

Analysis of Visual Isolated Words Using Geometric and Fourier Features for Audio Visual Speech Recognition

BORDE Prashant L.

Vision and Intelligent
System Lab

Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada
University, Aurangabad (MS) India.
borde.prashantkumar@gmail.com

MANZA Ramesh R.

Biomedical Image Processing Lab
Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada
University, Aurangabad (MS) India.
manzaramesh@gmail.com

KULKARNI Sadanand A.

Vision and Intelligent
System Lab

Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada
University, Aurangabad (MS) India.
sankalpsadanand.georai@gmail.com

YANNAWAR Pravin L.

Vision and Intelligent
System Lab

Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada
University, Aurangabad (MS) India.
pravinyannawar@gmail.com

Abstract: Automatic Speech Recognition (ASR) by machine is an attractive research topic in signal processing domain and has attracted many researchers to contribute in this area. In recent year, there have been many advances in automatic speech reading system with the inclusion of audio and visual speech features to recognize words under noisy conditions. The objective of audio-visual speech recognition system is to improve recognition accuracy. Considering an image sequence of a person pronouncing a word, a full image analysis solution would have to segment the mouth area, extract relevant features, and use them to be able to classify the word from those visual features. In this paper, we computed visual features using *Geometric Distance Features* and *Fourier Features* on 'vVISWa' (*Visual Vocabulary of Independent Standard Words*) dataset which contains collection of isolated set of city names of 10 speakers. The performance of recognition based on visual only features representation of isolated word's results in 82.60%.

Keywords: Lip tracking, Geometric Distance Method, Fourier's Transform, Euclidean Distance

1 Introduction

In the recent year, there are many automatic speech reading system proposed that combine audio as well as visual speech features. In computer speech recognition visual component of

speech is used for support of acoustic speech recognition. Design of an audio-visual speech recognizer is based on human lip-reading expert experiences. Hearing impaired people achieve recognition rate of 60-80% in dependence on lip-reading conditions. Most important conditions for good lip-reading are quality of visual speech of a speaker (proper articulation) and angle of view. Sometimes people, who are well understood from acoustic component may be not well lip-read but for hearing-impaired or even deaf people visual speech component is important source of information. Lip-reading (visual speech recognition) is used by people without disabilities, too. It helps in better understanding when the acoustic speech is less intelligible. Task of automatic speech recognition by a computer, when both acoustic and visual component of speech is used has attracted many researcher to contribute in automatic *Audio-Visual Speech Recognition* domain. This is challenging because of the visual articulations vary with speaker to speaker and can contain very less information as compared to acoustic signal therefore identification of robust features is still center of attraction of many researchers. The visual speech component is used by hearing-impaired people (for lip-reading), but is also used unconsciously by all people in common communication, especially in noisy environment [1].

An Automatic speech recognition (ASR) for well define applications like dictations and medium vocabulary transaction processing tasks are in relatively controlled environments have been designed. It is observed by the researchers that, the ASR performance was far from human performance in variety of tasks and conditions, indeed ASR to date is very sensitive to variations in the environmental channel (non-stationary noise sources such as speech babbled, reverberation in closed spaces such as car, multi-speaker environments) and style of speech (such as whispered etc.)[2]. Lip-reading is an auditory, imagery system as a source of speech and image information. It provides the redundancy with the acoustic speech signal but is less variable than acoustic signals; the acoustic signal depends on lip, teeth, and tongue position to the extent that significant phonetic information was obtainable using lip movement recognition alone [3][4]. The intimate relation between the audio and imagery sensor domains in human recognition can be demonstrated with McGurk Effect [5], [6]; where the perceiver “hears” something other than what was said acoustically due to the influence of conflicting visual stimulus. The current speech recognition technology may perform adequately in the absence of acoustic noise for moderate size vocabularies; but even in the presence of moderate noise it fails except for very small vocabularies [7], [8], [9], [10]. Humans have difficulty distinguishing between some consonants when acoustic signal is degraded. However, to date all automatic speech reading studies have been limited to very small vocabulary tasks and inmost of cases to very small number of speakers. In addition the numbers of diverse algorithms have been suggested in the literature for automatic speech reading and are very difficult to compare, as they are hardly ever tested on any common audio visual databases. Furthermore, most of such databases are very small duration thus placing doubts about generalization of reported results to large population and tasks. There is no specific answer to this but researchers are concentrating more on speaker independent audio-visual large vocabulary continuous speech recognition systems [11].

