comparison of recognition performance by combining shape based and image based features has shown considerable improvement in the recognition rate.

REFERENCES

Bagai, A., H. Gandhi and R. Goyal, 2009. Lip-reading using neural networks. Int. J. Comput. Sci. Networks Sec., 9: 108-111.

Ghayoori, A., F. Hendessi and A. Sheikh, 2005. Application of smooth Ergodic Hidden Markov model in text to speech systems. Proc. Int. J. Signal Proces., 2: 151-157.

Nguyen, Q.D. and M. Milgram, 2008. Multi features active shape models for lip contours detection. Proc. Int. Confer. Wavelet Anal. Pattern Recogn., 1: 172-176.

Ozeki, K., 1996. Likelihood normalization using an Ergodic HMM for continuous speech recognition. Proc. Int. Confer. Spoken Language Proces., 4: 2301-2304.

Ozgur, E., B. Yilmaz, H. Karabalkan, H. Erdogan and M. Unel, 2008. Lip segmentation using adaptive color space training. Proceedings of the International Conference on Auditory-Visual Speech Processing, Sept. 26-29, Tangalooma Wild Dolphin Resort, Moreton Island, Queensland, Australia, pp: 219-222.

Rabinar, L.R. and B.H. Juang, 1993. Fundamentals of Speech Recognition (Prentice Hall Signal Processing Series). Prentice Hall PTR., USA., ISBN-10: 0130 151572.

Sahoolizadeh, H., D. Sarikhanimoghadam and H. Dehghani, 2008. Face detection using gabor wavelets and neural networks. World Acad. Sci. Eng. Technol., 45: 552-554.

Terry, L., 2007. Ergodic Hidden Morkov models for visual-only isolated digit recognition. M.S. Thesis, Department of Electrical Engineering and Computer Science, Evanston, Northwestern University, pp: 90.

Werda, S., W. Mahdi and A.B. Hamadou, 2005. A spatio-temporal technique of viseme extraction: Application in speech recognition. Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems, Nov. 27-Dec. 1, Yaoundé, Cameroon, pp: 32-38.

Yaling, L. and D. Minghui, 2008. Lip contour extraction based on manifold. Proceedings of the International Conference on MultiMedia and Information Technology, Dec. 30-31, Three Gorges, China, pp: 229-232.

Zhang, S. and H. Yao, 2008. A novel feature-level multiple HMMs classifier for Lipreading based on Ada-Boost Gabor kernels selection. Proceedings of 11th Joint Conference on Information Sciences, Nov. 10-12, Atlanta Press, pp: 1-6.

Zhao, G., 2009. Mark burnard and matti pietikainen, lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia, 11: 1254-1265.