Many methods have been proposed by researchers in order to enhance speech recognition system by synchronization of visual information with the speech as improvement on automatic Lip-reading system which incorporates dynamic time warping, and vector quantization method applied on alphabets, digits and The recognition was restricted to isolated utterances and was speaker dependent [3]. Later *Christoph Bregler (1993)* had worked on how recognition performance in automated speech perception can be significantly improved & introduced an extension to existing Multi- State Time Delayed Neural Network architecture for handling both the modalities that is acoustics and visual sensor input [12]. Similar work have been done by *Yuhaset.al (1993)* & focused on neural network for vowel recognition and worked on static images [13]. *Paul Duchnowski et.al (1995)* worked on movement invariant automatic Lip-reading and speech recognition [14], *Juergen Luettin (1996)* used active shape model and hidden markov model for visual speech recognition [15],

K.L. Sum *et.al* (2001) proposed a new optimization procedure for extracting the point-based lip contour using active shape model [16], Capiler (2001) used Active shape model and Kalman filtering in spatiotemporal for noting visual deformations [17], Ian Matthews *et.al* (2002) has proposed method for extraction of visual features of Lip-reading for audio-visual speech recognition [18], Xiaopeng Hong *et.al* (2006) used PCA based DCT features Extraction method for Lip-reading [19], The redundancy in the visual cues in audio-visual speech recognition have been examined by Yannawar P L and *et.al* (2010) [20]. Takeshi Saito *et.al* (2008) has analyzed efficient Lip-reading method for various languages where they focused on limited set of words from English, Japanese, Nepalese, Chinese, Mongolian. The words in English and their translated words in above listed languages were considered for the experiment [21]; Meng Li *et.al* (2008) has proposed A Novel Motion Based Lip Feature Extraction for Lip-reading problems [22]. Prashant Borde *et.al* (2014) has Contribution of visual features computed through Zernike moments in association with MFCC [23]. Multi-pose AVSR is getting popularity because of its robustness and many researchers have attracted towards Multi-pose AVSR. Amarsinh Varpe *et.al* (2014) has discussed isolation of Region of Interest for Multi-pose AVSR and it was seen that skin color based detection of ROI was found better as compared with ‘Viola-Jones’ algorithm under multi-pose AVSR scenario [24].

This paper introduced mechanism of extraction of visual features for visual speech recognition. The content of the paper is organized in five section, section II deals with ‘vVISWa’ dataset, Section III deals with methodology adopted, section IV deal with experimental results, section V is conclusion of work followed acknowledgement and references.

2 ‘vVISWa’ Dataset

Many researchers have defined their own dataset and very few are available online freely. Indeed it is very difficult to distribute the data base freely on the web due to the size. The video sequences used for this study was collected in the laboratory in a closed environment. The ‘vVISWa’ (*Visual Vocabulary of Independent Standard Words*) database consists set of independent/isolated standard words from Marathi, Hindi and English script. The dataset of isolated city words like {‘Aurangabad’, ‘Beed’, ‘Hingoli’, ‘Jalgaon’, ‘Kolhapur’, ‘Latur’, ‘Mumbai’, ‘Osmanabad’, ‘Parbhani’, ‘Pune’, ‘Satara’, ‘Solapur’} in Marathi were considered for this experiments. Each visual utterance is recorded for 2 second. The database consists of 10 individuals speakers, 4 male and 6 female and each speaker speaking each word utterance for 10 times. Each individual has uttered word in close-open-close constraint without head movement. The database comprised of 1200 utterance (10*10*12) of these independent standard words. The figure 1 shows the experimental arrangement of acquisition of utterances from individual speaker.

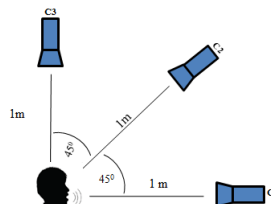


Figure 1 Acquisition of utterances

The visual utterance was recorded at resolution of 720 x 576 in (*.avi) format using high definition digital camera in three angle comprising full frontal using camera ‘C1’, 45°face using camera ‘C2’ and side pose using camera ‘C3’. This study was focused on full frontal profile of utterance. Lighting was controlled and a dark gray background was used. The

recognition of visual utterance of word, the data set of isolated city words was divided in to 70% known sample and 30% unknown sample. The System was trained over 70% known set and 30% unknown samples were tested on the known set and success recognition of isolated word based on visual utterance was evaluated.

3 Methodology

The typical audio visual speech recognition system accepts the audio and visual input as shown in figure 2. The audio input is captured with the help of standard audio mic and visual utterance is captured by using standard camera. The place between camera and individual speaker is kept constant in order to get proper visual utterance. Once the input is acquired, it will be preprocessed for acoustic feature extraction and visual feature extraction separately and further used for recognition and integration of utterance.

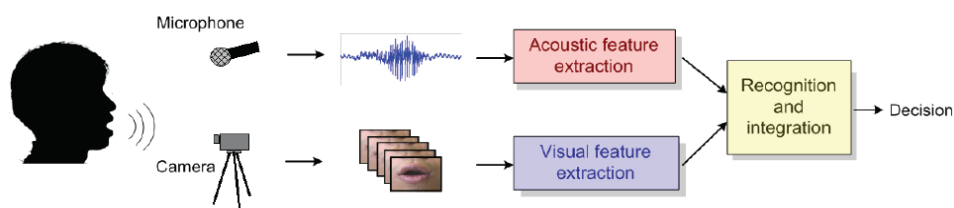


Figure 2 Organization of AVSR

Visual features can be grouped into three general categories: shape-based, appearance-based, and combinational approaches. All three types require the localization and tracking of Region of Interest (ROI). Region of interest for computation of visual feature will be concentrated towards the movement of lips (opening and closing of mouth) over the time frame which is very complex. In-view of this calculating good and discriminatory visual feature of mouth plays vital role in the recognition.

3.1 Region-of-Interest (ROI) Detection / Localization

The visual information relevant to speech is mostly contained in the motion of visible articulators such as lips, tongue and jaw. In order to extract this information from a sequence of video frames it is advantageous to track the complete motion of the face detection and mouth localization, this helps in visual feature extraction. In order to achieve robust and real-time face detection we used the 'Viola-Jones' detector based on 'AdaBoost' which is a binary classifier that uses cascades of weak classifiers to boost its performance [25],[26]. In our study we used detector to detect the face in each frame from the sequence and subsequently mouth portion of the face is detected. This is achieved by finding the median of the coordinates of the ROI object *bounding box* of frames. Finally a region-of-interest (ROI) is extracted by resizing the mouth bounding box to 120x120 pixels size as shown in figure 3.



Figure 3 Mouth Localization using Viola-Jones Algorithm

After isolating ROI Mouth frame, each frame from utterance was pre-processed so as to obtain good discriminating features. The preprocessing include, converting the RGB frame to luminance (Y), hue (I), and saturation (Q) information, filter the saturation (Q) image then calculating gray threshold range and converted frame into binary frame to

get the actual Region-of-Interest (ROI) of containing only the portion covered by lips as shown in figure 4 and in similar way the ROI for each frame of utterance is identified as shown *Table 1*.

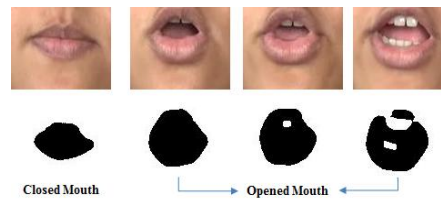


Figure 4 Isolation of ROI Identification

3.2 Visual Feature Extraction using Geometric Distance Method

A feature extraction system intended to be robust to all the sources of variability and must use as much knowledge about the scene as possible. Lip tracking method based on a single cue, about the scene are insufficient for robust and accurate tracking of lips. It was observed that, every speaker has different organization of lip shape therefore, it is necessary to form the outer lip contour so that the personal characteristic of lip shape can be easily measured. In the view of this scenario, we extract lip features points identified as *corner points*, *upper points*, *lower points* and *centre points* of *region-of-interest* (ROI). The marking of these points are as shown in figure 3.4. In order to build the contour the centre point of lips are defined as an intersection of upper and lower lips, as show in *figure 3.4(a)*, and the rest of lip shape eight key points P1 to P14 are isolated and labeled as [P1, P2] are the *corner points*, [P3,P4, P5,P6,P7,P8] are *upper points* of the outer contour, [P9, P10, P11,P12,P13,P14] are the *lower points* of outer contour and [C1,C2,C3,C4,C5,C6] are *centre point* of the lip upper and lower point, as shown in *figure 5.(c)*. To compute the distance between the contour points and center points, these points are joined and the respective distance have been computed for all isolated words as shown in *table 2* and considered as feature of *Region-of-Interest* as shown in *figure 5 (d)* respectively.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

Where, d is Distance, $x1, x2, y1, y2$ are Coordinate Point of lip Contour

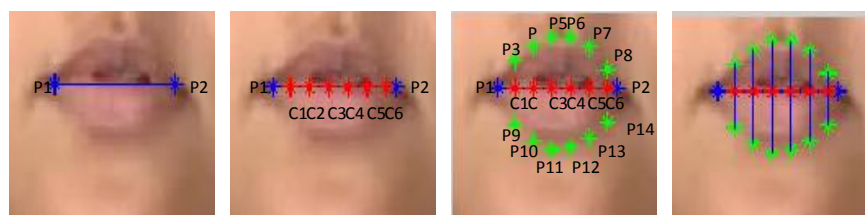


Figure 5 a) Corner Points b) Center Points of Contour c) Contour Points d) Joined Contour Points

Table 2. Geometric Distance features of word "Aurangabad"

Frame No	Point P1 to C1	Point P2 to C2	Point P3 to C3	Point P4 to C4	Point P5 to C5	Point P1 to P2
1	14	21	25	26	22	..	57
2	16	25	30	30	25	..	49
3	12	19	22	22	22	..	54
4	16	22	26	27	24	..	56
5	14	20	25	25	21	..	58

6	13	20	24	23	18	..	55
7	18	27	31	30	25	..	55
8	20	28	32	32	27	..	52
..

3.3 Fourier Feature

The geometric distance feature for each frame of isolated word utterance is represented in 13x1 size, which comprised of 13 features associated with outer contour points of lip. The visual utterance was captured for two seconds at sampling rate of 26fps it results in formation of 52 frames for the word. Therefore, the *Geometric Distance Features* for one visual utterance results in to 676x1 for single word. Fourier Transform was applied to these features for converting all geometric features in to frequency domain.

$$X_k = \sum_{n=0}^{N-1} e^{-2\pi i k - (\frac{n}{N}) x_n} \quad (2)$$

Where $k=0$ to $N-1$, $n= (n_1, \dots, n_d)$, d is dimension of vector.

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{i2\pi k n / N} \quad (3)$$

$\{X_k\}$ is sampled, $\{x_n\}$ must be periodic. A physical point of view, both sequence $\{X_k\}$ and $\{x_n\}$ are repeated indefinitely with period N [27].

After applying *Fourier Transform* on *Geometric Distance Features*, we obtained the Features matrix of size 338x1 for one visual utterance. Similarly all visual words belonging to city names from 'vVISWa' are passed for *Fourier feature* extraction which results in 338x72 size matrix. This feature set was called as 'Training Set'. The 'Fourier Feature Matrix' for the samples of 'Training set', are as shown in table 3 and 'Test set' are as shown in 4.

Table 3. Fourier features for training set

Known Set word No	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	..	Feature 338
1	17755	1096.32	283.677	70.03294	396.7405	..	76.55032
2	17256	1113.11	188.132	541.476	192.1658	..	66.07078
3	16760	1105.52	481.660	309.5822	303.8327	..	60.77191
4	17511	1448.70	529.377	155.3629	293.2735	..	6.040352
5	17542	1356.01	304.978	459.9021	383.7794	..	37.6694
6	15811	1095.39	681.102	904.6851	344.4622	..	16.0917
7	15189	1089.81	827.289	706.0779	426.5965	..	42.36466
8	15054	896.648	672.538	733.8816	200.9687	..	27.83091
..

Table 4. Fourier features for testing set

Unknown Set word No	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	..	Feature 338
1	17041	1853.89	250.747	500.2193	340.7124	..	8.013658
2	17151	1929.75	419.462	372.8522	378.5901	..	24.85433
3	16629	1537.46	728.050	167.7563	343.644	..	89.16924

4	17539	1591.00	8.66994	489.4532	285.6925	..	17.60512
5	17369	1857.23	193.910	389.6271	247.2157	..	10.36455
6	17041	1853.89	250.747	500.2193	340.7124	..	10.32413
7	17151	1929.75	419.462	372.8522	378.5901	..	36.58149
8	16629	1537.46	728.050	167.7563	343.644	..	11.39496
..

4 Experiment and Result

The recognition of visual only sample were carried out by *Geometric Distance Features* and *Fourier Features*. These features were calculated for all sample of training set and stored for recognition purpose. The entire data set of isolated city names (visual only) were divided into 70-30 ratio that is 70% (Training samples) and 30% (Test samples).

4.1 Visual Speech Recognition

Fourier Features for all visual samples from training set and test set were computed. The 'Euclidean' distance classifier was used for measuring the similarity between test samples and training sample. The 'Euclidean' distance provides information between each pair (one vector from test set and other vector from training set) of observations. The *table 5* shows confusion matrix for visual speech recognition based on *Fourier features*.

It was seen that out of 46 samples 38 samples were correctly recognized and 8 samples were misclassified. Samples of {'Hingoli', 'Kolhapur', 'Osmanabad', 'Parbhani' and 'Solapur'} were recognized completely. It was seen that for two samples of 'Aurangabad' one samples were correctly recognized and one sample was mated with 'Kolhapur', this result in 50% of recognition of word. Similar results were observed for isolated word {'Mumbai'}. For isolated word 'Latur' it was seen that out of two samples all samples were matched with word 'Hingoli'.

Table 6 provides the overall recognition rate for recognition of isolated words from 'vISWa' data set. This clearly indicates that the *Geometric Distance Features* with *Fourier Features* will helpful in recognition of words similar to the accuracy of recognition of words by hearing impaired peoples in dependence of lip-reading conditions. This approach will also be helpful in recognition of dictionary based words on visual only features when the acoustic signal is degraded due to noise.

Table 6. Performance of Recognition of System

Method	Result
Visual Only Speech Recognition based on Geometric Distance and Fourier features	82.60 %

5 Conclusion

This paper describes a visual speech recognition method using video without evaluating audio signals which include face and lip detection, features extraction and recognition. These experiment carried out using our own database, which was designed for evaluation purpose. We used a new technique for visual speech recognition that is *Geometric Distance Method* and *Fourier Transform*. The *Fourier Features* are found to be more reliable and accurate feature for visual speech recognition.

Acknowledgement

The Authors gratefully acknowledge support by the Department of Science and Technology (DST) for providing financial assistance for Major Research Project sanctioned under *Fast Track Scheme for Young Scientist*, vide sanction number SERB/1766/2013/14 and the

authorities of Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India, for providing the infrastructure for this research work.

References

1. ŽELEZNÝ, M., KRŇOUL, Z., CÍSAŘ, P., MATOUŠEK, J., "Design, Implementation and Evaluation of the Czech Realistic Audio-Visual Speech Synthesis", *Signal Processing*, vol. 86, no.12, December 2006, Elsevier Science (ISSN 0165-1684).
2. J R Deller, Jr. J G Proakis and J.H L Hansen, "Discrete-Time Processing of Speech Signals", Macmillan Publishing Company, Englewood cliffs, 1993.
3. Eric Petjan, Bradford Bischoff, and David Bodoff, "An Improved automatic Lip-reading system to enhance speech Recognition", Technical Report TM 11251-871012-11, AT&T Bell Labs, Oct. 1987.
4. Finn K.I, "An investigation of visible lip information to be used in automated speech recognition", Ph.D Thesis, George-Town University, 1986.
5. Macdonald J and H MacGurk, "Visual influences on speech perception process", *Perception and Psychophysics*, vol (24)pp 253-257, 1978.
6. MacGurk H and Macdonald J, "Hearing lips and seeing voices, *Nature*", vol (264), pp 746-748, Dec 1976.
7. Paul D.B, Lippmann R.P, Chen Y and Weinstein C.J, "Robust HMM based technique for recognition of speech Produced under stress and in noise", *Proceeding Speech Tech.* vol (87), pp 275-280, 1987.
8. Malkin F.J, "The effect on computer recognition of speech when speaking through protective masks, *proceedingSpeech*" Tech. vol (87), pp 265-268, 1986.
9. Meisel W.S, "A Natural Speech recognition system", *Proceeding Speech Tech.* vol (87), pp 10-13,1987.
10. Moody T, Joost M and Rodman R, "A Comparative Evaluation a speech recognizers", *Proceeding Speech Tech.* vol (87), pp 275-280, 1987.
11. Chalapathy Neti, et.al, "Audio-Visual Speech Recognition", Workshop 2000 Final report, Oct 2000.
12. Christopher Bergler, "Improving connected letter recognition by Lip-reading", IEEE, 1993.
13. B P Yuhas, M H Goldstien and T.J Sejnowski, "Integration of acoustic and visual speech signals using neural Networks", *IEEE Communication Magazine*.
14. Paul Duchnowski, "Toward movement invariant automatic Lip-reading and speech recognition", IEEE, 1995.
15. Juergen Luetin, "Visual Speech recognition using Active Shape Model and Hidden Markov Model", IEEE,1996.
16. K.L Sum, et.al, "A New Optimization procedure for extracting the point based lip contour using Active Shape Model", IEEE, 2001.
17. A Capiler, "Lip detection and tracking", 11th International Conference on Image Analysis and Processing (ICIAP 2001).
18. Ian Matthews, T F Cootes, J A Banbham, S Cox, Richard Harvey, "Extraction of Visual features of Lip- reading", *IEEETransaction on Pattern Analysis and Machine Intelligence*, vol 24, No 2, February 2002.
19. Xiaopeng Hong, et.al, "A PCA based Visual DCT feature extraction method for Lip-reading", International conference on Intelligent Information hiding and multimedia signal Processing, 2006.
20. Yannawar P.L Manza G R, Gawali B W, Mehrotra S C, "Detection of redundant frame in audio visual speech recognition using low level analysis", *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Coimbatore (TN), India 28-29 Dec. 2010, 2010. E-ISBN 978-1-4244-5967-4.

-
21. Takeshi Saitoh, Kazutoshi Morishita and Ryosuke Konishi, "Analysis of efficient Lip-reading method for various languages", 2008.
 22. Meng Li, Yiu-ming Cheung, "A Novel motion based Lip Feature Extraction for Lip-reading", IEEE International conference on Computational Intelligence and Security, pg.no 361-365, 2008.
 23. Borde, Prashant, Amarsinh Varpe, Ramesh Manza, and Pravin Yannawar. "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition." *International Journal of Speech Technology* (2014): 1-9.
 24. Amarsinh Varpe, Prashant Borde, Pallavi Pardeshi, Sadhana Sukale, Pravin Yannawar, "Analysis of Induced Color for Automatic Detection of ROI Multipose AVSR System", Springer International conference on Information System Design and Intelligent Application, 10.1007/978-81-322-2247-7_54, pp 525-538, 2015.
 25. G. Bradski and A. Kaehler. "*Learning OpenCV: Computer Vision with the OpenCV Library*". O'Reilly Media, 1st edition, September 2008.
 26. C. M. Bishop, "*Pattern Recognition and Machine Learning*". Springer, Heidelberg, 2006.
 27. Duhamel, P. and M. Vetterli, "Fast Fourier Transforms: A Tutorial Review and a State of the Art," *Signal Processing*, Vol. 19, April 1990, pp. 259-299

Table 1: Isolated Mouth from video sequence










No. of Frame	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	...	Frame 52
Processed Frame									...	

Table 5. Confusion Matrix for Visual Utterance Recognition

Words to Test	Visual Utterance	Training Sample												Recognition		
		Aurangabad	Beed	Hingoli	Jalgaon	Kolhapur	Latur	Mumbai	Osmanabad	Perbhani	Pune	Satara	Solapur	Recognized	Miss	Accuracy
2	Aurangabad	1	0	0	0	1	0	0	0	0	0	1	0	1	1	50%
3	Beed	0	2	0	0	0	0	1	0	0	0	0	0	2	1	66.66%
3	Hingoli	0	0	3	0	0	0	0	0	0	0	0	0	3	0	100%
5	Jalgaon	0	0	0	4	1	0	0	0	0	0	0	0	4	1	80%
3	Kolhapur	0	0	0	0	3	0	0	0	0	0	0	0	3	0	100%
2	Latur	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0%
2	Mumbai	0	0	0	0	0	0	1	0	1	0	0	0	1	1	50%
5	Osmanabad	0	0	0	0	0	0	0	5	0	0	0	0	5	0	100%
6	Perbhani	0	0	0	0	0	0	0	0	6	0	0	0	6	0	100%
3	Pune	0	0	0	0	0	0	1	0	0	2	0	0	2	1	66.66%
7	Satara	0	0	0	1	0	0	0	0	0	0	6	0	6	1	85.71%
5	Solapur	0	0	0	0	0	0	0	0	0	0	5	5	5	0	100%
Total Sample : 46													38	8	82.06%	
Overall Recognition													82.60%			